Bernardo Lucas Carvalho Pereira de Almeida

Discovery of novel mechanisms of centrosome amplification and their therapeutic value in cancer

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2017

Bernardo Lucas Carvalho Pereira de Almeida

Discovery of novel mechanisms of centrosome amplification and their therapeutic value in cancer

Master in Oncobiology – Molecular Mechanisms of Cancer

This work was done under the supervision of:

Nuno Barbosa Morais, Ph.D

Ana Teresa Maia, Ph.D

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2017

Discovery of novel mechanisms of centrosome amplification and their therapeutic value in cancer

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam na listagem de referências incluída.

I declare that I am the author of this work, that is original and unpublished. Authors and works consulted are properly cited in the text and included in the list of references."

(Bernardo Almeida)

"The people who are crazy enough to think they can change the world are the ones who do!"

Steve Jobs

AGRADECIMENTOS

A realização desta dissertação marca o fim de mais uma etapa académica, uma etapa marcante na minha vida científica e que, certamente, jamais esquecerei. No entanto, tudo isto não teria sido possível sem o apoio e amizade de inúmeras pessoas, às quais deixo um sentido e profundo agradecimento.

Um sincero obrigado ao Doutor Nuno Barbosa Morais, meu mentor e inspiração ao longo deste período, por me ter acolhido no seu grupo, ter confiado em mim, me ter proporcionado todos os meios para que eu me levasse ao limite, e, já agora, por me ter ajudado a chegar lá. Pela sua excelência e honestidade científica, pela exigência e profissionalismo que sempre demonstrou, pela independência que me proporcionou, mas pelo também apoio incondicional, o meu sincero agradecimento, será para sempre um exemplo a seguir.

Deixo um agradecimento muito especial também às duas colaboradoras neste projeto, Doutora Mónica Bettencourt Dias e Doutora Gaëlle Marteil, pela total disponibilidade ao longo deste período, por todas as discussões científicas que proporcionaram, e por me fazerem sentir parte integrante do seu grupo de investigação. Que a colaboração se mantenha e tenha muito sucesso.

Não poderia deixar de realçar duas pessoas marcantes neste meu percurso científico. À Professora Doutora Ana Teresa Maia, também minha coorientadora neste projeto, e à Doutora Joana Xavier, um muito obrigado por me terem introduzido "neste mundo" e pelo contínuo apoio e orientação.

Aos meus incríveis colegas, Lina, Marie, Teresa, Nuno, Mariana, Carolina e Luís, um muito obrigado pela forma como me acolheram, pela amizade e apoio, e pelas intensas trocas de ideias e conhecimentos. Não poderia pedir um melhor ambiente para trabalhar. Desejo-vos o maior sucesso do mundo, vocês merecem. Estendo também este agradecimento a todos os colegas do Instituto de Medicina Molecular, em particular aos jogadores da bola.

Muito obrigado a todos os meus amigos pela amizade e compreensão incondicional. Em especial à minha namorada, Sara Daniela, por ter aturado o meu entusiasmo ao longo deste projeto, pelo amor e ternura sempre demonstrados, e por ter vivido esta etapa ao meu lado.

O meu último, e mais importante, agradecimento é para toda a minha família, em particular pais, avós e irmão, pelo carinho e confiança, pelo constante sacrifício que fizeram para que eu pudesse chegar até aqui, e por fazerem de mim a pessoa que sou hoje.

ABSTRACT

Genomic instability is a hallmark of cancer cells that generates the genetic diversity that makes possible the acquisition of all the other hallmarks. Thus, the maintenance of genome stability is critical for proper cell function. Centrosomes, the major microtubule-organising centres of animal cells, are the main subcellular organelles implicated in the maintenance of genome stability. It is therefore not surprising that centrosome amplification (CA) – the presence of more than one centrosome in a cell – is a common feature in cancer. Recent work from the Bettencourt-Dias Lab has identified a new recurrent feature of cancer cells: centriole over-elongation (COE), which also promotes CA. Those abnormalities are specific features of cancer cells and hence appealing targets in cancer therapy. However, their origins and therapeutic value remain poorly understood, preventing their use in the clinic.

We have screened the NCI-60 panel of human cancer cell lines for centriole number and individual length to test their frequency and interdependence. We have thereby also generated a metric capturing each abnormality level per cell line that we then correlated with the publicly available molecular (particularly transcriptomic and proteomic) and drug-sensitivity quantitative profiles for that panel.

Our work showed lower frequency of COE compared to CA and lung and skin as the primary cancer tissues with higher centriole length heterogeneity. However, the two features are not independent, with overly-longer centrioles being more common in cells with CA. Our single-cell analyses have also suggested that cells apparently do not control their overall centriolar mass when the centriole number increases. Moreover, cancer cell lines with longer centrioles proliferated slower due to an accumulation of cells in G1 phase, suggesting that centriole length defects could lead to a cell cycle delay in G1. In addition, our original genome-wide approach highlighted putative mechanisms associated with both abnormalities in cancer, such as the PRKACA kinase promoting COE and the proteasome protecting cells from CA. Correlation with drug activity have both associated CA with higher sensitivity to compound activity and also identified some compounds as potential therapeutic options to selectively target cells with higher incidence of centriole abnormalities.

This work provides the first single-centriole-level portrait of centriole abnormalities in cancer and contributes to the understanding of their molecular origins, namely by revealing novel molecular mechanisms in cell cycle biology. Given the cancer-specificity of these abnormalities, the identified compounds will inspire the development of clinical applications based on selectively targeting these Achilles' heels of cancer cells.

Keywords: cancer; genomic instability; NCI-60 panel; centrosome amplification; centriole over-elongation; targeted therapies.

RESUMO

O cancro é um grupo de doenças complexas e heterogéneas caracterizadas por uma proliferação celular descontrolada. Uma propriedade intrínseca das células de cancro é a instabilidade genómica, cuja presença origina a diversidade genética que lhes possibilita a aquisição de outras propriedades cancerígenas. Deste modo, o controlo da estabilidade genómica é fundamental para um normal funcionamento celular.

Os centrossomas, constituídos por dois centríolos, são centros organizadores de microtúbulos nas células animais e os principais organelos celulares envolvidos na manutenção da estabilidade genómica. Não é, portanto, surpreendente que a amplificação centrossomal (CA) – a presença de mais do que um centrossoma numa célula – seja uma característica comum em cancro. De fato, a CA está descrita como o principal mecanismo subjacente ao desenvolvimento de mitoses multipolares, e consequente instabilidade genómica, e foi recentemente observada como sendo suficiente para provocar tumorigénese. Trabalho recente do grupo liderado pela investigadora Dr. Mónica Bettencourt-Dias identificou uma nova característica frequente em células cancerígenas: sobre-elongação centriolar (COE), a qual também promove CA. Ambas as anormalidades centriolares são específicas das células cancerígenas e, portanto, alvos promissores para o tratamento do cancro. No entanto, as suas origens e valor terapêutico continuam por explorar, impedindo o seu uso na prática clínica.

Para investigar estas questões, o grupo da Dr. Mónica Bettencourt-Dias quantificou o número e comprimento dos centríolos por célula num painel de 60 linhas celulares derivadas de diferentes cancros humanos (painel NCI-60, de acesso público). Este painel de linhas celulares foi, num estudo prévio, testado quanto à sua sensibilidade a drogas e extensamente caracterizado aos níveis genómico, transcriptómico e proteómico, entre outros. A integração dos dados resultantes de ambos os estudos proporciona um recurso sem precedentes para estudar anormalidades centriolares em cancro.

O presente projeto de tese teve como principais objetivos 1) caracterizar as anormalidades centriolares ao longo do painel NCI-60, 2) explorar as origens moleculares de tais anormalidades e 3) identificar novos compostos que tenham como alvo a CA.

Com vista a alcançar os objetivos propostos, começou-se por testar a frequência e interdependência de CA e COE no painel, tendo depois a sua frequência sido utilizada para gerar métricas que caracterizassem o grau de cada anormalidade por linha celular. Estas métricas foram depois correlacionadas com os diferentes perfis moleculares (nomeadamente transcriptómico e proteómico) e de sensibilidade às drogas disponíveis para aquele painel.

As nossas análises no painel NCI-60 confirmaram CA e COE como características prevalentes em linhas celulares de cancro associadas a diferentes tecidos de origem. Foi observada uma maior frequência de anormalidades no número de centríolos, comparativamente com anormalidades no seu comprimento, tendo CA sido também associada com os subtipos mais agressivos de cancro da mama e do cólon. Aproveitando a resolução ao nível de centríolos individuais, observou-se grande variabilidade tanto no número como no comprimento de centríolos por célula entre os diferentes tecidos de cancro, representados neste estudo por linhas celulares deles derivadas. Para além disso, verificou-se uma elevada heterogeneidade no comprimento centriolar entre linhas celulares com o mesmo tecido de origem, particularmente, com origem no pulmão e na pele.

A associação entre COE e CA foi validada no painel NCI-60 tanto ao nível de população de células, onde ambas as anormalidades estavam positivamente correlacionadas, como ao nível de centríolos individuais, onde foi observada uma maior proporção de centríolos longos em células com CA. Comparações entre a distribuição observada do número de centríolos longos em cada célula e a esperada, caso COE fosse um evento estocástico, revelaram que a COE é dependente do estado fisiológico da célula. Ademais, análises de células individuais de cancro revelaram que estas não controlam a sua massa centriolar quando o número de centríolos aumenta.

A associação entre as anormalidades centriolares e a proliferação celular foi avaliada usando o tempo de duplicação da população de células e a proporção de células em cada fase do ciclo celular. Linhas celulares com centríolos longos apresentaram uma menor taxa de proliferação devido a uma acumulação de células na fase G1, sugerindo que a COE pode levar a um atraso no ciclo celular em G1.

Relativamente aos mecanismos moleculares associados com a COE, análises de correlação entre a prevalência desta característica e a expressão, aos níveis transcricional e proteico, de virtualmente todos os genes presentes no genoma, juntamente com resultados de um estudo independente de potenciais reguladores do comprimento centriolar, revelaram a proteína PRKACA como potencial promotora de COE em cancro. Esta proteína codifica uma subunidade catalítica da quinase PKA, cuja localização celular é maioritariamente ao nível do centrossoma. Elevados níveis da proteína PKA foram previamente associados com comprimento elevado do cílio primário, mas não é conhecida nenhuma associação entre esta quinase e o comprimento dos centríolos, pelo que a associação encontrada em ambos os estudos necessita de validação experimental. Para além disso, as análises de correlação mencionadas acima identificaram vias de sinalização de interação com a matriz extracelular positivamente

associadas com COE. Em sentido oposto, a sub-expressão de genes envolvidos nos mecanismos de reparação do DNA foi associada com um aumento do comprimento dos centríolos.

Os mecanismos moleculares de CA em cancro foram também investigados através de análises de correlação entre a frequência desta característica e a expressão génica, das quais o proteassoma despontou como o principal candidato a explicar os níveis de CA observados nas linhas celulares do painel NCI-60: diversos genes que codificam para os diferentes componentes do proteassoma estavam negativamente associados com o número de centríolos por célula. Em particular, a proteína PSMD1, responsável pelo reconhecimento e ligação aos substratos, foi a que apresentou uma associação negativa mais forte. Estudos anteriores indicaram que o proteassoma está maioritariamente localizado no centrossoma, sendo responsável por controlar os níveis de várias proteínas centrossomais, incluindo as principais participantes na duplicação centriolar. Para além disso, estudos em *Drosophila* demonstraram que inibição do proteassoma induz CA. As observações no NCI-60 são concordantes com esta hipótese, na qual o proteassoma é proposto como um mecanismo de proteção das células para com a CA. De realçar, no entanto, que esta é a primeira evidência deste mecanismo em contexto de cancro, sugerindo um novo mecanismo molecular para a origem da CA.

Para além dos perfis moleculares das linhas celulares do painel NCI-60, estão disponíveis os seus perfis quantitativos de sensibilidade a cerca de 50,000 compostos. Estes perfis foram primeiramente utilizados com vista a investigar se a CA está globalmente associada a maior sensibilidade ou resistência a drogas. De fato, foi observada uma forte associação entre CA e uma maior sensibilidade às drogas, realçando o seu potencial como futuro alvo terapêutico. De seguida, análises de correlação entre a sensibilidade a estes compostos e os níveis de CA, bem como os de expressão do gene *PLK4*, gene que codifica para a principal proteína reguladora da duplicação centriolar, identificaram alguns compostos como potenciais opções terapêuticas para eliminar seletivamente as células com elevados níveis de CA e *PLK4*, respetivamente.

Este estudo apresenta a primeira caracterização, ao nível de centríolos individuais, de anormalidades centriolares em cancro e contribui para uma melhor compreensão das suas origens moleculares, revelando concomitantemente novos mecanismos biológicos envolvidos no ciclo celular. Visto que aquelas anormalidades são específicas das células de cancro, os compostos identificados neste projeto irão inspirar o desenvolvimento de novas terapias direcionadas em oncologia.

Palavras-chave: cancro; instabilidade genómica; painel NCI-60; amplificação centrossomal; sobre-elongação centriolar; terapias direcionadas.

INDEX OF CONTENTS

AGRADECIMENTOS vii

ABSTRACT ix

RESUMO xi

INDEX OF FIGURES xix

INDEX OF TABLES xxi

INDEX OF ANNEXES xxiii

LIST OF ABREVIATIONS xxv

CHAPTER 1 – INTRODUCTION 1

1.1 Cancer 1

1.1.1 Models of cancer development 1

1.1.1.1 Driver and passenger mutations 3

1.1.2 Epidemiology 3

1.1.2.1 Cancer: a disease of older people 4

1.1.2.2 Different types of cancer 5

1.1.3 Hallmarks of cancer 5

1.1.4 Genome instability and mutation 7

1.2 Centrosomes 7

1.2.1 Centrosome cycle 8

1.2.2 Centrosome amplification and cancer 10

1.2.2.1 Origins of centrosome amplification 11

1.2.2.2 Coping with centrosome amplification 12

1.2.2.3 Centrosome-targeting cancer therapies 13

1.3 Screen of centriole number and structure in the NCI-60 panel 14

CHAPTER 2 – AIMS 15

CHAPTER 3 – MATERIALS AND METHODS 17

3.1 NCI-60 panel of human cancer cell lines 17

3.1.1 Screen of centriole number and length 17

3.1.1.1 Centriole abnormality metrics 19

3.1.1.2 Centriolar mass 19

3.1.2 Centriole length regulators screen 19
3.1.3 Flow Cytometry analysis of cell cycle phases 20
3.1.4 Molecular and pharmacological data sets 21
3.1.4.1 Gene expression 21
3.1.4.2 Protein expression 22
3.1.4.3 Drug sensitivity 22
3.2 Transcriptomic alterations associated with centriole abnormalities 23
3.2.1 Centrosomal genes 23
3.2.2 Gene Set Enrichment Analyses 23
3.2.3 Linear regression models to decouple independent effects 24
3.3 Statistical hypothesis testing 25
3.3.1 Unpaired and paired two-sample statistical tests 26
3.3.2 Kruskal-Wallis rank sum test 27
3.3.3 Fligner-Killeen test of homogeneity of variances 27
3.3.4 Unsupervised hierarchical clustering 27
3.3.5 Spearman's rank correlation 27
3.3.6 Analyses of covariance (ANCOVA) 28
3.3.7 Pearson's chi-squared test 29
3.3.7.1 Chi-squared test of independence 29
3.3.7.2 Chi-squared goodness of fit test 29
3.3.7.3 Generate a random distribution of COE per cell 29
3.3.8 Binomial test 30
3.3.9 Correction for multiple testing 30

CHAPTER 4 – RESULTS 33
4.1 Profile of centriole abnormalities in the NCI-60 panel 33
4.1.1 Single-cell and single-centriole heterogeneity in cancer 34
4.1.1.1 Centriole number 35
4.1.1.2 Centriole length 38
4.1.2 Aggressive breast and colon cancer cell lines display high levels of CA 40
4.1.3 COE and CA are not independent 40
4.1.4 COE is not a stochastic event in cancer cells 43
4.1.5 COE is cell state-dependent 45
4.1.6 Total centriolar mass per cell is apparently not controlled 47

4.2 Discovery of novel molecular origins of centriole abnormalities in cancer 50
4.2.1 Cancer cell lines with overly-long centrioles have more cells in G1 50
4.2.2 PRKACA is a putative promoter of COE in cancer 53
4.2.2.1 Increased centriole length is associated with higher interaction with ECM and lower efficiency in DNA repair.....55
4.2.3 Cancer cells with less proteasome activity are more susceptible to CA 57
4.3 Identification of new compounds that target CA 60
4.3.1 CA is associated with higher sensitivity to compound activity 60
4.3.2 Compounds that selectively kill cancer cells with CA 61
4.3.3 Compounds that target CA-associated proteins 62

CHAPTER 5 – DISCUSSION 65
5.1 NCI-60 profile of centriole abnormalities.....65
5.2 Novel molecular mechanisms underlying centriole abnormalities in cancer 68
5.3 Therapeutic value of CA in cancer 72

CHAPTER 6 – CONCLUSION 75

BIBLIOGRAPHY 77

ANNEXES 89

INDEX OF FIGURES

Figure 1.1 Clonal evolution in cancer 2

Figure 1.2 Association between the year where cancer overtook CVD as the leading cause of death and the current ratio of cancer to CVD deaths 4

