Causes and consequences of large-scale copy number alterations in cancer evolution

by
Elle Loughran

A thesis submitted to
The University of Dublin
for the degree of



Doctor of Philosophy

Smurfit Institute of Genetics
Trinity College Dublin, The University of Dublin
January 2025

Supervisor: Aoife McLysaght

Declaration of Authorship

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the university's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and the Trinity College Library conditions of use and acknowledgement.

Elle Reidy

Signature

31 January 2025

Date

Acknowledgements

Firstly, thank you to my supervisor Aoife for convincing me to choose genetics, and for championing me all these years. Thank you to Máire for the incredibly useful insights, enthusiasm and encouragement—working with you has been inspiring. Thank you to Ross and Russell for guidance on all things GWAS—Ross, I think you are an actual genius. Thank you to my placement supervisors and labmates in Montpellier and Paris: Eric, Emmanuel, Andrei, Nicolas, Nikolai. Thank you also to my former supervisor Emilia, and to Shane Bergin for setting this whole thing up in the first place. Thank you to everyone in the McLysaght lab, particularly Alan, Zoe and Anthony for mentoring me back in undergrad, Laetitia (the world's friendliest *parisienne*), Peter for helpful thesis advice, and Róisín for being the platonic ideal of a labmate.

Thank you to Luke, Eoin and Andrew, the Commons crew who became a family—I'll never forget our tea times in House 2. Thanks also to my actual family, especially Dad. To Leon—I can never thank you enough for the kindness you showed me. To old and new friends, in no particular order: Anastasija, Christiana, Ofir, Ben, Ryan, Laoise, Claire, Caroline, Jerry, Catarina, Julian, Geoffrey, Athena, Cristina, Anjan, Ed. To Lion, winter solstice: thank you for the dancing. To the band: Oisín, Manus, Clem (and Trang)—it's an honour to do this with you.

To Dublin University Fencing Club: did you help me focus on my thesis? No. Did you give me a relaxing outlet for the stress? Definitely not. But you did fix my sleep schedule (...and completely change my life). A particular thank you to Colm for the time and wisdom, to my beloved WS and MS teammates for the camaraderie at 7 am training, to Christiana for the solidarity, to Cathal (my captain), and to Donncdha, David, Marco and Sarah for generally being rays of sunshine.

Thank you to the SFI Centre for Research Training in Genomics Data Science (and Aoife) for funding this PhD, to Sandra and all the long-suffering coordinators, and to my cohort-mates for the silliness at conferences. Thank you to the Naughton Foundation, the Laidlaw Foundation, the Foundation Scholarship and SUSI for allowing me to go to college at all. Trinity has been my spiritual home for eight years and my physical one for five—thank you to everyone who keeps this place running, especially the friendly faces at the Dining Hall. Thank you also to the genetics department staff, and especially to Russell, Aoife, Matt, Dan, Frank, Tony and more for the amazing lectures. I feel lucky I got to experience them. Thank you to Frances O'Regan and John Loughran for making me a scientist, and to Alice MacPherson from the Students' Union for convincing the Science Course Office to let me switch to biology in second year.

Finally, to Li—you were the best part of this PhD. Thank you for everything.

One last note to my little brother Jude, who went to school one day ten years ago and told everyone my science fair project had cured cancer: it had not done that, and this one isn't quite there either. Maybe one day.

Table of Contents

Introduction 1
1.1 Types of Copy Number Alterations 2
1.2 Consequences of Copy Number Alterations 4
1.2.1 Copy number alterations and the evolution of novelty 4
1.2.2 Copy number alterations in human disease 5
1.2.3 Dosage-sensitive genes 9
1.2.4 Dosage attenuation and dosage compensation 11
1.2.5 Bridging organismal and cancer evolution 13
1.3 Cancer Evolution 14
1.3.1 Somatic Evolution 14
1.3.2 Oncogenic Transformation 15
1.3.3 Dynamics of Selection in Cancer Evolution 17
1.4 Chromosomal Instability in Cancer 19
1.4.1 Patterns of aneuploidy in cancer 20
1.4.2 Chromosomal instability 21
1.4.3 Impact on tumour fitness 22
1.5 Motivation 24
1.6 Cancer Genomics 26
1.6.1 Development of Cancer Genomics 27
1.6.2 The Cancer Genome Atlas 28
1.6.3 Copy Number Profiling 30
1.6.4 Intra-Tumour Heterogeneity 32
1.7 Aims 34
Origins of extreme ploidy in cancer 35
2.1 Introduction 35
2.2 Methods 40
2.2.1 Identifying Polyploid and Hypodiploid Tumours 40
2.2.2 Germline and Somatic Mutations 43
2.2.3 Environmental Factors 48
2.2.4 Random Forest Model 49
2.3 Results 49
2.3.1 Identifying Polyploid and Hypodiploid Tumours 49
2.3.2 Relationship between rates of polyploidy and hypodiploidy 62
2.3.3 WGD rates across cancer types are partially explained by tissue replication rate 65
2.3.4 No association between germline mutations and extreme ploidy in patient tumours 70

2.3.5 Somatic mutations in TP53, ARID1A and PTEN are associated with ploidy 75
2.3.6 Hypoxia is strongly associated with extreme ploidy across tissues 81
2.3.7 Smoking is associated with hypoxia in lung cancer, but not with ploidy aberrations 86
2.3.8 Mutational signatures associated with smoking are enriched in polyploid tumours 90
2.3.9 Viral infection is nominally associated with polyploidy in a TP53-dependent manner 91
2.3.10 Integrating all factors 95
2.4 Discussion 102
2.4.1 Can the development of extreme tumour ploidy be explained? 102
2.4.2 Are factors identified in in vitro studies relevant within patient tumours? 103
2.4.3 Polyploidy and hypodiploidy: two sides of the same coin? 107
2.4.4 Proliferation 108
2.4.5 Additional factors that may influence rates of extreme ploidy 109
Karyotypic evolution by large-scale chromosome loss in hypodiploid and polyploid tumours 111
3.1 Introduction 111
3.2 Methods 115
3.2.1 Copy Number Profiles 115
3.2.2 Distinguishing Masked Hypodiploids from Hyperdiploids using Cytogenetic Data 116
3.2.3 Whole Genome Duplication Rate 116
3.2.4 Loss after Whole Genome Duplication 117
3.2.5 Chromosome Loss Patterns 118
3.2.6 Factors Associated with Loss Rate 119
3.2.7 Survival Analysis 120
3.3 Results 121
3.3.1 Identifying Masked Hypodiploid ALL from Cytogenetic Data 125
3.3.2 A Generalised Chromosomal Instability Phenotype: CIN at Multiple Scales 134
3.3.3 Loss and Retention after WGD 137
3.3.4 Patterns of Chromosome Loss in Hypodiploid and Polyploid Tumours 145
3.3.5 Prognostic Impact of Ploidy Alterations 158
3.4 Discussion 161
3.4.1 Two Classes of Hypodiploid Tumour 161
3.4.2 Patterns of chromosome loss 162
3.4.3 A generalised chromosomal instability phenotype 163
Haploinsufficiency and dosage responsiveness in tumour suppressor genes 166
4.1 Introduction 166
4.2 Methods 170

4.2.1 Data Sources and Processing 170
4.2.2 Classifying Gene Inactivation Status 170
4.2.3 Functional Readouts 173
4.2.4 Novel Haploinsufficient Tumour Suppressor Genes 175
4.2.4 Regulon Enrichment Analysis 177
4.2.5 Dosage Attenuation 177
4.3 Results 179
4.3.1 TSG Inactivation Events 179
4.3.2 Hallmarks of Cancer 181
4.3.3 One-Hit Tumour Suppressor Genes 184
4.3.4 Quantifying Dosage Responsiveness 191
4.3.5 Candidate Tumour Suppressor Genes 196
4.3.6 Regulon Enrichment Analysis 199
4.3.7 Understanding TSG Dosage Responsiveness using Regulon Enrichment Analysis 203
4.3.8 Attenuation of Dosage Alterations 206
4.3.9 Robustness 213
4.4 Discussion 215
Conclusions 221
Appendix 229
Bibliography 247

List of Figures

2.1	Comparison of tumour ploidy calculated using ASCAT vs ABSOLUTE	50
2.2	Comparison of WGD calls between MEDICC2 and alternatives	57
2.3	Loss of heterozygosity and ploidy distribution for each ploidy class	59
2.4	Ploidy distributions differ by cancer type	60
2.5	Visualisation of a near-haploid and a highly polyploid tumour	62
2.6	Correlation between polyploidy and hypodiploidy rate	64
2.7	Correlation between WGD rate and replication rate	67
2.8	Polyploid tumours have higher proliferative index	69
2.9	GWAS results	72
2.10	Power analysis	73
2.11	Estimates of SNP-based heritability	74
2.12	Total mutation count by ploidy class	77
2.13	Logistic regression of polyploidy on mutation status of cancer genes	79
2.14	Genes associated with both polyploidy and hypodiploidy	80
2.15	Ploidy-aberrant tumours are more hypoxic than near-diploids	83
2.16	Relationship between hypoxia and extreme ploidy across cancer types	85
2.17	Smoking is associated with increased hypoxia in lung adenocarcinoma	87
2.18	Relationship between smoking and copy number instability	89
2.19	Viral infection and ploidy alterations	92
2.20	The viral infection-polyploidy relationship depends on TP53 mutation	94
2.21	Performance of a random forest model in predicting polyploidy	97
2.22	Logistic regression of polyploidy on all predictors	100
2.23	Logistic regression of hypodiploidy on all predictors	101
3.1	Distribution of autosome counts among hypodiploid tumours	123

3.2	Visualisation of genome doubling in hypodiploid ALL	126
3.3	Rationale for the MH score heuristic	128
3.4	Confusion matrices for the MH score heuristic	129
3.5	Overlap of MH scores within the TCGA dataset	130
3.6	Ploidy and loss of heterozygosity in the TARGET-ALL-P2 study	131
3.7	Visualisation of genome doubling patterns in false negative cases	133
3.8	Rate of WGD by hypodiploidy class	134
3.9	Chromosomal instability at multiple scales	136
3.10	Loss of genomic material after WGD	138
3.11	Gene retention in post-WGD samples	140
3.12	Change in gene copy number proportions over time	141
3.13	Relationship between WGD rate and retention rate post-WGD	142
3.14	Gene copy number proportions over time	144
3.15	Chromosome arm loss rates in hypodiploid tumours	147
3.16	Correlations of chromosome arm loss rates between ploidy classes	148
3.17	Chromosome arm loss rates pre- vs post-WGD	150
3.18	Clustering of hypodiploid tumours based on chromosome losses	153
3.19	Linear regression of loss rate on chromosome features	157
3.20	Survival curves by ploidy class	160
4.1	Classification scheme for TSG inactivation status	172
4.2	TSG inactivation status	180
4.3	Hallmarks of cancer	183
4.4	Significant one-hit TSGs	186
4.5	One-hit TSGs associated with both survival and a hallmark	191
4.6	Dosage responsiveness ratios for significant one-hit TSGs	193
4.7	Ratio of hallmark levels for one- vs two-copy TSGs	194

4.8	Mixed evidence for CDKN2A as a haploinsufficient TSG	196
4.9	STOP genes as candidate TSGs	198
4.10	Correlations between activity scores from different regulons	200
4.11	RNA, protein and activity level correlations using different regulons	202
4.12	Relationship between TSG activity score and hallmark level	203
4.13	Effect of partial TSG inactivation on activity score	205
4.14	Attenuation of copy number alterations	208
4.15	Linear regression of attenuation coefficients on gene features	211
4.16	Dosage attenuation by hit status	212
6.1	Features of samples with discordant WGD calls	231
6.2	Tissue-specific comparison of PI between polyploids and diploids	233
6.3	Tissue-specific comparison of PI between hypodiploids and diploids	234
6.4	PCA to identify ancestry covariates for the GWAS	235
6.5	Tissue-specific mutation count difference in hypodiploids	236
6.6	Performance of random forest classifier using sample PI	237
6.7	Post-WGD gene losses in cancer vs human germline evolution	239
6.8	Relationship between loss rate and chromosome features	241
6.9	Copy number alteration rate of dosage-sensitive genes	242
6.10	Relevant driver TSGs in each cancer type	243
6.11	Comparison of activity scores using different inference methods	244
6.12	Similarity of DNA-RNA and RNA-Protein correlations across tissues	244
6.13	DNA-RNA and RNA-Protein correlations by hit status	245
6.14	Effect size ratios for partial and complete inactivations (regression)	245

List of Tables

1.1 Comparison of nomenclature in the organismal biology and cancer genomics literatures 2

2.1 Methods used to call WGD events in cancer genomics 53

2.2 Design of the GWAS 71

4.1 Novel one-hit TSGs identified by this screen 188

6.1 Cancer type abbreviations 229

Abbreviations

ALL	Acute Lymphoblastic Leukaemia
bp	base pair
CIN	Chromosomal instability
CMV	Cytomegalovirus
CNA	Copy Number Alteration
CN	Copy Number
CNH	Copy Number Heterogeneity
CNLOH	Copy-number Neutral Loss of Heterozygosity
CNV	Copy Number Variant
CPM	Counts Per Million
DR	Dosage Responsiveness
EBV	Epstein-Barr Virus
GDC	Genomic Data Commons
GWAS	Genome-Wide Association Study
HBV	Hepatitis B Virus
HeH-ALL	High-Hyperdiploid ALL
HPV	Human Papillomavirus
LDSC	Linkage Disequilibrium Score Regression
LH-ALL	Low-Hypodiploid ALL
LOH	Loss of Heterozygosity
MCN	Major Copy Number

mCN	Minor Copy Number
MH	Masked Hypodiploidy
NGS	Next Generation Sequencing
NH-ALL	Near-Haploid ALL
OG	Oncogene
PI	Proliferative Index
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
TCGA	The Cancer Genome Atlas
TPM	Transcripts Per Million
TSG	Tumour Suppressor Gene
VAF	Variant Allele Frequency
WGD	Whole Genome Duplication

Summary

Cancer genomes are characterised by genomic instability and frequent large-scale copy number alterations across the genome. The aim of this thesis was to investigate the consequences of these alterations from an evolutionary perspective, with a focus on the development of extreme ploidy and the impact of dosage sensitivity in cancer.

In chapter 2, I investigated genetic and environmental factors involved in ploidy alterations. While several factors are known to cause polyploidy *in vitro*, their relevance within patient cohorts is not clear. By analysing polyploid and hypodiploid patients across The Cancer Genome Atlas, I identified replication rate, TP53 mutations, hypoxia and HBV infection as significant predictors of polyploidy in patient tumours. Results from a random forest classifier based on these predictors support the idea that the rate of polyploidy in primary tumours is largely determined by the rate of occurrence of stochastic mitotic errors in cells with a permissive G1 checkpoint. TP53 mutations, replication rate and hypoxia were also significant predictors of hypodiploidy, indicating shared underlying factors between these two opposite ploidy states.

In chapter 3, I analysed patterns of chromosome loss in the establishment of hypodiploidy and in the evolution of post-WGD tumours. Multiple lines of evidence suggest two groupings of cancer types with frequent hypodiploidy: one where hypodiploids form a distinct subgroup with stereotyped chromosome loss patterns, and one in which chromosome losses are largely random and hypodiploid cases merely represent one end of a spectrum of chromosomal instability. Loss patterns are generally not explained by chromosome features, and in particular there is no evidence that hypodiploid tumours evolve to avoid monosomy of germline dosage-sensitive genes. I then explored the commonalities between extreme ploidy states further, finding evidence for a general chromosomal instability phenotype that contributes to both increases and decreases in ploidy and to both large-scale and small-scale copy number alterations. Additionally, I formalised and tested a clinical heuristic to identify a diagnostically-challenging subset of acute lymphoblastic leukaemias based on karyotype alone, which allowed me to calculate its prevalence and to rigorously estimate the rate of genome doubling in hypodiploid leukaemia.

In chapter 4, I performed a functional screen to identify haploinsufficient tumour suppressor genes based on observable changes in cell behaviour upon monoallelic TSG inactivation. This screen allowed me to identify novel haploinsufficient TSGs and to quantify TSG haploinsufficiency across tissues and functions. Our results support a model of continuous dosage-responsiveness, wherein TSGs are not cleanly divided into 'two-hit' vs 'haploinsufficient' classes but instead influence tumorigenesis depending on

both their degree of inactivation and their importance in regulating cell behaviour.

The results of this thesis emphasise the weakness of negative selection and the unifying role of chromosomal instability in cancer copy number evolution.

Chapter 1

Introduction

Cancer evolution is marked by the loss of constraints that govern normal cells and—in many cases—the development of extreme genomic instability and rapid accumulation of large-scale copy number alterations across the genome. These alterations reshape gene expression and cell behaviour, boost tumour evolvability, and pave the way for therapy resistance and relapse. Much recent work in the field of cancer genomics, including this thesis, aims to untangle the complex relationship between copy number instability and cancer evolution, with the aim of understanding, predicting and ultimately altering tumour progression.

1.1 Types of Copy Number Alterations

In this chapter, I will review current knowledge of the consequences of large-scale copy number alterations (CNAs) with reference to results and frameworks from both organismal biology and cancer genomics research. The nomenclature used to describe this class of alterations differs between the two fields; a list of correspondences between terms is presented in Table 1.1.

Description	Germline	Cancer
Change in the number of DNA copies of a genomic region	Copy number variant Duplication/deletion	Copy number alteration Gain/loss
Increase in copy number of a small genomic region (often a single gene)[1]	Small-scale duplication	Focal gain Focal amplification (when the number of copies is greatly increased, e.g. >8 copies)[2]
Change in copy number of a portion of a chromosome	Segmental duplication/deletion Microduplication/microdeletion	Segmental gain/loss
Gain or loss of an entire chromosome or chromosome arm	(Arm-level) aneuploidy	(Arm-level) aneuploidy
A region in two copies but with loss of one parental allele	Uniparental disomy	Copy-neutral loss of heterozygosity (CNLOH)
Duplication of the entire genome	Whole genome duplication, WGD	Whole genome duplication Genome doubling

Table 1.1 Comparison of nomenclature describing copy number changes in the organismal biology and cancer genomics literatures. This thesis generally aligns with the naming conventions of the cancer genomics literature.

[1] In the organismal evolution literature, a small-scale duplication may be operationalised as any duplicate that did not arise from a whole-genome duplication (Hakes et al, 2007; Fares et al, 2013)

[2] Threshold taken from Santarius et al (2010).

The smallest copy number alterations are 1000 bp long. Duplications or deletions of genomic material below this size are generally defined as indels (Tang & Amon, 2013; Murakami et al, 2020), although some authors include alterations as short as 50 bp (Alkan et al, 2011). The majority of copy number alterations in the human genome arise through one of three mechanisms: (i) non-allelic homologous recombination (ii) non-homologous end joining (iii) fork stalling and template switching (FoSTeS). These mechanisms all produce segmental duplications and deletions, but only the first creates recurrent alterations at the same locus (Gu et al, 2008). Non-homologous end joining and FoSTeS occur in response to double-strand and single-strand DNA damage, respectively (reviewed in Burssed et al, 2022). Whole-chromosome aneuploidy generally originates from improper chromosome segregation at meiosis (in the germline) or mitosis (in somatic cells); this can happen sequentially or in a concerted manner via multipolar mitosis (Woodward et al, 2023). Arm-level aneuploidy is common in cancer and can arise through several mechanisms, including dicentric chromosomes (Barra & Fachinetti, 2018), cohesion fatigue leading to mistimed separation of sister chromatids (Sapkota et al, 2018) and chromothripsis (Shorokhova et al, 2021).

Whole genome duplication (WGD) describes the simultaneous doubling of the entire chromosome complement, taking a cell from diploid to tetraploid or haploid to diploid in one concerted event. WGD typically occurs due to a failure of cell division such as mitotic slippage or cytokinesis failure (Lau & Poon, 2023), or due to repeated DNA replication without cell division (endoreduplication, Shu et

al, 2018). Cells may also become polyploid through cell fusion, which is an important mechanism of programmed polyploidy in certain normal somatic cells (Shabo et al, 2020). WGD events and polyploidy are particularly frequent in cancer due to (i) inactivating mutations in genome integrity genes such as TP53, which normally block proliferation of DNA-damaged cells (Marques & Kops, 2023) (ii) activating mutations in cell division-promoting genes such as cyclin E1 (Zeng et al, 2023) (iii) telomere dysfunction, which induces mitotic bypass (Davoli et al, 2010); and (iv) the action of cancer-associated viruses, which can both induce cell fusion and inactivate regulators of genome stability (Herbein & Nehme, 2020). Some rare tumours undergo the opposite ploidy alteration and become severely hypodiploid or near-haploid. The mechanism by which this occurs is not clear; it has been proposed that hypodiploidy results from sequential chromosome losses but near-haploidy is caused by sudden genome haploidisation, analogous to an inverted WGD (Walen, 2014).

1.2 Consequences of Copy Number Alterations

In this section, I discuss the consequences of copy number alterations at different scales based on insights gained from the study of organismal biology, genome evolution and human disease.

1.2.1 Copy number alterations and the evolution of novelty

Both copy number gains and copy number losses have been linked to adaptation across species, from amplification of detoxification enzymes for pesticide resistance in mosquitoes (Faucon et al, 2015) to metabolic adaptation in fruit

bats by loss of renal transporter genes (Sharma et al, 2018). Gene duplication in particular has long been heralded as an engine of evolutionary innovation (Ohno, 1970), creating the majority of new genes and facilitating a relaxed selective environment for the evolution of novel functions (neofunctionalisation) and the division of old ones (subfunctionalisation) (reviewed in Birchler & Yang, 2022). WGD is an extreme case, duplicating entire networks of interacting genes at once, and has been linked to adaptation to environmental change (Ebadi et al, 2023). After WGD, the genome undergoes a long process of rediploidisation, with massive loss of duplicate genes over tens of millions of years (Inoue et al, 2015). In the human lineage, duplicate genes (ohnologs) retained since the two WGD events at the base of vertebrates are enriched for roles in development, signalling and the regulation of transcription (Makino & McLysaght, 2010). Large-scale loss of non-duplicated genomic material is rarer, with notable exceptions including the evolution of parasites (Sun et al, 2018) and symbionts (Yin et al, 2024).

1.2.2 Copy number alterations in human disease

Copy number variation is an important cause of genetic disease in humans. At the single-gene level, examples include Charcot-Marie-Tooth disease, a muscle weakness disorder caused by duplication and overexpression of PMP22 (Boutary et al, 2021), heterozygous loss of MKRN3, which causes precocious puberty (Meader et al, 2020), and MECP2 duplication syndrome, which causes intellectual disability and severely reduced lifespan in males. Notably,

heterozygous deletions in the MECP2 gene result in another neurodevelopmental disorder causing intellectual disability, Rett syndrome (Ta et al, 2022).

Focusing on longer alterations, segmental copy number alterations and microdeletions/duplications tend to encompass multiple genes and thus cause syndromes with complex symptoms, frequently leading to intellectual disability (Nevado et al, 2014). They have also been implicated as causal factors in more common neurodevelopmental conditions such as autism and schizophrenia (Abedini et al, 2024; Rylaarsdam & Guemes-Gamboa, 2019; Kato et al, 2022), and as risk factors for ovarian cancer and ischaemic heart disease (Auwerx et al, 2024). Because they tend to arise in unstable hotspots of non-allelic homologous recombination, they are often recurrent and undergo both deletions and duplications (Gu et al, 2008). The reciprocal alteration to a known deletion/duplication syndrome sometimes causes another syndrome with opposite symptoms, suggesting a dosage effect. For example, Sotos syndrome, the symptoms of which include macrocephaly, childhood overgrowth and advanced bone age, is caused by a deletion on chromosome 5q; the reciprocal duplication causes a different syndrome featuring microcephaly, short stature and delayed bone age (Zilina et al, 2013). In other cases, deviations from disomy in either direction can result in similar syndromes: both deletions and duplications on chromosome 22q11.2 cause heart defects, velopharyngeal insufficiency and urogenital abnormalities (Leite et al, 2016). Similarly, both duplications and deletions that encompass the developmental regulator MBD5 on chromosome

2q23.1 are linked to autistic-like symptoms with motor delay and language impairment (Mullegama et al, 2014).

Uniparental disomy[3], in which both copies of a chromosome or chromosome segment come from one parent, can also result in disease if it affects an imprinted region. Genomic imprinting is an epigenetic silencing mechanism that results in parent of origin-specific expression of certain chromosome segments, so that a single-copy loss can result in total abolition of gene expression. Imprinting disorders can also show reciprocal patterns, as with chromosome 15q11-q13: loss of the maternal allele causes Angelman Syndrome, but loss of the paternal allele causes Prader-Willi syndrome (Eggermann et al, 2015). Uniparental disomy usually results from the 'rescuing' of a monosomy or trisomy during meiosis by chromosome duplication or loss, respectively (Shaffer et al, 2001).

Aneuploidy is severely pathogenic, accounting for 35% of spontaneous abortions and 4% of stillbirths, excluding those which arrest before implantation (Pacchierotti et al, 2007). Three aneuploidies are relatively mild: trisomy 21, which causes Down Syndrome, and the aneuploidies of the sex chromosomes, which cause disorders of sexual development. Two others (trisomy 13, Patau syndrome, and trisomy 18, Edwards syndrome) are sometimes compatible with survival to birth but have a life expectancy of approximately one week (Meyer et al, 2015). All other trisomies and every autosomal monosomy are non-viable. However, arm-length monosomies can be viable, including the 5p deletion that

[3] In the cancer context, this is called 'copy-neutral loss of heterozygosity'.

causes cri-du-chat syndrome (Nguyen et al, 2016), and mosaic cases (where only some cells are aneuploid) are generally milder. Indeed, a recent study using single-cell sequencing to detect minority cell populations found that the majority of good-quality IVF blastocysts have low-level mosaic aneuploidy (Chavli et al, 2024). -

Tetraploidy has also been reported in human embryos, making up ~7% of spontaneous abortions (Jia et al, 2015); survival to birth is rare, but the phenotype is variable and one affected child is reported to have lived more than four years (Okamura et al, 2024). Haploidy has not been reported and is likely non-viable due to genome-wide haploinsufficiency and disruption of imprinting, among other factors. However, haploid human embryonic stem cells have been developed and show significantly reduced cell size as well as relative increases in mitochondrial abundance and X chromosome expression (Sagi et al, 2016). Other ploidy abnormalities include androgenetic diploidy, which results in a non-viable hydatidiform mole (Lindor et al, 1992), and triploidy, which is generally non-viable but can occasionally lead to a live birth, including one case who survived for 10.5 months (Sherard et al, 1986).

While polyploidy causes a larger absolute change in genomic material, both clinical data and experiments in multiple species have shown that aneuploidy is significantly more deleterious, likely due to the gene dosage imbalances it creates (Torres et al, 2008). These dosage imbalances are expected to impair the functioning of a subset of 'dosage-sensitive' genes on the aneuploid

chromosome(s), and much effort has been devoted to identifying 'critical regions' on chromosome 21 that contain genes responsible for the Down Syndrome phenotype (Pelleri et al, 2019).

1.2.3 Dosage-sensitive genes

Dosage-sensitive genes[4] are defined as genes for which a change in copy number results in a phenotype, usually a deleterious one (Rice & McLysaght, 2017). A number of possible mechanisms have been put forward to explain why some genes are dosage-sensitive, including absolute gene dosage effects (e.g. developmental morphogens), haploinsufficiency (when a single copy is not enough to maintain gene function), accumulation of toxic protein aggregates (Bolognesi et al, 2016), off-target interactions between promiscuous intrinsically disordered proteins (Vavouri et al, 2009), and dosage balance constraint. Dosage balance constraint refers to pressure to maintain stoichiometry between interacting proteins, particularly in protein complexes, because deviations can slow assembly and lead to an increase in malformed complexes (Veitia, 2004; Holland & Johnson, 2018). Similarly, members of metabolic pathways may be dosage-balanced to prevent accumulation of disruptive byproducts. A striking example of a dosage-balanced gene pair is α- and ß-tubulin in budding yeast: one extra copy of ß-tubulin is lethal unless α-tubulin is also duplicated (Katz et al, 1990). The dosage balance hypothesis has also been employed to explain patterns in retention of duplicate genes created by WGD: genes which are retained in

[4] The terms 'dosage sensitivity' and 'dosage-sensitive gene' are sometimes used in the cancer genomics literature to refer to genes for which expression level follows copy number (Chang et al, 2020; Fehrmann et al, 2015). I instead refer to these genes as 'dosage-responsive'.

duplicate after the entire genome is duplicated together also tend not to undergo individual small-scale duplications, suggesting that their copy number can either change with the rest of the genome or not at all (Makino & McLysaght, 2010).

In general, dosage-sensitive or dosage-constrained genes are expected to have low copy number variability throughout evolution and across the human population, although the exact pattern may depend on their mode of dosage sensitivity. For example, some haploinsufficient genes may be able to undergo duplication without negative consequences and aggregation-prone genes may be more free to undergo deletion than duplication, whereas a dosage-balanced gene should avoid any copy number variation relative to the rest of the genome. This framework has been used to identify thousands of dosage-sensitive genes in the human genome based on the absence of loss-of-function mutations (Lek et al, 2016), conservation of gene copy number across mammalian species (Rice & McLysaght, 2017), or avoidance of small-scale duplications combined with preferential retention after WGD (Makino & McLysaght, 2010). These genes are frequently associated with disease, and those with no known disease associations are expected to be enriched for embryonic lethal mutant phenotypes (Minikel et al, 2020). Dosage-sensitive genes can also be identified directly using knockdown screens (Sarel-Galilly et al, 2022), induced aneuploidies (Ermakova et al, 2011) or overexpression (Makanae et al, 2013). The last of these studies showed that the large majority (80%) of protein-coding genes in yeast tolerated major increases in copy number (nearly 100-fold), and that dosage-sensitive genes were highly expressed and enriched in protein complexes, findings that have been

corroborated in humans using computational methods (Ni et al, 2019; Schuster-Böckler et al, 2010). These experimental studies are also useful for quantifying levels of dosage-sensitivity beyond alterations seen in nature: some of the dosage-sensitive genes identified by Makanae et al. could tolerate dosage increases of 5-10x, so a trisomy (1.5x) might not cause a phenotype.

1.2.4 Dosage attenuation and dosage compensation

The impact of copy number alterations is attenuated by the fact that gene expression does not always reflect gene copy number. The vast majority of genes have a non-zero correlation between DNA copy number and RNA expression and between RNA expression and protein abundance. However, these correlations are generally weak to moderate, with average $r \sim 0.3\text{-}0.4$ for DNA-RNA (Cheng et al, 2022) and 0.4-0.6 for RNA-protein correlations (Arad & Geiger, 2023). Additionally, genes with strong DNA-RNA correlations tend to have weaker RNA-protein correlations (Cheng et al, 2022). These correlations are also weaker in cancer than in normal tissues (Kosti et al, 2016), particularly in highly aneuploid tumours, and targeted uncoupling of expression from copy number has been suggested as an adaptive strategy in this context because large alterations can simultaneously amplify both tumour-suppressive and tumour-promoting genes (Mohanty et al, 2021). The relationship between DNA copy number and protein expression varies depending on copy number alteration length: focal amplifications in certain oncogenes (e.g. MYC) result in significant increases in protein abundance, whereas chromosome arm-level aneuploidies do not (Schukken & Sheltzer, 2022). This could occur if long copy number gains result

in the co-amplification of a gene and its transcriptional or translational repressor, or in increased competition for transcriptional or translational machinery.

A certain degree of dosage attenuation is inevitable due to noise in the process of gene expression. However, multiple mechanisms of targeted dosage compensation—regulation of gene expression to buffer copy number alterations—have also evolved in different biological contexts. Dosage compensation is most famous as a mechanism to equalise expression between the sex chromosomes and the autosomes in the heterogametic sex. In mammals, this is theorised to have evolved by a doubling of expression from the X chromosome, followed by near-total inactivation of one[5] of the X chromosomes in females (Ohno, 1966). This mechanism does not extend to the autosomal aneuploidies, with no evidence of dosage compensation in trisomy 21 (Hunter et al, 2023), 13 or 18 (Hwang et al, 2021). Other lineages have developed alternate means of dosage compensation: in chicken, for example, the Z chromosome does not show whole-chromosome dosage compensation but specific dosage-sensitive genes are compensated on a gene-by-gene basis (Zimmer et al, 2016).

Studies in yeast have suggested that most genes are not dosage-compensated (Torres et al, 2016). However, gene-specific targeted dosage compensation is possible given a sensor mechanism to detect and remedy copy number

[5] More precisely, $n - 1$, where n is the total number of X chromosomes in the cell. This is relevant in the case of sex chromosome aneuploidies, e.g. triple X syndrome (XXX) and Klinefelter's Syndrome (XXY). X-chromosome inactivation is not complete, with at least 12% of genes on the inactive X chromosome retaining expression (Balaton et al, 2015).

alterations (Bravo-Estupiñan et al, 2023). In the case of genes that directly or indirectly regulate their own expression, this can occur through negative feedback loops (Yang et al, 2021; Bader et al, 2015). Another kind of sensor is the stoichiometry between members of protein complexes: members produced in excess cannot join the complex and are rapidly degraded (Ishikawa et al, 2020; Sung et al, 2016), resulting in lower RNA-protein correlations for these genes (Cheng et al, 2022). Targeted dosage compensation may also occur through altered methylation of aneuploid regions (Batra et al, 2021).

1.2.5 Bridging organismal and cancer evolution

With the growth of bioinformatics and the development of the cancer genomics field, there has been increased interaction between molecular evolutionists and cancer biologists. This has resulted in the use of methods from organismal evolution to study the development and progression of cancer in new ways. For example, the dN/dS ratio, which measures positive and negative selection based on rates of nonsynonymous and synonymous mutations, has been adapted to characterise selective dynamics in cancer genomes (Martincorena et al, 2017); a phylogeographic approach has been applied to understand intra-tumour heterogeneity in liver cancer (Liu et al, 2024); and phylogenetic methods have been used extensively to trace metastatic seeding (Brown et al, 2017; Chroni et al, 2022). Some of the conceptual frameworks developed for copy number evolution in organisms may also be useful in the cancer context. However, their applicability is unclear because cancer evolution differs from germline evolution in a number of critical aspects.

1.3 Cancer Evolution

1.3.1 Somatic Evolution

The selective landscape of somatic cells differs dramatically from that of the germline. Only about 40-50% of human genes are ubiquitously expressed; these represent housekeeping genes required for the basic functioning of all cell types (Gu et al, 2023). Of the remaining genes, a somatic cell in a given tissue can dispense with any that are not related to its specific function. In line with this, negative selection is much weaker in somatic tissues (Martincorena et al, 2017), and mutation rates in somatic cells are two orders of magnitude higher than in the germline (Milholland et al, 2017).

Each tissue is thus a mosaic of slightly different genomes, with the frequency of individual mutations depending on the timing of the mutation (mutations earlier in development will be found in a larger fraction of cells) and on the variant's effect: mutations that increase proliferation result in clonal expansions. These mutations are often in cancer driver genes; it has been shown that samples of healthy oesophagus have mutations in NOTCH1 in 30% of cells and of TP53 in 5-20% of cells (Martincorena et al, 2018). While these gene mutations drive clonal expansions in healthy tissues just like in cancer cells, the clones in the former are smaller and restricted by tissue structure (Martincorena, 2019). Copy number alterations are thought to be much rarer than mutations in healthy tissues, excluding the programmed polyploidisation mechanisms known in hepatocytes and megakaryocytes, among others (Darmasaputra et al, 2024).

However, aneuploid subpopulations have recently been reported in healthy lymphocytes (Liu et al, 2022) and in healthy, morphologically normal breast tissues (~3% of cells), with clonal expansions of aneuploidies that are common in breast cancer (Lin et al, 2024).

1.3.2 Oncogenic Transformation

Cancer cells diverge further from the germline by the process of oncogenic transformation, adopting a set of abnormal characteristics that have been termed the 'hallmarks of cancer'. These include self-sufficiency in growth signals, insensitivity to growth-suppressive and apoptotic signals, unlimited replication with loss of contact inhibition, increased angiogenesis, and tissue invasion and metastasis (Hanahan & Weinberg, 2000). In a sense, cancer cells stop acting as if they are part of a multicellular organism: they neglect both chemical and mechanical signals from neighbouring cells, shift towards a transcriptional state reminiscent of unicellular organisms (Trigos et al, 2017; Chen et al, 2015), and may undergo dedifferentiation, losing tissue-specific markers and functions (dos Santos et al, 2023). The hallmarks of cancer were later updated to include evasion of immunosurveillance and reprogramming of energy metabolism, and to emphasise the underlying hallmark that enables the others: genomic instability and mutation (Hanahan & Weinberg, 2011).

Oncogenic transformation is thought to proceed by driver alterations in cancer genes, with more than seven hundred of these genes having been discovered to date (Sondka et al, 2018). Cancer driver genes typically have functions in

regulating cell proliferation, but they also influence cell death and senescence, genome stability, mutation rate and immune presentation, among other traits (Shortt & Johnstone, 2012; Roetman et al, 2022). They are traditionally divided into oncogenes, which positively regulate cell proliferation and promote oncogenic features, and tumour suppressor genes (TSGs), which function to restrain cell division and protect genome integrity. Driver mutations in oncogenes are activating and tend to involve clustered missense mutations in relevant domains or copy number amplifications; driver mutations in tumour suppressor genes are inactivating and typically involve nonsense or deleterious missense mutations and/or copy number losses (Davoli et al, 2013). Mutations in TSGs, unlike in oncogenes, have historically been considered to be recessive on the cellular level, requiring two 'hits' to inactivate the TSG and induce tumorigenesis (Knudson, 1971). This has been used to explain the highly elevated cancer risk in individuals who inherit heterozygous mutations in TSGs and the extreme rarity of germline mutations in oncogenes. However, there is now significant evidence for haploinsufficient or 'one-hit' behaviour across a wide range of TSGs (reviewed in Berger et al, 2011), including TP53 and PTEN. As well as copy number alterations and mutations, driver alterations can involve epigenetic mechanisms such as methylation of TSG promoters (Chatterjee et al, 2018) and loss of imprinting of oncogenes such as IGF2 (Damaschke et al, 2017).

Tumours frequently carry alterations in several different driver genes, and one seminal model of cancer initiation proposes that colorectal cancer develops from normal colonic epithelium in a multi-stage process involving sequential

alterations of APC (TSG), K-RAS (oncogene) and TP53 (TSG) (Fearon & Vogelstein, 1990). The effects of these genes extend beyond tumour initiation to drive continued evolution of cancer-related traits, and indeed some drivers with roles in immune evasion or metastasis may only become relevant in the later stages of cancer progression (Pineda & Bradley, 2024). These post-initiation effects contribute to differences in survival between cancer patients with alterations of different driver genes (Oh et al, 2019; Molina-Vila et al, 2014; Serebriiskii et al, 2023). Cancer driver genes also vary in strength and tissue specificity. Some drivers, such as TP53, CDKN2A, PIK3CA and MYC, have strong and near-universal effects on tumorigenesis; many others, such as BRCA1, BRCA2, VHL and APC, are strong drivers but only in specific cancer types (Schaefer & Serrano, 2016); and still others are only weakly linked to tumorigenesis (Zhang et al, 2021). Some drivers, such as EZH2, can act as oncogenes in some cancer types and as tumour suppressors in others (Wassef & Margueron, 2017). Tissue specificity is not generally explained by tissue-specific driver expression but may depend on the expression of other genes or external factors that can compensate for driver alterations (Schaefer & Serrano, 2016; Bianchi et al, 2020).

1.3.3 Dynamics of Selection in Cancer Evolution

The dominant model of cancer evolution is of a Darwinian process starting from a single cell and progressing through genetic diversification and selection of increasingly aggressive subclones (Nowell, 1976). Multiple models of clonal evolution have been proposed. These include *linear evolution*, where each

subsequent driver mutation induces a selective sweep and fully outcompetes all other subclones; *branching evolution*, in which multiple subclones expand in parallel, and may compete or cooperate with each other; *neutral evolution*, in which there is no selection and mutations accumulate neutrally over time, with frequency predicted by their age; and *punctuated evolution*, in which there is a flood of alterations at the beginning of the tumour's evolution that later gives way to stable clonal expansions. Given the ubiquity of both positive selection and intra-tumour heterogeneity in cancer, the branching model finds most support, but each model may be relevant in specific contexts, and—most relevant to this thesis—copy number evolution is often well-described by the punctuated model (Davis et al, 2017). Other models have explained heterogeneity between cancer cells through non-heritable means including cancer stem cells and phenotypic plasticity (Marusyk & Polyak, 2010).

In general, cancer evolution is characterised by positive selection on driver alterations and neutral evolution of the vast majority of other alterations. Purifying selection against mutations appears to be almost entirely absent, with exome-wide normalised dN/dS ratios suggesting that almost all coding mutations are tolerated, except those in hemizygous essential genes (Martincorena et al, 2017; Van den Eynden et al, 2016). Other work has found evidence of negative selection in specific cases, e.g. genes that might provoke an immune response (Zapata et al, 2018). There is also evidence of selection against homozygous deletions near essential genes (Pertesi et al, 2019), but the dynamics of selection on other copy number alterations are less clear. The extent of purifying selection

depends on the overall level of alterations in the genome: Tilk et al (2022) show that low-mutational burden tumours have much stronger evidence of negative selection on both mutations and copy number alterations than high-burden tumours (dN/dS ~ 0.56 vs ~1). They conclude that many of the mutations in cancer are in fact deleterious but cannot be efficiently removed by purifying selection due to Hill-Robertson interference caused by genome-wide linkage. Independent evidence from experimental evolution of transformed cells corroborates that individual copy number alterations (Zhang et al, 2019) and mutations (McFarland et al, 2017) are often strongly deleterious yet accumulate by Muller's ratchet regardless.

1.4 Chromosomal Instability in Cancer

High-burden tumours are the rule, not the exception, with alterations in guardians of genome stability and inefficient negative selection resulting in highly elevated mutation rates and widespread copy number alterations. The vast majority of solid tumours have at least one aneuploid chromosome (Taylor et al, 2018), and ~30% undergo WGD, making this the second most common alteration in cancer after TP53 mutation. In the MSK-IMPACT cohort of advanced cancers, WGD is associated with mutations in TP53 and in genes involved in the E2F-mediated G1 arrest pathway, and occurs more frequently in tissues with higher proliferation rates (Bielski et al, 2018). Tumours also exhibit loss of heterozygosity across large sections of the genome, occasionally reaching severely hypodiploid or near-haploid states (Gopal et al, 2018). The permissibility of trisomies and monosomies across the genome stands in stark

contrast to their catastrophic effects in the human germline. As well as numerical chromosome alterations, tumours frequently carry complex genomic rearrangements, including some resulting from chromosome shattering and random rearrangement (chromothripsis) (Cortes-Ciriano et al, 2020). The degree of copy number alteration varies significantly between cancer (sub)types, with some being classified as 'copy number-driven' while others are considered 'mutation-driven' (Levine et al, 2013). Steele et al (2022) recently established a set of 21 copy number signatures in cancer based on total copy number, loss of heterozygosity and the length of contiguous copy number segments. This grouped tumours into classes defined by the presence or absence of genome doubling and genome-wide loss of heterozygosity, and identified tumours with preferential loss of heterozygosity of particular drivers or evidence of homologous recombination deficiency. A separate analysis identified copy number signatures based on alternative features including copy number change between segments and breakpoint frequency, and linked them to replication stress and defects in NHEJ (Drews et al, 2022).

1.4.1 Patterns of aneuploidy in cancer

Aneuploidy rate varies between cancer types and chromosomes in a non-random manner. Some aneuploidies are recurrent across cancers and act as driver alterations by altering the copy number of cancer driver genes (e.g. 17p loss for TP53) (Lakhani et al, 2023). Others are drivers in specific cancer types, like the frequent loss of chromosome 10 and gain of chromosome 7 in gliomas, which are partially explained by selection for loss of PTEN and KLF6 and gain of EGFR

(Nair et al, 2024). Aneuploidy events may also affect cancer driver genes more indirectly: a common early event in breast cancer is trisomy of 1q, which suppresses TP53 by increasing the copy number of its inhibitor MDM4. Experiments in which aneuploid chromosomes were artificially returned to disomy have shown that tumours become dependent on their driver aneuploidies in a form of oncogene addiction (Girish et al, 2023). Aneuploidy patterns are also influenced by the number of weaker drivers, with chromosome arm gain rate increasing with the ratio of oncogenes to TSGs on the arm and vice versa for chromosome arm losses (Davoli et al, 2013). Aneuploidy patterns in cancer have also been suggested to imitate patterns of normal gene expression in their tissue of origin, potentially acting as a mechanism to 'hard wire' a gene expression profile favoured by that environment (Patkar et al, 2021), However, they are not tissue-specific enough to allow clustering of samples into their tissues of origin, unlike methylation, RNA expression, and protein abundance (Hoadley et al, 2018). Finally, aneuploidy patterns show some evidence of negative selection to avoid loss of essential genes (Shih et al, 2023) and gains of TSGs/losses of oncogenes (Jubran et al, 2024).

1.4.2 Chromosomal instability

In this work, I am primarily interested not in recurrent driver aneuploidies but in the large-scale ploidy alterations and dosage imbalances induced by chromosomal instability. Chromosomal instability (CIN) refers to an increased chromosome missegregation rate and the continuous accumulation of numerical and structural abnormalities (Potapova et al, 2013). CIN is widespread in

aneuploid tumours but not universal: certain cancer types develop characteristic aneuploidies and stably maintain them throughout cancer progression, as seen in hyperdiploid acute lymphoblastic leukaemia (Woodward et al, 2023). CIN tumours become more aneuploid over time and frequently undergo ploidy alterations, particularly via WGD. This is an accelerating, self-reinforcing process, as supernumerary chromosomes induce further chromosomal instability (Hintzen et al, 2022).

1.4.3 Impact on tumour fitness

The impacts of CIN and WGD on tumour fitness are complex and context-dependent. WGD increases proliferation in primary mammary tumours in mice but suppresses it in recurrent tumours (Newcomb et al, 2021), is an initiating event across cancer types but has been shown to block cellular transformation in multiple experiments (de Carcer et al, 2018), and is linked to worsened patient survival in most cancer types but better survival in others (Sanz-Gomez et al, 2023). Genome doubling is more commonly observed in metastases than in primary tumours, and CIN has been shown to enable relapse of k-Ras dependent lung cancer after k-Ras inhibition (Sotillo et al, 2010). Part of this complex relationship is explained by the overall level of aneuploidy: experiments in cells and mice have suggested a model wherein low levels of aneuploidy are tumorigenic but higher levels are tumour-suppressive (Weaver et al, 2007). This is supported by survival data in human cancer, where intermediate levels of copy number alterations are associated with worse prognosis than either very high or very low levels (Andor et al, 2015; Birkbak et al, 2011; Roylance et al, 2012). The

same pattern is found for mutation burden in some, but not all, studies, suggesting that very highly altered tumours may experience 'mutational meltdown' due to Muller's ratchet (Persi et al, 2018). Fitness also depends on the environmental context: experiments on colorectal cancer cells suggest that aneuploid cells have a proliferation advantage specifically in stressful conditions (Rutledge et al, 2016).

The direct negative impacts of CIN include an immediate increase in replication stress (Gemble et al, 2022) and proteotoxic stress (Ohashi et al, 2015), triggering of immune responses via cytosolic DNA and neoantigen production (Kuang & Li, 2022), and activation of senescence pathways (He et al, 2018), as well as an increased risk of homozygous deletion of essential genes and the creation of large-scale dosage imbalances across the genome. WGD, by drastically increasing the overall alteration rate in a punctuated manner, is also likely to reduce the efficiency of purifying selection. These effects are counterbalanced by the dramatically increased intra-tumour heterogeneity in CIN tumours, which allows rapid searching of karyotype space for oncogenic alterations (Hosea et al, 2024). On top of their high base rate of chromosome missegregation, genome-doubled tumours are also buffered against truncating mutations in essential genes in (formerly) hemizygous regions (Lopez et al, 2020) and against genomic imbalances caused by aneuploidy, resulting in a broader chromosome distribution of aneuploidies (Prasad et al, 2022). Additionally, WGD activates oncogenes through loss of chromatin segregation (Lambuta et al, 2023).

1.5 Motivation

The major obstacle to successful cancer treatment is the evolution of resistance, which is enabled by intra-tumour heterogeneity (Marusyk et al, 2020). CIN is an important mechanism in generating diversity, creating highly adaptable tumours with reservoirs of potentially-resistant subclones (Lukow & Sheltzer, 2022). For this reason, it is copy number heterogeneity—heterogeneity between tumours in their copy number profiles—-that is most strongly linked to poor prognosis across cancer types, more so than aneuploidy or polyploidy themselves (van Dijk et al, 2021). However, CIN also presents opportunities for targeted cancer therapies, which depend on the presence of differences between normal and cancerous cells.

One attractive target in this regard is KIF18A, a mitotic spindle kinesin that is essential in polyploid cells but dispensable in diploid cells due to delays at the spindle assembly checkpoint in the former (Quinton et al, 2021; Gliech et al, 2024). Small-molecule inhibitors of KIF18A have been shown to kill chromosomally unstable cancer cells and to induce regression of tumours in mice (Payton et al, 2024), and are currently in stage I/II clinical trials (Gallard et al, 2024). A more general strategy is to use genome-destabilising drugs that push chromosomally unstable tumours past their fitness peak and induce mutational meltdown, although this carries the risk of increasing CIN in normal cells and thus causing secondary cancers (Cunningham et al, 2019). Other approaches attack the compensatory mechanisms employed by aneuploid tumours, such as upregulation of the proteasome (Ippolito et al, 2024) or evasion of the immune

system. Successful chromosomally unstable tumours downregulate antigen presentation mechanisms (Burdett et al, 2024) and induce immune tolerance through chronic inflammatory signalling (Bakhoum et al, 2018). This suggests the possibility of targeting them by reactivating antigen presentation mechanisms to expose their neoantigen load (Balasubramanian et al, 2022) and restoring normal inflammatory signalling.

To turn CIN from a cancer driver to a liability, it is vital that we understand the effects of large-scale dosage alterations on the cancer cell and on cancer evolution. Despite their prevalence in cancer, and the vast (and increasing) amount of research in this field, the effects of these alterations are still poorly understood compared to similarly-frequent small-scale mutations. Some questions of interest include:

- What are the effects of driver aneuploidies beyond dosage changes in major driver genes? How does the length of a copy number alteration affect its impact? What is the importance of co-alteration of entire gene networks, or, conversely, imbalances between members of protein networks? For example, loss of multiple tumour suppressor genes on 8p cooperatively promotes cancer (Tschahargeneh et al, 2016), and gain of chromosome 7 rescues loss of chromosome 10 in gliomas (Nair et al, 2024) — how prevalent are these phenomena and what is their overall significance in cancer evolution?

- Is dosage sensitivity a relevant force in cancer? Tumours show avoidance of homozygous loss of essential genes, but tolerate hemizygous losses. Is copy number variation of dosage-sensitive genes deleterious in the cancer context? Is it ever removed by purifying selection, and if not, how much genetic load does it contribute to the tumour?
- In which contexts is CIN beneficial to a tumour, and in which is it detrimental? We know that it is beneficial in stressful environments, and there may be a 'switch' towards favouring CIN as the tumour metastasises (Bakhoum et al, 2018). Why do cancer types differ in CIN levels, and what determines the 'critical point' at which CIN becomes deleterious? Within the same cancer type, why do some patients develop chromosomally unstable tumours and others don't?
- How are different manifestations of CIN related to each other? Are polyploidy and hypodiploidy opposites or two sides of the same coin? Hypodiploidy has been linked to poor prognosis in acute lymphoblastic leukaemia (Schultz et al, 2014); does this apply across cancer types, and does it affect tumour fitness in the same way as WGD? What is the role of WGD in buffering vs creating CIN?

1.6 Cancer Genomics

This section briefly summarises the development of cancer genomics, then introduces some data sources and techniques relevant to this project.

1.6.1 Development of Cancer Genomics

Cancer is a disease of the genome, driven both by mutations in specific genes and by large-scale karyotypic changes. Cancer cytogenetics provided early genome-scale insights into cancer evolution, with advancements in chromosome banding techniques in the 1970s enabling identification of distinct chromosomal alterations and targetable gene fusions like the 9:22 translocation that causes chronic myelogenous leukaemia (Doroshow & Doroshow, 2019). To date, more than 80,000 cancer karyotypes have been deposited in the Mitelman Database of Chromosome Aberrations & Gene Fusions (Mitelman et al, 2025). Despite their utility for identifying broad chromosomal alterations, especially in haematological cancers, cytogenetic methods are limited by low resolution and cannot capture the full complexity of a cancer genome. The need for both broad and deep characterisation of cancer genomes formed part of the motivation for the Human Genome Project (Dulbecco, 1986); with that in place, and the beginning of the HapMap project promoting a focus on variation between genomes (Wheeler & Wang, 2013), the stage was set for the emergence of cancer sequencing projects.

In the last twenty years, large-scale multi-omics sequencing projects have transformed the landscape of cancer genomics by enabling standardised processing of large numbers of cancer genomes, with DNA sequencing supplemented by data on RNA expression, protein abundance and epigenetic alterations along with detailed clinical information. These projects generally employ next-generation sequencing of bulk tumour samples, with tradeoffs

between coverage and sample size: the Cancer Genome Atlas performed whole-exome sequencing of >11,000 tumours with paired normal samples, the Pan-Cancer Analysis of Whole Genomes project performed whole-genome sequencing of ~3,000 tumours, and MSK-IMPACT is a clinical project performing targeted sequencing on a more limited set of several hundred actionable cancer genes (Cheng et al, 2015), which allows greater sequencing depth and the detection of variants present in only a small fraction of cancer cells (Bewicke-Copley et al, 2019). Cancer sequencing projects continue to grow, with the ICGC-ARGO project aiming to sequence 100,000 cancer patients from around the world[6], and more data coming from the UK's 100,000 Genomes Project (Sosinsky et al, 2024). Another important project is the Cancer Cell Line Encyclopedia, which provides rich genomic and phenotypic data on over a thousand human cancer cell lines (Ghandi et al, 2019), including screens of drug sensitivity (Barretina et al, 2012) and gene essentiality information from genome-wide CRISPR knockout and RNAi knockdown screens (Tsherniak et al, 2017).

1.6.2 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) project ran from 2005 to 2013 and resulted in the multi-omics profiling of tumour and normal samples from more than 11,000 patients, covering 33 types of cancer. It has a central place in the field of cancer genomics, having provided the foundation for thousands of research papers on all aspects of cancer and spurred the development of both computational and

[6] http://platform.icgc-argo.org/

molecular tools for large-scale genomic analysis[7]. For example, TCGA has shown that some cancer types are better grouped by molecular profiles than by site of origin, paving the way for 'basket trials' which enroll patients based on specific genetic alterations irrespective of their cancer type (Hobbs et al, 2022). The size of the TCGA cohort has facilitated genome-wide association studies linking germline variants to immunosurveillance (Sayaman et al, 2021) and characterisation of driver alterations across cancer types (Bailey et al, 2018). It has additionally enabled the identification of dozens of 'mutational signatures' in cancer genomes, shedding light on endogenous mutational processes and on the genomic impacts of smoking and chemotherapy (Alexandrov et al, 2013; Levatic et al, 2022).

An important factor in TCGA's success is the Genomic Data Commons (GDC), which hosts and standardises the vast stores of sequencing data generated by TCGA and similar projects. The GDC creates and runs automated pipelines to transform raw sequencing data into quality-controlled mutation calls, copy number profiles, RNA read counts and more. Finally, it provides tools for the efficient upload and download of cancer genomic data and interfaces with the eRA Commons system to manage access to sensitive patient data (Zhang et al, 2021b).

[7] Information from the TCGA impact page, https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history

1.6.3 Copy Number Profiling

High-throughput methods used for copy number profiling of tumour samples include array comparative genomic hybridisation (array-CGH), SNP microarrays, and next-generation sequencing (Gordeeva et al, 2022). The latter two are capable of distinguishing between parental alleles and provide two useful data points for every locus in a sample: logR, the log ratio of signal intensity relative to a baseline, and the B-allele frequency (BAF), the relative signal (or read proportion) for the non-reference allele. A plethora of algorithms have sprung up to convert this information into absolute allele-specific copy number profiles, including ASCAT (van Loo et al, 2010), ABSOLUTE (Carter et al, 2012), FACETS (Shen & Seshan, 2016), CNVkit (Talevich et al, 2016), and PURPLE (Priestley et al, 2019). Two in particular have been widely adopted across the cancer genomics community and run on all TCGA samples by the GDC: ASCAT and ABSOLUTE. Both algorithms take in paired tumour-normal SNP array or NGS data, identify segments of contiguous copy number state, and simultaneously solve for ploidy and tumour purity. In ambiguous situations, ABSOLUTE additionally refers to a cancer type-specific dataset of recurrent aneuploidies (Carter et al, 2012).

Copy number profiles are usually presented as segment files: each segment is a contiguous genomic region in the same copy number state, with chromosomal coordinates and copy number information. The copy number information may be presented as logR ratios relative to the rest of the genome (Zhao & Baudis, 2024), as logR ratios relative to a paired normal sample, or as absolute integer copy

numbers. In the case of allele-specific absolute copy number data, the total copy number is split into the number of copies contributed by each parent based on germline polymorphisms. For any given segment, the major copy number is the copy number of the parental allele with the most copies at that locus, and vice versa for the minor copy number. A minor copy number of zero indicates loss of heterozygosity at that locus, i.e. deletion of one of the parental alleles.

Allele-specific absolute copy number profiles enable several methods of tracing cancer copy number evolution. One, called MEDICC2, calculates the minimum number of events between any two copy number profiles and uses this to construct parsimonious copy number evolution histories and call WGDs (Kaufmann et al, 2022). Another, GRITIC, enumerates all possible binary tree structures leading to a given copy number state and integrates mutation data to distinguish between routes, allowing it to detect non-parsimonious evolutionary trajectories. GRITIC has been applied to time the accumulation of complex copy number gains in the PCAWG and Hartwig databases, showing an increase in punctuated gains with WGD, weak correlations between WGD timing and the extent of chromosomal gains and losses post-WGD, and changes in chromosome arm loss patterns after WGD (Baker et al, 2024). Two useful heuristics used to time events in tumour evolution include (i) the biallelic infinite sites assumption (Demeulemeester et al, 2022), which is used to infer that duplicated mutations arise from copy number gains and thus must have occurred before the gain event (Gerstung et al, 2020; Kim et al, 2021), and (ii) the assumption that copy number losses that result in loss of heterozygosity occurred before genome duplication

(Baker et al, 2024; Kikutake & Suyama, 2023). These can be used to create a timeline of common alteration orderings in a given cancer type (Gerstung et al, 2020).

1.6.4 Intra-Tumour Heterogeneity

Bulk sequencing is not ideal for the understanding of tumour heterogeneity. It does provide information on the proportion of reads supporting a given mutation (its variant allele frequency, VAF), and with knowledge of the copy number state of each mutation, the VAFs can be converted to cancer cell fractions and used to infer the clonal composition of a tumour (Vandin, 2017). However, it struggles with subclonal copy number alterations (Tarabichi et al, 2021) and generally fails to detect rare subclones (Dentro et al, 2021). In cases of extreme intra-tumour heterogeneity or low tumour purity, even computation of overall copy number profiles as the weighted average of all clones in the tumour may become impossible (Carter et al, 2012). However, several methods have been developed that explicitly take advantage of the limitations of bulk sequencing to infer useful biological features of tumours. For example, the 'copy number heterogeneity' index referenced above is calculated based on deviation from integer copy numbers obtained by the ASCAT algorithm from bulk sequencing data (van Dijk et al, 2021), and several methods have been developed to estimate immune infiltration into tumours based on expression of immune markers within tumour samples (Chen et al, 2018; Yoshihara et al, 2013).

Recent studies have focused explicitly on characterising tumour heterogeneity through the use of multi-regional sampling. These include the TRACERx project, which combines multi-regional and longitudinal sampling to trace tumour evolution (Bailey et al, 2021), as well as individual studies of malignant pleural mesothelioma (Meiller et al, 2021), ovarian cancer (Lakis et al, 2020), basal cell carcinoma (Birl et al, 2023) and oral squamous cell carcinoma (Jie et al, 2021). Phylogenetic approaches have been used to understand inter-sample relationships, but as these are based on mixed populations they are more properly considered 'multiregional tumour trees' than phylogenetic trees (Alves et al, 2017). Single-cell sequencing permits full resolution of subclonal heterogeneity, but remains noisy, non-standardised and more expensive than bulk sequencing (Jia et al, 2022), which limits the numbers of patients analysed and the investigation of rare cancer subtypes or of statistical interactions between alterations. Several methods have been proposed to integrate single-cell with bulk sequencing data to mitigate the weaknesses of each (Malikic et al, 2019; Baghaarabani et al, 2021).

This work is based primarily on data from The Cancer Genome Atlas, including both germline and somatic mutation data, somatic copy number profiles, and RNA-seq expression data. In Chapter 3, I supplemented this with karyotypes from the Mitelman Database of Chromosome Aberrations & Gene Fusions in Cancer, and in Chapter 4 with protein abundance data from the Clinical Proteomic Tumor Analysis Consortium and the Cancer Cell Line Encyclopedia. I

also made use of data from the TARGET acute lymphoblastic leukaemia sequencing project, and of mutational signatures and WGD timing estimates from the Pan-Cancer Analysis of Whole Genomes project.

1.7 Aims

This thesis aims to further our understanding of the consequences of chromosomal instability in cancer, with a focus on the development of extreme ploidy states and the impact of dosage sensitivity. In Chapter 2, I search for genetic and environmental factors associated with the development of polyploidy and hypodiploidy in tumours from across The Cancer Genome Atlas, analysing the relative importance of each factor, the predictability of extreme ploidy, and the commonalities and differences between the two opposite ploidy states. In Chapter 3, I explore the karyotypic evolution of hypodiploid and polyploid tumours and analyse patterns of large-scale chromosome loss with respect to mechanical and dosage sensitivity-related factors. Finally, in Chapter 4, I investigate the importance of haploinsufficiency as a form of dosage sensitivity in cancer evolution, focusing on haploinsufficiency as a driver mechanism for inactivation of tumour suppressor genes. I take advantage of the detectable changes in cell behaviour caused by TSG inactivation to perform a functional screen for haploinsufficiency, and assess the relationship between dosage sensitivity and dosage attenuation/compensation in TSGs at the RNA, protein and activity levels.

Chapter 2

Origins of extreme ploidy in cancer

2.1 Introduction

Extreme ploidy—deviation from diploidy towards polyploidy or haploidy—is a common and consequential development across human cancers. The rate of extreme ploidy is highly variable between cancer types, subtypes and patients, ranging from three-quarters of adrenocortical carcinoma cases studied by TCGA (Zheng et al, 2016) to less than 1% of acute myeloid leukaemia cases (Kumar, 2011). Even within the same tissue, different cancer subtypes frequently show distinct ploidy profiles. For example, triple-negative breast cancer is more than twice as likely to undergo WGD as any hormone receptor-positive subtype (Wu et al, 2021); papillary thyroid carcinoma is almost always near-diploid, in stark

contrast to the Hürthle cell subtype, which is majority near-haploid (Gopal et al, 2018); and kidney chromophobe tumours are majority hypodiploid, while papillary kidney cancer is typically near-diploid or aneuploid with chromosome gains (Casuscelli et al, 2017; Kovac et al, 2015).

While polyploidy and hypodiploidy are in one sense two opposite extremes, they are both manifestations of extreme chromosomal instability and share several features, including high levels of aneuploidy and loss of heterozygosity, TP53 mutations (Bielski et al, 2018), and (in some cancer types) altered prognosis (Stengel et al, 2014). Additionally, hypodiploid cells have an increased tendency to undergo WGD (Carroll et al, 2019). What drives differences between cancer types in their tendency to adopt these highly aberrant ploidy states? More broadly, why do some tumours become polyploid or hypodiploid and others don't?

There are three levels at which this question may be usefully interrogated:

1) Mechanical: which physical factors affect the rate of mitotic errors leading to polyploidisation, haploidisation and aneuploidy?
2) Establishment of ploidy-aberrant cells: how does the cell manage its first few abnormal mitotic divisions without dying or being killed by failsafe mechanisms or the immune system?
3) Expansion: why does the ploidy-aberrant subclone so often expand to become the dominant component of the tumour? Single-cell sequencing of ovarian cancer indicated that the vast majority of cases were either polyploid or carried a small genome-doubled subclone (McPherson et al,

2024). Are there factors in the tumour environment that determine the extent to which a ploidy-aberrant subclone can expand? Both polyploidy and hypodiploidy have been suggested to promote chemoresistance, for example, making chemotherapy a possible selective factor (Schmidt et al, 2024; Molina et al, 2021).

A host of experiments have identified environmental factors capable of inducing polyploidy in cell lines. Infection with oncogenic viruses can cause polyploidisation through cell fusion (Hu et al, 2009) or via inactivation of the genome stability regulators p53 and pRB (Nehme et al, 2022) Experimentally-induced hypoxia also causes polyploidy in colon (Lopez-Sanchez et al, 2014) and glioma cell lines (Liu et al, 2022). Importantly, several classes of chemotherapeutics can directly induce polyploidisation by interfering with mitosis. These include well-established microtubule stabilisers such as docetaxel (Ogden et al, 2015) and DNA-damaging agents such as cisplatin (Delinassios et al, 2024). Inhibitors of Aurora B kinase (AURKB) are currently being investigated as potential chemotherapeutics (Borah & Reddy, 2021). They act by disrupting chromosome segregation during mitosis and overriding the spindle assembly checkpoint, directly inducing polyploidisation (Vora et al, 2025). Polyploidy may also be induced by radiotherapy (Zhang et al, 2021c). Experiments have also shown that successful establishment of polyploid cells depends on TP53 inactivation (Horri et al, 2015), and that polyploid cells are selectively eliminated by the immune system in immunocompetent mice (Senovilla et al, 2012).

In this chapter, I am interested in the factors that influence ploidy under physiological conditions and their relative contributions to differences in ploidy between patients and cancer types. Factors mechanistically linked to WGD in cell lines do not necessarily have the same effect in patients—for example, despite the demonstrated ability of AURKA inhibition to cause polyploidy, AURKA mutations are not associated with WGD in the MSK-IMPACT cohort (Bielski et al, 2018). For this reason, it is essential to analyse the effect of each factor specifically within patient cohorts. Recent studies indicate that the strongest predictor of polyploidy in patient tumours is somatic or germline mutation of core genome stability regulator TP53, which blocks proliferation of polyploid cells when intact (Kim et al, 2021b), although 46% of WGDs arise in TP53-wildtype tumours (Bielski et al, 2018). Bielski et al also found that almost half of the variability in polyploidy rate between cancer types can be explained by differences in tissue proliferation rates, suggesting that WGD rate is largely explained by the level of opportunity for random mitotic errors rather than specific genetic or external factors. Another analysis of somatic mutation rates in polyploid tumours found enrichment for mutations that promote centrosome clustering, and depletion for mutations related to microsatellite instability (Quinton et al, 2021). Hypodiploidy is less well-studied than polyploidy. TP53 mutations are strongly associated with hypodiploidy in acute lymphoblastic leukemia, but this is not seen in near-haploid cases, which instead show alterations in Ras signalling (Holmfeldt et al, 2013). This aligns with the suggestion that low-hypodiploid and near-haploid cases arise through different

mechanisms (Walen, 2014). Outside leukaemia, hypoxia has been linked to a copy number signature associated with defects in homologous recombination and with certain patterns of loss of heterozygosity (Steele et al, 2022). Homologous recombination defects have also been linked to polyploidy in breast cancer (Wu et al, 2021).

I first identify polyploid and hypodiploid cases in a cohort of 7,650 cancer patients from the Cancer Genome Atlas and then search for genetic and environmental factors that distinguish them from diploid tumours. I have three central questions:

1) Do extreme ploidy states develop at random, or can they be predicted?
2) Do environmental factors that are known to induce polyploidy *in vitro* (e.g. viruses, hypoxia) have the same effect at physiological levels in real patients? How much of a contribution does each factor make?
3) Are there shared factors between polyploidy and hypodiploidy?

2.2 Methods

2.2.1 Identifying Polyploid and Hypodiploid Tumours

Genome-wide allele-specific copy-number data for 10,364 patients with 33 cancer types sequenced by the Cancer Genome Atlas (TCGA) was obtained from the Genomic Data Commons. These absolute copy number profiles had been derived from SNP6-chip relative copy number profiles using two different algorithms: ASCAT (Van Loo et al, 2010) and ABSOLUTE (Carter et al, 2012; Taylor et al, 2018). I filtered both datasets to remove samples identified as likely contaminated or swapped by van Loo et al. I also removed formalin-fixed paraffin-embedded samples due to low DNA quality. Finally, I removed metastatic or recurrent tumours, and deduplicated samples to keep only one per patient.

I computed ploidy for each sample as:

$$P = \frac{\sum_{i=1}^{N} L_i C_i}{\sum_{i=1}^{N} L_i}$$

Eq. 2.1 Calculation of sample ploidy. L_i and C_i indicate the length and copy number of a copy number segment i; N indicates the number of copy number segments in the sample.

To make WGD status calls, I used the MEDICC2 (Minimum Event Distance for Intra-tumour Copy-number Comparisons) command-line tool (Kaufmann et al, 2022) to estimate the number of gains and losses required to reach a given

tumour's copy number profile from a diploid state. Tumours for which the minimum number of copy number alterations was decreased when a genome doubling was included as a possible step were considered WGD-positive. I ran MEDICC2 (v1.0.2) using both ASCAT-computed and ABSOLUTE-computed copy number profiles and removed samples where the WGD call was not concordant between the two datasets. Because MEDICC2 was designed to analyse multi-sample trees instead of single samples, I (in consultation with author Tom Kaufmann) created a duplicate of each sample and computed the length of the branch between the theoretical diploid ancestor cell and the split of the two duplicates.

Hypodiploidy has been best studied in acute lymphoblastic leukemia (ALL), where hypodiploid classes are defined as high-hypodiploid (40-45 chromosomes), low-hypodiploid (30-39 chromosomes) and near-haploid (< 30 chromosomes) (Harrison et al, 2004). Because hypodiploidy does not have an obvious saltatory cause analogous to whole genome duplication, I relied on these thresholds to define hypodiploids (low-hypodiploids and near-haploids) in all TCGA cancer types. Additionally, because reliable copy number information was unavailable for the sex chromosomes, I excluded them and defined ploidy groups based on autosome number (near-haploid: < 28, low-hypodiploid: 28-37). I calculated chromosome somy as the average copy number along the chromosome rounded to a whole number.

Due to the availability of heterozygosity information for the TCGA tumours, I was also able to identify former hypodiploids that had since undergone whole genome doubling. I defined a chromosome with loss of heterozygosity (LOH) over >90% of its length as 'lost' and capped its somy at 1 to calculate the minimum possible number of autosomes during the tumour's evolution:

$$min_somy = \begin{cases} \min(1, somy) & \text{if } lost \\ base_somy & \text{otherwise} \end{cases}$$

$$n_{min} = \sum_{1}^{22} somy_{min}$$

Eq. 2.2 Calculation of minimum number of chromosomes (or autosomes) during a tumour's evolution based on LOH. A chromosome with LOH over >90% of its length has its minimum somy capped at 1, taking the minimum of 1 and current somy to account for possible nullisomy. A chromosome with retained heterozygosity is assigned a minimum copy number of 2, except for the sex chromosomes in males. I excluded sex chromosomes in this work due to unreliable copy number data and counted the minimum number of autosomes instead. Minimum somy for each chromosome is then summed over all autosomes.

I assume that losses causing LOH occurred pre-WGD and thus that tumours with ≥7 LOH chromosomes, i.e. a minimum autosome count < 38, went through a hypodiploid stage, even if they are no longer hypodiploid due to a subsequent WGD. This assumption is based on the idea that pre-WGD losses are much more likely to lead to LOH than post-WGD losses, as post-WGD LOH requires at least two loss events affecting the same homologue of the same chromosome.

2.2.2 Germline and Somatic Mutations

I first searched for germline mutations associated with the development of extreme ploidy using a genome-wide association study of two related traits: WGD status and ploidy (mean copy number across the genome).

Germline Genotype Data Processing and Quality Control

Partially-processed genotype data for 10,128 TCGA patients was obtained from Carrot-Zhang et al (2020). This data came from the Affymetrix SNP6.0 chip and had already been partially quality-controlled by Carrot-Zhang et al to remove SNPs with heterozygosity > 3 standard deviations above the mean for their ancestry cluster, SNPs and individuals with > 5% missing data, and SNPs with minor allele frequency < 0.005. I used the Michigan Imputation Server (v1.2.4) to impute non-genotyped variants against the Haplotype Reference Commission's reference panel, bringing the dataset from 679,958 variants to 31,127,678 variants. I removed SNPs with an imputation quality score < 0.5 or with >2 alleles.

VCF files were converted to plink (v1.9) format and filtered to remove samples missing sex or phenotype information. To create a set of SNPs in linkage equilibrium with each other, I pruned SNPs with a pairwise correlation > 0.2 in 50 kb windows and also removed all SNPs in long-range linkage disequilibrium regions, the coordinates of which I obtained from Meyer's (2020) plinkQC Github. I calculated relatedness using the KING Table method (Manichaikul et al, 2010) on pruned, genotyped SNPs and removed 15 individuals who were related to one

or more individuals in the dataset with a coefficient > 0.05, which should remove third-degree relatives using KING's scaling.

I detected and removed variants with a frequency distribution strongly deviating from Hardy-Weinberg equilibrium expectations ($p < 1e-6$) in controls of European ancestry Subcluster 1 (N = 1,094), except for variants that had previously been identified as genome-wide significant hits in GWAS of cancer risk, following Sayaman et al (2021). I downloaded all cancer-associated variants from the GWAS catalog (https://www.ebi.ac.uk/gwas/), selected annotated variants with genome-wide significant hits ($p < 5e-8$) and used UCSC's LiftOver command-line tool to convert variant coordinates from human genome version hg38 to hg19. After all filtering was complete, I was left with 8,611,526 SNPs for 7,161 patients.

Covariates

Ancestry. I performed principal component analysis using quality-controlled, LD-pruned variants with MAF > 10% in plink. I performed several checks to ensure principal components reflect ancestry rather than technical artefacts. First, principal components were plotted against patient self-reported race. Second, I downloaded the 1000 Genomes Project genotyping files for individuals from 26 worldwide populations, MAF-filtered and pruned the variants as above, and performed PCA to visualise the similarity between 1000 Genomes Project and TCGA patient distributions in PC space. Finally, I projected 1000 Genomes population-labelled data onto the TCGA-derived principal component weights

and visually identified PCs that separated populations. This led me to include the first 9 PCs as covariates in the GWAS.

Genome Wide Association Study

I performed genome-wide association analyses for each SNP with MAF > 1% using plink. For the ploidy phenotype, which is quantitative and continuous, I used generalised linear regression (GLM), while for the case/control (binary) WGD status phenotype I used logistic regression with fallback Firth regression. Both analyses had age, sex, cancer type and the first nine principal components as covariates and standardized covariate variance. SNPs with p < 5e-8 were considered genome-wide significant.

To increase power, I also analysed associations between variants and phenotypes at the gene level. I mapped the 8,611,526 quality-controlled SNPs to 18,144 genes based on NCBI gene coordinates (human genome version hg19) and used MAGMA (v1.10) to test for gene-level phenotype associations, controlling for covariates and linkage disequilibrium. Genes with Bonferroni-corrected $p < 0.05$ were considered significantly associated with the phenotype. Manhattan and QQ plots were made using the ggmanh R package. GWAS power calculations were performed with the genpwr R package (v1.0.4). I calculated the effect size (continuous trait, ploidy) and odds ratio (case/control trait, WGD status) needed to have 90% power to detect an association at an alpha of 5e-8 (genome-wide significance) over a range of minor allele frequencies. I used a case/control ratio of 0.28 and a ploidy standard deviation of 0.88, and assumed additivity.

Heritability Analysis

I calculated genome-wide SNP-heritability (the proportion of phenotypic variance attributable to the genetic variance covered by all measured SNPs) of WGD status and ploidy using two orthogonal methods: linkage disequilibrium score regression (LDSC) and genomic-relatedness-based restricted maximum likelihood estimation (GCTA-GREML).

Twin- and family-based studies of trait heritability exploit the principle that, if a trait is heritable, more closely-related individuals should be more phenotypically similar. GCTA-GREML extends this idea to distantly-related individuals. I calculated a genetic relatedness matrix between all non-closely-related patients using genotyped SNPs and estimated genome-wide SNP-based heritability using GREML (v1.94.0), with ancestry principal components, sex, age, and cancer type as covariates. Heritability estimates were converted from the observed scale to the liability scale using a population prevalence (of tumour WGD among primary cancer patients) of 0.3.

LDSC is based on the idea that, if a phenotype is affected slightly by many SNPs across the genome (polygenicity), a SNP in higher linkage disequilibrium with other SNPs will have a higher effect size estimate due to having a higher chance of tagging a causal SNP. This allows us to distinguish an inflated QQ-plot caused by population stratification confounding from one caused by polygenicity, even if sample size is too low or effect size too small to identify any individual associated

SNPs. The relationship between SNP effect estimate and SNP LD Score is as follows:

$$E[X^2 | l_j] = \frac{Nh^2 l_j}{M} + Na + 1$$

Eq. 2.3 Linkage disequilibrium score regression $E[X^2 | l_j]$ is the expected association test statistic for SNP j, N is the sample size, M is the number of SNPs analysed, h^2 is the SNP-based heritability, and a is a measure of confounding due to population stratification or relatedness (Bulik-Sullivan et al, 2015).

Estimates of SNP effects (test statistic) adjusted for all covariates were obtained from the GWAS summary statistics for each phenotype. I calculated the LD score for each SNP and regressed each SNP effect against its LD score using the ldsc command line tool (v1.0.1) to estimate genome-wide SNP heritability and population stratification.

Somatic Mutations

mafCompare from the maftools R package (v2.22.0) was used to identify differentially mutated genes between ploidy groups. I also used logistic regression restricted to cancer genes, defined based on the COSMIC Cancer Gene Census (Tiers 1 and 2):

ploidy_status ~ mutated + total mutations + cancer type

Eq. 2.4 Logistic regression of ploidy status on somatic mutations., controlling for total mutations in the same and cancer type. Analysis was run separately for polyploidy and hypodiploidy. Each cancer gene was analysed independently and results were corrected for multiple testing using Benjamini-Hochberg correction.

For GO enrichment analysis, I used the enrichGO function from clusterProfiler (v4.14.3) on significant hits from the logistic regression analysis, with COSMIC genes as the gene universe. For mutational signatures, I used SigProfilerExtractor (v1.1.24) to calculate the activity of each COSMIC mutational signature in each sample based on mutational context matrices created by the PCAWG project.

Proliferative Index

I downloaded RNA-seq data (raw counts) for all TCGA samples from the Genomic Data Commons using the TCGAbiolinks R package (v2.34.0) and normalised it using the cpm() function in edgeR (v4.0.16). The proliferative index for each sample was calculated as the median expression (CPM [counts per million] or raw counts) of 117 genes included in the metaPCNA proliferation signature computed by Venet et al (2011) from the top 1% of genes correlated with the proliferation nuclear cell antigen across healthy samples.

2.2.3 Environmental Factors

Patient smoking history was obtained from the TCGA BCR biotabs, columns 'tobacco_smoking_history' (1 = never smoker, 2 = current smoker, 3 and 4 = former smokers) and 'number_pack_years_smoked'. For clinical data on viral infection, I used the 'hpv_status_by_ish_testing' and 'viral_hepatitis_serology' columns in the HNSC (head and neck squamous cell carcinoma) and LIHC

(hepatocellular carcinoma) datasets respectively. I downloaded hypoxia scores for TCGA tumours from Bhandari et al (2019) and viral abundance from Chen et al (2021). Virus levels were dichotomised into present/absent.

2.2.4 Random Forest Model

I built a random forest model to distinguish between polyploid and near-diploid samples using randomForest() in R. Missing hypoxia scores were imputed using the median. Smoking history was categorised into 'never smoker', 'ever smoker' and 'unknown'. Training and test sets were created at an 80/20 ratio using the caret R library (v7.0.1) and feature importance was calculated as the mean Gini (impurity) decrease.

2.3 Results

2.3.1 Identifying Polyploid and Hypodiploid Tumours

For 10,346 primary tumours sequenced by the Cancer Genome Atlas consortium, I calculated tumour ploidy (average copy number across the genome, excluding the sex chromosomes) based on pre-computed copy number profiles derived from two callers, ASCAT and ABSOLUTE. A comparison of ploidy estimates from the two callers revealed that they are only moderately correlated ($r = 0.54$, Fig. 2.1), with most of the discordance coming from samples that disagree fundamentally on the underlying ploidy from which each segment copy number is derived: 18.8% of samples have ploidy that differs by > 0.5 (more than half a genome)

between ASCAT and ABSOLUTE. When I filter to include only samples with a concordant MEDICC2 WGD call (either WGD or Non-WGD for both callers), this correlation improves to r = 0.94. I thus decided to take a consensus approach, restricting all analyses to these 7,650 concordant samples.

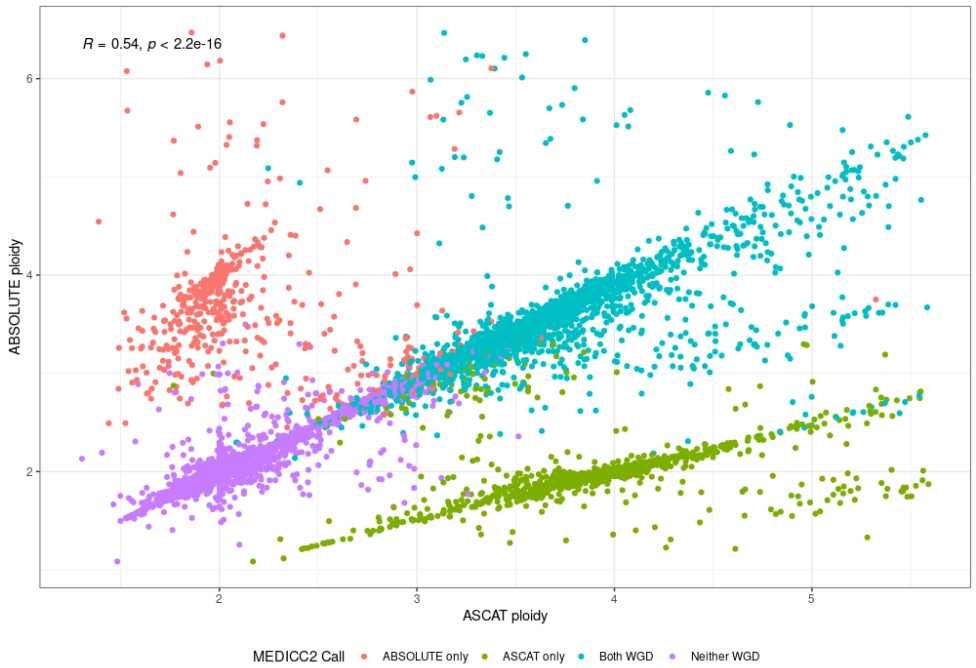


Fig. 2.1 Comparison of tumour ploidy calculated based on ASCAT vs ABSOLUTE copy number profiles. MEDICC2 was run separately on all samples to make WGD calls using copy number profiles from ASCAT and ABSOLUTE. Only samples that were concordant for WGD call based on both copy number profiles (either 'Both WGD' or 'Neither WGD') were included in my analyses.

The requirement for a consistent WGD call resulted in the exclusion of 1,589 samples with copy number data from both callers: 1,076 with a WGD call based on ASCAT but not ABSOLUTE copy numbers, and 513 with the reverse. To determine whether discordant samples were more likely to have chaotic, heterogeneous genomes, I compared copy number heterogeneity values as computed by van Dijk et al (2021) and found that discordant samples have significantly higher CNH than concordant samples as a whole (0.074 vs 0.047, p < 2.2e-16) but are intermediate between concordant non-WGD and concordant

WGD+ samples (Appendix Fig. 6.1a). Using another metric of intra-tumour copy number heterogeneity, discordant samples are called by ABSOLUTE as having a higher subclonal genome fraction (0.13 in discordant vs 0.07 in concordant, $p < 2.2e-16$, Appendix Fig. 6.1b). This is again intermediate between the values for concordant non-WGD and concordant WGD+ samples. Discordant samples also have slightly lower tumour purity (0.77 vs 0.79, $p = 2.1e-6$, Appendix Fig. 6.1c) based on consensus purity estimates from Aran et al (2015). I next considered whether these samples are over- or under-segmented, finding that they have a slightly higher number of segments than concordant samples (significant only using ASCAT segment count, median 150 in discordant samples vs 136 in concordant samples, $p = 1.5e-5$, Fig 6.1d-e).

Discordance occurred most frequently in low-grade glioma, adrenocortical carcinoma, kidney chromophobe, esophageal carcinoma and sarcoma (Appendix Fig. 6.1f). The latter four of these are known for chromosomal instability, and ACC, KICH and SARC in particular exhibit relatively frequent hypodiploidy. The top five most concordant cancer types (THCA, UVM, THYM, LAML and UCEC) all tend to be chromosomally stable, and there is a moderate negative correlation between the WGD rate in a cancer type (calculated using MEDICC on concordant samples) and the proportion of samples in that cancer type with concordant calls ($r = -0.4$, $p = 0.022$, Appendix Fig. 6.1g). The discordant samples have significantly higher median ploidy based on ASCAT but not ABSOLUTE copy number calls (Appendix Fig. 6.1h-i). While LOH is generally higher in WGD+ samples, and it is higher in concordant WGD+ samples than concordant

non-WGD samples, ASCAT-inferred LOH is lower in samples with an ASCAT-only WGD call than in samples with an ABSOLUTE-only WGD call (Appendix Fig. 6.1i-k). Of the 21 main copy number signatures identified by Steele et al (2022) based on ASCAT copy number profiles, CN1 (a signature of diploidy) is more common in concordant samples and CN2 (a signature of genome doubling) is more common in discordant samples, with no other signatures showing a significant difference (two-sided Wilcoxon test).

These analyses suggest that discordant samples are distinguished from concordant samples primarily by their heterogeneous genomes and high levels of chromosomal instability. I have chosen to exclude these samples in order to obtain conservative sets of confident diploid and ploidy-altered tumours, but I acknowledge that the requirement for concordance is likely to preferentially exclude chromosomally-unstable tumours and thus bias estimates of WGD and hypodiploidy rates downwards.

Our first step in classifying tumour ploidy alterations was to identify a reliable method of identifying WGD events that could be applied to allele-specific absolute copy number profiles derived from SNP arrays or from whole-exome sequencing. While genome-doubled tumours generally have high ploidy (typically near-triploid), it is not always trivial to distinguish a near-triploid karyotype derived from a WGD via chromosome losses from one that arises due to many independent chromosome gains. Since the TCGA project finished, a

preponderance of WGD-calling heuristics have accumulated in the literature. I have collected some popular methods in Table 2.1.

Name	Sample Implementations
Ploidy threshold	Ploidy > 2.5 (Kim et al, 2021) Ploidy > 2.7 (Lee et al, 2020) Ploidy > 3 (Zaccaria & Raphael, 2020)
Major allele copy number (MCN)	MCN > 1.5 on at least 50% of at least 11 autosomes (Priestley et al, 2019; Satguneseelan et al, 2022) MCN >= 2 in at least 50% of the genome (Bielski et al, 2018)
Dentro et al, 2021	Allelic copy number configuration (2,2) across >20% of the genome (plus manual checking)
Gerstung et al, 2020	Ploidy-LOH balance, plus (optionally) clustered timing of chromosome gains
GRITIC	Modal copy number of major allele > 1 and clustered timing of gains (Baker et al, 2024)
ABSOLUTE	High ploidy and major allele typically having an even copy number, plus (optionally) presence of duplicated mutations (Taylor et al, 2018)
MEDICC2	Minimum Event Distance between copy number profiles decreases when WGD is included as a possible event (Kaufmann et al, 2022)

Table 2.1 Methods used to call whole genome duplication events in cancer genomics. The major allele for a given copy number segment is whichever of the parental alleles has a higher somatic copy number in that segment. I note that some of the papers in which these methods were presented

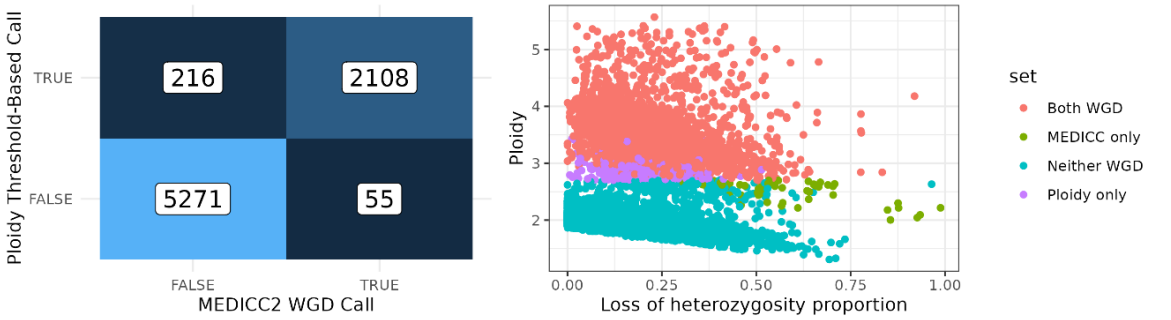
used manual checking (Dentro et al, 2021) or comparison with typical karyotypes for a given cancer type (ABSOLUTE) to confirm borderline calls.

In this research, I am interested in tumours with unusual karyotypes, including, for example, genome-doubled former hypodiploids. It is clear that some of these methods will not detect these cases, particularly the ploidy threshold-, (2,2) configuration- and (in more extreme cases) the MCN-based methods. The gain timing methods presented by Gerstung et al and Baker et al should be capable of handling hypodiploids, but are not applicable to samples without mutation information. Additionally, because I am interested in the evolutionary trajectories of post-polyploid and post-hypodiploid tumours, I would prefer a method that can give a detailed view of the evolutionary history of somatic copy number alterations in the tumour rather than relying on a heuristic of the end result. For this reason, I have chosen to identify WGD events using MEDICC2. MEDICC2 computes the minimum event distance between two copy number profiles (in this case, between a sample and a hypothetical diploid ancestor) assuming parsimony. I call a WGD if the minimum event distance with WGD included as an option is less than the minimum event distance without it. The assumption of parsimony is not infallible–it has been reported that many somatic copy number events do not follow parsimonious paths (Baker et al, 2024)–but I expect that it should suffice for identifying WGD. To characterise my choice of method, I compared WGD calls made using MEDICC2 to calls made using each of the first five methods presented in Table 2.1 (Fig. 2.2).

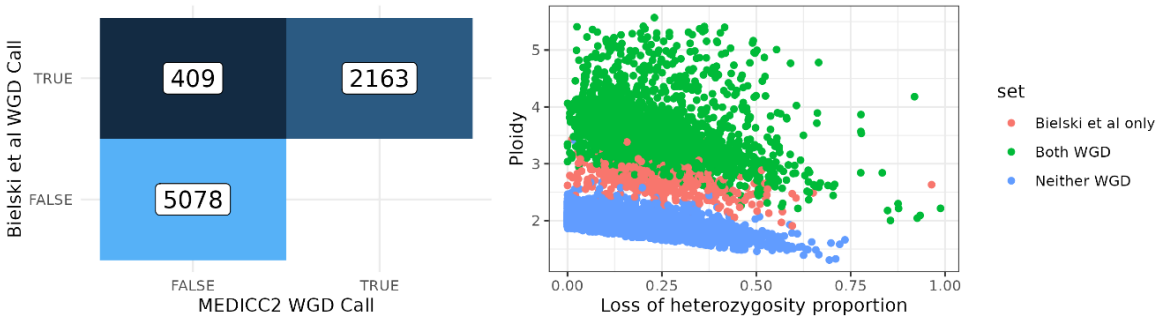
For simplicity, I report these results taking the MEDICC2 calls as reference, so a 'false positive' is a case that MEDICC called as non-WGD but the alternative method called as genome-doubled, and vice versa. This investigation showed that the simplest method, ploidy > 2.7, agreed with MEDICC2 on 96.4% of cases and was four times as likely to call a WGD not called by MEDICC2 than to miss one. The rare cases it missed tended, unsurprisingly, to have high LOH, i.e. to be former hypodiploids. The method used by Bielski et al (2018), a version of the major copy number heuristic, had a similar overall accuracy (94.7%) but was even more strongly biased towards false positives, giving zero false negatives but potentially classifying as genome-doubled many samples that had only experienced independent chromosome gains. Calls made using the (2,2) heuristic were the least similar to MEDICC2's, matching in 86.35% of cases. Unlike the previous two, this heuristic is very restrictive, with 1,039 false negatives vs 5 false positives. It missed formerly hypodiploid cases but also many cases with high levels of gains that exceeded the (2,2) configuration, although these were likely picked up with manual review. Additionally, it missed cases with intermediate ploidy that may represent older WGD events with lots of losses, i.e. allelic configurations of (2,1). Surprisingly, the calls from Gerstung et al and GRITIC had a lower match rate with the MEDICC2 calls than the ploidy-based method, at 92% and 90.6% respectively, although I note that these were based on much smaller samples (N = 600 for Gerstung, 547 for GRITIC). The disagreement between these last two methods and MEDICC2 seems to be primarily localised to intermediate ploidies (~2.5-3). Overall, most of these methods performed similarly to MEDICC2 in identifying WGDs, and the ploidy-

threshold method in fact had the highest agreement – but tended to miss a rare but important group for this project, the hypodiploids.

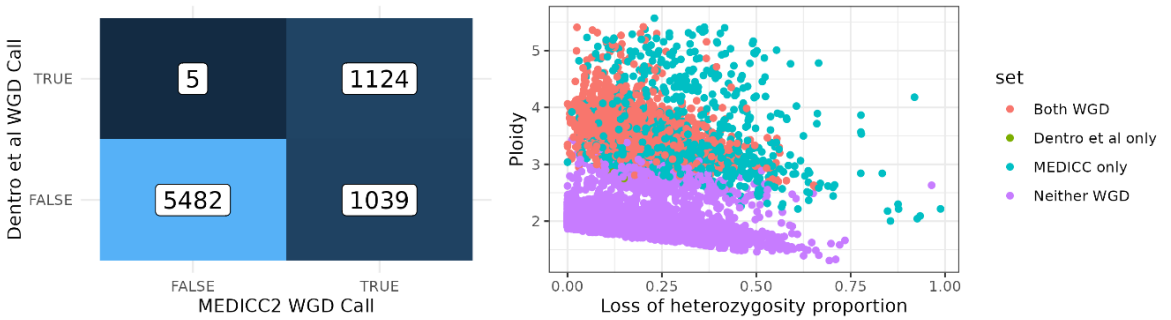
(a) Ploidy Heuristic



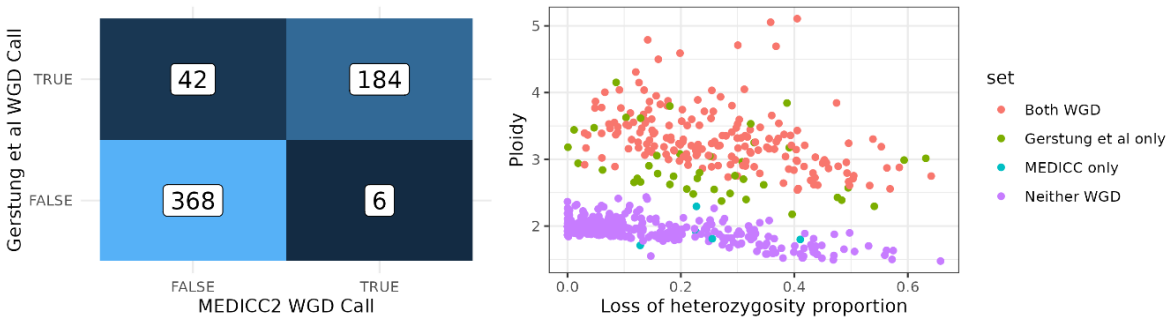
(b) MCN Heuristic



(c) Dentro et al Heuristic



(d) Gerstung Call



(e) GRITIC Call

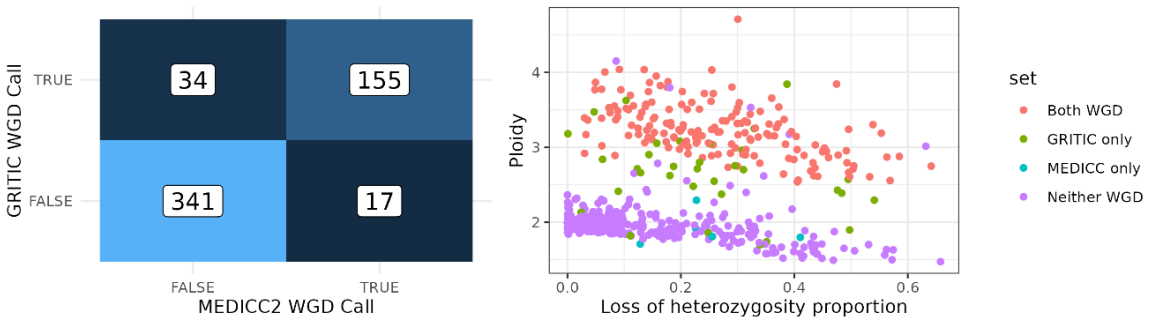


Fig. 2.2 Comparison of WGD calls between MEDICC2 and five alternative methods. MCN = Major Copy Number. ABSOLUTE's calls were excluded from this comparison because I used ABSOLUTE copy number profiles as part of my analysis. I note that this analysis, as with all of the analyses presented in this thesis, considers only cases with concordant MEDICC2 calls from both ASCAT and ABSOLUTE copy number profiles, so they are likely easier to call than the full set of all tumours. For the first three methods, I called WGD on the TCGA dataset using the stated heuristics; for the last two, I downloaded pre-computed calls for TCGA samples included in the PCAWG study.

I classified all cases into one of five ploidy groups based on a combination of WGD call, ploidy and loss of heterozygosity information (LOH). A case was classified as near-haploid if it was inferred to have had fewer than 28 autosomes at any point during its evolution and as low-hypodiploid if its minimum autosome count was between 28 and 37, using hypodiploidy thresholds from the literature (see Methods). All cases with a history of hypodiploidy, even if they had since gained chromosomes or undergone WGD, were classed as hypodiploid unless otherwise stated. Cases with a positive WGD call from MEDICC2 and no history of hypodiploidy were considered polyploid. A conservative set of near-diploid tumours was defined based on lack of hypodiploidy or WGD and a ploidy between 1.9 and 2.1. This classification scheme resulted in 13 near-haploids (0.17%), 310 low-hypodiploids (4%), 2,905 near-diploids (37.97%), 1,977 polyploids (25.84%) and 2,445 cases with intermediate ploidy ('Other', 31.96%) (Fig. 2.3). The intermediate cases are generally aneuploid, usually (61%) sub-diploid with a median ploidy of 1.88. They cover a broad range of loss of heterozygosity values but do not show evidence of loss of >7 chromosomes throughout their evolution (the low-hypodiploidy threshold). They also include samples with chromosome gains that bring ploidy above 2.1 but are not extreme enough to indicate a WGD event. These intermediate-ploidy samples were

excluded from binary comparisons between polyploid or hypodiploid and near-diploid tumours, but were included in all other analyses, e.g. those analysing ploidy as a continuous trait, examining smaller-scale copy number events, or estimating rates of WGD.

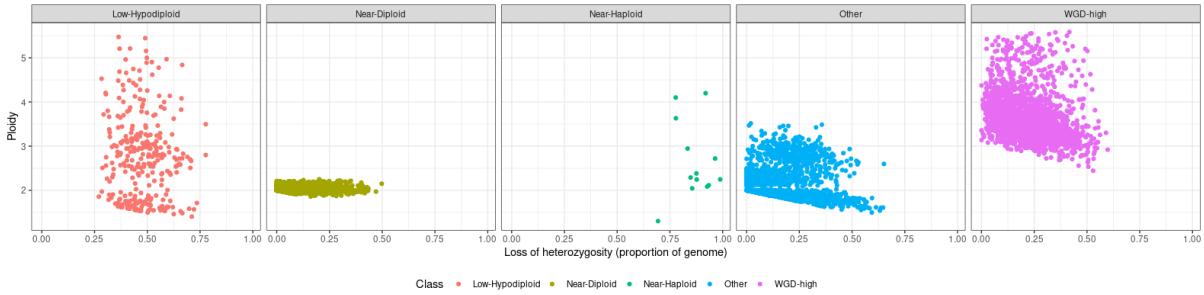


Fig. 2.3 Loss of heterozygosity and ploidy distributions for each of the five ploidy classes I define. 'WGD-high' indicates tumours that have undergone a WGD and do not have a hypodiploid history; these are also referred to as 'polyploid'.

The proportion of cases in each ploidy class varies dramatically between cancer types (Fig. 2.4). For example, 88% of thyroid cancer cases observed here are near-diploid, while only 2.78% of testicular germ cell tumours are. This is also apparent in the sample ploidy distribution of each cancer type. Some cancer types have a single narrow peak at 2 (diploidy, e.g. thyroid [THCA], acute myeloid leukaemia [LAML], thymoma [THYM], prostate [PRAD]). Others are bimodal with a peak at diploidy and another around 3-4, indicating a genome doubling event, and these vary in the relative size of the two peaks, from polyploidy being infrequent (e.g. low-grade glioma [LGG], liver [LIHC]) to it being about as common as diploidy (e.g. oesophageal, ESCA) or even more common (uterine [UCS]). A cancer type that stands out for its unusual ploidy distribution is adrenocortical carcinoma [ACC], which shows no clear distinction between peaks, likely due to its high rates of both hypodiploidy and genome

doubling. However, it should be noted that this dataset does not fully represent all subtypes of a given cancer type – with thyroid cancer, for example, TCGA focused on the papillary subtype but another subtype, Hurthle cell thyroid carcinoma, is much less chromosomally stable, exhibiting frequent near-haploidy (Gopal et al, 2018).

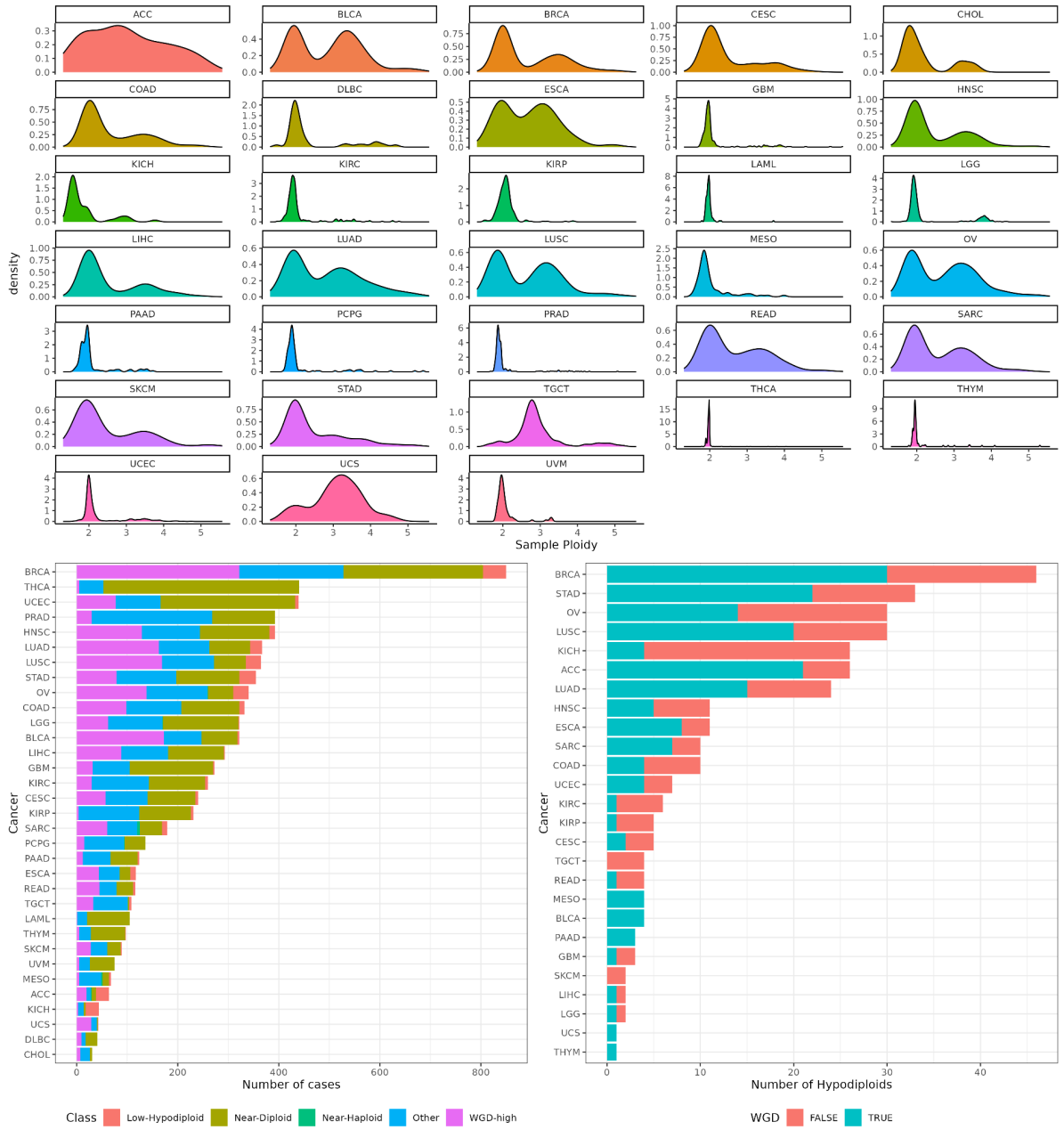


Fig. 2.4 Differences between cancer types in (top) sample ploidy (bottom left) ploidy class proportions and (bottom right) WGD rate among low-

hypodiploids. Cancer type abbreviations are taken directly from official TCGA project names and full names are provided in Appendix Table 1.

The samples at the two extremes of ploidy in this dataset are a near-haploid low-grade glioma with a ploidy of 1.3 and 26 autosomes (18/22 autosomes showing LOH), and a polyploid glioblastoma case with a ploidy of 5.57 and 118 autosomes, including amplification of chromosome 7 to ten copies and LOH of five chromosomes and 9p (Fig. 2.5). In both of these cases, copy number is mostly consistent across whole chromosomes. I note that tumours with significantly higher chromosome counts have been reported in the literature (Mitelman et al, 2025), but the copy number caller ASCAT is unlikely to call karyotypes beyond pentaploidy. Nullisomy is very rare: out of 168,300 autosomes analysed across 7,650 patients, only 7 appear to be nullisomic (copy number zero across at least 90% of their length), and all but one of these (a kidney renal cell carcinoma with nullisomy of chr9) have one or more small segments with a non-zero copy number, suggesting possible translocation (or sequencing error).

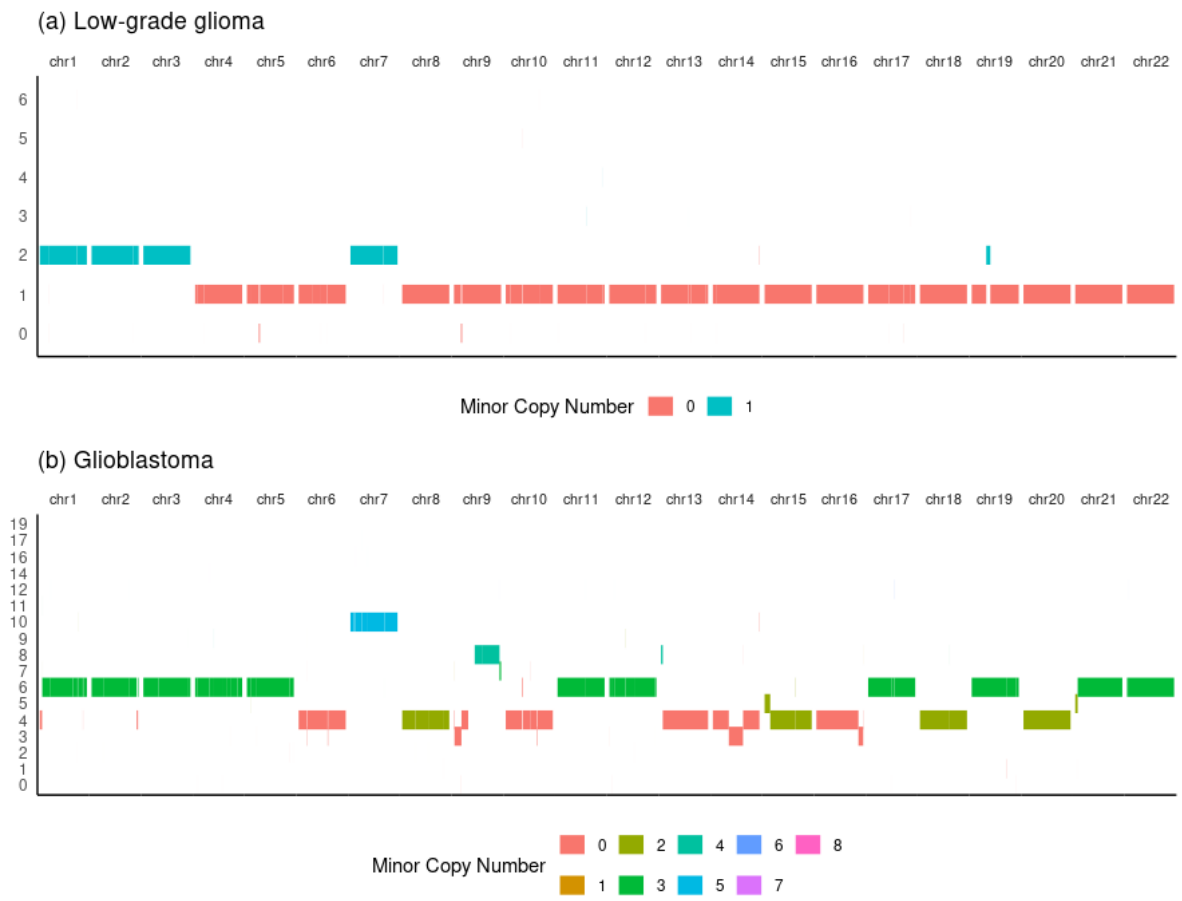


Fig. 2.5 Copy number segment-level visualisation of (a) a near-haploid and (b) a highly polyploid tumour. Total copy numbers are shown on the y-axis, with minor copy numbers shown via colour. A minor copy number of zero indicates loss of heterozygosity.

2.3.2 Relationship between rates of polyploidy and hypodiploidy

There is a significant weak-to-moderate correlation between polyploidy rate and hypodiploidy rate across tissues ($r = 0.38$, $p = 0.036$) when I exclude two cancer types with extremely high rates of hypodiploidy (kidney chromophobe, 59% hypodiploid; adrenocortical carcinoma, 43.8% hypodiploid) (Fig. 2.6). I note also that one outlying cancer type, mesothelioma, has an unusually high near-haploid:low-hypodiploid ratio (3 NH:4 LH, out of 68 total cases). As near-haploid cases have been proposed to arise via a different mechanism (near-

haploidisation), I also computed the correlation between polyploidy and low-hypodiploidy rate (excluding near-haploids) and find that it increases but only slightly ($r = 0.41$, $p = 0.023$), likely because near-haploids are too rare to make a major contribution to the overall hypodiploidy rate in most cancer types.

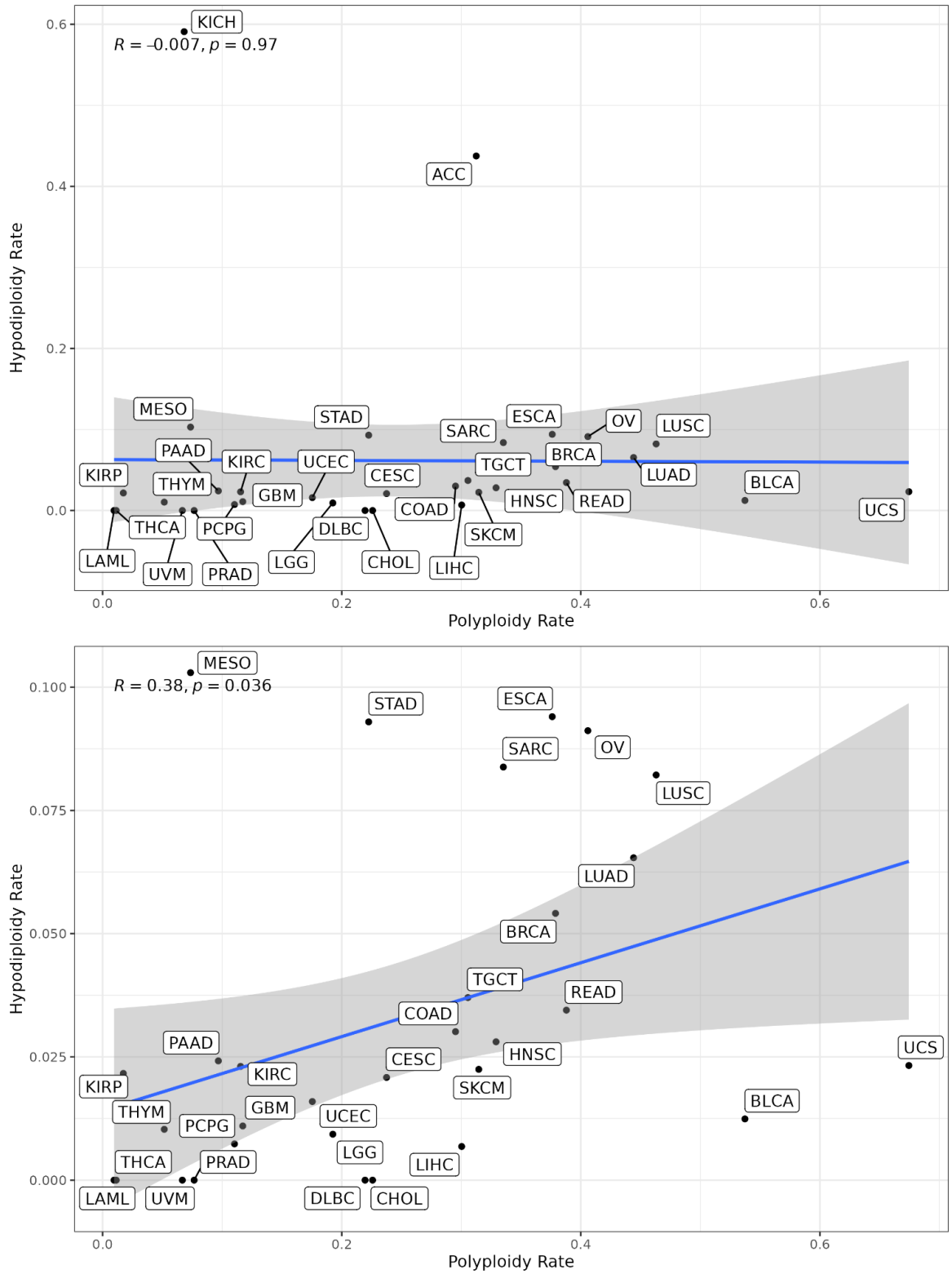


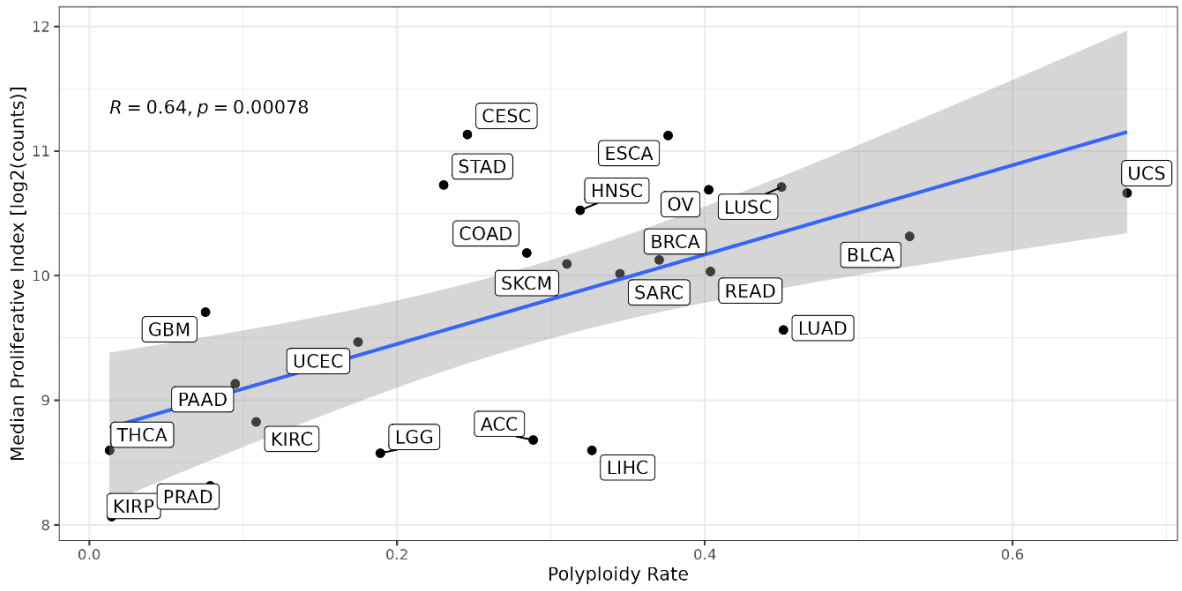
Fig. 2.6 Correlation between polyploidy and hypodiploidy rate across TCGA cancer types, with (top) and without (bottom) outliers. Former hypodiploids are excluded from the definition of polyploids.

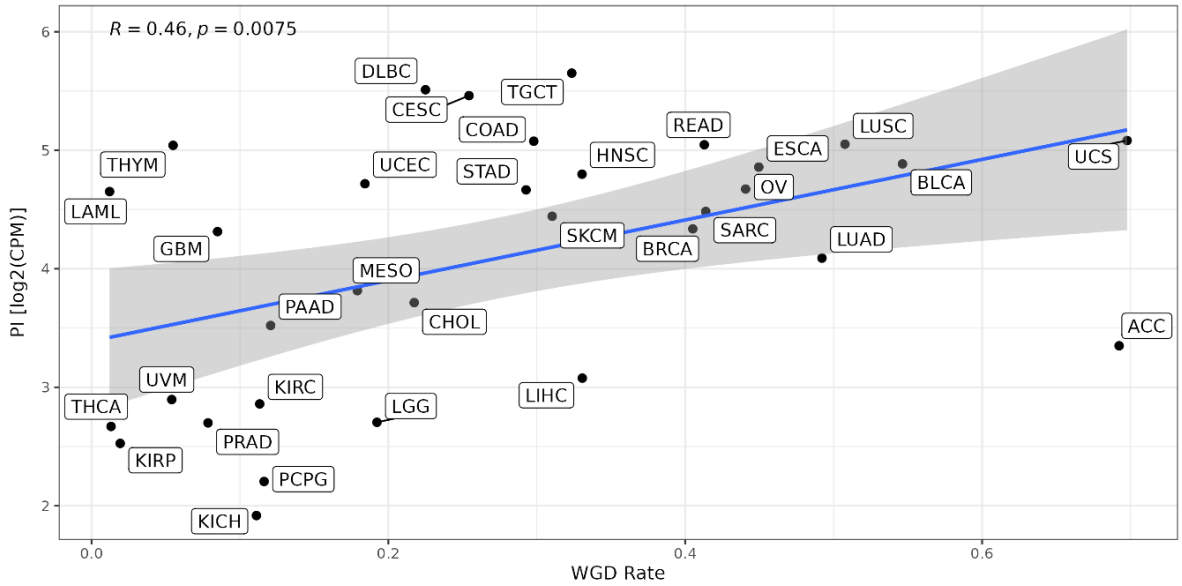
2.3.3 WGD rates across cancer types are partially explained by tissue replication rate

I now turn to the aim of this chapter: identifying the factors that distinguish tumours with extreme ploidy. It has previously been reported that variability in tissue proliferation rate can explain 42% of the variance ($r = 0.65$, $p = 0.0008$) between cancer types in rate of whole genome duplication, suggesting that differences in WGD rate are due largely to the level of opportunity for random mitotic errors and not to distinct antecedent genetic or environmental causes (Bielski et al, 2018). I begin by repeating this analysis, extending it to additional cancer types and to hypodiploids and investigating possible confounding factors.

The proliferative index (PI) of a sample is calculated as the median expression of the top 1% of genes correlated with the proliferation cell nuclear antigen marker (PCNA) in healthy tissues (Ramaker et al 2017, Venet et al 2011). After computing PI for each sample, I compared the median PI per tissue to the WGD rate in that tissue (Fig. 2.7). When I exclude former hypodiploids in order to more closely match Bielski et al's definition of WGD (at least 50% of the genome with a major copy number >= 2), include only the 25 TCGA cancer types they analysed, and examine raw counts, I obtain the same correlation ($r = 0.64$, $p = 0.00078$). When I include the additional eight cancer types available from TCGA and account for all genome doublings, including those on a hypodiploid background, the correlation is still significant but weaker ($r = 0.48$, $p = 0.0047$), suggesting that tissue proliferation rate explains a substantially lower 23% of the variance between tissues in WGD rate. Results are similar whether I analyse

log2(counts) or log2(CPM) ($r = 0.46$). The relationship with hypodiploidy is less clear: after excluding two outliers with extremely high hypodiploidy rates (kidney chromophobe and adrenocortical carcinoma) I find a significant correlation with median tissue PI based on raw counts, but not after normalisation (log2(counts): $r = 0.47$, $p = 0.0079$; log2(CPM): $r = 0.32$, $p = 0.078$).


R = 0.64, p = 0.00078
Median Proliferative Index [log2(counts)]
Polyploidy Rate
CESC
ESCA
STAD
HNSC
OV
LUSC
UCS
COAD
BRCA
BLCA
SKCM
SARC
READ
LUAD
GBM
UCEC
PAAD
THCA
KIRC
LGG
ACC
LIHC
KIRP
PRAD


R = 0.46, p = 0.0075
PI [log2(CPM)]
WGD Rate
DLBC
TGCT
CESC
COAD
READ
LUSC
ESCA
THYM
UCEC
STAD
HNSC
UCS
OV
BLCA
LAML
SKCM
SARC
GBM
BRCA
LUAD
MESO
PAAD
CHOL
ACC
UVM
KIRC
LIHC
THCA
LGG
PRAD
KIRP
PCPG
KICH

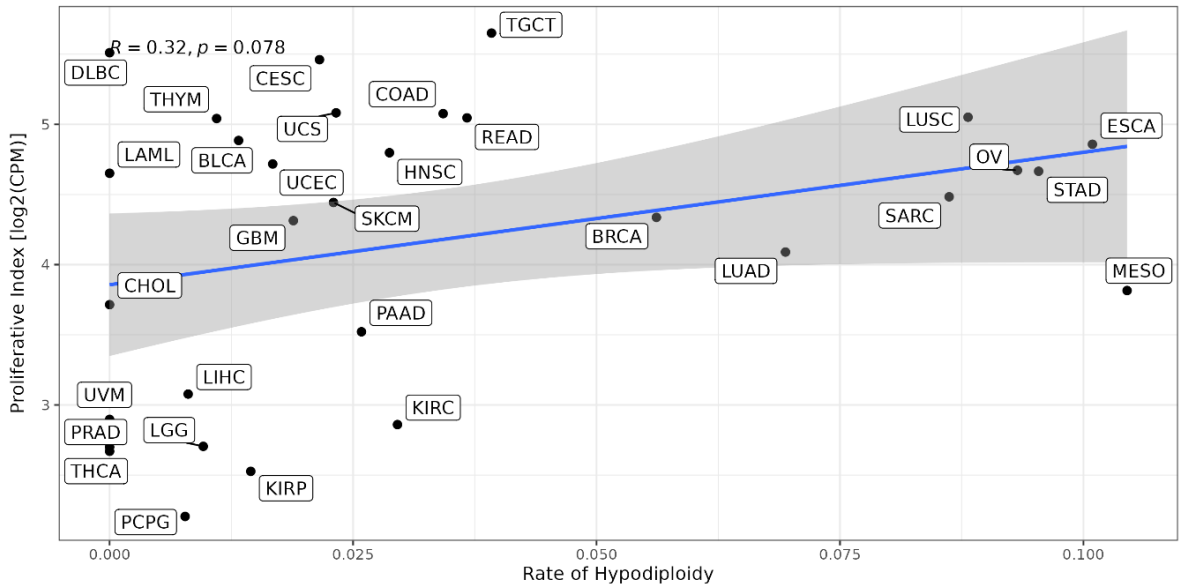

R = 0.32, p = 0.078
Proliferative Index [log2(CPM)]
Rate of Hypodiploidy
TGCT
DLBC
CESC
COAD
THYM
UCS
READ
LUSC
ESCA
LAML
BLCA
HNSC
OV
UCEC
STAD
SKCM
SARC
GBM
BRCA
LUAD
MESO
CHOL
PAAD
LIHC
UVM
KIRC
PRAD
LGG
THCA
KIRP
PCPG

Fig. 2.7 Relationship between tissue proliferation rate and WGD rate across cancer types. (top) To match the analysis in Bielski et al. more closely, I compared the rate of polyploidy to the median proliferative index (PI) for the relevant tissue, calculating PI based on raw counts. For other analyses, I normalised counts to CPM, included additional cancer types available in TCGA, and compared median PI to the rate of WGD **(centre)** and hypodiploidy **(bottom)**.

I next considered whether the link between polyploidy and median tissue PI could be confounded by the inclusion of PIs from polyploid cases in the calculation of the tissue median. If polyploid tumours have higher replication rates, then cancer types with many polyploids will have higher median PI values, even if their tissue of origin is not actually highly proliferative. Polyploids indeed have significantly higher PIs than near-diploids (Wilcoxon two-sided $p < 2.2e\text{-}16$, Fig. 2.8). This effect is significant within each of a wide range of cancer types, including bladder, breast, liver, lung adenocarcinoma, lung squamous, ovarian, prostate, sarcoma, stomach and endometrial cancer (Appendix Fig. 6.2). I find no cancer types in which PI is significantly lower in polyploids. Low-hypodiploid tumours also have significantly higher PI than near-diploids (Wilcoxon two-sided $p < 2.2e\text{-}16$), and in tissue-specific analyses this result is significant in 5/9 of the cancer types with at least ten each of low-hypodiploids and near-diploids (breast, lung adenocarcinoma, lung squamous, stomach, sarcoma; n.s. in colon, esophageal, head-and-neck and ovarian cancer) (Appendix Fig. 6.3).

To rule out confounding of the estimate of a tissue's average replication rate from the polyploid cases themselves, I recalculated median PI for each tissue using only near-diploid tumours and found that the relationship weakens: for polyploidy, r drops from 0.58 to 0.47 using log2(counts) and $r = 0.44$ ($p = 0.01$)

using log2(CPM); for all WGD events, the correlation is only marginally or no longer significant, at r = 0.35, p = 0.048 with log2(counts) and r = 0.32, p = 0.072 with log2(CPM).

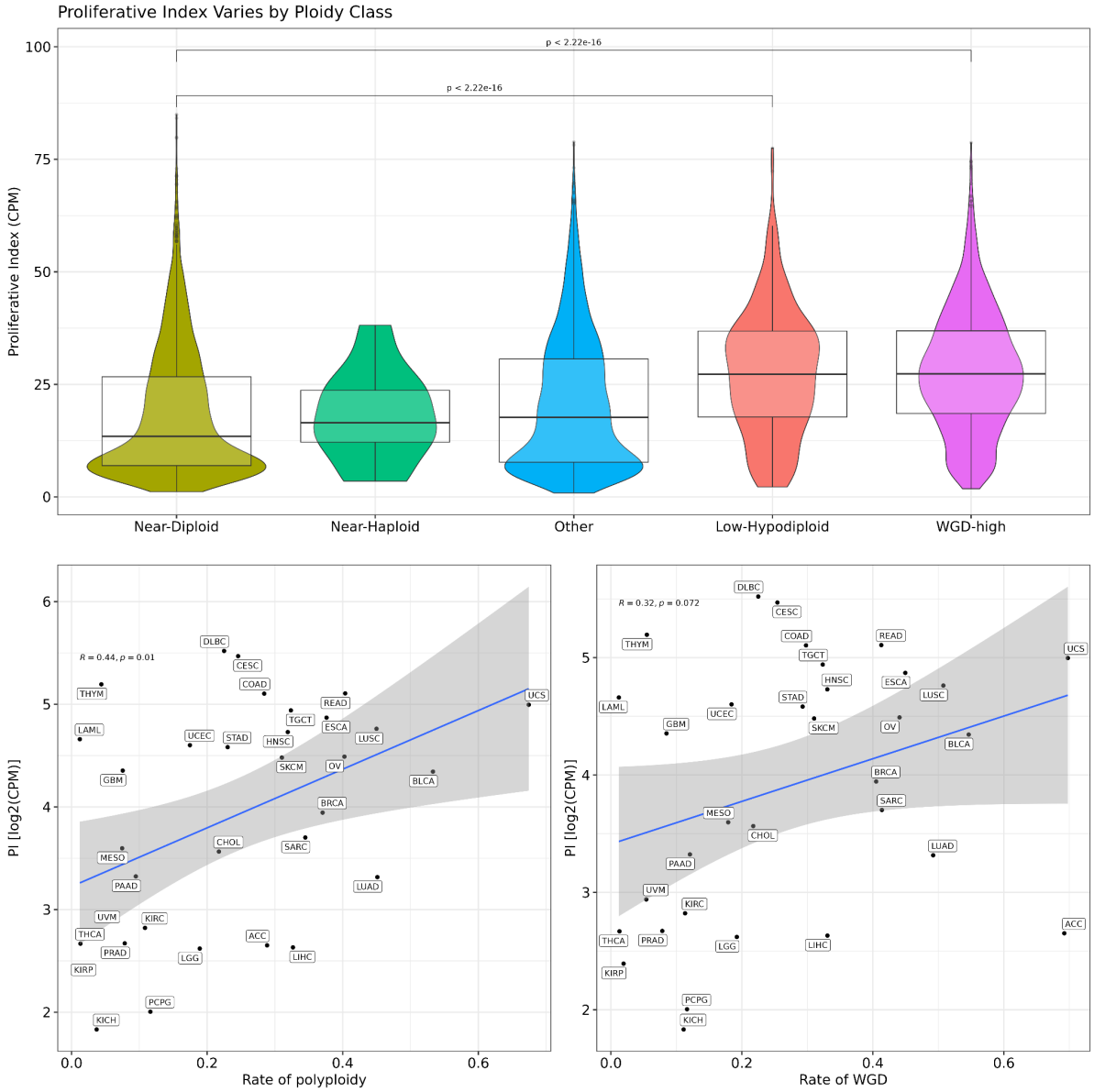


Fig. 2.8 Polyploid tumours have higher PI (top) Comparison of (per-sample) PI between ploidy classes: low-hypodiploid and polyploid tumours have significantly higher PI than near-diploid tumours **(bottom)** Correlation between **(left)** polyploidy rate or **(right)** WGD rate in a tissue and the median proliferative index of that tissue, but with median PI calculated based only on near-diploid tumours.

I hypothesise that part of the previously-reported correlation may be explained by (a) polyploid cells proliferating faster, either due to some polyploidy-intrinsic

effect, due to co-occurring factors (e.g. TP53 mutation) or due to arising at a different stage of tumour evolution, and (b) the PIs of polyploid tumours making a larger contribution to the median PI of tissues in which polyploidy is common. I note on this point that the polyploidy-PI relationship is not explained by TP53 mutation status: regressing PI (cpm) ~ ploidy class + TP53 mutation status + cancer type shows that both polyploidy (regression estimate +178, p = 2.83e-20) and TP53 mutation status (+259, 2.14e-39) are independent positive predictors of PI.

2.3.4 No association between germline mutations and extreme ploidy in patient tumours

If patients differ in their propensity to develop polyploid or hypodiploid tumours, some of that propensity might be reflected in the presence of predisposing germline variants. In the last five years, multiple genome-wide association studies examining traits within cohorts of cancer patients have been published, including two investigating common germline variants implicated in tumour immunosurveillance (Sayaman et al, 2021; Shahamatdar et al, 2020). Motivated partly by a report that the establishment of tumour polyploidy depends on the immune environment (Senovilla et al, 2012), I performed a genome-wide association study on two related traits: WGD status (binary) and tumour ploidy (quantitative). The study was multi-ethnic and pan-cancer due to the relatively limited sample size. I obtained genotype data from Sayaman et al (2021) and performed extensive quality control as described in 2.2. As covariates, I selected age, sex, cancer type and ancestry. WGD is slightly more common in tumours

from female patients (29.3% vs 27.1%, chi-squared p = 0.039), and patients with a WGD-positive tumour tend to be older at diagnosis (63 vs 60 years, Wilcoxon two-sided p < 2.2e-16). For ancestry, I projected population labels from the 1000 Genomes Project onto principal component weights derived from the TCGA data and observed visible population separation up to the 9th principal component (Appendix Fig. 6.3). This led me to include principal components 1-9 as ancestry covariates. The design of the GWAS is laid out in Table 2.2.

Trait	Method	Formula
WGD Status	Logistic regression with Firth fallback	WGD status ~ SNP + cancer type + sex + age + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9
Ploidy	Generalised linear regression	Ploidy ~ SNP + cancer type + sex + age + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9

Table 2.2 Setup of the GWAS for each trait.

I found no SNPs significantly associated with either trait (Fig. 2.9). To increase power, I used MAGMA to combine SNP effects within genes and pathways and found that no genes met the adjusted significance threshold (0.05/18,144) for either trait, nor did any biological pathways.

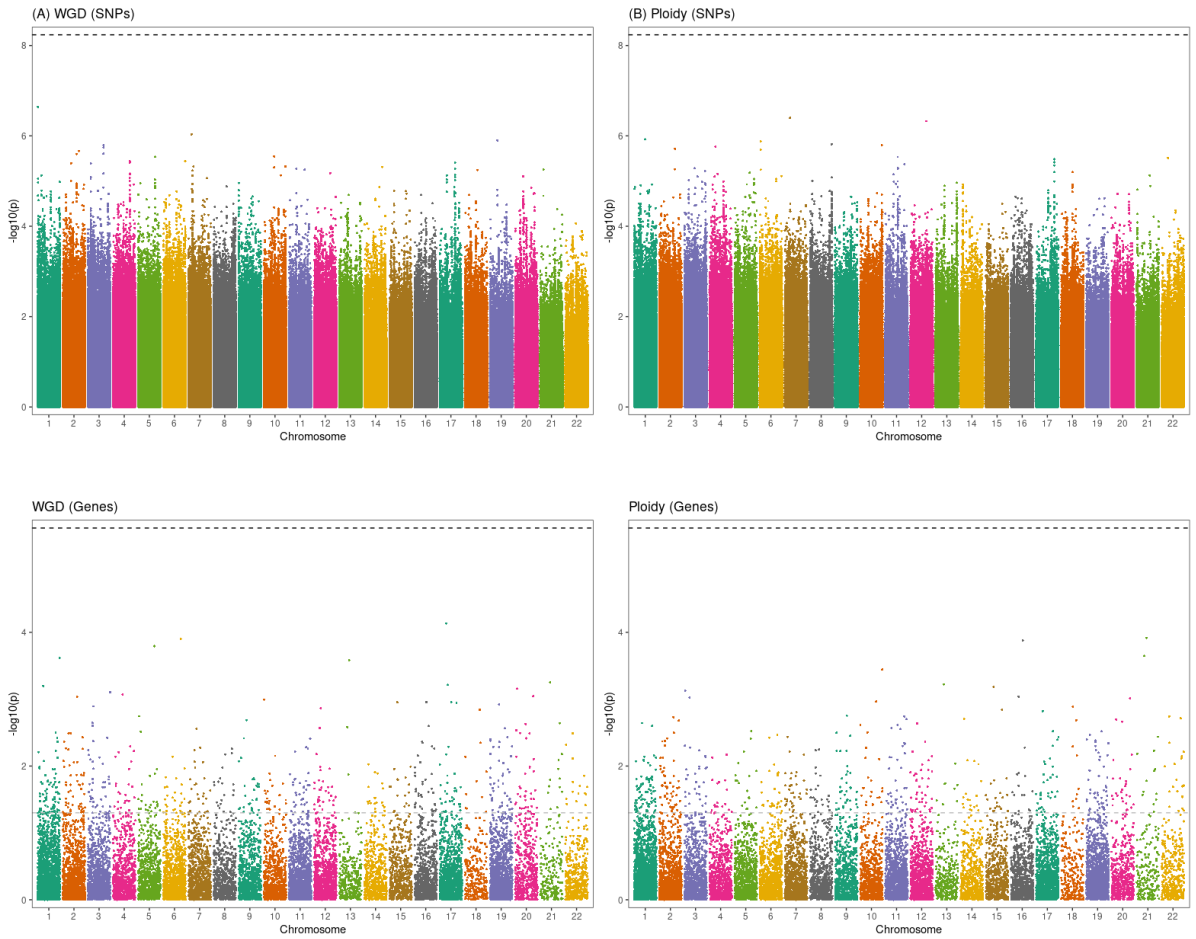


Fig. 2.9 GWAS Results at (top) SNP and (bottom) gene level. Dashed line indicates genome-wide significance threshold.

Quantile-quantile plots of the GWAS p-values show that they are largely uniformly distributed with a small amount of deflation, possibly due to low sample size (Fig. 2.10). Given the relatively limited sample size of this GWAS, I sought to assess how much statistical power there was to detect hits given the sample size using the genpwr R package. At this sample size, I should have 90% power to detect associations for additive SNPs with MAF >= 10% and OR >= 1.5, or effect size >= 0.15 (in units of ploidy). I suggest that there are unlikely to be SNPs (genotyped or in linkage disequilibrium with genotyped SNPs) meeting

these criteria, especially since no effects were found at the gene or pathway level either.

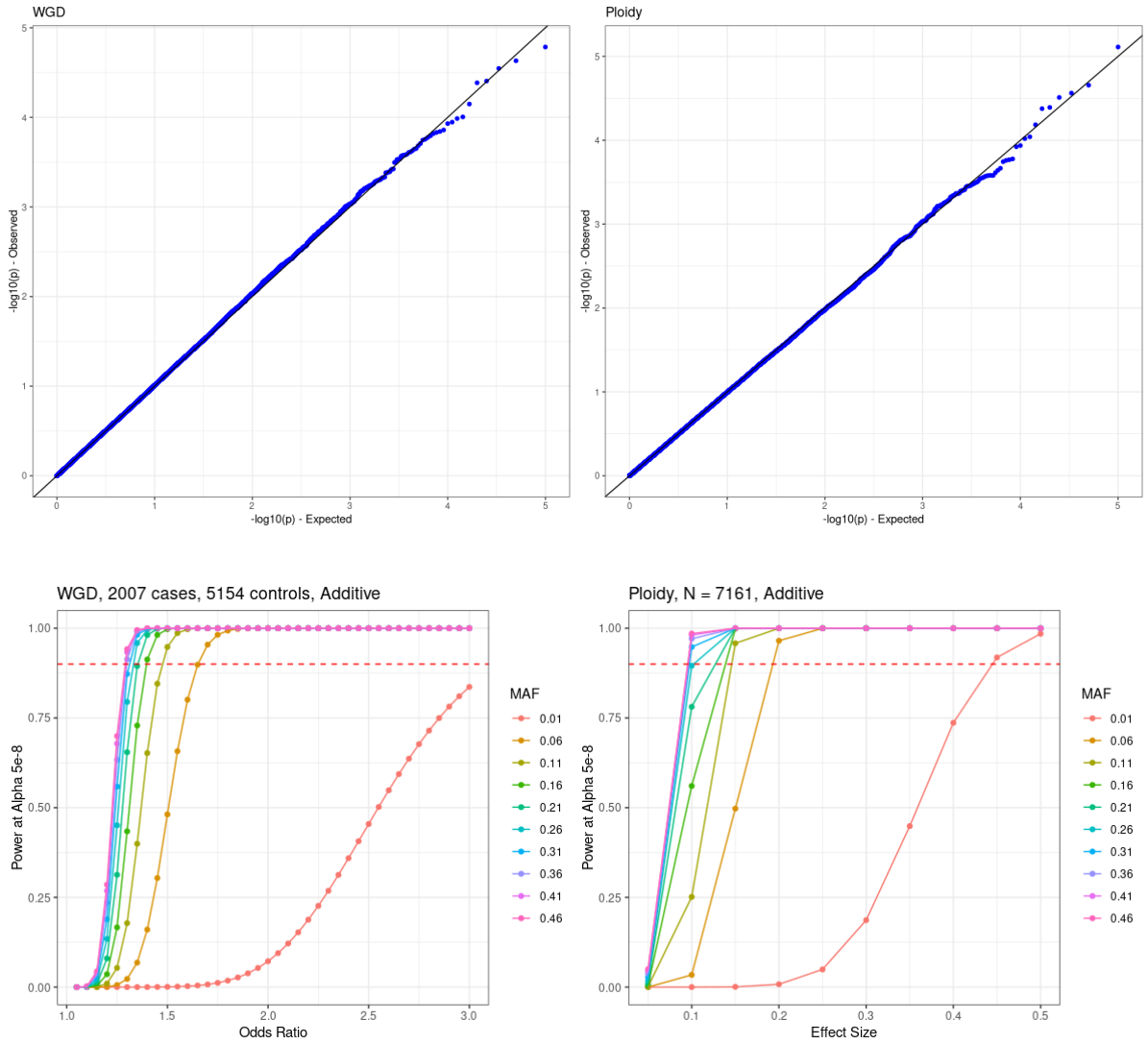


Fig. 2.10 Power analysis (top) quantile-quantile plots of observed GWAS p-values vs expected p-values for each trait **(bottom)** modelling of power to detect SNPs of different odds ratios (for WGD status) or effect sizes (for ploidy), assuming that SNPs have additive effects. MAF indicates minor allele frequency.

Corroborating this conclusion, I find that the common-variant heritability of WGD status and ploidy in tumour samples is essentially zero (Fig. 2.11). I do this using two methods, one based on the GWAS summary statistics (LDSC) and one based on a genetic relatedness matrix (GCTA-GREML), and their results are

concordant. I also note that the LD Score regression intercepts are very close to 1 (1.0034 and 0.9975 for WGD status and ploidy, respectively), suggesting that I successfully controlled for population stratification.

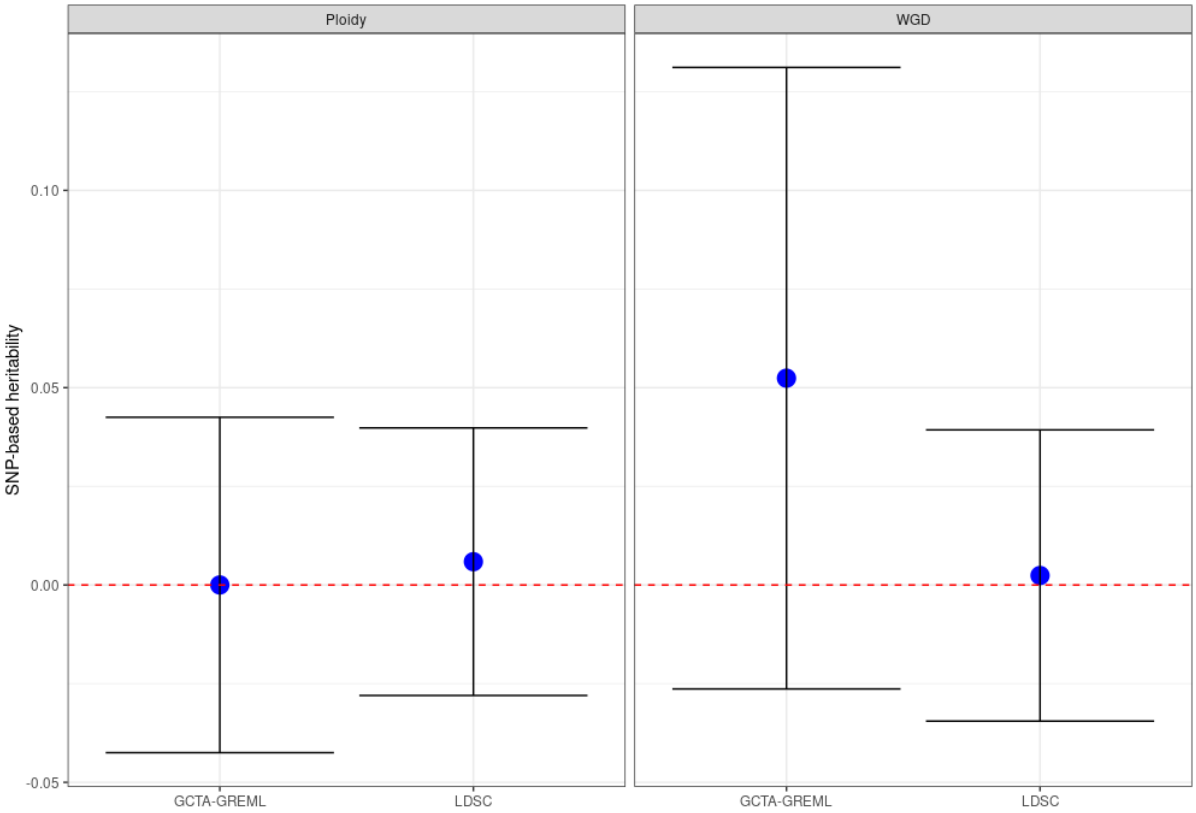


Fig. 2.11 Estimates of SNP-based heritability for tumour WGD status and ploidy among patients in The Cancer Genome Atlas using LD Score Regression and GCTA-GREML. Results for WGD status obtained using GCTA-GREML are presented on the underlying (liability) scale whereas from LDSC they are presented on the observed scale.

Finally, reasoning that mutations affecting ploidy are likely to be rare due to negative selection, I also searched for associations between rare germline mutations and WGD status using a set of prioritised cancer-related germline variants obtained from Huang et al (2018), but found no significant associations with polyploidy or hypodiploidy. Germline mutations in one gene, TP53, have previously been linked to polyploidy in advanced cancer patients (Bielski et al,

2018); it is possible I did not obtain the same result because there were relatively few germline TP53 mutations (Li-Fraumeni syndrome) in this dataset (0 for hypodiploids and 3 for polyploids).

2.3.5 Somatic mutations in TP53, ARID1A and PTEN are associated with ploidy

Next, I searched for associations between somatic mutations and ploidy. I first compared mutation rates for each gene between diploids, polyploids and hypodiploids pan-cancer, using mafCompare to conduct a Fisher's test on 2x2 contingency tables and including only genes with at least 5 mutations across both ploidy classes being tested. I find two genes significantly differentially mutated between hypodiploids and near-diploids after multiple testing correction: TP53, which is upmutated (OR 8.75, P_{adj} 8.2e-58), and BRAF, which is downmutated (OR = 0.169, P_{adj} 3.06e-5). The BRAF finding is likely to be an artefact: BRAF mutations are most common in papillary thyroid carcinoma, which is chromosomally stable. Indeed, in the tissue-specific analyses, only TP53 is significant, in BRCA (OR 45.45), STAD (OR 12.26) and LUAD (OR 12.17). I note that TP53 is likely to be relevant in other tissues also but limited by sample size among the hypodiploids.

In polyploids, I find thousands of genes differentially mutated (mostly downmutated) compared to diploids in both the pan-cancer and tissue-specific analyses. Suspecting that the simple Fisher's test is likely to be confounded by total sample mutation rate, I investigated this further and found that polyploid

tumours have significantly higher mutation rates: median 46 in near-diploids vs 103 in polyploid tumours (Wilcoxon two-sided $p < 2.2e-16$, Fig. 2.12). Polyploid tumours have significantly higher mutation counts in 11/21 cancer types, but there are notable counterexamples (UCEC and COAD), in which polyploid tumours have significantly lower total mutation counts.

Interestingly, low-hypodiploid tumours also have significantly higher mutation counts than near-diploids (median 80.5, Fig. 2.12), even those which have not undergone a WGD event (61, $p = 0.0048$). Near-haploids have low mutation counts (median 32), with intermediate-ploidy cases falling in between near-diploids and low-hypodiploids with 58. In the nine cancer types with at least ten each of near-diploid and hypodiploid cases, hypodiploids have higher mutation rates in four (BRCA, LUSC, LUAD, SARC) while in the others there is no significant difference. Restricting to the three tissues with at least ten non-doubled hypodiploids (breast, ovarian and stomach cancer), there is a significant difference only in breast cancer (higher mutation rate in hypodiploids, Wilcoxon two-sided $p = 0.0092$, Appendix Fig. 6.5).

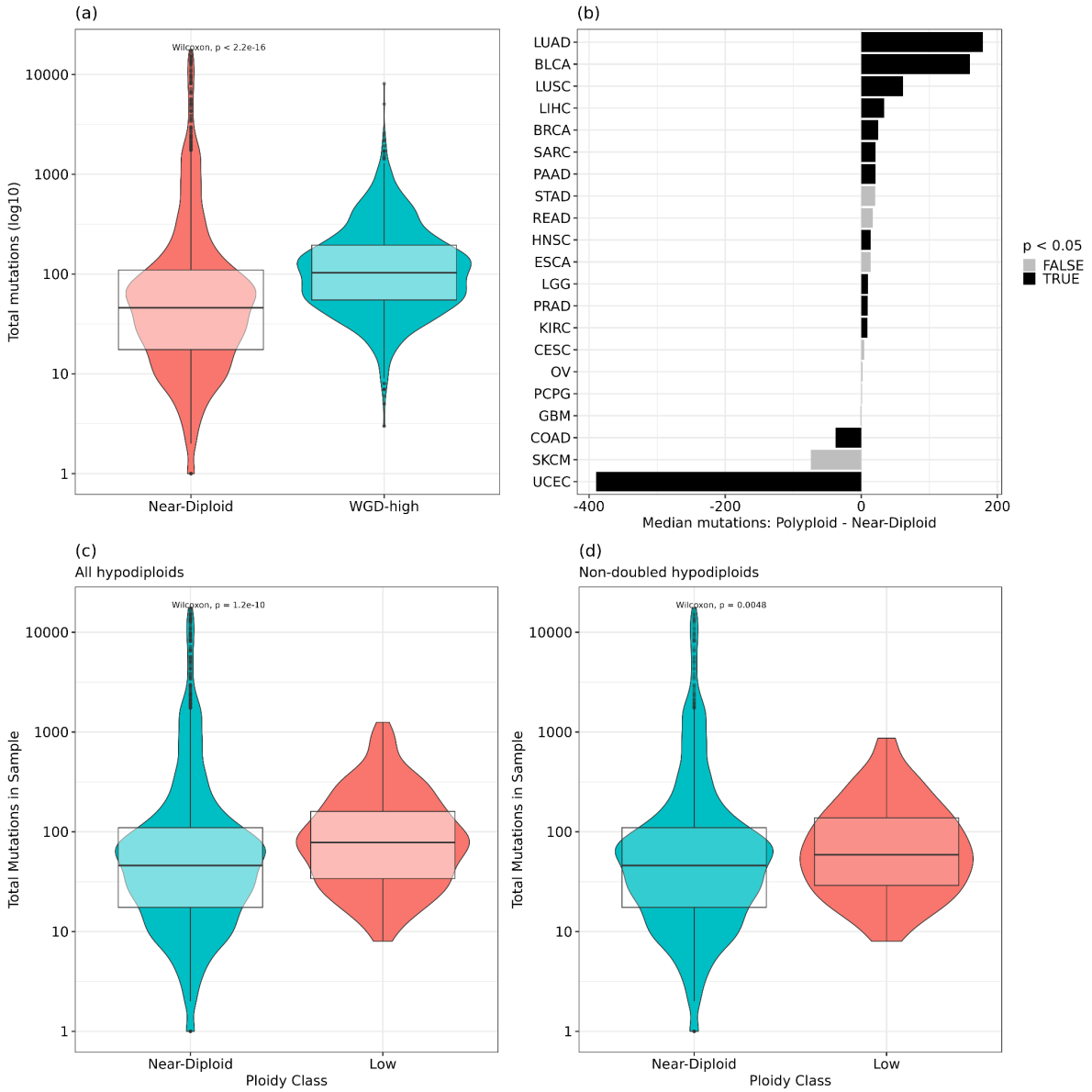


Fig. 2.12 Total mutation count is higher in polyploids and hypodiploids than diploids (a) Comparison of all polyploid vs all near-diploid samples **(b)** Tissue-specific comparison: for each tissue, median mutation count in near-diploids was subtracted from median mutation count in polyploids **(c)** Comparison of all hypodiploid (both low-hypodiploid and near-haploid) samples vs all near-diploid samples **(d)** excluding hypodiploid samples that had undergone WGD. The y-axes in (a), (c) and (d) are log10-transformed.

To account for differences in total mutation count by ploidy, I repeated these analyses using an alternative method inspired by Quinton et al (2021)'s analysis of somatic mutations in WGD+ tumours: logistic regression controlling for total

mutation count in the sample. Following Quinton et al, I restricted the analyses to known cancer driver genes, here using the COSMIC database (Tier 1 and 2), and included only genes with at least five mutations across the two ploidy classes (near-diploids + polyploids or near-diploids + hypodiploids). I performed logistic regression separately for each gene as follows:

$$\text{ploidy_class} \sim \text{mutated} + \text{total mutations} + \text{cancer type}$$

I find 23 genes differentially mutated between polyploids and near-diploids after multiple testing correction (Fig. 2.13a). Mutations in the majority of these genes are negatively associated with polyploidy, with the exceptions of TP53, CSMD3 and LRP1B. As expected, TP53 is the strongest hit by far (odds ratio 5.47, p_{adj} = 1.77e-76). Among the hypodiploids, only TP53 is significantly differentially mutated after correction (odds ratio 12.95, p_{adj} = 6.08e-34). There are 27 nominal associations, 20 upmutated (including LRP1B) and 7 downmutated, including four significantly downmutated in polyploids (PIK3CA, RPL22, ARID1A, CDH1).

I then performed this analysis tissue-specifically, i.e. for each gene and tissue:

$$\text{ploidy_status} \sim \text{mutated} + \text{total mutations}$$

Here I find two genes upmutated in polyploids (TP53 in seven cancer types and the serine/threonine phosphatase PPP2R1A in endometrial cancer) and three downmutated (ARID1A, PTEN and CTNNB1, all in endometrial cancer, Fig.

2.13b). These results are consistent with those of Quinton et al. Gene Ontology enrichment analysis for the (pan-cancer and tissue-specific) polyploidy hits shows no enriched categories when performed against a background of COSMIC genes. In hypodiploids, again, only TP53 is significantly differentially mutated after multiple testing correction (in breast and stomach cancer).

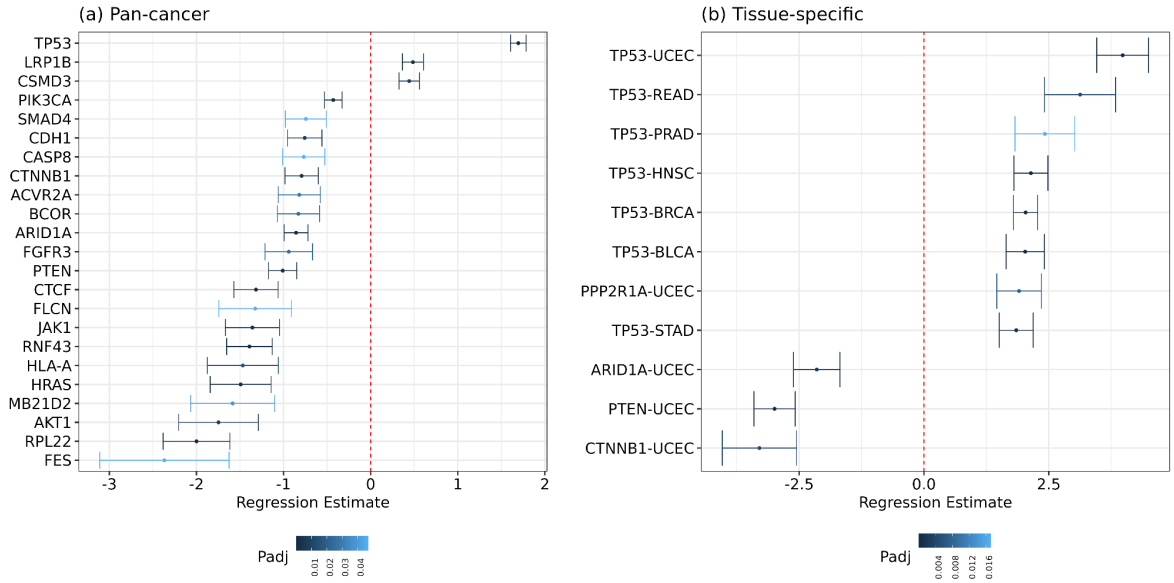


Fig. 2.13 Logistic regression of polyploidy on mutation status of cancer genes **(a)** pan-cancer **(b)** within each tissue. Cancer genes were taken from the COSMIC Cancer Gene Census (Tier 1 and 2); the regression controlled for total mutation count in the same and (in the pan-cancer analysis) cancer type. Genes with significant associations at a Benjamini-Hochberg FDR < 0.05 are shown.

Mutations in three genes appear as significantly associated with polyploidy across all four analyses: one positively (TP53) and two negatively (PTEN and ARID1A) (Fig. 2.14). Hypodiploidy and polyploidy share a consistent association with TP53 mutations across cancer types. ARID1A is consistently significantly downmutated in polyploids, and nominally downmutated ($p < 0.05$, $OR < 1$) in hypodiploids in three of the four analyses (pan-cancer Fisher's exact test, tissue-specific Fisher's exact test in STAD and UCEC, and pan-cancer logistic regression).