Figure 1.3 Cancer incidence increases with age 5

Figure 1.4 The hallmarks of cancer 6

Figure 1.5 The centrosome 8

Figure 1.6 The centrosome duplication cycle 9

Figure 1.7 PLK4 overexpression leads to centrosome amplification 10

Figure 3.1 NCI-60 panel of human cancer cell lines 17

Figure 3.2 Overview of the secondary screening of centriole alterations 18

Figure 3.3 Overview of the centriole length regulators screen 20

Figure 3.4 Distribution of cells along the cell cycle phases 21

Figure 3.5 Overview of Gene Set Enrichment Analyses 24

Figure 3.6 Comparing two regression slopes with ANCOVA 28

Figure 3.7 Generate a random distribution of COE per cell 30

Figure 4.1 Profile of centriole abnormalities in the NCI-60 panel 34

Figure 4.2 Distribution of the number of centrioles per cell across NCI-60 tissues of origin 35

Figure 4.3 Centriole number heterogeneity in cancer 37

Figure 4.4 Centriole length heterogeneity in cancer 38

Figure 4.5 Single-centriole length heterogeneity in individual cell lines for selected tissues of origin 39

Figure 4.6 NCI-60 cell lines from aggressive breast and colon cancer subtypes display high levels of centrosome amplification 40

Figure 4.7 Centriole over-elongation was positively correlated with centrosome amplification 41

Figure 4.8 Distribution of centriole length according to the number of centrioles per cell 42

Figure 4.9 Centriole over-elongation and centrosome amplification are not independent 43

Figure 4.10 Centriole length dysregulation does not affect all centrioles within the same cell 44

Figure 4.11 Centriole over-elongation is not a stochastic event in cancer cells 45

Figure 4.12 Centriole over-elongation is a cell state-dependent feature 47

Figure 4.13 Centriolar mass increased in association with the number of centrioles within the cell, while centriole length mean was maintained 49

Figure 4.14 Cancer cell lines with overly-long centrioles had lower proliferation rates 51

Figure 4.15 Cancer cell lines with overly-long centrioles have more cells in G1 52

Figure 4.16 TP53 status was not associated with the relation between centriole length defects and cell cycle delay in G1 53

Figure 4.17 PRKACA gene expression levels were positively correlated with centriole length 54

Figure 4.18 PRKACA knock down decreased centriole length.....55

Figure 4.19 KEGG pathways associated with centriole length abnormalities in cancer 56

Figure 4.20 Examples of KEGG pathways associated with the mean of centriole length in cancer 57

Figure 4.21 PSMD1 protein levels were negatively correlated with centriole number 58

Figure 4.22 Lower expression of proteasome components was associated with increased centriole number levels 59

Figure 4.23 Enrichment of positive correlations between compound activity and centrosome amplification prevalence 61

Figure 4.24 Correlation between activity of top compound NSC 633109 (z score) and percentage of cells with centrosome amplification.....62

Figure 4.25 Correlation between activity of top compound NSC 658364 (z score) and PLK4 transcript levels 63

Figure 5.1 Hypothesized portrait of the COE-CA association in cancer 67

Figure 5.2 Hypothesis of centriole over-elongation origins and consequences in the NCI-60 panel.....70

Figure 5.3 Hypothesis of centrosome amplification origin in the NCI-60 panel.....72

INDEX OF TABLES

Table 4.1 Compounds that have higher activity in cell lines with higher incidence of centrosome amplification.....61

Table 4.2 Compounds that target the centrosome amplification-associated protein PLK4 63

INDEX OF ANNEXES

Annex 1 Table with centriole abnormality metrics for each NCI-60 cell line 89

Annex 2 Correlation plot between cell lines' doubling time estimated by the U.S. National Cancer Institute and calculated on O'Connor et al., 1997 90

Annex 3 Distribution of gene expression variance across NCI60 samples 91

Annex 4 Distribution of protein expression variance across NCI60 samples 92

Annex 5 Distribution of drug activity variance across NCI60 samples 93

Annex 6 Tissue hierarchical clustering based on centriole heterogeneity 94

LIST OF ABREVIATIONS

3D - three-dimensions
AMP - adenosine monophosphate
ANCOVA - analysis of covariance
ANOVA - analysis of variance
ATP - Adenosine triphosphate
C - catalytic
CA - centrosome amplification
cAMP - cyclic AMP
CC - centrosome clustering
cDNA - complementary DNA
CIN - chromosomal instability
CNS - central nervous system
COE - centriole over-elongation
CSC - cancer stem cell
CVD - cardiovascular diseases
DNA - deoxyribonucleic acid
EACR - European Association for Cancer Research
ECM - extracellular matrix
ES - enrichment score
FACS - fluorescence-activated cell sorting
FDA - Food and Drug Administration
FDR - false discovery rate
GI50 - compound's concentration that causes 50% growth inhibition
GSEA - gene set enrichment analyses
HPV - human papillomavirus
mRNA - messenger RNA
MS - mass spectrometry
MSI-H - microsatellite instability
MT - mutated
MTOC - microtubule-organising centre
n.s. - not significant
NCI-60 - U.S. National Cancer Institute panel of 60 human cancer cell lines

NES - normalized enrichment score

nm - nanometre

PCM - pericentriolar material

R - regulatory

RNA - ribonucleic acid

RNAi - RNA interference

SID - substance accession Identifier

siRNAs - small interfering RNAs

SMILE - simplified molecular input line entry

UK - United Kingdom

UV - ultraviolet

WT - wild-type

CHAPTER 1 – INTRODUCTION

1.1 Cancer

Cancer is a group of complex and heterogeneous diseases characterized by uncontrolled proliferation of cells that can invade surrounding tissues and spread to distant organs, i.e. metastasize (Strachan and Read, 1996). This cellular condition is mainly generated by the loss of normal growth control that results from the accumulation of genetic alterations over time with, in some cases, also an inherited predisposition. Thus, cancer can be seen as a disease of the genome (Garraway and Lander, 2013; Macconaill and Garraway, 2015).

All cancers begin with defective cells. Normal cells have genes that directly or indirectly control cell proliferation. However, sometimes a change happens in the DNA sequence of those genes – called a genetic alteration – and abnormal cells lose the ability to control their proliferation. Mutations, defined as permanent alterations in the DNA sequence, are the most common alterations observed in cancer and they can arise by chance in proliferating cells, be it caused by the natural processes in our cells or by environmental perturbations such as tobacco smoke or UV radiation (Nowak and Waclaw, 2017; Tomasetti and Vogelstein, 2015). These mutations that occur at certain point during a person's life and are present only in certain cells are called somatic mutations (Campbell, 2016; Greenman et al., 2007). Some people can also inherit alterations (hereditary mutations) in particular genes that confer higher susceptibility to develop cancer (Nielsen et al., 2016; Rahman, 2014). Other common genetic alterations are chromosomal translocations, gene amplifications and gene deletions (Vogelstein et al., 2013).

Cancer-related genetic alterations occur mainly in four types of genes: oncogenes and tumour suppressor genes (genes that promote or inhibit cell division, respectively), DNA repair genes (that repair other damaged genes) and self-destruction genes (that promote cell death) (Garraway and Lander, 2013; Macconaill and Garraway, 2015).

1.1.1 Models of cancer development

Cancer starts with mutations in one cell or a small group of cells. If these changes confer proliferative advantage, the mutant clones will tend to overcome normal cells and take over the organism (Garraway and Lander, 2013; Greaves and Maley, 2012). Over time, sequential acquisition of mutations will generate sub-clone diversity, consequent cancer heterogeneity, and select the most capable/aggressive cancer cells. Thus, cancer can be seen as a natural evolutionary process (Nowell, 1976), a disease of Darwinian clonal evolution involving dynamic changes in the genome (Merlo et al., 2006, ***Figure 1.1***).

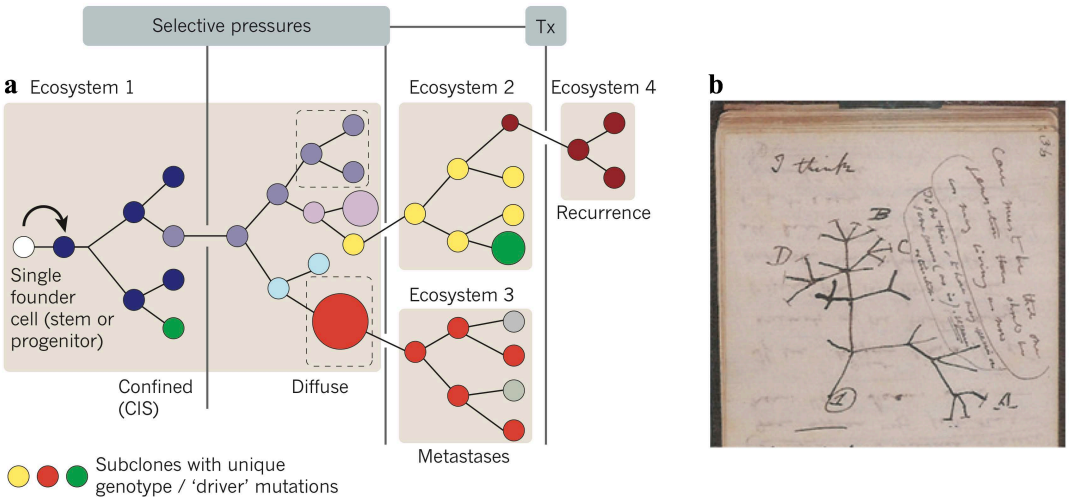


Figure 1.1 Clonal evolution in cancer. a) Different mutations create genetically distinct subclones (represented with different colours) which will compete when faced with some selective pressure (vertical lines), e.g. therapy (Tx). Selective pressure allows some subclones to expand while others become extinct and/or quiescent, selecting for the most aggressive cancer cells in each particular environment. Ecosystems 1–4 (boxes) represent the different tissue ecosystems and consequently different kinds of selective pressures, selecting for different subclones - the most capable for each ecosystem. CIS, carcinoma in situ. **b)** Branching evolutionary tree of speciation from Darwin's 1837 notebook (adapted from Greaves and Maley, 2012).

This clonal evolution model was proposed by Peter Nowell over 40 years ago (Nowell, 1976) and attempted to explain the origin of cancer and its heterogeneity - one of the greatest puzzles for cancer researchers. Although this evolutionary model is considered a *bona fide* scientific theory, surviving to 40 years of empirical observation and testing (Greaves and Maley, 2012), a new theory has emerged in the last years – the cancer stem cell (CSC) model (Kreso and Dick, 2014).

The CSC hypothesis was developed as a result of transplantation experiments with leukemic cells and proposes that cancer is clonally derived by small subpopulations of CSCs (Dick, 2008; Reya et al., 2001). The results of these experiments showed that mice were developing leukaemia only when they were injected with a specific group of cells – the leukemic stem cells. The same was already observed in solid cancers, with the first identification of CSCs achieved in human breast cancer over ten years ago (Al-Hajj et al., 2003). The CSC model suggests that many cancers may be hierarchically organized like normal tissues, where stem cells differentiate into diverse progeny with limited proliferative potential. Thus, to characterize and eliminate cancers that follow this model, it is necessary to focus on the small subpopulations of tumourigenic cells (Shackleton et al., 2009).

However, these models are not mutually exclusive, as CSCs and their progeny are also expected to evolve by clonal evolution. The existence of CSCs and the inherently Darwinian

character of cancer seem to be the main causes of therapeutic failure but perhaps also hold the key to more effective cancer control (Kreso and Dick, 2014; Shackleton et al., 2009)

1.1.1.1 Driver and passenger mutations

All cancers arise as a result of hereditary or somatically acquired mutations, but that does not mean that all changes present in a cancer genome are involved in the oncogenic process. Genetic mutations occur by chance with respect to adaptation but we can distinguish between driver and passenger mutations. A driver mutation is causally implicated in cancer development - it confers growth advantage to cancer cells, is positively selected along cancer lineages and, therefore, is usually observed in a greater proportion of cancer samples than what would be expected by chance. Passenger mutations are the ones that occur along the way but do not confer growth advantage, therefore not contributing to cancer development. These somatic mutations without functional consequences are the majority of alterations (Lawrence et al., 2013; Stratton et al., 2009).

1.1.2 Epidemiology

Cancer is one of the leading causes of morbidity and mortality worldwide, with 14.1 million new cancer cases, 8.2 million cancer deaths and 32.5 million people living with cancer (within 5 years of diagnosis) in 2012 (Ferlay et al., 2015). In Portugal, 49,174 people were diagnosed and 24,112 died from cancer, also in 2012 (Ferlay et al., 2013a).

Currently, cancer is still the second leading cause of death worldwide (Global Burden of Disease Cancer Collaboration, 2015) but it is predicted to overtake cardiovascular diseases (CVD) in the early future (Townsend et al., 2016). Indeed, this is already happening in Europe, with cancer being responsible for more deaths than CVD in 12 European countries (Townsend et al., 2016). Moreover, the sooner the overtaking happened, the higher is the current ratio of cancer to CVD deaths, suggesting this ratio keeps increasing even in those countries (Townsend et al., 2016; ***Figure 1.2***).

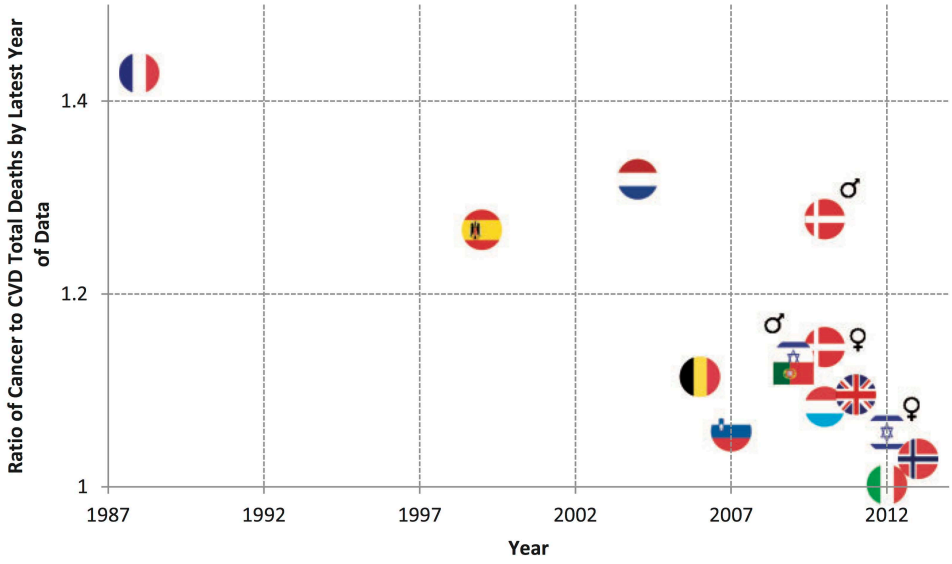


Figure 1.2 Association between the year where cancer overtook CVD as the leading cause of death and the current ratio of cancer to CVD deaths, by sex and European country. Data for countries without a gender symbol are for men only (adapted from Townsend et al., 2016).

The incidence of cancer is also a growing problem across the world, being expected to increase to 24 million new cancer cases/year by 2035 (Ferlay et al., 2013b). For instance, recent estimates from Cancer Research UK suggest that one in two British adults born after 1960 will develop cancer in their lifetime, with tendency to rise in future generations due to increasing life expectancy (Ahmad et al., 2015). What is behind this association?

1.1.2.1 Cancer: a disease of older people

The human body experiences about 10,000 trillion cell divisions in a lifetime, each of them with very small chance of error in DNA replication and segregation and consequent genetic/genomic alteration (Quammen, 2008). The development of cancer is a multistep process that usually requires two to eight mutational driver events to happen in the same cell (Vogelstein et al., 2013), which is very unlikely. However, with increasing age people have more time to acquire these genomic alterations and therefore the small chances add up, making cancer a disease of older people (Cancer Research UK; ***Figure 1.3***). Indeed in the UK, between 2012 and 2014, 50% of all diagnosed cancer cases were in people aged 70 and over (Cancer Research UK), whereas in the United States 86% were diagnosed in people 50 years of age or older in 2016 (American Cancer Society, 2016), confirming age as the greatest risk factor for cancer development.

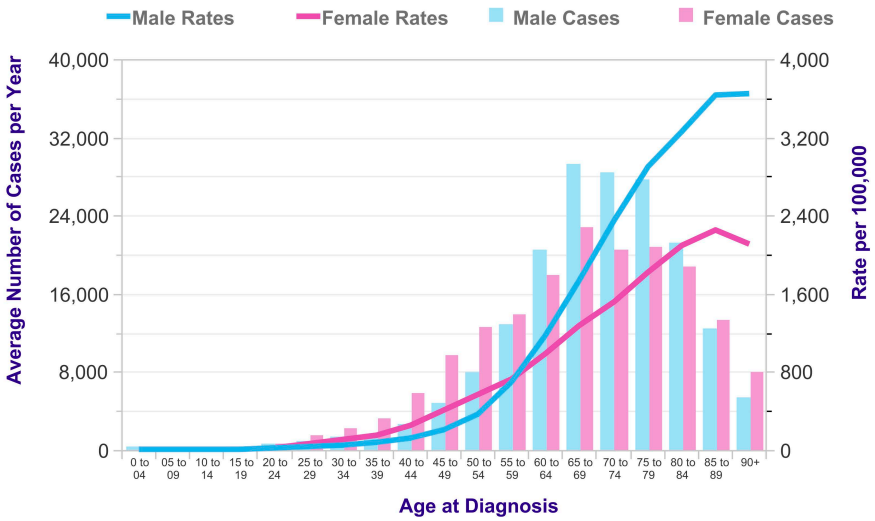


Figure 1.3 Cancer incidence increases with age. Average number of new cases per year and age-specific incidence rates per 100,000 inhabitants in the UK. Data for all cancers, excluding Non-Melanoma Skin Cancer, from 2012 to 2014 (adapted from Cancer Research UK).