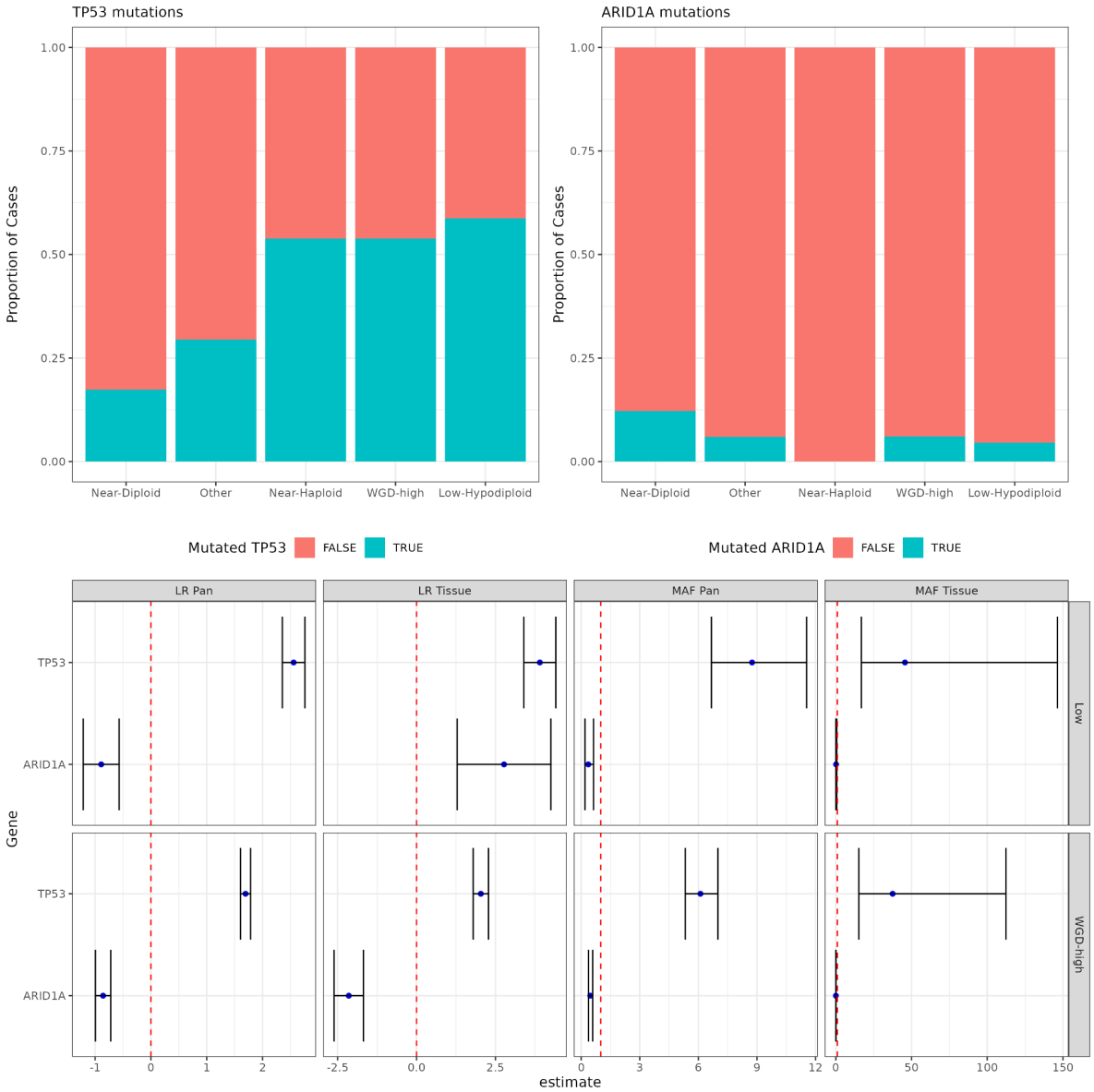


Fig. 2.14 Genes associated with both polyploidy and hypodiploidy across multiple analyses: TP53 (positively) and ARID1A (negatively). (top) Proportion of cases in each ploidy class with TP53 or ARID1A mutations **(bottom)** results of the four different association tests. LR Pan: pan-cancer logistic regression controlling for total mutations in sample and cancer type; LR Tissue: tissue-specific logistic regression controlling for total mutations in sample; MAF Pan: pan-cancer mafCompare; MAF Tissue: tissue-specific mafCompare. For the logistic regressions, x-axis values are regression estimates for the ploidy term and the red line is at 0, with the error bars indicating standard error; for the mafCompare results, the values are odds ratios, with the red line at 1 and the error bars indicating confidence intervals.

2.3.6 Hypoxia is strongly associated with extreme ploidy across tissues

Hypoxia has been experimentally shown to induce cell fusion and/or polyploidy in colon cancer cells, melanoma cells and healthy cardiomyocytes, and is correlated with the proportion of the genome involved in copy number alterations across multiple TCGA tumours (Bhandari et al, 2019). To determine whether this association applies to both polyploid and hypodiploid tumours, I obtained hypoxia scores for each TCGA sample calculated by Bhandari et al using expression signatures from Buffa et al (2010). I was able to assign ploidy calls to 5,384 of these samples across 20 cancer types, including 2,011 near-diploid cases, 1,492 polyploids, 172 hypodiploids and 1,709 intermediate-ploidy cases. As expected, ploidy-aberrant tumours are much more hypoxic than near-diploid or intermediate-ploidy tumours (Fig. 2.15a). In terms of continuous ploidy, hypoxia is lowest in tumours with a ploidy of exactly 2 (diploidy) and increases with aneuploidy in both directions. Interestingly, hypoxia is lower among tetraploids than among near-triploids or near-pentaploids (Fig. 2.15b). This suggests a link with smaller-scale chromosomal instability and not just clean ploidy changes, which aligns with the known role of hypoxia in suppressing homology-directed repair (Kaplan & Glazer, 2019). Among the ploidy-aberrant tumours, hypodiploid tumours are significantly more hypoxic than polyploid tumours (Wilcoxon two-sided $p = 0.013$). Near-haploid tumours appear to have low hypoxia levels, but there are only two samples available so I have excluded them from this analysis.

In a tissue-specific analysis, I find that polyploids are significantly more hypoxic than near-diploids (Wilcoxon two-sided $p_{adj} < 0.05$, Fig. 2.15c) in 8/20 cancer types and significantly less hypoxic in one, ovarian cancer ($p_{ad} = 0.0429$). Of the four cancer types with hypoxia data and at least 10 low-hypodiploid cases, the low-hypodiploid tumours are more hypoxic than near-diploid tumours (Wilcoxon two-sided $p_{adj} < 0.05$) in three (BRCA, LUAD, LUSC, Fig. 2.15d). Interestingly, hypoxia is decreased in both polyploid and low-hypodiploid ovarian cancers, but the difference is not significant among low-hypodiploids ($p = 0.14$) despite a reasonable sample size ($N = 30$).

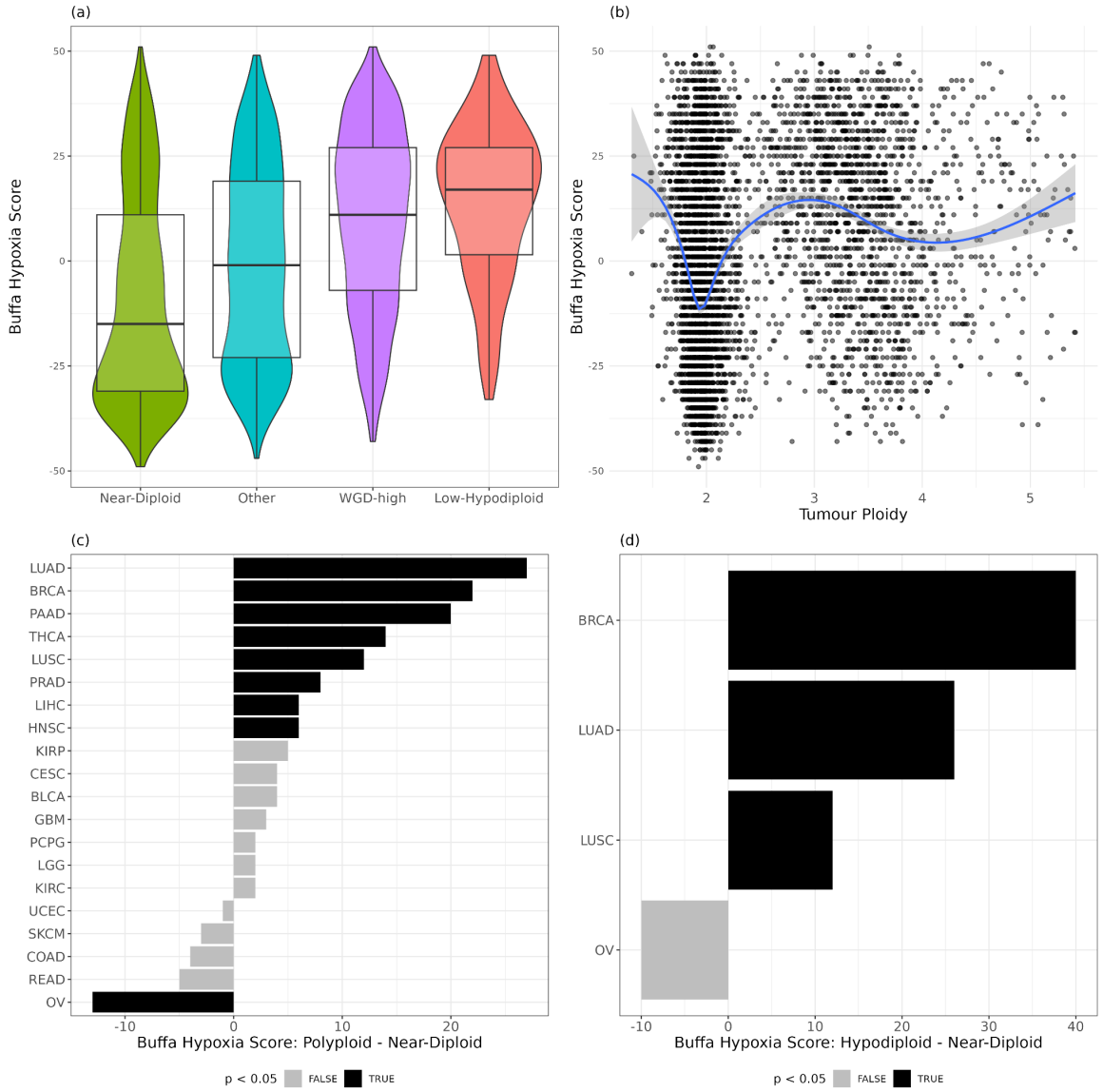


Fig. 2.15 Ploidy-aberrant tumours are more hypoxic than near-diploid tumours **(a)** Comparison between hypoxia levels across ploidy classes **(b)** Hypoxia varies continuously with tumour ploidy **(c)** Tissue-specific comparison of median hypoxia score in polyploids vs near-diploids **(d)** Tissue-specific comparison of median hypoxia scores in low-hypodiploids vs near-diploids. Hypoxia scores were computed for the TCGA dataset by Bhandari et al (2019) based on the Buffa hypoxia expression signature.

Because I am interested in assessing hypoxia as a potential cause of extreme ploidy in patient tumours, I sought to exclude the possibility that these results can be explained by reverse causality (ploidy-aberrant tumours increasing hypoxia in their environment or overexpressing hypoxia-related genes for other reasons). To do this, I calculated the median hypoxia score in each tissue and

compared it to the rates of polyploidy and hypodiploidy in that tissue, finding that they are strongly correlated (polyploidy r = 0.63, p = 0.0027; hypodiploidy r = 0.51, p = 0.023, Fig. 2.16). The correlations decrease only slightly after excluding polyploid or hypodiploid tumours (respectively) from the calculation of median hypoxia score per tissue (r = 0.6, r = 0.49). To assess whether this finding is robust to the choice of expression signature used to calculate hypoxia (Buffa), I obtained hypoxia scores based on seven additional signatures from Bhandari et al and recomputed the correlations for each one. The correlation is significant and positive in 5/8, including all three signatures analysed by Bhandari et al (Buffa, Ragnum and Winter), and the median correlation is 0.59 for polyploidy and 0.52 for hypodiploidy. Based on the Ragnum signature, which has been externally validated using the tumour hypoxia marker pimonidazole (Ragnum et al, 2015), the correlation for polyploidy rate reaches r = 0.754 (0.726 after exclusions) and for hypodiploidy rate r = 0.628 (0.629). Thus I find that hypoxia is strongly linked to both polyploidy and hypodiploidy, and that–while this is an associative analysis–the association cannot be explained by reverse causality.

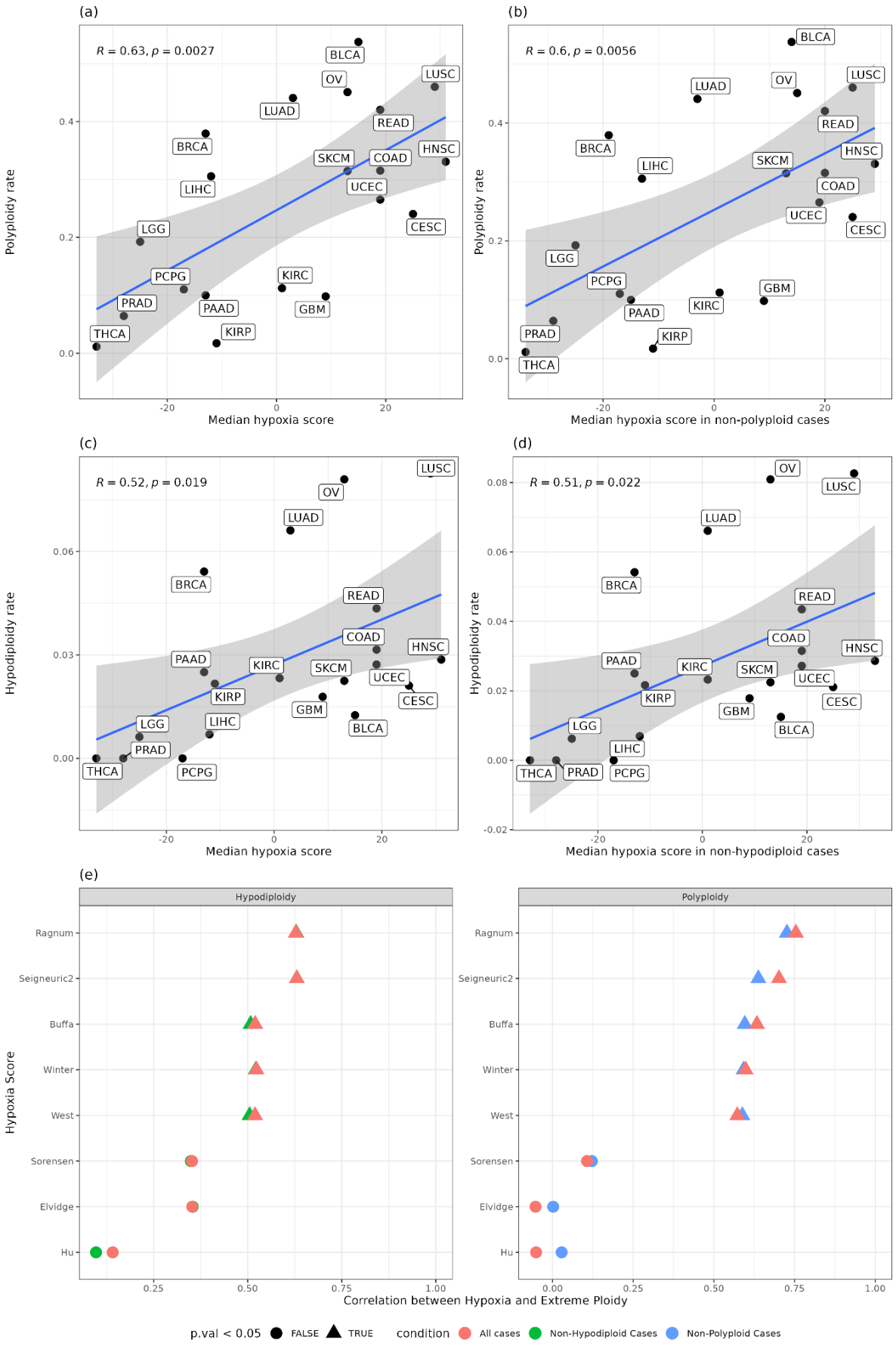

(a)
R = 0.63, p = 0.0027
Polyploidy rate
Median hypoxia score
(b)
R = 0.6, p = 0.0056
Polyploidy rate
Median hypoxia score in non-polyploid cases
(c)
R = 0.52, p = 0.019
Hypodiploidy rate
Median hypoxia score
(d)
R = 0.51, p = 0.022
Hypodiploidy rate
Median hypoxia score in non-hypodiploid cases
(e)
Hypodiploidy
Polyploidy
Hypoxia Score
Ragnum
Seigneuric2
Buffa
Winter
West
Sorensen
Elvidge
Hu
Correlation between Hypoxia and Extreme Ploidy
p.val < 0.05 FALSE TRUE
condition All cases Non-Hypodiploid Cases Non-Polyploid Cases

Fig. 2.16 Relationship between hypoxia and extreme ploidy across cancer types (a) Correlation between polyploidy rate and median hypoxia score (Buffa gene signature) in the relevant cancer type **(b)** adjusted to remove polyploid samples from the calculation of median hypoxia score **(c)** correlation between low-hypodiploidy rate and median hypoxia score (Buffa) **(d)** adjusted to remove low-hypodiploid samples from calculation of the median **(e)** Correlations between median tissue hypoxia and rate of hypodiploidy (left) or polyploidy (right) based on each of the eight hypoxia scores computed by Bhandari et al from different hypoxia gene signatures; significant correlations are indicated by triangles, adjusted hypoxia scores are indicated by blue for polyploids and green for low-hypodiploids.

2.3.7 Smoking is associated with hypoxia in lung cancer, but not with ploidy aberrations

I hypothesised that smoking might lead to ploidy aberrations, either by inducing hypoxia or—as has been suggested by Was et al (2022)—by causing DNA damage. Smoking has previously been linked to chromosomal aberrations in blood lymphocytes from healthy subjects, with both smoker status and smoking intensity predicting an increase in chromosomal aberrations (Farkas et al, 2021). For eight smoking-related cancers that also had hypoxia data available (bladder, cervical, head-and-neck squamous, kidney (clear-cell and papillary cell), lung (adenocarcinoma and squamous cell) and prostate), I downloaded smoking history and compared hypoxia levels between current smokers and never smokers. I find that smoking is nominally associated with increased hypoxia in both lung adenocarcinoma and lung squamous cell carcinoma, but this effect persists after multiple testing correction only in lung adenocarcinoma (Wilcoxon two-sided $p_{adj} = 0.036$, Fig. 2.17). I considered that intensity of smoking might predict a further increase in hypoxia, but there is no significant correlation between hypoxia and pack-years (among ever-smokers) in any cancer type.

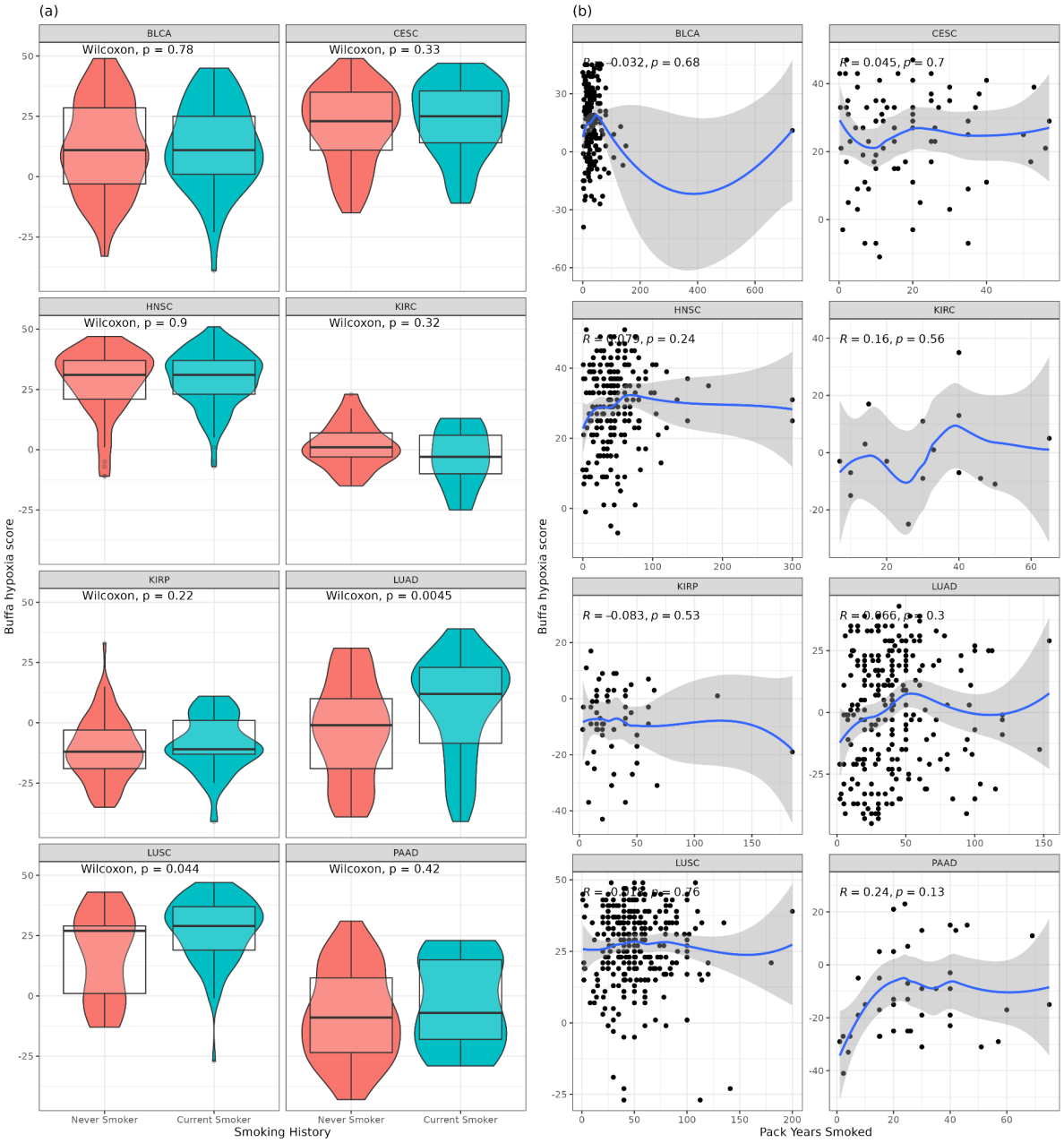


Fig. 2.17 Smoking is associated with increased hypoxia in lung adenocarcinoma (a) Comparison between hypoxia scores in tumours from current smokers vs never smokers **(b)** No relationship between tumour hypoxia and intensity of smoking as measured by pack years smoked.

Does the association with hypoxia translate to a link between smoking and aberrant ploidy? I find that rates of polyploidy (but not hypodiploidy) are higher in smokers in bladder, cervical, head-and-neck, and lung cancers, increasing e.g. in lung squamous cell carcinoma from 30.8% to 49.5%. However, these

differences are not statistically significant (chi-squared $p > 0.05$ for each cancer type). Smoking is also not significantly related to overall sample ploidy (Wilcoxon two-sided $p > 0.05$ for each cancer type). When I regress polyploid or hypodiploid status pan-cancer on smoking history (current vs never smoker) and control for age, gender and self-reported race, I find that the tumours of current smokers are 2.1x as likely to be polyploid as those of never smokers, while being only 0.55x as likely to be hypodiploid. However, both of these effects disappear when I control for cancer type. The inclusion of hypoxia score in the regression does not make a difference.

I then considered, based on the results of Farkas et al (2021), that smoking may be linked not to gross ploidy abnormalities but to lower-level chromosomal instability or smaller copy number alterations. I thus counted the number of copy number segments per sample as a proxy for the number of copy number alteration events or the "brokenness" of the chromosomes (Fig. 2.18c) and found that it is increased among smokers in lung adenocarcinoma ($p_{adj} = 0.0053$) and head and neck squamous cell carcinoma ($p_{adj} = 0.0053$). Linear regression controlling for age, gender, self-reported race and cancer type confirms this finding: smokers have 16.7 more segments than never-smokers ($p = 0.00612$), against an overall median of 153 segments in the tumours included in this analysis. As with hypoxia, however, pack-years are not predictive of the number of copy number segments. In this analysis, inclusion of hypoxia as a covariate renders smoking status non-significant ($p = 0.0632$), although the direction remains consistent (+12 segments), and I note that filtering to samples with

hypoxia data reduces the significance level of smoking status even before hypoxia is added as a covariate (estimate +14.6, p = 0.0246).

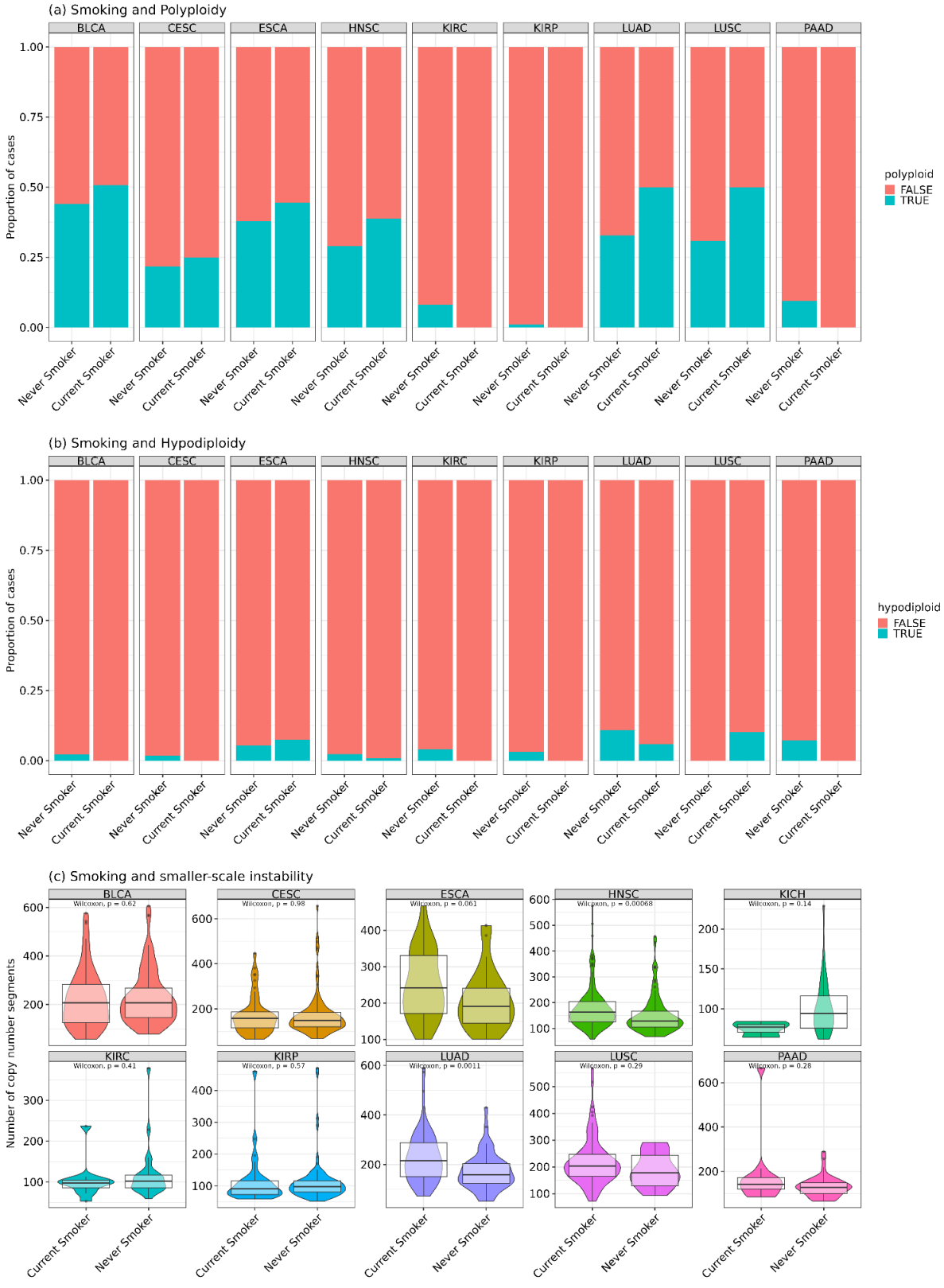


Fig. 2.18 Relationship between smoking and copy number instability. Smoking is not significantly associated with **(a)** polyploidy or **(b)** hypodiploidy, but is associated with the total number of copy number segments in the sample in lung adenocarcinoma and head and neck squamous cell carcinoma.

2.3.8 Mutational signatures associated with smoking are enriched in polyploid tumours

Mutational signature analysis offers insight into environmental exposures I have not specifically tested for here. I used SigProfilerExtractor to assign COSMIC mutational signature activities to each sample based on mutational context matrices created by the PCAWG project. Logistic regression controlling for patient age, cancer type and total mutation count revealed that polyploid tumours are enriched for three mutational signatures (SBS40a, SBS4, and SBS25) and depleted for two (SBS1 and SBS44). SBS1 is a clock-like signature and SBS44 is associated with defective mismatch repair and microsatellite instability; depletion for this signature matches the known mutual exclusivity between WGD and microsatellite instability found in advanced cancers (Bielski et al, 2018) and in this TCGA dataset (Quinton et al, 2021). A tissue-specific analysis additionally found that SBS5 is enriched in polyploid breast cancer. SBS4 is the canonical smoking-associated signature, and SBS5 has also been linked to smoking, albeit more strongly with age (Huang et al, 2022). SBS25 has an unclear aetiology, but the COSMIC database notes that it is associated with chemotherapy exposure in Hodgkin's lymphoma cells, which might track with the ability of certain chemotherapy drugs to induce polyploidy in cell lines. However, it is no longer significantly associated with polyploidy when I additionally control for gender and race. The aetiology of signature SBS40a is

unknown, and could potentially represent a new mutational process particular to polyploid cells or an unknown exposure that promotes polyploidy. No mutational signatures were significantly associated with hypodiploidy after multiple testing correction.

2.3.9 Viral infection is nominally associated with polyploidy in a TP53-dependent manner

Finally, I investigate the extent to which viral infection explains polyploidy in patient tumours from TCGA and test whether this link extends to hypodiploidy. For my initial analysis, I obtained data on the presence of five viruses (HPV16, HPV18, HBV, CMV, EBV) in cervical, head-and-neck, liver, and stomach cancer from Chen et al (2021). Logistic regression controlling for age, sex and race resulted in no significant associations except a negative one between HPV16 and polyploidy in head-and-neck squamous cell carcinoma (Fig. 2.19, $p = 0.004$ after multiple testing correction).

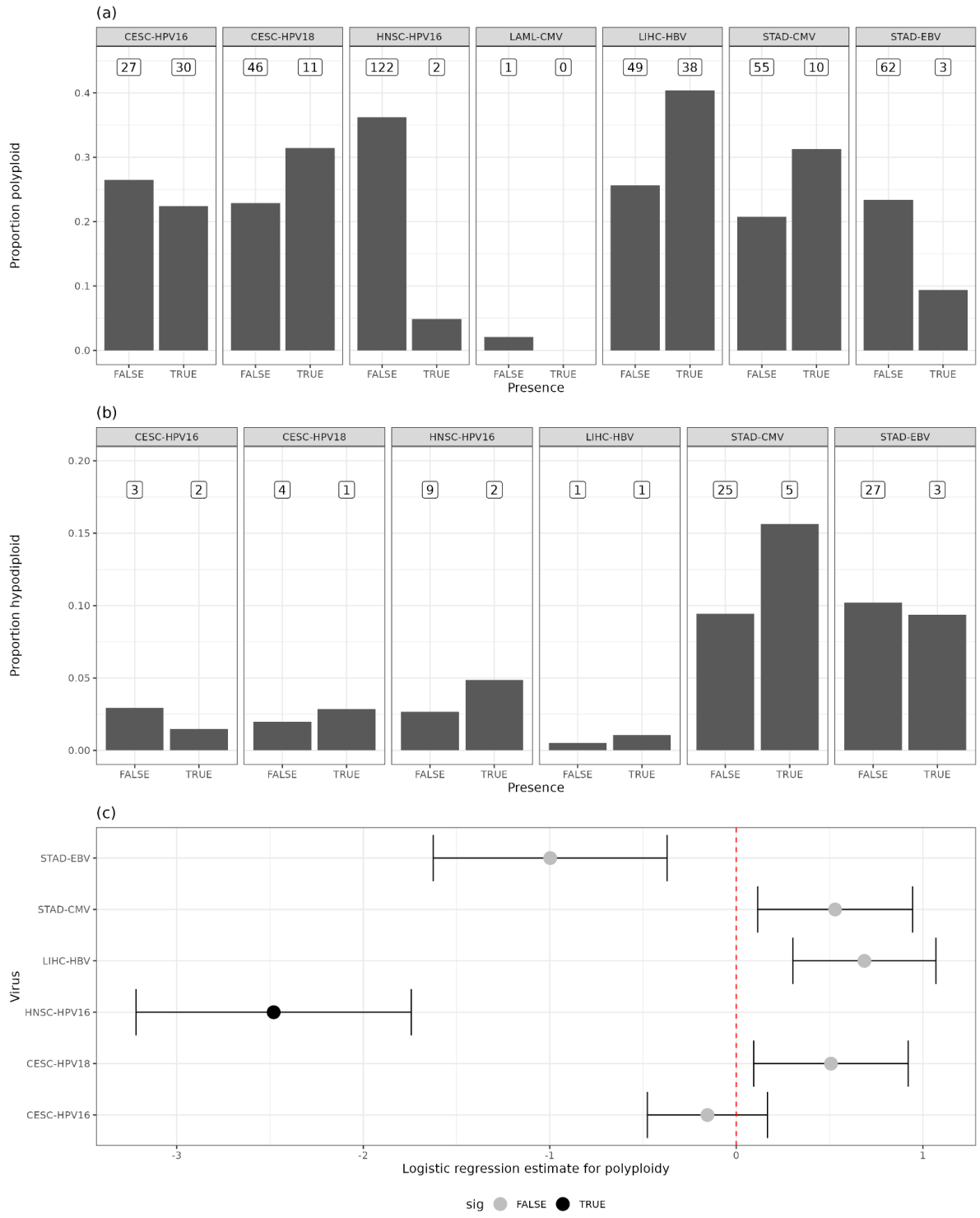


Fig. 2.19 Viral infection and ploidy alterations. Rates of **(a)** polyploidy and **(b)** hypodiploidy by viral infection status; **(c)** results of logistic regression of polyploidy status on viral infection, controlling for age, sex and self-reported race. Viral infection was quantified by Chen et al (2021) using the presence of viral sequences in the transcriptome, and here dichotomised into present/absent.

This was surprising, as HPV proteins have been shown to induce polyploidy in vitro via endoreduplication and lagging chromosomes (Patel et al, 2004) and upon DNA damage (Fan et al, 2013). Given the known mutual exclusivity between TP53 mutation and HPV infection in this cancer type (Shi et al, 2021), I hypothesised that these contradictory results could come from a selection effect with TP53. In this dataset, there are 264 cases with a TP53 mutation, but zero of these are HPV16-infected, while 41 of the 73 non-TP53-mutated samples are HPV16-infected. There is also significant mutual exclusivity between TP53 mutations and Epstein-Barr virus infection in stomach adenocarcinoma (TP53 mutation rate 46.8% in EBV-negative STAD vs 18.8% in EBV-positive STAD, Fisher's exact $p_{adj} = 0.0077$). Interestingly, p53 mutations were not mutually exclusive with HBV infection in liver cancer, and indeed the rate of TP53 mutation increases from 23% in HBV-negative LIHC to 40.4% when HBV is present ($p_{adj} = 0.0078$). An increase in TP53 mutations in HBV-related liver cancer has been reported (Gao et al, 2019), but this does contrast with significant mutual exclusivity found between HBV-LIHC and TP53 mutations by the PCAWG consortium (Zapatka et al, 2020).

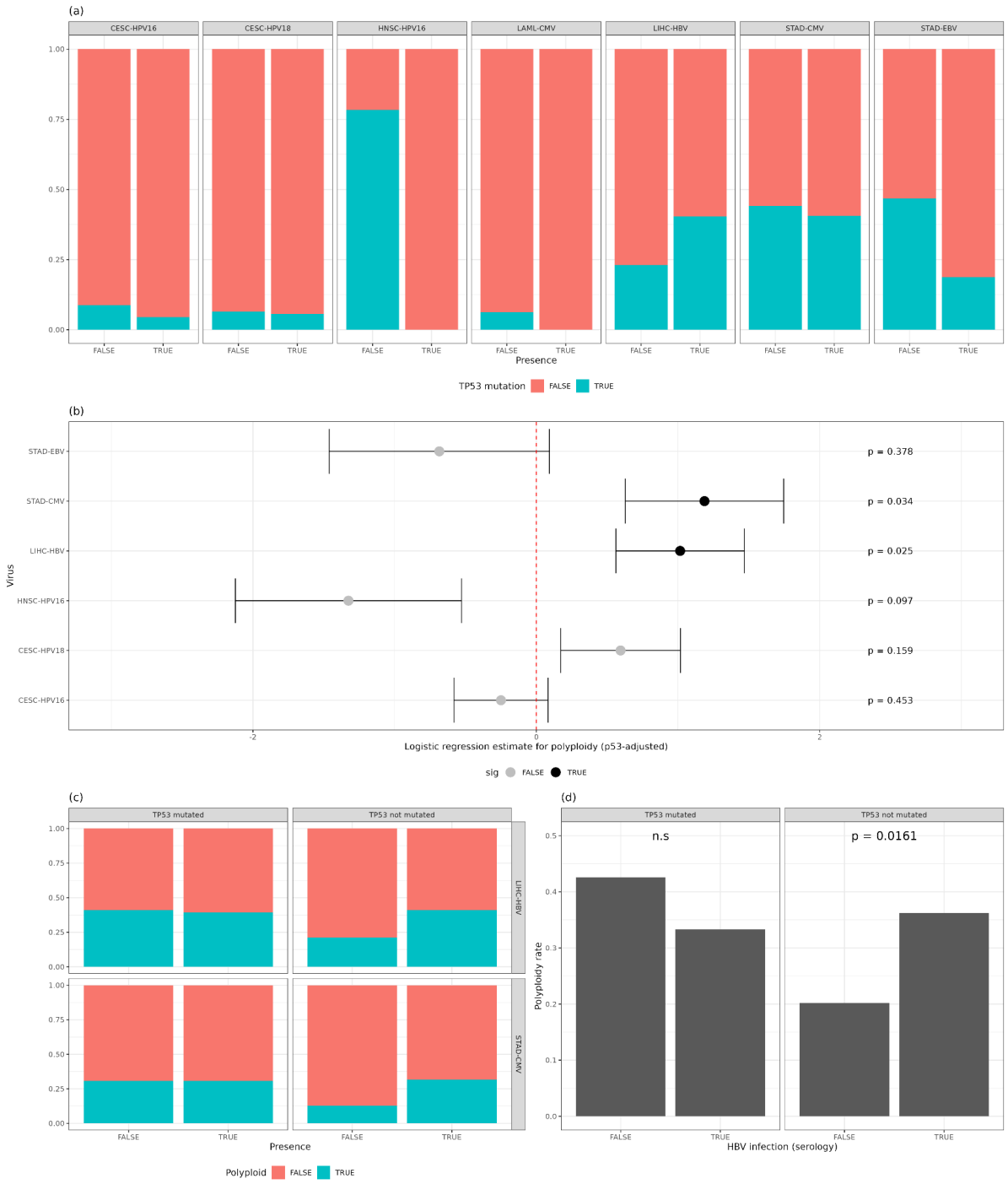


Fig. 2.20 Associations between viral infection and polyploidy status depend on TP53 mutation status (a) Mutual exclusivity between viral infection and TP53 mutation in HPV-HNSC and (to a lesser extent) EBV-STAD **(b)** Updated logistic regression of polyploidy on viral infection status, adding TP53 mutation status and TP53 mutation:viral infection interaction as covariates **(c)** Increase in polyploidy rate upon viral infection in liver and stomach cancer specifically when TP53 is not mutated **(d)** Increase in polyploidy rate in HBV-infected liver cancers, with HBV infection status determined by serology instead of transcriptomics.

When I add TP53 mutation status to the logistic regression as a covariate, the negative association between HPV16 and polyploidy in HNSC becomes non-significant. When I additionally add an interaction term (p53*virus presence), I also observe positive associations between polyploidy and infection with HBV in liver cancer and CMV in stomach adenocarcinoma (Fig. 2.20), although these are not significant after multiple testing correction. Further investigation shows that in both HBV-LIHC and CMV-STAD, viral infection is associated with an increased risk of polyploidy exclusively when TP53 is intact—in cases with mutated TP53, viral infection does not increase the rate of polyploidy. I confirmed these findings with an alternate source of viral infection information: results from ISH testing for HPV in HNSC patients (no association) and seropositivity data for HBV in liver cancer patients: an increase in polyploidy with viral infection, specifically in non TP53-mutant samples ($p = 0.016$).

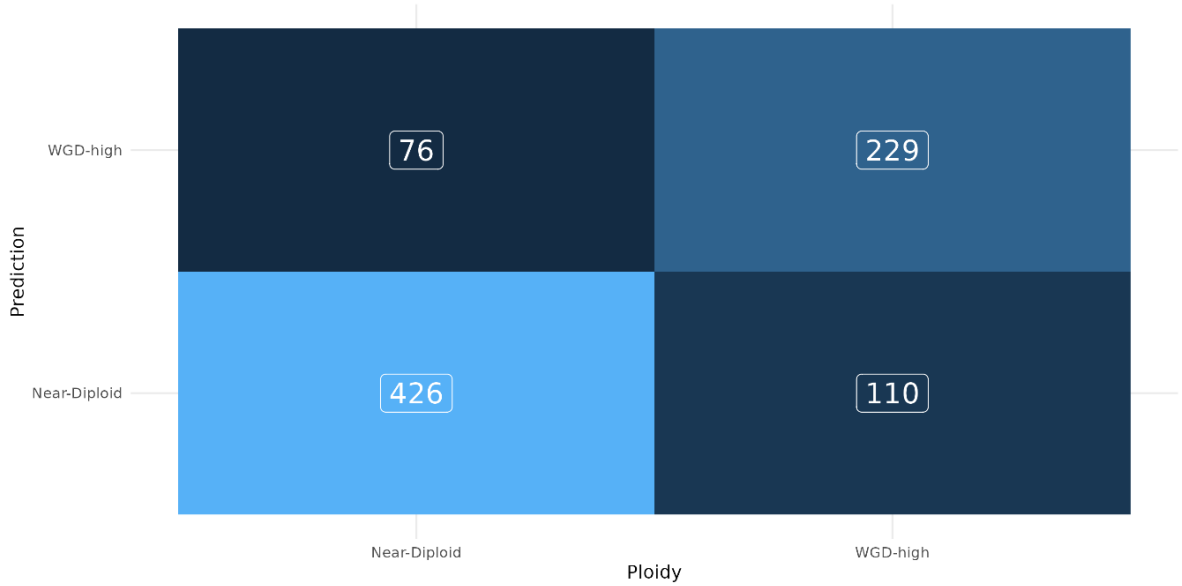
2.3.10 Integrating all factors

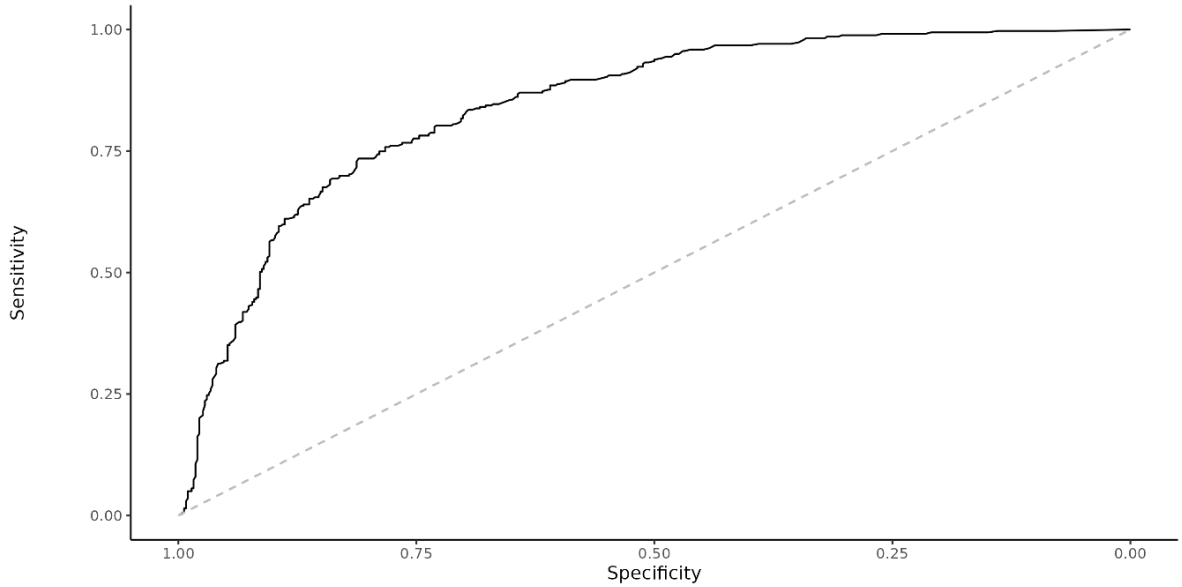
I have now analysed the extent to which a range of genetic and environmental influences are associated with tumour ploidy within patient samples, but several questions remain:

- How important is each factor?
- Do they interact with each other?
- Based on these factors, can patient tumours be classified by their propensity towards extreme ploidy?

To attempt to answer these questions, I built a random forest model and used it to classify TCGA tumours into polyploid and near-diploid groups. I included fourteen factors of interest: proliferative index, the mutational status of three genes (TP53, ARID1A, PTEN), hypoxia measured by the Ragnum score, smoking status, infection status for three viruses (HPV, HBV, CMV) and normalised activity values for five mutational signatures (SBS4, SBS5, SBS1, SBS40a, SBS44). Our data was reasonably well-balanced, with the minority class (polyploids) making up 40.3% of the sample. After splitting the data 80/20 into training and test groups, I had 1,357 polyploids and 2,009 near-diploids in the training set and 339 polyploids and 502 near-diploids in the test set. Because I wanted to assess the ability of my factors of interest to predict ploidy class on their own, I did not include age, sex, or cancer type as predictors. Proliferative index was defined as the median PI among diploids in hopes of capturing the overall tissue replication rate (i.e. a potential causal factor) and not simply the high proliferative index of polyploid tumours themselves.

The model had good predictive value overall, with an AUROC (area under the receiver operating curve) of 0.84 (Fig. 2.21). The specificity (0.84) is significantly better than the sensitivity (0.68), suggesting that the model is more likely to incorrectly assign a polyploid case as diploid than vice versa. The most important features for ploidy classification, as measured by Mean Gini Decrease, are (1) hypoxia (2) TP53 mutation status (3) tissue median proliferative index. The least important factors are the three viral infection features, but this may be because they are only relevant to a small subset of the data (e.g. liver cancer for HBV).


76
229
426
110
WGD-high
Near-Diploid
Prediction
Near-Diploid
WGD-high
Ploidy


1.00
0.75
0.50
0.25
0.00
Sensitivity
1.00
0.75
0.50
0.25
0.00
Specificity

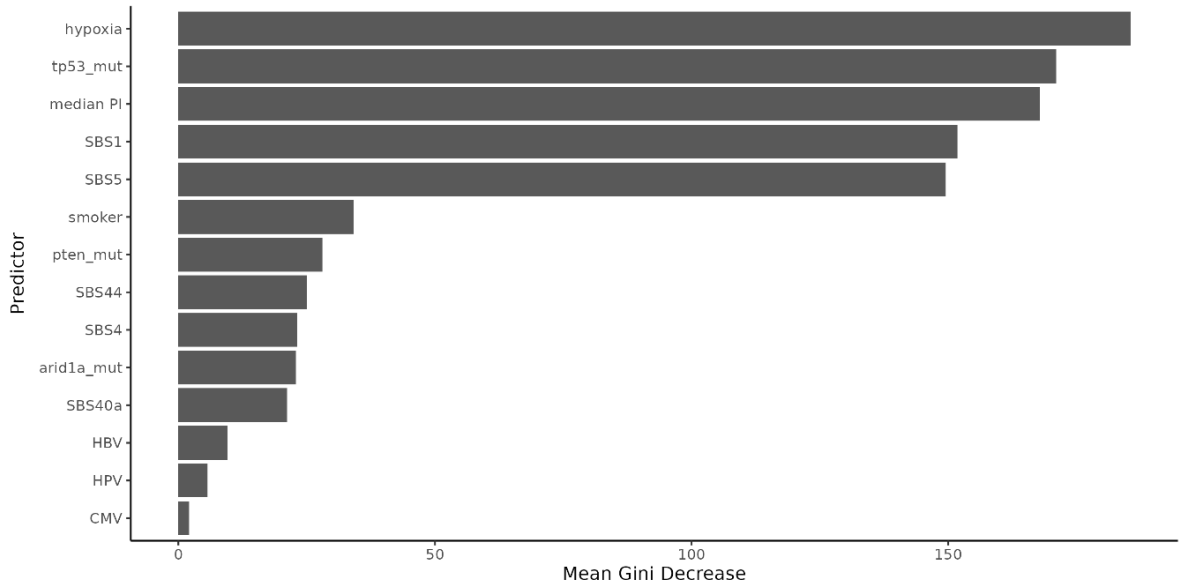

hypoxia
tp53_mut
median PI
SBS1
SBS5
smoker
pten_mut
SBS44
SBS4
arid1a_mut
SBS40a
HBV
HPV
CMV
Predictor
0
50
100
150
Mean Gini Decrease

Fig. 2.21 Performance of a random forest model in predicting polyploidy. I included 14 features based on the analyses in this chapter: mutations in TP53, ARID1A and PTEN, hypoxia, smoking status, viral infection, significant mutational signatures and tissue median proliferative index. **(top)** confusion matrix, with observed ploidy status on the x-axis and the model's prediction on the y-axis **(centre)** ROC curve showing the relationship between model sensitivity and false positive rate **(bottom)** importance of each feature to the model's performance, as measured by Mean Gini Decrease.

As TP53 is a well-established predictor of ploidy status in cancer, a minimal version of the model was tested in which TP53 mutation status was the only predictor. In this model, the AUROC decreased from 0.84 to 0.72, with sensitivity decreasing to 0.63 and specificity to 0.8. When the activity of the SBS1 mutational signature was included to account for age, the AUROC increased slightly to 0.74. These ancillary analyses indicate that TP53 mutation status alone (or with age) is a strong predictor of polyploidy, but does not capture the total predictive value of the full model.

Interestingly, adding cancer type as a feature only increased the AUC slightly, to 0.854. This might indicate that most of the information provided by cancer type is already being accounted for by these predictors, but it may also be that the 'median proliferative index' variable is acting as a proxy for cancer type. In another version of the model, I instead used the proliferative index of each individual sample (and excluded cancer type). The AUC decreased to 0.8183 but remained within the 'good' predictive value category (0.8-0.9). However, as I am interested in assessing the potential causal factors behind polyploidy in cancer, not in *predicting* polyploidy per se, it is problematic that the individual sample proliferative index is affected by reverse causality. In this version of the model,

the most important factors are similar to the original version: (1) proliferative index (2) TP53 mutation status (3) hypoxia score (Appendix Fig. 6.6).

I also integrated the various features in a multiple logistic regression model to obtain a more interpretable view of predictor relationships with ploidy status and to include hypodiploidy, for which I did not build a random forest model due to low sample size. Age, gender and cancer type were included in the regressions but not in the random forest model because here I am not attempting to assess overall predictive power.

The overall regression for polyploidy finds that the following features are significantly positively associated with polyploidy across the TCGA dataset: TP53 mutation status, HBV infection, proliferative index (individual), hypoxia, SBS5 and SBS1, male sex and higher age at diagnosis (Fig. 2.22). Three variables were negatively associated with polyploidy status after taking into account all other predictors: SBS44, PTEN mutation and (somewhat surprisingly) HPV infection. In the full model, ARID1A mutation and the activities of the SBS4 and SBS40a mutational signatures become non-significant.

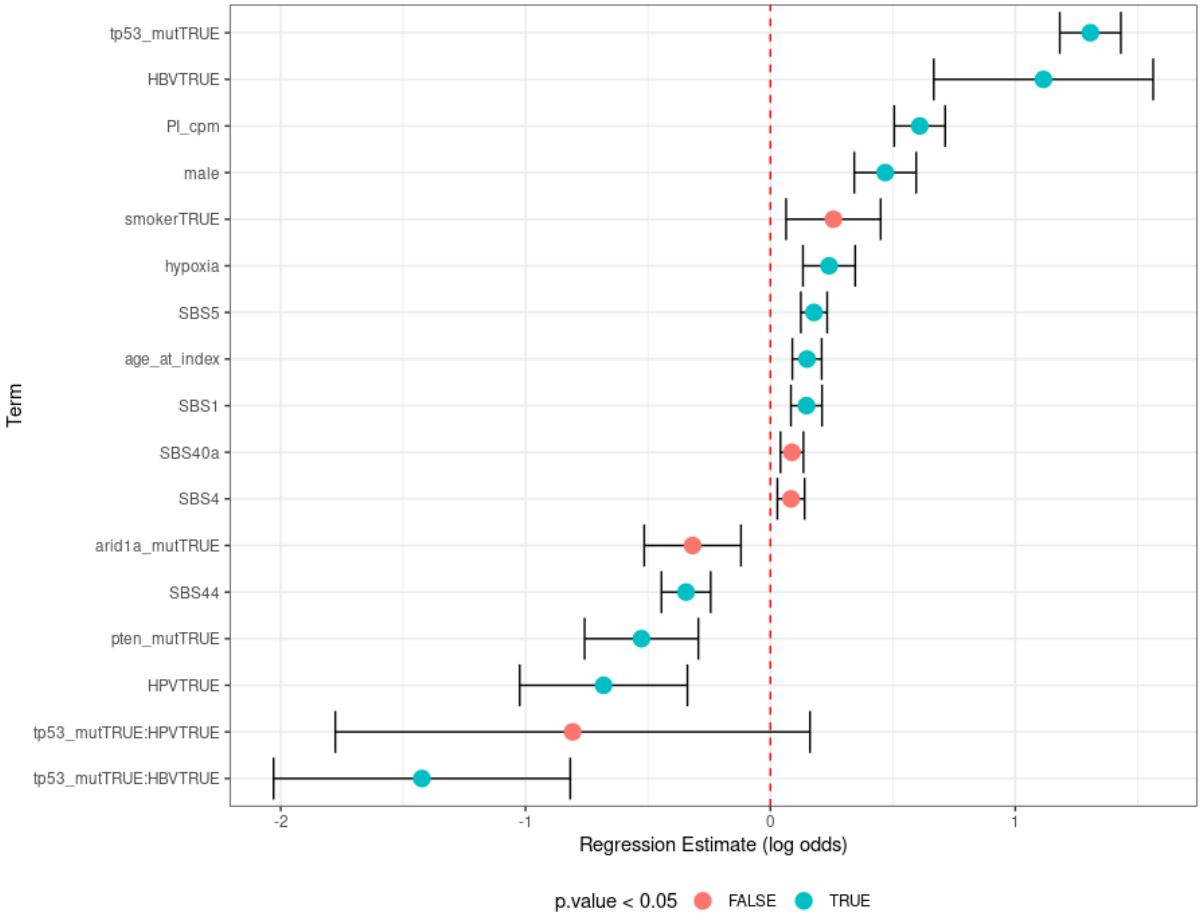


Fig. 2.22 Logistic regression of all investigated predictors, polyploidy vs near-diploidy. Numerical predictors have been centered and scaled. Cancer type ($p < 0.05$ in 26/32), smoking status 'unknown' (n.s.) and self-reported race (n.s.) were also included in the regression but are not shown. The bottom two rows show interaction terms for TP53 mutation status with HPV/HBV infection.

Turning to hypodiploids, in the full model the only significant predictors were

TP53 mutation status, proliferative index and age at diagnosis (Fig. 2.23).

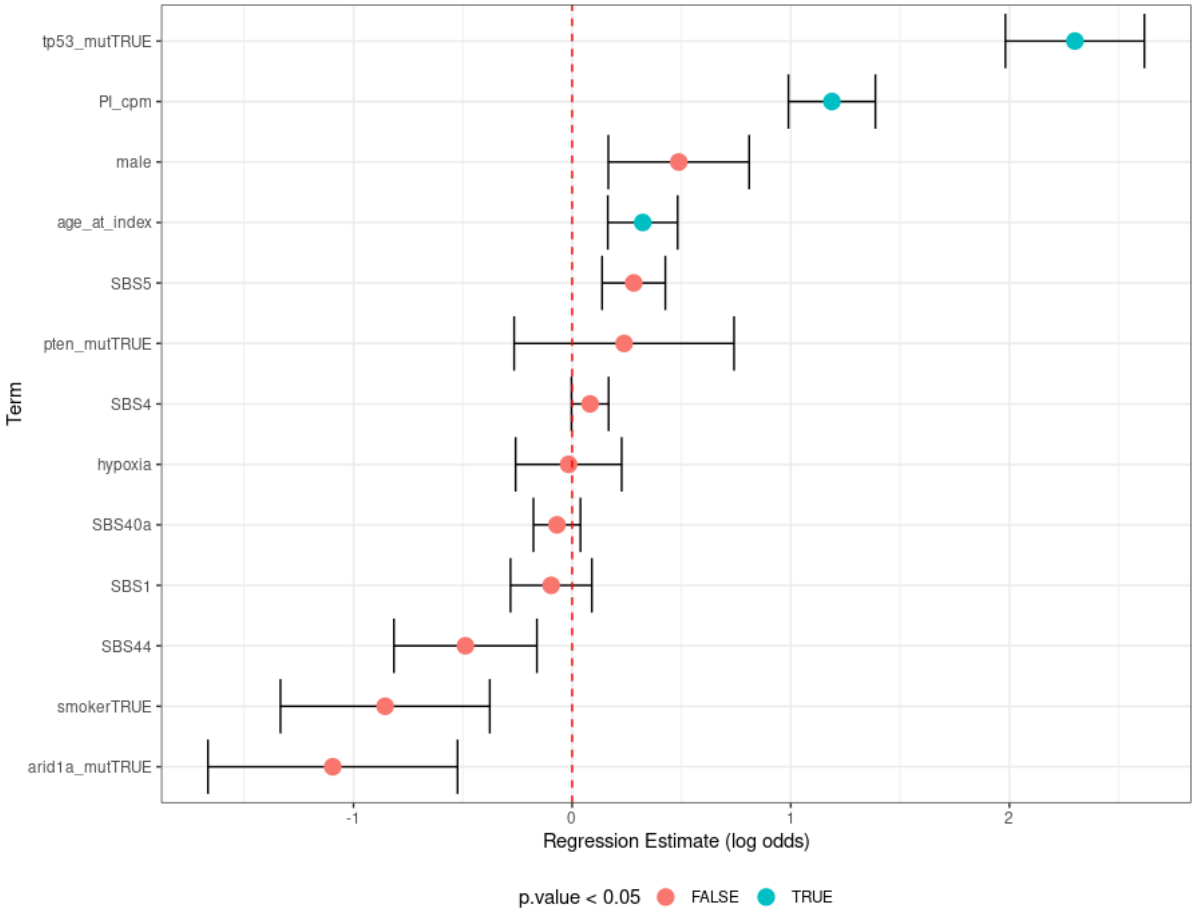


Fig. 2.23 Logistic regression of all investigated predictors, hypodiploidy vs near-diploidy. Numerical predictors have been centered and scaled. Cancer type, smoking status 'unknown' and self-reported race were again included but are not shown. HPV, HBV and CMV infection status were not included in this regression due to low sample size in the hypodiploids.

2.4 Discussion

2.4.1 Can the development of extreme tumour ploidy be explained?

An understanding of factors predisposing towards ploidy alterations might help predict tumour progression and shed light on the complex relationship between ploidy alterations and prognosis. I used a combination of random forest models and logistic regression to understand the extent to which each investigated factor predicts ploidy. The regression models were used to understand the direction of relationships between each predictor and ploidy, whereas the random forest model provided estimates of feature importance and an overall estimate of classification accuracy while implicitly handling non-linear relationships and interactions between features. The model performed quite well overall, but its recall was notably lower than its precision, i.e. it missed many polyploid cases. While the classes were reasonably balanced (minority class proportion: 0.4), it is possible I could improve recall by either undersampling the diploid cases while training or using a tweaked method such as Balanced Random Forest. My interpretation of the model could also be improved by using permutation-based importance values instead of mean error reduction values. It would also be interesting to examine the impact of comparing polyploids to perfectly diploid cases (exactly 22 autosomes) rather than near-diploids.

My goal with this analysis was not to predict tumour ploidy *per se*—why measure TP53 mutation status and hypoxia level when you can just karyotype the tumour?—but rather to assess the importance of my chosen features in discriminating between polyploid and diploid samples. However, if the model is applicable to samples that are diploid at diagnosis but later polyploidise, then it could theoretically be useful to anticipate likely polyploidisation. If one considers two tumours containing a small polyploid clone that is not yet detectable by sequencing, the tumour positive for factors that support the expansion of polyploid clones is likely to have higher intra-tumour heterogeneity in the future and to be more resistant to chemotherapy. This could be prognostic, although of course many of the factors analysed here are already prognostic independently of their association with ploidy. The usefulness of a method to predict polyploidy depends not only on the model's accuracy but also on whether the prediction would change the treatment plan, e.g. if more aggressive treatment/combination therapy might be pursued, or if a ploidy-targeted drug such as a KIF18A inhibitor (Payton et al, 2023) were under consideration.

2.4.2 Are factors identified in *in vitro* studies relevant within patient tumours?

Even in cases where a clear mechanistic link has been established between the presence of a given factor (e.g. viral infection) and polyploidy, it is not guaranteed that differences between patients in ploidy status will be attributable to that factor. It is plausible, for example, that viral infection could attract the attention of the immune system and increase policing/immunoediting of

polyploid cells, potentially cancelling out or even reversing the polyploidy-promoting effect of the viral proteins themselves. Additionally, real tumours face combinations of exposures not experienced *in vitro* which could overwhelm the contribution of any given factor. While all five of the viruses I analysed have been shown to induce polyploidy in cell culture, only two appeared as a (nominally) significant positive predictor of polyploidy in patient tumours: CMV in stomach cancer and HBV in liver cancer. The association with liver cancer is interesting because hepatocytes regularly undergo programmed polyploidisation, which is generally protective against tumorigenesis (Zhang et al, 2018) and thus polyploidy in liver cancer cells might be ignored by the immune system. I did not find significant associations between EBV infection in stomach cancer or HPV in cervical cancer and ploidy, and in head-and-neck squamous cell carcinoma HPV16 infection is negatively associated with polyploidy in the full model, contrasting with cell line studies.

I was generally able to confirm the established link with hypoxia in patient tumours, including in the full regression model for polyploidy. In hypodiploids, however, hypoxia was not a significant predictor in the full model despite its strong association in the individual model. This could be because hypoxia is tightly linked with replication rate (fast-growing tumours tend to outgrow their vascularisation capacity and become hypoxic) and is explained by that or by other factors correlated with replication rate (e.g. the mutation status of individual tumour suppressor genes). I found that smoking is associated with hypoxia in lung cancer but not with ploidy. However, it is strongly associated

with smaller copy number alterations, measured by the number of copy number segments per sample, consistent with previous research (Farkas et al, 2021).

The lack of common-variant heritability for ploidy is not entirely surprising: the ability to maintain chromosome stability across cell divisions is fundamental to organismal survival, so variants affecting it should be under strong negative selection and thus rare. Somatic mutations do not face this constraint to the same degree, and indeed I found a set that are significantly associated with polyploidy. The strongest hit by far, for both polyploidy and hypodiploidy, is TP53, which is well-established as a central regulator of genome stability. It has been shown in both the TCGA and MSK-IMPACT cohorts that TP53 mutations precede WGD in the vast majority (>90%) of cases amenable to timing (McGranahan et al, 2015; Bielski et al, 2018). Analyses of these and other datasets across a range of cancer types indicate that driver mutations tend to occur before WGD in general, with WGD being an intermediate event in tumour evolution (Gerstung et al, 2020; Lu et al, 2024; Yates et al, 2017; Offin et al, 2019; Satgunaseelan et al, 2022). This has been characterised in detail for TP53 in particular: lineage tracing in mice with heterozygous Trp53 mutations revealed a consistent pattern of chromosomal deletions before WGD (Baslan et al, 2022). The delay between the occurrence of oncogenic mutations and WGD suggests that these mutations (at least in the case of TP53) enable WGD but do not necessarily immediately precipitate it.

ARID1A mutations were negatively associated with both polyploidy and (in some analyses) hypodiploidy, but were not significant independent predictors in the combined regression model. ARID1A inactivation is known to cause mitotic errors and directly induce polyploidy, yet cancer types with a high frequency of ARID1A mutations are notable for their chromosomal stability. This is explained by a loss of telomere cohesion in ARID1A-deficient tumours that renders chromosomally-unstable cells incapable of proliferating (Zhao et al, 2019).

Aneuploidy appears to be targeted by the immune system (Senovilla et al, 2012), and successful polyploid clones have activated immune evasion pathways (Burdett et al, 2024). As well as actively evading the immune system, it is possible that polyploid clones are more likely to arise in the first place in immune-cold microenvironments. I have not seen any evidence of this so far: no common germline variants in immune genes were significantly associated with tumour ploidy in the GWAS, and a preliminary analysis of immune infiltration based on CIBERSORT analysis of GTEx data suggests that polyploid tumours are not more common in tissues with lower average levels of infiltration of aggressive immune cells. This suggests that—while polyploid tumours are more likely to successfully establish themselves in an immunodeficient mouse than an immunocompetent one—patients in the TCGA cohort do not differ enough in their strength of immunosurveillance to create a noticeable difference in polyploidy rate. It might be interesting to calculate polygenic risk scores for patients based on summary statistics from GWASs of immune traits (e.g. autoimmunity) and compare these to ploidy status, but given the finding that the

SNP-based heritability of tumour ploidy is zero, it does seem like differences in polyploidy rate are not related to differences in immunosurveillance between patients, at least not differences that can be proxied by germline variants or tissue averages.

2.4.3 Polyploidy and hypodiploidy: two sides of the same coin?

There is a weak to moderate correlation between rates of polyploidy and hypodiploidy across most cancer types, except two with extremely high rates of hypodiploidy (adrenocortical carcinoma and kidney chromophobe). These two cancer types were also identified by Steele et al (2022) as the only TCGA tumours that did not show a linear relationship between number of copy number segments and amount of loss of heterozygosity across the genome. I speculate that there may be a common basis partially underlying polyploidy and hypodiploidy rates in general (e.g. a chromosomal instability phenotype or tissue replication rate) while the two outlying cancers are driven specifically towards hypodiploidy by some other factor that does not promote polyploidy (to the same extent). The only significant predictors shared between polyploidy and hypodiploidy in the final regression were TP53 mutation status and proliferative index, but some of the difference may be due to low sample size among the hypodiploids. I conclude that polyploid and hypodiploid tumours share some factors: TP53 mutations, hypoxia, high replication rates – but may be distinct in their relationship with others, such as viral infection and other environmental exposures proxied by mutational signatures. Due to their very limited sample

size (13, 0.17%), I was not able to study the near-haploids specifically. It would be interesting to determine the rate of sudden near-haploidisation (Walen, 2014) as opposed to gradual chromosome loss, but this is difficult because we cannot time losses the way we time gains.

2.4.4 Proliferation

Polyploid and hypodiploid cells have significantly higher proliferation rates than diploids, even within the same cancer type. One interpretation of this result is that the (pre-ploidy change) ancestors of these cells were a more proliferative, aggressive subset and they have simply inherited this predisposition. Another interpretation is that polyploidy itself increases proliferation rate. Experimentally, polyploidy has been shown to increase proliferation in primary, but not recurrent, tumours in an engineered mouse model of breast cancer (Newcomb et al, 2021). This is somewhat surprising because polyploid cells are known for entering a (temporary) quiescent state (Pienta et al, 2022), and because polyploidy slows the process of mitosis (Bloomfield et al, 2021). However, (a) their higher evolvability could unlock other oncogenic alterations that increase proliferation, and (b) some of the polyploids I studied have lost large amounts of genetic material and have ploidies closer to diploid than tetraploid. It might be possible to distinguish between these two interpretations by using time series biopsies to understand whether there is a change in proliferative index after the establishment of poly/hypodiploidy.

2.4.5 Additional factors that may influence rates of extreme ploidy

One factor that I was not able to include in this work but would be a valuable line of future research is chemotherapy exposure. I did not investigate it here because the TCGA samples were almost all sequenced pre-treatment, but chemotherapy drugs including docetaxel (Ogden et al, 2015) and cisplatin (Adibi et al, 2023) have been shown to induce polyploidy in cell culture. One could analyse whether ploidy aberrations in cancer patients can be attributed to chemotherapy exposure using a combination of clinical data and mutational signatures in a dataset that includes treated cancers, e.g. the Hartwig database. It would be particularly interesting to see whether the known higher rate of polyploidy in advanced cancers (Bielski et al, 2018) is partially attributable to the influence of chemotherapy. Given the poor prognosis typically associated with polyploidy, it would be valuable to know if treatment decisions are increasing its likelihood, and if there are alternate drugs available for a given cancer (sub)type that do not have this effect and could be substituted.

Chapter 3

Karyotypic evolution by large-scale chromosome loss in hypodiploid and polyploid tumours

3.1 Introduction

While most aneuploid tumours are hyperploid, ~1% of tumours will lose one copy of many or most of their chromosomes and become severely hypodiploid, with widespread loss of heterozygosity (LOH) across the genome. Widespread monosomy creates potentially targetable vulnerabilities, e.g. the essentiality of the homologous recombination repair gene RAD51B in near-haploid leukaemia cells (Liu-Lupo et al, 2023). However, it could also create a sensitised background for tumour suppressor gene mutations, cause oncogenic dosage imbalances in gene signalling networks, or induce

chromosomal instability, which has been associated with poor prognosis across cancer types (van Dijk et al, 2021).

Hypodiploidy has been best studied in acute lymphoblastic leukaemia (ALL), where near-haploid (<30 chromosomes) and low-hypodiploid (30-39 chromosomes) cases form a rare subgroup with distinct mutational profiles (Holmfeldt et al, 2013), altered response to treatment (McNeer et al, 2019) and very poor prognosis (reviewed in Safavi & Paulsson, 2017). Near-haploid or low-hypodiploid subsets have also been observed in other cancers including chondrosarcoma (Olsson et al, 2011), giant cell glioblastoma (Baker et al, 2020), malignant pleural mesothelioma (Hmeljak et al, 2018) and oncocytic follicular thyroid carcinoma (Corver et al, 2012). The mechanism by which these karyotypes arise is unclear, but could involve successive losses or multipolar mitosis (Molina et al, 2021). In at least some tumour types, hypodiploidy can arise early in tumour development (Bovée et al, 2000) and be stable throughout metastatic progression (Gopal et al, 2018).

In the hypodiploid cancer types studied thus far, chromosomes are not lost at random (Mandahl et al, 2012). Hypodiploid leukemias almost universally retain both copies of chromosome 21, typically alongside chromosomes 14, 18 and X, while chromosomes 5 and 7 are preferentially retained in chondrosarcoma (Olsson et al, 2011) and oncocytic follicular thyroid carcinoma (Gopal et al, 2018). The cause of these patterns is largely unknown. In general, cancer aneuploidy patterns seem to act to alter the copy number of oncogenes and tumour suppressors (Davoli et al, 2013) and/or to “hard-wire” typical tissue-specific expression levels (Patkar et al, 2021).

Chromosome loss and retention patterns may be partially determined by dosage balance constraint. Across cancer types, small chromosomes with fewer genes are more likely to be lost (Duijf et al, 2013) despite their lower initial missegregation frequency (Worrall et al, 2018), but previous studies have found no link between chromosome size and loss rate in hypodiploidy (Mandahl et al, 2012). Chromosome 7 is notably unlikely to show loss of heterozygosity across solid tumours, which may be due to the presence of the oncogenes EGFR and BRAF or the six essential imprinted genes scattered across the chromosome (Boot et al, 2016). Using translocations and partial deletions, an earlier study identified an essential region at 7q32-7qter in peripheral chondrosarcomas (Bovée et al, 2000). However, chromosome 7 retention is not universal even in solid tumours, and indeed monosomy 7 is one of the most common abnormalities in acute myeloid leukaemia and B-cell precursor ALL (Lundin-Strom et al, 2021). Thus it is unclear to what extent the presence of imprinted or otherwise dosage-sensitive genes affects the likelihood of a chromosome becoming monosomic, especially on a hypodiploid background.

Whole genome duplication (WGD) in cancers is intimately linked with loss of genomic material. While hypodiploidy is perhaps the most obvious manifestation of large-scale chromosome loss, this is also a notable feature of post-WGD evolution in polyploid tumours (Bielski et al, 2018; Prasad et al, 2022; Baker et al, 2024), with the diverse, highly aneuploid karyotypes that characterise many forms of cancer arising via losses from an “unstable tetraploid intermediate” (Ganem et al, 2007). Additionally, there is evidence that WGD is selected to buffer the effect of deleterious mutations in regions of loss of heterozygosity (Lopez et al, 2020). Many studies have reported elevated rates of WGD in hypodiploid tumours, resulting in hyperdiploid clones with homozygous disomies and heterozygous tetrasomies. While WGD is rare in ALL in general, it has

been estimated to occur in 64% of near-haploid and 44% of low-hypodiploid ALL cases and may benefit the tumour by restoring a more normal gene dosage (Safavi & Paulsson, 2017). The doubled and non-doubled clones can coexist, with the doubled clone often dominating at diagnosis and the hypodiploid founder clone at relapse (Molina et al, 2021). In an unknown proportion of cases, only the doubled clone remains, creating "masked hypodiploids". In ALL, masked hypodiploids share the poor prognosis of their founder clones (Nachman et al, 2007) but can be mistaken for good-prognosis hyperdiploids. In the interest of accurate prognostication and appropriate treatment, it is thus essential to distinguish masked hypodiploids from hyperdiploids at diagnosis. This typically requires allele-specific copy number information from SNP chips or sequencing (Safavi & Paulsson, 2017) rather than standard diagnostic cytogenetics (Wan et al, 2014; Liehr et al, 2015).

In this work, I identify hypodiploid tumours across a range of haematological and solid cancers and test a simple method for distinguishing hyperdiploid ALL from masked hypodiploid ALL using only cytogenetic data. I then investigate the relationship between hypodiploidy, genome duplication and smaller-scale forms of chromosomal instability. I analyse the amount and timing of losses post-WGD in tumours with and without a hypodiploid history. Finally, I compare tissue-specific chromosome loss patterns between hypodiploids, post-WGD polyploids and near-diploid tumours and attempt to explain chromosome (arm) loss rates with reference to chromosome features including size, nuclear location, expression and the presence of cancer drivers and dosage-constrained genes.

3.2 Methods

3.2.1 Copy Number Profiles

I downloaded all karyotypes from the Mitelman Database of Chromosome Aberrations & Gene Fusions in Cancer and used CytoConverter (Wang et al, 2019), bedops (v2.4.37) partition and bedtools (v2.27.1) map to calculate net copy number alterations for each genomic region in all samples annotated with a morphology of 'acute lymphoblastic leukemia/lymphoblastic lymphoma'. I removed cases with errors from CytoConverter or with unspecified total chromosome number, then calculated copy number as 2 + net copy number alteration. The sex chromosomes were excluded because I could not reliably ascertain their copy numbers using CytoConverter. Cases with no numerical (only structural) abnormalities in the karyotype were inserted separately and assumed to be fully diploid. I processed the TCGA data and defined ploidy classes as described in the Methods section for Chapter 2. I obtained intra-tumour copy number heterogeneity estimates from van Dijk et al (2021) and quantified small-scale copy number instability in each sample using (i) the total number of copy number segments (ii) the proportion of each chromosome covered by its longest copy number segment, where a 'segment' is a contiguous region with one copy number configuration.

3.2.2 Distinguishing Masked Hypodiploids from Hyperdiploids using Cytogenetic Data

I calculated a masked hypodiploidy (MH) score for each sample as $N_4 - N_3$, i.e. number of tetrasomic chromosomes - number of trisomic chromosomes. Only chromosomes with ≥90% of their length having a consistent copy number were considered, and MH scores were not calculated for samples with fewer than ten chromosomes that met this criterion. To test the ability of the score to distinguish masked hypodiploids from hyperdiploids, I gathered two datasets of "true positives" (hypodiploids) and "true negatives" (hyperdiploids). For the positives, I took Mitelman ALL cases that had both a hypodiploid and a hyperdiploid clone, built copy number phylogenetic trees with MEDICC2 (with the --cn parameter set to 'total copy numbers') and selected cases where the hyperdiploid clone was inferred to be genome-doubled but had at most 74 autosomes (N = 123). For the negatives, I obtained allele-specific hyperdiploid copy number data from Woodward et al (2023) (N = 577) and counted chromosomes labelled as "Tri" vs "Tetra_2_2" (balanced tetrasomy) or "Tetra_3_1" (unbalanced tetrasomy) after excluding the sex chromosomes. I downloaded ASCAT-based allele-specific copy number profiles for a separate cohort of ALL cases from the TARGET-ALL Phase II project using GDCquery from the TCGAbiolinks package in R.

3.2.3 Whole Genome Duplication Rate

ALL cases with a masked hypodiploidy score > 0 were redefined as hypodiploid, classified as near-haploid or low-hypodiploid based on their

autosome count divided by two, and counted as genome-doubled when calculating WGD rate by ploidy class. They were not used for chromosome-specific analyses because I cannot be sure of their exact chromosome count pre-doubling.

3.2.4 Loss after Whole Genome Duplication

I calculated the minimum ploidy of a genome-doubled tumour immediately after the WGD event as (4*proportion of genome with retained heterozygosity) + (2*proportion of genome with loss of heterozygosity) + (0*proportion of genome in zero copies), which is close to 4 - 2*(LOH proportion). This allowed me to classify tumours into four groups: post-WGD losses outweigh all gains (current ploidy < minimum ploidy), total losses outweigh total (non-WGD) gains (current ploidy < 4), total losses equal total (non-WGD) gains (ploidy = 4), and non-WGD gains outweigh total losses (current ploidy > 4). To classify losses as having occurred before or after WGD, I assume a certain level of parsimony, i.e. that loss of heterozygosity always results from pre-WGD losses. However, it is also possible for LOH to arise through multiple post-WGD losses, or for a segment to be lost pre-WGD without LOH if it had previously been gained. I defined a post-WGD loss as a copy number of 2 or 3 with LOH across <20% of the chromosome arm and a pre-WGD loss as LOH across >90% of the chromosome arm. When calculating rates of post-WGD loss, I removed samples with a pre-WGD loss in that chromosome arm from the denominator. I investigated loss timing using chronological WGD timing estimates based on the presence of duplicated

mutations across the genome from Gerstung et al (2020). I also analysed timing estimates from GRITIC (Baker et al, 2024) and confirmed that they were highly correlated with the Gerstung et al. estimates.

3.2.5 Chromosome Loss Patterns

I created a binary loss/retention profile for each chromosome arm in each sample, defining a chromosome arm as lost if it had >90% loss-of-heterozygosity (TCGA samples) or was monosomic for >90% of its length (Mitelman samples). The sex chromosomes and acrocentric arms (13p, 14p, 15p, 21p, 22p) were excluded due to lack of data (in TCGA, for the acrocentric arms). The average loss rate for each chromosome arm was calculated separately by tissue, except for the TCGA near-haploid cases which were grouped due to low sample size. I clustered all samples by their chromosome loss profiles using the Jaccard distance (1 - number of shared lost chromosome arms divided by total number of lost chromosome arms between the two samples), and calculated the median Jaccard similarity (1 - D) within and between each cancer type.

$$D = 1 - \frac{S_1 \cap S_2}{S_1 \cup S_2}$$

Eq. 3.1 Jaccard distance between samples S_1 and S_2 is calculated as 1 - the size of the intersection between the two sets divided by the size of the union of the two sets, where sets here are lost chromosome arms.

The heatmap was created with ComplexHeatmap (v2.22.0) in R. To assess tissue specificity of chromosome loss profiles, I performed PAM (Partitioning Around Medoids) clustering on the Jaccard distance matrices, with k set to

the number of cancer types being analysed. PAM is agnostic to the method of distance matrix computation, unlike k-means which relies on Euclidean distance and is not suitable for binary data. I then compared cluster assignment to true cancer type labels using the adjusted Rand index from the mclust R package (v6.1.1), which quantifies the similarity between two labellings after adjusting for chance.

3.2.6 Factors Associated with Loss Rate

I collated four sets of genes previously identified as being enriched for dosage-sensitive genes and counted the number of these genes on each chromosome arm. These included: genes involved in pathogenic CNVs in the human population or whose copy number has remained unchanged throughout mammalian evolution (Rice & McLysaght, 2017), haploinsufficient essential genes (Sarel-Galilly et al, 2022), and imprinted genes from the GeneImprint database. Candidate oncogenes and tumour suppressor genes for each tissue were obtained from Davoli et al (2013); I calculated the driver gene imbalance score as number of oncogenes - number of tumour suppressor genes with p-value < 0.05 per chromosome arm, including for each tissue all pan-cancer driver genes plus its matched tissue-specific drivers. For chromosome arm-level gene expression data in healthy individuals, I downloaded the GTEx median gene-level TPM by tissue and calculated the median TPM for all coding genes on each chromosome arm. Davoli et al (2013) and GTEx tissue types were manually matched to cancer types as shown in Appendix Table 1. I obtained GRCh38 chromosome arm lengths from the D3GB R package and

classified chromosome nuclear locations as central, peripheral or other based on Parl et al (2023)'s quartiles. Multivariate linear regression was performed in R for each cancer type, regressing chromosome arm loss rate against (i) driver gene imbalance score (ii) arm-wide expression in healthy individuals (iii) # essential haploinsufficient genes (iv) # imprinted genes (v) # genes involved only in pathogenic CNVs and with conserved copy number in mammals (vi) arm length (vii) nuclear location. For (iii), (iv) and (v), I included only genes that are normally expressed in the relevant tissue (TPM > 2 in the GTEx median gene-level TPM dataset). I chose this threshold by visualising the distribution of log2-transformed TPMs for various tissues and selecting a value that fell between the two peaks, as suggested by Cockrum et al (2020).

3.2.7 Survival Analysis

I obtained clinical data for all TCGA patients using GDCquery from the TCGAbiolinks R package and used Cox proportional hazards regression to compare overall survival between groups, controlling for cancer type, sex, age at diagnosis and self-reported race. This was done using the survival (3.7.0) and survminer (v0.5.0) R packages.

3.3 Results

I begin by expanding my study of hypodiploids to a non-TCGA cancer type where hypodiploidy is well-established and clinically relevant: acute lymphoblastic leukaemia. Using the Mitelman Database of Chromosome Abnormalities & Gene Fusions in Cancer, I obtained karyotype strings for 15,301 clones from 12,861 ALL patients and computed low-resolution copy number profiles using CytoConverter and bedtools. Many of these karyotype strings were incomplete: after excluding clones with incomplete or unparseable karyotypes, I was left with 7,926 clones from 6,910 patients. Most patient samples had one karyotyped clone, but 789 (11.4%) had two clones, 85 (1.2%) had three clones, 15 (0.2%) had four clones and 3 (0.04%) had five. I identified hypodiploid cases using standard hypodiploidy thresholds derived from the ALL literature: cases with < 30 chromosomes are defined as near-haploid (NH), while cases with 31-39 chromosomes are defined as low-hypodiploid (LH). However, because I could not reliably infer sex chromosome numbers, I reduced the standard hypodiploidy thresholds by two and counted autosomes. This allowed me to identify 301 near-haploid (<28 autosomes) and 291 low-hypodiploid (28-37) clones from 291 patients (4.2%).

With this data, I was able to recapitulate the known bimodal distribution of chromosome (autosome) counts in hypodiploid ALL: cases are split roughly evenly into 163 near-haploid and 138 low-hypodiploid clones. The peaks are

quite distinct, with only 5.3% of LH/NH cases falling into the 28-31 autosome range between the two peaks. This pattern is not seen in the hypodiploid TCGA cases, in which autosome numbers are exponentially distributed, case frequency dropping gradually with degree of hypodiploidy apart from a small group of near-haploids (Fig. 3.1). This is not an artefact of combining different cancer types: the ALL-style bimodal pattern is not seen within individual TCGA cancer types such as breast and stomach cancer either. However, hypodiploid sample size is limited in each TCGA cancer type, and many of the near-haploids in particular have since undergone genome doubling and are thus excluded from Fig. 3.1.

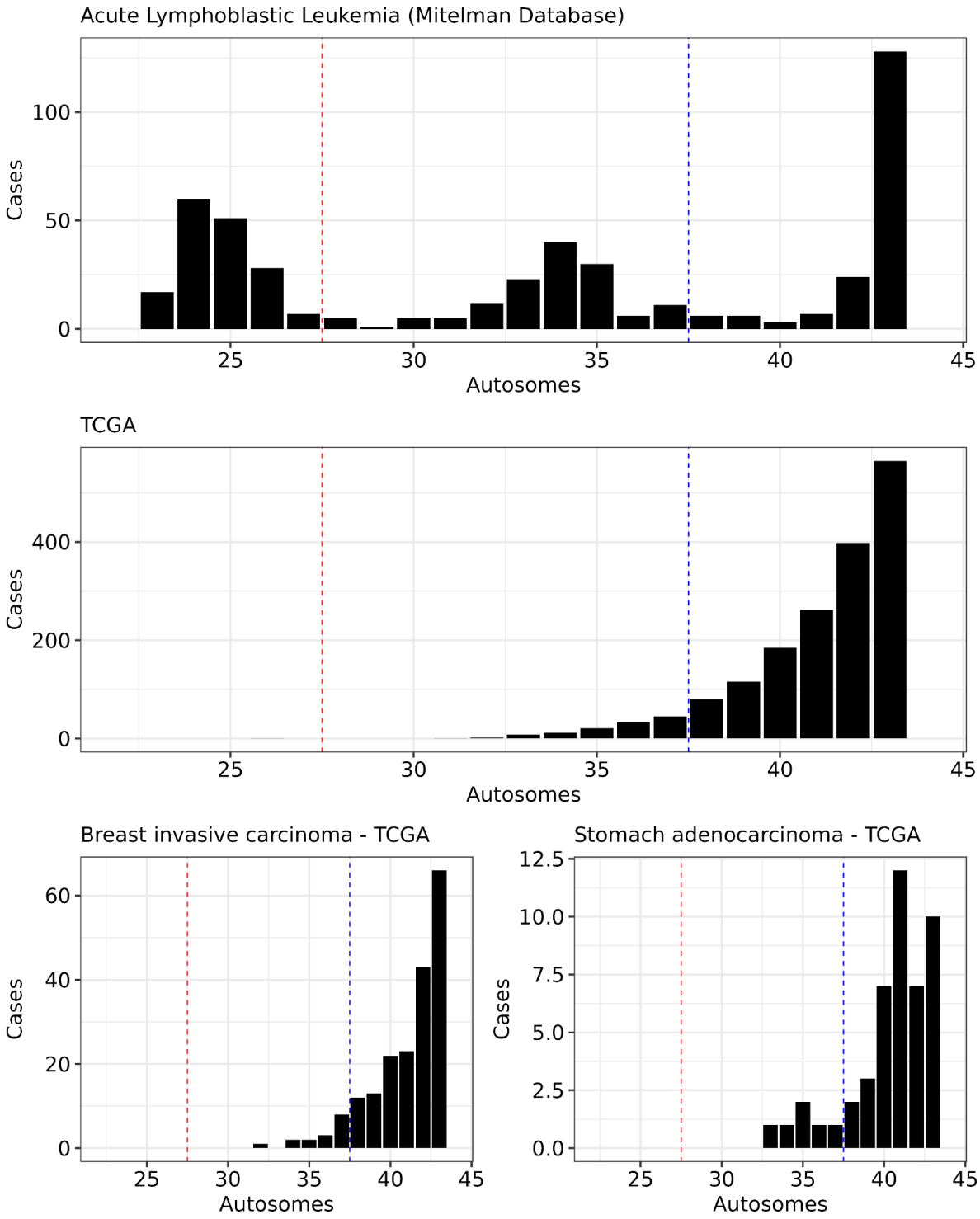


Fig. 3.1 Distribution of autosome counts among hypodiploid tumours: acute lymphoblastic leukemia cases from the Mitelman Database of Chromosome

Aberrations & Fusions in Cancer **(top),** all currently hypodiploid cases from The Cancer Genome Atlas **(middle)**, and specifically for TCGA breast cancer **(bottom left)** and stomach cancer **(bottom right)** cases. Dashed lines mark the literature-derived thresholds for low-hypodiploidy (blue) and near-haploidy (red). Only tumours which are currently hypodiploid, i.e have not since undergone large-scale chromosome gains or WGD, are included in this visualisation. Tumours with chromosome counts between 40 and 45 (autosome counts 38-43, on average), would be considered high-hypodiploid and are not considered in the following analyses.

Using loss of heterozygosity information from TCGA, I was able to identify former hypodiploids and note that only 1 of 13 near-haploids (a low-grade glioma with 26 autosomes) retained a near-haploid autosome count, with the others boasting autosome counts between 45 and 93. These included five sarcomas, three mesotheliomas, two adrenocortical carcinomas, one ovarian serous cystadenocarcinoma and one pheochromocytoma/paraganglioma. In total, I identified 186 (formerly) hypodiploid TCGA tumours that had undergone WGD and 137 that had not. Interestingly, the total proportion of hypodiploid cases is very similar between the TCGA tumours and ALL at 4.2%, but the ratio of LHs to NHs in the TCGA tumours is much higher at 310:13. Thus, even including former hypodiploids, there is a strong imbalance between near-haploidy and low-hypodiploidy in the TCGA dataset. Additionally, when near-haploids do appear, it is not generally in the TCGA cancer types with large numbers of hypodiploids.

Does the relative paucity of near-haploid cases in TCGA reflect a difference between cancer types (ALL vs TCGA tumours), or is it an artefact of different sampling strategies? ALL has by far the highest sample size for any cancer type considered here, at 6,910 patients vs 850 patients in the largest TCGA

group, breast cancer, and the ALL data includes karyotypes for multiple clones per sample, which might be hidden in bulk sequencing. More pressingly, the Mitelman Database is derived from the literature and collects exclusively chromosomally-aberrant tumours. That should include both low-hypodiploid and near-haploid ALL, but could lead to an overrepresentation of near-haploidy if near-haploid ALL has received more focus in the literature. From the literature, near-haploidy is about as common as low-hypodiploidy in paediatric leukemia but rare-to-nonexistent in adult cases (Molina et al, 2021). The TCGA cases should represent a relatively unbiased cross-section of (American) cancer diagnoses, but this data could potentially underrepresent certain unusual karyotypes if they were uncallable or less likely to be concordant between different copy number callers. To address this concern, I examined chromosome counts for some Mitelman cases of TCGA cancer types. Despite representing 16.2% of all Mitelman cases, ALL has 51.4% of low-hypodiploids and 81.6% of near-haploid cases. Out of 1,012 breast tumours, 6 are low-hypodiploid and 3 are near-haploid; of 139 stomach tumours, 1 is low-hypodiploid and none are near-haploid. The Mitelman Database annotates 266 tumour morphologies; ranking the top ten by NH proportion shows ALL at #6, with sarcomas and muscle tumours making up six of the nine others (alongside carcinoid tumours, mesothelioma and malignant peripheral nerve sheath tumours). Sarcomas and mesotheliomas have the most near-haploid cases of all TCGA tumours (5 [2.79%] and 3 [4.41%], respectively) suggesting that broad tumour type differences can be captured despite very different sampling strategies in the two databases.

3.3.1 Identifying Masked Hypodiploid ALL from Cytogenetic Data

Thus far, I have only considered hypodiploid ALL cases that remain hypodiploid, because the Mitelman Database lacks LOH information that can be used to infer a hypodiploid history post-genome doubling. However, WGD is both common and clinically relevant in hypodiploid ALL. Doubling of the hypodiploid chromosome complement creates a hyperdiploid or near-triploid karyotype, where disomic chromosomes become tetrasomic and monosomic chromosomes become disomic with loss of heterozygosity (Fig. 3.2). These "masked hypodiploids" may be mistaken for true good-prognosis high hyperdiploids (HeH-ALL) if the original hypodiploid clone is lost, but they retain the extremely poor prognosis of hypodiploids, potentially leading to improper treatment and prognostication.

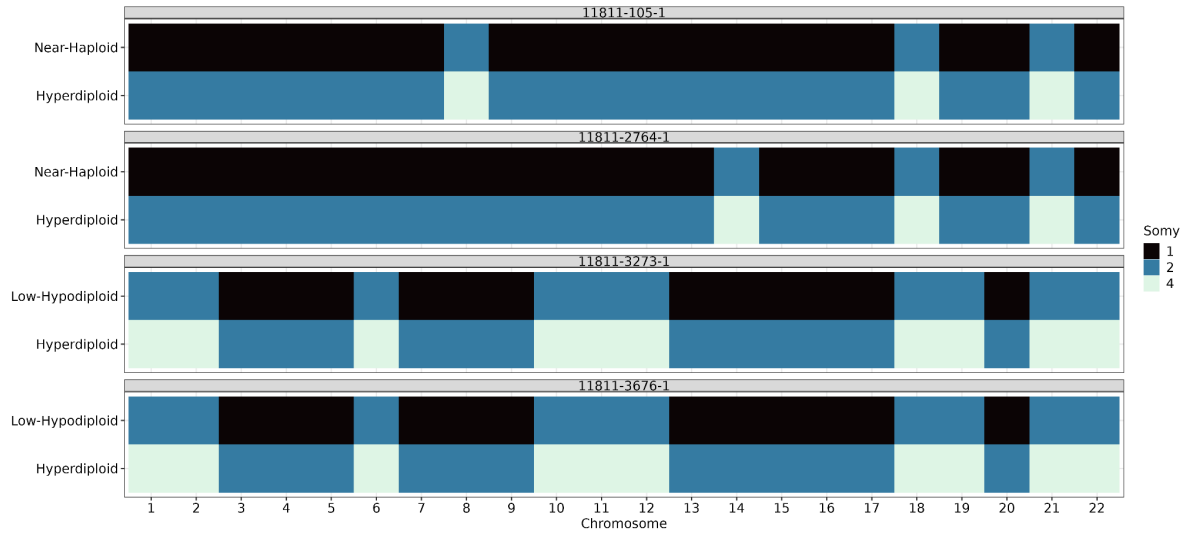


Fig. 3.2 Visualisation of genome doubling in hypodiploid ALL. Four representative cases are shown, all drawn from multi-clone Mitelman samples.The first two cases contain a near-haploid clone and a doubled hyperdiploid clone; the last two have a low-hypodiploid clone and a doubled hyperdiploid/near-triploid clone. Colours indicate copy numbers, showing perfect doubling of all chromosomes. Retained chromosomes (particularly chr21) are typical of hypodiploid ALL, and no trisomies are present.

The most reliable way to distinguish these two cases is to search for widespread loss of heterozygosity using SNP chips or sequencing, but most existing data and many routine clinical investigations are cytogenetics-based, i.e. allele non-specific. To this end, it is of both clinical and scientific interest to be able to distinguish masked hypodiploids from true hyperdiploids using only karyotype information. Previous attempts have focused on the copy number ratios of specific chromosomes: Creasey et al (2021) trained a classifier to distinguish masked low-hypodiploid from HeH ALL based on the genome-relative dosage of chromosomes 1, 7 and 14, but their model consistently misclassified near-haploid cases as HeH because chromosome 1 has a low relative dosage in both near-haploid ALL and HeH. Similarly, Gupta et al (2021) relied on ratios of ETV6:RUNX1 FISH probes and monosomies of chromosomes 7, 15 and 17. These approaches suffer from the fact that there is no chromosome consistently retained in duplicate in hypodiploid ALL that is not also gained across other ploidy subgroups.

Retreating to the intuition that masked hypodiploids will contain disomies and tetrasomies (vs disomies and trisomies in HeH-ALL) (Carroll et al, 2019), I investigated whether they could be systematically distinguished by the simple test # *tetrasomies* - # *trisomies*, where a masked hypodiploidy (MH) score > 0 indicates masked hypodiploids and a score ≤ 0 demarcates HeH. To test this heuristic, I used MEDICC2 to construct copy number phylogenies for all multi-clone samples in the Mitelman Database and identified 123 masked hypodiploids. I also obtained 577 true HeH-ALL cases using chromosome copy

numbers inferred from allele-specific sequencing by Woodward et al (2023). I show first that in HeH-ALL, the number of trisomies rises with the total number of chromosomes (r = 0.76). This relationship is much weaker in masked hypodiploids, where regardless of total autosome number the number of trisomies remains flat at < 3, and usually exactly zero, in 90% of cases (Fig. 3.3a). Visualisation of the distribution of MH scores shows that there is little overlap between the two sets (Fig. 3.3b).

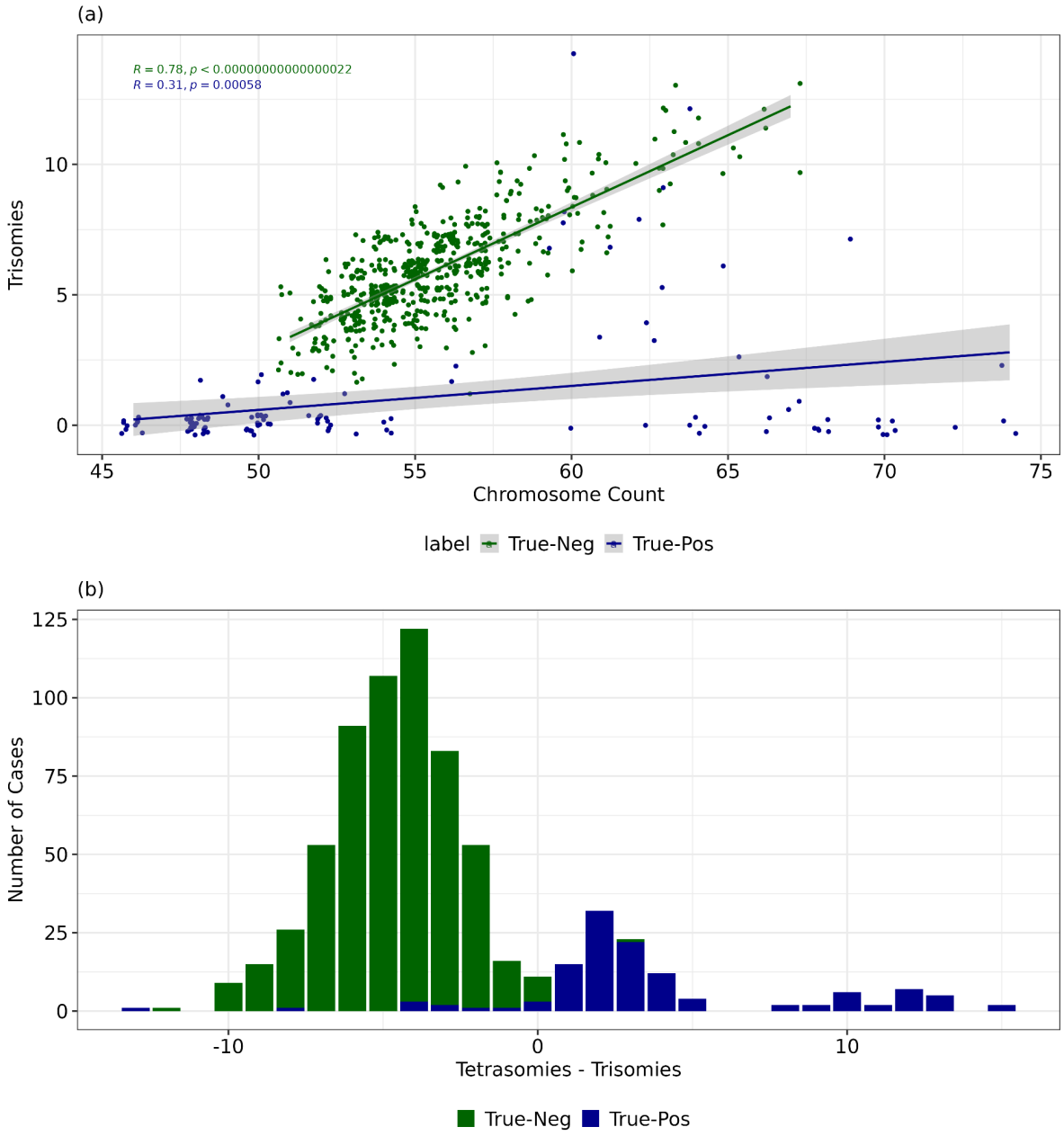


Fig. 3.3 Rationale for the MH score heuristic. (a) Increase in number of trisomies with increasing chromosome (autosome) count in HeH-ALL ('true-neg') vs masked hypodiploids ('true-pos'). A jitter effect has been added to the points for visibility. **(b)** Distribution of MH scores in each set.

Overall, I find that this test identifies masked hypodiploids with a specificity of 99.8% and a sensitivity of 90.2%. An absolute version of the test, where masked hypodiploids are defined by hyperdiploid/near-triploid karyotypes with fewer than three trisomies, has slightly weaker performance (specificity 97.9%, sensitivity 87.8%) (Fig. 3.4).

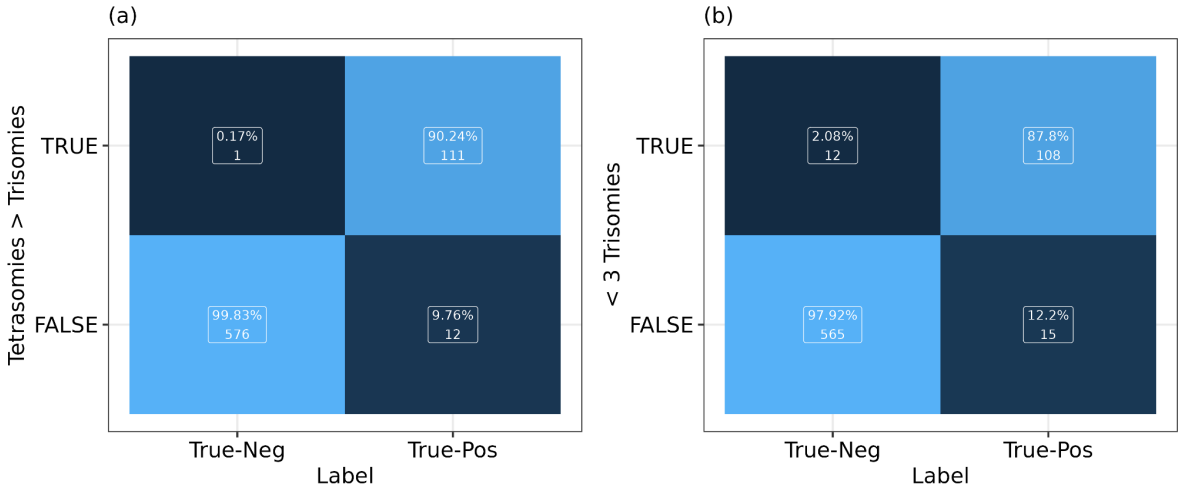


Fig. 3.4 Confusion matrices for the MH score heuristic, (a) relative version (fewer trisomies than tetrasomies) **(b)** absolute version (fewer than three trisomies). Colour indicates proportion of each set (column) with each test outcome.

This test is simple and applies to both near-haploid and low-hypodiploid ALL. However, its sensitivity is dependent on the chromosomal stability and clean doubling seen in ALL; it has no discriminative power for highly chromosomally unstable tumours, because they tend to experience additional chromosome gains and losses besides the initial hypodiploidy and WGD, with

less than half (41/100, 40%) of doubled hypodiploid TCGA tumours having a score > 0 (Fig. 3.5). Within specific cancer types with at least ten cases for which I could calculate MH score, sensitivity rises only slightly to 54.5% for adrenocortical carcinoma and 60% for breast cancer. Interestingly, however, the specificity remains relatively high (94.77%) in the TCGA tumours, confirming the intuition that it is rare to arrive at a disomies and tetrasomies-dominant hyperdiploid karyotype without a hypodiploid intermediate, even in highly chromosomally unstable cancer types.

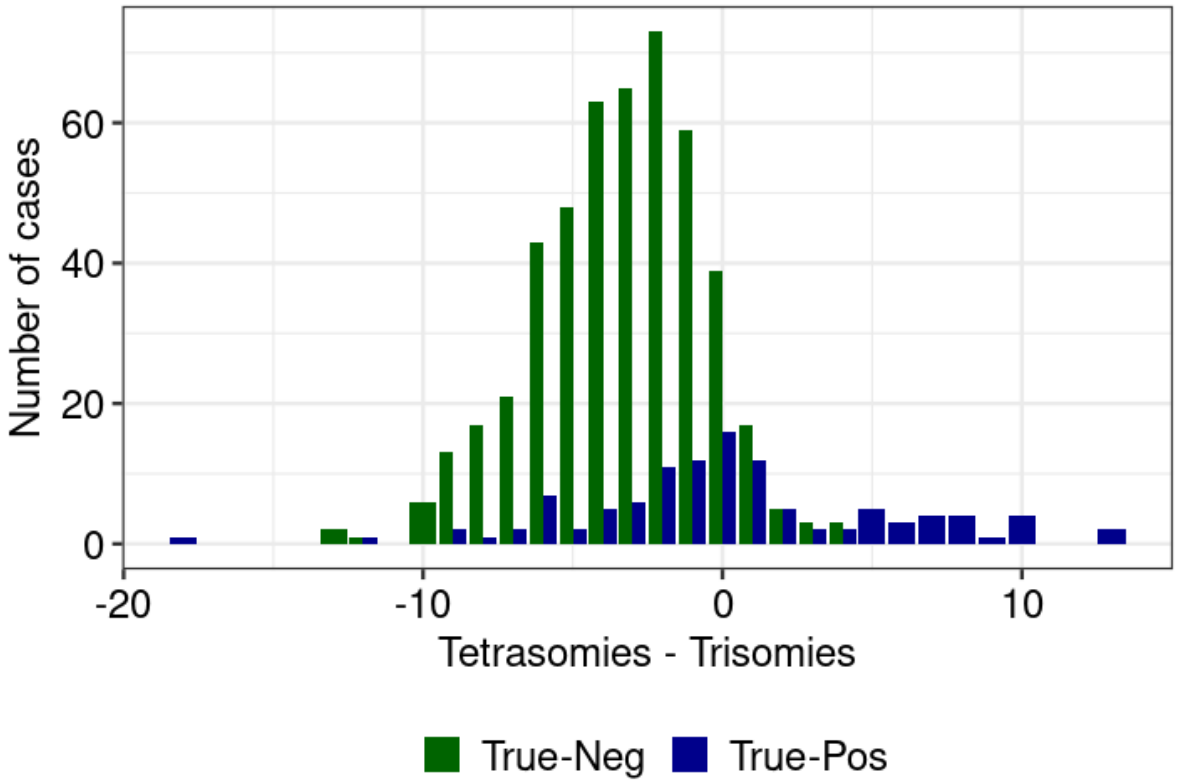


Fig. 3.5 Overlap of MH scores within the TCGA dataset.

Additionally, the heuristic works specifically in the clinically-relevant hyperdiploid to near-triploid range, so it will not detect hypodiploids that have undergone two WGDs and are now hypertetraploid, which could be

problematic if the likelihood of a first WGD and of a second are not independent. However, using allele-specific data for 294 ALL cases from the TARGET-ALL-P2 project, I could not find any examples of this scenario: 0/16 cases in the near-tetraploid range had widespread loss of heterozygosity suggestive of a low-hypodiploid stage (Fig. 3.6).

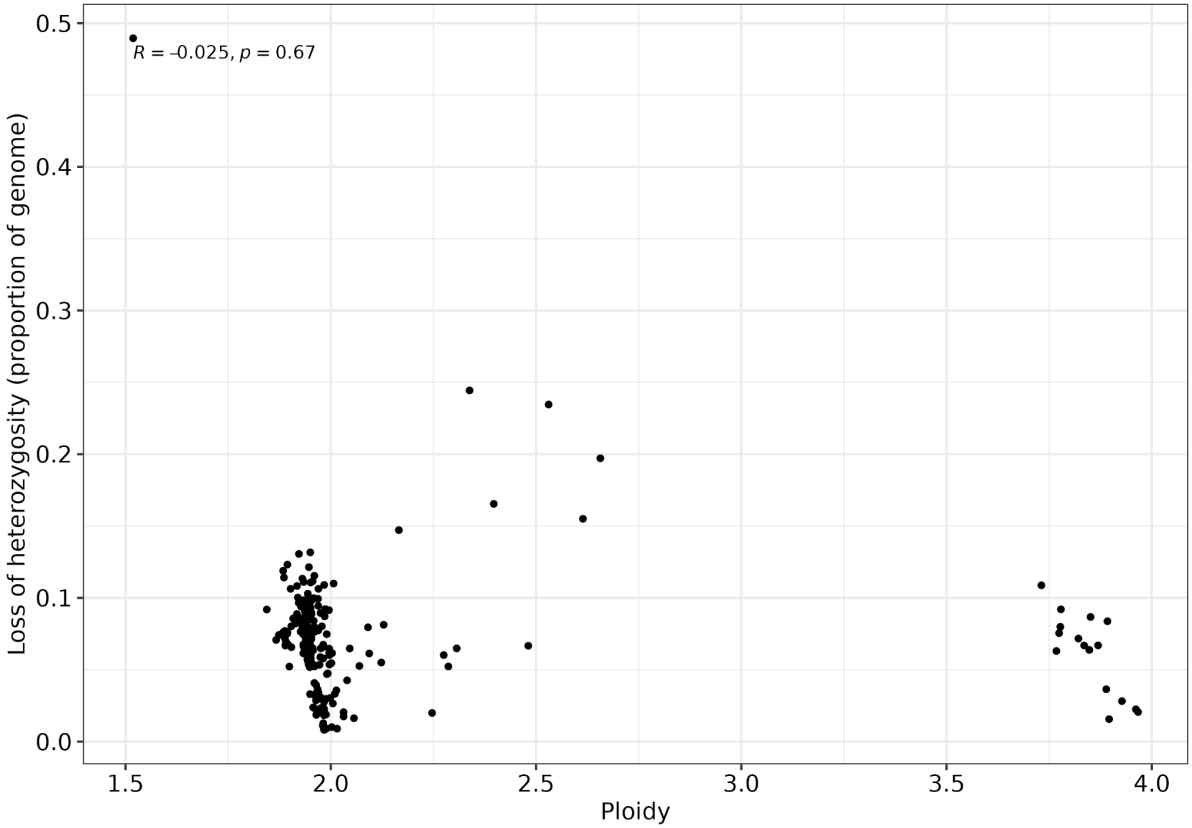


Fig. 3.6 Ploidy and loss of heterozygosity of ALL cases in the TARGET-ALL-P2 study. Former hypodiploids that have undergone two WGDs would be visible in the top right corner.

Finally, sensitivity depends on the set of masked hypodiploids included. Some masked near-haploids with very low chromosome numbers end up barely hyperdiploid (46-48 autosomes), so they don't actually reach the clinically-relevant hyperdiploid range and would not be at risk of receiving the corresponding favourable prognosis. When I exclude these cases, sensitivity

falls to 86.9%, and when I additionally exclude near-triploid cases it falls to 82.5%. The false negatives–cases with both a hypodiploid clone and a WGD-called hyperdiploid clone, but with relative MH score ≤ 0–generally deviate strongly from the criterion, having on average 7.5 trisomies (Fig. 3.7), i.e. this is not strongly threshold-dependent. These cases could constitute a subset of masked hypodiploid ALL that is chromosomally unstable after the WGD event, or, in some cases, samples that were misassigned to the 'masked hypodiploid' test set, i.e. where the doubled clone fails to show a 2:1 copy number relationship to the hypodiploid clone because the two clones are not actually related.

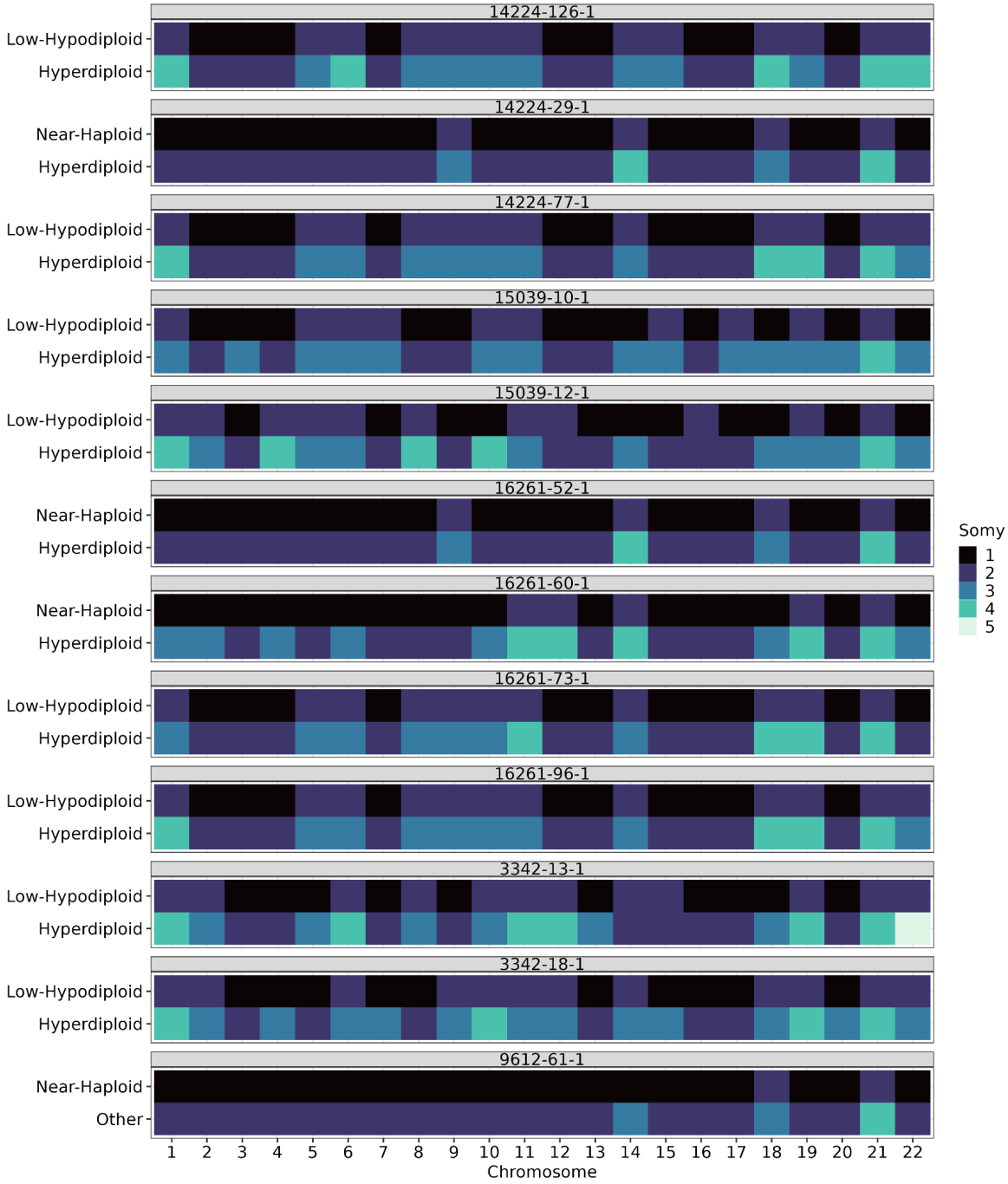


Fig. 3.7 Visualisation of genome doubling patterns in false negative cases. Each facet displays one sample with two clones, with the hypodiploid clone (near-haploid or low-hypodiploid) in the top row and its doubled counterpart below. As before, colours indicate chromosome copy numbers.

Using the MH score, I was able to identify 19 fully-masked hypodiploids—cases which I infer to have been hypodiploid but which now retain only the doubled clone–in the Mitelman database, which represents 6.2% of all

hypodiploid ALL cases. 45% of cases have a hypodiploid and non-hypodiploid (often doubled) clone coexisting, with the remaining 48.7% of cases being exclusively hypodiploid.

3.3.2 A Generalised Chromosomal Instability Phenotype: CIN at Multiple Scales

Reassigning the 19 'rescued' cases to the doubled hypodiploid group, I can estimate the rate of genome doubling in hypodiploid ALL. WGD is rare in ALL in general, occurring in only 1.2% of non-hypodiploid cases in this dataset (5% of cases in TARGET-ALL-P2), but its frequency is dramatically increased to 39% in low-hypodiploids and further still to 58.8% in near-haploids. This is likely a slight underestimate, given that I expect the test to miss 10% of doubled hypodiploids. The same pattern appears in the TCGA tumours: 27% of non-hypodiploids undergo genome doubling, vs 57% of low-hypodiploids and 85% of near-haploids (Fig. 3.8).

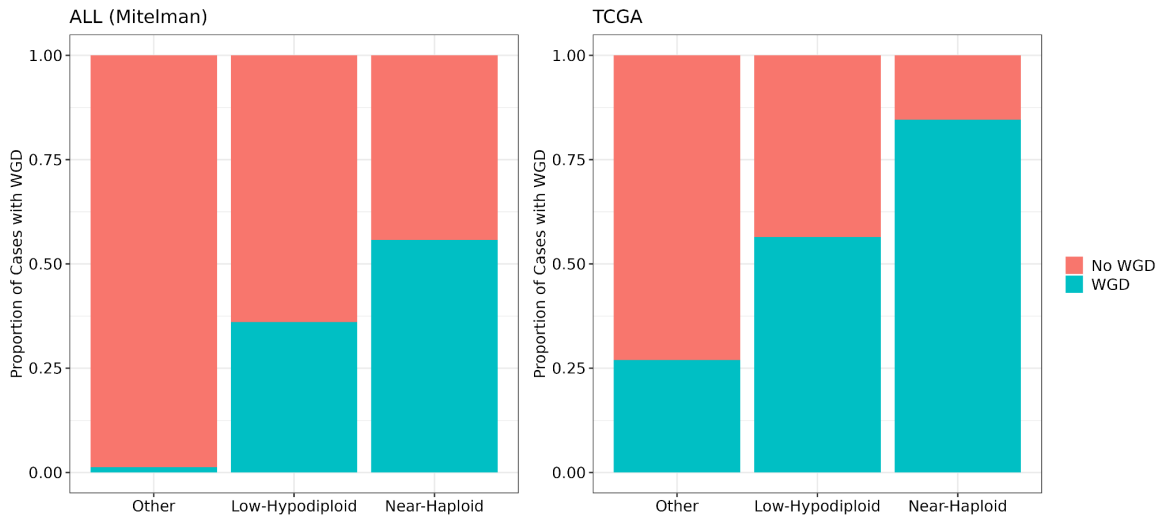


Fig. 3.8 Rate of genome doubling by hypodiploidy class in ALL (left) and TCGA (right). Genome doubling was inferred using MEDICC2, with total copy numbers and multi-clone phylogenies for ALL and allele-specific copy numbers for single bulk samples for TCGA.

This increasing rate of WGD with degree of hypodiploidy is in line with the idea that WGD is positively selected in hypodiploids to restore gene dosage (Safavi & Paulsson, 2017). However, it has been observed clinically in ALL that doubled hypodiploids do not differ in prognosis from their stemlines, and often the hypodiploid clone is the only one found at relapse (Molina et al, 2022). An alternate possibility is that the increased WGD rate simply reflects a higher underlying level of chromosomal instability. I compared levels of intra-tumour copy number heterogeneity (CNH) calculated by van Dijk et al (2021) across ploidy classes and found that both polyploids and hypodiploids have significantly higher CNH than near-diploid tumours (Fig. 3.9a). This is not exclusively due to large-scale events such as ploidy alterations and chromosome missegregation: individual chromosomes in ploidy-aberrant tumours are also more broken up by small-scale copy number alterations. This is true whether intrachromosomal CNAs are quantified by the total number of copy number segments per sample (Fig. 3.9b) or by the proportion of each chromosome in its longest copy number segment (Fig. 3.9c). These effects are robust to the removal of genome-doubled hypodiploids and to the inclusion of cancer type as a covariate.

Do polyploidy and hypodiploidy themselves induce copy number instability, or is the relationship due to a latent chromosomal instability phenotype that

causes hypodiploidy, polyploidy, high rates of chromosome missegregation and small-scale intrachromosomal copy number alterations? The answer appears to be both: there is significant support in the literature for the former explanation (e.g. Hintzen et al, 2022, Passerini et al, 2016), but I also find some evidence for the latter, complementary possibility: cancer types with high rates of WGD tend to have high CNH (Fig. 3.9d) and more intrachromosomal CNAs (Fig. 3.9e-f) even in non-ploidy aberrant cases.

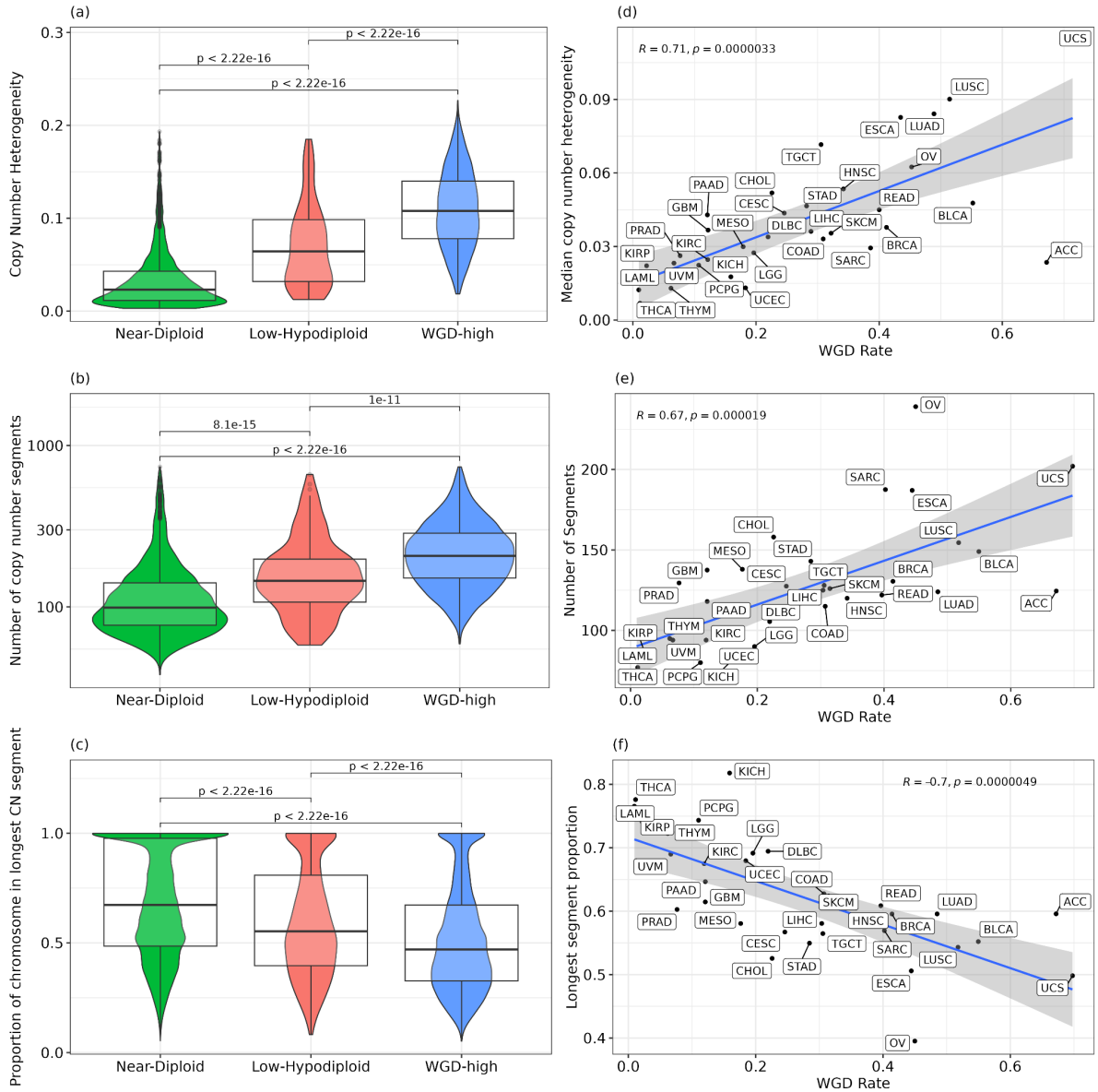


Fig. 3.9 Relationship between ploidy aberrations, copy number heterogeneity and intrachromosomal copy number alterations. In **(a)** to **(c)**, genome-doubled hypodiploids are excluded; inclusion of all (current and former) hypodiploids pushes them further from the diploid median towards the polyploids. **(a)** uses the Copy Number Heterogeneity score computed by van Dijk et al (2021) based on distance from integer ASCAT copy number profiles; this estimates diversity of copy number profiles within a single tumour. **(b)** shows the total number of copy number segments in each sample, where a segment is a contiguous region with one copy number. The data for **(c)** is computed at the chromosome (not sample) level and shows the longest copy number segment on each chromosome as a proportion of that chromosome's length. In **(d)** to **(f),** the WGD rate in each cancer type is correlated with the median of each of the values from a-c, including in the median only those samples which are neither polyploid nor hypodiploid.

3.3.3 Loss and Retention after WGD

Polyploid tumours are usually (83.4%) sub-tetraploid, with a median ploidy of 3.49 suggesting that loss predominates over gain in post-WGD evolution. However, this is partially due to pre-WGD losses: high levels of LOH have previously been linked to WGD (Lopez et al, 2020), and the median polyploid tumour (WGD-high, no hypodiploidy) has LOH across 20.6% of the genome, which after WGD would lead to a ploidy of 3.59. To put a lower bound on the amount of loss since WGD in each tumour, I used LOH information to calculate the minimum ploidy of each tumour immediately after genome doubling, if all LOH events occurred pre-WGD and there were no copy number gains pre- or post-WGD. Post-WGD losses outweighed both pre- and post-WGD gains in a majority (58.8%) of polyploids and a large minority (46.2%) of doubled hypodiploids (Fig. 3.10). The median difference was small, at 7.62% of a genome. This ranges from 33% loss-dominant in diffuse large B-cell lymphoma to 100% in AML (1 case) and uveal melanoma (5 cases) and 86.7% in pancreatic adenocarcinoma.

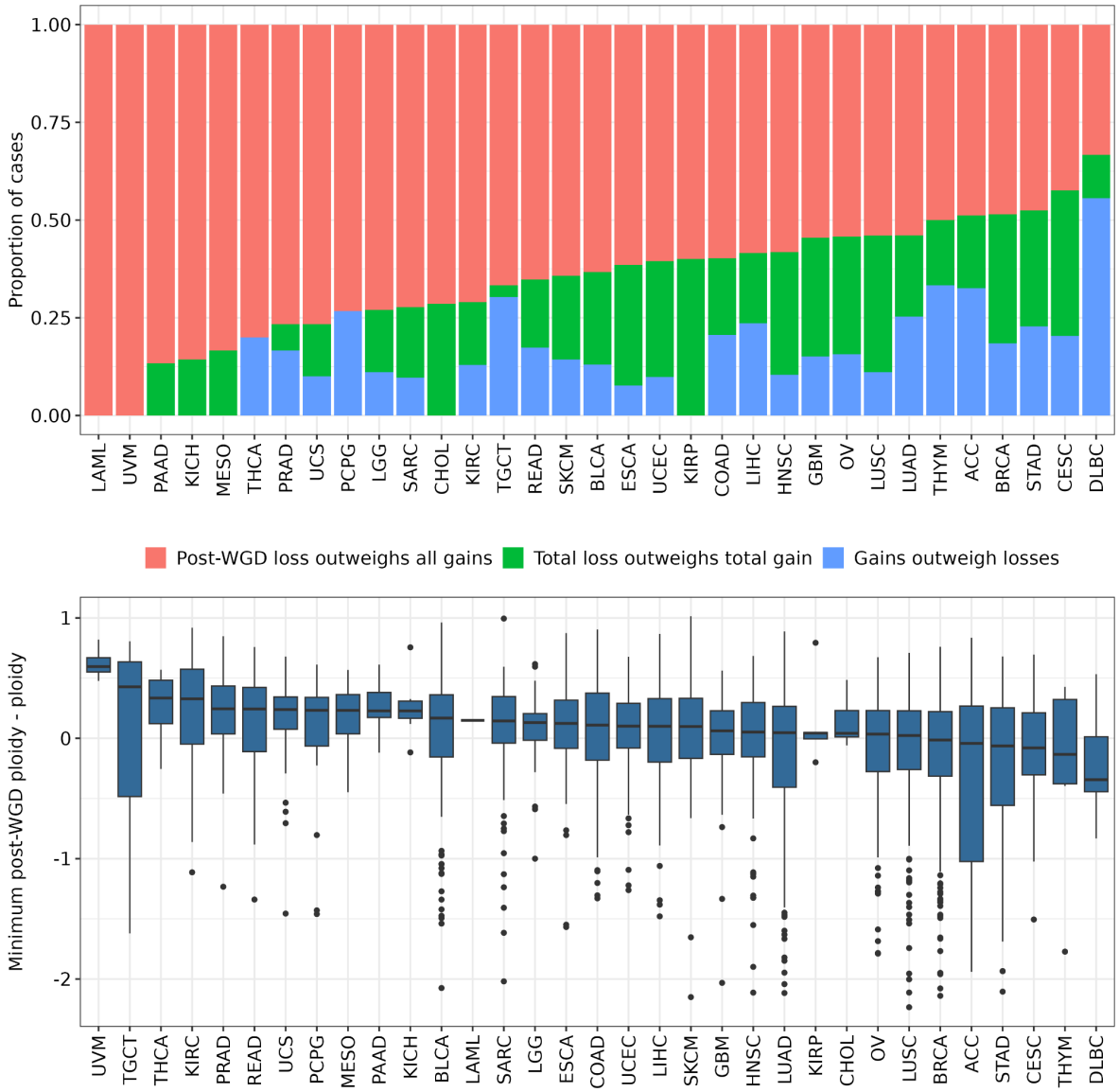


Fig. 3.10 Loss of genomic material after WGD. I calculated the minimum post-WGD ploidy for each sample, which is approximately 4 minus twice the genome-wide LOH proportion (adjusted for zero-copy segments), based on the assumptions that pre-WGD losses cause LOH and all LOH comes from pre-WGD losses. I classified samples with ploidy lower than this value as 'post-WGD loss outweighs all gains', samples with ploidy between this value and 4 as 'total loss outweighs total gain' and samples with ploidy > 4 as 'gains outweigh losses'. There were no cases with ploidy of exactly 4, although this may be a technical artefact as ASCAT cannot distinguish perfect tetraploids with no gains or losses from diploids. The bottom panel shows the difference between the minimum possible post-WGD ploidy and the current ploidy for samples in each cancer type. A value of one indicates that the amount of loss post-WGD exceeds all gains pre- and post-WGD by one genome equivalent, while a value of -1 means that total gains exceed post-WGD losses by one genome equivalent. The WGD event itself is not counted as a gain. All genome-doubled samples (polyploid and former hypodiploid) are included in this visualisation.

I calculated the proportion of genes at each copy number in polyploid tumours and found that a slight plurality (31%, on average) are in three copies, closely followed by 29.9% of genes in four copies. 14.72% of genes are in two copies, with 11.8% being two-copies-with-LOH and 2.9% showing no LOH (Fig. 3.11). 7% have increased to five copies and 3% to six, with the proportion in greater than six copies (2.4%) exceeding the proportion in one or zero (0.3%). Some of these are likely to be from tumours that have undergone multiple WGD events, which I do not explicitly address here. 90% of three-copy genes in polyploid tumours have retained heterozygosity, suggesting that the vast majority of three-copy genes represent single losses from four copies post-WGD rather than gains following pre-WGD losses. In doubled low-hypodiploids, a plurality of genes are in two copies with LOH (representing pre-WGD loss) but genes in three copies are still more likely to have been lost after WGD than before; in doubled near-haploids, two-copy genes with LOH form an outright majority and even genes in three copies tend to show LOH. Genes in four copies, which have almost universally retained heterozygosity (99.2%), suggest stability both before and after WGD. The genes retained in four copies after WGD in cancer are not enriched for genes retained in duplicate in the human lineage from the vertebrate WGDs (Appendix Fig. 6.7).

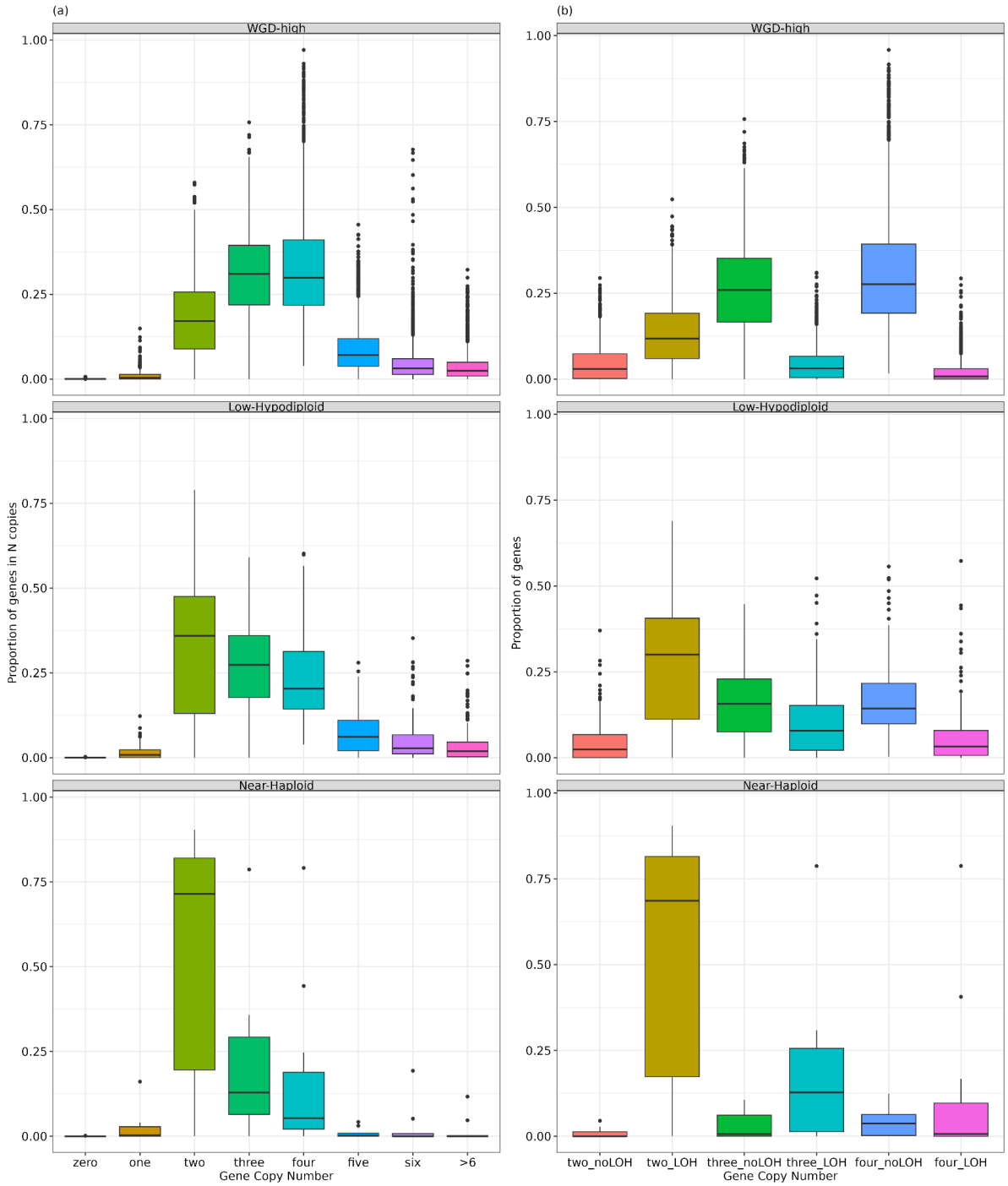


Fig. 3.11 Proportions of genes retained in different copy numbers in post-WGD samples: **(a)** total copy numbers **(b)** breakdown by LOH status. The top row is based on polyploid samples with no hypodiploid history ('WGD-high') while the bottom two rows are the two classes of former hypodiploids.

Using WGD timing estimates calculated by Gerstung et al (2020) based on duplicated mutations, I examined how post-WGD gene copy number proportions change over time. The proportion of genes in four copies in polyploids decreases non-linearly with time since the WGD event, with a period of rapid copy number change followed by stasis. The copy number change is mostly driven by losses to three or two copies–there is no clear relationship between WGD timing and gains to five or six copies (Fig. 3.12). However, there is significant variation: one lung adenocarcinoma case had retained 40.3% of genes in four copies more than 43 years after the WGD event, but a cervical cancer case that underwent WGD eight months before sequencing had only 13.4% of genes in four copies. We lack insight into the very first stages of post-WGD tumour evolution because all TCGA tumours for which there is timing data underwent WGD at least three months before sequencing.

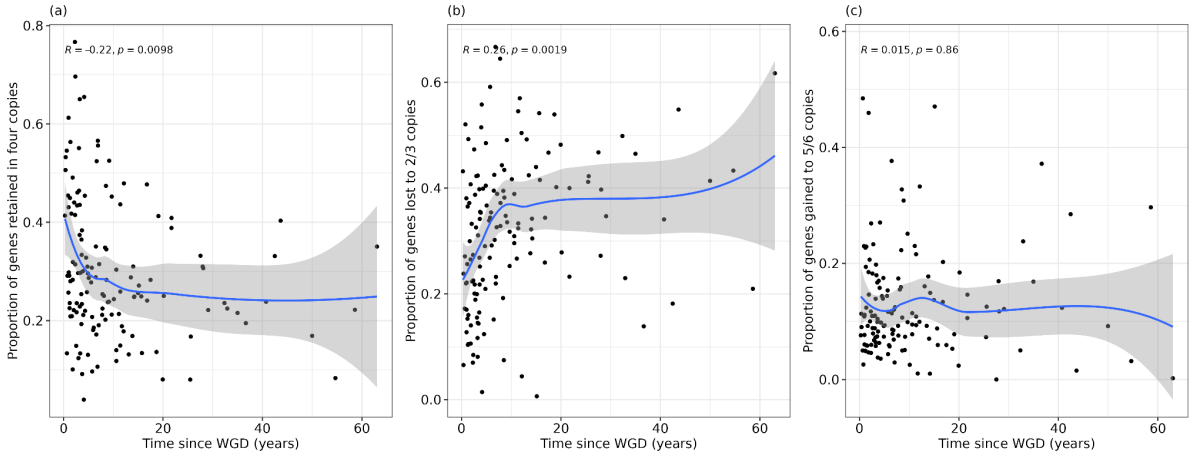


Fig. 3.12 Change in gene copy number proportions over time: **(a)** proportion of genes in four copies **(b)** proportion lost to two or three copies, excluding genes with LOH indicating pre-WGD loss **(c)** proportion gained to five or six copies. Genes in five or six copies are used as a measure of post-WGD gain, although in some cases they will have been gained pre-WGD.

In general, gene copy number proportions vary significantly between cancer types. The median proportion of genes retained in four copies, for example, ranges from 24.2% in lung squamous cell carcinoma to 90.5% in acute myeloid leukaemia (1 case) or 78.3% in low-grade glioma. Based on my results in Fig. 3.9, I hypothesised that this might be due to cancer type-specific levels of chromosomal instability. I thus compared the proportion of genes in four copies in polyploid tumours to the WGD rate in the corresponding cancer type and found a moderate to strong inverse correlation (r = -0.6), indicating that less WGD-prone cancer types tend to retain more genes in four copies after WGD (Fig. 3.12a). The correlation persists (r = -0.57, p = 0.0011) after removing cancer types with fewer than five polyploids. This has the interesting implication that tetraploidy in low-WGD tumours lasts longer — i.e., rare polyploids are stable polyploids. As a corollary, I correlated WGD rate with median post-WGD gene loss (Fig. 3.13b) and gain (Fig. 3.13c). As expected, this correlation is in the opposite direction in both cases, albeit non-/marginally significant.

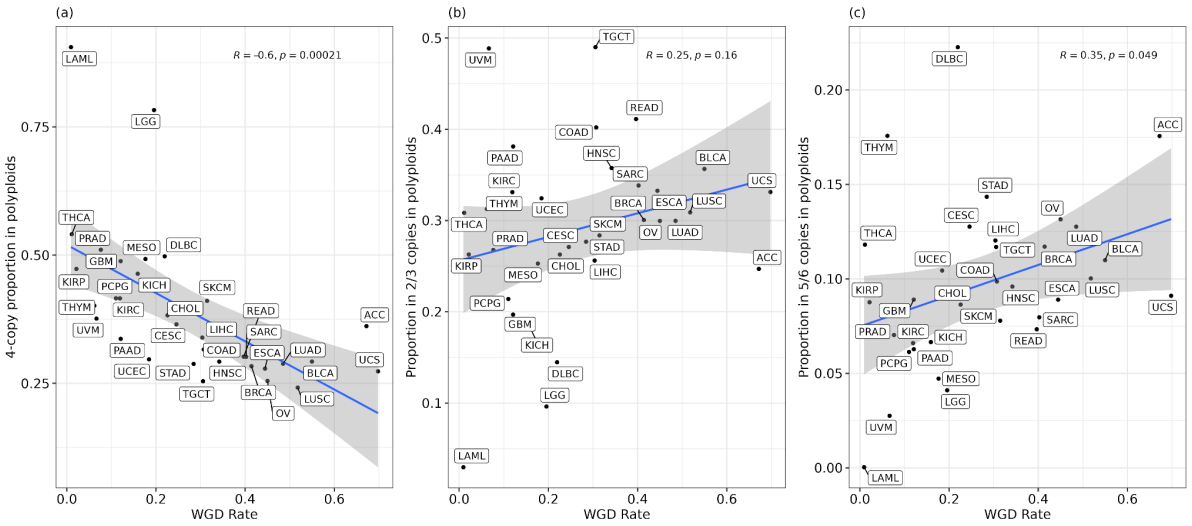


Fig. 3.13 Relationship between the WGD rate in a cancer type and the median proportion of genes (a) retained in four copies **(b)** lost to two or three

copies (without LOH) **(c)** gained to five or six copies after WGD. Only polyploid tumours with no hypodiploid history are included.

One possible alternative explanation for this correlation between WGD rate and gene retention in four copies is a link between WGD prevalence and WGD timing. If WGDs in a given cancer type tend to occur early in tumour development, they are likely to be present and detectable at diagnosis (sequencing time) in a larger subset of cases. Simultaneously, if copies are lost or gained over time, then cancers with more recent WGDs will also tend to have retained more of the genome in four copies. On the other hand, late WGDs allow more time for pre-WGD gains and losses, which after genome doubling could result in gene copy numbers of six or two, respectively, rather than exactly four. When I regress four-copy proportion on WGD rate and include median WGD timing as a covariate, WGD rate remains a significant predictor ($p = 4.92e\text{-}5$). Additionally, there is no significant correlation between the average timing of WGD events in a cancer type and the rate of WGD in that cancer type ($r = -0.058$, $p = 0.82$). However, sample sizes of timed WGDs in each cancer type are low (median 8, range 1-38) and so the estimates of average WGD timing may not be reliable. Another possible confounder is pre-WGD loss rate, in that tumours with high levels of pre-WGD loss will necessarily have a lower proportion of genes in four copies immediately after WGD regardless of post-WGD loss dynamics. I regressed median four-copy proportion post-WGD on WGD rate and median LOH across the genome and found that median LOH is indeed negatively associated with median four-copy proportion, but WGD rate remains a significant negative predictor ($p = 0.0194$). The regression coefficient is -0.306 (vs -0.47 before

adjusting for LOH), indicating that a 10% increase in tissue WGD rate is associated with a 4.7% decrease in median four-copy proportion after WGD.

If rare polyploids are stable polyploids, we may be able to use these timing estimates to directly observe this stability (or lack thereof) over time. Visualising three cancer types with at least 15 polyploid cases, I note first that there do appear to be differences in rate of copy number change: slow in breast cancer, intermediate in lung squamous cell carcinoma, rapid in endometrial cancer (Fig. 3.14). However, as there are only three cancer types with >= 15 WGD-timed polyploid samples (eleven cancer types with >= 5), there are not enough samples to directly test the relationship between WGD rate and duration of duplicate gene retention across cancers.

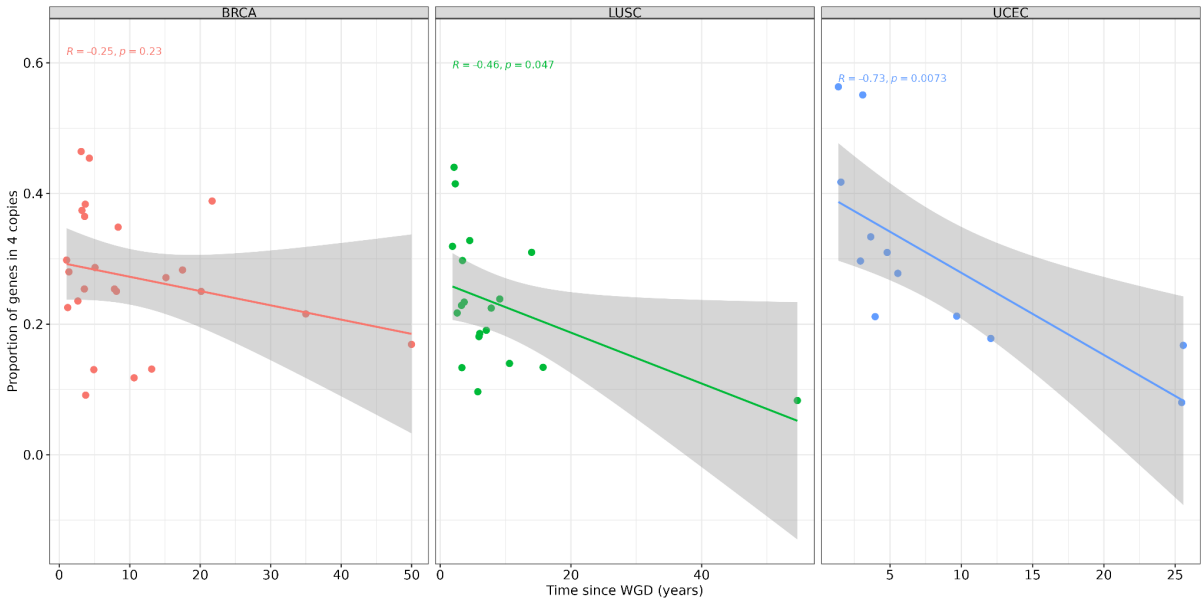


Fig. 3.14 Proportion of genes in four copies over time since WGD in three cancer types: breast, lung squamous cell, and endometrial. WGD timing estimates are taken from Gerstung et al (2020) and are based on numbers of duplicated mutations across the genome.

3.3.4 Patterns of Chromosome Loss in Hypodiploid and Polyploid Tumours

While hypodiploid and polyploid tumours are highly chromosomally unstable and chaotic in general, some (particularly ALL) have patterns in which chromosomes or chromosome arms tend to be lost. In this section, I characterise loss rates in hypodiploids (pre-WGD) and polyploids (post-WGD) and search for factors that explain differential chromosome loss rates. I am interested in understanding (i) whether chromosomes in these highly unstable tumours are lost at random (ii) whether there are similarities between hypodiploids and polyploids in terms of which chromosomes are lost (iii) whether loss patterns can be explained by selective factors such as the presence of essential haploinsufficient genes or of cancer driver genes, or mechanical ones such as chromosome location in the nucleus.

On the first point, I find that chromosomes are lost at non-uniform rates in hypodiploid tumours (chi-square test of uniformity, $p < 1e\text{-}5$ for each cancer type). The most extreme deviations from uniformity are seen in low-hypodiploid ALL (LH-ALL), kidney chromophobe cancer and especially near-haploid ALL (NH-ALL), in which the top three most-retained chromosomes (21, 14, 18) make up 77% of all retained chromosomes. Kidney chromophobe cancer (KICH) is known for characteristic loss of seven chromosomes (1, 2, 6, 10, 13, 17, 21) (Casuscelli et al, 2017; Davis et al, 2014; Speicher et al, 1994): their loss rates as calculated here range from 100% (chr17) to 76.9% (chr21), with the next highest loss rate being in chr8 at only 30.8%. Some cross-tissue

patterns can be seen in the hypodiploid TCGA tumours, with chromosome 7 being retained in >90% of cases across 6/7 cancer types in line with previous reports, but there remains significant variation both intra- (Fig. 3.15a) and inter-tissue (Fig. 3.15b). For example, chromosome 1 is retained in 94.2% of LH-ALL but 0% of NH-ALL cases.

Loss rates are generally similar between arms of the same chromosome. In adrenocortical carcinoma, kidney chromophobe and ovarian serous cystadenocarcinoma, as well as both near-haploid and low-hypodiploid ALL, inter-arm loss rate correlations are >= 0.95, suggesting that whole-chromosome losses are much more common than arm-level losses in hypodiploid tumours. The correlation is less strong in BRCA (R = 0.8), LUAD (0.71) and STAD (0.62), and weakest in LUSC (0.45). These cancer types have several outlying chromosomes with imbalanced losses between their arms, including chromosome 3 in LUSC (p-arm lost in 83.3% of cases, q-arm lost in only 0.067%) and chromosome 5, which loses its q arm significantly more often than its p arm in LUSC, STAD and OV (Fig. 3.15c). This is likely caused by oncogenic LOH of regions containing tumour suppressor genes such as APC (5q) and VHL (3p). LOH in these regions is common in multiple types of cancer including lung (e.g. Mendes-da Silva 2000, Zabarovsky et al 2002) and may occur as an initiating event, before the onset of general chromosomal instability and large-scale chromosome loss.

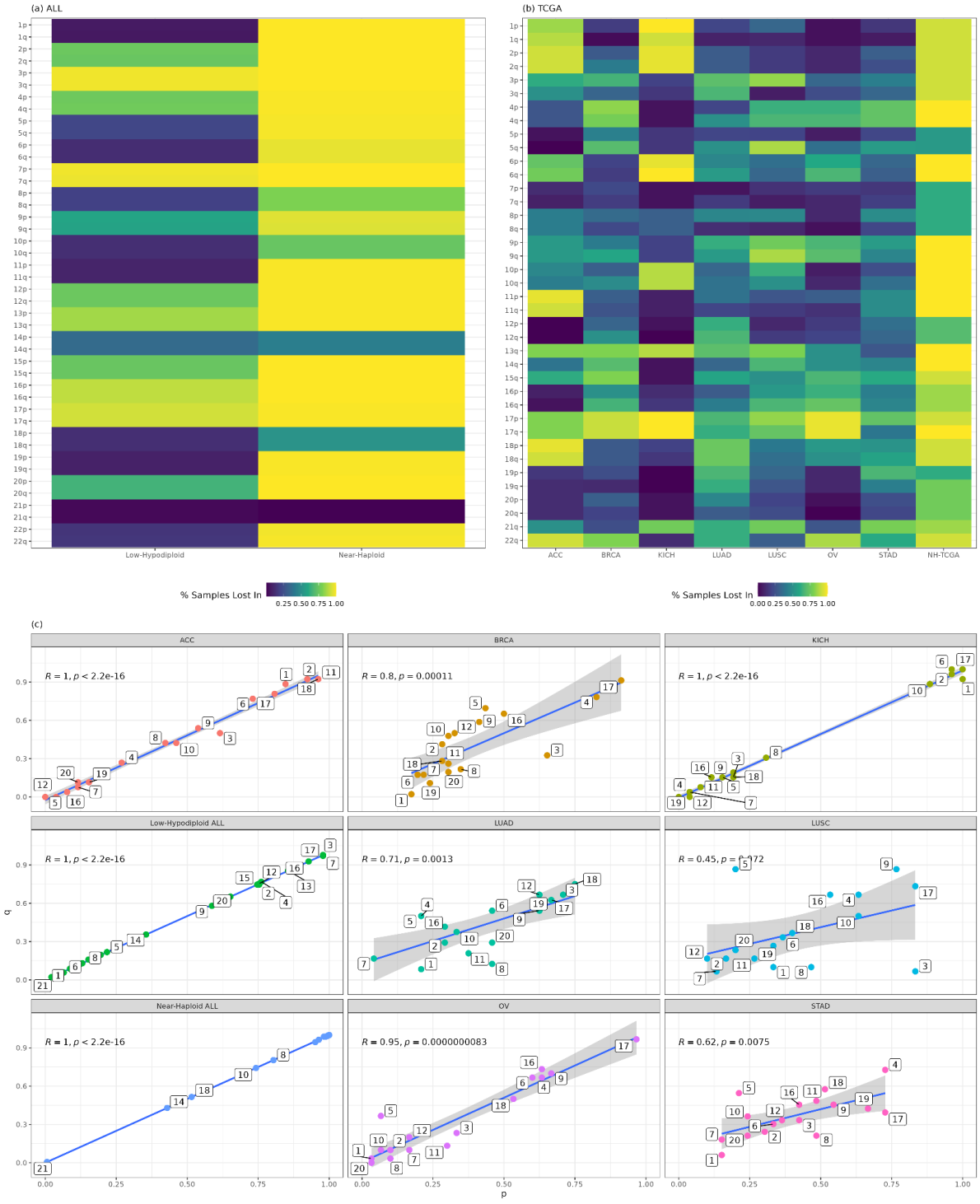


Fig. 3.15 Chromosome arm loss rates in hypodiploids by tissue Proportion of ALL cases from **(a)** the Mitelman database and **(b)** TCGA in which each chromosome arm is lost. In (a), only current hypodiploids are shown; in (b), loss is inferred by LOH and includes former hypodiploids. Loss rates were calculated among cancer types with at least 15 hypodiploids. **(c)** Correlation between loss rates of the p- and q-arm on each chromosome.

Comparing chromosome arm loss rates between hypodiploids and other classes, I find notable similarities: there are moderate-to-strong correlations between loss (LOH) rates in hypodiploids and in near-diploids (r = 0.54, Fig. 3.16a) and intermediate-ploidy cases, which have too many CNAs to be classed as near-diploid but are not classed as polyploid or (low-)hypodiploid (r = 0.71, Fig. 3.16b). They are also correlated with loss rates in polyploids, although moreso with pre-WGD losses (r = 0.67) than post-WGD (r = 0.42).

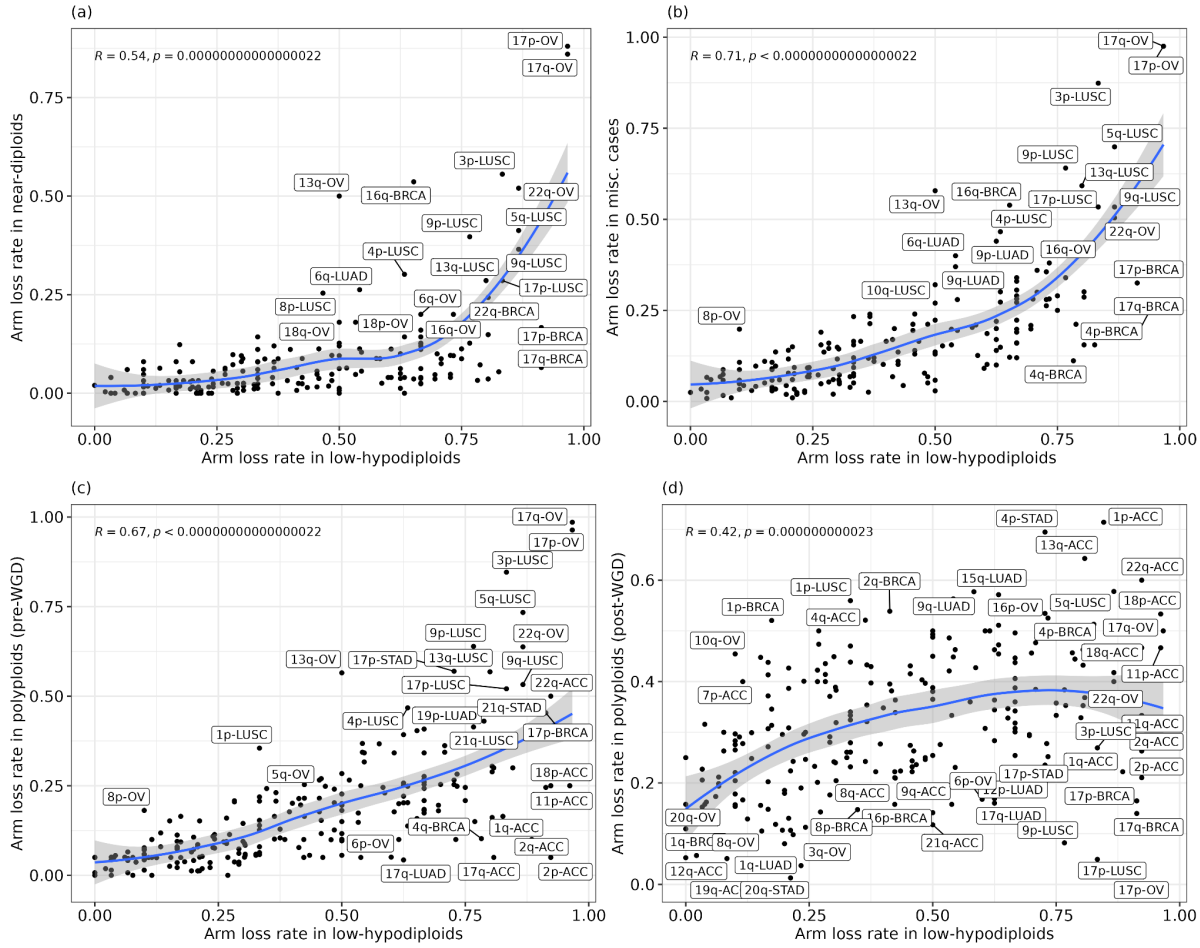


Fig. 3.16 Correlations of chromosome arm loss rates between ploidy classes: low-hypodiploids vs **(a)** near-diploids **(b)** intermediate-ploidy cases **(c)** polyploids (pre-WGD losses only) **(d)** polyploids (post-WGD losses only). In polyploids, a pre-WGD loss was inferred from LOH across >90% of the chromosome arm (as with losses in the other ploidy class), and a post-WGD loss was identified when the chromosome arm was in two or three copies with LOH across <20% of its length. Because a post-WGD loss and a pre-WGD loss are mutually exclusive in this accounting, I corrected post-WGD loss rates by removing cases with pre-WGD loss

from the denominator. Only cancer types with at least 15 cases for each of the classes being compared were included.