1.1.2.2 Different types of cancer

Cancer can result from abnormal proliferation of any of the different types of human cells, so, in theory, there are over 200 different types of cancer in humans, with very different behaviour and response to treatment (Cooper, 2000). Indeed, the incidence and mortality varies a lot across cancers in different tissues.

The most common cancers arise from epithelial tissue cells and are called carcinomas. The tissues with higher incidence in 2012, according to GLOBOCAN 2012, were lung (1.82 million), breast (1.67 million), and colorectal (1.36 million). The most common causes of cancer death were lung (1.6 million), liver (745.000), and stomach cancer (723.000) (Ferlay et al., 2015).

Cancer incidence also varies between genders. The most common types of cancer in men are lung, prostate and colorectal cancer, while breast, colorectal, and cervix cancer are the most common among women (Ferlay et al., 2015).

1.1.3 Hallmarks of cancer

The transformation from a normal cell into an abnormal one, and therefore into cancer, is a complex and multistage process. The initial mass of abnormal cells is called a tumour, but not all tumours develop into a malignant/invasive one – denominated cancer. In 2000, Douglas Hanahan and Robert Weinberg published a seminal review that has influenced the study of cancer and the development of new therapeutics in oncology. They have proposed six fundamental properties (basic capabilities) that are acquired during the multistep development

of human tumours (tumourigenesis) and which are required for the development of cancer. These "hallmarks of cancer" include (Hanahan and Weinberg, 2000):

1. Sustaining proliferative signalling;
2. Evading growth suppressors;
3. Activating invasion and metastasis;
4. Enabling replicative immortality;
5. Inducing angiogenesis;
6. Resisting cell death.

However, care must be taken with the definition of a hallmark of cancer. A comment article in *Nature Reviews Cancer* in 2010 (Lazebnik, 2010) pointed out that five of the hallmarks were also characteristic of benign tumours, not classified as cancer because they do not metastasize to other parts of the body, and the only hallmark exclusive of cancer was the ability to invade and metastasize. Why were those five features considered to be in the same "league" as tissue invasion and metastasis? Well, the authors suggest that a potential and worrying explanation could be that "in many publications, including the article under discussion, the terms tumour and cancer are used interchangeably, perhaps because in the minds of many basic scientists these terms now mean the same thing" (Lazebnik, 2010).

In an update published in 2011 ("Hallmarks of cancer: the next generation"), Hanahan and Weinberg have revisited, refined, and extended this concept of cancer hallmarks proposing two new emerging hallmarks - deregulating cellular energetics and avoiding immune destruction - and two characteristics that enable the acquisition of all the previous hallmark capabilities: tumour-promoting inflammation and genome instability and mutation (Hanahan and Weinberg, 2011; ***Figure 1.4***).

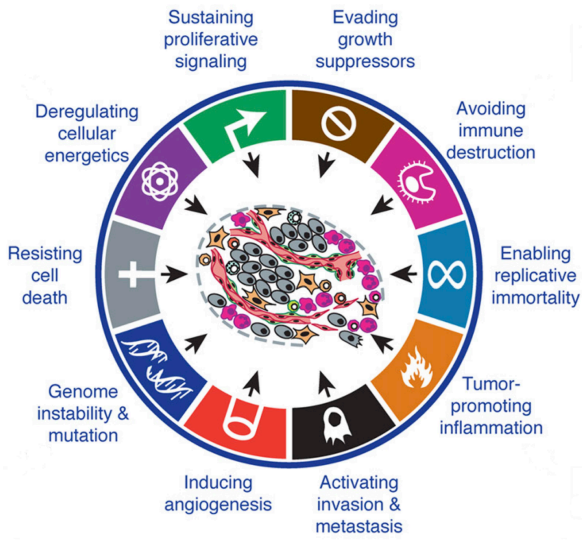


Figure 1.4 The hallmarks of cancer. Douglas Hanahan and Robert Weinberg have suggested ten functional capabilities acquired during cancer development, providing a useful conceptual framework for understanding the complex biology of this condition (adapted from Hanahan and Weinberg, 2011).

1.1.4 Genome instability and mutation

Genome instability and mutation is a particular hallmark, not being a functional capability of cancer *per se*, but a property that generates the genetic diversity that makes possible the acquisition of all the other hallmarks (Hanahan and Weinberg, 2011; Negrini et al., 2010). Supporting this view, tumours harbour too many mutations to be explained by anything other than underlying genomic instability (Sieber et al., 2003). Genomic instability is then defined as an increased propensity for DNA alterations, from single nucleotide to whole chromosome rearrangements, that usually is generated by compromising the surveillance systems that normally monitor genomic integrity (Pikor et al., 2013; Shen, 2011).

Hereditary cancers are often characterized by the presence of mutations in DNA repair genes, such as *BRCA1*, *BRCA2*, *MSH2* and *MYH*, which leads to genomic instability and consequent acquisition of mutations in oncogenes and tumour suppressor genes. Oppositely, in sporadic cancers the first alteration usually happens in oncogenes and tumour suppressor genes, promoting abnormal cell proliferation, with the resulting DNA replication stress (broadly defined as inefficient DNA replication characterized by DNA synthesis slow down and/or replication fork stalling) being responsible for the presence of genomic instability therein (Gaillard et al., 2015; Macheret and Halazonetis, 2015; Negrini et al., 2010). Genomic instability is therefore not only a hallmark of but also a driving force for tumourigenesis, promoting the acquisition of further DNA alterations, clonal evolution, and tumour heterogeneity (Sieber et al., 2003).

Thus, cells need to keep their genome unharmed for proper cell function, resorting in four main mechanisms: high-fidelity DNA replication in S-phase, precise chromosome segregation in mitosis, error free repair of sporadic DNA damage, and a coordinated cell cycle progression (Shen, 2011). One subcellular organelle in animal cells critically implicated in the maintenance of genome stability is the centrosome (Lerit and Poulton, 2016).

1.2 Centrosomes

Centrosomes were identified more than one century ago by Edouard Van Beneden (Van Beneden and Neyt, 1887) and, almost simultaneously, by Theodor Boveri (Boveri, 1887). These organelles are the major microtubule-organising centre (MTOC) in animal cells, hence being pivotal for several fundamental cellular processes, including signalling, cell polarity and migration (Bettencourt-Dias and Glover, 2007; Stevens et al., 2007; Vinogradova et al., 2012; Wang et al., 2009). Furthermore, centrosome function in the organization of the spindle poles during mitosis is crucial for chromosome segregation and successful cell division (Bettencourt-

Dias, 2013). Therefore, strict control of centrosome number and structure is critical to appropriate cell function.

Centrosomes are found in most animal cells and comprise two centrioles, a mother and a daughter, surrounded by a complex proteinaceous structure, the pericentriolar material (PCM), which confers the microtubule nucleation capacity (Bettencourt-Dias and Glover, 2007; Gould and Borisy, 1977; ***Figure 1.5***). Centrioles are small microtubule-based cylinders with a normal length ranging from 400 to 500 nm in human cycling cells. In addition to being the core centrosomal components, they function as basal bodies for the formation of cilia and flagella on the cell surface, which have crucial roles in physiology, development and disease (Badano et al., 2005; Carvalho-Santos et al., 2011; Praetorius and Spring, 2005). Normal cells in G0 or G1 phase of the cell cycle have a single centrosome that undergoes duplication once, and only once, in every cell cycle so that its number remains stable, like the genetic material of the cell (Nigg and Stearns, 2011).

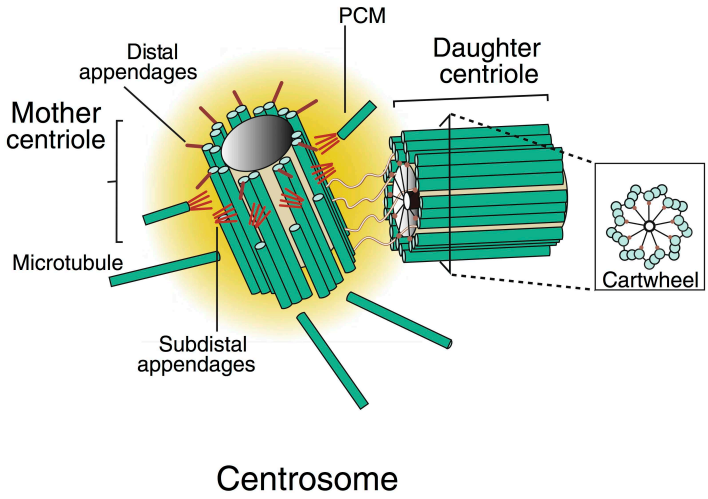


Figure 1.5 The centrosome. Centrosomes are composed by two, a mother (older) and a daughter (younger), centrioles, positioned in an orthogonal configuration, and surrounded by a complex proteinaceous structure called the pericentriolar material (PCM). The mother centriole shows subdistal and distal appendages where microtubules and the cell membrane are docked, respectively (adapted from Bettencourt-Dias, 2013).

1.2.1 Centrosome cycle

Centrosome duplication occurs at S phase in coordination with the cell cycle, where two new procentrioles (the future daughter centrioles) start forming adjacent to both the original mother centriole (now the grandmother centriole) and the original daughter centriole (now the mother centriole), and subsequently elongate until mitosis. There, the two newly formed centrosomes separate and migrate to opposite poles, allowing the bipolar spindle formation and faithful chromosome segregation. In the end of mitosis, centrioles disengage and, after cytokinesis, each of daughter cells has only one centrosome again, composed by one grandmother/mother (old) and one daughter (new) centriole (Bettencourt-Dias, 2013; ***Figure 1.6***). Interestingly, these centrioles differ in age and maturity and thus have different functions.

For instance, only the older centrioles recruit PCM and have centriolar appendages, and can hence initiate the assembly of the primary cilium, whereas daughter centrioles become competent only in the ensuing cell cycle (Hoyer-Fender, 2010). Moreover, the sister cell inheriting the grandmother usually grows the primary cilia faster (Anderson and Stearns, 2009).

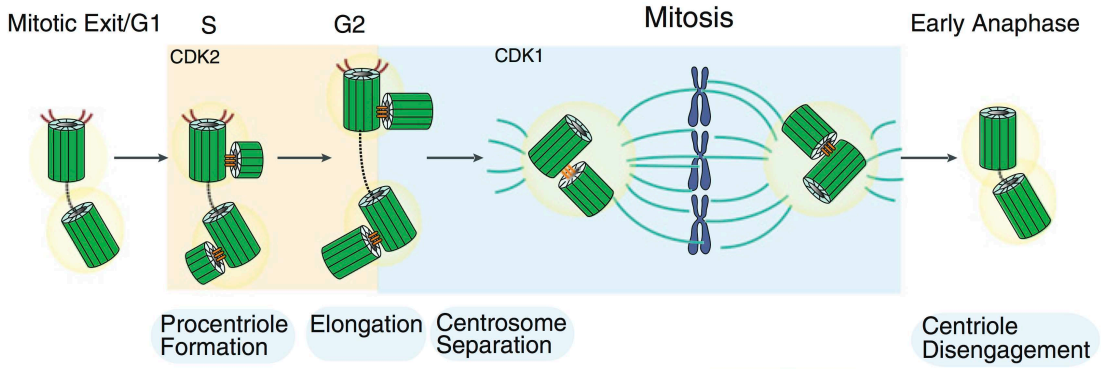


Figure 1.6 The centrosome duplication cycle. The centrosome cycle is tightly regulated in normal cells, to guarantee that each cell only has one centrosome. Procentriole formation begins in S phase, orthogonally to the proximal end of its mother centriole, and the subsequently elongation and maturation occurs during G2 phase. At mitosis, the newly formed centrosomes separate and direct the bipolar spindle assembly, segregating the chromosomes equally to the two daughter cells. Then the centrioles within each centrosome disengage, and after cytokinesis each of daughter cells will have one centrosome again, composed by one grandmother/mother (old) and one daughter (new) centriole (adapted from Bettencourt-Dias, 2013).

Recently, the advent of sensitive proteomics and RNA interference (RNAi)-based screens have proved invaluable in identifying and studying the critical components of the centrosome and its duplication cycle (Andersen et al., 2003; Balestra et al., 2013; Bauer et al., 2016; Jakobsen et al., 2011). Although these analyses have revealed hundreds of proteins and considerable complexity, forward genetic and RNAi screens in *C. elegans* uncovered just five proteins essential for procentriole formation (Strnad and Gönczy, 2008).

Among those proteins, the serine/threonine-protein kinase Polo-like Kinase 4 (PLK4, also known as SAK in *Drosophila melanogaster* and ZYG1 in *C. elegans*) has been demonstrated to be the master regulator of centriole biogenesis: in its absence centrioles fail to form, while its excess leads to centrosome amplification (CA) – the presence of more than one (or two, in mitosis) centrosome in a cell (Bettencourt-Dias et al., 2005; Habedanck et al., 2005; Kleylein-Sohn et al., 2007; O'Connell et al., 2001; Peel et al., 2007; Rodrigues-Martins et al., 2007; ***Figure 1.7a***). Proper centriole duplication is ensured by the regulation of PLK4 protein levels, mostly through SCF/Slimb ubiquitin-dependent proteolysis (Cunha-Ferreira et al., 2009; Rogers et al., 2009). Together with PLK4, a module comprising also the two proteins STIL, a substrate of PLK4, and SAS-6 has been shown to stay at the core of centriole duplication (Arquint and Nigg, 2016).

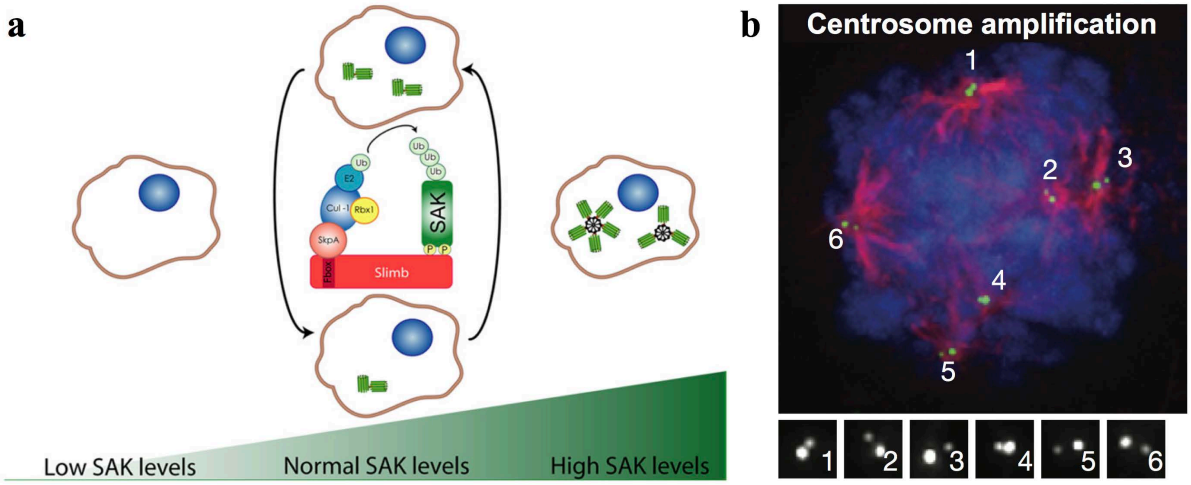


Figure 1.7 PLK4 overexpression leads to centrosome amplification. a) SAK/PLK4 is the master regulator of centriole biogenesis and therefore its protein levels must be controlled, which is done mostly through SCF/Slimb ubiquitin-dependent proteolysis (Cunha-Ferreira et al., 2009; Rogers et al., 2009). Low levels of SAK/PLK4 will lead to cells without centrosomes and, on the other hand, SAK/PLK4 accumulation, both by its overexpression or depletion of the SCF/Slimb complex, will result in centrosome amplification through simultaneous multiple procentriole formation (adapted from Cunha-Ferreira et al., 2009). **b)** Image illustrating centrosome amplification and consequent mitotic spindle multipolarity. Image of HeLa cell stably co-expressing EGFP–centrin-2 (green, reflecting centriole position) and stained for α-tubulin (red, reflecting microtubules) and DNA (blue). Insets are magnifications of the centrosomes (EGFP–centrin-2 signal) in the numbered poles (adapted from Maiato and Logarinho, 2016).