I also directly compared loss patterns in polyploids before and after WGD (Fig. 3.17), and found that they are moderately correlated ($r = 0.32$, $p < 2.2e-16$), with notable outliers including 17p (which hosts TP53) and 9p (CDKN2A). As reported recently by Baker et al (2024), these chromosome arms are disproportionately likely to be lost pre-WGD. 17p is extreme: in four cancer types, its loss rate falls from 52-96% pre-WGD to <5% post-WGD, even though I adjusted post-WGD rates to include in the denominator only cases without a pre-WGD loss. If 17p loss rates are primarily driven by TP53, it seems plausible that there is no longer selection on TP53 loss post-WGD because cancers that manage to undergo WGD without losing TP53 have already found an alternate route to allowing chromosomal instability. Other outliers favouring pre-WGD loss included 3p (VHL) in kidney papillary and lung squamous cell carcinoma, 1p in low-grade glioma and 8p in prostate adenocarcinoma. In the other direction, I identified outliers that are disproportionately likely to be lost post-WGD: of the 30 outlying chromosome arm-tissue combinations, eight are in chromosome 4 and six in chromosome 10. Different cancer types can have opposite timings: despite being lost early in other cancer types, 1p and 9p are lost disproportionately late in adrenocortical carcinoma and kidney clear-cell carcinoma, respectively. Chromosomes preferentially lost late might carry tumour suppressor genes that only act in concert with previous losses, or represent arms under dosage constraint, where losses on a tetraploid background are more tolerable than on a diploid background.

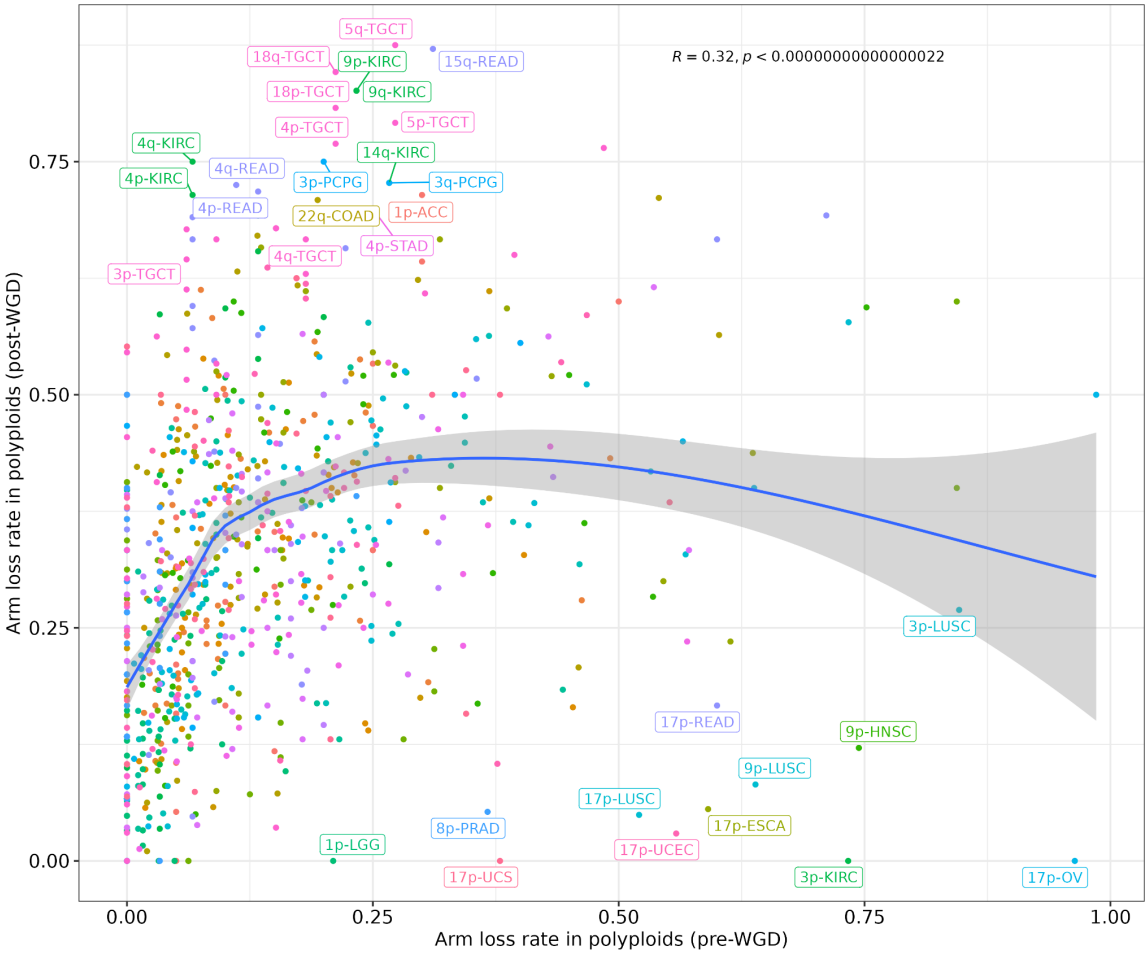


Fig. 3.17 Chromosome arm loss rates in polyploid tumours, pre- vs post-WGD. Loss timing inferred as described in Fig. 3.16. Labels indicate outliers (defined as arm-tissue combinations with absolute standardised residual > 2).

Taken together, these results suggest that there are common patterns of chromosome loss shared between different ploidy classes or chromosomal instability regimes, i.e. that while the overall level of aneuploidy (and total number of chromosomes lost) differs between classes, there is a shared rank order determining which chromosomes tend to be lost or retained. However, this is to some extent driven by particular chromosomes under strong selection for loss: within the [0, 0.25] range the trend is weak, so there is not necessarily a specific rank order for every chromosome.

Given that aneuploidy patterns are at least to some degree tissue-specific, I next attempted to determine whether chromosome loss profiles in hypodiploids cluster by tissue. Hierarchical clustering of TCGA samples by aneuploidy patterns has shown some relationship with tissue of origin, particularly for squamous tumours (Taylor et al, 2018). However, when all ploidy classes are considered together, clustering tends to be based on the overall level of aneuploidy (Hoadley et al, 2018). WGD makes some difference to the identity of chromosomes lost: aneuploidy patterns are more even across the genome in tumours that have undergone WGD (Prasad et al, 2022), suggesting that additional aneuploidies can occur on top of the characteristic positively-selected ones for a given cancer type either because polyploids are more permissive of aneuploidy or because of their high level of chromosomal instability. Hypodiploid tumours are similarly chromosomally unstable, and it is unclear to what extent their loss patterns are characteristic of their tissue of origin. To address this question, I created binary chromosome arm loss profiles for each hypodiploid sample and performed hierarchical clustering based on the pairwise Jaccard distance between samples. I find that samples of LH-ALL, NH-ALL, kidney chromophobe cancer and adrenocortical carcinoma form visible clusters, while the other cancer types are intermixed (breast, lung, ovarian, stomach) (Fig. 3.18). This is confirmed by computation of the median Jaccard similarity (1 - distance, scale between 0 and 1) within and between cancer types: NH-ALL has the highest median similarity at 0.912, with kidney chromophobe at 0.688, low-hypodiploid ALL at 0.591 and adrenocortical at 0.56. In breast, lung adenocarcinoma, lung squamous,

stomach, and ovarian cancer, the median Jaccard similarity for pairs of samples within a cancer type is not higher than the similarity with other cancer types. To measure the ability of chromosome loss profiles to cluster hypodiploid samples into cancer types overall, I performed PAM clustering with k = 9 (the seven TCGA low-hypodiploid groups, plus LH-ALL and NH-ALL) and obtained an adjusted Rand index (ARI) of 0.4. The ARI measures how much better a clustering is than chance, with 0 indicating random assignment to clusters and 1 indicating a perfect match between clusters and labels; 0.4 suggests a weak-to-moderate agreement. However, this is strongly dependent on the highly divergent and consistent patterns seen in NH-ALL: when I remove NH-ALL and include only low-hypodiploid groups, the ARI (k = 8) falls to a weak 0.218. This indicates that hypodiploid samples do not generally cluster well into tissue types based on their chromosome loss profiles. Interestingly, chromosome loss profiles within near-diploid samples also do not cluster by tissue (ARI 0.038), corroborating the point in Hoadley et al (2018) that aneuploidy-based clustering is mainly based on large-scale changes and so is not effective within ploidy strata.

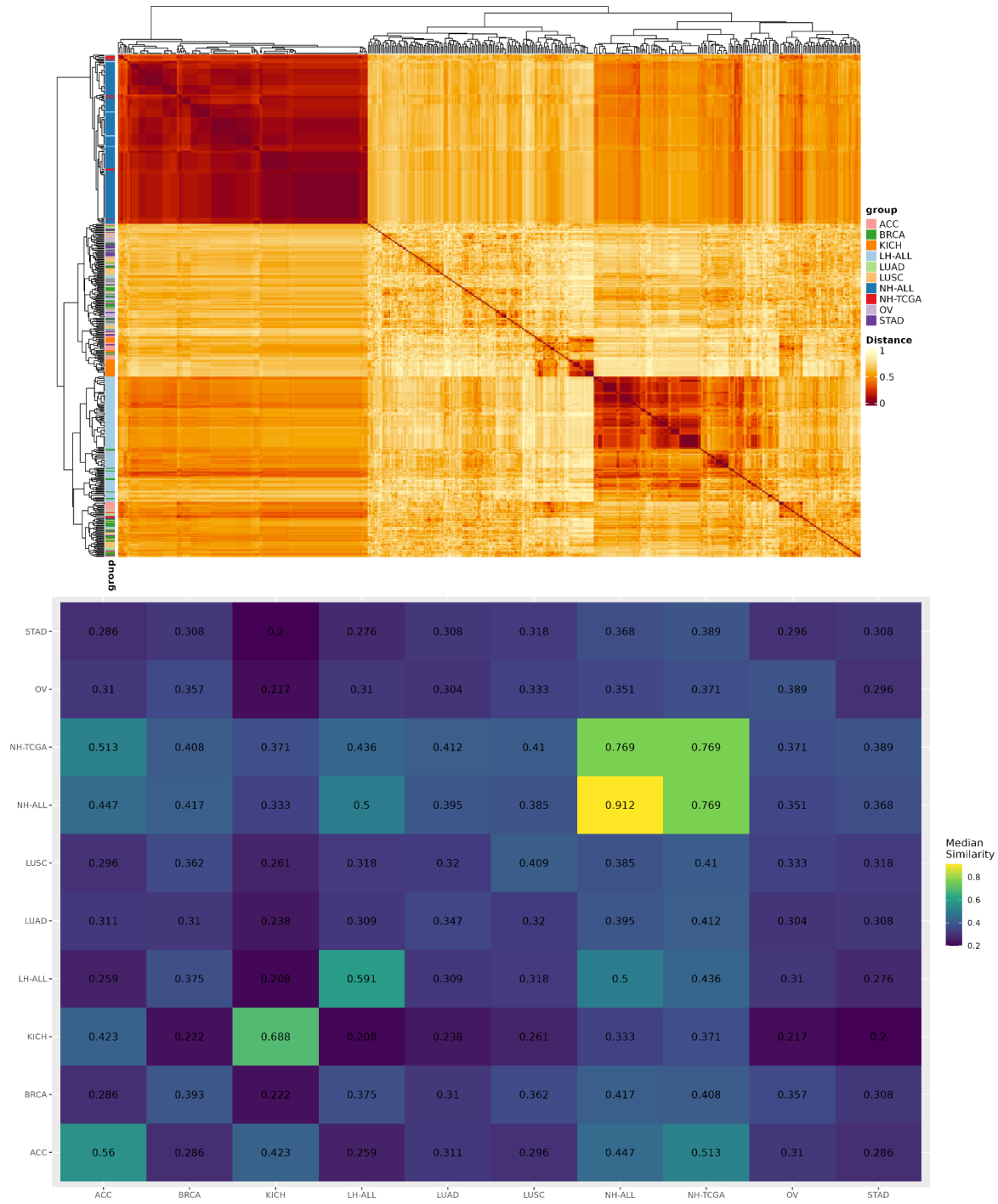


Fig. 3.18 Clustering of hypodiploid tumours based on chromosome losses. I computed the Jaccard distance between binary chromosome arm loss profiles pairwise for every sample and performed hierarchical clustering **(top)**; colours in the left bar indicate cancer types. I then calculated the median Jaccard similarity within and between cancer types **(bottom)**. NH-TCGA indicates a grouping of the 13 near-haploid TCGA tumours from six different cancer types.

Can one explain these loss patterns with respect to features of each chromosome (arm)? Focusing on hypodiploids, I collated a set of features that have been suggested to or could plausibly affect loss rate, including:

- length: Duijf et al (2014) report that small chromosomes are more frequently lost, while Mandahl et al (2012) find no relationship between chromosome size and loss rate in near-haploid tumours, and Klaasen & Kops (2022) report the opposite relationship
- location (central/peripheral/other) in the cell nucleus (Klaasen et al, 2022) as inferred from interchromosomal translocation frequency (Parl, 2023)
- driver genes on the chromosome (OG - TSG score) (e.g. Davoli et al, 2013; Watkins et al, 2020)
- expression level of genes on that chromosome in healthy tissues (Patkar et al, 2021)
- presence of essential imprinted genes (Boot et al, 2016)
- number of dosage-sensitive genes, which in the germline are constrained against copy number alterations due to e.g. haploinsufficiency (e.g Rice & McLysaght, 2017)

I performed linear regression of chromosome arm loss rate against arm length, chromosome location in the nucleus (peripheral/central/other), median expression from that chromosome arm in normal tissues, and the OG - TSG score, i.e. the difference between the number of oncogenes and the number of tumour suppressors on the chromosome arm (Fig. 3.19). I find that the OG -

TSG score is associated with a decrease in loss rate in hypodiploid adrenocortical carcinoma (ACC) ($p_{adj} = 0.016$) and is nominally significant in the same direction in kidney chromophobe (KICH), but the opposite is seen in low-hypodiploid ALL. Longer chromosomes are significantly more likely to be lost in ACC, but show no significant relationship in either direction in other cancer types. Median expression from the chromosome arm was not significantly associated with loss rate in any cancer type. Peripherally-located chromosomes had a nominally-significant increased loss rate in lung adenocarcinoma but showed the opposite pattern in near-haploid ALL.

While it has been suggested that chromosome 7's high retention rate in solid tumours is due to the imprinted essential genes it carries (Boot et al, 2016), I find that the number of imprinted genes on a chromosome arm does not predict its loss rate. Indeed, chromosome 15q is the arm with the most imprinted essential genes (6), and it is lost in the majority (56%) of TCGA hypodiploids. This may be possible due to loss of imprinting, which has been observed for specific essential imprinted genes (e.g. IGF2) in cancer, or because most germline essential imprinted genes are not essential in the cancer context.

I also included two measures of dosage sensitivity in the regression: genes under strong copy number constraint in the human germline and throughout mammalian evolution (Rice & McLysaght, 2017), and genes identified as being essential and haploinsufficient via a screen in human embryonic stem

cells (Sarel-Galilly et al, 2022). To account for genes not expected to be relevant in cancer or with tissue-specific importance, I included copy-constrained genes only if they were (a) expressed in the relevant tissue in GTEx, and (b) essential in cancer cells based on CRISPR screens (Pacini et al, 2021). Neither of the dosage sensitivity-related features were significantly associated with loss rate, suggesting that avoidance of monosomies affecting dosage-sensitive genes is not a major factor in the evolution of hypodiploid tumours. I note that germline copy-constrained genes are enriched for cancer genes, with 4.89% of copy-constrained genes listed in the COSMIC Cancer Gene Census as opposed to 0.84% of non-copy-constrained genes (chi-square $p < 2.2e-16$, Fig. 3.19c). However, the copy number alteration rate of these genes in cancer does not significantly differ from that of non-copy constrained genes (Appendix Fig. 6.9), and the results of the linear regression are unchanged when I remove cancer genes from the set of copy-constrained genes.

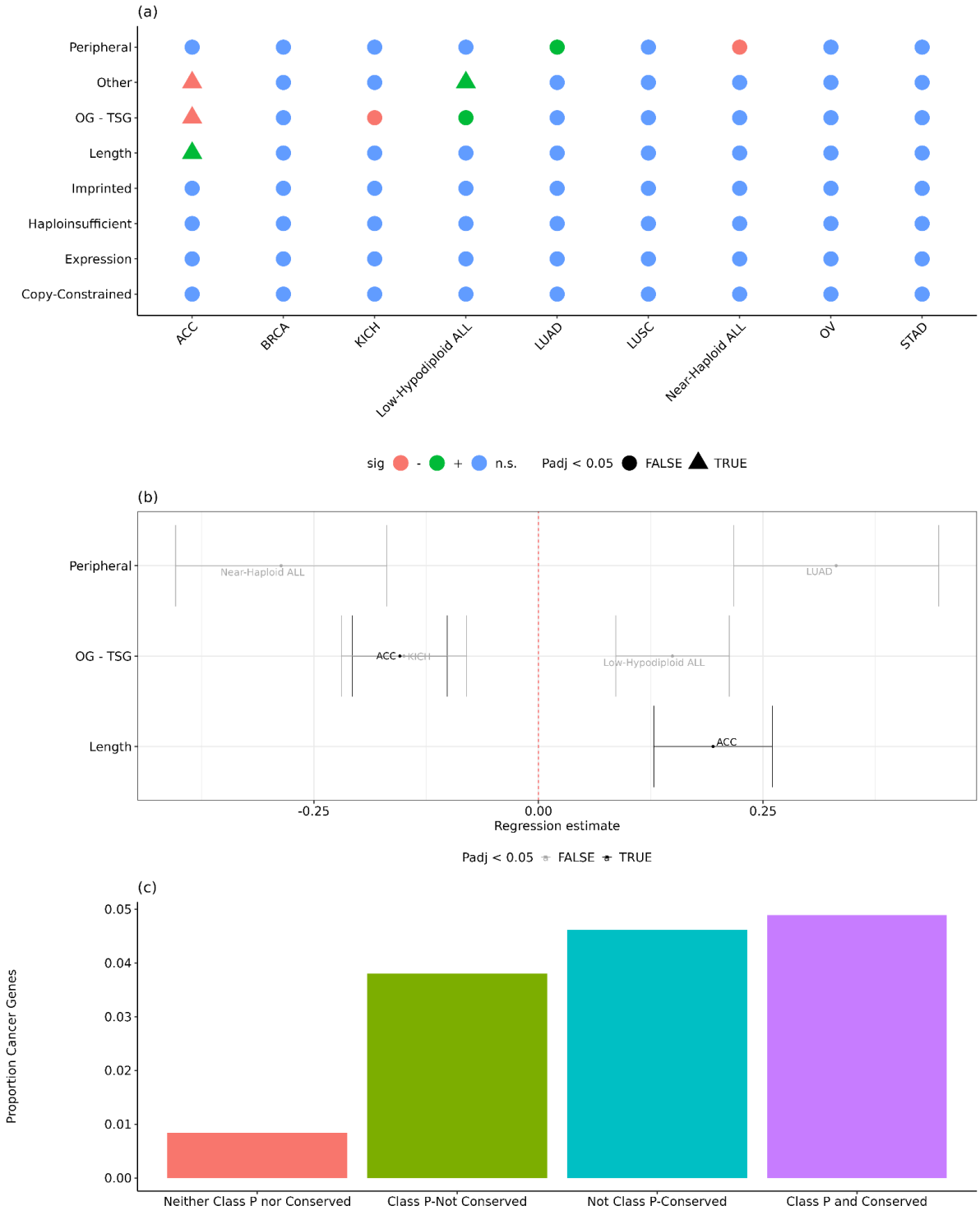


Fig. 3.19 Linear regression of chromosome arm loss rate on chromosome features (a) All results. Points are coloured by nominal significance and direction of association (green for increased loss rate, pink for decreased loss rate). Triangles indicate features significant after Benjamini-Hochberg correction **(b)** Regression estimates and standard errors for significantly-associated features, excluding 'other' nuclear location. Nominally-significant associations are in grey. Detailed visualisations of relationships between loss rate and chromosome arm length,

chromosome location, OG - TSG score and median chromosome arm expression are presented in Appendix Fig. 6.8 **(c)** Enrichment of cancer genes among germline copy-constrained genes. Class P indicates genes found in pathogenic CNVs in the human population; Conserved indicates genes with conserved copy number across thirteen mammalian species. Cancer genes were obtained from the COSMIC Cancer Gene Census (Tiers 1 and 2).

3.3.5 Prognostic Impact of Ploidy Alterations

It is clear that ploidy-aberrant tumours follow distinct evolutionary paths from near-diploids, with higher levels of chromosomal instability at both small (segmental CNAs) and large (WGD) scales. Do these ultimately alter the fitness of the tumour and the survival of the patient? Both WGD and – particularly – intra-tumour chromosomal heterogeneity have been linked to poor prognosis across cancer types (e.g. Bielski et al, 2018; van Dijk et al, 2021), and hypodiploidy is associated with much poorer prognosis in ALL, but some questions remain.

- Does hypodiploidy confer poor prognosis in non-ALL cancer types?
- Do masked hypodiploids differ in prognosis from non-doubled hypodiploids?
- Does the duration or stability of the altered ploidy state matter – i.e. if a hypodiploid tumour later undergoes WGD, does the timespan during which it was hypodiploid make a difference to its prognosis? For polyploids, does the rate at which duplicates are lost (and ploidy reduced) matter?

Considering all cancer types together, the most chromosomally unstable ploidy classes have the worst prognosis, with a gradient from near-diploid to

other to low-hypodiploid to polyploid (Fig. 3.20). However, when this is controlled for cancer type along with age, sex and race in a Cox proportional hazards regression, neither polyploids nor hypodiploids have significantly worse prognosis when compared to all ploidy classes. When compared specifically to near-diploids, both polyploidy and WGD status are significantly negatively associated with survival ($p = 0.02$). This association persists when TP53 mutation status and microsatellite instability status are included for covariates. However, there is no longer a significant survival difference when tumour stage and grade are included as covariates (suggesting that WGD does not alter prognosis within a stage). The difference also shrinks or disappears when the van Dijk et al (2021) CNH score and aneuploidy score are included as covariates, likely because one of the tumorigenic effects of WGD is to increase CNH and aneuploidy. Hypodiploidy is not significantly associated with prognosis across the TCGA dataset, and genome doubling does not affect this. Near-haploids appear to have very poor prognosis but their low sample size precludes further investigation.

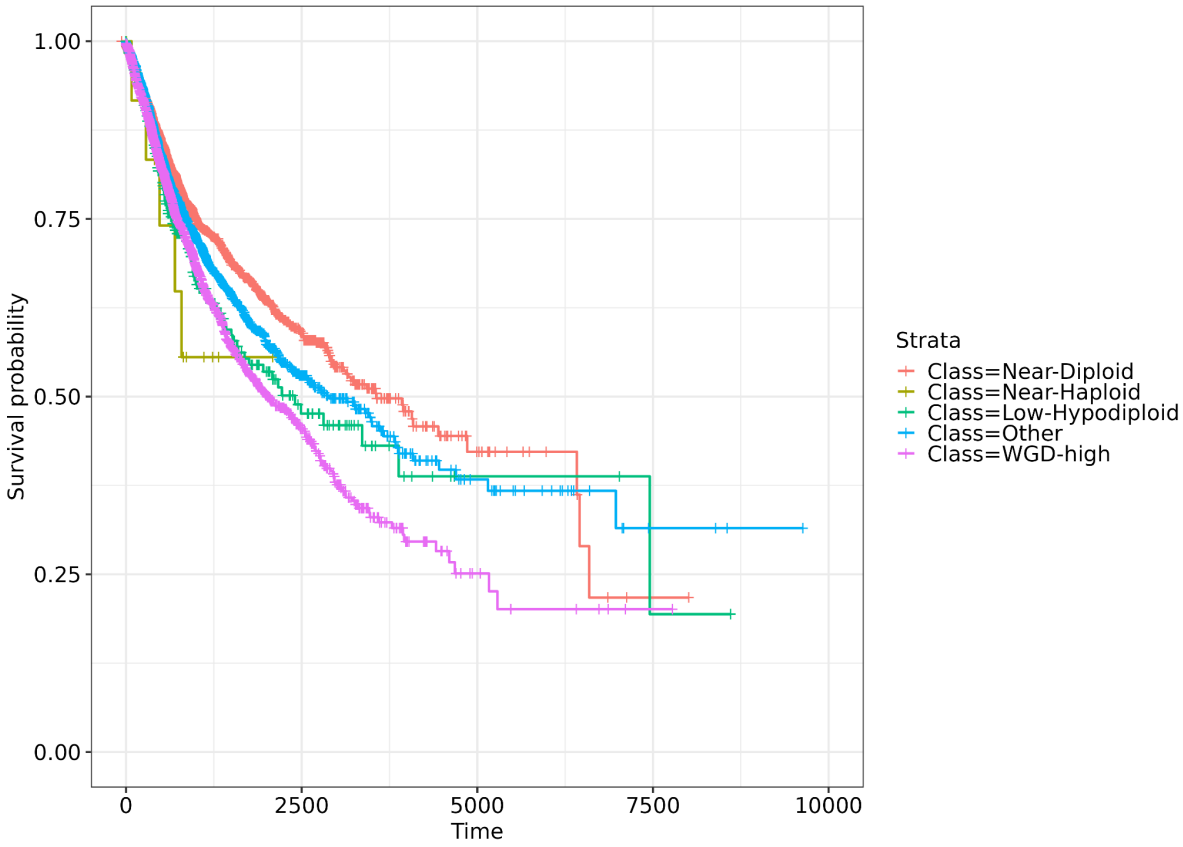


Fig. 3.20 Kaplan-Meier survival curves for all TCGA tumours by ploidy class. Time is in days since diagnosis.

Finally, I investigated the relationship between duplicate retention after WGD and survival and found that polyploid tumours that retain a higher proportion of genes in four copies have better overall survival when controlled for cancer type, sex, and age at diagnosis ($p = 0.04$). This holds (marginally) when I exclude cancer types with fewer than 15 polyploid cases ($p = 0.049$).

3.4 Discussion

3.4.1 Two Classes of Hypodiploid Tumour

Most TCGA cancer types do not show the clearly separated hypodiploid subgroups defined for ALL, and near-haploids are almost entirely absent. However, using those definitions as an arbitrary but useful threshold, it is apparent that low-hypodiploidy is widespread across solid tumours and not vanishingly rare, appearing in 4% of cases across 27/33 cancer types. Chromosome count distributions and results from the clustering analyses support the idea that there are particular cancer types with distinct hypodiploid subgroups (ALL, kidney chromophobe, possibly adrenocortical carcinoma) and others in which hypodiploidy seems to arise merely as a manifestation of extreme chromosomal instability. In line with this distinction, measurements of intra-tumour copy number heterogeneity and small-scale chromosomal instability within kidney chromophobe and (to a lesser extent) adrenocortical carcinoma reveal that they are notably less chromosomally unstable than expected compared to hypodiploids from other cancer types. ALL is an extreme case of this: its chromosome loss patterns are highly stereotyped, and even when hypodiploid ALL undergoes WGD, the doublings tend to be 'clean', i.e. create exact copies without additional aneuploidies. Near-haploidy is less likely to arise via successive chromosome losses from diploidy and does not connect to the continuous distribution of hypodiploid chromosome counts/spectrum of chromosomal instability seen in the TCGA tumours; if it has a different origin, as suggested by Molina et al

(2021), one might expect to see relatively stereotyped chromosome loss/retention profiles in non-ALL near-haploids too.

3.4.2 Patterns of chromosome loss

Chromosome loss patterns in hypodiploids are statistically non-random and bear significant similarities to loss patterns in other ploidy classes, including near-diploids, intermediate-ploidy cases, and polyploids. The correlation is weaker with post-WGD than pre-WGD losses in polyploids, suggesting that the WGD itself (and not just the initial chromosomal instability of polyploids-to-be) alters the loss landscape. This is borne out by direct comparison of pre- and post-WGD losses in polyploid tumours in this work and others (Baker et al, 2024; Boisselier et al, 2018). However, while hypodiploids do retain some of the loss patterns of their non-hypodiploid counterparts, I could not in general explain loss patterns with reference to chromosome features. In particular, I found no evidence that hypodiploid tumours evolve to avoid monosomy of germline dosage-constrained genes. However, it is interesting to note that the associations I did detect (with driver gene score, length and chromosome location) are mostly (7/8) in the tumours with characteristic hypodiploid loss patterns: ALL, ACC and KICH.

It must be acknowledged that my gene counting-based analysis of factors in chromosome loss rates is quite coarse-grained: the presence of a single strong tumour suppressor gene (e.g. TP53) on a chromosome may well outweigh the presence of many weak oncogenes. There are also much subtler effects that I

could not address here; it has been proposed that the consistently high relative copy number of chromosome 21 in ALL, for example, promotes oncogenic imbalance between specific isoforms of the RUNX1 gene (Gialesaki et al, 2023). Overall, however, the partial similarity with loss patterns in non-ploidy-aberrant tumours, together with the weakness of the evidence for specific selective effects, suggests a model wherein unstable hypodiploid tumours initially lose chromosome (arms) according to selective demands in the relevant tissue, and then—perhaps after losing a stability-maintaining tumour suppressor—lose further chromosomes rapidly and semi-randomly. Selection for or against each individual loss is likely to weaken as the overall burden of aneuploidy increases, in line with the Hill-Robertson explanation for the weakness of negative selection in cancer (Tilk et al, 2018).

3.4.3 A generalised chromosomal instability phenotype

I have shown that hypodiploidy is associated with WGD in both the Mitelman ALL and TCGA datasets, and that the rate of WGD increases with the degree of hypodiploidy. This extends to a unified chromosomal instability phenotype: besides the obvious (high levels of chromosome missegregation), hypodiploid TCGA tumours tend to be chromosomally unstable at both large and small scales, from WGD to intrachromosomal copy number alterations. Interestingly, monosomies themselves (unlike trisomies) do not induce a chromosomal instability phenotype (Hintzen et al, 2022), and instead this finding seems to point to an underlying predisposition towards chromosomal

instability in each tumour which can manifest as any or all of small-scale copy number alterations, chromosome missegregation events or ploidy changes. This factor also shapes the karyotypic evolution of polyploid tumours: in polyploids from cancer types that tend to be stable, a higher proportion of genes are retained in four copies after WGD. This implies an 'easy come, easy go' dynamic, where polyploids in rare-WGD cancer types tend to remain polyploid for longer. It also demonstrates that there are significant differences in chromosomal instability even within polyploids.

The stability of the ploidy-aberrant state has potential implications for tumour evolution: a sample that becomes hypodiploid and then rapidly undergoes WGD spends a shorter period of time sensitized to tumour suppressor gene mutations, for example, but is also less likely to sustain a deleterious mutation in an essential gene. I find that more stable polyploids have (marginally) better prognosis than unstable ones, which suggests that high ploidy itself does not necessarily increase tumour fitness. Instead, it reflects and increases the level of chromosomal instability and intra-tumour copy number heterogeneity in the tumour, which itself is strongly linked to poor prognosis (van Dijk et al, 2021).

While I would expect this factor to apply to hypodiploid tumours too, I did not find a significant relationship between hypodiploidy and prognosis, or between WGD and prognosis within hypodiploid tumours, although this may be due to low sample size among the TCGA tumours. Even if an analysis with

increased sample size did find an association between genome doubling and increased fitness/worsened prognosis of hypodiploid tumours, however, it would be difficult to untangle whether this was due to the proposed dosage-restoration effect or to an increase in chromosomal instability and intra-tumour heterogeneity. This could be partially, but not fully, mitigated by time-series subclone frequency data or experimentally-induced WGDs in hypodiploid cell lines. With the aid of my test to identify masked hypodiploid ALL cases, however, we can at least see that the doubled clone rarely fully outcompetes the hypodiploid, confirming that hypodiploid clones can survive alone or coexist with non-hypodiploid clones despite their aberrant ploidy and genome-wide dosage imbalances. This, along with the demonstrated ability of hypodiploid tumours across cancer types to tolerate widespread monosomy of dosage-sensitive genes, is an illustrative example of the extreme freedom from dosage constraint under which cancers evolve.

Chapter 4

Haploinsufficiency and dosage responsiveness in tumour suppressor genes

4.1 Introduction

What's in a copy number alteration? Some CNAs, like the recurrent HER2 amplification in breast cancer, drive tumorigenesis. Others are passengers arising from general chromosomal instability, but still have downstream effects: altering gene expression, activating stress pathways (Li & Zhu, 2022), sensitizing tumour suppressor regions to mutations, or creating oncogenic mutant allele imbalance (Bielski et al, 2018b). Still others—plausibly the

majority of CNAs in cancer—may have almost no effect on the cancer cell at all. This could arise because the genes involved are not dosage-sensitive (Rice & McLysaght, 2017), because they are not relevant in cancer cells in a given tissue, or because dosage changes are compensated or attenuated by the noisy process of gene expression and not reflected in protein abundance (Arad & Geiger, 2023; Kosti et al, 2016).

Under Knudson's seminal model of RB1 mutations and retinoblastoma inheritance (Knudson, 1971), tumour suppressor gene (TSG) losses are thought to be recessive at the cellular level, requiring two separate 'hits' to inactivate the tumour suppressor and promote tumorigenesis. This framework has been extremely influential in the development of cancer genetics: the idea that (most) TSGs are two-hit has historically been used to identify TSGs based on the presence of biallelic alterations (Ryland et al, 2015; Tomlinson et al, 2002), and provides a foundation for cancer evolution studies to this day (Wang et al, 2022; Zucker et al, 2024). One important implication of the model is that TSGs are dosage-insensitive: one intact copy is sufficient to maintain function.

However, in the fifty years since Knudson's seminal paper, evidence has mounted that many tumour suppressors are not 'classical' TSGs, i.e. their complete inactivation is unnecessary for tumorigenesis. TP53 is an important example: while it is frequently biallelically inactivated across cancer types, its monoallelic inactivation by either dominant-negative mutation (Willis et al,

2004) or single-copy loss is also tumorigenic, even when expression of the wild-type allele is maintained (Venkatachalam et al, 2001). Dozens of TSGs have now been reported to be haploinsufficient, with more recent examples including BRCA1 and BRCA2 (reviewed in Minnello & Carreira, 2024), WRN (Shih et al, 2023), SETD2 (Leung et al, 2022) and BARD1 (Randall et al, 2024), among others. Even the prototypical two-hit tumour suppressor RB1 is haploinsufficient for maintenance of genome stability (Coschi et al, 2014), particularly when TP53 is inactivated (Ishak & Dick, 2014). Furthermore, some TSGs are obligate haploinsufficient: partial inactivation is tumorigenic, but complete inactivation is more weakly tumorigenic or even tumour-suppressive or non-viable (Berger et al, 2011). Obligate haploinsufficient TSGs include PTEN, for which even a 20% decrease in expression is tumorigenic (Alimonti et al, 2010) and p120-catenin in intestinal neoplasia (Short et al, 2017), both of which show obligate haploinsufficiency conditional on the inactivation status of another TSG (TP53/APC).

It is not clear what proportion of TSGs are haploinsufficient, although it has been proposed based on mutational patterns that the vast majority of sporadic tumour suppressors are (Davoli et al, 2013). Very few TSGs have been positively confirmed to be haplo*sufficient*, with notable exceptions including LATS1/2 (Li et al, 2020) and POLD1 (Andrianova et al, 2024). Other large scale mutational analyses indicate that many TSGs can be inactivated by both one-hit and two-hit mechanisms (Besedina & Supek,

2024) and that two-hit genes can become one-hit through alterations elsewhere in the genome (Park et al, 2021).

Despite the growing mountain of reports on the haploinsufficiency of individual TSGs, there remain many TSGs for which the functional impact of partial inactivation is unknown. Jia & Zhao (2019) developed a method to classify TSG inactivation events and measured their *cis* and *trans* effects on the transcriptome, but they did not directly address haploinsufficiency or account for loss of heterozygosity (LOH), germline mutations or epigenetic silencing in their classification. Another screen, this one genome-wide, identified hundreds of 'STOP genes' whose knockdown resulted in increased proliferation in non-transformed mammary epithelial cells; these genes were enriched for TSGs, and recurrent deletions in cancers tend to encompass clusters of them (Solimini et al, 2012). Our work combines these two approaches in a computational screen to understand TSG haploinsufficiency across the TCGA dataset.

In this chapter, I classify TSG inactivation events and develop a new method to screen for and quantify TSG haploinsufficiency using cancer hallmarks as readouts of gene function. I identify novel candidate haploinsufficient genes among known TSGs and prioritise candidate TSGs among known haploinsufficient genes. I perform regulon enrichment analysis across the TCGA dataset and analyse attenuation of dosage alterations at the RNA, protein and functional output levels. Finally, I investigate whether

differential haploinsufficiency between TSGs can be explained by dosage attenuation.

4.2 Methods

4.2.1 Data Sources and Processing

I obtained copy number and mutation data with CLONETv2-based clonality information for 4,950 TCGA tumours from Ciani et al (2022), and ASCAT copy number data from the Genomic Data Commons. For the dosage compensation analyses, tumour and matched-normal CPTAC-3 proteomics data, upper quartile-normalised RNA-seq data and gene-level relative copy number profiles for OV, UCEC, HNSC, RCC and COAD (full names in Appendix Table 1) were downloaded from www.linkedomics.com. Proteomic and transcriptomic data for LUAD were obtained using the cptac python package.

4.2.2 Classifying Gene Inactivation Status

Genes were classified into the following five groups based on the presence of copy number alterations and the presence, deleteriousness and variant allele frequency (VAF) of mutations:

1) Zero intact copies — homozygous deletion (CN = 0), hemizygous deletion (CN = 1) with one or more deleterious mutations, or copy-

neutral loss of heterozygosity (CNLOH) with at least one deleterious mutation at variant allele frequency (VAF) ≥ 0.75. This VAF threshold was chosen based on Donehower et al (2019).

2) One intact copy — hemizygous deletion with no deleterious mutations, normal copy number (CN = 2) with one deleterious mutation, or CNLOH with one deleterious mutation at VAF < 0.75.
3) Two intact copies — normal copy number with no loss of heterozygosity or deleterious mutations
4) Miscellaneous copy number — CN > 2 or CNLOH with no deleterious mutations
5) Unclassified — CN > 1 with multiple deleterious variants (no phasing), CN > 2 with at least one deleterious mutation (because the VAF threshold is less useful at high copy numbers), or CNLOH with an indel (VAF unavailable).

Single nucleotide variants were defined as deleterious if they had a SIFT annotation of 'deleterious' and a PolyPhen annotation of 'probably damaging' or 'possibly damaging'. Indels were defined as deleterious if they were labelled as 'high impact'. A list of pathogenic germline variants in TCGA samples was obtained from Huang et al (2018). If a pathogenic germline mutation was present in a gene-patient combination, its classification was adjusted as follows:

1) Zero intact copies or unclassified → unchanged

2) One intact copy — if hemizygous, zero intact copies; if CN > 1, unclassified
3) Two intact copies → one intact copy
4) Miscellaneous CN → unclassified
5) Multiple germline mutations in a gene with CN ≥ 2 → unclassified

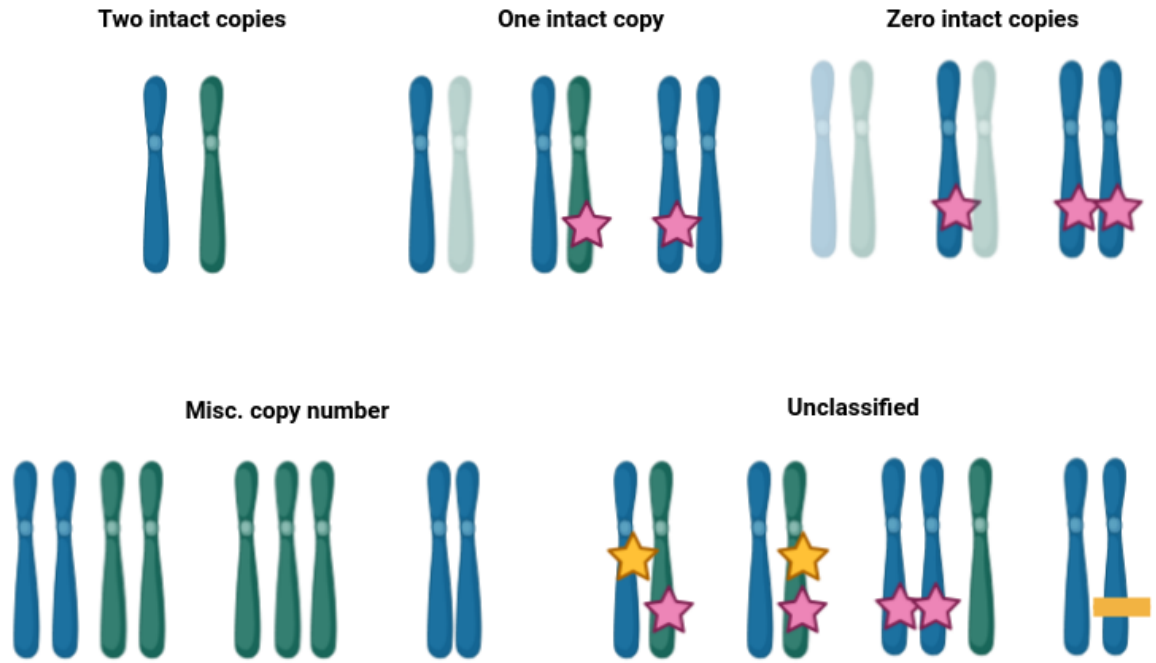


Fig. 4.1 Classification scheme for TSG status. Stars indicate mutations, the yellow bar designates an indel, chromosome colours indicate parent of origin and faded colours represent copy number losses (not necessarily of the entire chromosome). Copy number and RNA expression data were taken from the GDC TCGA data portal, while mutation calls and variant allele frequencies were obtained from Ciani et al (2022). Only the first three cases were considered in this analysis: two intact copies ('Intact'), one intact copy ('One' or partial loss) and zero intact copies ('Zero' or complete loss). This classification scheme was partially inspired by Jia and Zhao (2019), with some modifications including the addition of germline and RNA information and stringent handling of ambiguous cases.

I cross-checked the Ciani et al copy number calls against the ASCAT calls from the GDC portal, filtering out any instances with CN ≤ 2 where either the

major or minor copy numbers did not match. I also removed gene-tissue combinations with ≥ 25% missing data. Only genes in the first three categories (zero, one or two intact copies) were included in the following analyses. For the functional analyses, I removed instances where a gene was classified as having one or two intact copies in a sample but was not expressed (defined as raw read count < 10), with the aim of handling both (strong) epigenetic silencing and tissue-specific expression.

I finally integrated data on promoter methylation into the classifications to assess the robustness of my results to undetected hypermethylation. HM450k beta values were downloaded for all available TCGA samples and filtered to select probes located in the promoter region of a TSG or STOP gene, with the promoter region being defined as the transcription start site ± 500 bp. A gene's promoter was defined as hypermethylated if its mean promoter beta value was > 0.8 or if it had at least 3 promoter probes with beta > 0.8. Genes with hypermethylated promoters were defined as Unclassified and excluded from the methylation-aware analysis unless they were already classified as fully inactivated ('Zero').

4.2.3 Functional Readouts

I inferred the activity of a TSG using a set of readouts based on some of the hallmarks of cancer: proliferation rate, angiogenesis, mutation rate and copy number instability. Proliferation rate was computed from the expression (CPM) of the top 1% of genes correlated with the proliferating cell nuclear

antigen marker as described in Chapter 2. Angiogenesis was quantified by the expression (CPM) of the established angiogenesis markers VEGF-A and PECAM1/CD31 (Schluter et al, 2018). I obtained RNA counts and calculated CPM as described in Chapter 1. Mutation rate was measured as the total number of mutations per sample (including silent mutations) divided by sample ploidy. Copy number instability was measured using the intra-tumour copy number heterogeneity score computed by van Dijk et al (2021). As an ultimate readout, survival analysis was performed using Cox proportional hazards regression controlling for age at diagnosis and sex. Hallmark association tests were done using one-sided Wilcoxon tests and Benjamini-Holm-corrected for both the number of genes and the number of hallmarks analysed, with an FDR threshold of 0.05. I analysed only gene-cancer type combinations with at least 10 cases each with one intact copy and two intact copies.

I also compared hallmark levels between Intact and Zero TSG status. For TSG-cancer type-hallmark combinations with at least ten samples with each status, the extent of haploinsufficiency was quantified as a 'dosage-responsiveness ratio':

$$DR\ ratio = \frac{H_{One} - H_{Intact}}{H_{Zero} - H_{Intact}}$$

Eq. 4.1 Dosage-responsiveness ratio. H indicates the level of a given hallmark (proliferation rate, angiogenesis, CNH, mutation rate).

Linear regression was also performed as in Eq. 4.2. Dosage-responsiveness ratios were computed from this analysis as the regression estimate for partial inactivation divided by the regression estimate for complete inactivation.

```
lm(H ~ inactivation_status + gender + age)
```

Eq. 4.2 Linear regression of hallmark level on activation status. H indicates the level of a given hallmark; gender was excluded in sex-specific cancers. Inactivation status was coded as a factor with Intact as its base level.

4.2.4 Novel Haploinsufficient Tumour Suppressor Genes

I downloaded the COSMIC Cancer Gene Census dataset and selected genes annotated as either Tier 1 or Tier 2 TSGs. Genes that can act as either TSGs or oncogenes were also included (e.g. TP53, which is annotated as 'oncogene, TSG, fusion'). I computed tissue-specific dN/dS values for every TSG using the dndscv R package. A TSG was defined as *a priori* relevant in a given cancer type if its dndscv q-value for all mutation types combined was < 0.05 (evidence of positive selection, Martincorena et al 2017) or its complete inactivation (zero intact copies) rate was $\geq 1\%$.

Novel Tumour Suppressor Genes

I manually searched each TSG hit on Google Scholar with the term "haploinsufficient" to determine whether it had previously been reported as a haploinsufficient TSG, classifying each gene as definitely haploinsufficient, possibly haploinsufficient (for genes with weak or mixed evidence), conditionally haploinsufficient (e.g. depending on the status of other TSGs or

environmental conditions), no evidence of haploinsufficiency (genes for which I could not find a paper reporting them to be haploinsufficient), or not haploinsufficient (for genes specifically reported to be haplosufficient). I excluded TSGs that were only reported to be haploinsufficient in the germline for a non-cancer disease and papers that used unrelated definitions of 'haploinsufficiency'. I then re-ran the hallmark and survival analyses for all TSGs after excluding any copy number segments on which a known haploinsufficient TSG was hemizygously or homozygously deleted to identify genes whose apparent haploinsufficiency could not be explained by the proximity of a known haploinsufficient TSG. To compare the novel hits to other screens, I downloaded a list of genes shown to be essential-haploinsufficient from Sarel-Galilly et al (2022), filtered a list of STOP genes from Solimini et al (2012) (see below), and extracted one-hit vs two-hit cluster assignments from Park et al (2021).

STOP genes

A large-scale shRNA-based knockdown screen in immortalised mammary epithelial cells identified hundreds of STOP genes whose knockdown increases cell proliferation (Solimini et al, 2012). I defined genes with a $\log_2$ ratio ≥ 2 (i.e. doubled proliferation upon knockdown) for at least 4 shRNAs as STOP genes for my analysis, following one of the thresholds used by Solimini et al. Any genes that fell into both the STOP and TSG datasets were classified as TSGs. I searched for associations between single-copy inactivation of STOP genes and functional readouts in the same way as for

TSGs. Because in my design nearby genes will have similar associations due to co-deletions, I controlled for confounding by analysing only deletions that did not include a TSG that was significantly associated with the readout in question.

4.2.4 Regulon Enrichment Analysis

ARACNE regulons were obtained from the aracne.networks R package, DoRothEA regulons (classes A-D) from the DoRothEA R package, and TRRUST regulons from the TRRUST website (https://www.grnpedia.org/trrust/). The TRRUST regulons were filtered for transcriptional regulators with ≥5 targets with a known direction of regulation (activation/repression). I processed RNA expression matrices from the CPTAC projects and inferred activity scores for each transcriptional regulator by running VIPER with pleiotropy and scale set to TRUE, and ULM with center set to TRUE. Downstream analyses used VIPER to infer activity scores across the TCGA datasets based on the TRRUST regulons.

4.2.5 Dosage Attenuation

I computed DNA-RNA and RNA-protein correlations for each gene in each cancer type using the relative copy number and normalised RNA expression and protein abundance data from CPTAC. To create the cumulative attenuation visualisation, I multiplied the value in the previous stage (e.g. at the RNA level) by the absolute value of the correlation between stages (e.g.

the RNA-protein correlation) to get the degree of attenuation at the next stage (e.g. the protein level). To investigate factors associated with DNA-RNA, RNA-protein and protein abundance-activity score correlations, I performed multivariate linear regression separately for each expression stage and cancer type using the glm() function in R as follows:

```
glm(correlation ~ complex + TSG + OG + copyconserved +
  total_ptms + EHI + aggregated_reproducibility_rank)
```

Eq. 4.3 Linear regression of correlation (DNA-RNA, RNA-protein or protein abundance-functional activity) on gene features.

The covariates were (i) complex membership, based on human protein complexes downloaded from CORUM 5.0 (https://mips.helmholtz-muenchen.de/corum/download), (ii) TSG and oncogene status, based on COSMIC annotations ('role in cancer'), (iii) conservation of gene copy number in the 13 mammal species analysed by Rice & McLysaght (2017), (iv) essential haploinsufficient status in human embryonic stem cells (hESCs), based on Sarel-Galilly et al (2022)'s screen, (v) number of post-translational modifications (phosphorylation, acetylation and methylation events in PhosphoSitePlus, Hornbeck et al, 2014) and (vi) reproducibility of protein abundance measurements for each gene, as calculated by Upadhya & Ryan (2022).

4.3 Results

4.3.1 TSG Inactivation Events

As described in Methods, I classified TSG status into one of five categories based on the presence of copy number alterations and the variant allele frequency and deleteriousness of any somatic or germline mutations. In total, I counted 98,353 partial or complete TSG inactivations in 3,363 of 4,485 TCGA samples from 26 cancer types. The vast majority of partial TSG inactivation events involved hemizygous losses, which frequently include multiple genes at once; complete losses were relatively rare (1,697 instances) and were more evenly split between hemizygous deletion with a deleterious mutation (50.8%), homozygous deletion (37.7%) and copy-neutral loss of heterozygosity (CNLOH) with a deleterious variant at high VAF (11.5%). As expected, mechanisms of inactivation varied significantly by TSG. The TSGs with the highest rates of complete inactivation were CDKN2A, TP53, PTEN, RB1 and NF1 (Fig. 4.2). Complete inactivations of CDKN2A were always by homozygous deletion, while for TP53 they were usually by hemizygous deletion with a deleterious mutation on the remaining allele (62.5% of complete deletions) or CNLOH with a deleterious variant at high VAF (36.2%), with only 1.3% being due to homozygous deletion. PTEN, RB1 and NF1 were between these two extremes. TP53 was also the TSG most likely to be marked as unclassified, in 6.4% of all samples (vs 0.3% of all gene-sample combinations overall) due to a high frequency of indels, for which I did not have VAF information, in CNLOH regions.

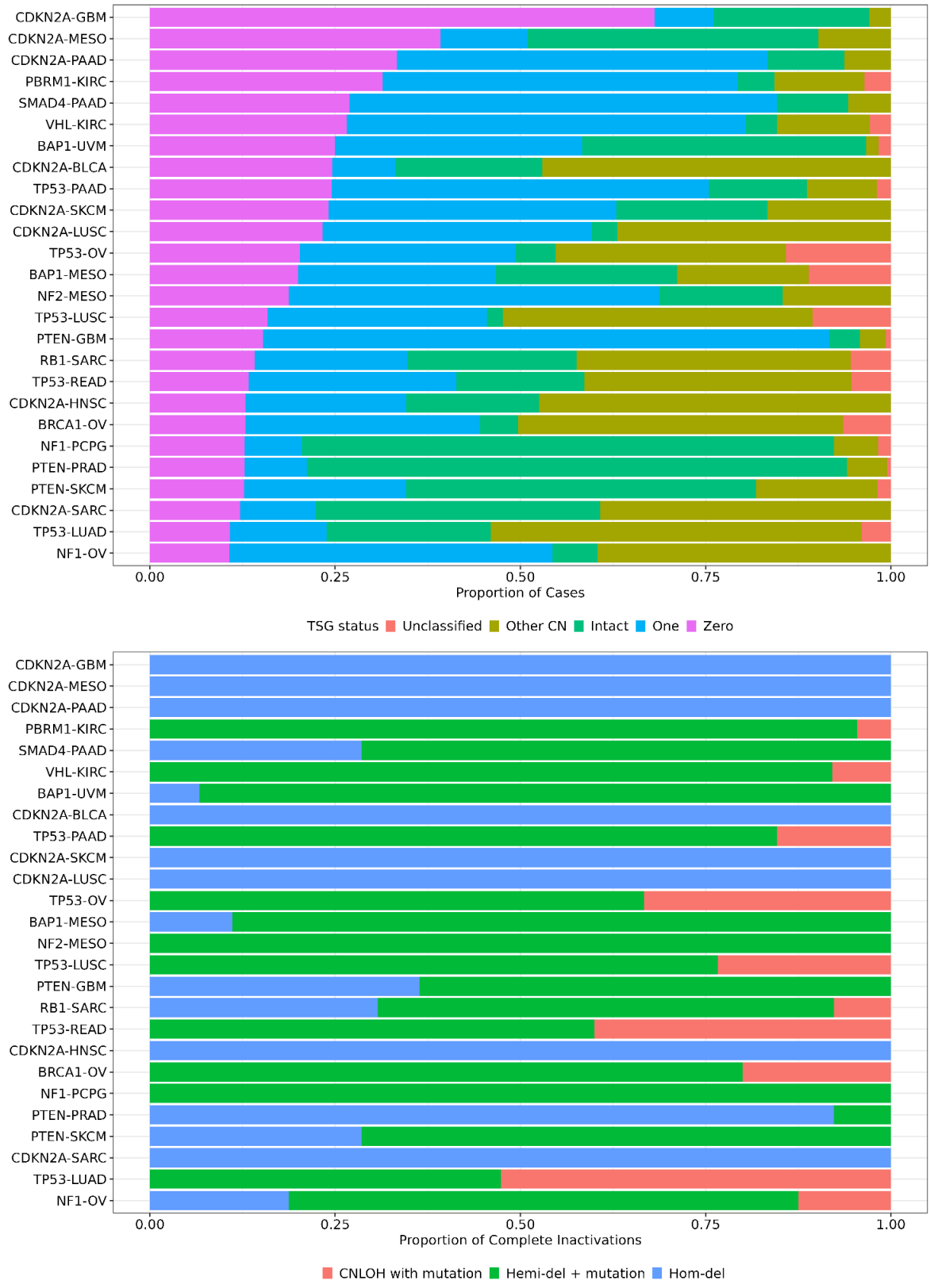


Fig. 4.2 TSG classifications (top) proportion of cases in a given cancer type in which TSG has each status **(bottom)** proportion of fully-inactivated cases lost by each mechanism. Hemi-del = hemizygous deletion, hom-del = homozygous deletion. Only TSG-cancer type combinations with a complete inactivation frequency ≥ 10% are shown.

4.3.2 Hallmarks of Cancer

Cancer cells are conceptually distinguished from normal cells by a set of hallmarks, including sustained proliferative signalling and evasion of growth suppression and cell death, induction of angiogenesis, and metastasis (Hanahan & Weinberg, 2000). This list was later expanded to include other traits, such as immune evasion and genomic instability and mutation (Hanahan & Weinberg, 2011). Because the inactivation of tumour suppressor genes (by definition) promotes oncogenic transformation, I propose to infer the activity of TSGs, and thus assess the effect of partial inactivation, by analysing the activity of cancer hallmarks in each sample. I aim to detect decreases in TSG activity upon partial inactivation as an increase in a relevant hallmark (e.g. mutation rate for BRCA1 in ovarian cancer) or a decrease in survival.

The four hallmarks I have chosen (proliferation, angiogenesis, mutation rate and copy number heterogeneity [CNH]) are largely independent of each other: the most correlated pair of hallmarks is CNH and proliferative index, as expected from previous analyses, at $r = 0.231$ (median of within-tissue correlations). The most negative median correlation is $r = -0.07$ for CNH and mutation rate, which is also somewhat unsurprising given the normalisation method. The strongest individual correlation is between CNH and proliferative index in KICH ($r = 0.813$), and the strongest negative one is between proliferative index and mutation rate in thymoma ($r = -0.59$). I also

analysed the relationship between each hallmark and survival using Cox proportional hazards regression controlling for sex and age at diagnosis (Fig. 4.3). Proliferative index is the most predictive hallmark, associated with worse prognosis in 9/31 cancer types; after multiple testing correction, these associations persist in adrenocortical, mesothelioma, liver, low-grade glioma, pancreatic and kidney papillary cancer, generally corresponding to the results of Ramaker et al (2017). Copy number heterogeneity was also associated with worse prognosis in several cancers, as reported by van Dijk et al (2021). Mutation rate is associated with better prognosis in bladder cancer after correction, which has also been previously reported (Voutsadakis, 2022). The level of angiogenesis was nominally significantly associated with worse prognosis in liver, cervical and colon cancer and glioblastoma, but these associations were not significant after multiple testing correction.

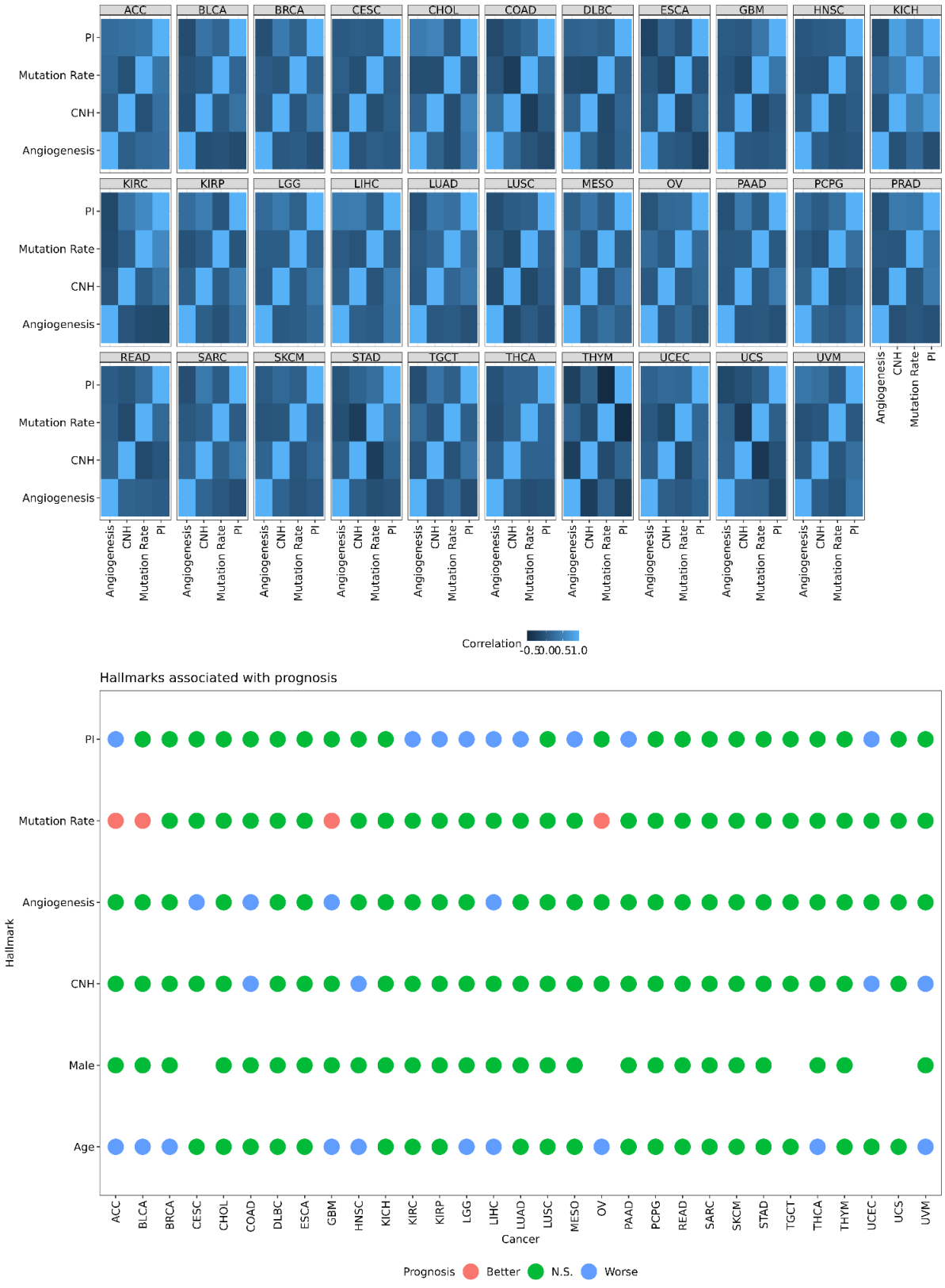


Fig. 4.3 Selected hallmarks of cancer: **(top)** correlation between value of hallmarks within each cancer type **(bottom)** association of each hallmark with overall survival. Points are coloured by nominal p-value. PI = proliferative index, CNH = copy number heterogeneity (van Dijk et al, 2021), N.S. = non-significant. Age

and sex are not hallmarks but were included in the Cox proportional hazards regression as covariates. Sex was excluded as a covariate for sex-specific cancers. These analyses were performed on the full dataset (for which I had data on all hallmarks), not restricted to cases with TSG mutation information. Prostate cancer was excluded because all patients for whom there was hallmark data survived.

4.3.3 One-Hit Tumour Suppressor Genes

Overall, I found 491 TSG-tissue combinations where a decrease from two to one intact copies was associated with a significant increase in a cancer hallmark after multiple testing correction, out of 6,920 total gene-tissue-hallmark combinations tested (7%). These were most commonly associated with proliferation (192), copy number heterogeneity (153), then mutation rate (109) and more rarely with angiogenesis (37). Because large deletions can partially inactivate many TSGs at once, there are islands of co-located genes sharing statistical signal (Fig. 4.4). This is particularly notable in the angiogenesis hallmark: of its 37 hits, 18 are on chromosome 3 in uveal melanoma and 16 are on chromosome 10 in low-grade glioma. This latter group includes the obligate haploinsufficient glial cancer driver PTEN, but the strongest hit is KLF6, which is a haploinsufficient tumour suppressor in glioblastoma (Masilamani et al, 2017) and inhibits angiogenesis when upregulated (Shi et al, 2024, Gallardo-Vara et al, 2016). In the other three hallmarks, hits are more spread out across the genome and mostly seen in breast cancer, possibly due to its large sample size. For mutation rate, the strongest hit was PD-L1 in low-grade glioma, and as expected both BRCA1 and BRCA2 emerged as haploinsufficient. For copy number heterogeneity, the strongest hits were B2M (an immune gene) and CASP3 (a caspase) in breast

cancer. TP53, which is known to be haploinsufficient, appears as a hit for CNH in bladder and stomach cancer, and for CNH, mutation rate and proliferative index in breast cancer. Most hits across both the hallmark and survival analyses were in breast cancer or low-grade glioma, which were the two largest cohorts after all filters were applied. There were 66 hits for survival, including a large group of 16 genes on chromosome 10 in low-grade glioma, again featuring PTEN and KLF6.

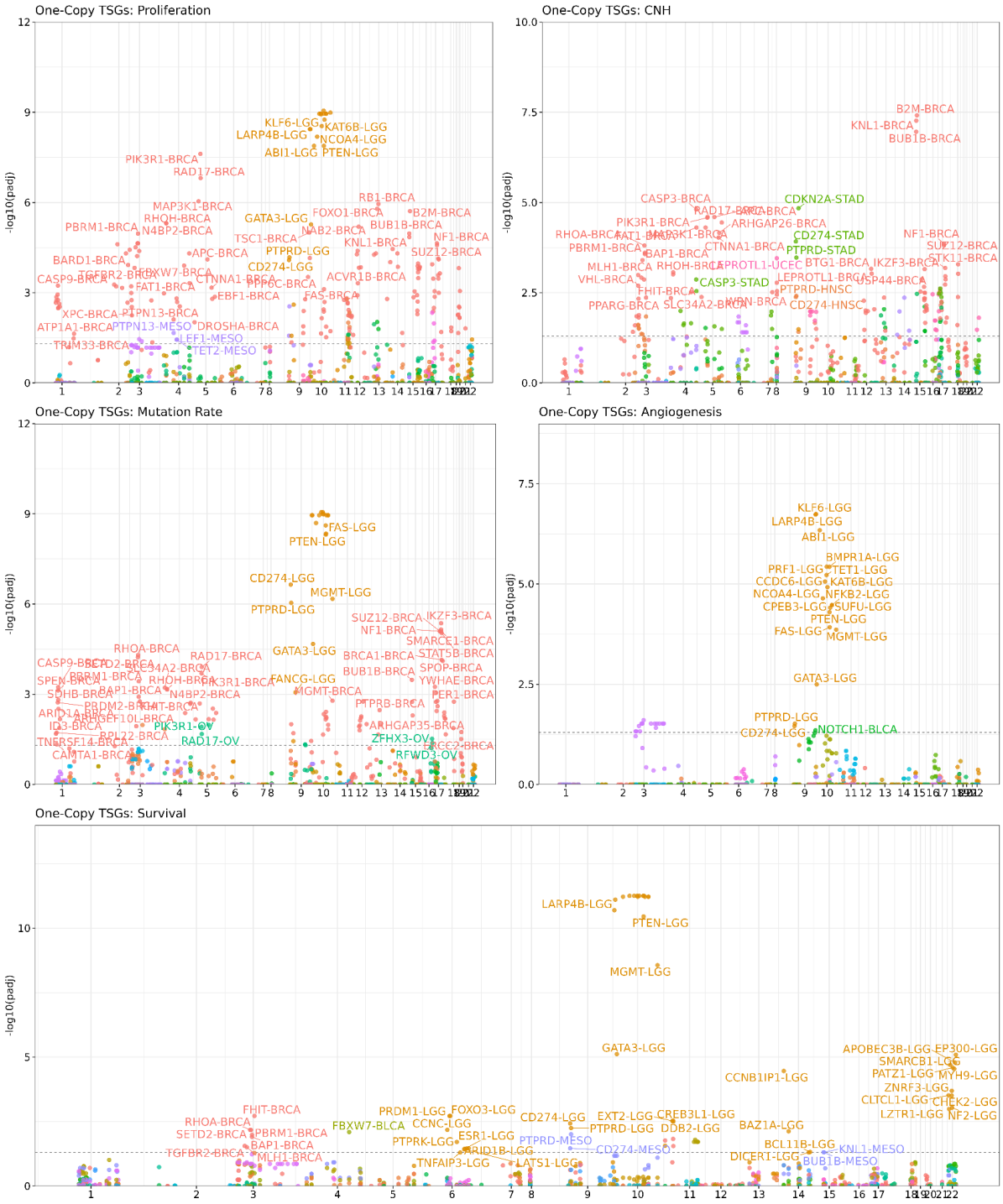


Fig. 4.4 TSGs significantly associated with increased hallmark activity or worsened prognosis when in one intact copy vs. two. One-sided Wilcoxon tests and Benjamini-Hochberg correction were performed for every combination of gene, cancer type and hallmark. Survival analysis was done using Cox proportional hazards regression controlling for age at diagnosis and sex. Colours separate cancer types.

The strongest hit in the survival analysis was for the bone morphogenetic protein receptor BMPR1A. BMPR1A, which was also significantly associated with increased proliferation and mutation rate in breast cancer and glioma, and increased angiogenesis in glioma, is involved in the TGFß pathway and can function as either a TSG or an oncogene (Bach et al, 2018). However, BMPR1A shares its hits with PTEN, which is 1.1 MB away on chromosome 10 and tends to be co-deleted.

To determine which of the candidate haploinsufficient TSGs identified by this screen were novel, I performed a manual literature search for each hit. This revealed that 69/209 genes, responsible for 32.8% of the hallmark hits and 42.6% of the survival hits, had previously been reported to be haploinsufficient or one-hit, including some with limited evidence (e.g. ARHGEF10, Xue et al 2012) or that show haploinsufficiency in cooperation with other genes (e.g. BTG1 with IKZF1 in leukaemia, Scheijen et al 2017). To address co-deletions, I re-ran the analyses after removing all known haploinsufficient TSGs and all copy number segments on which those TSGs were deleted from the dataset (for tissue-hallmark combinations where that TSG was a significant hit). This is somewhat conservative, because it removes genuinely haploinsufficient TSGs that happen to be located near established ones, but it allowed me to identify 18 novel hits that could not be explained by the presence or proximity of a known haploinsufficient TSG (Table 4.1).

gene	readout	cancer	chromosome	start	p.adj	STOP	EHI	hits
N4BP2	CNH	BRCA	4	40.1	0.01028	No	No	
RHOH	CNH	BRCA	4	40.2	0.01028	No	No	
LEPROTL1	CNH	UCEC	8	30.1	0.01028	No	No	
B2M	CNH	LUAD	15	44.7	0.046286	No	No	
STK11	CNH	BRCA	19	1.2	0.01028	No	No	C2
TCF3	CNH	BRCA	19	1.6	0.01028	No	No	
SLC34A2	Mutation Rate	BRCA	4	25.6	0.01028	No	No	
N4BP2	Mutation Rate	BRCA	4	40.1	0.01028	No	No	
RHOH	Mutation Rate	BRCA	4	40.2	0.01028	No	No	
ACVR2A	Proliferation	BRCA	2	147.8	0.01028	Yes	No	C1
SLC34A2	Proliferation	BRCA	4	25.6	0.010534	No	No	
N4BP2	Proliferation	BRCA	4	40.1	0.00011	No	No	
RHOH	Proliferation	BRCA	4	40.2	0.00011	No	No	
FANCG	Proliferation	LGG	9	35.1	0.01028	No	No	
CCNB1IP1	Proliferation	BRCA	14	20.3	0.008884	No	No	
BAZ1A	Proliferation	BRCA	14	34.8	0.01028	Yes	No	
SIRPA	Proliferation	BRCA	20	1.9	0.01028	No	No	
CCNB1IP1	Survival	LGG	14	20.3	0.00379	No	No	

Table 4.1 Novel one-hit TSGs identified by this screen. The screen was performed as described in Fig. 4.4, but the dataset was pre-filtered to remove any copy number segments on which an established one-hit TSG that was significant in the relevant tissue and readout had a deletion. The last three columns indicate the results of other haploinsufficiency-related screens for these genes (non-tissue-specific): the Solimini et al. (2012) screen for 'STOP genes' which increase cell proliferation when knocked down, the Sarel-Galilly et al (2022) screen of haploinsufficient essential genes in human embryonic stem cells, and the Park et al (2021) screen of TSGs with statistically significant co-occurrence between mutations and copy number alterations. For this last screen, 'C1' indicates a gene that was part of the 'one-hit' cluster and 'C2' a gene that was in the 'two-hit' cluster; NA demarcates a gene that was not included in their clustering analysis. CNH = copy number heterogeneity, chr = chromosome, BRCA = breast cancer, UCEC = endometrial, LGG = low-grade glioma, LUAD = lung adenocarcinoma.

Four novel candidate one-hit TSGs appear multiple times: N4BP2 and RHOH for proliferation, mutation rate and copy number heterogeneity in breast cancer, SLC34A2 for proliferation and mutation rate in breast cancer, and CCNB1IP1 for survival in low-grade glioma and proliferation in breast cancer. Most of the candidates are established tumour suppressors, but four are putative (COSMIC Tier 2): N4BP2, LEPROTL1, BAZ1A and SIRPA. While none have to my knowledge been previously identified as haploinsufficient or one-hit TSGs, some are haploinsufficient in the germline for other functions, e.g. germline monoallelic TCF3 mutation causes immunodeficiency via both dominant-negative and haploinsufficiency mechanisms (Boast et al, 2023). These genes are almost exclusively (partially) inactivated by hemizygous deletions, with hemizygous deletion rate in the relevant tissue ranging from 3.74% (SIRPA) to 17.4% (LEPROTL1). Due to their low mutation rate, they would not generally be picked up as a relevant driver by mutation-based methods such as dndscv ($q > 0.05$ for all novel candidates except B2M in LUAD). They also do not undergo homozygous deletions, raising the intriguing possibility that they are obligate haploinsufficient. However, they have not been identified as essential in cancer based on CRISPR-Cas9 dependency screens (Pacini et al, 2021).

Comparing to the results of other haploinsufficiency-related screens, two of the novel candidates (ACVR2A and BAZ1A), both significantly associated with breast cancer proliferation in my screen, increase the proliferation of mammary epithelial cells when knocked down (Solimini et al, 2012).

Interestingly, ACVR2A was also classified into the 'one-hit TSG' cluster by Park et al (2021) due to its lack of significant co-occurrence between mutations and copy number losses. STK11, on the other hand, which I found to be associated with increased copy number heterogeneity in breast cancer when partially inactivated, did have a significant co-occurrence between mutations and copy number losses in that analysis.

Returning to the full set of hits, I searched for TSGs whose partial inactivation was associated with both overall survival and a cancer hallmark, in the interest of finding potential routes through which these TSGs might influence tumour fitness and ultimately patient prognosis. Of the 66 gene-cancer types combinations significantly associated with worsened overall survival, 28 (42%) were also a significant hit for at least one hallmark. This proportion rises to 180/225 (80%) when I consider genes with nominally significant associations for both survival and a hallmark. Genes associated with both survival and any hallmark are almost always associated with multiple hallmarks, particularly proliferation and mutation rate (Fig. 4.5). This suggests that partial inactivation of these genes contributes to a general aggressiveness phenotype with increases in proliferation, mutation and angiogenesis.

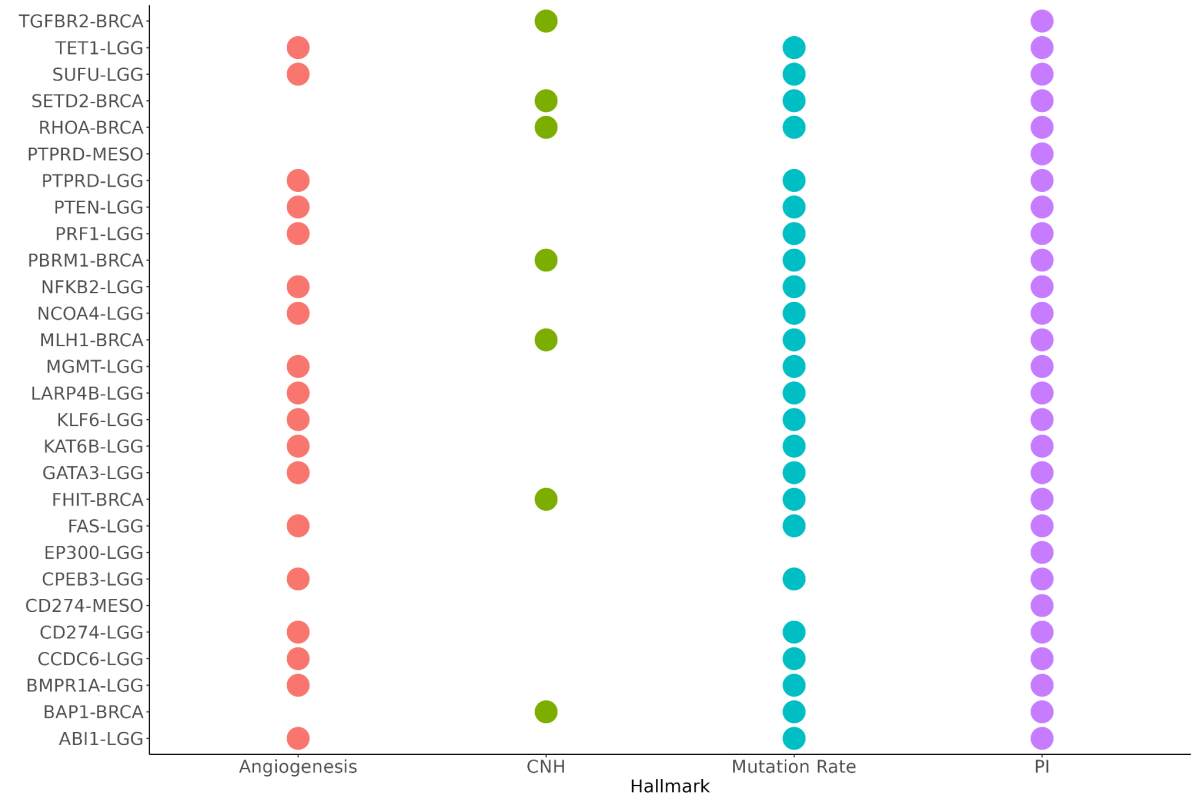


Fig. 4.5 TSGs whose partial inactivation is associated with both survival and at least one hallmark. CNH = copy number heterogeneity, PI = proliferative index.