1.2.2 Centrosome amplification and cancer

Faithful control of centrosome number is dysregulated in a growing list of human tumours, particularly in more aggressive ones, leading to CA (Chan, 2011). Indeed, centrosomal abnormalities have been observed in several solid tumours, such as breast, prostate, colon, ovarian and pancreatic (Hsu et al., 2005; Lingle et al., 1998; Pihan et al., 1998; Sato et al., 1999), as well as haematological malignancies (Giehl et al., 2005; Krämer et al., 2005). This is somewhat surprising given their deleterious effects on cell proliferation and that they are poorly tolerated by non-transformed cells (Ganem et al., 2009; Sluder and Nordberg, 2004). Thus, these observations suggest CA as a candidate "hallmark" of tumour cells (Chan, 2011).

Over a 100 years ago, Theodor Boveri proposed for the first time that aneuploidy (the presence of an abnormal number of chromosomes in a cell) induced by CA promotes tumourigenesis (Boveri, 2008; Holland and Cleveland, 2009). Supporting his hypothesis, supernumerary centrosomes (***Figure 1.7b***) have been observed early in the development of several tumours, are implicated as the major mechanism underlying the generation of multipolar mitosis (***Figure 1.7b***), aneuploidy and genomic instability (a known hallmark of cancer), and often correlate with advanced tumour grade and poor clinical outcome (Ganem et

al., 2009; Godinho, 2015; Godinho and Pellman, 2014; Gonczy, 2015; Nigg, 2006; Nigg and Raff, 2009). Moreover, extra centrosomes have been shown to promote invasiveness, an important feature of tumourigenesis (Godinho et al., 2014).

Such observations suggest CA could promote the initial stages of tumour development. However, whether this is a cause or consequence of tumourigenesis remains unclear. Recently, Levine and co-workers have tried to answer this question by chronically inducing CA in mice, by PLK4 overexpression, where they demonstrated that CA is sufficient to promote chromosome missegregation, aneuploidy, and the development of tumours in multiple tissues, supporting a direct causal relationship (Levine et al., 2017). This result gives more robustness to those of previous studies showing that CA can initiate tumourigenesis in flies overexpressing *PLK4* (Basto et al., 2008) and *PLK4* overexpression accelerates tumourigenesis in p53-deficient mice (Coelho et al., 2015; Serçin et al., 2015). Together, they highlight centrosome abnormalities as critical promoters of tumourigenesis.

1.2.2.1 Origins of centrosome amplification

To clinically exploit CA, it is crucial to identify its aetiology. A number of mechanisms are known to experimentally induce supernumerary centrosomes, including cytokinesis failure, mitotic slippage (exiting mitosis without cell division), cell–cell fusion, and deregulation of the centrosome duplication machinery, leading to overduplication of centrioles and *de novo* centriole assembly (Godinho and Pellman, 2014; Godinho et al., 2009). However, little is known regarding their relative contribution to cancer, the existence of other origins or even their underlying molecular mechanisms.

Until now, disruption of the centrosome duplication cycle appears as a major route to CA in cancer, namely through dysregulation of centriolar components both at protein or mRNA level (Godinho and Pellman, 2014). One good example are the high-risk human papillomavirus (HPV)-associated tumours, whose overexpression of the HPV-16 viral E7 oncoprotein induce CA through a process that involves increased *PLK4* transcription levels (Korzeniewski et al., 2011). Loss of p53 is not sufficient to generate CA (Marthiens et al., 2013) but it was shown to contribute to it. p53 might negatively regulate *PLK4* mRNA levels (Li et al., 2005) and most importantly it monitors centrosome number through its checkpoint function, whose mechanism is not yet well understood (Fava et al., 2017; Fukasawa et al., 1996). In colon cancer, it was suggested that the frequently observed amplification of the Aurora-A gene leads to CA, as its overexpression leads to abnormal cytokinesis and supernumerary centrosomes (Ghadimi et al., 2000; Lentini et al., 2007; Meraldi et al., 2002).

Dysregulation of PCM components was also shown to play a role in promoting CA in cancer, for example by overexpression of pericentrin (Loncarek et al., 2008), an integral component of the PCM, or loss of the tumour suppressor BRCA1 (Starita et al., 2004), which regulates the PCM component γ-tubulin.

The absence of more extensive analyses of CA-associated mechanisms in cancer, largely due to associated technical challenges, is limiting our understanding of the CA role in cancer and preventing its use in the clinic. Therefore, it is crucial to integrate systematic surveys of centrosome structure with large-scale *omic* profiles.

1.2.2.2 Coping with centrosome amplification

CA promotes chaotic mitoses that pose a challenge for cell viability. How cancer cells control the detrimental consequences of CA is a very important question in cancer research, since the resulting aneuploidy degree will determine the outcome: moderate levels of genomic instability can induce tumourigenesis, whereas high levels can suppress it (Weaver et al., 2007). While most non-transformed cells normally die or stop proliferating after multipolar mitosis, cancer cells can somehow cope with this abnormality and divide successfully, with an apparent "normal" bipolar mitotic spindle and controlled genome instability (Ganem et al., 2009; Lingle and Salisbury, 1999; Ring et al., 1982).

To supress multipolar divisions and cell death, cancer cells can inactivate, expel, segregate or cluster extra centrosomes (Godinho et al., 2009). Despite little being known regarding these mechanisms and their relative contribution to the "management" of CA, the most prevalent seems to be centrosome clustering (CC), where cancer cells cluster supernumerary centrosomes into two poles to assemble pseudo-bipolar spindles (Brinkley, 2001; Godinho et al., 2009; Nigg, 2002; Zyss and Gergely, 2009). Previous studies, including two genome-wide RNAi screens (Kwon et al., 2008; Leber et al., 2010), uncovered, as the main classes of genes important for CC, i) microtubule-associated proteins, such as the motor protein HSET (encoded by *KIFC1*; Kwon et al., 2008), ii) proteins that induce spindle assembly checkpoint activation, providing a delay on mitosis that allows clustering to occur, and iii) the sub-cortical actin clouds, which create pulling forces on the centrosome via astral microtubules, a process dependent on Myosin 10 (Kwon et al., 2015).

1.2.2.3 Centrosome-targeting cancer therapies

Altogether these findings highlight CA and associated cancer-specific adaptation mechanisms as potential targets for cancer therapy. Accordingly, there are currently in clinical trials drugs that prevent CA, such as PLK4 inhibitors (Mason et al., 2014). Given that normal cells do not rely on CC, pharmacological inhibition of this mechanism promises to selectively target tumour cells and HSET inhibitors are therefore under development (Watts et al., 2013; Wu et al., 2013; Yang et al., 2014). Furthermore, a previous screen has identified 14 compounds that specifically induced multipolar spindles, by CC inhibition, in a breast cancer cell line containing extra centrosomes (Kawamura et al., 2013).

Although CA-targeting therapies are appealing due to their potential selectivity, tumour heterogeneity should be considered. Tumours are heterogeneous populations of cells with and without CA and to date it is unclear how these approaches will affect tumour progression (Godinho and Pellman, 2014).

Furthermore, it will be necessary to identify the patients who will benefit from such therapies. Immunostaining for centriolar components is currently the best laboratory method to quantify CA levels. However, although immunohistochemistry assays are already established in the clinic, particularly for breast cancer subtype classification (Blows et al., 2010), their employment in CA-based patient stratification is still far from ready (Chan, 2011). A promising approach would be to identify a gene-expression-based signature that reflects the tumour CA levels. Indeed, Ogden and co-workers have recently developed a proof-of-principle score based on CA-associated gene expression levels that showed a prognostic value in breast tumours (Ogden et al., 2017). However, in order to bring such scores to the clinic, it will be necessary to characterize the transcriptomic changes associated with centrosome abnormalities and experimentally test if these scores really reflect the CA levels in the tumour.

For all the aforementioned reasons, a comprehensive study of the molecular mechanisms behind centrosome abnormalities will allow to identify ways of selectively targeting this Achilles' heel of cancer cells.

1.3 Screen of centriole number and structure in the NCI-60 panel

Despite the recognized potential of CA in several aspects of cancer, such as aetiology and therapeutics, its incidence, origins and implications remain poorly understood. A recent study from Dr Mónica Bettencourt-Dias' lab, led by Dr Gaëlle Marteil, addressed these questions by screening the NCI-60 panel of human cancer cell lines (Shoemaker, 2006), derived from 9 distinct tissues, for centriole number and length, at a single-cell level (Marteil et al., *manuscript in preparation*). They have confirmed CA as a widespread phenomenon in cancer and have identified a new recurrent feature of cancer cells: centriole over-elongation (COE). Overly-long centrioles generate over-active centrosomes that nucleate more microtubules, a known cause of invasiveness, and induce chromosomal instability. Furthermore, they showed that COE promotes CA through both centriole fragmentation and ectopic procentriole formation, thus identifying novel causes of that abnormality in cancer (Marteil et al., *manuscript in preparation*).

The NCI-60 panel was developed in the late 1980s as an anticancer drug screen, with the main goal of identifying and characterizing novel compounds with potential anticancer activity on these cell lines, and has served the global cancer research community for more than 20 years. This panel has data on cell line sensitivity for around 50.000 compounds, being the largest public database of anticancer drug activity. Moreover, these cell lines have been extensively characterized at genomic (including *TP53* and ploidy status), transcriptomic and proteomic levels (Gholami et al., 2013; Leroy et al., 2014; Liu et al., 2010; Nishizuka et al., 2003; Park et al., 2010; Roschke et al., 2003; Scherf et al., 2000; Shankavaram et al., 2007), allowing the study of diverse biological questions by the cancer community.

Beyond the discovery of COE as a novel centriole abnormality in cancer and promoter of CA, the mentioned centriole screen provides for the first time a rigorous quantification of CA levels in a publicly characterized cell line panel. This unique resource will allow further insights on the CA-associated molecular mechanisms in cancer and the development of clinical applications based on targeting such aberrations.

CHAPTER 2 – AIMS

The centriole profiles uncovered in the Bettencourt-Dias Lab, combined with the publicly available data of the NCI-60 cell lines, allow to both study the molecular origins of COE and CA and identify new compounds that selectively target cells with such abnormalities. Therefore, in my thesis project, I set up to:

1. **Profile the centriole abnormalities along the panel**, going deeper into a single-cell and single-centriole level;
2. **Explore the molecular origins of such centriole abnormalities**, correlating their prevalence along NCI-60 cell lines with the publicly available molecular data for the panel;
3. **Identify new compounds that target CA**, either by selectively killing cancer cells with higher incidence of this abnormality or by targeting genes involved in its origins.

CHAPTER 3 – MATERIALS AND METHODS

3.1 NCI-60 panel of human cancer cell lines

The U.S. National Cancer Institute panel of 60 human cancer cell lines (NCI-60; *https://dtp.cancer.gov/discovery_development/nci-60/*) is composed by 60 cell lines derived from cancers from nine different human tissues/organs (blood, breast, central nervous system (CNS), colon, kidney, lung, ovaries, prostate and skin; ***Figure 3.1***). These cell lines are very well characterized with publicly available information, including their tissue of origin, ploidy, *TP53* status and doubling time (hours). Other parameters, such as sex and age of the donors, prior treatments and the histology of the respective tumour types, are also available.

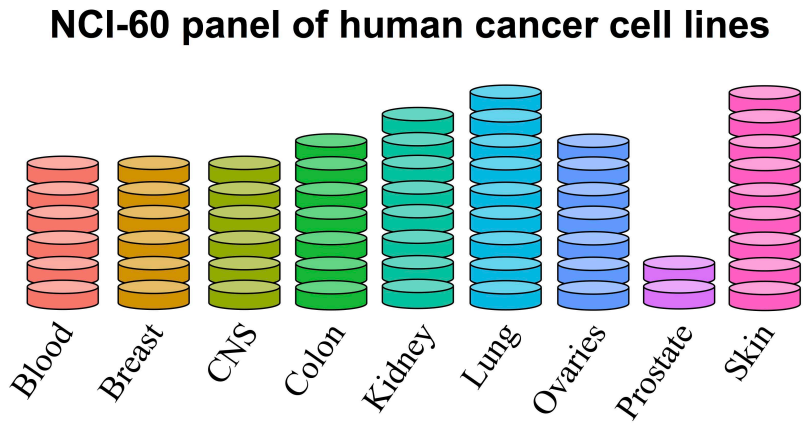


Figure 3.1 NCI-60 panel of human cancer cell lines. Each disc depicts a cell line. CNS, central nervous system.

3.1.1 Screen of centriole number and length

The centriole profiles of the NCI-60 cancer cell lines were obtained from a semi-automated and systematic two-step survey, performed in Dr Mónica Bettencourt-Dias' lab, aiming to quantify both centriole number and length (Marteil et al., *manuscript in preparation*; ***Figure 3.2***). Importantly, as centriole number and length vary throughout cell cycle progression (***Figure 1.6***), the screen was restricted to mitotic cells to limit data variability, since these cells have four fully elongated centrioles (each around 500nm long in human cells). Quantifications were performed in three-dimensions (3D), given the small size of centrioles, and in at least 50 mitotic cells, for each of the cell lines tested.

All the 60 cell lines were analysed in a primary screening in which centrin was used to identify, count and measure centrioles. This marker localises very early to the distal end of centrioles, therefore maximizing their detection. In addition, centrin staining reflects centriole size, since the diameter of centrin dots increases as centrioles elongate, and this marker has

been successfully used in different centriole-related screens (Balestra et al., 2013; Loncarek et al., 2008; Piel et al., 2000; White et al., 2000; Zyss and Gergely, 2009). The primary screening allowed for the distinction of cell lines displaying CA and/or COE from the non-defective ones. However, centrin is also present in small electron-dense cytoplasmic granules called centriolar satellites that might have generated false positives in the primary screening (Dammermann and Merdes, 2002; Löffler et al., 2012; Van de Mark et al., 2015). To validate its results, all the cell lines displaying higher abnormality levels, as well as some less-defective ones (to test the presence of false negatives), i.e. a total of 52 cell lines, were then processed in a secondary screening. Here, centrin was used in combination with a second centriolar marker, CP110, to specifically label *bona fide* centrioles (structures positive for both markers), ensuring the absence of false positives. This secondary screening provided us with the centriole profiles (number and respective length) of 52 cell lines. For the remaining eight cell lines that were not incorporated in the secondary screening (low amplification and low elongation), we used their centriole profiles from the primary screening. In summary, for each of these 60 cell lines screened (eight from the primary and 52 from the secondary screening) we had the centriole number and respective length for at least 50 mitotic cells.

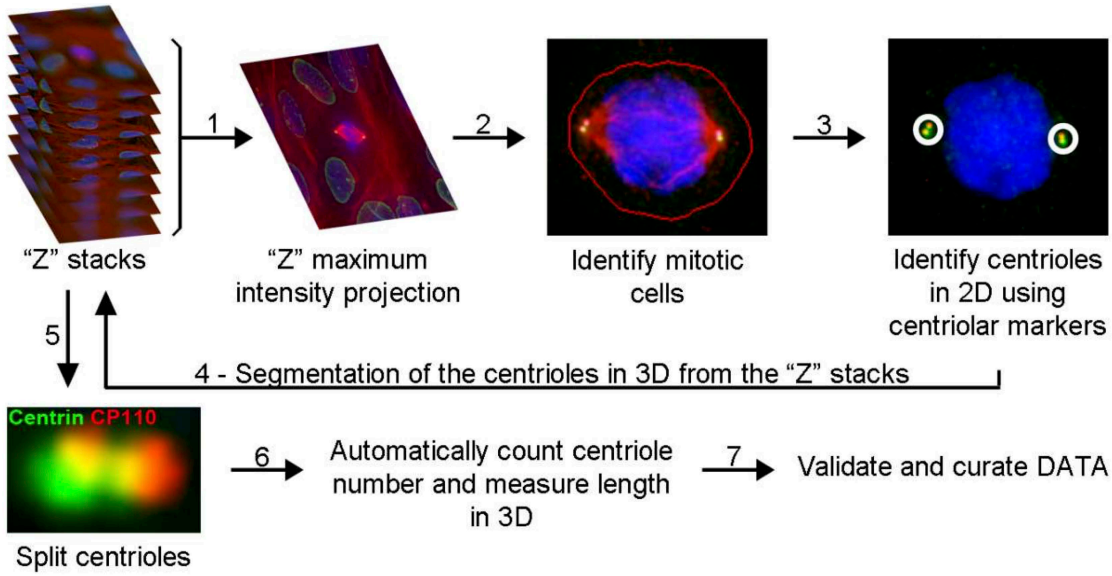


Figure 3.2 Overview of the secondary screening of centriole alterations. First, the microscope was programmed to automatically acquire "Z" stack images, using a 100x objective, that were used to create maximum intensity projections (step 1). Mitotic cells were then identified, in a background of interphasic cells, based on DAPI and α-tubulin signals, both of which are brighter in mitotic cells (step 2). Afterwards, centrioles were individually identified using centriolar markers (centrin and CP110 in the secondary screening; step 3) and segmented in three-dimensions (3D) using "Z" stacks (step 4). Finally, all centrioles were automatically split into individual centrioles and measured in 3D (step 5 and 6). Centriole number and respective length per mitotic cell were stored, together with a gallery of annotated images. These galleries were manually curated twice at steps 3 and 5 (step 7), whereas all the remaining steps were automatically performed by the developed algorithm. Adapted from Marteil et al., *manuscript in preparation*.

Primary screening images were acquired on an Applied Precision DeltaVision CORE system, mounted on an Olympus inverted fluorescence microscope, using a 100x 1.4 NA Oil immersion objective. Secondary screening confocal image stacks were acquired on a Yokogawa CSU-X1 Spinning Disk confocal scan head, coupled to a Nikon Ti confocal microscope using a 100x 1.49 NA Oil immersion objective with 1.5x auxiliary magnification. The later has a pinhole system (a spinning disk with little holes) that minimize the photons coming from planes up and down the focal plane, giving images with more contrast than the ones obtained with the epifluorescence microscope (Marteil et al., *unpublished data*).

3.1.1.1 Centriole abnormality metrics

To profile the centriole abnormality level of each of the NCI-60 cell lines, we used two metrics: the percentage of cells with CA (more than four centrioles) and the percentage of cells with COE (at least one overly-long – longer than 500nm, twice the normal length measured using centrin staining – centriole; ***Annex 1***).

For single-cell and single-centriole analyses, we used only the centriole number and length quantifications resulting from the secondary screening (data available for only 52 cell lines: 12,927 centrioles in 2,842 cells), given the difference in the two screening methodologies. These quantifications were also used to generate two other metrics to characterize the centriole structure per cell line: the arithmetic mean of centriole number and length per cell (***Annex 1***).

3.1.1.2 Centriolar mass

The centriolar mass of each individual cell was calculated as the sum of its centrioles' length.

3.1.2 Centriole length regulators screen

Our analyses of the NCI-60 panel identify one gene, *PRKACA*, that might regulate centriole length. Interestingly, this gene was already tested in an independent screen for putative centriole length regulators led by Dr Mariana Faria in Dr Mónica Bettencourt-Dias' lab (*unpublished data*). In the human osteosarcoma U2OS cell lines, overexpressing *PLK4*, 187 genes were individually knocked down, using three independent small interfering RNAs (siRNAs) per gene. In each siRNA experiment, the CEP135 and centrin protein intensity levels were measured (***Figure 3.3a***). CEP135 concentrates at the centriolar proximal ends and its intensity reflects centriole number (Fu et al., 2016), whereas centrin localises to the distal end

and its intensity increases as centrioles elongate (Loncarek et al., 2008). A decrease/increase in centrin intensity but no change in CEP135 intensity would reflect a decrease/increase in centriole length (***Figure 3.3c***). Variation in intensity levels was quantified in robust z-scores – number of median absolute deviation below or above the control (Scramble) median.

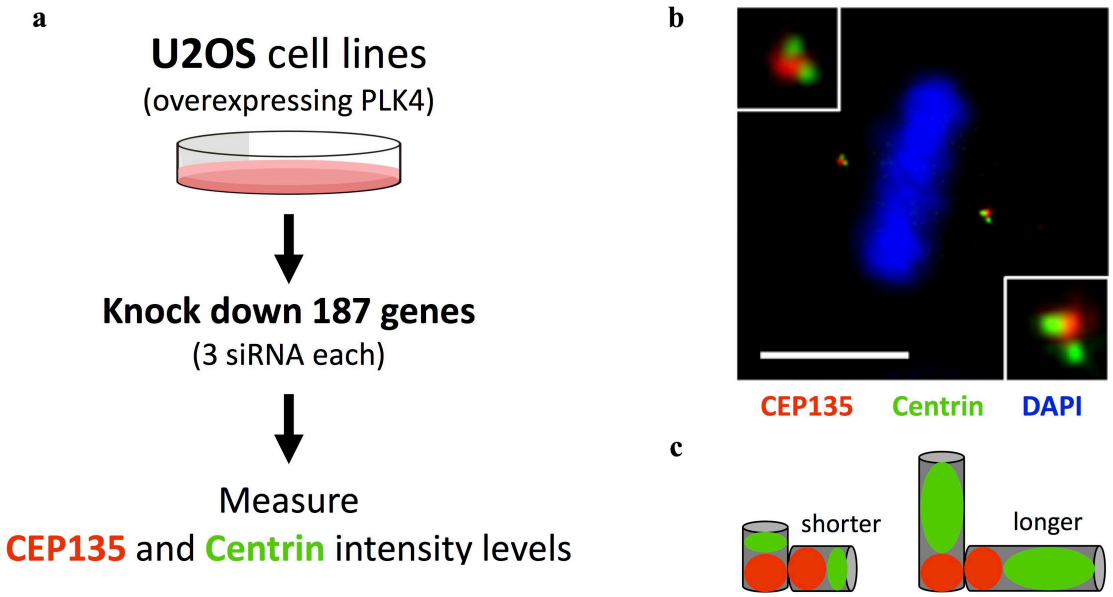


Figure 3.3 Overview of the centriole length regulators screen. a) Briefly, human osteosarcoma U2OS cell lines, overexpressing *PLK4*, were knocked down for 187 genes individually, using three independent small interfering RNAs (siRNAs) per gene. CEP135 and centrin protein intensity levels were measured to reflect the centriole length. **b)** Example of HeLa cells immunostained with CEP135 (red), centrin (green) and DAPI (blue, for DNA) antibodies (adapted from Seo et al., 2015). **c)** Illustration of how a decrease/increase in centrin intensity but no change in CEP135 intensity would reflect centrioles shorter/longer than the control.

3.1.3 Flow Cytometry analysis of cell cycle phases

The cell cycle profile of the NCI-60 cell lines was analysed by *O'Connor et al., 1997*, through fluorescence-activated cell sorting (FACS). In this approach, cellular DNA is labelled and cells are then sorted according to the fluorescence intensity, which reflects the amount of DNA within the cell, allowing to distribute them in three categories encompassing the four cell cycle stages: G1, S and G2/Mitosis (G2/M; Jayat and Ratinaud, 1993; ***Figure 3.4***). The authors characterized each cell line for the cell cycle distribution in percentage of cells in each of the G1, S and G2/M phases. They also calculated the cell lines' doubling time, which we compared with the recent doubling time estimates from the U.S. National Cancer Institute to confirm reproducibility of an experiment done twenty years ago. Indeed, estimated cell lines' doubling times from both studies were correlated (Spearman correlation coefficient: 0.96, $p < 0.0001$; ***Annex 2***).

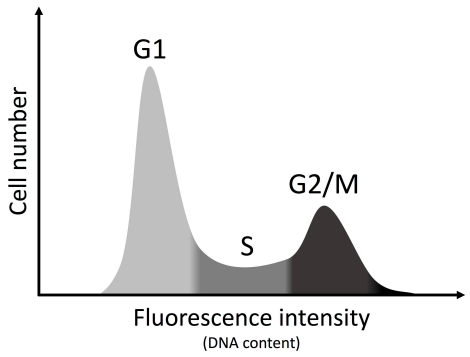


Figure 3.4 Distribution of cells along the cell cycle phases. Measuring DNA content (using fluorescence intensity) allows to determine the percentage of cells in G1, S and G2/Mitosis (G2/M). The first peak represents diploid cells, before DNA replication, in G1 phase. Cells that are undergoing DNA replication at S phase have an increased amount of DNA, representing the space between the two peaks. After DNA replication, during G2 and mitosis, cells have the double of the G1 cells' DNA and therefore the double fluorescence intensity, reflected on the second peak.

3.1.4 Molecular and pharmacological data sets

The NCI-60 panel has been largely characterized at the molecular and pharmacological levels, with the resulting information having been made publicly available in several databases. Most of the data sets can be accessed through the CellMiner web application (*https://discover.nci.nih.gov/cellminer/loadDownload.do*; Reinhold et al., 2012; Shankavaram et al., 2009), while others need to be accessed via the original papers, such as the global proteome analysis (Gholami et al., 2013). In this work, we focused on gene and protein expression and drug activity analyses.

3.1.4.1 Gene expression

The expression level of a gene can be measured using different technologies. One of the most important are microarrays – a collection of gene-specific nucleic acids probes attached to a solid surface at defined locations. The mRNA complement of a biological sample is converted into complementary DNA (cDNA), labelled (usually with a fluorescent dye), and then allowed to hybridize with the gene-specific probes on the array. Gene expression levels can thereby be quantified genome-wide by the fluorescence intensities measured across all spots on the array, assumed to be proportional to the amount of nucleic acid hybridized to the respective probes (John Quackenbush, 2001).

Gene expression in NCI-60 cell lines has been profiled using different microarray platforms. Normalized gene expression values (averaged probe intensities combined from five microarray platforms) for the 60 cell lines were downloaded from the CellMiner. Probes containing missing values in more than 5% of cell lines (i.e. either with low sensitivity or associated with non-expressed genes) or a variance across samples lower than 0.1 (i.e. biologically non-informative; ***Annex 3***) were removed from the analysis. After this quality control step, 19,676 genes remained for further analysis.

3.1.4.2 Protein expression

The identification and quantification of the proteins expressed in a sample can already be done through high-throughput techniques. Mass spectrometry (MS) is one of the most commonly used and is based on the mass-to-charge ratio of ions. In this technique, proteins are enzymatically digested and the resultant peptide masses and fragment ions charges measured, allowing peptide identification through database searching. Finally, proteins are quantified from the correctly identified peptides that constitute them (Kolker et al., 2006).

The NCI-60 global proteome was retrieved from Gholami et al., 2013, where 59 cell lines were analysed by MS-based proteomics. Relative protein abundance across samples was calculated from intensity-based label-free quantification. Peptides not quantified in more than 50% of cell lines (more relaxed cut-off, compared with gene expression, given the lower MS detection efficiency), as well as probes with a variance across samples lower than 0.025 (***Annex 4***), were removed from the analysis for the reasons described in section 3.1.4.1. 3,328 of the initial 8,113 proteins remained for further analysis.

3.1.4.3 Drug sensitivity

Activity data for 21,121 compounds were downloaded (on 6/1/2017; this list is frequently updated with new compounds) from CellMiner in z-transformed negative log10 of GI50 (the compound's concentration that causes 50% growth inhibition) values, with higher values corresponding to higher sensitivity of cell lines to the respective drugs. These GI50 scores were measured based on a screen at five concentration levels (0.01, 0.1, 1, 10 and 100 μM) in each of the 60 cell lines (screening methodology explained in https://dtp.cancer.gov/discovery_development/nci-60/methodology.htm). All these compounds have passed the NCI-60 quality control steps for data reproducibility and minimum variability across cell lines, and they include 158 Food and Drug Administration (FDA) approved and 79 clinical trial drugs. Furthermore, each of these drugs is annotated with its FDA status, mechanism of action, PubChem SID (Substance accession Identifier) and SMILE (Simplified Molecular Input Line Entry).

Compounds with activity data for a minimum of 35 informative cell lines (as used in CellMiner: NCI-60 Analysis Tools, https://discover.nci.nih.gov/cellminer/analysis.do) and with a variance higher than 0.99 (***Annex 5***) were kept, for the reasons described in section 3.1.4.1, accounting for 14,005 compounds.

3.2 Transcriptomic alterations associated with centriole abnormalities

3.2.1 Centrosomal genes

To identify genes putatively associated with centriole abnormalities, we focused on centrosomal genes retrieved from the CentrosomeDB - Centrosomal Proteins Database (Alves-Cruzeiro et al., 2014; Nogales-Cadenas et al., 2009).

3.2.2 Gene Set Enrichment Analyses

Gene Set Enrichment Analyses (GSEA; Mootha et al., 2003; Subramanian et al., 2005) is a computational method that determines if an a priori defined gene set (for instance, a group of genes that share a biological function) shows statistically significant enrichment at the top (or bottom) of an ordered gene list (***Figure 3.5a***). This approach has the main advantage of considering information from all the studied genes, not only those above an arbitrary cut-off as commonly done by other knowledge-based functional enrichment analysis methods. GSEA also allows the users to create their own custom gene sets.

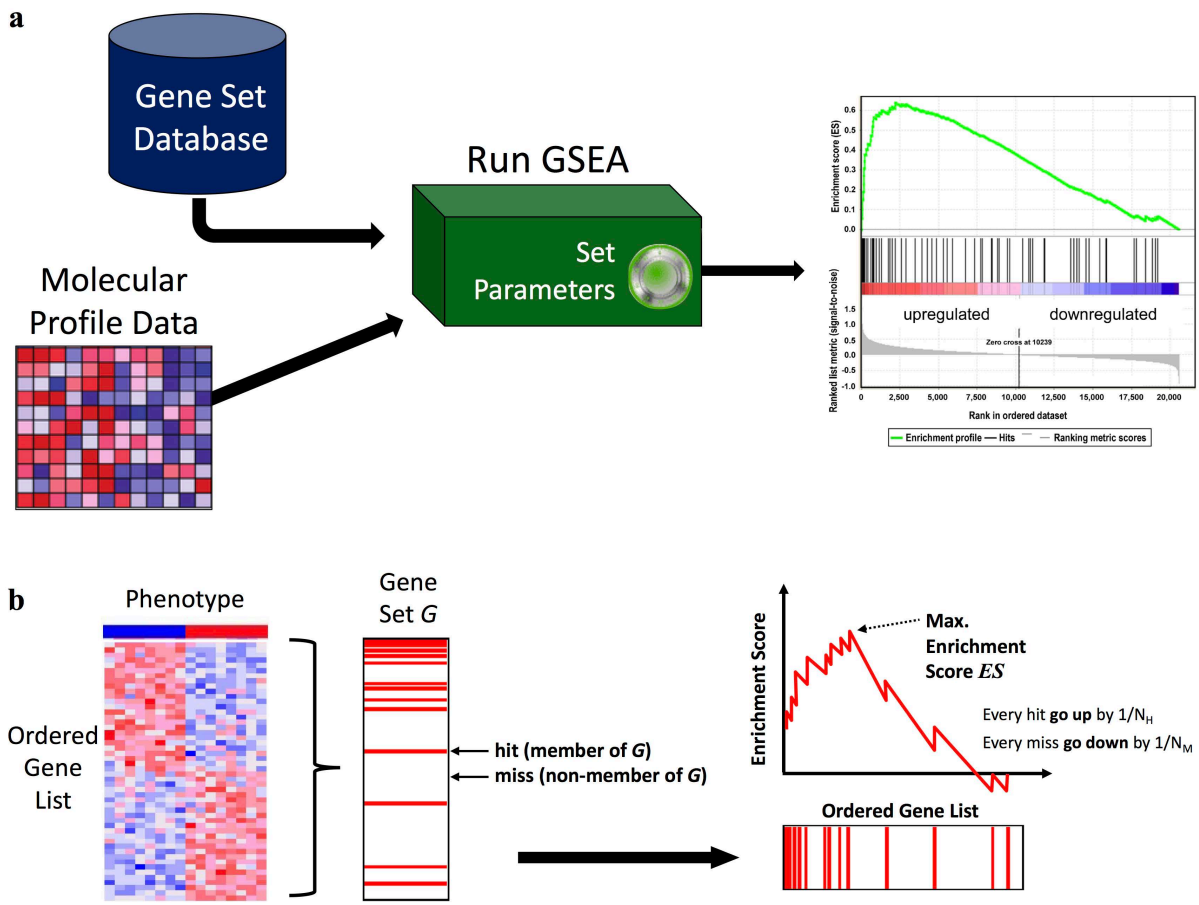


Figure 3.5 Overview of Gene Set Enrichment Analyses. a) GSEA uses molecular profile data (e.g. ordered gene list according with some phenotype) and a gene set database as input. It then tests the enrichment of each gene set at the top (or bottom) of the ranked list. An example of a gene set positively enriched, e.g. enriched on genes upregulated between two conditions, is shown (gene set result retrieved from Kammerer et al., 2011). **b)** An ordered gene list, ranked by the association with a phenotype (e.g. difference in gene expression between two groups, blue and red) is used in GSEA. For each gene set *G*, its members are considered *hits*, while non-member genes are considered *misses*. Then GSEA walks down the ordered gene list increasing a weighted running-sum statistic when it finds a *hit* (sums $1/N_h$, where N_h is the number of hits) and decreasing when it finds a *miss* (subtracts $1/N_m$, where N_m is the number of misses), so that the running-sum statistic starts and finishes at 0. Finally, the enrichment score (ES) is defined as the maximum deviation from zero (adapted from Subramanian et al., 2005).

Briefly, the GSEA software (Mootha et al., 2003; Subramanian et al., 2005) sorts the list of genes in descending order, according to a chosen ranking criterion (e.g. fold-change of gene expression between two groups or correlation coefficient between gene expression and some phenotype), and walks it down, incrementing a running-sum statistic when a gene is in the gene set and decreasing when it is not. The increments are weighted according to the number of genes in the gene set, since the running-sum statistic should start at 0 and finish also at 0. Then, an enrichment score (ES) is defined as the maximum deviation from zero encountered when walking the list. Thus, gene sets enriched at the top of the gene list will have a higher positive ES, while gene sets enriched at the bottom will have a higher negative ES (***Figure 3.5b***).

The significance of an observed ES (nominal p-value) is estimated by comparing it with a null distribution of ESs computed with randomly reordered gene lists. To allow the comparison of results across gene sets, the ESs of all gene sets are normalized to account for their sizes, yielding a normalized enrichment score (NES) for each set reflecting the observed enrichment's effect size. The probability of a given gene set being a false positive finding is estimated by calculating the false discovery rate (FDR; Benjamini and Hochberg, 1995).