4.3.4 Quantifying Dosage Responsiveness

While the one-hit TSGs identified in this screen are all associated with an increase in a hallmark or a decrease in survival upon partial inactivation, the magnitude of the change varies. I calculated a 'dosage-responsiveness ratio' (DR ratio) as follows:

$$DR\ ratio = \frac{hallmark_{one} - hallmark_{intact}}{hallmark_{zero} - hallmark_{intact}}$$

for the 16 gene-cancer type (34 gene-cancer type-hallmark) combinations that were significant hits in the one-hit TSG screen and had at least 5 cases where the TSG was fully inactivated (zero intact copies). For 21/34 combinations, the DR ratio was between 0 and 1, indicating that both partial and complete inactivation were associated with an increase in hallmark level, with complete inactivation having a stronger effect. These ranged from 0.21 for CDKN2A and proliferation in low-grade glioma, implying that partial inactivation was associated with only 20% as much of an increase in proliferation as complete inactivation, to 0.899 for PTEN and mutation rate in breast cancer, indicating that partial and complete inactivation were almost equivalent. Seven of these combinations, including BRCA1 for CNH, TP53 for mutation rate and RB1 for multiple phenotypes in breast and bladder cancer, had DR ratios close to 0.5, suggesting additivity. Ten combinations had a DR ratio > 1, suggesting that partial inactivations were associated with stronger increases than complete inactivations, but none of these differences were statistically significant, so this might indicate complete haploinsufficiency but does not provide evidence of obligate haploinsufficiency. Fully obligate haploinsufficient TSGs are unlikely to be amenable to this analysis because they rarely undergo complete inactivation; I considered an approach to detect them based on depletion of complete inactivations relative to the expected rate based on partial inactivations, but this is complicated by selection against homozygous deletions that encompass multiple genes. The remaining three combinations, with negative values of the DR ratio indicating a decrease in hallmark level with complete

inactivation, involved NCOR1, for which partial inactivations were associated with increased proliferation, mutation and copy number heterogeneity and complete inactivations were not significantly different from intact cases. When I restrict to combinations with ≥10 complete inactivations, 8/9 have DR ratios between 0 and 1 (Fig. 4.6).

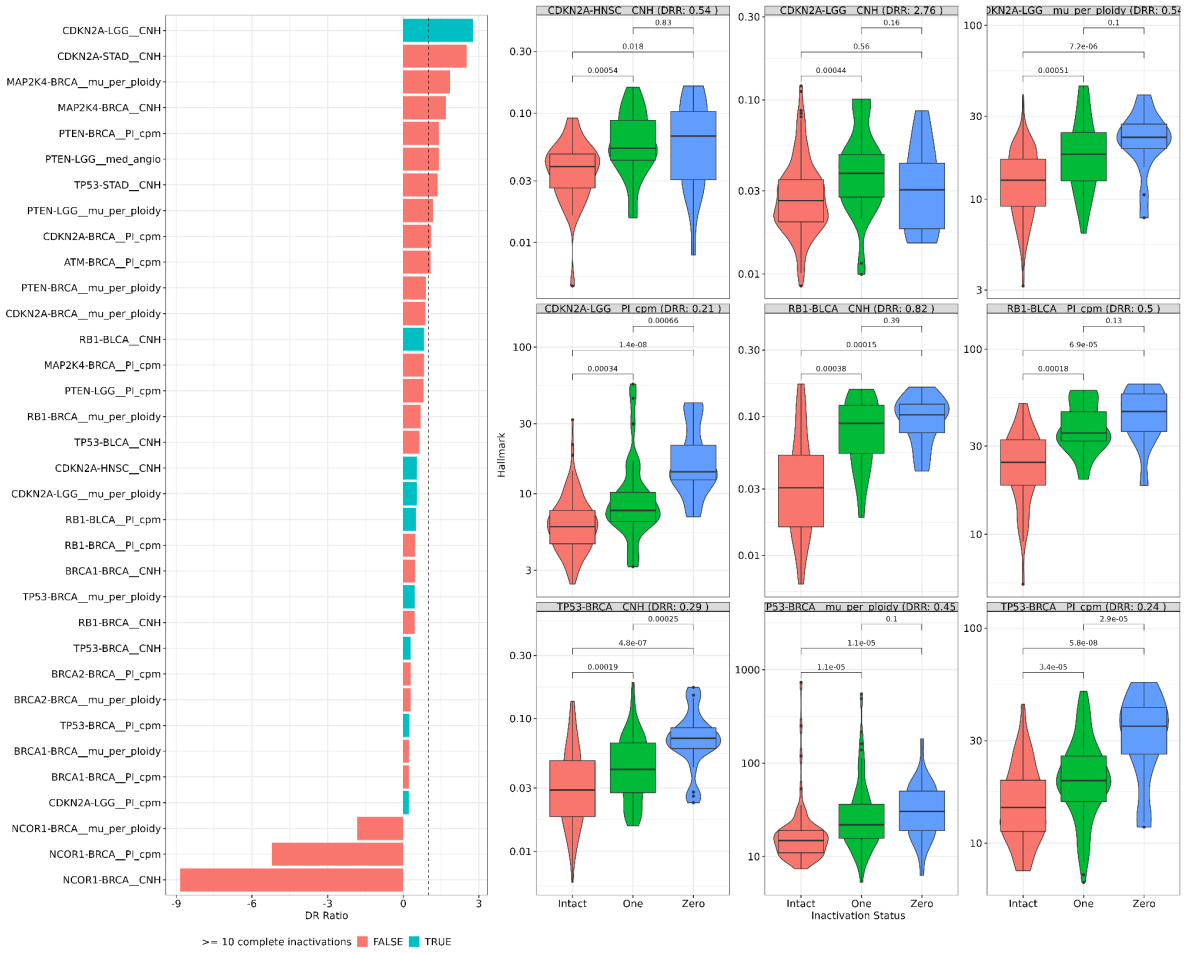


Fig. 4.6 Dosage-responsiveness ratios for significant one-hit TSGs. (left) All DR ratios for hits with ≥5 complete inactivations. Dashed line at 1 indicates the boundary between complete vs partial inactivation being associated with a stronger difference from Intact status. Colours indicate whether the combination had ≥10 complete inactivations. CNH = copy number heterogeneity, mu_per_ploidy = mutation rate, PI_cpm = proliferation rate, med_angio = angiogenesis. **(right)** Hallmark levels and DR ratios for the 9 significant hits with ≥10 complete inactivations. The y-axis is $\log_{10}$-transformed; p-values shown are from two-sided Wilcoxon tests.

The vast majority of gene-cancer type combinations do not have enough complete inactivation events to compute DR ratios, but it is still possible to quantify their dosage-responsiveness by computing the ratio of median hallmark level for TSGs in one intact copy to that of TSGs in two intact copies (Fig. 4.7). Significant hits in this analysis tend to have ratios around 1.81, suggesting that they are associated with an 81% increase in the level of the hallmark. This is highest for CNH at 1.92, followed by angiogenesis at 1.83, mutation rate at 1.76 and PI at 1.69. Non-hits, many of which may not be relevant in a given cancer type or for a given hallmark, tend to have ratios close to 1, as expected. However, there is some inflation (median ratios 1.03, 1.09, 1.12 and 1.14 for angiogenesis, proliferation, mutation rate and CNH, respectively) and notable variation around the median, which could indicate a group of weaker haploinsufficient TSGs that were not picked up by this analysis. The smallest ratio for which I could detect a significant difference after multiple testing correction was for CBL and proliferation in lung squamous carcinoma, at 1.21 (a 21% increase); in other hallmarks, the minimum detectable ratios were 1.36, 1.39 and 1.41 for angiogenesis, copy number heterogeneity and mutation rate respectively.

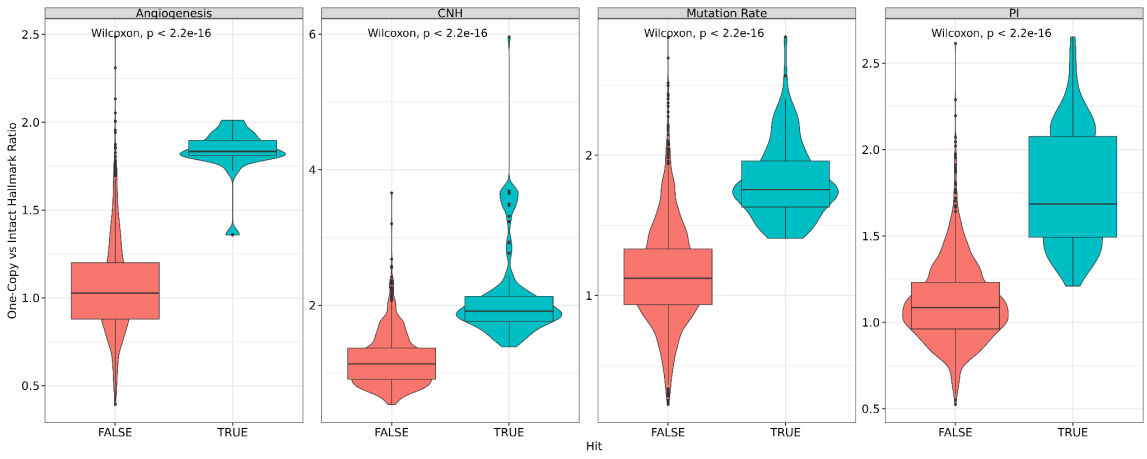


Fig. 4.7 Ratio of median hallmark level when TSG is in one intact copy to median hallmark level when TSG is in two intact copies. Data is presented for all 6,920 gene-cancer type-hallmark combinations tested, i.e. all have at least ten cases with each inactivation status. The y-axis is independent between plots. The x axis and colours indicate hit status, showing the distributions of ratios for non-hits (left) and hits (right).

What proportion of TSGs are haploinsufficient or otherwise one-hit? Of the fourteen combinations where a complete inactivation increases hallmark levels significantly, seven are also significant in the same tissue and hallmark when partially inactivated. The others (BAP1, TP53 and CDKN2A) are significant hits in at least one other tissue or hallmark, and have all been previously suggested to be haploinsufficient (e.g. Perkail et al, 2020). CDKN2A's effect on angiogenesis in low-grade glioma is perhaps the strongest candidate for haplo*sufficiency* in this dataset: partial inactivation is not significantly different from intact CDKN2A even at nominal $p < 0.05$, the DR ratio is fairly low at 0.17, and CDKN2A has very high rates of homozygous deletion. However, even CDKN2A in low-grade glioma does appear to be haploinsufficient for other hallmarks (CNH, proliferation, mutation rate), with a low DR ratio for proliferation (0.21) but moderate to high ones for mutation rate (0.53) and CNH (2.74). Additionally, patients with each of the three CDKN2A inactivation statuses have distinct prognoses (Fig. 4.8), although this is only nominally significant in the Cox proportional hazards regression controlling for age and sex ($p = 0.015$). I tested whether this association could be confounded by copy number instability (both increasing the rate of CDKN2A deletion and worsening prognosis), but CNH is not a significant predictor of survival in LGG and partial inactivation of

CDKN2A retains its (nominally) significant association with survival when CNH is included as a covariate. CDKN2A has a known role in preventing angiogenesis (Baruah et al, 2015), but also other functions such as promoting senescence (Zhao et al, 2016). It is possible that CDKN2A is haplosufficient for one role (suppressing angiogenesis) but haploinsufficient for others.

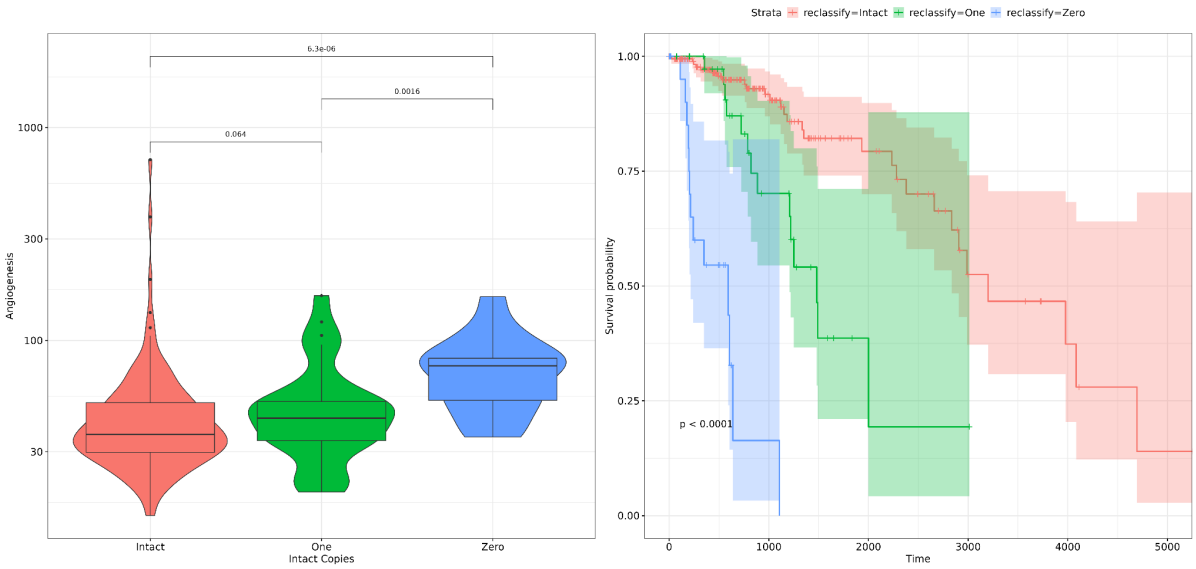


Fig. 4.8 Mixed evidence for CDKN2A as a haplosufficient tumour suppressor. (left) angiogenesis in low-grade glioma by CDKN2A inactivation status, y axis $\log_{10}$-transformed **(right)** patient prognosis (overall survival) by CDKN2A inactivation status. P-value shown is not adjusted for covariates.

4.3.5 Candidate Tumour Suppressor Genes

Finally, I investigated whether this screen could provide or strengthen the evidence that some known haploinsufficient proliferation-related genes are tumour suppressor genes. I performed the same analysis as above on 692 STOP genes (genes whose knockdown promotes proliferation in mammary epithelial cells) identified by Solimini et al (2012) that were not listed as either Tier 1 or Tier 2 TSGs in the COSMIC database. To account for confounding from nearby co-deleted TSGs, I removed all segments on which a

significant hit from the TSG analysis was deleted, which reduced the number of STOP hits from 1,010 (361 for proliferation, 282 for CNH, 188 for mutation rate, 57 for angiogenesis, 122 for survival) to 19 (all survival). The 19 STOP genes significant for survival were grouped into seven colocated clusters: RAD18, ARF4, FLNB and C3orf14 on chromosome 3 in head and neck; ERI1, SOX7, TUSC3 and MTUS1 on chromosome 8 in head and neck; HES1 on chromosome 3, SSR1, SNRNP48, MYLIP, FAM8A1 and NRSN1 on chromosome 6, and TEK and SMU1 on chromosome 9 in low-grade glioma; and SMU1 on chromosome 9 and NUDT5 and CDC123 on chromosome 10 in mesothelioma (Fig. 4.9). None of the STOP genes had a significant association with a hallmark after Benjamini-Hochberg correction, although 13/13 of the significant survival hits that were included in the hallmark analysis had nominal $p < 0.05$ for at least one hallmark (as opposed to 49.96% of all gene-cancer type combinations in the STOP hallmark analysis). In general, the STOP hits have very low rates of complete inactivation (none except in TUSC3-HNSC and TEK-LGG, with 2 and 5 respectively vs 73 and 48 partial inactivations), and they are always partially inactivated by hemizygous deletion rather than mutation. Four of them, representing 5/19 hits, are essential genes in cancer: ARF4, SNRNP48, SMU1 and CDC123.


One-Copy STOP genes: Survival
-log10(padj)
HES1-LGG
SNRNP48-LGG
SSR1-LGG
NRSN1-LGG
MYLIP-LGG
FAM8A1-LGG
ARF4-HNSC
SMU1-MESO
ERI1-HNSC
SMU1-LGG
RAD18-HNSC
C3orf14-HNSC
FLNB-HNSC
SOX7-HNSC
TEK-LGG
CDC123-MESO
TUSC3-HNSC
MTUS1-HNSC
NUDT5-MESO
RAD18-HNSC
ARF4-HNSC
FLNB-HNSC
C3orf14-HNSC
ERI1-HNSC
p = 0.0026
p = 0.0046
p = 0.0046
p = 0.0027
p = 0.00013
SOX7-HNSC
TUSC3-HNSC
MTUS1-HNSC
HES1-LGG
SSR1-LGG
p = 0.00021
p = 0.00076
p = 0.00022
p < 0.0001
p < 0.0001
SNRNP48-LGG
MYLIP-LGG
FAM8A1-LGG
NRSN1-LGG
TEK-LGG
p < 0.0001
p < 0.0001
p < 0.0001
p < 0.0001
p < 0.0001
SMU1-LGG
SMU1-MESO
NUDT5-MESO
CDC123-MESO
p < 0.0001
p = 0.00048
p = 0.0012
p = 0.0012
Survival probability
Time
Number at risk
Strata

Fig. 4.9 STOP genes associated with worse prognosis when in one intact copy vs two, displayed via Manhattan plot **(top)** and Kaplan-Meier plots **(bottom)**. A list of STOP genes was obtained from Solimini et al (2012). I ran Cox proportional hazards regression controlling for age and sex after removing all segments on which a significant TSG was hemizygously or homozygously deleted. Only genes with at least 10 partial inactivations in the relevant cancer type were included. TUSC3-HNSC and TEK-LGG had some complete inactivations (2 and 5, respectively). Each Kaplan-Meier plot includes a table of the number of patients with each STOP gene inactivation status at different timepoints.

4.3.6 Regulon Enrichment Analysis

Thus far, I have used cancer hallmarks and patient prognosis data as indicators of TSG function. This approach allowed me to identify candidate haploinsufficient TSGs, but is biased towards strong TSGs with major impacts on cell behaviour and patient survival. Some weaker TSGs may be strongly dosage-responsive in that a decrease in (intact) copy number results in a significant decrease in their functioning, but if they have only a minor impact on cell behaviour then this may not appear as a significant difference in the hallmark or survival analyses. Conversely, even a minor decrease in the functioning of a strong TSG could alter cell behaviour. For this reason, it would be beneficial to have a more proximal measure of TSG functioning.

Regulon enrichment analysis is used to estimate the activity of a protein (a transcriptional regulator, specifically) within a sample by measuring the expression of the genes under its control, i.e. its regulon. Different methods exist both for identifying a gene's regulon and for estimating its activity based on the regulon's expression. I compared the results of two inference methods (VIPER and ULM) and three regulon databases (ARACNE, DoRothEA,

TRRUST) on the CPTAC project cancers before computing TSG activity scores across the TCGA database. Briefly, I found that the two inference methods gave very similar results (r between 0.91 and 0.97, Appendix Fig. 6.11), but the same is not true for different regulons: scores inferred by VIPER using ARACNE regulons are only correlated at r ~ 0.2 with scores based on DoRothEA, and r ~ 0.1 with scores based on TRRUST (Fig. 4.10). Scores based on TRRUST and DoRothEA are more similar to each other, although the correlation is still not strong (r ~ 0.4).

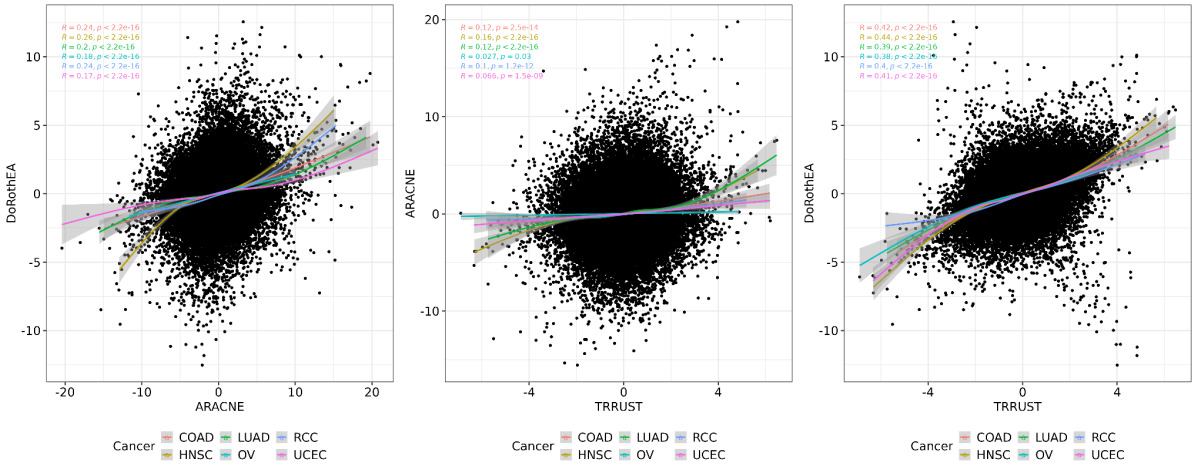


Fig. 4.10 Correlations between gene activity (output) scores inferred by VIPER using different regulons: ARACNE, DoRothEA and TRRUST. Each plot used the same set of genes analysed with all regulons. DoRothEA classes A-D were included.

The three regulon datasets were constructed using different approaches. ARACNE regulons are based on RNA-RNA relationships and are available in tissue-specific versions for each TCGA cancer type. The TRRUST database comes from manual curation of regulatory relationships mined from Pubmed articles, while DoRothEA combines both ARACNE-style inference from gene

expression and literature-derived networks with TF binding data derived from ChIP-seq and promoter sequence analyses.

To investigate the differences between regulons further, I compared their predicted activity scores to (a) protein abundance and (b) RNA expression for each gene (Fig. 4.11). ARACNE-based score predictions had notably stronger correlations with protein abundance (median r = 0.19) than scores based on DoRothEA (0.0598) or TRRUST (0.0327). Surprisingly, for ARACNE, RNA expression was much more predictive of activity score than protein abundance was (median r = 0.521). I expect that activity score should be more closely related to protein abundance than RNA expression because these are protein-coding genes, so their protein stage is more proximal to their ultimate function. With DoRothEA and TRRUST, the RNA-activity correlation also exceeded the protein-activity correlation, but to a much lesser extent (median RNA-activity score correlation 0.132 and 0.0569, respectively). While the small differences might be due to difficulty in accurately measuring protein abundance, I hypothesise that the large difference in ARACNE could be due to RNA-RNA correlations that do not reflect direct regulatory relationships. This idea is supported by the fact that in ARACNE, genes with strong protein abundance-activity correlations tend to have strong RNA-protein correlations (r = 0.58), an effect that is much weaker in DoRothEA (r = 0.19) and weaker still in TRRUST (r = 0.12, p = n.s. within cancer types). This pattern persists when I restrict to the 85 genes included in all three regulons (r = 0.48 for ARACNE and 0.2 for DoRothEA), although the correlation increases somewhat to r = 0.19 for TRRUST. This suggests that the set of

transcriptional regulators included in each regulon has an effect, but does not explain the pattern. Because DoRothEA would also be somewhat vulnerable to this problem, I decided to use the TRRUST regulons for these analyses.

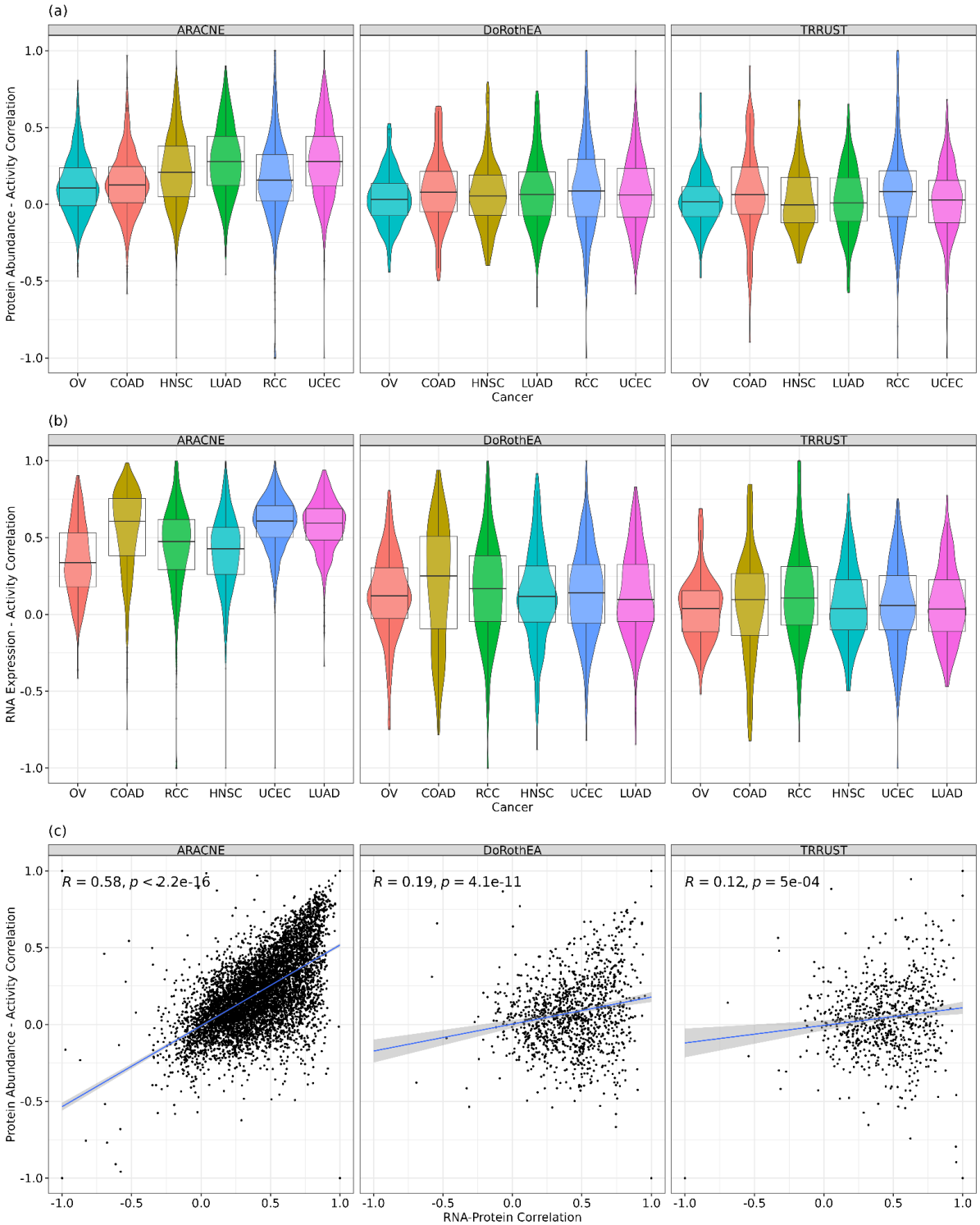


Fig. 4.11 Comparing the results of regulon enrichment analysis using different regulons (a) Correlations between protein abundance and gene output **(b)** correlations between RNA expression and gene output score **(c)** correlation

between RNA-protein correlation and protein abundance-gene output correlation. Regulon enrichment analysis was performed tissue-specifically on each transcriptional regulator using VIPER, for a total of 1,362 (ARACNE), 277 (DoRothEA) and 183 (TRRUST) genes. Each point in (c) represents a gene-cancer type combination. RNA and protein data were obtained from CPTAC.

4.3.7 Understanding TSG Dosage Responsiveness using Regulon Enrichment Analysis

Having calculated activity scores for each TSG across 29 TCGA cancer types using VIPER and the TRRUST regulons, I first investigated whether TSG-hallmark associations were visible at the activity score level. I expect that the activity score of a TSG should be negatively correlated with the level of the cancer hallmarks it influences. I find that the correlations between hallmark and activity score are generally very weak, but that hits (gene-cancer type-hallmark combinations) have a significantly higher proportion of significant negative correlations (33.8% vs 12.6%, p = 3.01e-6). This difference is visible in Fig. 4.12b as a hump around r = -0.5 among hits.

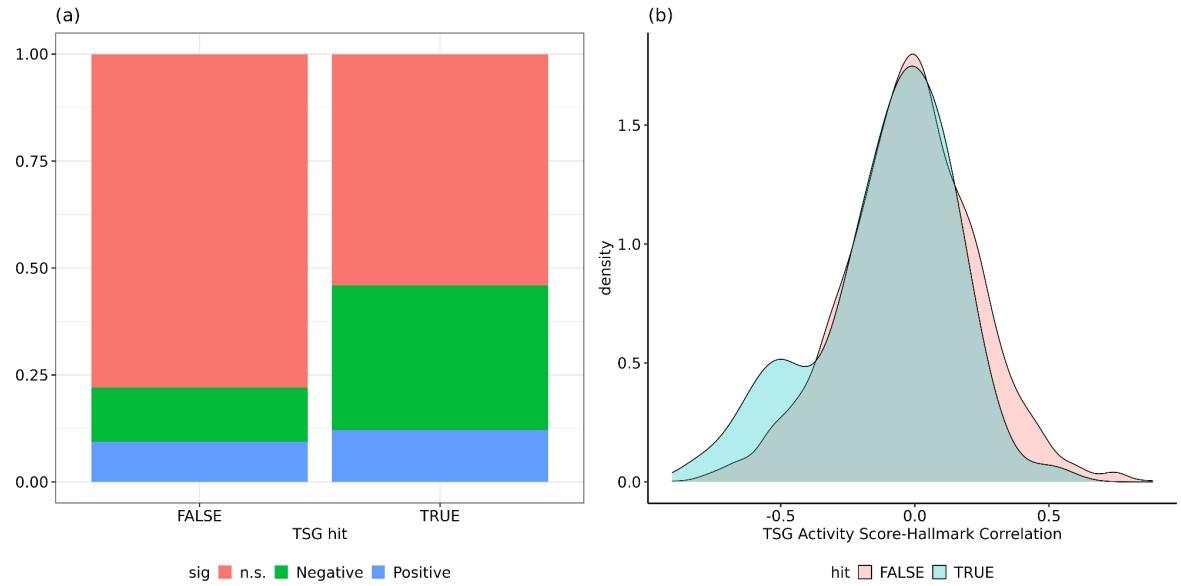


Fig. 4.12 Relationship between TSG Activity Score (output) and hallmark level. Hits are gene-tissue-hallmark combinations with Bonferroni-Holm-adjusted p < 0.05. **(a)** Proportion of gene-cancer type-hallmark combinations where TSG output

and hallmark level are positively/negatively correlated at nominal $p < 0.05$ **(b)** Distribution of correlations between TSG output and hallmark level, divided by hit status.

In order to understand whether dosage-responsiveness at the hallmark level is related to dosage-responsiveness at the gene output level, I computed ratios of the median hallmark level when each TSG was partially inactivated vs intact and compared them to ratios of the median activity score. I expect that if TSG haploinsufficiency status in this analysis is driven by the responsiveness of the TSG's activity to dosage changes, there should be a negative correlation between the activity score ratio and the dosage ratio.

I find that the plurality (357/732, 48.8%) of gene-cancer type-hallmark combinations had the expected hallmark ratios > 1 and activity score ratios < 1 when partially inactivated. This increased only slightly, to 53.8%, when I filtered to driver genes, i.e. TSGs identified as relevant in the cancer type being considered (Appendix Fig. 6.10). In general, there is little to no correlation between hallmark ratio and activity score ratio ($r = 0.02$); there is a slight negative correlation within driver genes, but it is not significant ($r = -0.119$, $p = 0.138$; r reaches -0.227 for CNH and -0.159 for mutation rate but is -0.07 for angiogenesis and 0.016 for proliferative index). There is also no significant relationship when I filter to the 25 gene-cancer type-hallmark combinations from Fig. 4.12 for which hallmark level is significantly negatively associated with activity score. Additionally, hits do not significantly differ in activity score ratio from non-hits, i.e. their activity level

is not more responsive to dosage alterations than that of other genes (Fig. 4.13).

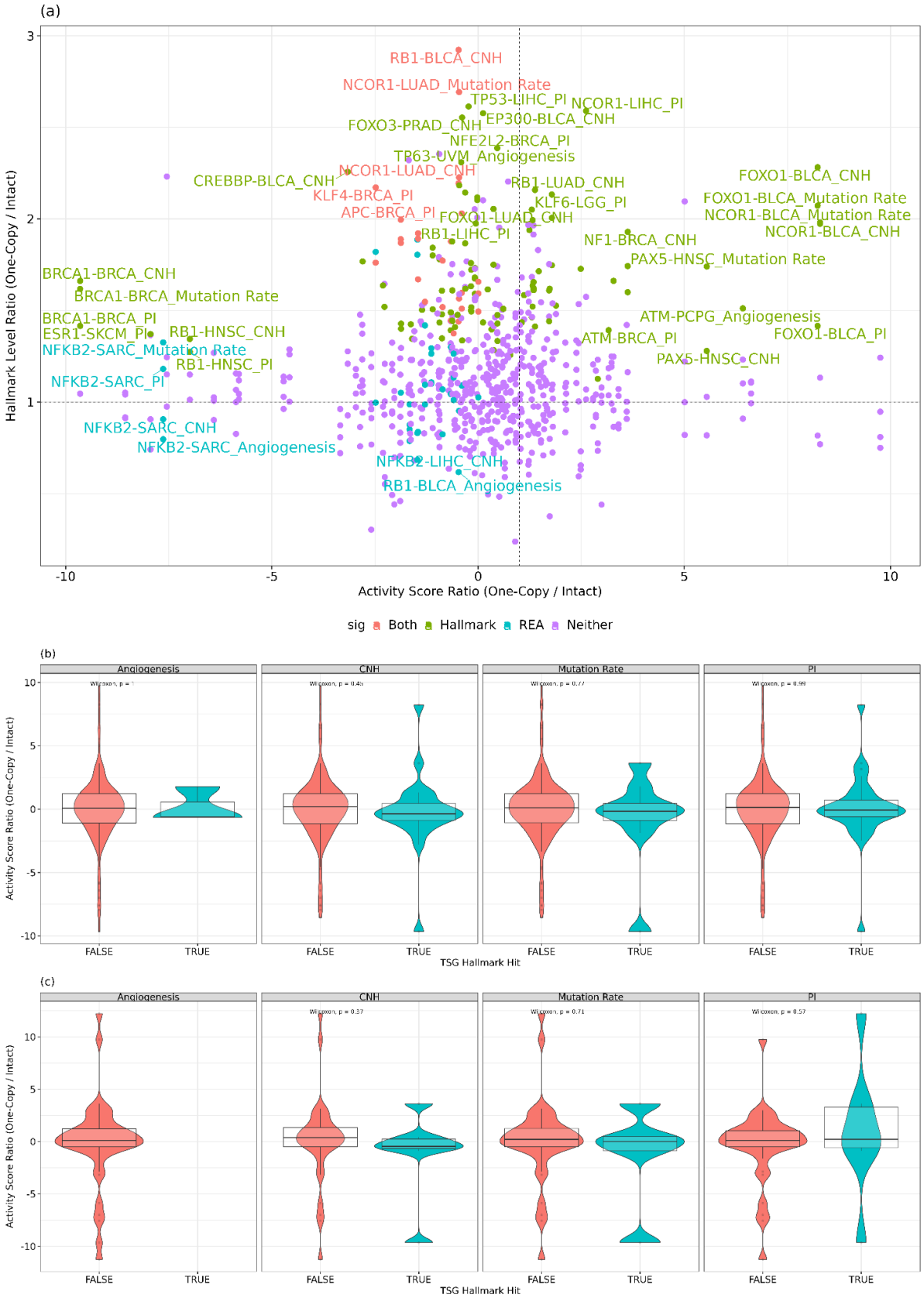


Fig. 4.13 Effect of partial TSG inactivation on Activity Score (output) and comparison to hallmark test results. (a) Ratio between activity scores when gene is partially inactivated vs intact compared to ratio between hallmark levels. To assess significance, one-sided Wilcoxon tests were performed; I tested if partial inactivation was associated with higher hallmark levels and with lower activity scores. Activity score values are repeated for different hallmarks. **(b)** Distribution of activity score ratios, comparing hallmark hits to non-hits. The angiogenesis hallmark had only 3 hits that were included in the TRRUST regulons database. In (a) and (b), regulon activity scores were constrained to (-10, 10). **(c)** As in (b), but only for genes that would be considered drivers in the relevant cancer type, defining a driver as having either dndscv global q-value < 0.05 or a frequency of complete inactivation ≥1%. I only computed ratios for gene-cancer type combinations with at least 10 partially inactivated and at least 10 intact cases.

4.3.8 Attenuation of Dosage Alterations

The partial inactivation events I study here occur at the DNA level and typically involve copy number losses. In general, for a copy number loss in a TSG to alter the behaviour of a tumour, it should affect the RNA expression, protein abundance, and ultimately functional output of the protein encoded by the TSG. However, the effects of copy number alterations are attenuated by the fact that RNA and protein levels do not perfectly follow DNA copy number. It has been shown repeatedly that DNA-RNA and RNA-protein correlations are only weak to moderate (Arad & Geiger, 2023), and lower in cancer than in healthy tissues (Kosti et al, 2016). These correlations depend on both gene-specific features, such as complex membership (Cheng et al, 2022), and other factors such as CNA length (Schukken & Sheltzer, 2022). I touched on the relationship between protein abundance and gene activity score (output) in 4.3.6, but here I investigate the idea of dosage attenuation further to gain some insight into possible mechanisms of haplo(in)sufficiency. Specifically, I will examine attenuation throughout the gene expression

process, analyse factors in gene attenuation levels, and finally test whether copy number alterations in haploinsufficient TSGs are less attenuated.

The attenuation of dosage changes from the DNA level to the protein activity level can be visualised as a cumulative loss of information related to the product of the correlations between each stage (Fig. 4.14). I find that the median DNA-RNA correlation is 0.32 and the median RNA-Protein correlation is 0.46, in agreement with previous studies, but the median Protein-Output correlation is minute at 0.026. Under this (simplified) model, the information a gene's expression level carries about its copy number thus decreases rapidly throughout the gene expression process, averaging 32% at the RNA level, 12% at the protein level and 1.3% at the level of gene functional output. Accounting for the genes with negative correlations, these values decrease slightly to 32% at the RNA level, 11.9% at the protein level and 0.05% at the output level. Some rare genes buck this trend to an extent, like IRF7 in head-and-neck cancer which retains 21% at the gene functional output level.

Do genes have a consistent level of attenuation between different stages? In other words, do genes with high DNA-RNA correlations also have high correlations between RNA and protein and protein abundance and activity score? As reported by Cheng et al (2022), DNA-RNA and RNA-protein correlations are actually (very weakly) negatively correlated (r between -0.043 and -0.19, Fig. 4.14b), and this is significant in all six cancer types analysed ($p < 0.001$). The relationship between RNA-protein correlations and protein

abundance-output correlations, on the other hand, is weakly positive albeit less significant due to smaller sample size, with correlations ranging from 0.091 to 0.25.

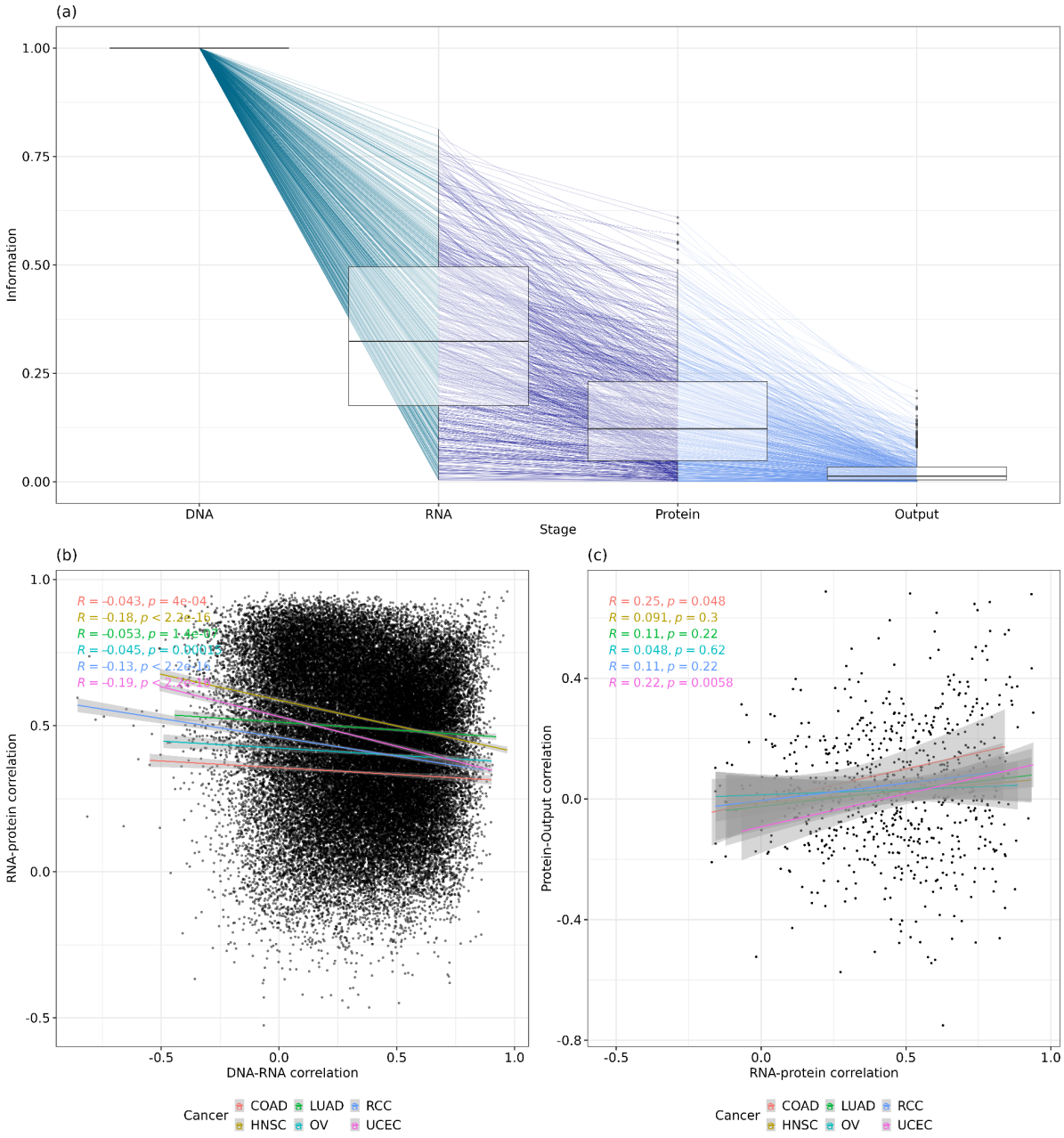


Fig. 4.14 Attenuation of dosage alterations between the DNA level and the gene's functional output. (a) Visualisation of cumulative information loss from copy number changes based on absolute correlations between DNA copy number and RNA expression, RNA expression and protein abundance, and protein abundance and VIPER-inferred-i gene output. Each line is a gene-cancer type combination; boxplots show the average information retained at each stage. The DNA level is set to 1. I do not consider gene mutation and assume it to be negligible for this visualisation—the median mutation frequency of a gene in the TRRUST database in

cancer is 1%. **(b)** Correlation between DNA-RNA correlation and RNA-protein correlation. Regression lines and correlation coefficients computed separately for each cancer type. **(c)** Correlation between RNA-protein correlation and protein abundance-functional output correlation. (a) and (c) include only transcriptional regulators included in the TRRUST database.

It has previously been shown that DNA-RNA and RNA-protein correlations can be affected by both gene-specific features such as complex membership (Cheng et al, 2022) and sample-specific features like CNA length (Schukken & Sheltzer, 2022). Linear regression (Eq. 4.3) was employed to analyse some gene features that are of interest in this study (cancer genes) or could plausibly influence attenuation levels at the DNA-RNA, RNA-Protein or Protein-Output stages (Fig. 4.15). These included complex membership (from CORUM) and two measures of dosage constraint. I also included the number of post-translational modifications known for a gene, using data on protein phosphorylation, acetylation and methylation from PhosphoSitePlus. The final term in the model was a score of protein abundance measurement reproducibility calculated by Upadhya & Ryan (2022), because they showed that the reproducibility of abundance measurements for a given protein increases its RNA-protein correlation and that this can create artefactual RNA-protein correlation differences between biological pathways.

As expected, complex membership was associated with increased RNA-protein and decreased DNA-protein correlations, as was the protein measurement reproducibility score. Interestingly, I also found that non-dosage constrained genes (which vary in copy number between species) had significantly lower RNA-protein correlations in 5/6 cancer types, and this was

also seen at the DNA-RNA stage in colon and kidney cancer. This suggests the intriguing possibility that some dosage-constrained genes are dosage-constrained because they are more faithfully expressed, whereas other genes with the same importance or protein dosage constraint might escape negative selection on dosage variation because they are dosage-compensated, i.e. the copy number change is not felt. This is partially supported by the fact that genes essential-haploinsufficient for proliferation in embryonic stem cells (Sarel-Galilly et al, 2022), another aspect of dosage constraint, had higher DNA-RNA correlations in 4/6 cancer types, although this effect was not seen at the RNA-Protein stage. Oncogenes and tumour suppressor genes generally did not differ from other genes in their attenuation levels, but RNA-protein correlations were higher for TSGs and oncogenes in head-and-neck squamous cell carcinoma and, for TSGs, lung adenocarcinoma. No gene features were significantly associated with Protein-Output correlation, including number of post-translational modifications. However, there were far fewer genes available for this stage than for the previous two (179 vs 12,422). Also, while it would make sense for the protein abundance-activity relationship of a gene to be influenced to some degree by the complexity of its post-translational modifications, a simple count of their number may not be sufficiently fine-grained.

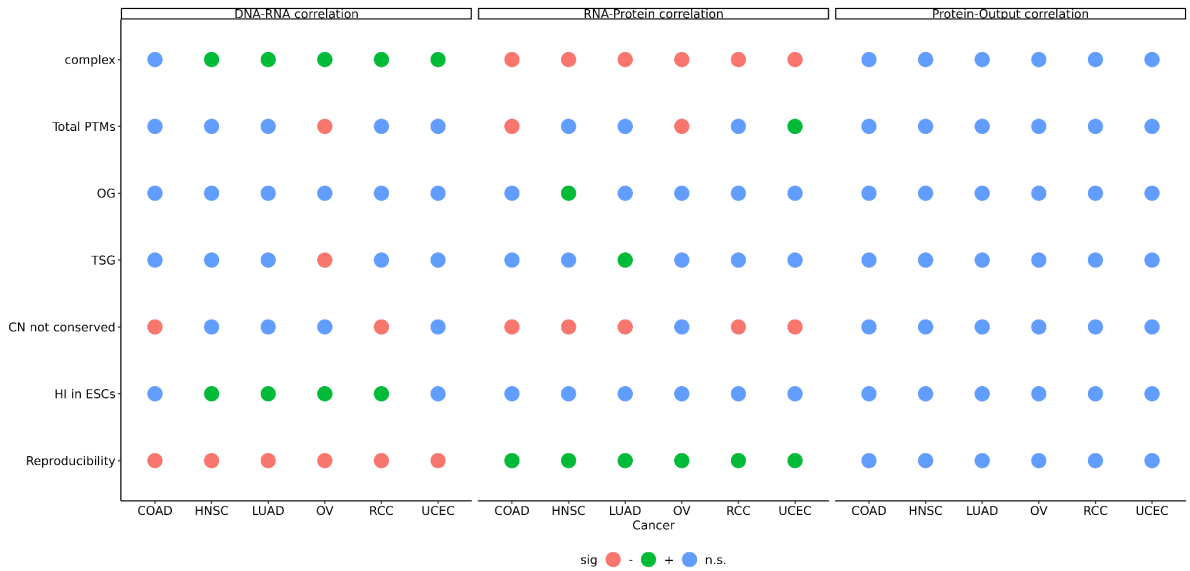


Fig. 4.15 Linear regression of attenuation coefficients on gene features. PTMs = post-translational modifications from PhosphoSitePlus (phosphorylation, acetylation and methylation); OG = oncogene; TSG = tumour suppressor gene; complex = complex membership status; CN not conserved = gene without conserved copy number across mammalian evolution, Class P CNV = gene involved only in pathogenic CNVs (Rice & McLysaght, 2017). Reproducibility is the gene-specific aggregated protein measurement reproducibility score from Upadhya & Ryan (2022). Regression values were Benjamini-Holm corrected for multiple testing across all combinations of feature, cancer and attenuation stage. DNA-RNA and RNA-protein correlations used all available data; protein abundance-output correlations used only the genes included in the TRRUST regulon dataset.

Finally, I asked whether the one-hit TSGs I have identified might be haploinsufficient because their dosage changes are less attenuated. I find that hits do not have significantly higher DNA-RNA or RNA-protein correlations than non-hits, especially when I analyse only genes that would have been included in the haploinsufficiency tests (≥10 partial inactivations) or genes identified as drivers in the relevant cancer type based on mutational patterns or complete inactivations (Fig. 4.16). To account for noise introduced by 'passenger TSGs' in hit islands, I filtered the hits to genes independently identified as haploinsufficient in the literature and found the same pattern (Appendix Fig. 6.13). This analysis is limited to the six CPTAC cancer types

for which I could compute RNA-protein correlations, which includes 52 hits for Fig. 4.16a-b and only 7 for Fig. 4.16c. Still, it suggests that the haploinsufficient TSGs detectable using this method are not distinguished from other TSGs by their attenuation levels, and may instead represent haploinsufficient behaviour within strong tumour suppressors.

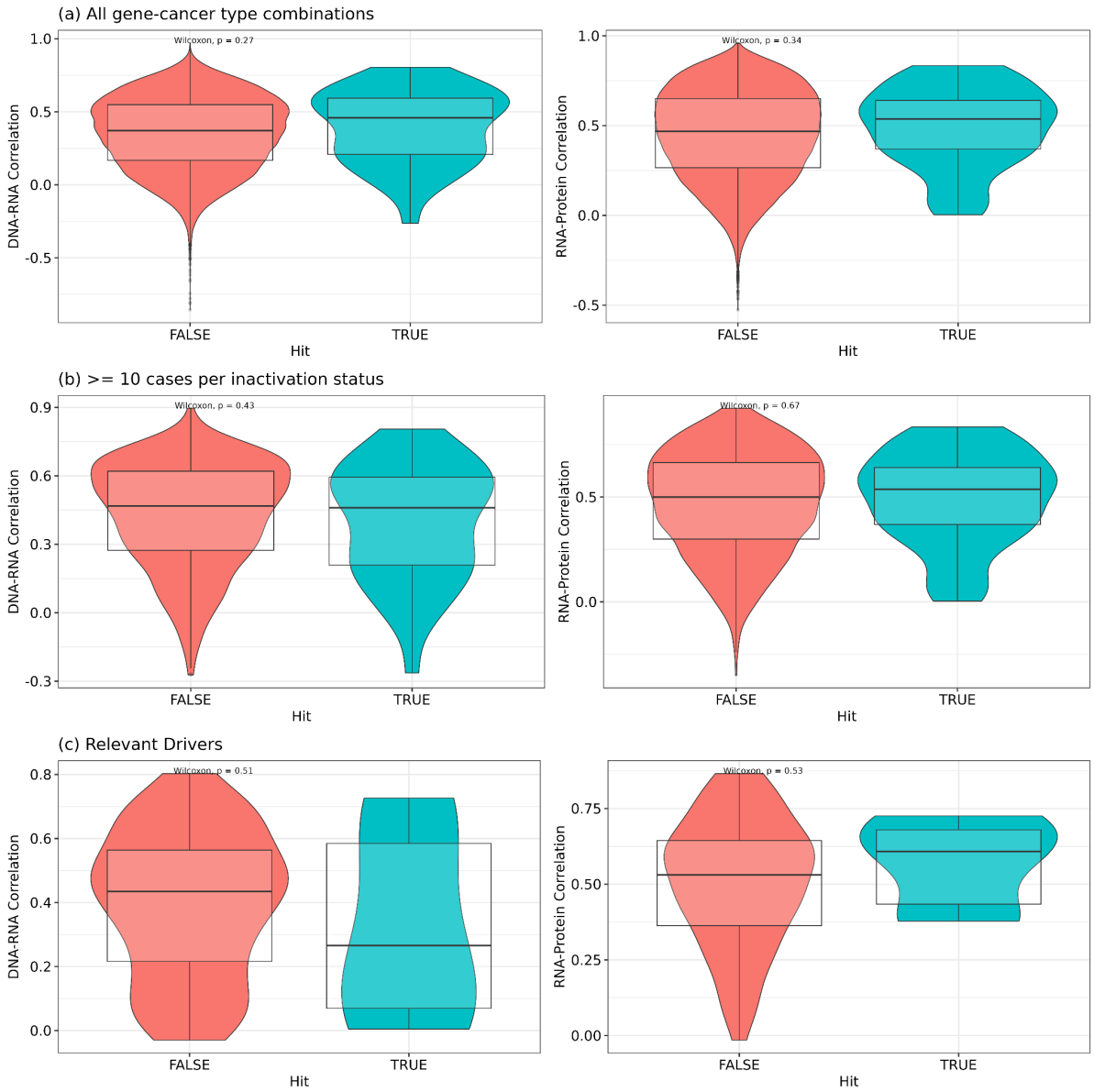


Fig. 4.16 DNA-RNA and RNA-Protein correlations by hit status (a) all gene-cancer combinations with copy number, RNA and protein abundance data available from CPTAC **(b)** only combinations with at least 10 partially inactivated cases and at least 10 intact cases for the relevant gene **(c)** only combinations identified as

relevant drivers based on mutations (dndscv global $q < 0.05$) or complete inactivation rate ≥1% (visualised in Appendix Fig. 6.10).

4.3.9 Robustness

One important mechanism of TSG inactivation that I have not fully addressed thus far is promoter hypermethylation. Undetected promoter methylation could lead to the classification of a fully-inactivated TSG as partially-inactivated or intact. I have attempted to mitigate this issue by requiring non-fully inactivated TSGs to maintain some expression as detected by RNA-seq, but in this section I will incorporate TCGA methylation data to directly examine the robustness of my results to promoter hypermethylation. Because the effect of promoter hypermethylation is not consistent across all genes, genes defined as hypermethylated in a sample (see Methods 4.2.2) were excluded. Of the 1,595,298 gene-patient combinations included in the original analysis, 10.3% of those originally classified as intact and 14.4% of those classified as partially inactivated did not have methylation data available, while 12.1% and 10.6%, respectively, were excluded for hypermethylation. This resulted in 1,513 of the original 5,958 gene-cancer type combinations failing to meet sample size thresholds. Of the 491 original TSG hallmark hits, 120 were excluded in the methylation analysis due to insufficient data; of the remainder, the majority (78.4%) were recaptured. The others may be due to previously undetected methylation or to decreases in sample size. The lost hits include some of the novel hits identified in 4.3.3 (RHOH for mutation rate and CNH in BRCA, N4BP2 for mutation rate in BRCA). There were also 11 new hits detected in the methylation analyses,

including PAX5 for proliferation and mutation rate in BRCA, and ROBO2 for CNH in BRCA and HNSC. Of the survival hits, a large majority were recaptured in the methylation-aware analysis (86.4% of those available for analysis, and 51/66 original hits), and six new hits were found (KLF6-MESO, MGMT-MESO, TSC1-KIRC, SFPQ-BRCA, PTPRD-KIRC, CD274-KIRC). In the original analysis, no STOP genes were significantly associated with a hallmark after multiple testing correction; in the methylation-aware analysis, five were: DUSP8 in STAD and FBX025, TUSC3, ERI1 and SOX7 in BRCA). This is likely due at least in part to the loss of TSG hits, because deletions involving significant TSGs were removed from the STOP genes analysis as potential confounders. The number of STOP genes associated with worsened prognosis when partially inactivated decreased from 19 to 11, although only 7/11 were hits in the initial analysis.

The hallmark association tests were supplemented with a linear regression controlling for age and sex (Eq. 4.2), with the three genotypes (intact, partial inactivation, complete inactivation) encoded as factors. This was performed using the methylation-aware classification system and resulted in the recapturing of the vast majority of hits from the association tests (294/312, 94.2%), This comparison included only significant hits associated with an increase (rather than a decrease) in a hallmark, which made up 83.4% (891/1,068) of all significant hits in this analysis, suggesting that partial inactivation of a non-negligible minority of TSGs is associated with a decrease in one or more cancer hallmarks, at least in some cancer types. The similarity

to the binary association tests may be explained by the fact that in most cases, only two genotypes (intact and partial inactivation) were present. Of the 52 gene-cancer type combinations for which complete TSG inactivation is significantly associated with an increase in a hallmark, partial inactivation is also associated with a significant increase in 32 of them (61.5%). In these cases, the effect size estimates of partial inactivations relative to complete inactivations ranged from 0.11 to 1.79, with 30/32 (93.75%) falling between 0 and 1, indicating a stronger effect for complete inactivation than partial inactivation. The median effect size ratio was 0.397, suggesting that, for these genes, partial inactivation has somewhat less than half the effect of complete inactivation (Appendix Fig. 6.14).

4.4 Discussion

In this work, I developed a computational approach to screen for alterations in tumour suppressor gene function upon monoallelic inactivation. I identified 209 TSGs involved in genomic islands whose partial inactivation was associated with significant increases in proliferation, copy number heterogeneity, mutation rate or angiogenesis or with a significant decrease in patient survival, including 12 novel haploinsufficient TSGs. I also quantified dosage responsiveness levels for each TSG and found evidence supporting TSG status for 19 STOP genes previously linked to proliferation in non-cancerous cells. Finally, I integrated DNA → RNA → protein correlations and

regulon enrichment analysis to show that hits in this screen are not distinguished from non-hits by their dosage attenuation profiles, i.e. they are not preferentially genes for which dosage changes have large effects on protein abundance and gene activity. This suggests the alternative explanation that the screen tends to favour haploinsufficient behaviour in strong TSGs with major effects on tumour hallmarks, and not necessarily TSGs with high intrinsic dosage sensitivity.

There are several limitations to my approach. The specificity of this screen for haploinsufficient tumour suppressors, and particularly the quantification of dosage-responsiveness ratios, is sensitive to misclassification of complete inactivation events as partial inactivation events. For this reason, I was as stringent as possible with the classification system, integrating both germline and somatic mutations and copy number alterations along with RNA-seq data and excluding any ambiguous cases. I did not directly include data on epigenetic silencing via promoter methylation in the main analyses because this does not consistently predict gene silencing (Jesionek-Kupnicka et al, 2014), and even in cases where it does it is not clear what methylation threshold one could choose to define a silencing event. I did attempt to account for (strong) epigenetic silencing by requiring that partially-inactivated genes retain some RNA expression, but in some cases this criterion could be satisfied by a TSG expressing only a mutant allele. However, a robustness analysis suggested that a large majority (~80%) of hits were recaptured when methylation data was integrated into the classification

system. There are also additional mechanisms of TSG inactivation that I could not address here, such as sequestration (e.g. SMAD4, Xiao et al, 2024) and overexpression of TSG inhibitors (e.g. MDM2/4 and TP53, Valentin-Vega et al, 2007).

Secondly, there is potential for confounding in the copy number heterogeneity hallmark because chromosomally unstable tumours are more likely to have a deletion encompassing any given TSG. It is unclear how strong this effect is, given that the vast majority of TSGs and all of the STOP genes were not hits for CNH despite being inactivated primarily by copy number loss, and I decided to include CNH as a hallmark regardless because maintenance of chromosomal stability is an important role for many tumour suppressors. Still, results for this hallmark must be interpreted with caution. I investigated whether this might affect other readouts, such as survival, by reanalysing the survival hits after including CNH as a covariate in the Cox proportional hazards regression. 57/66 survival hits remained significant ($p_{adj} < 0.05$); in the other 9, the survival association could be confounded or the TSG's survival effect might be via its effect on CNH. Confounding could also occur with the mutation rate hallmark, but this is a lesser concern because the vast majority of partial inactivations occur via copy number loss, not mutation. The paucity of mutations also precluded an investigation into dominant negative mechanisms of TSG inactivation.

In general, however, I have tried to err on the side of caution, and it seems likely that the results of this screen are conservative, i.e. that there are significant numbers of haploinsufficient TSGs I couldn't detect. This is partially due to the stringent classification system, which resulted in my excluding a non-negligible fraction of TSG mutation events, and possibly to the multiple testing correction method, which can inflate the false-negative rate when a large proportion of genes are true hits. The analysis is also biased towards TSGs that operate through one of the hallmarks measurable using data in The Cancer Genome Atlas; I could not include other important hallmarks such as apoptosis and senescence, which are better detected using other assays like caspase cleavage (Zhang et al, 2013) and SA-ß-gal activity (Itahana et al, 2007). This screen was performed exclusively within cancer samples, so it does not account for any TSGs that are only relevant in the first stages of oncogenic transformation and do not affect at least one of survival, proliferation, mutation rate or genomic instability within cancer cells. Additionally, my method of removing the influence of known (haploinsufficient) TSGs from my secondary screens (by excluding any copy number segment with a deletion of a known hit), while necessary, is very conservative: it removes genes located near significant TSGs even if they would be genuine hits, including genes which cooperate to form islands of compound haploinsufficiency (Xue et al, 2012; Chen et al, 2019). Finally, while I was able to identify tissue- and hallmark-specific haploinsufficiency, this analysis was underpowered to analyse other contextual factors such as genetic background (compound haploinsufficiency), ploidy (is a one-copy TSG

as effective in a tetraploid as in a diploid?) and copy number alteration size (are these associations affected by dosage compensation of long deletions?).

One major aspect of TSG dosage responsiveness that I have not addressed here is their sensitivity to copy number gains. Several haploinsufficient TSGs are known to show oncogenic behaviour in certain contexts, including TP53 (Soussi & Wiman, 2015), KMT2D (Lv et al, 2017) and KLF4 (Qi et al, 2018). Certain copy number gains in tumour suppressors have recently been shown to be under positive selection, likely due to increased dosage of dominant negative alleles (Besedina & Supek, 2024), but this could also arise due to dosage sensitivity: alterations in a highly dosage-sensitive TSG that requires an exact stoichiometry with its interaction partners could theoretically promote tumorigenesis by lowering *or* raising its dosage. TSG dosage-responsiveness could also manifest in the opposite way, with a decrease in TSG dosage promoting proliferation and an increase suppressing it below normal levels; this is likely more common, as TSG density on a chromosome is negatively correlated with amplification frequency in cancer (Davoli et al, 2013).

I close with the question of defining haploinsufficiency. While it is generally accepted to mean the inability of a single intact allele to maintain 'normal function', it is unclear to what degree function must be impaired for the gene to be considered haploinsufficient. When Andrianova et al (2024) identified POLD1 as haplosufficient in humans, for example, they noted that its partial

inactivation was associated with an increase of 'less than 15%' in mutation rate. I suggest that even small differences in function can have important implications for cancer development, e.g. for genes like BRCA1, where monoallelic inactivation leads to an increase in genomic instability and accelerates the evolution of the tumour (Konishi et al, 2011). The design of my screen allowed me to quantify haploinsufficiency based on the level of hallmark activity in different TSG inactivation states. A complementary approach could be to calculate '%-sufficiency', i.e. the level of gene expression required to maintain function at normal levels. Along these lines, it has been shown using a hypomorphic allelic series that PTEN requires at least 80% expression for tumour suppression (Carracedo et al, 2011), and that shRNA knockdown of the essential gene PRIM1 begins to impair cell viability once expression falls below 20% (Nichols et al, 2020). One future direction could be to compare the results of CRISPR knockouts and RNAi knockdowns from Project Achilles (Tsherniak et al, 2017): highly dosage-sensitive genes might have more similar results between the two screens.

In summary, this study adds to the growing evidence that TSG haploinsufficiency has important functional effects on tumour development and may be the rule, not the exception. It also supports a model of continuous dosage-responsiveness, wherein TSGs are not cleanly divided into 'two-hit' vs 'haploinsufficient' classes but instead drive tumorigenesis depending on a combination of their degree of inactivation and their (cancer type- and genetic background-specific) importance in regulating cell behaviour.

Chapter 5

Conclusions

The aim of this thesis was to investigate the consequences of large-scale copy number alterations across cancer types from an evolutionary perspective, with a focus on the development of extreme ploidy and the impact of dosage sensitivity in cancer. In chapter 2, I identified polyploid and hypodiploid tumours across The Cancer Genome Atlas and analysed genetic and environmental factors associated with the development of extreme ploidy in cancer. Factors identified as significantly associated with polyploidy included proliferative index, mutations in TP53, lack of mutations in ARID1A and PTEN, hypoxia, HBV infection (in liver cancer), and five mutational signatures. Random forest modelling indicated that the development of

polyploidy can be predicted with reasonable accuracy by this small set of factors, and that the most important factors by far are replication rate, TP53 mutation and hypoxia. The development of hypodiploidy was also significantly associated with TP53 mutations, replication rate and hypoxia, indicating some commonality between these two opposite ploidy states.

For a cancer cell to become (detectably) polyploid or hypodiploid, it must pass through three stages: (i) occurrence of the mitotic error(s) (ii) initial establishment of the ploidy-aberrant subclone (iii) expansion. The factors significantly associated with ploidy in this study were generally related to the first two of these stages: rapid replication and hypoxia increasing the likelihood of mitotic errors, and TP53 mutation (or inactivation via HBV infection) allowing the ploidy-aberrant subclone to survive. This, combined with the predictive ability of the random forest model, supports the idea put forth by Bielski et al (2018) that the rate of polyploidy in patient tumours is largely determined by the rate of occurrence of stochastic mitotic errors in cells with a defective G1 checkpoint. However, this work focused on primary, untreated tumours and the story may be different in treated or metastatic cancers, as both the rapidly changing environment of a metastasising tumour and the presence of chemotherapeutics are likely to act as selective factors affecting the expansion of ploidy-aberrant clones. In particular, the effect of exposure to specific chemotherapeutics on ploidy alteration rate in patient tumours could have clinical implications and is an important direction for future research.

In chapter 3, I analysed patterns of chromosome loss in the establishment of hypodiploidy and in the evolution of post-WGD tumours. I began by expanding the dataset of hypodiploids to include karyotypes for acute lymphoblastic leukaemia (ALL) cases. I formalised a clinical heuristic to detect masked (genome-doubled) hypodiploid ALL and built multi-sample trees of hypodiploid and genome-doubled subclones to test its performance across the Mitelman Database of Chromosome Aberrations & Gene Fusions in Cancer, finding very good specificity (>99%) and reasonable sensitivity (~90%). This allowed me to estimate the rate of fully masked hypodiploidy in ALL at 6-7% of hypodiploid cases, suggesting that this prognostically-challenging subgroup makes up a very minor fraction of total leukaemia cases (~0.06%). I could also estimate the rate of WGD in near-haploid and low-hypodiploid ALL, finding dramatic increases in WGD rate with degree of hypodiploidy, consistent with previous results (Safavi & Paulsson, 2017). This pattern additionally extended to the solid tumours in The Cancer Genome Atlas.

Why are hypodiploid tumours more likely to undergo WGD? Survival analysis does not support the idea that WGD increases the fitness of the hypodiploid clone by alleviating dosage imbalances. Instead, there appears to be a general chromosomal instability phenotype that explains both increases and decreases in ploidy, and both large-scale and small-scale (intrachromosomal) copy number alterations. This is based on two observations: (i) polyploid and

hypodiploid tumours have higher intra-tumour copy number heterogeneity and more intrachromosomal copy number alterations, and (ii) cancer types with higher rates of polyploidy also have higher rates of hypodiploidy, higher copy number heterogeneity and more intrachromosomal alterations even in non-polyploid samples, so the elevated chromosomal instability is not (exclusively) an effect of the ploidy alteration itself. This common underlying factor suggests that the factors identified as explaining rates of polyploidy in chapter 2 may also be relevant to other kinds of copy number alteration. The relationship between WGD rate and general chromosomal instability also has interesting implications for the stability of polyploidy in cancer. Post-WGD cancer evolution is generally characterised by large-scale loss of genomic material, with initial rapid losses followed eventually by stasis. There are notable differences between cancer types in their level of retention of duplicate copies after WGD, and this rate is strongly inversely correlated with the rate of WGD in the relevant cancer type. This suggests an 'easy-come, easy-go' model for polyploidy in cancer, wherein polyploids in cancer types that rarely undergo WGD are more stable and thus persist as polyploids for longer.

Finally, I examined patterns of chromosome loss in hypodiploid tumours. Clustering of chromosome loss rates suggested two groupings of cancer types with frequent hypodiploidy: one where hypodiploids form a distinct subgroup with stereotyped chromosome loss patterns (ALL, kidney chromophobe, adrenocortical carcinoma) and one in which chromosome losses are largely

random and hypodiploid cases merely represent the high end of a spectrum of chromosomal instability. This conclusion is supported by the significantly lower levels of small-scale copy number alterations and intra-tumour heterogeneity in the first group. Patterns of chromosome loss in hypodiploid tumours show some similarity to patterns in near-diploid tumours of the corresponding cancer type, but are not generally explained by selective or mechanical chromosome features. This suggests a model wherein the signal from (likely early) driver losses is overwhelmed by subsequent massive losses across the genome. In particular, there is no evidence that chromosome losses in hypodiploid tumours avoid monosomy of dosage-sensitive genes. This is in keeping with the general inefficiency of negative selection in cancer, a conclusion further supported by the fact that post-WGD gene losses do not follow the dosage balance hypothesis, and that germline dosage-sensitive genes are not copy number-constrained even in less-chaotic diploid tumours. An interesting direction of future research would be to investigate whether monosomies of any germline dosage-sensitive genes are in fact deleterious in cancer, and if so how much genetic load they create in hypodiploid tumours.

In chapter 4, I studied haploinsufficiency in cancer from the opposite perspective: not as a source of dosage constraint but as a mechanism for positively-selected driver alterations in tumour suppressor genes (TSGs). I designed a screen based on observable changes in cancer cell behaviour upon TSG inactivation to identify haploinsufficient TSGs. This approach resulted in the identification of 12 novel candidate haploinsufficient TSGs, aided in

prioritising known haploinsufficient genes as potential TSGs, and provided information on the tissue-specificity of haploinsufficiency in established haploinsufficient TSGs as well as on their functional roles in tumour suppression. It also enabled the quantification of TSG haploinsufficiency based on the magnitude of changes in cancer hallmarks when a TSG was monoallelically vs biallelically inactivated. Having identified a set of haploinsufficient TSGs, I then analysed the level of dosage attenuation for each gene and found that effective TSG haploinsufficiency is not predicted by the faithfulness of its expression but may instead depend on the potency of the TSG, i.e. on its importance for tumour suppression. Taken together, these results emphasise the widespread relevance of TSG haploinsufficiency in cancer and support the emerging consensus that TSGs are not generally 'two-hit' or 'one-hit' but have a continuous relationship with tumour suppression depending on their importance and their level of inactivation.

This screen was somewhat limited by the fact that the vast majority of the partial inactivation events occurred via large chromosome loss events covering many genes, making it difficult to identify individual hits from genomic 'islands' of significantly-associated genes. However, the monoallelic inactivation of multiple nearby TSGs by large losses is a biologically significant phenomenon and a valuable area of study in its own right**.** A single large loss can simultaneously inactivate multiple cooperating tumour suppressor genes, increasing the potency of weak TSGs or even potentially converting tumorigenesis from a multi-step process into a saltatory event.

Additionally, the size of a loss may affect its likelihood of being compensated. It has been shown that large-scale gains of at least some oncogenes are dosage-compensated, while focal gains are not (Schukken & Sheltzer, 2022). Future work should investigate this phenomenon for copy number losses in tumour suppressor genes at multiple levels. Are large losses also more likely to be dosage-compensated at the RNA expression and protein abundance levels? Does this translate to differences in TSG activity within the cell, and to differences in cell behaviour or cancer hallmarks? If so, what are the mechanisms by which larger alterations promote dosage compensation?

Finally, my analysis of dosage attenuation in cancer showed that genes dosage-constrained in the human germline tend to have higher RNA-protein and DNA-RNA correlations, i.e. to be more dosage-responsive than other genes. If this pattern extends to non-cancer cells, it suggests the possibility that these genes are dosage-constrained at least in part because their dosage changes are less attenuated. As a corollary, there are likely to be genes with expected dosage-critical functions that are nevertheless not under dosage constraint in germline evolution because their dosage changes are strongly attenuated or compensated. Transcription factors are a particularly interesting gene category in this regard because of their capacity to regulate their own expression, and a useful direction of future research could be to test whether non-dosage constrained transcription factors are more likely to have autoregulatory negative feedback loops. Modelling of energy homeostasis supports this idea, suggesting that metabolic enzymes tend to be dosage-

insensitive because of the presence of negative feedback loops in metabolic pathways (Dong et al, 2023).

Taken together, the results of this thesis emphasise the weakness of negative selection and the unifying role of chromosomal instability in cancer copy number evolution. They also support a bidirectional relationship between experimental and computational validation, with mechanistic insights from cell line studies being tested within patient cohorts and contextualised in tissue-specific functional screens. Future work would benefit enormously from increased resolution of tumour copy number heterogeneity via single-cell sequencing and from the development of methods to time copy number losses. From a translational perspective, it will be essential to characterise in detail the genetic load contributed by aneuploidy in cancer, with the ultimate aim of finding and targeting the 'tipping points' where chromosomal instability becomes tumour-suppressive.

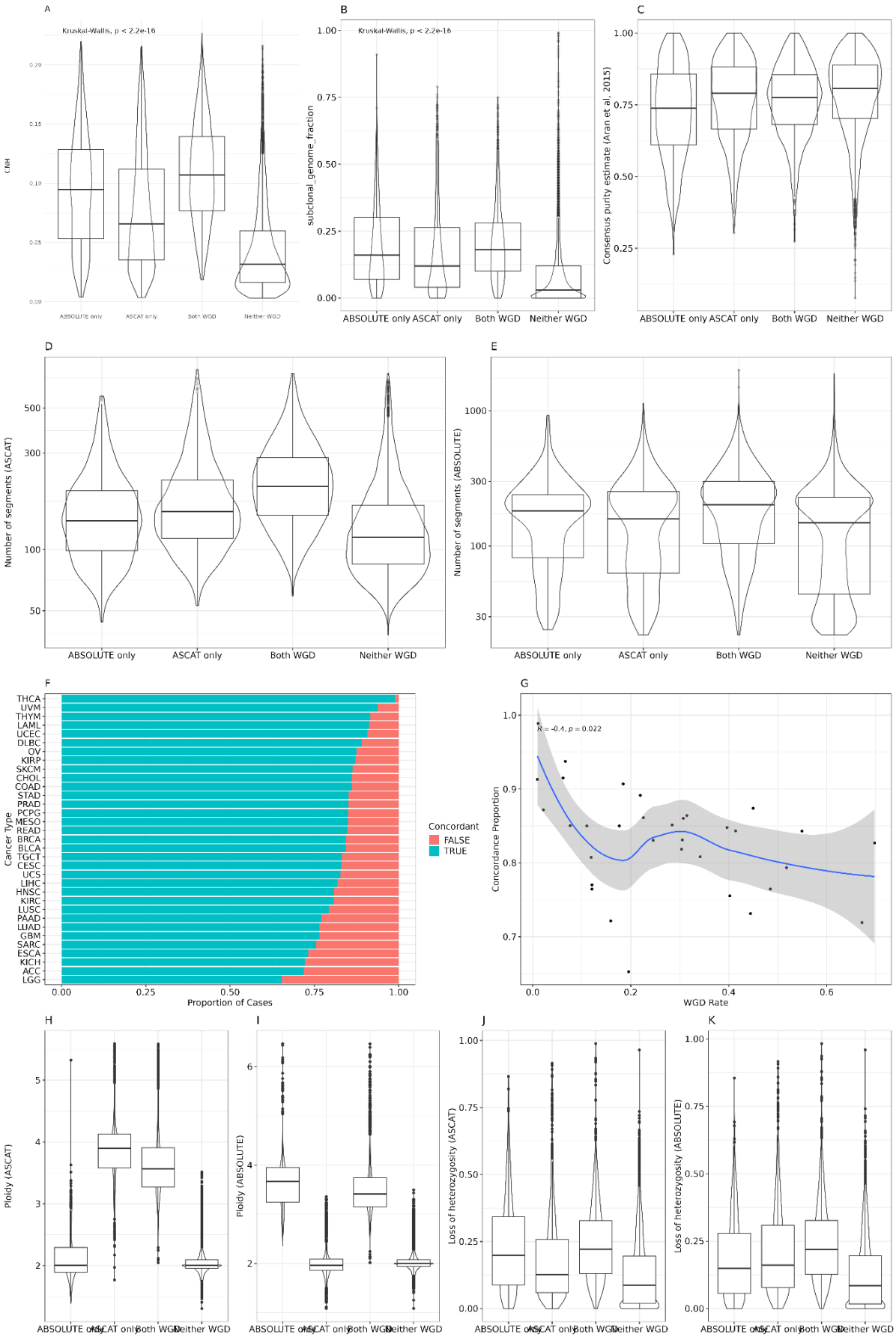
Appendix

Code	Name	TUSON	GTEx
LAML	Acute Myeloid Leukemia		
ACC	Adrenocortical carcinoma	N/A (assigned pan-cancer)	Adrenal gland
BLCA	Bladder Urothelial Carcinoma		
LGG	Brain Lower Grade Glioma		
BRCA	Breast invasive carcinoma	Breast	Breast mammary tissue
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma		
CHOL	Cholangiocarcinoma		
COAD	Colon adenocarcinoma		
ESCA	Esophageal carcinoma		
GBM	Glioblastoma multiforme		
HNSC	Head and Neck squamous cell carcinoma		
KICH	Kidney Chromophobe	N/A (assigned pan-cancer)	Kidney cortex
KIRC	Kidney renal clear cell carcinoma		
KIRP	Kidney renal papillary cell carcinoma		
LIHC	Liver hepatocellular carcinoma		
LUAD	Lung adenocarcinoma	Lung adenocarcinoma	Lung
LUSC	Lung squamous cell carcinoma	Lung squamous cell	Lung
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma		
MESO	Mesothelioma		
OV	Ovarian serous cystadenocarcinoma	Ovarian	Ovary
PAAD	Pancreatic adenocarcinoma		
PCPG	Pheochromocytoma and Paraganglioma		
PRAD	Prostate adenocarcinoma		
READ	Rectum adenocarcinoma		

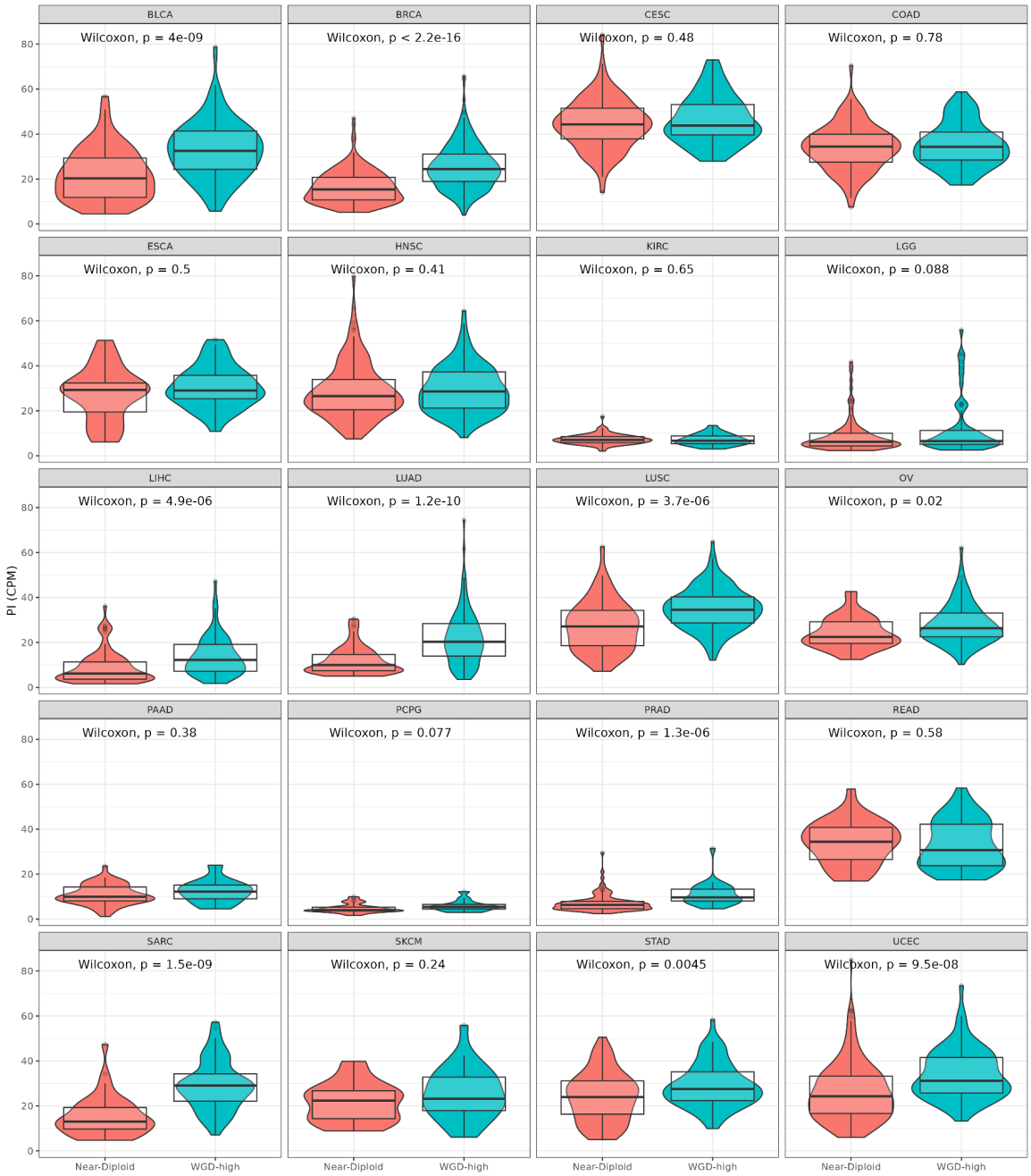
SARC	Sarcoma		
SKCM	Skin Cutaneous Melanoma		
STAD	Stomach adenocarcinoma	Stomach	Stomach
TGCT	Testicular Germ Cell Tumors		
THYM	Thymoma		
THCA	Thyroid carcinoma		
UCS	Uterine Carcinosarcoma		
UCEC	Uterine Corpus Endometrial Carcinoma		
UVM	Uveal Melanoma		
NH-ALL	Near-Haploid Acute Lymphoblastic Leukaemia	All hematopoietic	Whole blood
LH-ALL	Low-Hypodiploid Acute Lymphoblastic Leukaemia	All hematopoietic	Whole blood

Appendix Table 6.1 Cancer type abbreviations. All abbreviations and names except NH-ALL and LH-ALL are taken from the TCGA Study Abbreviations page[8], Columns 3 and 4 show the tissue groupings from Davoli et al and GTEx to which I matched each hypodiploid cancer type in Chapter 3. The TUSON tissues are the groupings for which Davoli et al (2013) identified tissue-specific cancer drivers using TUSON. TCGA cancer types that didn't have a close match in the tissue-specific TUSON analysis were assigned driver genes from the pan-cancer TUSON calls.