The GSEA software (version 2.2.2) was run with gene sets retrieved from the KEGG database (Kanehisa and Goto, 2000; Kanehisa et al., 2016), that contains a collection of pathway maps associated with diverse biological functions. We used the *GseaPreranked* tool, that takes as input a ranked list of genes, with the default parameters. The input gene lists were ranked according to the Spearman correlation coefficients, ordered from 1 to -1, between gene/protein expression and centriole abnormalities levels. Gene sets with a FDR lower than 5% were considered significant.

3.2.3 Linear regression models to decouple independent effects

When performing gene expression analyses, it is important to consider all sources of variation to ensure that results correspond to true biological events, and are not related to other

variables partially confounded with the phenotype of interest. One strategy to separate independent effects is to use linear regression models i.e. to test the hypothesized linear relationship between a dependent and one or more explanatory variables (simple or multiple linear regression, respectively; Lai et al., 1979).

A model for a multiple linear regression that relates a dependent variable Y with n explanatory variables (X_1, X_2, ..., X_n), given k observations, can be stated as

$$Y_i = \beta + \beta X_{1,i} + \beta X_{2,i} + \ldots + \beta X_{n,i} + \varepsilon \quad \text{for } i = 1,2,\ldots,k$$

where β (also called the y–intercept) is the value of Y when all explanatory variables are null, β (beta coefficients) represent the strength of the effect of each individual explanatory variable on the dependent variable, and ε is the error term.

Linear regression models are frequently used in gene expression analyses, where Y represents the expression of a gene in k samples. In this context, the explanatory variables are usually phenotypic features that can be associated with changes in gene expression. Multiple linear regression models are used to decompose the independent effect of those phenotypes on gene expression. Afterwards, approaches as GSEA can be run on gene lists ranked according to the impact of each independent effect on gene expression. Linear regression models were implemented using the *limma* R package (Ritchie et al., 2015) and genes were ranked by the moderated t-statistic for differential expression. For each gene, the moderated t-statistic was calculated as the ratio between the base 2 logarithm of the fold-change in expression and the moderated standard error of the base 2 logarithm of expression across samples. Standard errors were moderated across genes using an empirical Bayesian model that borrows information from the ensemble of genes to help inference about each individual gene (Smyth, 2004).

3.3 Statistical hypothesis testing

The main aim of statistics is to test a hypothesis (Witte and Witte, 2013). A hypothesis is a proposed explanation, often called an "educated guess", for a phenomenon, that should be testable, either by experiment or observation. Hypothesis testing compares two hypotheses: the null hypothesis encodes for "nothing happening", whereas the alternative hypothesis states that there is relation between phenomena, which is usually the working hypothesis of the researcher. Hypothesis testing has three main steps: state the null hypothesis, choose the appropriate statistical test to perform, and in the end assess if there is enough evidence to reject the null hypothesis, usually based on a p-value. The p-value (or probability value) is the probability of rejecting the null hypothesis when it is true. Thus, the lower the p-value, the

higher the confidence in rejecting the null hypothesis and accepting the researcher's working one (Greenland et al., 2016; Sedgwick, 2014; Shaw and Proschan, 2013; Witte and Witte, 2013).

Statistical tests can be parametric or non-parametric, their choice mainly depending on how the observed data are distributed. A parametric test makes assumptions about the data distribution (e.g. the data is normally distributed) whereas a non-parametric one makes no such assumptions. The latter is less powerful because it uses less information, since it sorts the data by magnitude and replaces observed values with their ranks, but it is also less biased by outliers and misassumptions about distributions (Conover, 1971; Sedgwick, 2015).

In the present work, different statistical tests were used, according to the different null hypotheses under examination. All the statistical analyses and graphics were performed using the R free software environment (R Core Team, 2017). R is a language and environment that provides a wide variety of statistical and graphical techniques, and is highly extensible due to a vast range of freely available packages. Most of the graphics displayed in this work were generated using the *ggplot* function from the *ggplot2* R package (Wickham, 2009).

3.3.1 Unpaired and paired two-sample statistical tests

The Wilcoxon signed-rank test (Wilcoxon, 1945) is a non-parametric test used to compare repeated measurements on a single sample (one-sample) or two paired samples (i.e. each measurement in one sample is uniquely paired with a measurement in the other sample, resulting in pairs of observations). The one-sample version is the analogue of the parametric independent one sample t-test and tests the null hypothesis "the median of the sample is equal to a known standard value (i.e. theoretical value)". The paired version is used for the null hypothesis "the median difference between pairs of observations is zero". The last is the alternative to the parametric paired Student's t-test, where the null hypothesis is "the mean difference between pairs is zero".

When comparing two unpaired samples, a Wilcoxon rank-sum test (also known as the Mann-Whitney U test) should be used. It is the non-parametric alternative to the unpaired Student's t-test and tests the null hypothesis "the median of two samples are equal" or "both samples are from populations having the same distribution".

Both the Wilcoxon signed-rank and rank-sum tests were implemented using the *wilcox.test* function provided by the *stats* R package (R Core Team, 2017).

3.3.2 Kruskal-Wallis rank-sum test

The Kruskal-Wallis rank-sum test (Kruskal and Wallis, 1952) is an extension of the Wilcoxon rank-sum test for more than two samples. It is a non-parametric method for testing whether different samples originate from populations with the same distribution, where the null hypothesis is that the samples' medians are the same.

This test was performed using the *kruskal.test* function provided by the *stats* R package.

3.3.3 Fligner-Killeen test of homogeneity of variances

The Fligner-Killeen non-parametric test (Conover et al., 1981) was used to assess the assumption of equality of variances between two or more groups, under the null hypothesis that the groups' variances are equal. It is very robust against departures from normality and outliers.

This test was performed using the *fligner.test* function provided by the *stats* R package.

3.3.4 Unsupervised hierarchical clustering

Unsupervised hierarchical clustering is an exploratory data analysis technique used to group similar objects into clusters. Similarity is calculated based on inter-object distance measures, including Euclidean (square root of the sum of the squares of the differences between coordinates) and correlation-based (calculated by subtracting from 1 the coefficient of correlation between objects) ones, where shorter distances reflect more similar samples. Each object is initially considered as a cluster of its own. Then, the two most similar clusters are successively combined until all objects are in the same cluster. The result is a tree-based representation of the objects, also known as dendrogram, that shows the hierarchy of the different clusters (Everitt, 1974).

Unsupervised hierarchical clustering was computed using the *heatmap.2* function provided by the *gplots* R package (Warnes et al., 2016).

3.3.5 Spearman's rank correlation

Spearman's rank correlation (Spearman, 1904) is a non-parametric measure of the statistical dependence between two paired variables that assesses their monotonic relationship. In short, it ranks the values from smallest to largest within each variable and then compares paired ranks. Under the null hypothesis that "the ranks of two variables do not covary", the test calculates Spearman's correlation coefficient, that varies between -1 (negative correlation) and 1 (positive), and the respective p-value.

This test was executed using the *cor.test* function provided by the *stats* R package.

3.3.6 Analyses of covariance (ANCOVA)

Analysis of covariance (ANCOVA) combines analysis of variance (ANOVA) and linear regressions. It is used to compare two or more regression lines by testing the effect of a categorical factor (i.e. different groups) on a linear relationship between a continuous covariate (x-var) and a response variable (y-var), under the null hypothesis that slopes of different regression lines are equivalent, i.e. regression lines are parallel between groups (Borm et al., 2007; Miller and Chapman, 2001).

To compare two or more regression lines, the relation between variables is split into several linear equations, according to the categorical factor levels. Then, the interaction of the categorical factor on that relation is measured. As exemplified in ***Figure 3.6***, when the response follows the same trend in two groups, the group factor does not affect the association between variables, i.e. there is no interaction (***Figure 3.6a***). If the interaction is significantly different from zero, it means that we can reject the null hypothesis: the categorical factor group affects the relation between the covariate and the response variable and, therefore, regression lines have different slopes between groups (***Figure 3.6b***; Fuchs, 2011).

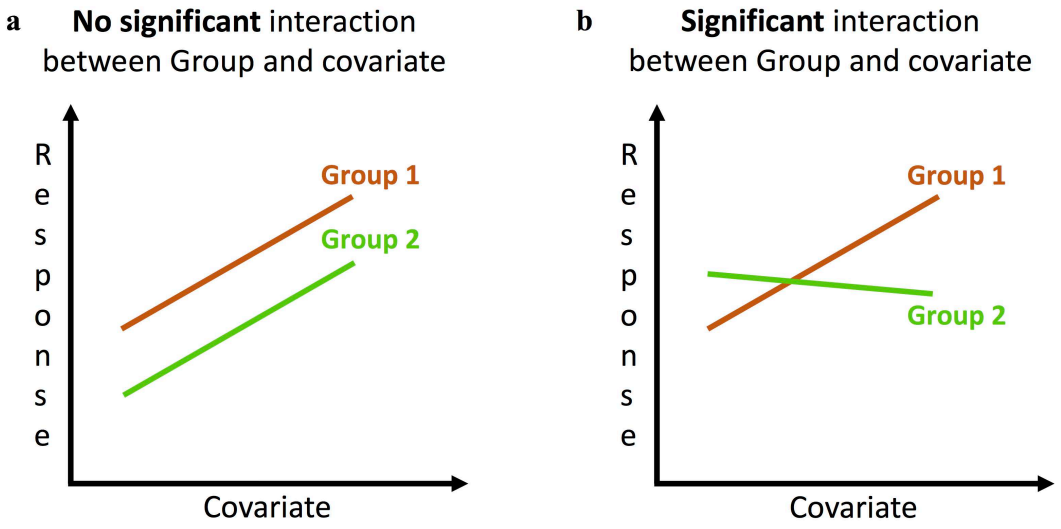


Figure 3.6 Comparing two regression slopes with ANCOVA. Analysis of covariance (ANCOVA) tests the null hypothesis that slopes of different regression lines, that model the relation between a covariate and a response variable in different categorical factor levels (groups), are equivalent. **a)** If the factor Group does not influence the linear relationship between variables, it means that there is no significant interaction and both slopes are parallel. **b)** When the effect of the covariate on the response variable depends on the sample group, the interaction is significantly different from zero and the regression lines have different slopes (adapted from Fuchs, 2011).

In the context of this work, the continuous variables are CA and COE per cell line, and the levels of the categorical factor are the different NCI-60 cancer tissues. This test was done using the *aov* function provided by the *stats* R package.

3.3.7 Pearson's chi-squared test

3.3.7.1 Chi-squared test of independence

Pearson's chi-squared test is a non-parametric method commonly used to determine whether there is a significant association between two categorical variables, each with two or more levels. It converts categorical variables into a 2-dimensional contingency table and tests the null hypothesis "the two variables are independent". If those variables are independent, the expected value for each cell of the contingency table is: the sum of all cells in its row multiplied by the sum of all cells in its column, then divided by the sum of all cells in the table. The test compares the observed values with the expected ones and, if the null hypothesis is rejected, it means that those variables are not independent and there is a relation between them (Mchugh, 2013; Yates, 2012)

This test was performed using the *chisq.test* function provided by the *stats* R package.

3.3.7.2 Chi-squared goodness of fit test

Pearson's chi-squared goodness of fit (Chernoff and Lehmann, 1954) is a non-parametric test used to compare the observed frequency distribution with that expected, when analysing a single categorical variable with two or more levels. In this case, the null hypothesis being tested is whether the observed proportions equal the expected ones.

This test was performed using the *chisq.test* function provided by the *stats* R package.

3.3.7.3 Generate a random distribution of COE per cell

A distribution of COE per cell in the NCI-60 panel is generated by counting the number of cells (within the total 2,842) with each amount of longer centrioles (e.g. 0, 1, 2, ..., n) and can be illustrated by a histogram.

To generate a random (i.e. expected by chance) distribution of COE per cell, we performed 1000 permutations on centriole length data (for 12,947 centrioles), randomizing all centriole lengths across all cells, and, for each permutation, we counted the number of cells with each amount of longer centrioles. Finally, we summarized the results of 1000

permutations, using the mean and respective standard error for each amount, in an overall random distribution of COE per cell (***Figure 3.7***).

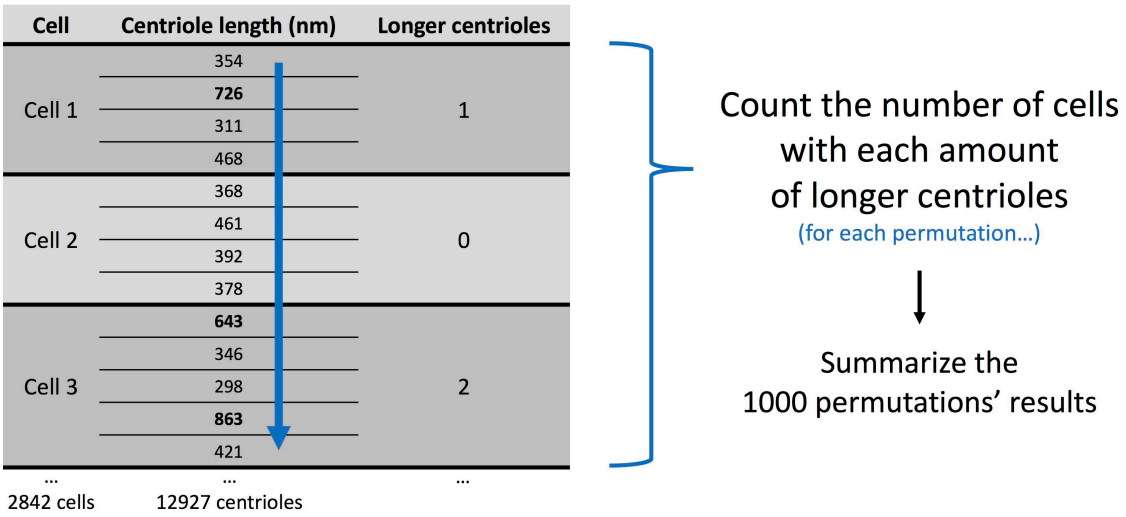


Figure 3.7 Generate a random distribution of COE per cell. A table with all centriole lengths per cell was generated, with data for 12,947 centrioles distributed by 2,842 cells. The longer centrioles (more than 500nm) are highlighted in bold (column 2) and the number of longer centrioles, per cell, is represented in column three. To generate the expected-by-chance distribution, the column with centriole length information was randomly permutated 1000 times (blue arrow) and, for each permutation, the counts of column three were recalculated. Then, for each permutation, the number of cells with each amount of longer centrioles (e.g. 0, 1, 2, ..., n) was counted. Finally, the results of those 1000 permutations were summarized in an overall distribution, using the mean and respective standard error for each amount of longer centrioles.

3.3.8 Binomial test

The binomial test is used when an experiment has two possible outcomes (i.e. success/failure) and compares the observed proportion of successes with a hypothesized probability of success. It tests the null hypothesis that the observed probability of success is equal to the hypothesized one (Conover, 1971).

This test was implemented using the *binom.test* function provided by the *stats* R package.

3.3.9 Correction for multiple testing

Statistical analysis can involve the simultaneous testing of many hypotheses. The more hypotheses are tested, the higher the chances of getting significant results just by chance. For example, a p-value of 0.05 means a probability of 5% of rejecting the null hypothesis when it is true. Thus, when performing 100 statistical tests, and assuming for all of them the null hypothesis is true, one expects about five of them to be significant at that p-value, with the null hypothesis being rejected just due to chance (the so-called false positives). Therefore, if the

interest is to find observations yielding enough evidence to reject the null hypothesis, p-values should be corrected for multiple testing to adjust the associated statistical confidence according with the number of tests performed (Noble, 2009).

We used the false discovery rate (FDR) method (Benjamini and Hochberg, 1995), that controls for the expected proportion of false positives (i.e. rejections of the null hypothesis when it actually holds) in the correction of nominal p-values for multiple testing. Each FDR-corrected p-value is calculated by multiplying the nominal p-value by the number of tests and then dividing it by its rank amongst all the nominal p-values sorted in ascending order (smallest to largest).

This approach was implemented using the *p.adjust* function provided by the *stats* R package.

CHAPTER 4 – RESULTS

4.1 Profile of centriole abnormalities in the NCI-60 panel

The NCI-60 screen of centriole number and length provided a comprehensive landscape of centriole abnormalities in different cancer cell lines and tissues. To compare the frequencies of centriole defects in different cancer types, we chose two metrics that capture abnormality levels in their number and length: the percentage of cells with CA and the percentage of cells with COE in each cell line.

CA was more frequent than COE in cancer cell lines (means equal to 17% and 5% of cells, respectively; paired Wilcoxon signed-rank test $p < 0.0001$) and none of these centriole abnormalities was specific for any primary tumour type from which the cell lines were derived. Cell lines from all different primaries exhibited some degree of abnormality both at number and length levels (***Figure 4.1a,b***). Yet, lung cancer was the primary histology with the highest variability in both abnormalities.

All cell lines had some level of CA, except the control UO-31 (***Figure 4.1c***). In contrast, only 41 cell lines (68%) had length abnormalities (***Figure 4.1d***). The skin cancer cell line MDA-MB-435 was the one with the most extreme COE within the panel.

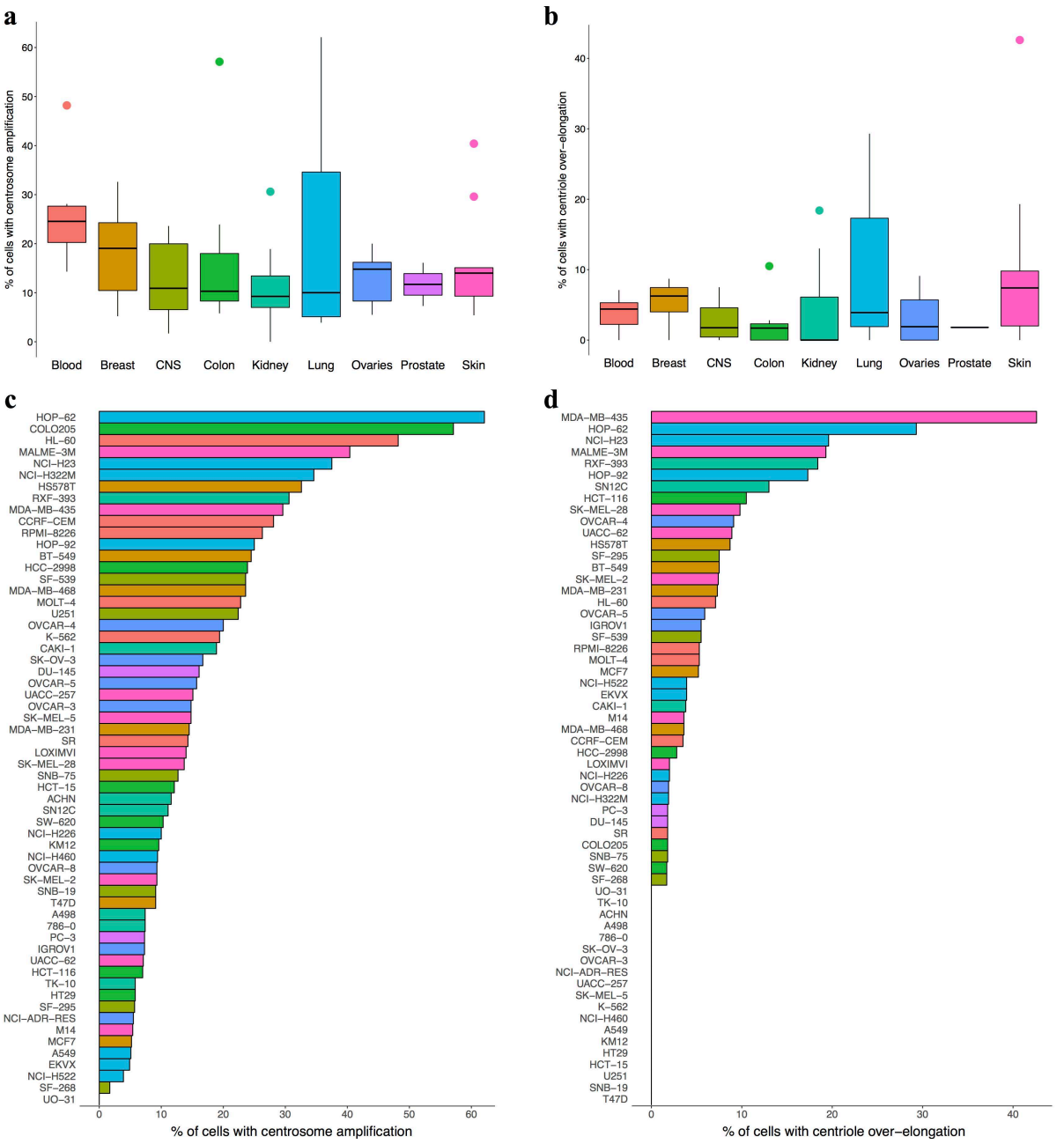


Figure 4.1 Profile of centriole abnormalities in the NCI-60 panel. Percentage of cells with **(a,c)** CA and **(b,d)** COE for each **(a,b)** primary cancer histology and **(c,d)** cell line, coloured by tissue of origin and ordered by abnormality level (in the case of cell lines).

4.1.1 Single-cell and single-centriole heterogeneity in cancer

Studies on centriole abnormalities are usually done at a cell population level, whereas the prevalence and heterogeneity of these abnormalities in cancer at single-cell (for centriole number), and even single-centriole (for centriole length), level remain unknown. Taking advantage of the single-centriole resolution of the NCI-60 screen, we went deeper in the profiling of those abnormalities in that cancer cell line panel.

4.1.1.1 Centriole number

Although cell lines from all origins presented cells with different amounts of centrioles, we observed different variances of number of centrioles per cell between primary cancer tissues (Fligner-Killeen p < 0.0001; ***Figure 4.2***).

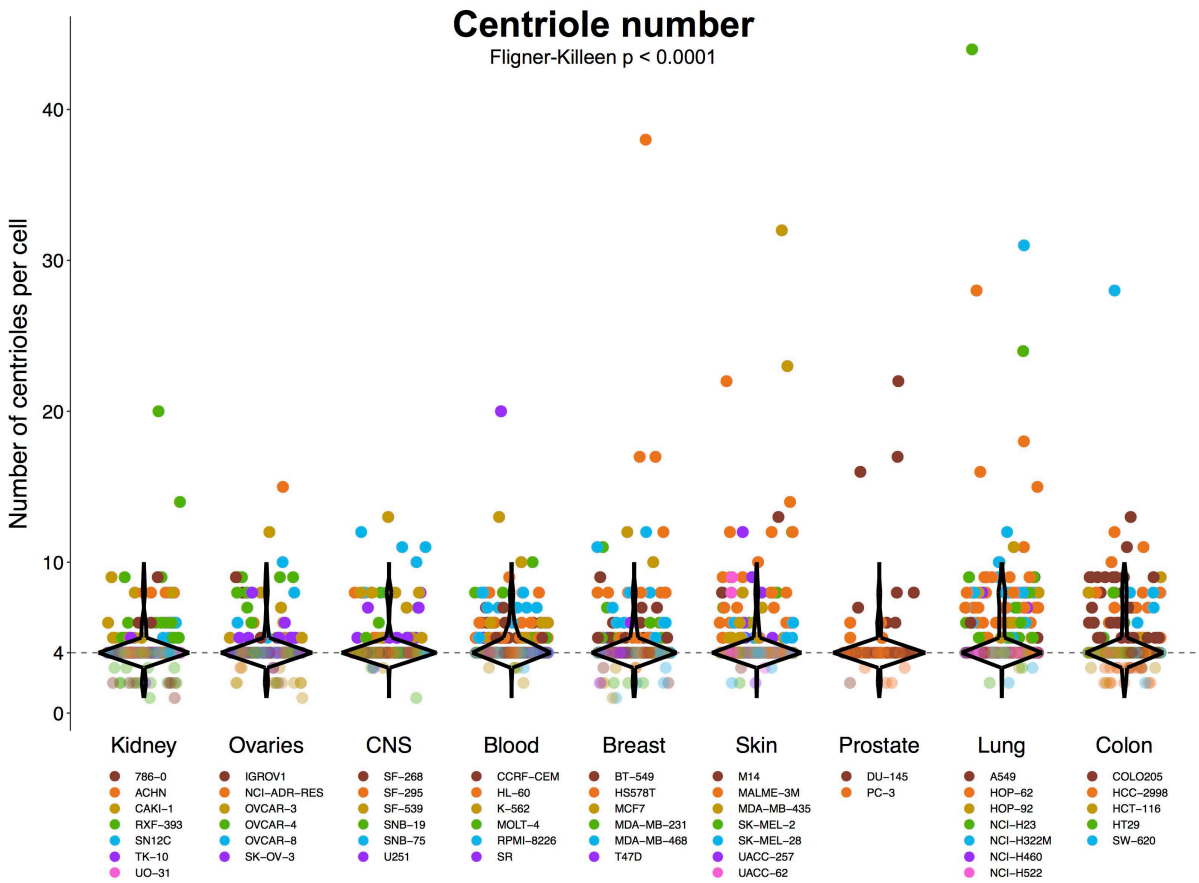


Figure 4.2 Distribution of the number of centrioles per cell across NCI-60 tissues of origin, coloured by cancer cell line. Each dot corresponds to a cell. Violin plots were created based on segments connecting frequencies at each integer (from one to ten centrioles/cell), given that centriole number is a discrete variable.

However, for each primary tissue, about 70% of cells had four centrioles, making it difficult to define a metric for single-cell CA heterogeneity. Given that difficulty, we have characterized each primary tissue's heterogeneity based on the cumulative distribution of the number of centrioles per cell across cells (***Figure 4.3a***). Tissues with higher CA heterogeneity were therefore expected to exhibit lower cumulative distribution values for more than four centrioles. Indeed, despite blood cancer cell lines being those with the largest proportion of cells with CA, colon and lung cancer ones presented overall the highest CA dispersion among the panel (***Figure 4.3a***).

To group tissues according to their CA heterogeneity, we compared centriole number variance in all combinations of pairs of tissues of origin, using the Fligner-Killeen non-parametric test. Unsupervised hierarchical clustering based on the resulting p-values (more precisely, minus their base 10 logarithm) suggests three main groups: colon and lung cancer tissues with higher heterogeneity; prostate, skin, breast and blood with an intermediary dispersion level; CNS, ovaries and kidney with lower CA dispersion (***Figure 4.3b*** and ***Annex 6a***). Remarkably, we observed in all tissues more cells with eight than with seven centrioles (***Figure 4.3b***).

Despite those differences across tissues, centriole number distributions were similar between cell lines with the same tissue of origin (data not shown).

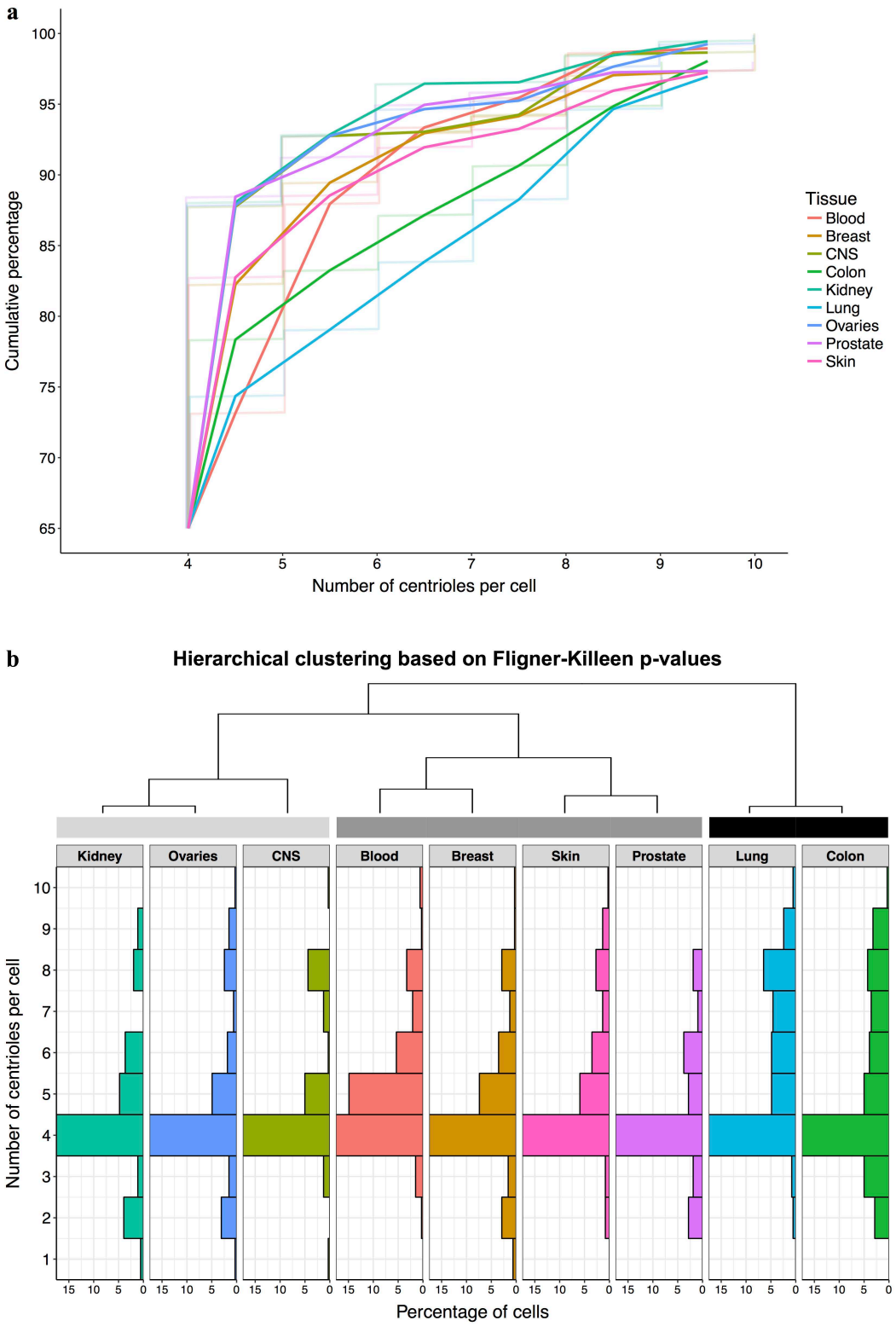


Figure 4.3 Centriole number heterogeneity in cancer. **a)** Cumulative distribution of the number of centrioles per cell, with cells grouped by tissue of origin (fainter lines). Dark lines connect the "jumps" in the respective discrete cumulative distributions. **b)** Histograms (percentage of cells) of the number of centrioles per cell, grouped and coloured by tissue of origin. Only percentages for cells with one to ten centrioles are shown but they were calculated based on all the cells for each tissue of origin. Tissues are hierarchically clustered (Euclidean distances calculated based on Fligner-Killeen p-values across all combinations of tissue pairs) and the three main clusters are highlighted. The associated heatmap of distances is shown in *Annex 6a*.

4.1.1.2 Centriole length

Centriole length variability was also observed across different tissues of origin (Fligner-Killeen p < 0.0001; ***Figure 4.4***). Like centriole number, we performed unsupervised hierarchical clustering based on Fligner-Killeen p-values across all possible combinations of tissue pairs. We observed three main clusters, with lung and skin constituting that of higher length heterogeneity and ovaries the one with lower heterogeneity (***Figure 4.4*** and ***Annex 6b***). It should be noted that we observed some extremely long centrioles in the HOP-62 lung cancer cell line, including the longest centriole within the screen, which was 5,339nm long (more than ten times longer than a normal-length centriole).

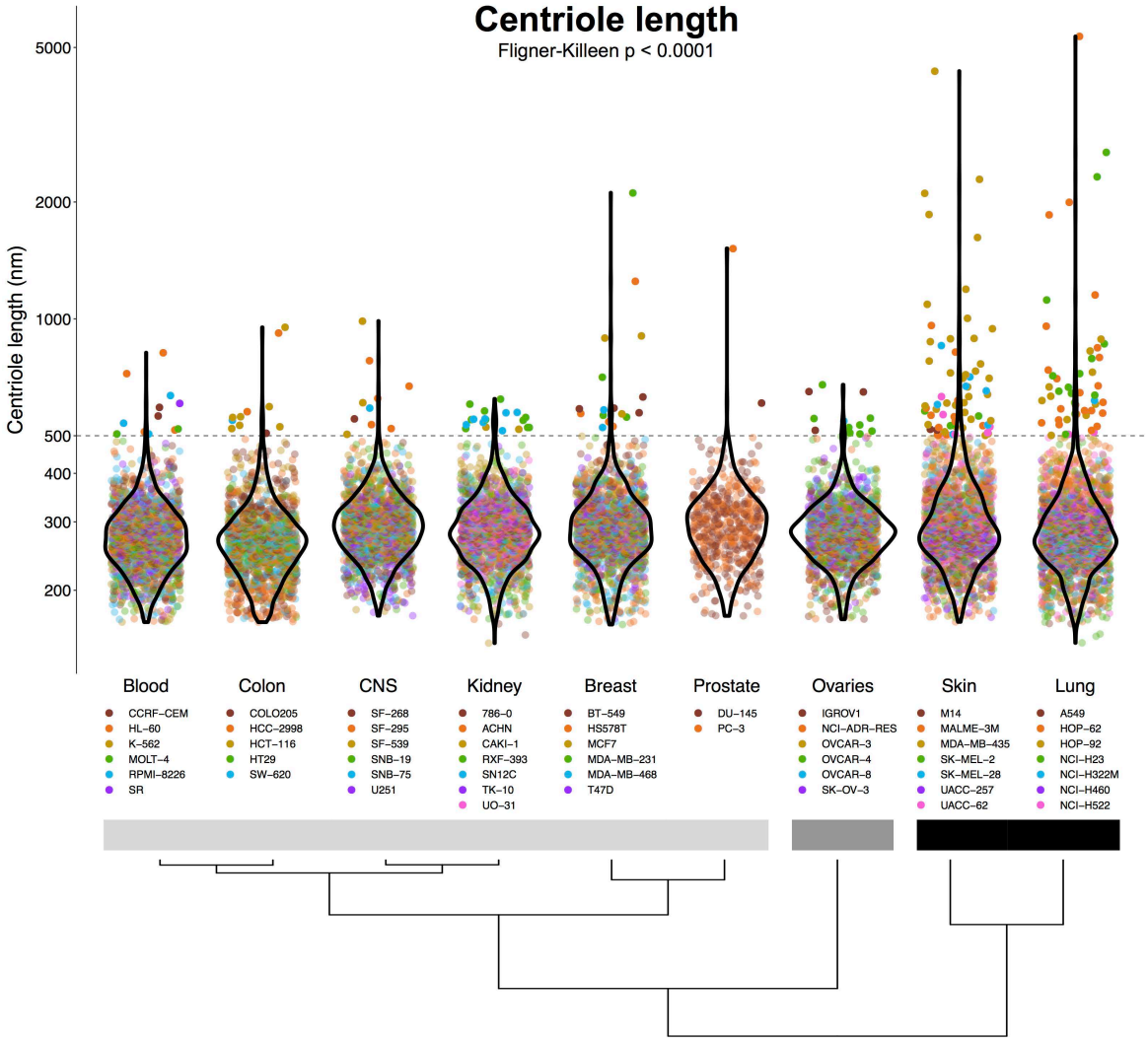


Figure 4.4 Centriole length heterogeneity in cancer. Length of individual centrioles (nm) distributed across the NCI-60 tissues of origin, coloured by cancer cell line. Violin plots of the respective distributions are shown. Tissues are hierarchically clustered (Euclidean distances calculated based on Fligner-Killeen p-values across all combinations of tissue pairs) and the three main clusters are highlighted with bars in different shades of grey. The associated heatmap of distances is shown in *Annex 6b*. The Y-axis is in logarithmic scale. Horizontal dashed line at 500nm represents the stipulated maximal length of a normal-length centriole.