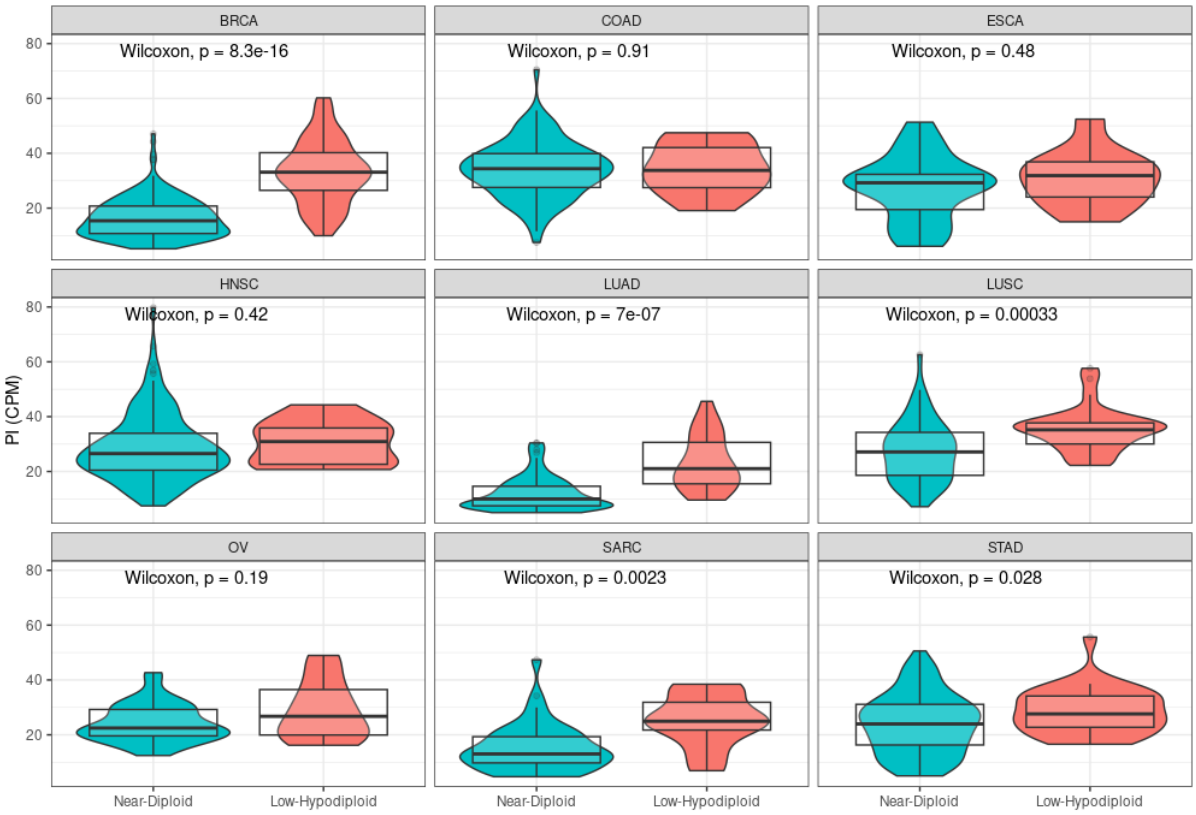
[8] https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations


A
Kruskal-Wallis, p < 2.2e-16
CNH
ABSOLUTE only
ASCAT only
Both WGD
Neither WGD
B
Kruskal-Wallis, p < 2.2e-16
subclonal_genome_fraction
ABSOLUTE only ASCAT only Both WGD Neither WGD
C
Consensus purity estimate (Aran et al, 2015)
ABSOLUTE only ASCAT only Both WGD Neither WGD
D
Number of segments (ASCAT)
ABSOLUTE only
ASCAT only
Both WGD
Neither WGD
E
Number of segments (ABSOLUTE)
ABSOLUTE only
ASCAT only
Both WGD
Neither WGD
F
Cancer Type
THCA
UVM
THYM
LAML
UCEC
DLBC
OV
KIRP
SKCM
CHOL
COAD
STAD
PRAD
PCPG
MESO
READ
BRCA
BLCA
TGCT
CESC
UCS
LIHC
HNSC
KIRC
LUSC
PAAD
LUAD
GBM
SARC
ESCA
KICH
ACC
LGG
Proportion of Cases
Concordant
FALSE
TRUE
G
R = -0.4, p = 0.022
Concordance Proportion
WGD Rate
H
Ploidy (ASCAT)
I
Ploidy (ABSOLUTE)
J
Loss of heterozygosity (ASCAT)
K
Loss of heterozygosity (ABSOLUTE)

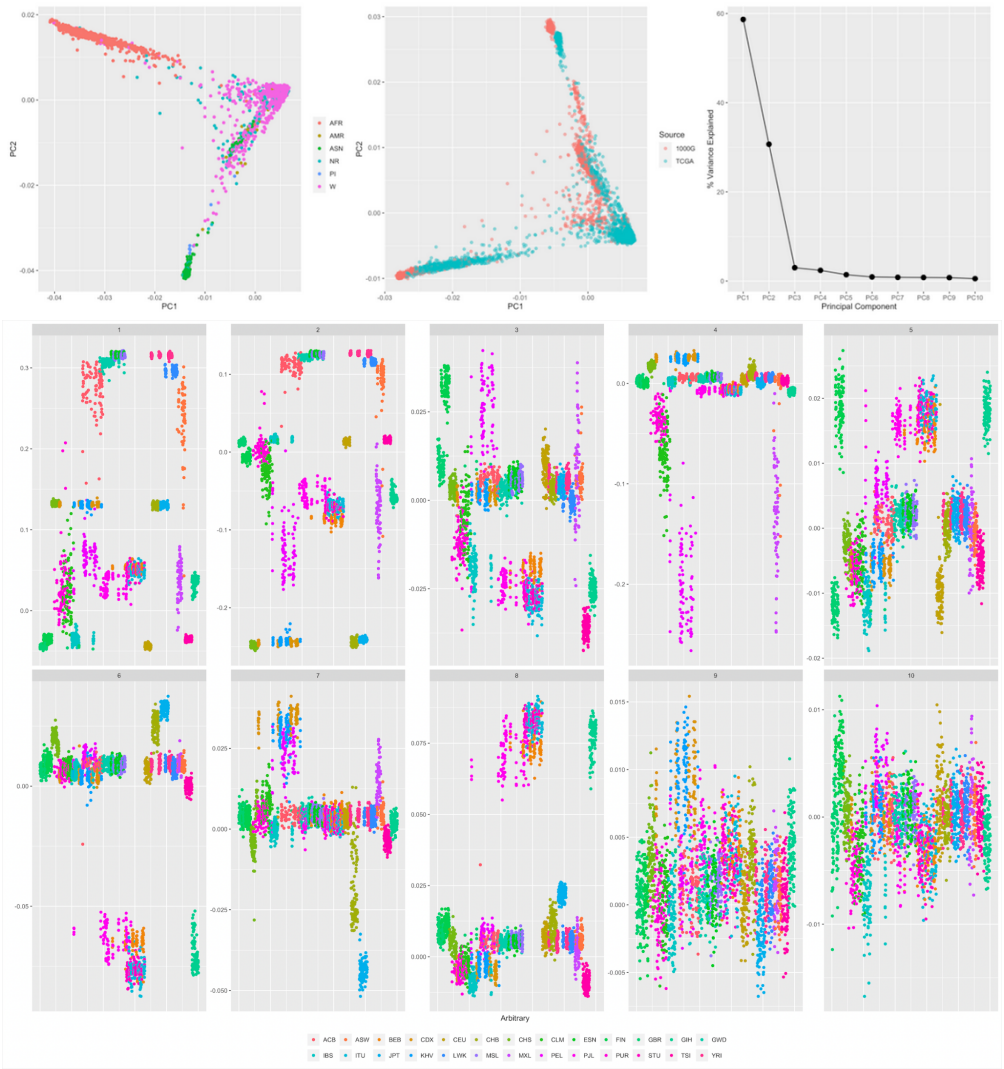
Appendix Fig. 6.1 Comparison of features of samples with concordant vs discordant WGD calls. Calls were made using MEDICC2 on copy number profiles from both ASCAT and ABSOLUTE and classified into 'both WGD' (WGD call from both copy number profiles on the same sample), 'neither WGD', WGD call only for the ASCAT copy number profile ('ASCAT only'), and WGD call only for the ABSOLUTE copy number profile ('ABSOLUTE only'). Samples with a call of 'Both WGD' or 'Neither WGD' were classified as concordant. **(a)** Copy number heterogeneity score from van Dijk et al (2021) **(b)** Subclonal genome fraction computed by ABSOLUTE **(c)** Tumour purity based on consensus purity estimate from Aran et al (2015) **(d)** Segment counts from ASCAT **(e)** Segment counts from ABSOLUTE **(f)** Proportion of concordant samples by cancer type **(g)** Proportion of concordant samples in a cancer type vs WGD rate (computed using MEDICC2 on concordant samples) **(h)** Ploidy computed by ASCAT **(i)** Ploidy computed by ABSOLUTE **(j)** Loss of heterozygosity computed by ASCAT **(k)** Loss of heterozygosity computed by ABSOLUTE.



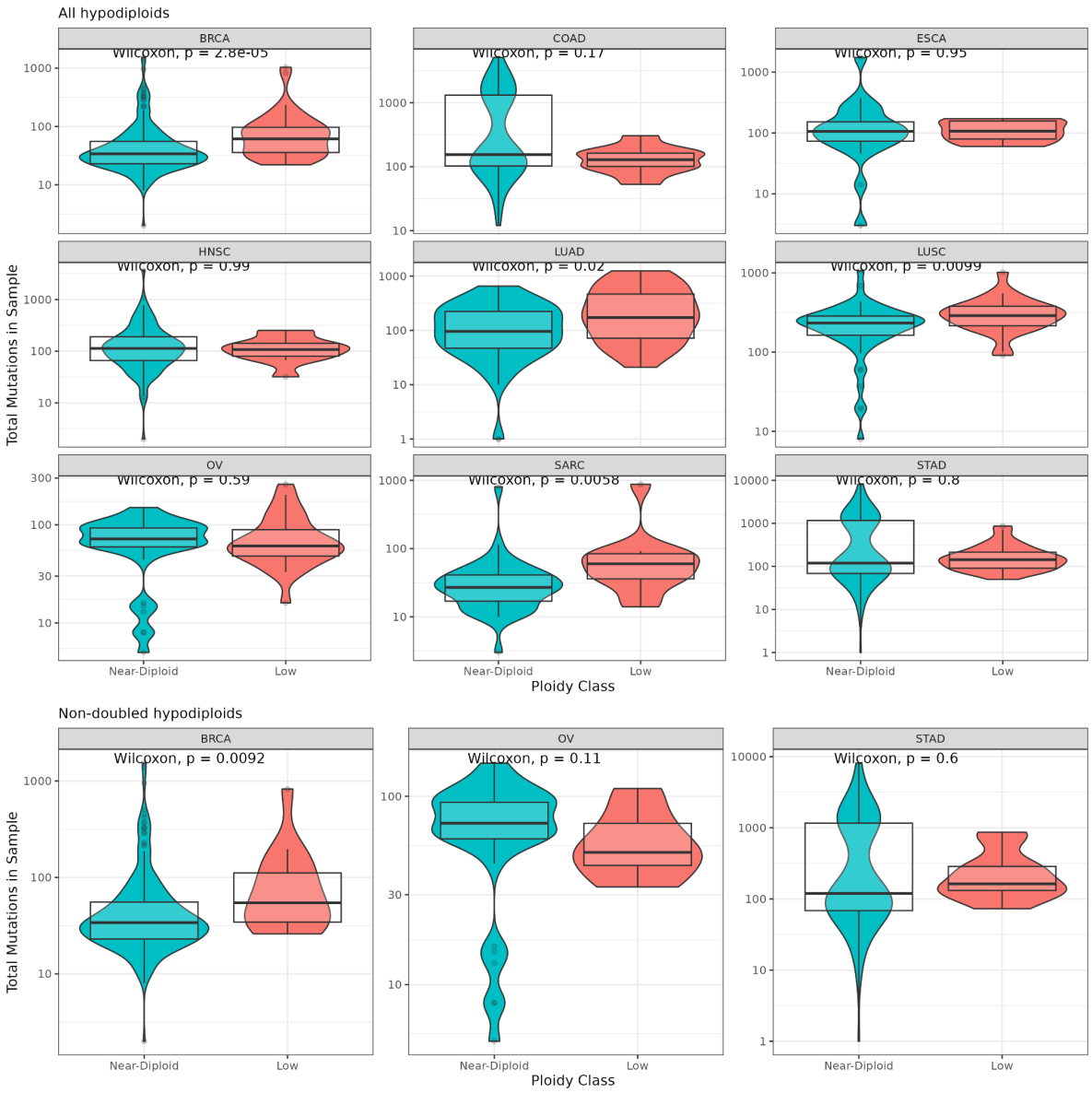
Appendix Fig. 6.2 Tissue-specific comparisons of proliferative index between near-diploid and polyploid samples. Only cancer types with at least 10 samples in each of the two ploidy categories were included.



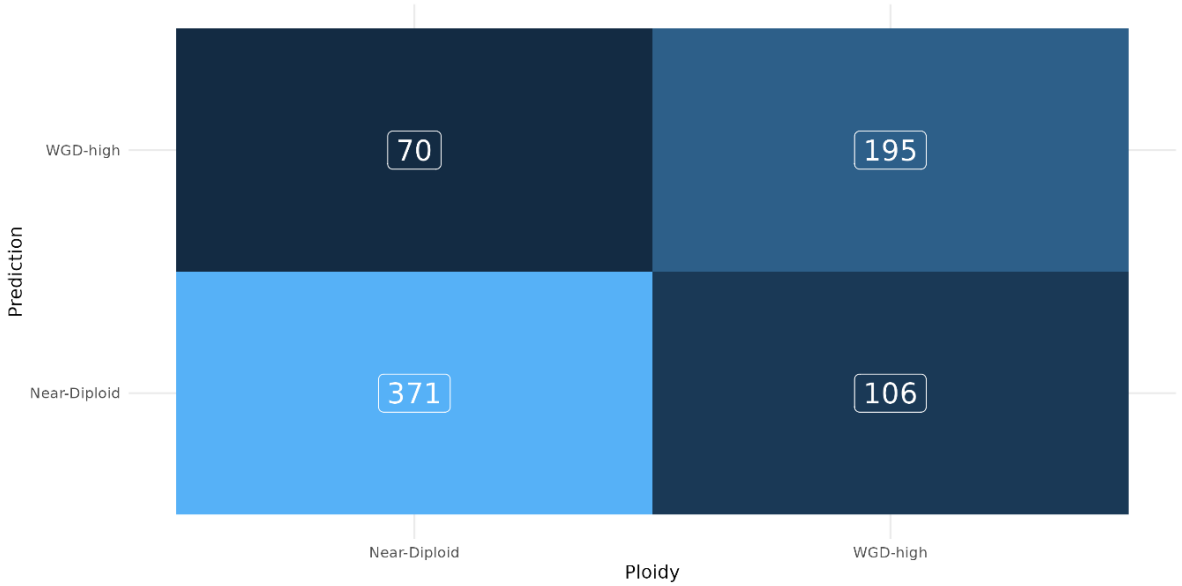
Appendix Fig. 6.3 Tissue-specific comparisons of proliferative index between near-diploid and low-hypodiploid samples. Only cancer types with at least 10 samples in each of the two ploidy categories were included.

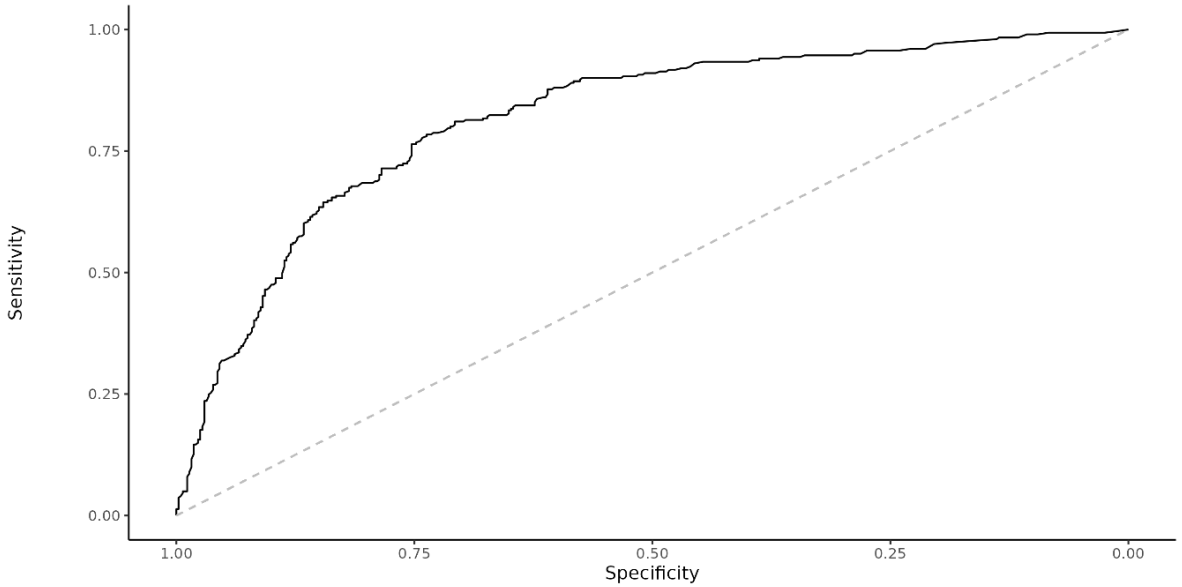


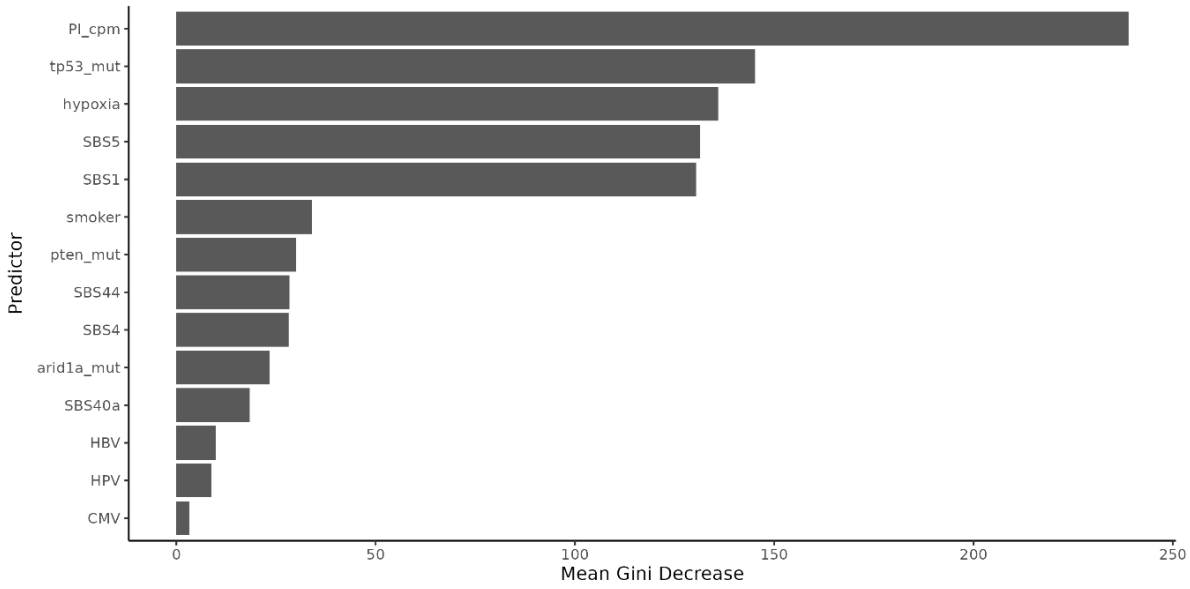
Appendix Fig. 6.4 Principal component analysis to identify ancestry covariates for the GWAS. **(top left)** the first two principal components of the TCGA genotype data, coloured by patient's self-reported race **(top, centre)** the first two principal components of the joint PCA-1000 Genomes Project genotype data, coloured by data source **(top right)** eigenvalues of (genotypic variance explained by) the first ten principal components **(bottom)** projection of population labels from the 1000 Genomes Project onto principal component weights derived from the TCGA data, coloured by population label.



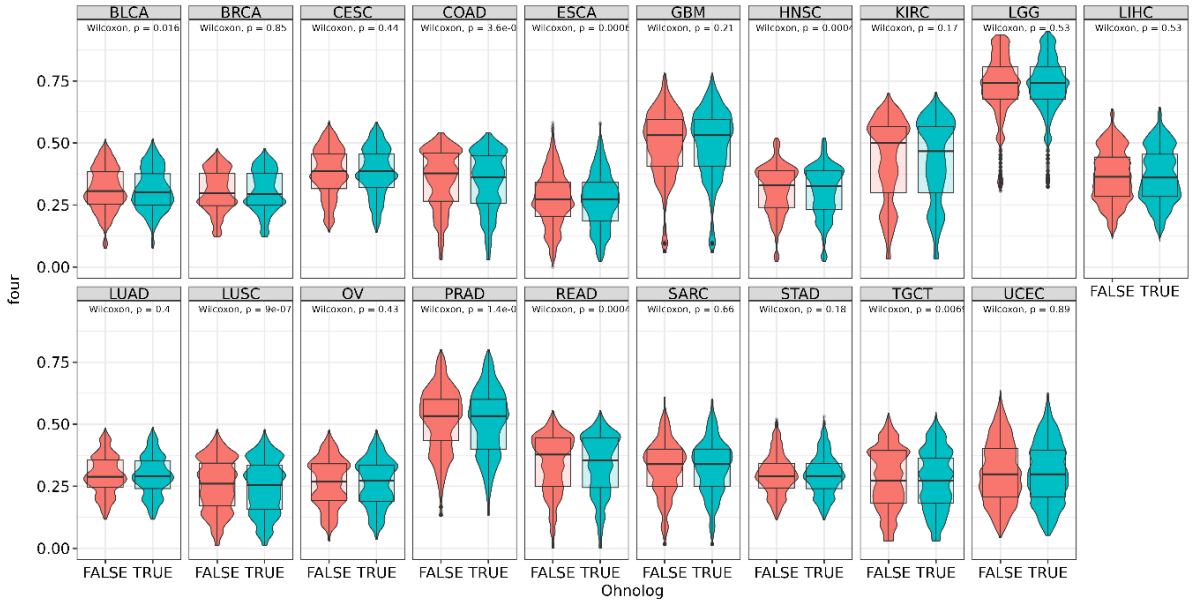
Appendix Fig. 6.5 Tissue-specific comparison of total mutation count between (top) near-diploids and all hypodiploids and **(bottom)** near-diploids and non-doubled hypodiploids. Only cancer types with at least ten samples in each of the relevant conditions are shown.

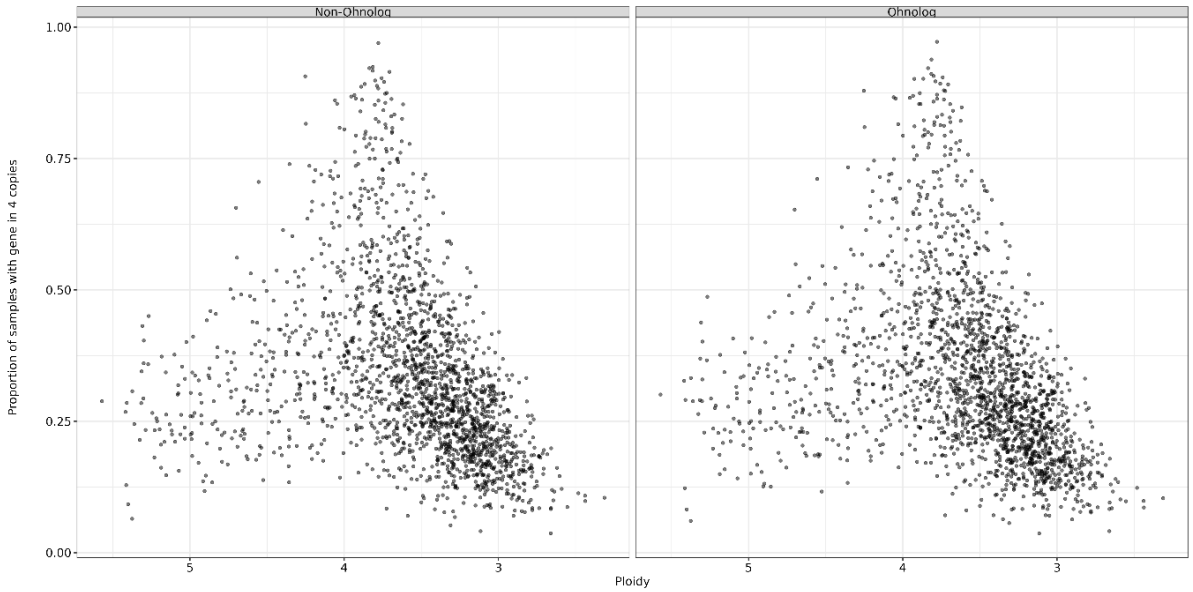

70
195
371
106
WGD-high
Near-Diploid
Prediction
Near-Diploid
WGD-high
Ploidy

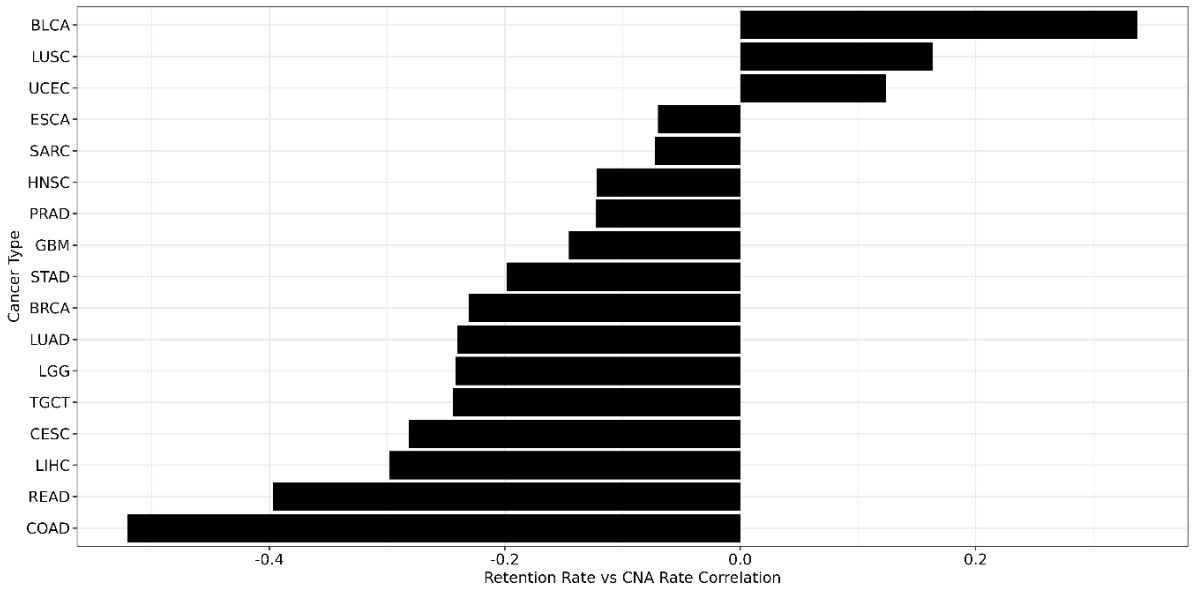

1.00
0.75
0.50
0.25
0.00
Sensitivity
1.00
0.75
0.50
0.25
0.00
Specificity


PI_cpm
tp53_mut
hypoxia
SBS5
SBS1
smoker
pten_mut
SBS44
SBS4
arid1a_mut
SBS40a
HBV
HPV
CMV
Predictor
0
50
100
150
200
250
Mean Gini Decrease

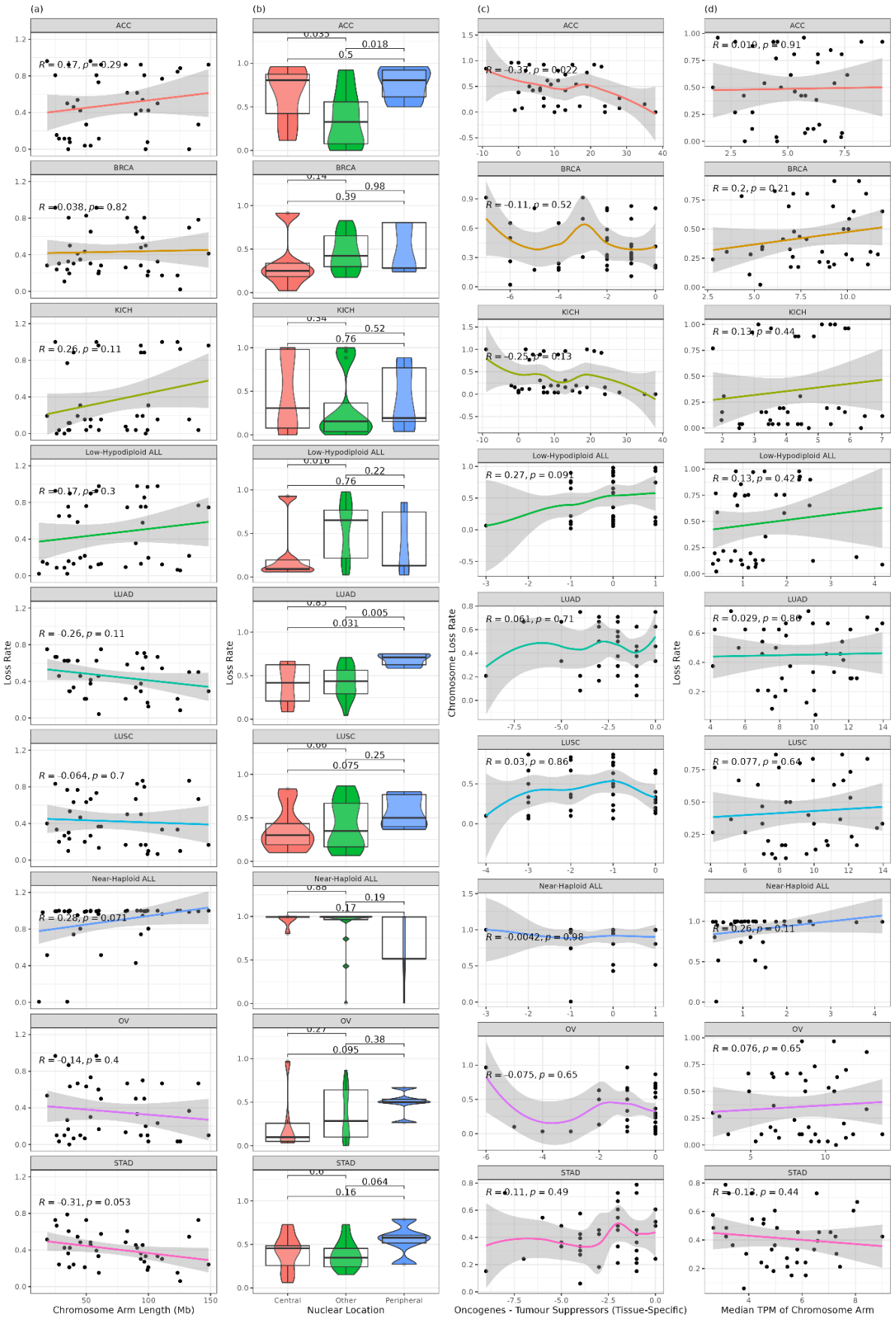
Appendix Fig. 6.6 Performance of random forest model after replacing tissue median proliferative index with individual sample proliferative index (top) confusion matrix, with observed ploidy status on the x-axis and the model's prediction on the y-axis **(centre)** ROC curve showing the relationship between model sensitivity and false positive rate **(bottom)** importance of each feature to the model's performance, as measured by Mean Gini Decrease.


BLCA
BRCA
CESC
COAD
ESCA
GBM
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PRAD
READ
SARC
STAD
TGCT
UCEC
four
0.00
0.25
0.50
0.75
FALSE TRUE
Ohnolog

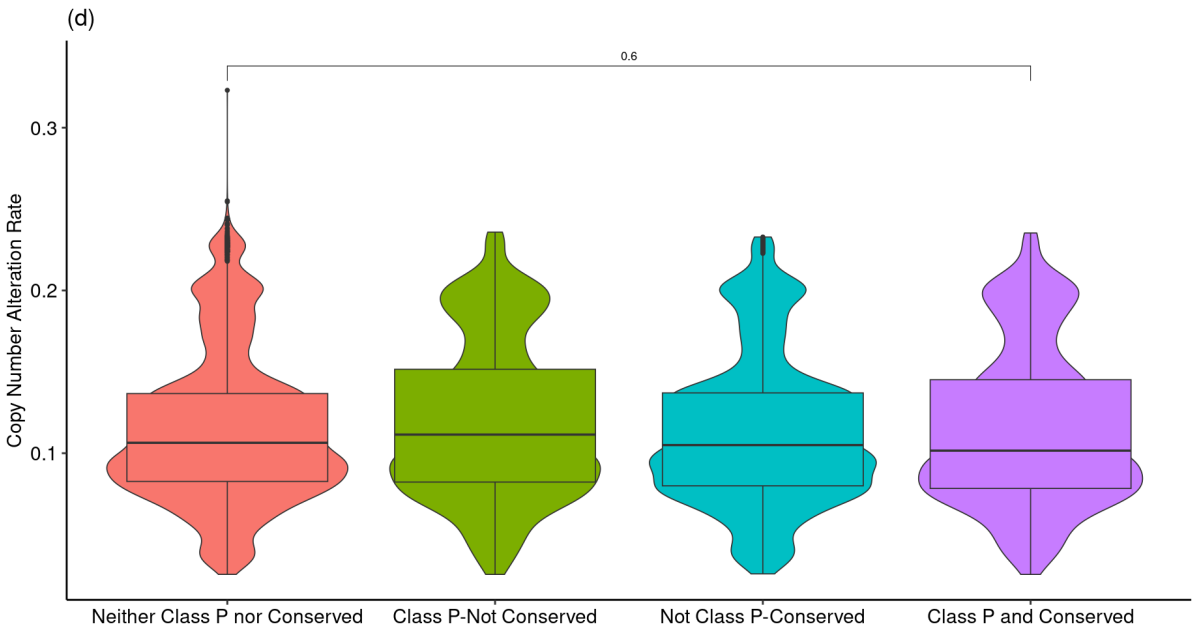

Non-Ohnolog
Ohnolog
Proportion of samples with gene in 4 copies
0.00
0.25
0.50
0.75
1.00
5
4
3
Ploidy


Cancer Type
BLCA
LUSC
UCEC
ESCA
SARC
HNSC
PRAD
GBM
STAD
BRCA
LUAD
LGG
TGCT
CESC
LIHC
READ
COAD
-0.4
-0.2
0.0
0.2
Retention Rate vs CNA Rate Correlation

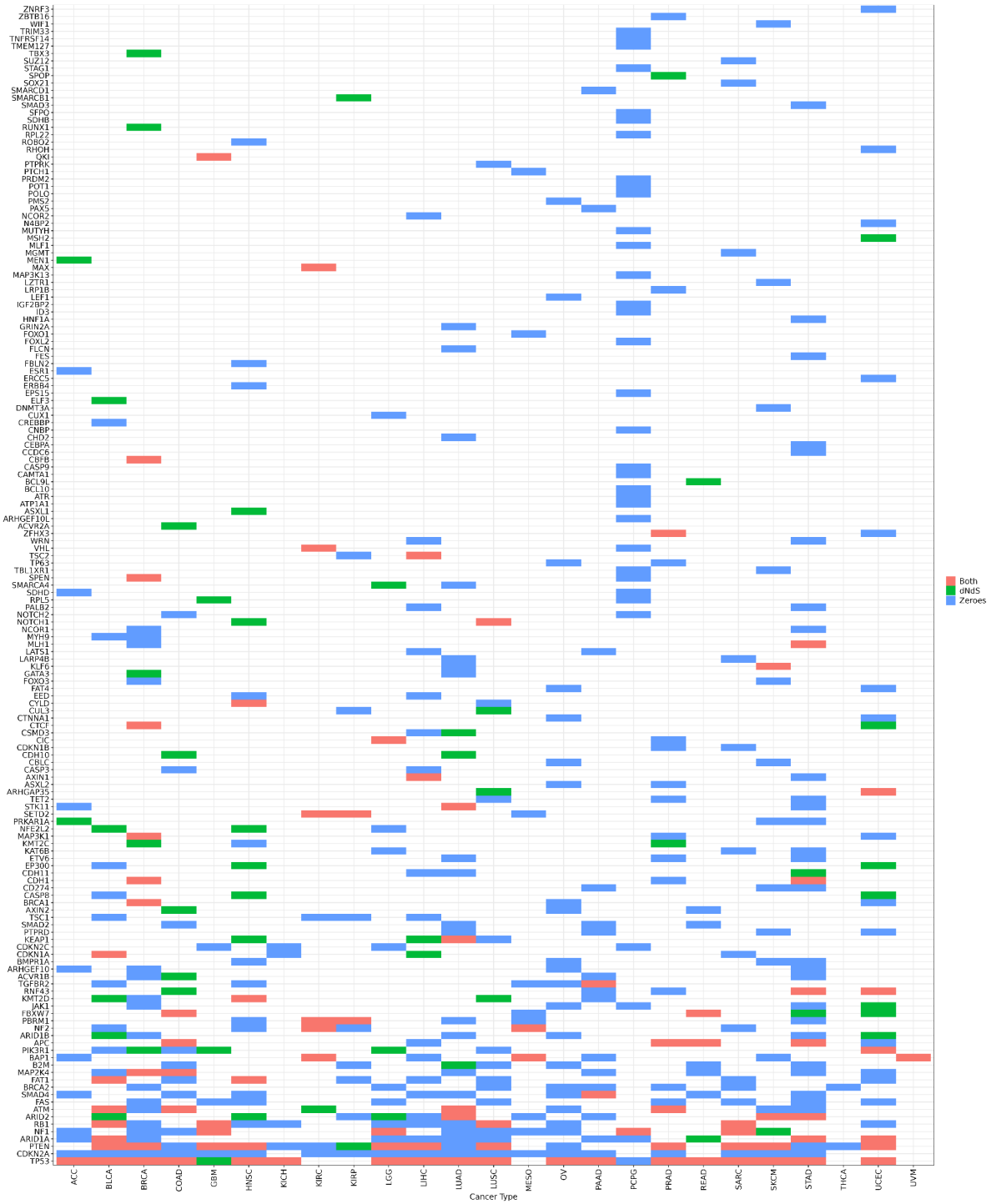
Appendix Fig. 6.7 Post-WGD gene losses in cancer vs human germline evolution (top) comparison between ohnologs and non-ohnologs in rate of gene retention in four copies after WGD **(centre)** loss trajectories of ohnologs and non-ohnologs with overall ploidy **(bottom)** searching for 'ohnolog-like' behaviour in cancer: tissue-specific correlations between rate of gene retention in four copies post-WGD and rate of any copy number alteration in the gene in near-diploid samples. The dosage balance hypothesis suggests that dosage-sensitive genes should not undergo copy number changes alone, but should remain duplicated after a whole genome duplication, which would predict a negative correlation between retention in duplicate after WGD and rate of copy number alterations in a diploid setting. An alternate explanation for a negative correlation is regional instability. Retention rate was adjusted for pre-WGD losses by removing those cases from the denominator.


(a)
(b)
(c)
(d)
ACC
BRCA
KICH
Low-Hypodiploid ALL
LUAD
LUSC
Near-Haploid ALL
OV
STAD
Loss Rate
Chromosome Loss Rate
Chromosome Arm Length (Mb)
Central
Other
Peripheral
Nuclear Location
Oncogenes - Tumour Suppressors (Tissue-Specific)
Median TPM of Chromosome Arm

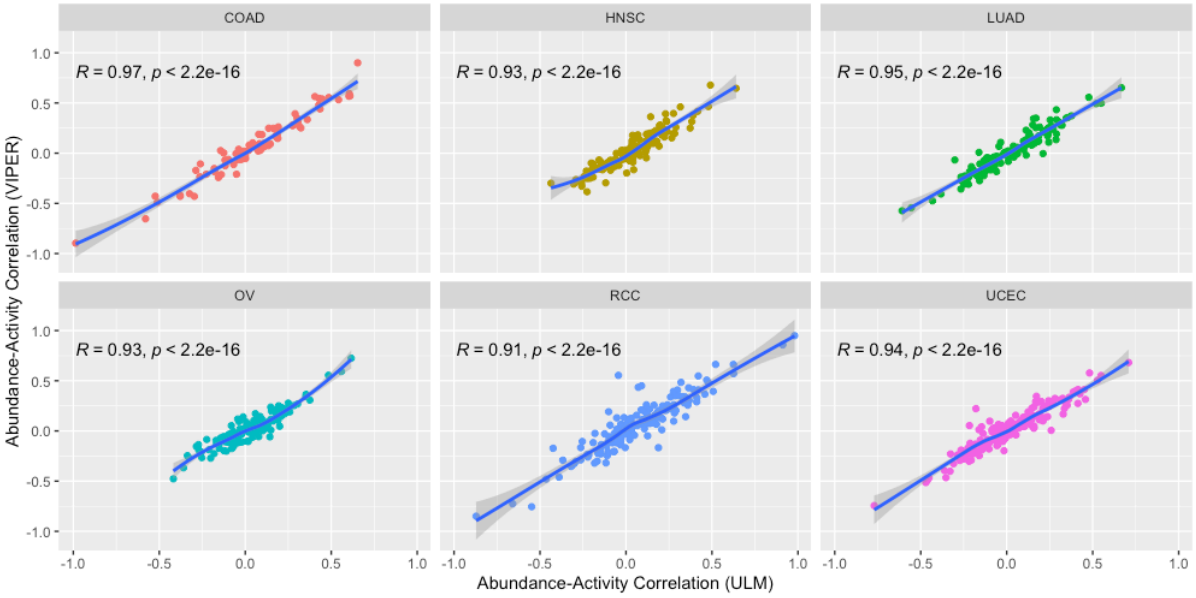
Appendix Fig. 6.8 Relationship between loss rate and chromosome features: (a) arm length **(b)** location in the nucleus (central/peripheral/other) **(c)** difference between number of oncogenes and number of tumour suppressor genes, using a dataset of tissue-specific driver genes from Davoli et al 2013 **(d)** median expression (TPM) of all genes on the chromosome arm in the relevant healthy tissue from GTEx.



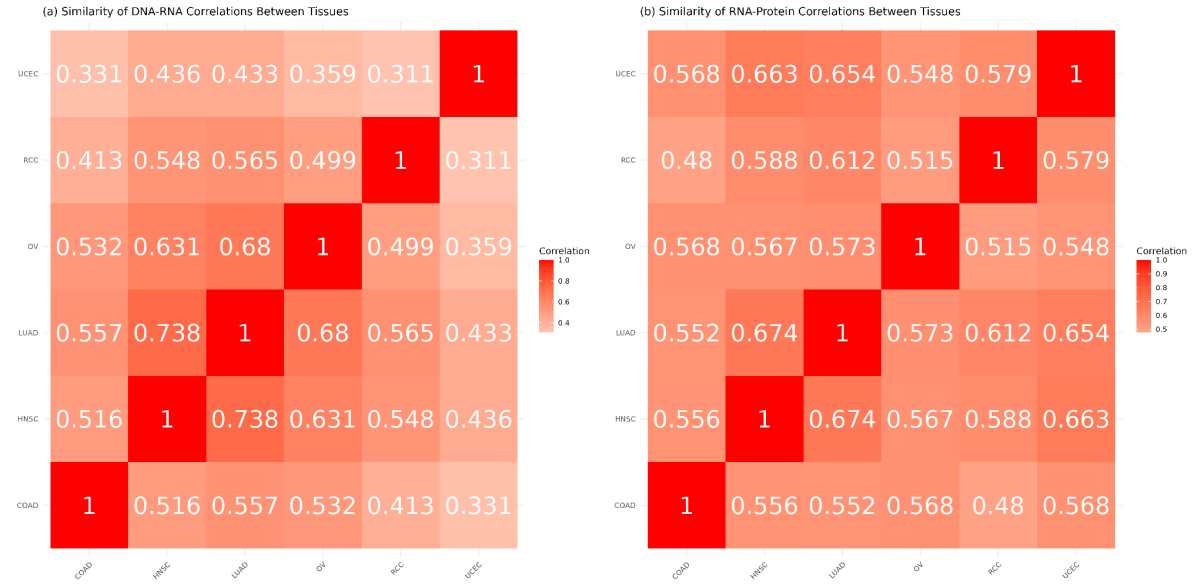
Appendix Fig. 6.9 Copy number alteration rate of genes by germline dosage sensitivity classification. Copy number alteration rate was calculated as the proportion of cases in which the gene was not in two copies in near-diploid tumours. In a tissue-specific analysis, copy-constrained genes ('Class P and Conserved') had more CNAs than non-copy constrained genes ('Neither Class P nor Conserved') in 6 cancer types, fewer in 14 cancer types, and no significant difference (Wilcoxon two-sided p >= 0.05) in 13 cancer types.



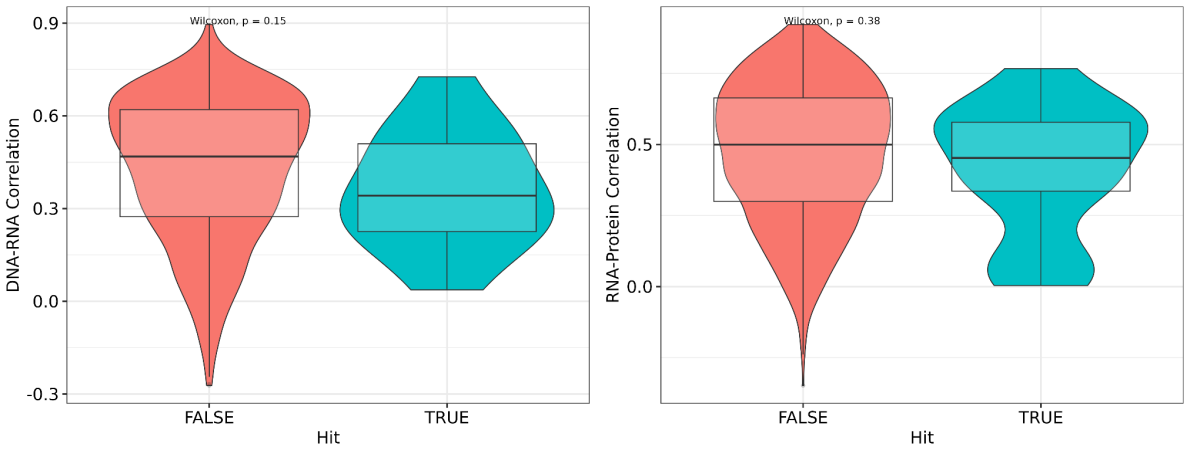
Appendix Fig. 6.10 Relevant driver TSGs (y-axis) in each cancer type (x-axis), defined by either a significant dndscv value after multiple testing correction (global q-value < 0.05), green, a frequency of complete inactivation >= 1%, blue, or both, pink.



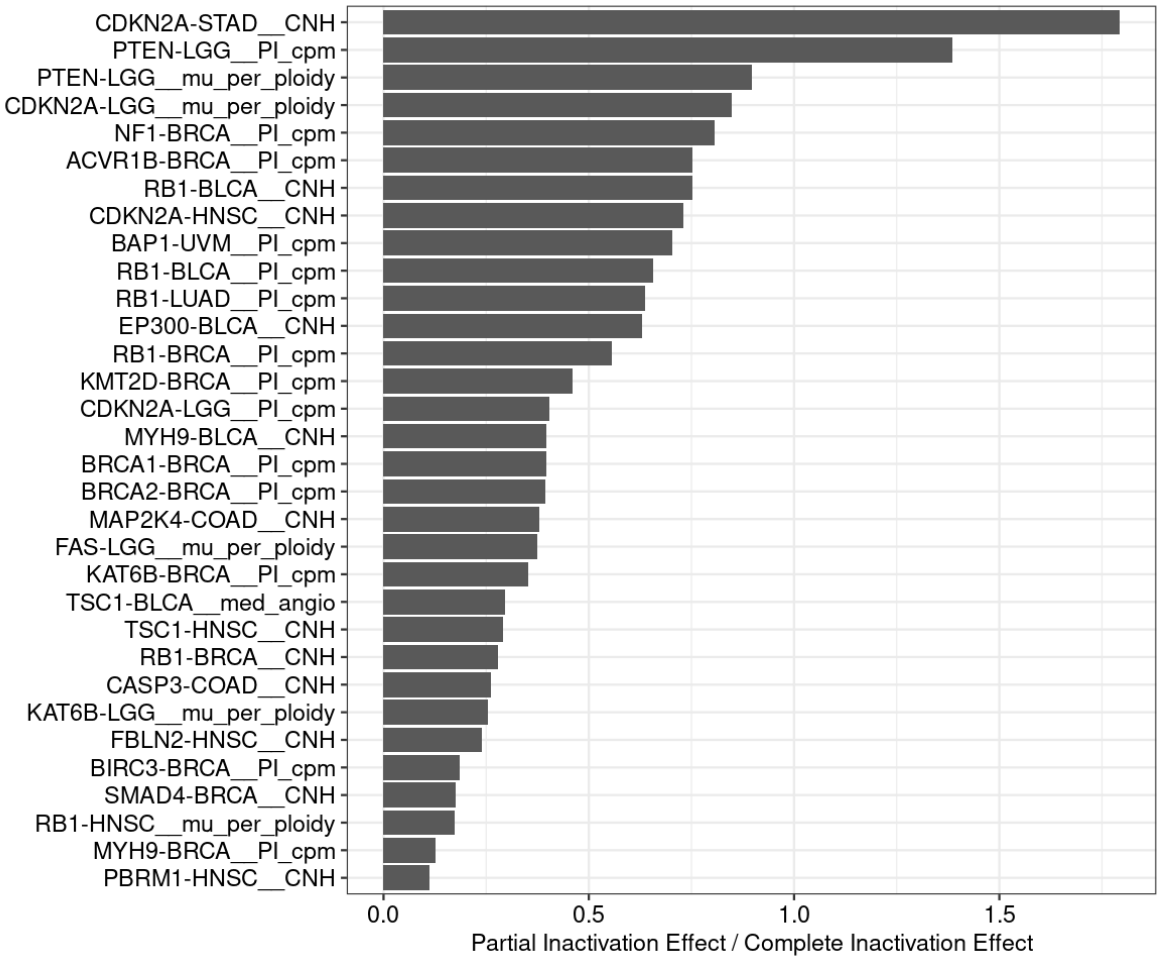
Appendix Fig. 6.11 Results of regulon enrichment analysis using different inference methods on the same regulons (VIPER and ULM)



Appendix Fig. 6.12 Similarity of (a) DNA-RNA and (b) RNA-protein correlations between tissues.



Appendix Fig. 6.13 DNA-RNA and RNA-protein correlations by hit status, limited to gene-cancer type combinations with at least 10 partially inactivated and 10 intact cases for the relevant TSG, and with hits filtered to genes independently reported as haploinsufficient in the literature.



Appendix Fig. 6.14 Effect size ratios for partial vs complete TSG inactivations. Results are based on linear regression controlling for age and sex. Ratios were calculated as partial inactivation regression estimate / complete

inactivation regression estimate. The y-axis shows gene-cancer type-hallmark combinations for which both partial and complete inactivation were associated with a significant increase in hallmark level. BH correction was performed considering both levels of inactivation together.

Bibliography

Abedini, S. S., Akhavantabasi, S., Liang, Y., Heng, J. I., Alizadehsani, R., Dehzangi, I., Bauer, D. C., & Alinejad-Rokny, H. (2024). A critical review of the impact of candidate copy number variants on autism spectrum disorder. *Mutation Research. Reviews in Mutation Research*, *794*. https://doi.org/10.1016/j.mrrev.2024.108509

Adibi, R., Moein, S., & Gheisari, Y. (2023). Cisplatin-Resistant Ovarian Cancer Cells Reveal a Polyploid Phenotype with Remarkable Activation of Nuclear Processes. *Advanced Biomedical Research*, *12*. https://doi.org/10.4103/abr.abr_348_21

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421.

Alimonti, A., Carracedo, A., Clohessy, J. G., Trotman, L. C., Nardella, C., Egia, A., Salmena, L., Sampieri, K., Haveman, W. J., Brogi, E., Richardson, A. L., Zhang, J., & Pandolfi, P. P. (2010). Subtle variations in Pten dose determine cancer susceptibility. *Nature Genetics*, *42*(5), 454–458.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural

variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363.

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Belinda Ding, B., Hilda Ye, B., & Califano, A. (2016). Network-based inference of protein activity helps functionalize the genetic landscape of cancer. *Nature Genetics*, *48*(8), 838.

Alves, J. M., Prieto, T., & Posada, D. (2017). Multiregional Tumor Trees Are Not Phylogenies. *Trends in Cancer*, *3*(8). https://doi.org/10.1016/j.trecan.2017.06.004

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., & Maley, C. C. (2015). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, *22*(1), 105–113.

Andrianova, M. A., Seplyarskiy, V. B., Terradas, M., Sánchez-Heras, A. B., Mur, P., Soto, J. L., Aiza, G., Borràs, E., Kondrashov, F. A., Kondrashov, A. S., Bazykin, G. A., & Valle, L. (2024). Discovery of recessive effect of human polymerase δ proofreading deficiency through mutational analysis of POLD1-mutated normal and cancer cells. *European Journal of Human Genetics : EJHG*, *32*(7), 837–845.

Arad, G., & Geiger, T. (2023). Functional Impact of Protein-RNA Variation in Clinical Cancer Analyses. *Molecular & Cellular Proteomics : MCP*, *22*(7), 100587.

Auwerx, C., Jõeloo, M., Sadler, M. C., Tesio, N., Ojavee, S., Clark, C. J., Mägi, R., Reymond, A., & Kutalik, Z. (2024). Rare copy-number variants as modulators of common disease susceptibility. *Genome Medicine*, *16*(1).

https://doi.org/10.1186/s13073-023-01265-5

Bach, D.-H., Park, H. J., & Lee, S. K. (2018). The Dual Role of Bone Morphogenetic Proteins in Cancer. *Molecular Therapy Oncolytics*, *8*, 1–13.

Bader, D. M., Wilkening, S., Lin, G., Tekkedil, M. M., Dietrich, K., Steinmetz, L. M., & Gagneur, J. (2015). Negative feedback buffers effects of regulatory variants. *Molecular Systems Biology*, *11*(1), 785.

Badia-I-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., & Saez-Rodriguez, J. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances*, *2*(1), vbac016.

Baghaarabani, L., Goliaei, S., Foroughmand-Araabi, M. H., Shariatpanahi, S. P., & Goliaei, B. (2021). Conifer: clonal tree inference for tumor heterogeneity with single-cell and bulk sequencing data. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04338-7

Bailey, C., Black, J. R. M., Reading, J. L., Litchfield, K., Turajlic, S., McGranahan, N., Jamal-Hanjani, M., & Swanton, C. (2021). Tracking Cancer Evolution through the Disease Course. *Cancer Discovery*, *11*(4). https://doi.org/10.1158/2159-8290.CD-20-1559

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Ding, L. (2018). Comprehensive Characterization of

Cancer Driver Genes and Mutations. *Cell*, *173*(2). https://doi.org/10.1016/j.cell.2018.02.060

Baker, T. G., Alden, J., Dubuc, A. M., Welsh, C. T., Znoyko, I., Cooley, L. D., Farooqi, M. S., Schwartz, S., Li, Y. Y., Cherniack, A. D., Lindhorst, S. M., Gener, M., Wolff, D. J., & Meredith, D. M. (2020). Near haploidization is a genomic hallmark which defines a molecular subgroup of giant cell glioblastoma. *Neuro-Oncology Advances*, *2*(1), vdaa155.

Baker, T. M., Lai, S., Lynch, A. R., Lesluyes, T., Yan, H., Ogilvie, H. A., Verfaillie, A., Dentro, S., Bowes, A. L., Pillay, N., Flanagan, A. M., Swanton, C., Spellman, P. T., Tarabichi, M., & Van Loo, P. (2024). The History of Chromosomal Instability in Genome-Doubled Tumors. *Cancer Discovery*, *14*(10), 1810–1822.

Bakhoum, 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., Duran, M., Pauli, C., Shaw, C., Chadalavada, K., Rajasekhar, V. K., Genovese, G., Venkatesan, S., Birkbak, N. J., McGranahan, N., … Cantley, L. C. (2018). Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature*, *553*(7689). https://doi.org/10.1038/nature25432

Balasubramanian, A., John, T., & Asselin-Labat, M. L. (2022). Regulation of the antigen presentation machinery in cancer and its implication for immune surveillance. *Biochemical Society Transactions*, *50*(2). https://doi.org/10.1042/BST20210961

Barra, V., & Fachinetti, D. (2018). The dark side of centromeres: types, causes and consequences of structural abnormalities implicating

centromeric DNA. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06545-y

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607.

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B (2025). *data.table: Extension of 'data.frame'*. R package version 1.16.99, https://Rdatatable.gitlab.io/data.table, https://github.com/Rdatatable/data.table, https://r-datatable.com.

Baruah, P., Lee, M., Wilson, P. O. G., Odutoye, T., Williamson, P., Hyde, N., Kaski, J. C., & Dumitriu, I. E. (2015). Impact of p16 status on pro- and anti-angiogenesis factors in head and neck cancers. *British Journal of Cancer*, *113*(4), 653–659.

Baslan, T., Morris, J. P., Zhao, Z., Reyes, J., Ho, Y.-J., Tsanov, K. M., Bermeo, J., Tian, S., Zhang, S., Askan, G., Yavas, A., Lecomte, N., Erakky, A., Varghese, A. M., Zhang, A., Kendall, J., Ghiban, E., Chorbadjiev, L., Wu, J., … Lowe, S. W. (2022). Ordered and deterministic cancer genome evolution after p53 loss. *Nature*, *608*(7924), 795–802.

Batra, R. N., Lifshitz, A., Vidakovic, A. T., Chin, S.-F., Sati-Batra, A., Sammut, S.-J., Provenzano, E., Ali, H. R., Dariush, A., Bruna, A., Murphy, L., Purushotham, A., Ellis, I., Green, A., Garrett-Bakelman, F.

E., Mason, C., Melnick, A., Aparicio, S. A. J. R., Rueda, O. M., … Caldas, C. (2021). DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. *Nature Communications*, *12*(1), 5406.

Berger, A. H., Knudson, A. G., & Pandolfi, P. P. (2011). A continuum model for tumour suppression. *Nature*, *476*(7359), 163–169.

Berl, A., Shir-az, O., Genish, I., Biran, H., Mann, D., Singh, A., Wise, J., Kravtsov, V., Kidron, D., Golberg, A., Vitkin, E., Yakhini, Z., & Shalom, A. (2023). Exploring multisite heterogeneity of human basal cell carcinoma proteome and transcriptome. *PLOS ONE*, *18*(11), e0293744.

Besedina, E., & Supek, F. (2024). Copy number losses of oncogenes and gains of tumor suppressor genes generate common driver mutations. *Nature Communications*, *15*(1), 6139.

Bewicke-Copley, F., Arjun, K. E., Palladino, G., Korfi, K., & Wang, J. (2019). Applications and analysis of targeted genomic sequencing in cancer studies. *Computational and Structural Biotechnology Journal*, *17*. https://doi.org/10.1016/j.csbj.2019.10.004

Bhandari, V., Hoey, C., Liu, L. Y., Lalonde, E., Ray, J., Livingstone, J., Lesurf, R., Shiah, Y.-J., Vujcic, T., Huang, X., Espiritu, S. M. G., Heisler, L. E., Yousif, F., Huang, V., Yamaguchi, T. N., Yao, C. Q., Sabelnykova, V. Y., Fraser, M., Chua, M. L. K., … Bristow, R. G. (2019). Molecular landmarks of tumor hypoxia across cancer types. *Nature Genetics*, *51*(2), 308–318.

Bianchi, J. J., Zhao, X., Mays, J. C., & Davoli, T. (2020). Not all cancers

are created equal: Tissue specificity in cancer genes and pathways. *Current Opinion in Cell Biology*, *63*, 135.

Bielski, C. M., Donoghue, M. T. A., Gadiya, M., Hanrahan, A. J., Won, H. H., Chang, M. T., Jonsson, P., Penson, A. V., Gorelick, A., Harris, C., Schram, A. M., Syed, A., Zehir, A., Chapman, P. B., Hyman, D. M., Solit, D. B., Shannon, K., Chandarlapaty, S., Berger, M. F., & Taylor, B. S. (2018). Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell*, *34*(5), 852–862.e4.

Bielski, C. M., Zehir, A., Penson, A. V., Donoghue, M. T. A., Chatila, W., Armenia, J., Chang, M. T., Schram, A. M., Jonsson, P., Bandlamudi, C., Razavi, P., Iyer, G., Robson, M. E., Stadler, Z. K., Schultz, N., Baselga, J., Solit, D. B., Hyman, D. M., Berger, M. F., & Taylor, B. S. (2018). Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature Genetics*, *50*(8), 1189–1195.

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. (2011). Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Cancer Research*, *71*(10), 3447–3452.

Bloomfield, M., Chen, J., & Cimini, D. (2021). Spindle Architectural Features Must Be Considered Along With Cell Size to Explain the Timing of Mitotic Checkpoint Silencing. *Frontiers in Physiology*, *11*. https://doi.org/10.3389/fphys.2020.596263

Boast, B., Goel, S., González-Granado, L. I., Niemela, J., Stoddard, J., Edwards, E. S. J., Seneviratne, S., Spensberger, D., Quesada-Espinosa, J.

F., Allende, L. M., McDonnell, J., Haseley, A., Lesmana, H., Walkiewicz, M. A., Muhammad, E., Bosco, J. J., Fleisher, T. A., Cohen, S., Holland, S. M., … Rosenzweig, S. D. (2023). TCF3 haploinsufficiency defined by immune, clinical, gene-dosage, and murine studies. *The Journal of Allergy and Clinical Immunology*, *152*(3), 736–747.

Boisselier, B., Dugay, F., Belaud-Rotureau, M.-A., Coutolleau, A., Garcion, E., Menei, P., Guardiola, P., & Rousseau, A. (2018). Whole genome duplication is an early event leading to aneuploidy in -wild type glioblastoma. *Oncotarget*, *9*(89), 36017–36028.

Bolognesi, B., Lorenzo Gotor, N., Dhar, R., Cirillo, D., Baldrighi, M., Tartaglia, G. G., & Lehner, B. (2016). A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Reports*, *16*(1), 222–231.

Boot, A., Oosting, J., de Miranda, N. F., Zhang, Y., Corver, W. E., van de Water, B., Morreau, H., & van Wezel, T. (2016). Imprinted survival genes preclude loss of heterozygosity of chromosome 7 in cancer cells. *The Journal of Pathology*, *240*(1), 72–83.

Borah, N. A., & Reddy, M. M. (2021). Aurora Kinase B Inhibition: A Potential Therapeutic Strategy for Cancer. *Molecules (Basel, Switzerland)*, *26*(7). https://doi.org/10.3390/molecules26071981

Boutary, S., Echaniz-Laguna, A., Adams, D., Loisel-Duwattez, J., Schumacher, M., Massaad, C., & Massaad-Massade, L. (2021). Treating PMP22 gene duplication-related Charcot-Marie-Tooth disease: the past, the present and the future. *Translational Research : The Journal of*

Laboratory and Clinical Medicine, *227*, 100–111.

Bovée, J. V., van Royen, M., Bardoel, A. F., Rosenberg, C., Cornelisse, C. J., Cleton-Jansen, A. M., & Hogendoorn, P. C. (2000). Near-haploidy and subsequent polyploidization characterize the progression of peripheral chondrosarcoma. *The American Journal of Pathology*, *157*(5), 1587–1595.

Bravo-Estupiñan, D. M., Aguilar-Guerrero, K., Quirós, S., Acón, M.-S., Marín-Müller, C., Ibáñez-Hernández, M., & Mora-Rodríguez, R. A. (2023). Gene dosage compensation: Origins, criteria to identify compensated genes, and mechanisms including sensor loops as an emerging systems-level property in cancer. *Cancer Medicine*, *12*(24), 22130–22155.

Brown, D., Smeets, D., Székely, B., Larsimont, D., Szász, A. M., Adnet, P.-Y., Rothé, F., Rouas, G., Nagy, Z. I., Faragó, Z., Tőkés, A.-M., Dank, M., Szentmártoni, G., Udvarhelyi, N., Zoppoli, G., Pusztai, L., Piccart, M., Kulka, J., Lambrechts, D., … Desmedt, C. (2017). Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nature Communications*, *8*(1), 1–13.

Buffa, F. M., Harris, A. L., West, C. M., & Miller, C. J. (2010). Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British Journal of Cancer*, *102*(2). https://doi.org/10.1038/sj.bjc.6605450

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide

association studies. *Nature Genetics*, *47*(3), 291–295.

Burdett, N. L., Willis, M. O., Pandey, A., Twomey, L., Alaei, S., Bowtell, D. D. L., & Christie, E. L. (2024). Timing of whole genome duplication is associated with tumor-specific MHC-II depletion in serous ovarian cancer. *Nature Communications*, *15*(1), 1–13.

Burssed, B., Zamariolli, M., Bellucco, F. T., & Melaragno, M. I. (2022). Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*, *15*(1), 23.

Carracedo, A., Alimonti, A., & Pandolfi, P. P. (2011). PTEN level in tumor suppression: how much is too little? *Cancer Research*, *71*(3), 629–633.

Carroll, A. J., Shago, M., Mikhail, F. M., Raimondi, S. C., Hirsch, B. A., Loh, M. L., Raetz, E. A., Borowitz, M. J., Wood, B. L., Maloney, K. W., Mattano, L. A., Jr, Larsen, E. C., Gastier-Foster, J., Stonerock, E., Ell, D., Kahwash, S., Devidas, M., Harvey, R. C., Chen, I.-M. L., … Heerema, N. A. (2019). Masked hypodiploidy: Hypodiploid acute lymphoblastic leukemia (ALL) mimicking hyperdiploid ALL in children: A report from the Children's Oncology Group. *Cancer Genetics*, *238*, 62–68.

Carrot-Zhang, J., Chambwe, N., Damrauer, J. S., Knijnenburg, T. A., Robertson, A. G., Yau, C., Zhou, W., Berger, A. C., Huang, K.-L., Newberg, J. Y., Mashl, R. J., Romanel, A., Sayaman, R. W., Demichelis, F., Felau, I., Frampton, G. M., Han, S., Hoadley, K. A., Kemal, A., … Beroukhim, R. (2020). Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell*, *37*(5), 639–654.e6.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T.,

Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, *30*(5), 413–421.

Casuscelli, J., Weinhold, N., Gundem, G., Wang, L., Zabor, E. C., Drill, E., Wang, P. I., Nanjangud, G. J., Redzematovic, A., Nargund, A. M., Manley, B. J., Arcila, M. E., Donin, N. M., Cheville, J. C., Thompson, R. H., Pantuck, A. J., Russo, P., Cheng, E. H., Lee, W., … Hsieh, J. J. (2017). Genomic landscape and evolution of metastatic chromophobe renal cell carcinoma. *JCI Insight*, *2*(12). https://doi.org/10.1172/jci.insight.92688

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7.

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., & Alizadeh, A. A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods in Molecular Biology (Clifton, N.J.)*, *1711*, 243–259.

Chen, H., Lin, F., Xing, K., & He, X. (2015). The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nature Communications*, *6*, 6367.

Chen, M., Zhang, J., Berger, A. H., Diolombi, M. S., Ng, C., Fung, J., Bronson, R. T., Castillo-Martin, M., Thin, T. H., Cordon-Cardo, C., Plevin, R., & Pandolfi, P. P. (2019). Compound haploinsufficiency of Dok2 and Dusp4 promotes lung tumorigenesis. *The Journal of Clinical Investigation*, *129*(1), 215–222.

Chen, S., Lai, H., Zhao, J., Chen, B., Li, Y., Li, Y., Li, Q., Zheng, Q., Huang, S., & Zhu, X. (2021). The viral expression and immune status in human cancers and insights into novel biomarkers of immunotherapy. *BMC Cancer*, *21*, 1183.

Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z. Y., Won, H. H., Scott, S. N., Brannon, A. R., O'Reilly, C., Sadowska, J., Casanova, J., Yannes, A., Hechtman, J. F., Yao, J., Song, W., Ross, D. S., … Berger, M. F. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of Molecular Diagnostics : JMD*, *17*(3). https://doi.org/10.1016/j.jmoldx.2014.12.006

Cheng, P., Zhao, X., Katsnelson, L., Camacho-Hernandez, E. M., Mermerian, A., Mays, J. C., Lippman, S. M., Rosales-Alvarez, R. E., Moya, R., Shwetar, J., Grun, D., Fenyo, D., & Davoli, T. (2022). Proteogenomic analysis of cancer aneuploidy and normal tissues reveals divergent modes of gene regulation across cellular pathways. *eLife*, *11*. https://doi.org/10.7554/eLife.75227

Chroni, A., Miura, S., Hamilton, L., Vu, T., Gaffney, S. G., Aly, V., Karim, S., Sanderford, M., Townsend, J. P., & Kumar, S. (2022). Clone Phylogenetics Reveals Metastatic Tumor Migrations, Maps, and Models. *Cancers*, *14*(17). https://doi.org/10.3390/cancers14174326

Ciani, Y., Fedrizzi, T., Prandi, D., Lorenzin, F., Locallo, A., Gasperini, P.,

Franceschini, G. M., Benelli, M., Elemento, O., Fava, L. L., Inga, A., & Demichelis, F. (2022). Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. *Cell Systems*, *13*(2), 183–193.e7.

Cockrum, C., Kaneshiro, K. R., Rechtsteiner, A., Tabuchi, T. M., & Strome, S. (2020). A primer for generating and using transcriptome data and gene sets. *Development (Cambridge, England)*, *147*(24). https://doi.org/10.1242/dev.193854

Comprehensive Analysis of Ubiquitously Expressed Genes in Humans from A Data-driven Perspective. (2023). *Genomics, Proteomics & Bioinformatics*, *21*(1), 164–176.

Cortés-Ciriano, I., Lee, J. J.-K., Xi, R., Jain, D., Jung, Y. L., Yang, L., Gordenin, D., Klimczak, L. J., Zhang, C.-Z., Pellman, D. S., PCAWG Structural Variation Working Group, Park, P. J., & PCAWG Consortium. (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, *52*(3), 331–341.

Corver, W. E., Ruano, D., Weijers, K., den Hartog, W. C. E., van Nieuwenhuizen, M. P., de Miranda, N., van Eijk, R., Middeldorp, A., Jordanova, E. S., Oosting, J., Kapiteijn, E., Hovens, G., Smit, J., van Wezel, T., & Morreau, H. (2012). Genome haploidisation with chromosome 7 retention in oncocytic follicular thyroid carcinoma. *PloS One*, *7*(6), e38287.

Coschi, C. H., Ishak, C. A., Gallo, D., Marshall, A., Talluri, S., Wang, J., Cecchini, M. J., Martens, A. L., Percy, V., Welch, I., Boutros, P. C.,

Brown, G. W., & Dick, F. A. (2014). Haploinsufficiency of an RB-E2F1-Condensin II complex leads to aberrant replication and aneuploidy. *Cancer Discovery*, *4*(7), 840–853.

Creasey, T., Enshaei, A., Nebral, K., Schwab, C., Watts, K., Cuthbert, G., Vora, A., Moppett, J., Harrison, C. J., Fielding, A. K., Haas, O. A., & Moorman, A. V. (2021). Single nucleotide polymorphism array-based signature of low hypodiploidy in acute lymphoblastic leukemia. *Genes, Chromosomes & Cancer*, *60*(9), 604–615.

Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. (2013). *Cell*, *155*(4), 948–962.

Cunningham, C. E., MacAuley, M. J., Yadav, G., Vizeacoumar, F. S., Freywald, A., & Vizeacoumar, F. J. (2019). Targeting the CINful genome: Strategies to overcome tumor heterogeneity. *Progress in Biophysics and Molecular Biology*, *147*, 77–91.

Damaschke, N. A., Yang, B., Bhusari, S., Avilla, M., Zhong, W., Blute, M. L., Jr, Huang, W., & Jarrard, D. F. (2017). Loss of Gene Imprinting in Murine Prostate Promotes Widespread Neoplastic Growth. *Cancer Research*, *77*(19), 5236–5247.

Darmasaputra, G. S., van Rijnberk, L. M., & Galli, M. (2024). Functional consequences of somatic polyploidy in development. *Development (Cambridge, England)*, *151*(5). https://doi.org/10.1242/dev.202392

Davoli, T., Denchi, E. L., & de Lange, T. (2010). Persistent telomere damage induces bypass of mitosis and tetraploidy. *Cell*, *141*(1). https://doi.org/10.1016/j.cell.2010.01.031

de Cárcer, G., Venkateswaran, S. V., Salgueiro, L., El Bakkali, A., Somogyi, K., Rowald, K., Montañés, P., Sanclemente, M., Escobar, B., de Martino, A., McGranahan, N., Malumbres, M., & Sotillo, R. (2018). Plk1 overexpression induces chromosomal instability and suppresses tumor development. *Nature Communications*, *9*(1), 3012.

de Leeuw, C. A., Mooij, J. M., & Heskes, T. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, *11*(4), e1004219.

Delinassios, J. G., Hoffman, R. M., Koumakis, G., Palitskaris, D., Poulatsidou, K.-N., & Delinasios, G. J. (2024). Sub-toxic cisplatin concentrations induce extensive chromosomal, nuclear and nucleolar abnormalities associated with high malignancy before acquired resistance develops: Implications for clinical caution. *PloS One*, *19*(12), e0311976.

Demeulemeester, J., Dentro, S. C., Gerstung, M., & Van Loo, P. (2022). Biallelic mutations in cancer genomes reveal local mutational determinants. *Nature Genetics*, *54*(2), 128–133.

Dentro, S. C., Leshchiner, I., Haase, K., Tarabichi, M., Wintersinger, J., Deshwar, A. G., Yu, K., Rubanova, Y., Macintyre, G., Demeulemeester, J., Vázquez-García, I., Kleinheinz, K., Livitz, D. G., Malikic, S., Donmez, N., Sengupta, S., Anur, P., Jolly, C., Cmero, M., … PCAWG Evolution and Heterogeneity Working Group and the PCAWG Consortium. (2021). Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, *184*(8), 2239–2254.e39.

Dong, D., Shen, H., Wang, Z., Liu, J., Li, Z., & Li, X. (2023). An RNA-informed dosage sensitivity map reflects the intrinsic functional nature of genes. *American Journal of Human Genetics*, *110*(9), 1509–1521.

Doroshow, D. B., & Doroshow, J. H. (2019). Genomics and the History of Precision Oncology. *Surgical Oncology Clinics of North America*, *29*(1), 35.

dos Santos, G. A., Chatsirisupachai, K., Avelar, R. A., & de Magalhães, J. P. (2023). Transcriptomic analysis reveals a tissue-specific loss of identity during ageing and cancer. *BMC Genomics*, *24*(1), 1–13.

Drews, R. M., Hernando, B., Tarabichi, M., Haase, K., Lesluyes, T., Smith, P. S., Morrill Gavarró, L., Couturier, D.-L., Liu, L., Schneider, M., Brenton, J. D., Van Loo, P., Macintyre, G., & Markowetz, F. (2022). A pan-cancer compendium of chromosomal instability. *Nature*, *606*(7916), 976–983.

Duijf, P. H. G., Schultz, N., & Benezra, R. (2013). Cancer cells preferentially lose small chromosomes. *International Journal of Cancer*, *132*(10), 2316–2326.

Ebadi, M., Bafort, Q., Mizrachi, E., Audenaert, P., Simoens, P., Van Montagu, M., Bonte, D., & Van de Peer, Y. (2023). The duplication of genomes and genetic networks and its potential for evolutionary adaptation and survival during environmental turmoil. *Proceedings of the National Academy of Sciences*, *120*(41), e2307289120.

edgeR. (n.d.). Bioconductor. Retrieved January 29, 2025, from http://bioconductor.org/packages/edgeR/

Eggermann, T., Perez de Nanclares, G., Maher, E. R., Temple, I. K., Tümer, Z., Monk, D., Mackay, D. J. G., Grønskov, K., Riccio, A., Linglart, A., & Netchine, I. (2015). Imprinting disorders: a group of congenital disorders with overlapping patterns of molecular changes affecting imprinted loci. *Clinical Epigenetics*, *7*(1), 1–18.

Ermakova, O., Piszczek, L., Luciani, L., Cavalli, F. M. G., Ferreira, T., Farley, D., Rizzo, S., Paolicelli, R. C., Al-Banchaabouchi, M., Nerlov, C., Moriggl, R., Luscombe, N. M., & Gross, C. (2011). Sensitized phenotypic screening identifies gene dosage sensitive region on chromosome 11 that predisposes to disease in mice. *EMBO Molecular Medicine*, *3*(1), 50–66.

Exploration of inhibitors targeting KIF18A with ploidy-specific lethality. (2024). *Drug Discovery Today*, *29*(10), 104142.

Fan, X., Liu, Y., Heilman, S. A., & Chen, J. J. (2013). Human Papillomavirus E7 Induces Rereplication in Response to DNA Damage. *Journal of Virology*, *87*(2), 1200.

Fares, M. A., Keane, O. M., Toft, C., Carretero-Paulet, L., & Jones, G. W. (2013). The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of Saccharomyces cerevisiae Genes. *PLOS Genetics*, *9*(1), e1003176.

Faucon, F., Dusfour, I., Gaude, T., Navratil, V., Boyer, F., Chandre, F., Sirisopa, P., Thanispong, K., Juntarajumnong, W., Poupardin, R., Chareonviriyaphap, T., Girod, R., Corbel, V., Reynaud, S., & David, J. P. (2015). Identifying genomic changes associated with insecticide resistance in the dengue mosquito Aedes aegypti by deep targeted

sequencing. *Genome Research*, *25*(9). https://doi.org/10.1101/gr.189225.115

Fearon, E. R., & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*(5). https://doi.org/10.1016/0092-8674(90)90186-i

Firke S (2025). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.2.

Gallardo-Vara, E., Blanco, F. J., Roqué, M., Friedman, S. L., Suzuki, T., Botella, L. M., & Bernabeu, C. (2016). Transcription factor KLF6 upregulates expression of metalloprotease MMP14 and subsequent release of soluble endoglin during vascular injury. *Angiogenesis*, *19*(2), 155–171.

Ganem, N. J., Storchova, Z., & Pellman, D. (2007). Tetraploidy, aneuploidy and cancer. *Current Opinion in Genetics & Development*, *17*(2), 157–162.

Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., Liu, Q., Ma, L., Wang, X., Zhou, J., Liu, Y., Boja, E., Robles, A. I., Ma, W., Wang, P., … Fan, J. (2019). Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell*, *179*(2). https://doi.org/10.1016/j.cell.2019.08.052

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8), 1363–1375.

GCTA: A Tool for Genome-wide Complex Trait Analysis. (2011). *The*

American Journal of Human Genetics, *88*(1), 76–82.

Gemble, S., Wardenaar, R., Keuper, K., Srivastava, N., Nano, M., Macé, A.-S., Tijhuis, A. E., Bernhard, S. V., Spierings, D. C. J., Simon, A., Goundiam, O., Hochegger, H., Piel, M., Foijer, F., Storchová, Z., & Basto, R. (2022). Genetic instability from a single S phase after whole-genome duplication. *Nature*, *604*(7904), 146–151.

Gene Copy-Number Alterations: A Cost-Benefit Analysis. (2013). *Cell*, *152*(3), 394–405.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, I., Haase, K., Jerman, L., Sengupta, S., Macintyre, G., … PCAWG Consortium. (2020). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128.

ggmanh. (n.d.). Bioconductor. Retrieved January 30, 2025, from http://bioconductor.org/packages/ggmanh/

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, *569*(7757). https://doi.org/10.1038/s41586-019-1186-3

Gialesaki, S., Bräuer-Hartmann, D., Issa, H., Bhayadia, R., Alejo-Valle, O., Verboon, L., Schmell, A.-L., Laszig, S., Regényi, E., Schuschel, K.,

Labuhn, M., Ng, M., Winkler, R., Ihling, C., Sinz, A., Glaß, M., Hüttelmaier, S., Matzk, S., Schmid, L., … Klusmann, J.-H. (2023). RUNX1 isoform disequilibrium promotes the development of trisomy 21-associated myeloid leukemia. *Blood*, *141*(10), 1105–1118.

Giorgi FM (2024). *aracne.networks: ARACNe-inferred gene networks from TCGA tumor datasets*. R package version 1.32.0.

Girish, V., Lakhani, A. A., Thompson, S. L., Scaduto, C. M., Brown, L. M., Hagenson, R. A., Sausville, E. L., Mendelson, B. E., Kandikuppa, P. K., Lukow, D. A., Yuan, M. L., Stevens, E. C., Lee, S. N., Schukken, K. M., Akalu, S. M., Vasudevan, A., Zou, C., Salovska, B., Li, W., … Sheltzer, J. M. (2023). Oncogene-like addiction to aneuploidy in human cancers. *Science (New York, N.Y.)*, *381*(6660). https://doi.org/10.1126/science.adg4521

Gliech, C. R., Yeow, Z. Y., Tapias-Gomez, D., Yang, Y., Huang, Z., Tijhuis, A. E., Spierings, D. C., Foijer, F., Chung, G., Tamayo, N., Bahrami-Nejad, Z., Collins, P., Nguyen, T. T., Plata Stapper, A., Hughes, P. E., Payton, M., & Holland, A. J. (2024). Weakened APC/C activity at mitotic exit drives cancer vulnerability to KIF18A inhibition. *The EMBO Journal*, *43*(5), 666–694.

Gopal, R. K., Kübler, K., Calvo, S. E., Polak, P., Livitz, D., Rosebrock, D., Sadow, P. M., Campbell, B., Donovan, S. E., Amin, S., Gigliotti, B. J., Grabarek, Z., Hess, J. M., Stewart, C., Braunstein, L. Z., Arndt, P. F., Mordecai, S., Shih, A. R., Chaves, F., … McFadden, D. G. (2018). Widespread Chromosomal Losses and Mitochondrial DNA Alterations as

Genetic Drivers in Hürthle Cell Carcinoma. *Cancer Cell*, *34*(2), 242–255.e5.

Gordeeva, V., Sharova, E., & Arapidi, G. (2022). Progress in Methods for Copy Number Variation Profiling. *International Journal of Molecular Sciences*, *23*(4), 2143.

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585.

Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, *1*(1), 1–17.

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)*, *32*(18), 2847–2849.

Gupta, T., Arun, S. R., Babu, G. A., Chakrabarty, B. K., Bhave, S. J., Kumar, J., Radhakrishnan, V., Krishnan, S., Ghara, N., Arora, N., Mishra, D. K., & Parihar, M. (2021). A Systematic Cytogenetic Strategy to Identify Masked Hypodiploidy in Precursor B Acute Lymphoblastic Leukemia in Low Resource Settings. *Indian Journal of Hematology & Blood Transfusion : An Official Journal of Indian Society of Hematology and Blood Transfusion*, *37*(4), 576–585.

Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., & Robertson, D. L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology*, *8*(10). https://doi.org/10.1186/gb-2007-8-10-r209

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.

Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H. N., Jung, H., Nam, S., Chung, M., Kim, J. H., & Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1). https://doi.org/10.1093/nar/gkx1013

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1). https://doi.org/10.1016/s0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674.

HannahVMeyer. (n.d.). *meyer-lab-cshl/plinkQC: plinkQC 0.3.2*. https://doi.org/10.5281/zenodo.3934294

Harrison, C. J., Moorman, A. V., Broadfield, Z. J., Cheung, K. L., Harris, R. L., Reza Jalali, G., Robinson, H. M., Barber, K. E., Richards, S. M., Mitchell, C. D., Eden, T. O. B., Hann, I. M., Hill, F. G. H., Kinsey, S. E., Gibson, B. E. S., Lilleyman, J., Vora, A., Goldstone, A. H., Franklin, I. M., … Childhood and Adult Leukaemia Working Parties. (2004). Three distinct subgroups of hypodiploidy in acute lymphoblastic leukaemia. *British Journal of Haematology*, *125*(5), 552–559.

He, Q., Au, B., Kulkarni, M., Shen, Y., Lim, K. J., Maimaiti, J., Wong, C. K., Luijten, M. N. H., Chong, H. C., Lim, E. H., Rancati, G., Sinha, I., Fu, Z., Wang, X., Connolly, J. E., & Crasta, K. C. (2018). Chromosomal instability-induced senescence potentiates cell non-autonomous tumourigenic effects. *Oncogenesis*, *7*(8), 1–18.

Herbein, G., & Nehme, Z. (2020). Polyploid Giant Cancer Cells, a

Hallmark of Oncoviruses and a New Therapeutic Challenge. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.567116

Hintzen, D. C., Soto, M., Schubert, M., Bakker, B., Spierings, D. C. J., Szuhai, K., Lansdorp, P. M., Kluin, R. J. C., Foijer, F., Medema, R. H., & Raaijmakers, J. A. (2022). The impact of monosomies, trisomies and segmental aneuploidies on chromosomal stability. *PloS One*, *17*(7), e0268579.

Hmeljak, J., Sanchez-Vega, F., Hoadley, K. A., Shih, J., Stewart, C., Heiman, D., Tarpey, P., Danilova, L., Drill, E., Gibb, E. A., Bowlby, R., Kanchi, R., Osmanbeyoglu, H. U., Sekido, Y., Takeshita, J., Newton, Y., Graim, K., Gupta, M., Gay, C. M., … Ladanyi, M. (2018). Integrative Molecular Characterization of Malignant Pleural Mesothelioma. *Cancer Discovery*, *8*(12), 1548–1565.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., … Laird, P. W. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, *173*(2), 291–304.e6.

Hobbs, B. P., Pestana, R. C., Zabor, E. C., Kaizer, A. M., & Hong, D. S. (2022). Basket Trials: Review of Current Practice and Innovations for Future Trials. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *40*(30). https://doi.org/10.1200/JCO.21.02285

Holland, D. O., & Johnson, M. E. (2018). Stoichiometric balance of protein copy numbers is measurable and functionally significant in a protein-protein interaction network for yeast endocytosis. *PLoS Computational Biology*, *14*(3), e1006022.

Holmfeldt, L., Wei, L., Diaz-Flores, E., Walsh, M., Zhang, J., Ding, L., Payne-Turner, D., Churchman, M., Andersson, A., Chen, S.-C., McCastlain, K., Becksfort, J., Ma, J., Wu, G., Patel, S. N., Heatley, S. L., Phillips, L. A., Song, G., Easton, J., … Mullighan, C. G. (2013). The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nature Genetics*, *45*(3), 242–252.

Horii, T., Yamamoto, M., Morita, S., Kimura, M., Nagao, Y., & Hatada, I. (2015). p53 suppresses tetraploid development in mice. *Scientific Reports*, *5*, 8907.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, *43*(Database issue), D512–D520.

Hosea, R., Hillary, S., Naqvi, S., Wu, S., & Kasim, V. (2024). The two sides of chromosomal instability: drivers and brakes in cancer. *Signal Transduction and Targeted Therapy*, *9*(1), 1–30.

Hu, L., Plafker, K., Vorozhko, V., Zuna, R. E., Hanigan, M. H., Gorbsky, G. J., Plafker, S. M., Angeletti, P. C., & Ceresa, B. P. (2009). Human papillomavirus 16 E5 induces bi-nucleated cell formation by cell-cell fusion. *Virology*, *384*(1), 125–134.

Huang, K.-L., Mashl, R. J., Wu, Y., Ritter, D. I., Wang, J., Oh, C.,

Paczkowska, M., Reynolds, S., Wyczalkowski, M. A., Oak, N., Scott, A. D., Krassowski, M., Cherniack, A. D., Houlahan, K. E., Jayasinghe, R., Wang, L.-B., Zhou, D. C., Liu, D., Cao, S., … Ding, L. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, *173*(2), 355–370.e14.

Huang, Z., Sun, S., Lee, M., Maslov, A. Y., Shi, M., Waldman, S., Marsh, A., Siddiqui, T., Dong, X., Peter, Y., Sadoughi, A., Shah, C., Ye, K., Spivack, S. D., & Vijg, J. (2022). Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nature Genetics*, *54*(4), 492.

Hunter, S., Hendrix, J., Freeman, J., Dowell, R. D., & Allen, M. A. (2023). Transcription dosage compensation does not occur in Down syndrome. *BMC Biology*, *21*(1), 228.

Hwang, S., Cavaliere, P., Li, R., Zhu, L. J., Dephoure, N., & Torres, E. M. (2021). Consequences of aneuploidy in human fibroblasts with trisomy 21. *Proceedings of the National Academy of Sciences*, *118*(6), e2014723118.

Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., & Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences*, *112*(48), 14918–14923.

Ippolito, M. R., Zerbib, J., Eliezer, Y., Reuveni, E., Viganò, S., De Feudis, G., Shulman, E. D., Savir, K. A., Slutsky, R., Chang, T., Campagnolo, E. M., Taglietti, S., Scorzoni, S., Gianotti, S., Martin, S., Muenzner, J., Mülleder, M., Rozenblum, N., Rubolino, C., … Ben-David, U. (2024).

Increased RNA and Protein Degradation Is Required for Counteracting Transcriptional Burden and Proteotoxic Stress in Human Aneuploid Cells. *Cancer Discovery*, *14*(12). https://doi.org/10.1158/2159-8290.CD-23-0309

Ishak, C. A., & Dick, F. A. (2014). Conditional haploinsufficiency of the retinoblastoma tumor suppressor gene. *Molecular & Cellular Oncology*, *2*(1). https://doi.org/10.4161/23723548.2014.968069

Ishikawa, K., Ishihara, A., & Moriya, H. (2020). Exploring the Complexity of Protein-Level Dosage Compensation that Fine-Tunes Stoichiometry of Multiprotein Complexes. *PLoS Genetics*, *16*(10), e1009091.

Islam, S. M. A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E. N., He, Y., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Steele, C. D., Otlu, B., … Alexandrov, L. B. (2022). Uncovering novel mutational signatures by extraction with SigProfilerExtractor. *Cell Genomics*, *2*(11), None.

Itahana, K., Campisi, J., & Dimri, G. P. (2007). Methods to detect biomarkers of cellular senescence: the senescence-associated beta-galactosidase assay. *Methods in Molecular Biology (Clifton, N.J.)*, *371*, 21–31.

Jesionek-Kupnicka, D., Szybka, M., Malachowska, B., Fendler, W., Potemski, P., Piaskowski, S., Jaskolski, D., Papierz, W., Skowronski, W., Och, W., Kordek, R., & Zawlik, I. (2014). TP53 promoter methylation in

primary glioblastoma: relationship with TP53 mRNA and protein expression and mutation status. *DNA and Cell Biology*, *33*(4), 217–226.

Jia, C.-W., Wang, L., Lan, Y.-L., Song, R., Zhou, L.-Y., Yu, L., Yang, Y., Liang, Y., Li, Y., Ma, Y.-M., & Wang, S.-Y. (2015). Aneuploidy in Early Miscarriage and its Related Factors. *Chinese Medical Journal*, *128*(20), 2772.

Jia, P., & Zhao, Z. (2019). Characterization of Tumor-Suppressor Gene Inactivation Events in 33 Cancer Types. *Cell Reports*, *26*(2), 496–506.e3.

Jia, Q., Chu, H., Jin, Z., Long, H., & Zhu, B. (2022). High-throughput single-cell sequencing in cancer research. *Signal Transduction and Targeted Therapy*, *7*(1), 1–20.

Jie, W., Bai, J., Yan, J., Chi, Y., & Li, B. B. (2021). Multi-Site Tumour Sampling Improves the Detection of Intra-Tumour Heterogeneity in Oral and Oropharyngeal Squamous Cell Carcinoma. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.670305

Jubran, J., Slutsky, R., Rozenblum, N., Rokach, L., Ben-David, U., & Yeger-Lotem, E. (2024). Machine-learning analysis reveals an important role for negative selection in shaping cancer aneuploidy landscapes. *Genome Biology*, *25*, 95.

Kaplan, A. R., & Glazer, P. M. (2020). Impact of hypoxia on DNA repair and genome integrity. *Mutagenesis*, *35*(1), 61–68.

Kassambara A, Kosinski M, Biecek P (2024). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.5

Kassambara A (2023). *ggpubr: 'ggplot2' Based Publication Ready Plots*.

R package version 0.6.0, https://rpkgs.datanovia.com/ggpubr/.

Kato, H., Kimura, H., Kushima, I., Takahashi, N., Aleksic, B., & Ozaki, N. (2023). The genetic architecture of schizophrenia: review of large-scale genetic studies. *Journal of Human Genetics*, *68*(3). https://doi.org/10.1038/s10038-022-01059-4

Katz, W., Weinstein, B., & Solomon, F. (1990). Regulation of tubulin levels and microtubule assembly in Saccharomyces cerevisiae: consequences of altered tubulin gene copy number. *Molecular and Cellular Biology*, *10*(10), 5286–5294.

Kaufmann, T. L., Petkovic, M., Watkins, T. B. K., Colliver, E. C., Laskina, S., Thapa, N., Minussi, D. C., Navin, N., Swanton, C., Van Loo, P., Haase, K., Tarabichi, M., & Schwarz, R. F. (2022). MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome Biology*, *23*(1), 241.

Kikutake, C., & Suyama, M. (2023). Pan-cancer analysis of whole-genome doubling and its association with patient prognosis. *BMC Cancer*, *23*(1). https://doi.org/10.1186/s12885-023-11132-6

Kim, J. E., Choi, J., Sung, C.-O., Hong, Y. S., Kim, S. Y., Lee, H., Kim, T. W., & Kim, J.-I. (2021). High prevalence of TP53 loss and whole-genome doubling in early-onset colorectal cancer. *Experimental & Molecular Medicine*, *53*(3), 446–456.

Klaasen, S. J., & Kops, G. J. P. L. (2022). Chromosome Inequality: Causes and Consequences of Non-Random Segregation Errors in Mitosis and Meiosis. *Cells*, *11*(22). https://doi.org/10.3390/cells11223564

Klaasen, S. J., Truong, M. A., van Jaarsveld, R. H., Koprivec, I., Štimac, V., de Vries, S. G., Risteski, P., Kodba, S., Vukušić, K., de Luca, K. L., Marques, J. F., Gerrits, E. M., Bakker, B., Foijer, F., Kind, J., Tolić, I. M., Lens, S. M. A., & Kops, G. J. P. L. (2022). Nuclear chromosome locations dictate segregation error frequencies. *Nature*, *607*(7919), 604–609.

Knudson, A. G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(4), 820–823.

Konishi, H., Mohseni, M., Tamaki, A., Garay, J. P., Croessmann, S., Karnan, S., Ota, A., Wong, H. Y., Konishi, Y., Karakas, B., Tahir, K., Abukhdeir, A. M., Gustin, J. P., Cidado, J., Wang, G. M., Cosgrove, D., Cochran, R., Jelovac, D., Higgins, M. J., … Park, B. H. (2011). Mutation of a single allele of the cancer susceptibility gene BRCA1 leads to genomic instability in human breast epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(43), 17773–17778.

Kosti, I., Jain, N., Aran, D., Butte, A. J., & Sirota, M. (2016). Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Scientific Reports*, *6*, 24799.

Kovac, M., Navas, C., Horswell, S., Salm, M., Bardella, C., Rowan, A., Stares, M., Castro-Giner, F., Fisher, R., de Bruin, E. C., Kovacova, M., Gorman, M., Makino, S., Williams, J., Jaeger, E., Jones, A., Howarth, K., Larkin, J., Pickering, L., … Tomlinson, I. (2015). Recurrent chromosomal gains and heterogeneous driver mutations characterise papillary renal

cancer evolution. *Nature Communications*, *6*, 6336.

Kuang, X., & Li, J. (2022). Chromosome instability and aneuploidy as context-dependent activators or inhibitors of antitumor immunity. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.895961

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*, 1–26.

Kumar, C. C. (2011). Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes & Cancer*, *2*(2), 95–107.

Lakhani, A. A., Thompson, S. L., & Sheltzer, J. M. (2023). Aneuploidy in Human Cancer: New Tools and Perspectives. *Trends in Genetics : TIG*, *39*(12), 968.

Lambuta, R. A., Nanni, L., Liu, Y., Diaz-Miyar, J., Iyer, A., Tavernari, D., Katanayeva, N., Ciriello, G., & Oricchio, E. (2023). Whole-genome doubling drives oncogenic loss of chromatin segregation. *Nature*, *615*(7954), 925–933.

Lau, T. Y., & Poon, R. Y. C. (2023). Whole-Genome Duplication and Genome Instability in Cancer Cells: Double the Trouble. *International Journal of Molecular Sciences*, *24*(4). https://doi.org/10.3390/ijms24043733

Lee, C., Fernandez, K. J., Alexandrou, S., Sergio, C. M., Deng, N., Rogers, S., Burgess, A., & Caldon, C. E. (2020). Cyclin E2 Promotes Whole Genome Doubling in Breast Cancer. *Cancers*, *12*(8), 2268.

Leite, A. J. C., Pinto, I. P., Cunha, D. M. da C. E., Ribeiro, C. L., da Silva, C. C., da Cruz, A. D., & Minasi, L. B. (2016). The Identification of

Microdeletion and Reciprocal Microduplication in 22q11.2 Using High-Resolution CMA Technology. *BioMed Research International*, *2016*, 7415438.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.

Leung, W., Teater, M., Durmaz, C., Meydan, C., Chivu, A. G., Chadburn, A., Rice, E. J., Muley, A., Camarillo, J. M., Arivalagan, J., Li, Z., Flowers, C. R., Kelleher, N. L., Danko, C. G., Imielinski, M., Dave, S. S., Armstrong, S. A., Mason, C. E., & Melnick, A. M. (2022). SETD2 Haploinsufficiency Enhances Germinal Center-Associated AICDA Somatic Hypermutation to Drive B-cell Lymphomagenesis. *Cancer Discovery*, *12*(7), 1782–1803.

Levatić, J., Salvadores, M., Fuster-Tormo, F., & Supek, F. (2022). Mutational signatures are markers of drug sensitivity of cancer cells. *Nature Communications*, *13*(1), 1–19.

Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73.

Li, R., & Zhu, J. (2022). Effects of aneuploidy on cell behaviour and function. *Nature Reviews. Molecular Cell Biology*, *23*(4). https://doi.org/10.1038/s41580-021-00436-9

Li, Z.-G., Fu, X.-M., Chai, C.-Y., Sun, F.-F., Xiao, F.-F., Huang, Y.-X., Yao, K., Chen, J.-P., & Hou, Y. (2020). A single copy of large tumor suppressor 1 or large tumor suppressor 2 is sufficient for normal hematopoiesis. *Chinese Medical Journal*, *133*(16), 1943.

Liehr, T., Othman, M. A. K., Rittscher, K., & Alhourani, E. (2015). The current state of molecular cytogenetics in cancer diagnosis. *Expert Review of Molecular Diagnostics*, *15*(4), 517–526.

Lin, Y., Wang, J., Wang, K., Bai, S., Thennavan, A., Wei, R., Yan, Y., Li, J., Elgamal, H., Sei, E., Casasent, A., Rao, M., Tang, C., Multani, A. S., Ma, J., Montalvan, J., Nagi, C., Winocour, S., Lim, B., … Navin, N. (2024). Normal breast tissues harbour rare populations of aneuploid epithelial cells. *Nature*, *636*(8043). https://doi.org/10.1038/s41586-024-08129-x

Lindor, N. M., Ney, J. A., Gaffey, T. A., Jenkins, R. B., Thibodeau, S. N., & Dewald, G. W. (1992). A genetic review of complete and partial hydatidiform moles and nonmolar triploidy. *Mayo Clinic Proceedings*, *67*(8). https://doi.org/10.1016/s0025-6196(12)60805-2

Liu, L., Chen, H., Sun, C., Zhang, J., Wang, J., Du, M., Li, J., Di, L., Shen, J., Geng, S., Pang, Y., Luo, Y., Wu, C., Fu, Y., Zheng, Z., Wang, J., & Huang, Y. (2022). Low-frequency somatic copy number alterations in normal human lymphocytes revealed by large-scale single-cell whole-genome profiling. *Genome Research*, *32*(1), 44.

Liu, X., Zhang, K., Kaya, N. A., Jia, Z., Wu, D., Chen, T., Liu, Z., Zhu, S., Hillmer, A. M., Wuestefeld, T., Liu, J., Chan, Y. S., Hu, Z., Ma, L., Jiang,

L., & Zhai, W. (2024). Tumor phylogeography reveals block-shaped spatial heterogeneity and the mode of evolution in Hepatocellular Carcinoma. *Nature Communications*, *15*(1), 1–14.

Liu, Y., Shi, Y., Wu, M., Liu, J., Wu, H., Xu, C., & Chen, L. (2022). Hypoxia-induced polypoid giant cancer cells in glioma promote the transformation of tumor-associated macrophages to a tumor-supportive phenotype. *CNS Neuroscience & Therapeutics*, *28*(9), 1326–1338.

Liu-Lupo, Y., Ham, J. D., Jeewajee, S. K. A., Nguyen, L., Delorey, T., Ramos, A., Weinstock, D. M., Regev, A., & Hemann, M. T. (2023). Integrated multi-omics analyses reveal homology-directed repair pathway as a unique dependency in near-haploid leukemia. *Blood Cancer Journal*, *13*(1), 92.

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., Birkbak, N. J., Wilson, G. A., Van Loo, P., Jamal-Hanjani, M., TRACERx Consortium, Swanton, C., & McGranahan, N. (2020). Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nature Genetics*, *52*(3), 283–293.

Lopez-Sánchez, L. M., Jimenez, C., Valverde, A., Hernandez, V., Peñarando, J., Martinez, A., Lopez-Pedrera, C., Muñoz-Castañeda, J. R., De la Haba-Rodríguez, J. R., Aranda, E., & Rodriguez-Ariza, A. (2014). CoCl2, a mimic of hypoxia, induces formation of polyploid giant cells with stem characteristics in colon cancer. *PloS One*, *9*(6), e99143.

Lu, W.-T., Zalmas, L.-P., Bailey, C., Black, J. R. M., Martinez-Ruiz, C.,

Pich, O., Gimeno-Valiente, F., Usaite, I., Magness, A., Thol, K., Webber, T. A., Jiang, M., Saunders, R. E., Liu, Y.-H., Biswas, D., Ige, E. O., Aerne, B., Grönroos, E., Venkatesan, S., … Swanton, C. (2024). TRACERx analysis identifies a role for FAT1 in regulating chromosomal instability and whole-genome doubling via Hippo signalling. *Nature Cell Biology*, *27*(1), 154–168.

Lukow, D. A., & Sheltzer, J. M. (2022). Chromosomal instability and aneuploidy as causes of cancer drug resistance. *Trends in Cancer*, *8*(1). https://doi.org/10.1016/j.trecan.2021.09.002

Lundin-Ström, K. B., Biloglav, A., Lazarevic, V., Behrendtz, M., Castor, A., & Johansson, B. (2021). Parental origin of monosomy 7 in acute leukaemia. *British Journal of Haematology*, *192*(5), e132–e135.

Lv, S., Ji, L., Chen, B., Liu, S., Lei, C., Liu, X., Qi, X., Wang, Y., Lai-Han Leung, E., Wang, H., Zhang, L., Yu, X., Liu, Z., Wei, Q., & Lu, L. (2018). Histone methyltransferase KMT2D sustains prostate carcinogenesis and metastasis via epigenetically activating LIFR and KLF4. *Oncogene*, *37*(10), 1354–1368.

Makanae, K., Kintaka, R., Makino, T., Kitano, H., & Moriya, H. (2013). Identification of dosage-sensitive genes in Saccharomyces cerevisiae using the genetic tug-of-war method. *Genome Research*, *23*(2), 300–311.

Makino, T., & McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9270–9274.

Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C., & Beerenwinkel, N. (2019). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature Communications*, *10*(1), 1–12.

Mandahl, N., Johansson, B., Mertens, F., & Mitelman, F. (2012). Disease-associated patterns of disomic chromosomes in hyperhaploid neoplasms. *Genes, Chromosomes & Cancer*, *51*(6), 536–544.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, *7*(1), 1–15.

Marques, J. F., & Gjpl, K. (2023). Permission to pass: on the role of p53 as a gatekeeper for aneuploidy. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *31*(4). https://doi.org/10.1007/s10577-023-09741-9

Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R. J., Abascal, F., Hall, M. W. J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., Fitzgerald, R. C., Handford, P. A., Campbell, P. J., Saeb-Parsy, K., & Jones, P. H. (2018). Somatic mutant clones colonize the human esophagus with age. *Science (New York, N.Y.)*, *362*(6417), 911–917.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2018). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *173*(7), 1823.

Martincorena, I. (2019). Somatic mutation and clonal expansions in

human tissues. *Genome Medicine*, *11*(1), 35.

Marusyk, A., & Polyak, K. (2010). Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta*, *1805*(1). https://doi.org/10.1016/j.bbcan.2009.11.002

Marusyk, A., Janiszewska, M., & Polyak, K. (2020). Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell*, *37*(4), 471–484.

Masilamani, A. P., Ferrarese, R., Kling, E., Thudi, N. K., Kim, H., Scholtens, D. M., Dai, F., Hadler, M., Unterkircher, T., Platania, L., Weyerbrock, A., Prinz, M., Gillespie, G. Y., Harsh, G. R., IV, Bredel, M., & Carro, M. S. (2017). KLF6 depletion promotes NF-κB signaling in glioblastoma. *Oncogene*, *36*(25), 3562–3575.

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756.

McFarland, C. D., Yaglom, J. A., Wojtkowiak, J. W., Scott, J. G., Morse, D. L., Sherman, M. Y., & Mirny, L. A. (2017). The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Research*, *77*(18). https://doi.org/10.1158/0008-5472.CAN-15-3283-T

McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, *7*(283), 283ra54.

McNeer, J. L., Devidas, M., Dai, Y., Carroll, A. J., Heerema, N. A.,

Gastier-Foster, J. M., Kahwash, S. B., Borowitz, M. J., Wood, B. L., Larsen, E., Maloney, K. W., Mattano, L., Winick, N. J., Schultz, K. R., Hunger, S. P., Carroll, W. L., Loh, M. L., & Raetz, E. A. (2019). Hematopoietic Stem-Cell Transplantation Does Not Improve the Poor Outcome of Children With Hypodiploid Acute Lymphoblastic Leukemia: A Report From Children's Oncology Group. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *37*(10), 780–789.

McPherson, A., Vázquez-García, I., Myers, M. A., Zatzman, M., Al-Rawi, D., Weiner, A., Freeman, S., Mohibullah, N., Satas, G., Williams, M. J., Ceglia, N., Zhang, A. W., Li, J., Lim, J. L. P., Wu, M., Choi, S., Havasov, E., Grewal, D., Shi, H., … Shah, S. P. (2024). Ongoing genome doubling promotes evolvability and immune dysregulation in ovarian cancer. In *bioRxiv*. https://doi.org/10.1101/2024.07.11.602772

Meader, B. N., Albano, A., Sekizkardes, H., & Delaney, A. (2020). Heterozygous Deletions in MKRN3 Cause Central Precocious Puberty Without Prader-Willi Syndrome. *The Journal of Clinical Endocrinology and Metabolism*, *105*(8), 2732–2739.

Meiller, C., Montagne, F., Hirsch, T. Z., Caruso, S., de Wolf, J., Bayard, Q., Assié, J. B., Meunier, L., Blum, Y., Quetel, L., Gibault, L., Pintilie, E., Badoual, C., Humez, S., Galateau-Sallé, F., Copin, M. C., Letouzé, E., Scherpereel, A., Zucman-Rossi, J., … Jean, D. (2021). Multi-site tumor sampling highlights molecular intra-tumor heterogeneity in malignant pleural mesothelioma. *Genome Medicine*, *13*(1).

https://doi.org/10.1186/s13073-021-00931-w

Mendes-da-Silva, P., Moreira, A., Duro-da-Costa, J., Matias, D., & Monteiro, C. (2000). Frequent loss of heterozygosity on chromosome 5 in non-small cell lung carcinoma. *Molecular Pathology : MP*, *53*(4), 184–187.

Meyer, R. E., Liu, G., Gilboa, S. M., Ethen, M. K., Aylsworth, A. S., Powell, C. M., Flood, T. J., Mai, C. T., Wang, Y., & Canfield, M. A. (2016). Survival of children with trisomy 13 and trisomy 18: A multi-state population-based study. *American Journal of Medical Genetics. Part A*, *170A*(4). https://doi.org/10.1002/ajmg.a.37495

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., & Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, *8*, 15183.

Minello, A., & Carreira, A. (2024). BRCA1/2 Haploinsufficiency: Exploring the Impact of Losing one Allele. *Journal of Molecular Biology*, *436*(1), 168277.

Minikel, E. V., Karczewski, K. J., Martin, H. C., Cummings, B. B., Whiffin, N., Rhodes, D., Alföldi, J., Trembath, R. C., van Heel, D. A., Daly, M. J., Schreiber, S. L., & MacArthur, D. G. (2020). Evaluating drug targets through human loss-of-function genetic variation. *Nature*, *581*(7809), 459–464.

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2025). Mitelman F, Johansson B and Mertens F (Eds.), https://mitelmandatabase.isb-cgc.org

Mohanty, V., Wang, F., Mills, G. B., CTD2 Research Network, & Chen, K.

(2021). Uncoupling of gene expression from copy number presents therapeutic opportunities in aneuploid cancers. *Cell Reports. Medicine*, *2*(7), 100349.

Molina, O., Bataller, A., Thampi, N., Ribera, J., Granada, I., Velasco, P., Fuster, J. L., & Menéndez, P. (2021). Near-Haploidy and Low-Hypodiploidy in B-Cell Acute Lymphoblastic Leukemia: When Less Is Too Much. *Cancers*, *14*(1). https://doi.org/10.3390/cancers14010032

Molina-Vila, M. A., Bertran-Alamillo, J., Gascó, A., Mayo-de-las-Casas, C., Sánchez-Ronco, M., Pujantell-Pastor, L., Bonanno, L., Favaretto, A. G., Cardona, A. F., Vergnenègre, A., Majem, M., Massuti, B., Morán, T., Carcereny, E., Viteri, S., & Rosell, R. (2014). Nondisruptive p53 mutations are associated with shorter survival in patients with advanced non-small cell lung cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *20*(17). https://doi.org/10.1158/1078-0432.CCR-13-2391

Moore, C. M., Jacobson, S. A., & Fingerlin, T. E. (2019). Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Human Heredity*, *84*(6). https://doi.org/10.1159/000508558

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., & Papaleo, E. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Computational Biology*, *15*(3), e1006701.

Mullegama, S. V., Rosenfeld, J. A., Orellana, C., van Bon, B. W. M.,

Halbach, S., Repnikova, E. A., Brick, L., Li, C., Dupuis, L., Rosello, M., Aradhya, S., Stavropoulos, D. J., Manickam, K., Mitchell, E., Hodge, J. C., Talkowski, M. E., Gusella, J. F., Keller, K., Zonana, J., … Elsea, S. H. (2014). Reciprocal deletion and duplication at 2q23.1 indicates a role for MBD5 in autism spectrum disorder. *European Journal of Human Genetics : EJHG*, *22*(1), 57–63.

Murakami, F., Tsuboi, Y., Takahashi, Y., Horimoto, Y., Mogushi, K., Ito, T., Emi, M., Matsubara, D., Shibata, T., Saito, M., & Murakami, Y. (2021). Short somatic alterations at the site of copy number variation in breast cancer. *Cancer Science*, *112*(1), 444–453.

Nachman, J. B., Heerema, N. A., Sather, H., Camitta, B., Forestier, E., Harrison, C. J., Dastugue, N., Schrappe, M., Pui, C.-H., Basso, G., Silverman, L. B., & Janka-Schaub, G. E. (2007). Outcome of treatment in children with hypodiploid acute lymphoblastic leukemia. *Blood*, *110*(4), 1112–1115.

Nair, N. U., Schäffer, A. A., Gertz, E. M., Cheng, K., Zerbib, J., Sahu, A. D., Leor, G., Shulman, E. D., Aldape, K. D., Ben-David, U., & Ruppin, E. (2024). Chromosome 7 Gain Compensates for Chromosome 10 Loss in Glioma. *Cancer Research*, *84*(20), 3464–3477.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics (Oxford, England)*, *28*(14), 1919–1920.

Nevado, J., Mergener, R., Palomares-Bralo, M., Souza, K. R., Vallespín, E., Mena, R., Martínez-Glez, V., Mori, M. Á., Santos, F., García-Miñaur, S., García-Santiago, F., Mansilla, E., Fernández, L., de Torres, M. L., Riegel, M., & Lapunzina, P. (2014). New microdeletion and microduplication syndromes: A comprehensive review. *Genetics and Molecular Biology*, *37*(1 Suppl). https://doi.org/10.1590/s1415-47572014000200007

Newcomb, R., Dean, E., McKinney, B. J., & Alvarez, J. V. (2021). Context-dependent effects of whole-genome duplication during mammary tumor recurrence. *Scientific Reports*, *11*(1), 14932.

Nguyen, J. M., Qualmann, K. J., Okashah, R., Reilly, A., Alexeyev, M. F., & Campbell, D. J. (2015). 5p deletions: Current knowledge and future directions. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, *169*(3). https://doi.org/10.1002/ajmg.c.31444

Ni, Z., Zhou, X. Y., Aslam, S., & Niu, D. K. (2019). Characterization of Human Dosage-Sensitive Transcription Factor Genes. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01208

Nichols, C. A., Gibson, W. J., Brown, M. S., Kosmicki, J. A., Busanovich, J. P., Wei, H., Urbanski, L. M., Curimjee, N., Berger, A. C., Gao, G. F., Cherniack, A. D., Dhe-Paganon, S., Paolella, B. R., & Beroukhim, R. (2020). Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. *Nature Communications*, *11*(1), 2517.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations.

Science (New York, N.Y.), *194*(4260), 23–28.

Offin, M., Chan, J. M., Tenet, M., Rizvi, H. A., Shen, R., Riely, G. J., Rekhtman, N., Daneshbod, Y., Quintanal-Villalonga, A., Penson, A., Hellmann, M. D., Arcila, M. E., Ladanyi, M., Pe'er, D., Kris, M. G., Rudin, C. M., & Yu, H. A. (2019). Concurrent RB1 and TP53 alterations define a subset of EGFR-mutant lung cancers at risk for histologic transformation and inferior clinical outcomes. *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer*, *14*(10), 1784.

Ogden, A., Rida, P. C. G., Knudsen, B. S., Kucuk, O., & Aneja, R. (2015). Docetaxel-induced polyploidization may underlie chemoresistance and disease relapse. *Cancer Letters*, *367*(2), 89–92.

Oh, H. J., Bae, J. M., Wen, X., Jung, S., Kim, Y., Kim, K. J., Cho, N. Y., Kim, J. H., Han, S. W., Kim, T. Y., & Kang, G. H. (2019). p53 expression status is associated with cancer-specific survival in stage III and high-risk stage II colorectal cancer patients treated with oxaliplatin-based adjuvant chemotherapy. *British Journal of Cancer*, *120*(8). https://doi.org/10.1038/s41416-019-0429-2

Ohashi, A., Ohori, M., Iwai, K., Nakayama, Y., Nambu, T., Morishita, D., Kawamoto, T., Miyamoto, M., Hirayama, T., Okaniwa, M., Banno, H., Ishikawa, T., Kandori, H., & Iwata, K. (2015). Aneuploidy generates proteotoxic stress and DNA damage concurrently with p53-mediated post-mitotic apoptosis in SAC-impaired cells. *Nature Communications*, *6*, 7668.

Ohno, S. (n.d.). *Evolution by Gene Duplication*. Springer Berlin Heidelberg.

Okamura, T., Yoshimoto, J., Morimoto, D., Wanatabe, H., & Washio, Y. (2024). Long-term survival of an infant with complete tetraploidy: A case report. *American Journal of Medical Genetics. Part A*, *194*(11), e63717.

Olsson, L., Paulsson, K., Bovée, J. V. M. G., & Nord, K. H. (2011). Clonal evolution through loss of chromosomes and subsequent polyploidization in chondrosarcoma. *PloS One*, *6*(9), e24977.

org.Hs.eg.db. (n.d.). Bioconductor. Retrieved January 30, 2025, from http://bioconductor.org/packages/org.Hs.eg.db/

Pacchierotti, F., Adler, I. D., Eichenlaub-Ritter, U., & Mailhes, J. B. (2007). Gender effects on the incidence of aneuploidy in mammalian germ cells. *Environmental Research*, *104*(1). https://doi.org/10.1016/j.envres.2006.12.001

Pacini, C., Dempster, J. M., Boyle, I., Gonçalves, E., Najgebauer, H., Karakoc, E., van der Meer, D., Barthorpe, A., Lightfoot, H., Jaaks, P., McFarland, J. M., Garnett, M. J., Tsherniak, A., & Iorio, F. (2021). Integrated cross-study datasets of genetic dependencies in cancer. *Nature Communications*, *12*(1), 1661.

Pan-genome insights into adaptive evolution of bacterial symbionts in mixed host-microbe symbioses represented by human gut microbiota Bacteroides cellulosilyticus. (2024). *Science of The Total Environment*, *927*, 172251.

Park, S., Supek, F., & Lehner, B. (2021). Higher order genetic

interactions switch cancer genes from two-hit to one-hit drivers. *Nature Communications*, *12*(1), 7051.

Parl, F. F. (2023). Different Tumor Types Share a Common Nuclear Map of Chromosome Territories. *Cancer Informatics*, *22*, 11769351221148592.

Passerini, V., Ozeri-Galai, E., de Pagter, M. S., Donnelly, N., Schmalbrock, S., Kloosterman, W. P., Kerem, B., & Storchová, Z. (2016). The presence of extra chromosomes leads to genomic instability. *Nature Communications*, *7*, 10754.

Patel, D., Incassati, A., Wang, N., & McCance, D. J. (2004). Human papillomavirus type 16 E6 and E7 cause polyploidy in human keratinocytes and up-regulation of G2-M-phase proteins. *Cancer Research*, *64*(4). https://doi.org/10.1158/0008-5472.can-03-2917

Patkar, S., Heselmeyer-Haddad, K., Auslander, N., Hirsch, D., Camps, J., Bronder, D., Brown, M., Chen, W.-D., Lokanga, R., Wangsa, D., Wangsa, D., Hu, Y., Lischka, A., Braun, R., Emons, G., Ghadimi, B. M., Gaedcke, J., Grade, M., Montagna, C., … Ried, T. (2021). Hard wiring of normal tissue-specific chromosome-wide gene expression levels is an additional factor driving cancer type-specific aneuploidies. *Genome Medicine*, *13*(1), 93.

Payton, M., Belmontes, B., Hanestad, K., Moriguchi, J., Chen, K., McCarter, J. D., Chung, G., Ninniri, M. S., Sun, J., Manoukian, R., Chambers, S., Ho, S.-M., Kurzeja, R. J. M., Edson, K. Z., Dahal, U. P., Wu, T., Wannberg, S., Beltran, P. J., Canon, J., … Hughes, P. E. (2023). Small-molecule inhibition of kinesin KIF18A reveals a mitotic

vulnerability enriched in chromosomally unstable cancers. *Nature Cancer*, *5*(1), 66–84.

Pedersen T (2025). patchwork: The Composer of Plots. R package version 1.3

Pelleri, M. C., Cicchini, E., Petersen, M. B., Tranebjaerg, L., Mattina, T., Magini, P., Antonaros, F., Caracausi, M., Vitale, L., Locatelli, C., Seri, M., Strippoli, P., Piovesan, A., & Cocchi, G. (2019). Partial trisomy 21 map: Ten cases further supporting the highly restricted Down syndrome critical region (HR-DSCR) on human chromosome 21. *Molecular Genetics & Genomic Medicine*, *7*(8), e797.

Perkail, S., Andricovich, J., Kai, Y., & Tzatsos, A. (2020). BAP1 is a haploinsufficient tumor suppressor linking chronic pancreatitis to pancreatic cancer in mice. *Nature Communications*, *11*(1), 3018.

Persi, E., Wolf, Y. I., Leiserson, M. D. M., Koonin, E. V., & Ruppin, E. (2018). Criticality in tumor evolution and clinical outcome. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(47). https://doi.org/10.1073/pnas.1807256115

Pertesi, M., Ekdahl, L., Palm, A., Johnsson, E., Järvstråt, L., Wihlborg, A.-K., & Nilsson, B. (2019). Essential genes shape cancer genomes through linear limitation of homozygous deletions. *Communications Biology*, *2*, 262.

Pienta, K. J., Hammarlund, E. U., Austin, R. H., Axelrod, R., Brown, J. S., & Amend, S. R. (2022). Cancer cells employ an evolutionarily conserved polyploidization program to resist therapy. *Seminars in Cancer*

Biology, *81*. https://doi.org/10.1016/j.semcancer.2020.11.016

Pineda, J. M. B., & Bradley, R. K. (2024). DUX4 is a common driver of immune evasion and immunotherapy failure in metastatic cancers. *eLife*, *12*. https://doi.org/10.7554/eLife.89017

Polyploid giant cancer cells, EZH2 and Myc upregulation in mammary epithelial cells infected with high-risk human cytomegalovirus. (2022). *eBioMedicine*, *80*, 104056.

Potapova, T. A., Zhu, J., & Li, R. (2013). Aneuploidy and chromosomal instability: a vicious cycle driving cellular evolution and cancer genome chaos. *Cancer Metastasis Reviews*, *32*(0), 10.1007/s10555–013 – 9436–6.

Prasad, K., Bloomfield, M., Levi, H., Keuper, K., Bernhard, S. V., Baudoin, N. C., Leor, G., Eliezer, Y., Giam, M., Wong, C. K., Rancati, G., Storchová, Z., Cimini, D., & Ben-David, U. (2022). Whole-Genome Duplication Shapes the Aneuploidy Landscape of Human Cancers. *Cancer Research*, *82*(9), 1736–1752.

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., Voda, M., Bloemendal, H. J., Tjan-Heijnen, V. C. G., van Herpen, C. M. L., Labots, M., Witteveen, P. O., Smit, E. F., Sleijfer, S., … Cuppen, E. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*, *575*(7781), 210–216.

Qi, X.-T., Li, Y.-L., Zhang, Y.-Q., Xu, T., Lu, B., Fang, L., Gao, J.-Q., Yu, L.-S., Zhu, D.-F., Yang, B., He, Q.-J., & Ying, M.-D. (2019). KLF4 functions as an oncogene in promoting cancer stem cell-like

characteristics in osteosarcoma cells. *Acta Pharmacologica Sinica*, *40*(4), 546–555.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842.

Quinton, R. J., DiDomizio, A., Vittoria, M. A., Kotýnková, K., Ticas, C. J., Patel, S., Koga, Y., Vakhshoorzadeh, J., Hermance, N., Kuroda, T. S., Parulekar, N., Taylor, A. M., Manning, A. L., Campbell, J. D., & Ganem, N. J. (2021). Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature*, *590*(7846), 492–497.

Ragnum, H. B., Vlatkovic, L., Lie, A. K., Axcrona, K., Julin, C. H., Frikstad, K. M., Hole, K. H., Seierstad, T., & Lyng, H. (2014). The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. *British Journal of Cancer*, *112*(2), 382.

Ramaker, R. C., Lasseigne, B. N., Hardigan, A. A., Palacio, L., Gunther, D. S., Myers, R. M., & Cooper, S. J. (2017). RNA sequencing-based cell proliferation analysis across 19 cancers identifies a subset of proliferation-informative cancers with a common survival signature. *Oncotarget*, *8*(24), 38668–38681.

Ramos M, Schiffer L, Waldron L (2024). TCGAutils: TCGA utility functions for data management. R package version 1.26.

Randall, M. P., Egolf, L. E., Vaksman, Z., Samanta, M., Tsang, M., Groff, D., Evans, J. P., Rokita, J. L., Layeghifard, M., Shlien, A., Maris, J. M.,

Diskin, S. J., & Bosse, K. R. (2024). BARD1 germline variants induce haploinsufficiency and DNA repair defects in neuroblastoma. *Journal of the National Cancer Institute*, *116*(1), 138–148.

Rice, A. M., & McLysaght, A. (2017). Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nature Communications*, *8*, 14366.

Robertson, S. A., & Richards, R. I. (2024). Single-cell sequencing shows mosaic aneuploidy in most human embryos. *The Journal of Clinical Investigation*, *134*(6). https://doi.org/10.1172/JCI179134

Robinson, D. (2014). *broom: An R Package for Converting Statistical Analysis Objects Into Tidy Data Frames*. http://arxiv.org/abs/1412.3565

Roetman, J. J., Apostolova, M. K. I., & Philip, M. (2022). Viral and cellular oncogenes promote immune evasion. *Oncogene*, *41*(7), 921–929.

Roylance, R., Endesfelder, D., Gorman, P., Burrell, R. A., Sander, J., Tomlinson, I., Hanby, A. M., Speirs, V., Richardson, A. L., Birkbak, N. J., Eklund, A. C., Downward, J., Kschischo, M., Szallasi, Z., & Swanton, C. (2011). Relationship of Extreme Chromosomal Instability with Long-term Survival in a Retrospective Analysis of Primary Breast Cancer. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *20*(10), 2183.

Rutledge, S. D., Douglas, T. A., Nicholson, J. M., Vila-Casadesús, M., Kantzler, C. L., Wangsa, D., Barroso-Vilares, M., Kale, S. D., Logarinho, E., & Cimini, D. (2016). Selective advantage of trisomic human cells

cultured in non-standard conditions. *Scientific Reports*, *6*(1), 1–12.

Rylaarsdam, L., & Guemez-Gamboa, A. (2019). Genetic Causes and Modifiers of Autism Spectrum Disorder. *Frontiers in Cellular Neuroscience*, *13*, 385.

Ryland, G. L., Doyle, M. A., Goode, D., Boyle, S. E., Choong, D. Y. H., Rowley, S. M., Li, J., Australian Ovarian Cancer Study Group, Bowtell, D. D. L., Tothill, R. W., Campbell, I. G., & Gorringe, K. L. (2015). Loss of heterozygosity: what is it good for? *BMC Medical Genomics*, *8*, 45.

Safavi, S., & Paulsson, K. (2017). Near-haploid and low-hypodiploid acute lymphoblastic leukemia: two distinct subtypes with consistently poor prognosis. *Blood*, *129*(4), 420–423.

Sagi, I., Egli, D., & Benvenisty, N. (2016). Identification and propagation of haploid human pluripotent stem cells. *Nature Protocols*, *11*(11), 2274–2286.

Sanz-Gómez, N., González-Álvarez, M., De Las Rivas, J., & de Cárcer, G. (2023). Whole-Genome Doubling as a source of cancer: how, when, where, and why? *Frontiers in Cell and Developmental Biology*, *11*, 1209136.

Sapkota, H., Wasiak, E., Daum, J. R., & Gorbsky, G. J. (2018). Multiple determinants and consequences of cohesion fatigue in mammalian cells. *Molecular Biology of the Cell*, *29*(15), 1811–1824.

Sarel-Gallily, R., Golan-Lev, T., Yilmaz, A., Sagi, I., & Benvenisty, N. (2022). Genome-wide analysis of haploinsufficiency in human embryonic stem cells. *Cell Reports*, *38*(13).

https://doi.org/10.1016/j.celrep.2022.110573

Satgunaseelan, L., Strbenac, D., Willet, C., Chew, T., Sadsad, R., Wykes, J., Low, T. (hubert), Cooper, W. A., Soon Lee, C., Palme, C. E., Yang, J. Y. H., Clark, J. R., & Gupta, R. (2022). Whole genome duplication in oral squamous cell carcinoma in patients younger than 50 years: implications for prognosis and adverse clinicopathological factors. *Genes, Chromosomes & Cancer*, *61*(9), 561.

Sayaman, R. W., Saad, M., Thorsson, V., Hu, D., Hendrickx, W., Roelands, J., Porta-Pardo, E., Mokrab, Y., Farshidfar, F., Kirchhoff, T., Sweis, R. F., Bathe, O. F., Heimann, C., Campbell, M. J., Stretch, C., Huntsman, S., Graff, R. E., Syed, N., Radvanyi, L., … Bedognetti, D. (2021). Germline genetic contribution to the immune landscape of cancer. *Immunity*, *54*(2), 367–386.e8.

Schaefer, M. H., & Serrano, L. (2016). Cell type-specific properties and environment shape tissue specificity of cancer genes. *Scientific Reports*, *6*, 20707.

Schauberger P, Walker A (2025). *openxlsx: Read, Write and Edit xlsx Files*. R package version 4.2.8, https://github.com/ycphs/openxlsx, https://ycphs.github.io/openxlsx/index.html.

Scheijen, B., Boer, J. M., Marke, R., Tijchon, E., van Ingen Schenau, D., Waanders, E., van Emst, L., van der Meer, L. T., Pieters, R., Escherich, G., Horstmann, M. A., Sonneveld, E., Venn, N., Sutton, R., Dalla-Pozza, L., Kuiper, R. P., Hoogerbrugge, P. M., den Boer, M. L., & van Leeuwen, F. N. (2017). Tumor suppressors BTG1 and IKZF1 cooperate during mouse leukemia development and increase relapse risk in B-cell

precursor acute lymphoblastic leukemia patients. *Haematologica*, *102*(3), 541–551.

Schlüter, A., Weller, P., Kanaan, O., Nel, I., Heusgen, L., Höing, B., Haßkamp, P., Zander, S., Mandapathil, M., Dominas, N., Arnolds, J., Stuck, B. A., Lang, S., Bankfalvi, A., & Brandau, S. (2018). CD31 and VEGF are prognostic biomarkers in early-stage, but not in late-stage, laryngeal squamous cell carcinoma. *BMC Cancer*, *18*(1), 272.

Schmidt, M. J., Naghdloo, A., Prabakar, R. K., Kamal, M., Cadaneanu, R., Garraway, I. P., Lewis, M., Aparicio, A., Zurita-Saavedra, A., Corn, P., Kuhn, P., Pienta, K. J., Amend, S. R., & Hicks, J. (2024). Polyploid cancer cells reveal signatures of chemotherapy resistance. *Oncogene*. https://doi.org/10.1038/s41388-024-03212-z

Schukken, K. M., & Sheltzer, J. M. (2022). Extensive protein dosage compensation in aneuploid human cancers. *Genome Research*, *32*(7), 1254–1270.

Schuster-Böckler, B., Conrad, D., & Bateman, A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. *PloS One*, *5*(3). https://doi.org/10.1371/journal.pone.0009474

Scrucca, L., Fraley, C., Brendan Murphy, T., & Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. CRC Press.

Senovilla, L., Vitale, I., Martins, I., Tailler, M., Pailleret, C., Michaud, M., Galluzzi, L., Adjemian, S., Kepp, O., Niso-Santano, M., Shen, S., Mariño, G., Criollo, A., Boilève, A., Job, B., Ladoire, S., Ghiringhelli, F., Sistigu,

A., Yamazaki, T., … Kroemer, G. (2012). An immunosurveillance mechanism controls cancer cell ploidy. *Science (New York, N.Y.)*, *337*(6102), 1678–1684.

Serebriiskii, I. G., Pavlov, V. A., Andrianov, G. V., Litwin, S., Basickes, S., Newberg, J. Y., Frampton, G. M., Meyer, J. E., & Golemis, E. A. (2023). Source, co-occurrence, and prognostic value of PTEN mutations or loss in colorectal cancer. *Npj Genomic Medicine*, *8*(1), 1–12.

Shabo, I., Svanvik, J., Lindström, A., Lechertier, T., Trabulo, S., Hulit, J., Sparey, T., & Pawelek, J. (2020). Roles of cell fusion, hybridization and polyploid cell formation in cancer metastasis. *World Journal of Clinical Oncology*, *11*(3), 121.

Shaffer, L. G., Agan, N., Goldberg, J. D., Ledbetter, D. H., Longshore, J. W., & Cassidy, S. B. (2001). American College of Medical Genetics Statement on Diagnostic Testing for Uniparental Disomy. *Genetics in Medicine*, *3*(3), 206–211.

Shahamatdar, S., He, M. X., Reyna, M. A., Gusev, A., AlDubayan, S. H., Van Allen, E. M., & Ramachandran, S. (2020). Germline Features Associated with Immune Infiltration in Solid Tumors. *Cell Reports*, *30*(9), 2900–2908.e4.

Sharma, V., Hecker, N., Roscito, J. G., Foerster, L., Langer, B. E., & Hiller, M. (2018). A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature Communications*, *9*(1), 1–9.

Shen, R., & Seshan, V. E. (2016). FACETS: allele-specific copy number

and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, *44*(16). https://doi.org/10.1093/nar/gkw520

Sherard, J., Bean, C., Bove, B., DelDuca, V., Jr, Esterly, K. L., Karcsh, H. J., Munshi, G., Reamer, J. F., Suazo, G., & Wilmoth, D. (1986). Long survival in a 69,XXY triploid male. *American Journal of Medical Genetics*, *25*(2), 307–312.

Shi, C., Liu, S., Tian, X., Wang, X., & Gao, P. (2021). A TP53 mutation model for the prediction of prognosis and therapeutic responses in head and neck squamous cell carcinoma. *BMC Cancer*, *21*(1), 1–13.

Shi, J., Jing, W., He, Y., & Huang, Y. (2024). Decreased expression of KLF6 in ectopic endometrial stromal cells contributes to endometriosis progression by targeting CTNNB1. *Cellular Signalling*, *120*, 111230.

Shih, J., Sarmashghi, S., Zhakula-Kostadinova, N., Zhang, S., Georgis, Y., Hoyt, S. H., Cuoco, M. S., Gao, G. F., Spurr, L. F., Berger, A. C., Ha, G., Rendo, V., Shen, H., Meyerson, M., Cherniack, A. D., Taylor, A. M., & Beroukhim, R. (2023). Cancer aneuploidies are shaped primarily by effects on tumour fitness. *Nature*, *619*(7971), 793–800.

Shorokhova, M., Nikolsky, N., & Grinchuk, T. (2021). Chromothripsis-Explosion in Genetic Science. *Cells*, *10*(5). https://doi.org/10.3390/cells10051102

Short, S. P., Kondo, J., Smalley-Freed, W. G., Takeda, H., Dohn, M. R., Powell, A. E., Carnahan, R. H., Washington, M. K., Tripathi, M., Payne, D. M., Jenkins, N. A., Copeland, N. G., Coffey, R. J., & Reynolds, A. B.

(2017). p120-Catenin is an obligate haploinsufficient tumor suppressor in intestinal neoplasia. *The Journal of Clinical Investigation*, *127*(12), 4462–4476.

Shortt, J., & Johnstone, R. W. (2012). Oncogenes in cell survival and cell death. *Cold Spring Harbor Perspectives in Biology*, *4*(12). https://doi.org/10.1101/cshperspect.a009829

Shu, Z., Row, S., & Deng, W. M. (2018). Endoreplication: The Good, the Bad, and the Ugly. *Trends in Cell Biology*, *28*(6). https://doi.org/10.1016/j.tcb.2018.02.006

Slowikowski K (2024). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. https://ggrepel.slowkow.com/, https://github.com/slowkow/ggrepel.

Smoking, chromosomal aberrations, and cancer incidence in healthy subjects. (2021). *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *867*, 503373.

Solimini, N. L., Xu, Q., Mermel, C. H., Liang, A. C., Schlabach, M. R., Luo, J., Burrows, A. E., Anselmo, A. N., Bredemeyer, A. L., Li, M. Z., Beroukhim, R., Meyerson, M., & Elledge, S. J. (2012). Recurrent Hemizygous Deletions in Cancers May Optimize Proliferative Potential. *Science (New York, N.Y.)*, *337*(6090), 104.

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews. Cancer*, *18*(11). https://doi.org/10.1038/s41568-018-0060-1

Sosinsky, A., Ambrose, J., Cross, W., Turnbull, C., Henderson, S., Jones, L., Hamblin, A., Arumugam, P., Chan, G., Chubb, D., Noyvert, B., Mitchell, J., Walker, S., Bowman, K., Pasko, D., Buongermino Pereira, M., Volkova, N., Rueda-Martin, A., Perez-Gil, D., … Murugaesu, N. (2024). Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nature Medicine*, *30*(1), 279–289.

Sotillo, R., Schvartzman, J.-M., Socci, N. D., & Benezra, R. (2010). Mad2-induced chromosome instability leads to lung tumor relapse after oncogene withdrawal. *Nature*, *464*(7287), 436.

Soussi, T., & Wiman, K. G. (2015). TP53: an oncogene in disguise. *Cell Death and Differentiation*, *22*(8), 1239–1249.

Starodub, A., O'Neil, B., Mathews, C. A., Dumbrava, E. E., Friedman, C. F., Swiecicki, P., Oh, C., Andreu, C., Lee, L., Campbell, J., Bowler, T. G., Drutman, S., & LoRusso, P. (2024). A phase I/II, first-in-human study of VLS-1488, an oral KIF18A inhibitor, in patients with advanced cancer. *Journal of Clinical Oncology*.

https://doi.org/10.1200/JCO.2024.42.16_suppl.TPS3181

Steele, C. D., Abbasi, A., Islam, S. M. A., Bowes, A. L., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P., McLatchie, A., Lechner, M., Light, N., Shlien, A., Malkin, D., Feber, A., Proszek, P., Lesluyes, T., Mertens, F., … Pillay, N. (2022). Signatures of copy number alterations in human cancer. *Nature*, *606*(7916), 984–991.

Stengel, A., Schnittger, S., Weissmann, S., Kuznia, S., Kern, W.,

Kohlmann, A., Haferlach, T., & Haferlach, C. (2014). TP53 mutations occur in 15.7% of ALL and are associated with MYC-rearrangement, low hypodiploidy, and a poor prognosis. *Blood*, *124*(2), 251–258.

Sun, G., Xu, Y., Liu, H., Sun, T., Zhang, J., Hettenhausen, C., Shen, G., Qi, J., Qin, Y., Li, J., Wang, L., Chang, W., Guo, Z., Baldwin, I. T., & Wu, J. (2018). Large-scale gene losses underlie the genome evolution of parasitic plant Cuscuta australis. *Nature Communications*, *9*(1), 1–8.

Sung, M.-K., Reitsma, J. M., Sweredoski, M. J., Hess, S., & Deshaies, R. J. (2016). Ribosomal proteins produced in excess are degraded by the ubiquitin-proteasome system. *Molecular Biology of the Cell*, *27*(17), 2642–2652.

Ta, D., Downs, J., Baynam, G., Wilson, A., Richmond, P., & Leonard, H. (2022). A brief history of MECP2 duplication syndrome: 20-years of clinical understanding. *Orphanet Journal of Rare Diseases*, *17*(1), 131.

Talevich, E., Hunter Shain, A., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology*, *12*(4), e1004873.

Tarabichi, M., Salcedo, A., Deshwar, A. G., Leathlobhair, M. N., Wintersinger, J., Wedge, D. C., Van Loo, P., Morris, Q. D., & Boutros, P. C. (2021). A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nature Methods*, *18*(2), 144.

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., Schumacher, S. E., Wang, C., Hu, H., Liu, J., Lazar, A. J., Cancer

Genome Atlas Research Network, Cherniack, A. D., Beroukhim, R., & Meyerson, M. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*, *33*(4), 676–689.e3.

The Multiple Facets of PRC2 Alterations in Cancers. (2017). *Journal of Molecular Biology*, *429*(13), 1978–1993.

The Polyploid State Plays a Tumor-Suppressive Role in the Liver. (2018). *Developmental Cell*, *44*(4), 447–459.e5.

Therneau T (2024). *A Package for Survival Analysis in R*. R package version 3.7-0, https://CRAN.R-project.org/package=survival.

Tilk, S., Tkachenko, S., Curtis, C., Petrov, D. A., & McFarland, C. D. (2022). Most cancers carry a substantial deleterious load due to Hill-Robertson interference. *eLife*, *11*. https://doi.org/10.7554/eLife.67790

Tomlinson, I. P. M., Lambros, M. B. K., & Roylance, R. R. (2002). Loss of heterozygosity analysis: practically and conceptually flawed? *Genes, Chromosomes & Cancer*, *34*(4), 349–353.

Torres, E. M., Springer, M., & Amon, A. (2016). No current evidence for widespread dosage compensation in S. cerevisiae. *eLife*, *5*, e10996.

Torres, E. M., Williams, B. R., & Amon, A. (2008). Aneuploidy: cells losing their balance. *Genetics*, *179*(2), 737–746.

Trigos, A. S., Pearson, R. B., Papenfuss, A. T., & Goode, D. L. (2017). Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. *Proceedings of the National Academy of Sciences of the United States of*

America, *114*(24), 6406–6411.

Tschaharganeh, D. F., Bosbach, B., & Lowe, S. W. (2016). Coordinated Tumor Suppression by Chromosome 8p. *Cancer Cell*, *29*(5), 617.

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., … Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, *170*(3), 564–576.e16.

Tsitsiridis, G., Steinkamp, R., Giurgiu, M., Brauner, B., Fobo, G., Frishman, G., Montrone, C., & Ruepp, A. (2023). CORUM: the comprehensive resource of mammalian protein complexes-2022. *Nucleic Acids Research*, *51*(D1), D539–D545.

Tumor evolution: Linear, branching, neutral or punctuated? (2017). *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1867*(2), 151–161.

Upadhya, S. R., & Ryan, C. J. (2022). Experimental reproducibility limits the correlation between mRNA and protein abundances in tumor proteomic profiles. *Cell Reports Methods*, *2*(9), 100288.

Valentin-Vega, Y. A., Barboza, J. A., Chau, G. P., El-Naggar, A. K., & Lozano, G. (2007). High levels of the p53 inhibitor MDM4 in head and neck squamous carcinomas. *Human Pathology*, *38*(10), 1553–1562.

Van den Eynden, J., Basu, S., & Larsson, E. (2016). Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution. *PLoS Genetics*, *12*(12), e1006506.

van Dijk, E., van den Bosch, T., Lenos, K. J., El Makrini, K., Nijman, L. E., van Essen, H. F. B., Lansu, N., Boekhout, M., Hageman, J. H., Fitzgerald, R. C., Punt, C. J. A., Tuynman, J. B., Snippert, H. J. G., Kops, G. J. P. L., Medema, J. P., Ylstra, B., Vermeulen, L., & Miedema, D. M. (2021). Chromosomal copy number heterogeneity predicts survival rates across cancers. *Nature Communications*, *12*(1), 3188.

Vandin, F. (2017). Computational Methods for Characterizing Cancer Mutational Heterogeneity. *Frontiers in Genetics*, *8*, 259369.

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., & Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(39), 16910–16915.

Vavouri, T., Semple, J. I., Garcia-Verdugo, R., & Lehner, B. (2009). Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, *138*(1), 198–208.

Veitia, R. A. (2004). Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics*, *168*(1), 569–574.

Venet, D., Dumont, J. E., & Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*, *7*(10). https://doi.org/10.1371/journal.pcbi.1002240

Venkatachalam, S., Tyner, S. D., Pickering, C. R., Boley, S., Recio, L.,

French, J. E., & Donehower, L. A. (2001). Is p53 haploinsufficient for tumor suppression? Implications for the p53+/- mouse model in carcinogenicity testing. *Toxicologic Pathology*, *29 Suppl*, 147–154.

Vora, S., Chatterjee, S., Andrew, A., Kumar, R. P., Proctor, M., Zeng, Z., Bhatt, R., Nazareth, D., Fernando, M., Jones, M. J. K., He, Y., Hooper, J. D., McMillan, N. A. J., Urosevic, J., Saeh, J., Travers, J., Cimini, D., Chen, J., & Gabrielli, B. (2025). Aurora B inhibition induces hyper-polyploidy and loss of long-term proliferative potential in RB and p53 defective cells. *Cell Death & Disease*, *16*(1), 7.

Voutsadakis, I. A. (2022). Urothelial Bladder Carcinomas with High Tumor Mutation Burden Have a Better Prognosis and Targetable Molecular Defects beyond Immunotherapies. *Current Oncology (Toronto, Ont.)*, *29*(3), 1390–1407.

Walen, K. H. (2014). Haploidization of human diploid metaphase cells: Is this genome reductive mechanism opperational in near-haploid leukemia? *Journal of Cancer Therapy*, *05*(01), 101–114.

Wan, T. S. K. (2014). Cancer cytogenetics: methodology revisited. *Annals of Laboratory Medicine*, *34*(6), 413–425.

Wang, J., & LaFramboise, T. (2019). CytoConverter: a web-based tool to convert karyotypes to genomic coordinates. *BMC Bioinformatics*, *20*(1), 467.

Wang, X., Hu, W., Li, X., Huang, D., Li, Q., Chan, H., Zeng, J., Xie, C., Chen, H., Liu, X., Gin, T., Wang, M. H., Cheng, A. S. L., Kang, W., To, K.-F., Plewczynski, D., Zhang, Q., Chen, X., Chan, D. C. W., … Wu, W. K. K.

(2022). Single-Hit Inactivation Drove Tumor Suppressor Genes Out of the X Chromosome during Evolution. *Cancer Research*, *82*(8), 1482–1491.

Was, H., Borkowska, A., Olszewska, A., Klemba, A., Marciniak, M., Synowiec, A., & Kieda, C. (2022). Polyploidy formation in cancer cells: How a Trojan horse is born. *Seminars in Cancer Biology*, *81*, 24–36.

Watkins, T. B. K., Lim, E. L., Petkovic, M., Elizalde, S., Birkbak, N. J., Wilson, G. A., Moore, D. A., Grönroos, E., Rowan, A., Dewhurst, S. M., Demeulemeester, J., Dentro, S. C., Horswell, S., Au, L., Haase, K., Escudero, M., Rosenthal, R., Bakir, M. A., Xu, H., … Swanton, C. (2020). Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, *587*(7832), 126–132.

Weaver, B. A., Silk, A. D., Montagna, C., Verdier-Pinard, P., & Cleveland, D. W. (2007). Aneuploidy acts both oncogenically and as a tumor suppressor. *Cancer Cell*, *11*(1). https://doi.org/10.1016/j.ccr.2006.12.003

Website. (n.d.). https://academic.oup.com/plcell/article/34/7/2466/6542320

Wheeler, D. A., & Wang, L. (2013). From human genome to cancer genome: the first decade. *Genome Research*, *23*(7). https://doi.org/10.1101/gr.157602.113

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Springer.

Wilke C, Wiernik B (2022). _ggtext: Improved Text Rendering Support for 'ggplot2'_. R package version 0.1.2, <https://CRAN.R-project.org/package=ggtext>

Willis, A., Jung, E. J., Wakefield, T., & Chen, X. (2004). Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene*, *23*(13). https://doi.org/10.1038/sj.onc.1207396

Woodward, E. L., Yang, M., Moura-Castro, L. H., van den Bos, H., Gunnarsson, R., Olsson-Arvidsson, L., Spierings, D. C. J., Castor, A., Duployez, N., Zaliova, M., Zuna, J., Johansson, B., Foijer, F., & Paulsson, K. (2023). Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukaemia. *Nature Communications*, *14*(1), 1658.

Worrall, J. T., Tamura, N., Mazzagatti, A., Shaikh, N., van Lingen, T., Bakker, B., Spierings, D. C. J., Vladimirou, E., Foijer, F., & McClelland, S. E. (2018). Non-random Mis-segregation of Human Chromosomes. *Cell Reports*, *23*(11), 3366–3380.

Wu, C.-H., Hsieh, C.-S., Chang, Y.-C., Huang, C.-C., Yeh, H.-T., Hou, M.-F., Chung, Y.-C., Tu, S.-H., Chang, K.-J., Chattopadhyay, A., Lai, L.-C., Lu, T.-P., Li, Y.-H., Tsai, M.-H., & Chuang, E. Y. (2021). Differential whole-genome doubling and homologous recombination deficiencies across breast cancer subtypes from the Taiwanese population. *Communications Biology*, *4*(1), 1052.

Xiao, M., Wang, F., Chen, N., Zhang, H., Cao, J., Yu, Y., Zhao, B., Ji, J., Xu, P., Li, L., Shen, L., Lin, X., & Feng, X.-H. (2024). Smad4 sequestered in SFPQ condensates prevents TGF-ß tumor-suppressive signaling. *Developmental Cell*, *59*(1), 48–63.e8.

Xue, W., Kitzing, T., Roessler, S., Zuber, J., Krasnitz, A., Schultz, N., Revill, K., Weissmueller, S., Rappaport, A. R., Simon, J., Zhang, J., Luo, W., Hicks, J., Zender, L., Wang, X. W., Powers, S., Wigler, M., & Lowe, S. W. (2012). A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(21), 8212–8217.

Yang, J., Lee, J., Land, M. A., Lai, S., Igoshin, O. A., & St-Pierre, F. (2021). A synthetic circuit for buffering gene dosage variation between individual mammalian cells. *Nature Communications*, *12*(1), 1–13.

Yates, L. R., Knappskog, S., Wedge, D., Farmery, J. H. R., Gonzalez, S., Martincorena, I., Alexandrov, L. B., Van Loo, P., Haugland, H. K., Lilleng, P. K., Gundem, G., Gerstung, M., Pappaemmanuil, E., Gazinska, P., Bhosle, S. G., Jones, D., Raine, K., Mudie, L., Latimer, C., … Campbell, P. J. (2017). Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell*, *32*(2). https://doi.org/10.1016/j.ccell.2017.07.005

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B., & Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, *4*, 2612.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). *clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters*. https://doi.org/10.1089/omi.2011.0118

Yu G (2023). _ggplotify: Convert Plot to 'grob' or 'ggplot' Object_. R package version 0.1.2, <https://CRAN.R-project.org/package=ggplotify>.

Zabarovsky, E. R., Lerman, M. I., & Minna, J. D. (2002). Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene*, *21*(45), 6915–6935.

Zaccaria, S., & Raphael, B. J. (2020). Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature Communications*, *11*(1), 1–13.

Zapata, L., Pich, O., Serrano, L., Kondrashov, F. A., Ossowski, S., & Schaefer, M. H. (2018). Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biology*, *19*(1), 67.

Zapatka, M., Borozan, I., Brewer, D. S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sültmann, H., Moch, H., Cooper, C. S., Eils, R., Ferretti, V., & Lichter, P. (2020). The landscape of viral associations in human cancers. *Nature Genetics*, *52*(3), 320–330.

Zeng, J., Hills, S. A., Ozono, E., & Diffley, J. F. X. (2023). Cyclin E-induced replicative stress drives p53-dependent whole-genome duplication. *Cell*, *186*(3). https://doi.org/10.1016/j.cell.2022.12.036

Zhang, C., Jiang, M., Zhou, N., Hou, H., Li, T., Yu, H., Tan, Y.-D., & Zhang, X. (2021). Use tumor suppressor genes as biomarkers for

diagnosis of non-small cell lung cancer. *Scientific Reports*, *11*(1), 1–13.

Zhang, J., Wang, X., Cui, W., Wang, W., Zhang, H., Liu, L., Zhang, Z., Li, Z., Ying, G., Zhang, N., & Li, B. (2013). Visualization of caspase-3-like activity in cells using a genetically encoded fluorescent biosensor activated by protein cleavage. *Nature Communications*, *4*, 2157.

Zhang, Y., Li, Y., Li, T., Shen, X., Zhu, T., Tao, Y., Li, X., Wang, D., Ma, Q., Hu, Z., Liu, J., Ruan, J., Cai, J., Wang, H. Y., & Lu, X. (2019). Genetic Load and Potential Mutational Meltdown in Cancer Cell Populations. *Molecular Biology and Evolution*, *36*(3). https://doi.org/10.1093/molbev/msy231

Zhang, Z., Feng, X., Deng, Z., Cheng, J., Wang, Y., Zhao, M., Zhao, Y., He, S., & Huang, Q. (2021). Irradiation-induced polyploid giant cancer cells are involved in tumor cell repopulation via neosis. *Molecular Oncology*, *15*(8), 2219–2234.

Zhang, Z., Hernandez, K., Savage, J., Li, S., Miller, D., Agrawal, S., Ortuno, F., Staudt, L. M., Heath, A., & Grossman, R. L. (2021). Uniform genomic data analysis in the NCI Genomic Data Commons. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21254-9

Zhao, B., Lin, J., Rong, L., Wu, S., Deng, Z., Fatkhutdinov, N., Zundell, J., Fukumoto, T., Liu, Q., Kossenkov, A., Jean, S., Cadungog, M. G., Borowsky, M. E., Drapkin, R., Lieberman, P. M., Abate-Shen, C. T., & Zhang, R. (2019). ARID1A promotes genomic stability through protecting telomere cohesion. *Nature Communications*, *10*(1), 4067.

Zhao, H., & Baudis, M. (2024). labelSeg: segment annotation for tumor

copy number alteration profiles. *Briefings in Bioinformatics*, *25*(2). https://doi.org/10.1093/bib/bbad541

Zhao, R., Choi, B. Y., Lee, M.-H., Bode, A. M., & Dong, Z. (2016). Implications of Genetic and Epigenetic Alterations of CDKN2A (p16(INK4a)) in Cancer. *EBioMedicine*, *8*, 30–39.

Zheng, S., Cherniack, A. D., Dewal, N., Moffitt, R. A., Danilova, L., Murray, B. A., Lerario, A. M., Else, T., Knijnenburg, T. A., Ciriello, G., Kim, S., Assie, G., Morozova, O., Akbani, R., Shih, J., Hoadley, K. A., Choueiri, T. K., Waldmann, J., Mete, O., … Verhaak, R. G. W. (2016). Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. *Cancer Cell*, *29*(5), 723–736.

Žilina, O., Reimand, T., Tammur, P., Tillmann, V., Kurg, A., & Õunap, K. (2013). Patient with dup(5)(q35.2-q35.3) reciprocal to the common Sotos syndrome deletion and review of the literature. *European Journal of Medical Genetics*, *56*(4). https://doi.org/10.1016/j.ejmg.2013.01.008

Zimmer, F., Harrison, P. W., Dessimoz, C., & Mank, J. E. (2016). Compensation of Dosage-Sensitive Genes on the Chicken Z Chromosome. *Genome Biology and Evolution*, *8*(4), 1233–1242.

Zucker, M., Perry, M. A., Gould, S. I., Elkrief, A., Safonov, A., Thummalapalli, R., Mehine, M., Chakravarty, D., Brannon, A. R., Ladanyi, M., Razavi, P., Donoghue, M. T. A., Murciano-Goroff, Y. R., Grigoriadis, K., McGranahan, N., Jamal-Hanjani, M., Swanton, C., Chen, Y., Shen, R., … Bandlamudi, C. (2024). Pan-cancer analysis of biallelic inactivation in tumor suppressor genes identifies KEAP1 zygosity as a

predictive biomarker in lung cancer. *Cell*.

https://doi.org/10.1016/j.cell.2024.11.010