Interestingly, even within individual tissues of origin there were big discrepancies in length variability across cell lines (***Figure 4.4***). Indeed, within skin and lung cancers (those with higher centriole length heterogeneity), different cell lines had different centriole length variances (Fligner-Killeen test $p < 0.0001$ in both cases; ***Figure 4.5***).

Length heterogeneity in skin cancer (***Figure 4.5a***) was mainly due to cell line MDA-MB-435 (statistically significant higher variance than those of all other skin cancer cell lines: Fligner-Killeen test FDR-adjusted p-value < 0.001), also the one with the highest penetrance of COE in the panel (***Figure 4.1d***). Particularly, and contrasting with the higher abnormality levels of the MDA-MB-435 cell line, UACC-257 did not have any overly-long centriole and its centriole length distribution is narrower.

For lung cancer cell lines (***Figure 4.5b***), we observed four cell lines with almost no overly-long centrioles (A549 (0% of cells with COE), NCI-H460 (0%), NCI-H322M (1.9%) and NCI-H522 (3.9%); ***Figure 4.1d***) and three with high COE levels (HOP-92 (17.3%), NCI-H23 (19.6%) and HOP-62 (29.3%); ***Figure 4.1d***), explaining the higher overall COE variability observed in lung cancer cell lines (***Figure 4.1b***). Particularly, cell line HOP-62 was the one showing the highest length heterogeneity (statistically significant higher variance than those of all other lung cancer cell lines: Fligner-Killeen test FDR-adjusted p-value < 0.001; ***Figure 4.5b***).

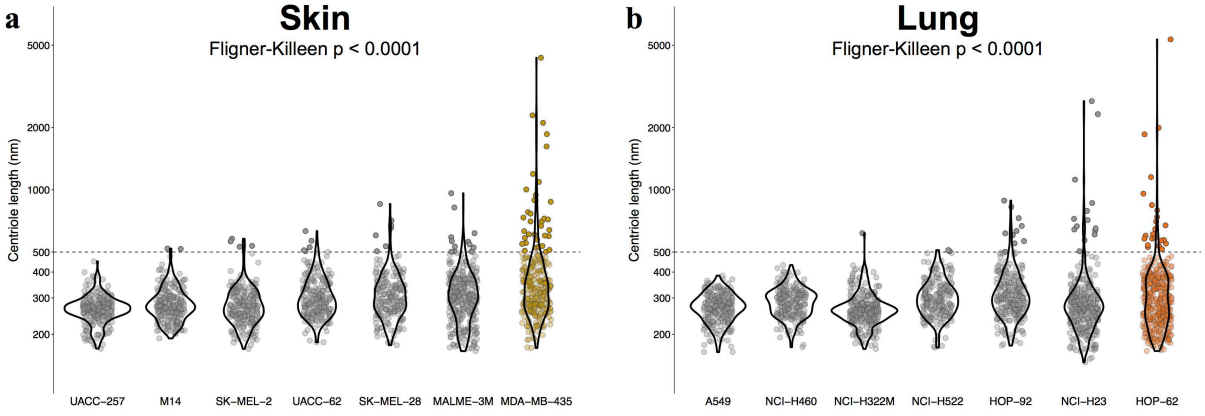


Figure 4.5 Single-centriole length heterogeneity in individual cell lines for selected tissues of origin. Distributions of centriole length (nm) for different cell lines derived from **(a)** skin and **(b)** lung cancer tissues. The respective violin plots are shown. Cell lines are ordered by centriole length variance and the ones with higher length heterogeneity (i.e. showing statistically significant higher variance than those of the remaining cell lines) are highlighted with respective colours from *Figure 4.4*. Y-axis and horizontal dashed line as in *Figure 4.4*.

4.1.2 Aggressive breast and colon cancer cell lines display high levels of CA

We noticed that CA was more prevalent in specific aggressive subtypes of breast and colon cancer (Kocarnik et al., 2015; Parker et al., 2009; Phipps et al., 2015; Sørlie et al., 2001). Cell lines from the basal breast cancer molecular subtype displayed higher CA than luminal ones (***Figure 4.6a***) and, similarly, CA was more frequent in the most common subset of colon carcinoma, CIN (chromosomal instable, microsatellite stable), than in MSI-H cell lines (microsatellite instable, hyper-mutated; ***Figure 4.6b***). Breast cancer findings were subsequently validated in human tissue samples, in Dr Joana Paredes' lab (Instituto de Investigação e Inovação em Saúde; Marteil et al., *manuscript in preparation*), supporting the results of this systematic survey and suggesting that CA specifically occurs in more aggressive tumour subtypes.

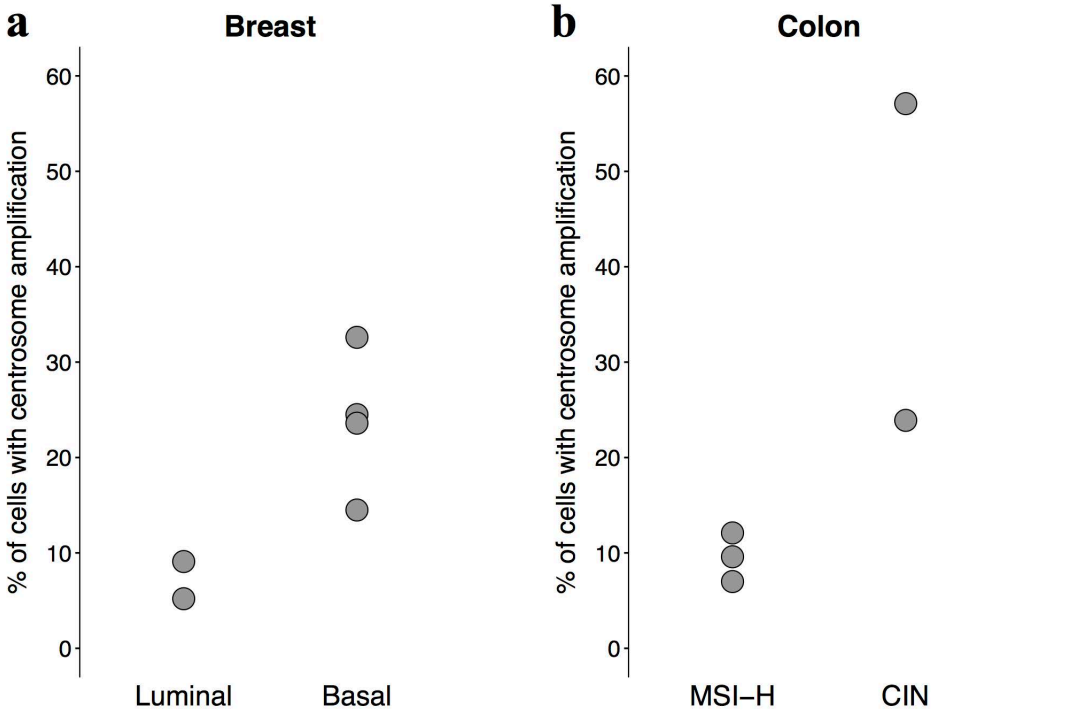


Figure 4.6 NCI-60 cell lines from aggressive breast and colon cancer subtypes display high levels of centrosome amplification. Percentage of cells with CA for each cell line from different **(a)** breast (Luminal and Basal) and **(b)** colon (MSI-H, microsatellite instability, and CIN, chromosomal instability) cancer molecular subtypes. Given the low sample size, no statistical test was performed between subtypes.

4.1.3 COE and CA are not independent

Since COE were shown to promote CA *in vitro* (Marteil et al., *manuscript in preparation*), we then tested the association between both abnormalities in the NCI-60 panel at the cell line level. Indeed, COE was significantly positively correlated with CA (Spearman

correlation coefficient: 0.4, $p < 0.01$; ***Figure 4.7a***). We have also observed the existence of cell lines with high levels of CA but reduced COE (bottom right corner of ***Figure 4.7a***), contrasting with the absence of cell lines with high COE but low CA (top left corner of ***Figure 4.7a***).

However, this association was different between tissues of origin (ANCOVA $p < 0.05$; ***Figure 4.7b***), likely because only skin, lung and kidney cancer tissues had enough variability on COE levels (***Figure 4.1b***). In summary, for tissues with variability at both number and length abnormality levels, COE was associated with CA.

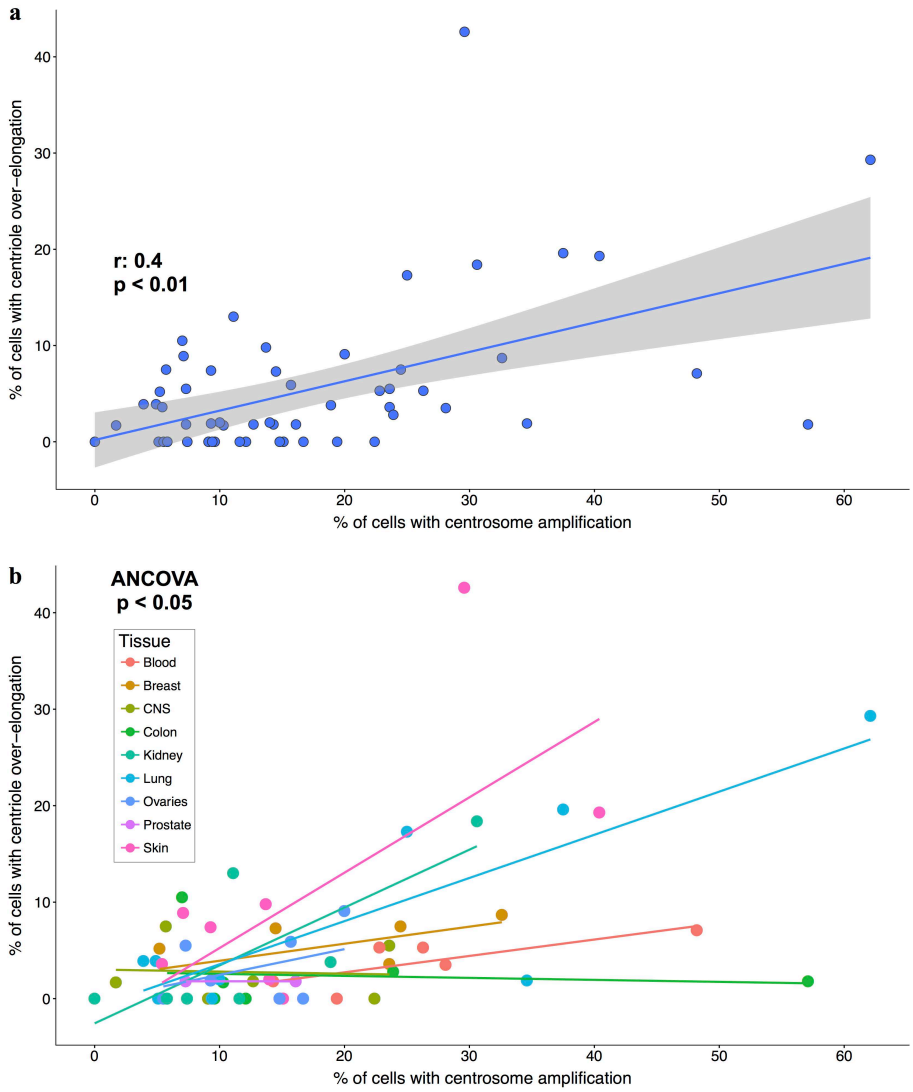


Figure 4.7 Centriole over-elongation was positively correlated with centrosome amplification. **a)** Comparison between percentage of cells with CA and percentage of cells with COE on NCI-60 cell lines (Spearman correlation coefficient: 0.4, $p < 0.01$). The grey shade around the blue linear regression line represents its 95% confidence interval. **b)** The same as a) but with cell lines coloured according to their tissue of origin and a regression line represented for each tissue. Regressions were significantly different between tissues (ANCOVA test $p < 0.05$).

We also examined if there was an association between both abnormalities at the single-centriole level. Indeed, we observed statistically significant differences in centriole length between cells with different numbers of centrioles (Kruskal-Wallis rank-sum test $p < 0.0001$; ***Figure 4.8***).

One cell with 32 centrioles, from the skin cell line MDA-MB-435, exibhited a particular enrichment of overly-long centrioles. This cell had a centriole length median of 420nm and 34% (11 out of 32) of overly-long centrioles, contrasting respectively with 280nm (Wilcoxon rank-sum test $p < 0.0001$) and 1.54% (Chi-squared test of independence $p < 0.0001$) throughout the dataset.

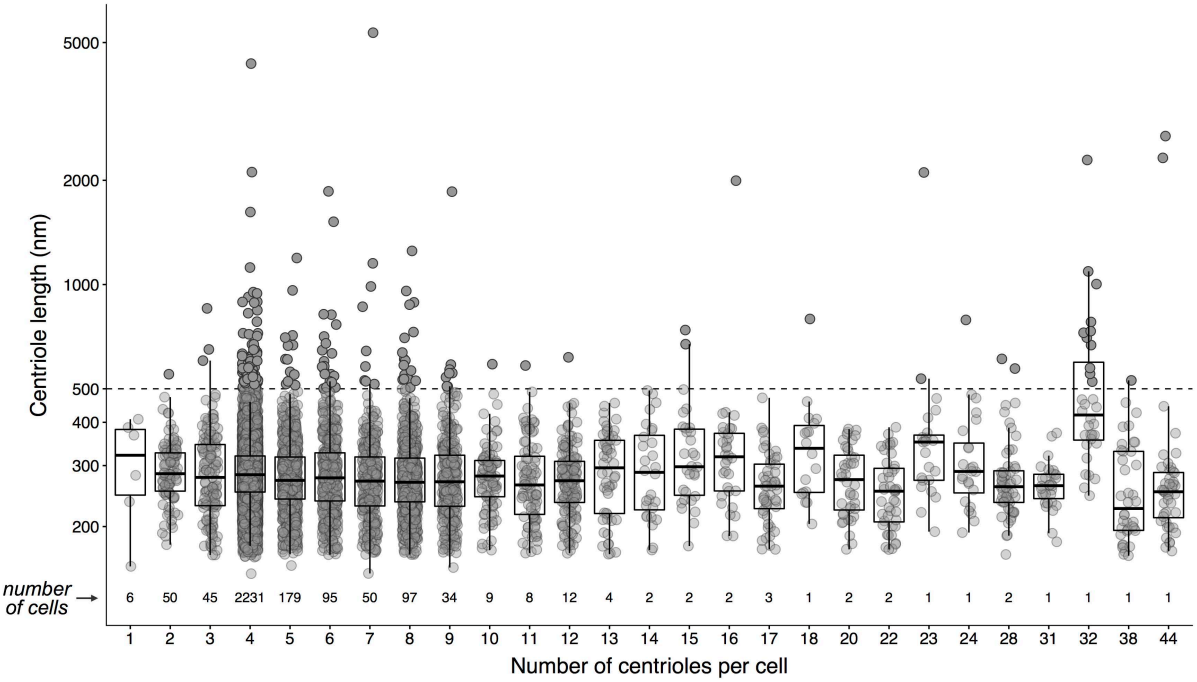


Figure 4.8 Distribution of centriole length according to the number of centrioles per cell. Comparison of single-centriole length (nm) across cells with different numbers of centrioles (12,927 centrioles from 2,842 cells; Kruskal-Wallis rank-sum test $p < 0.0001$). Box plots and the number of cells per group are shown. Y-axis and horizontal dashed line as in *Figure 4.4*.

We then hypothesized that CA and COE were related at single-centriole level and tested the null hypothesis of independence between abnormalities, where one would expect the same proportion of overly-long centrioles within cells with or without CA (1.54% of centrioles in each group). However, we observed a statistically significant higher proportion of overly-long centrioles in cells with CA (2.53%, compared with 1.13% in cells without CA; Chi-squared test of independence $p < 0.0001$; ***Figure 4.9***). This result led us to reject the null hypothesis

and accept the alternative one that COE and CA are not independent, i.e. there is a relation between those abnormalities.

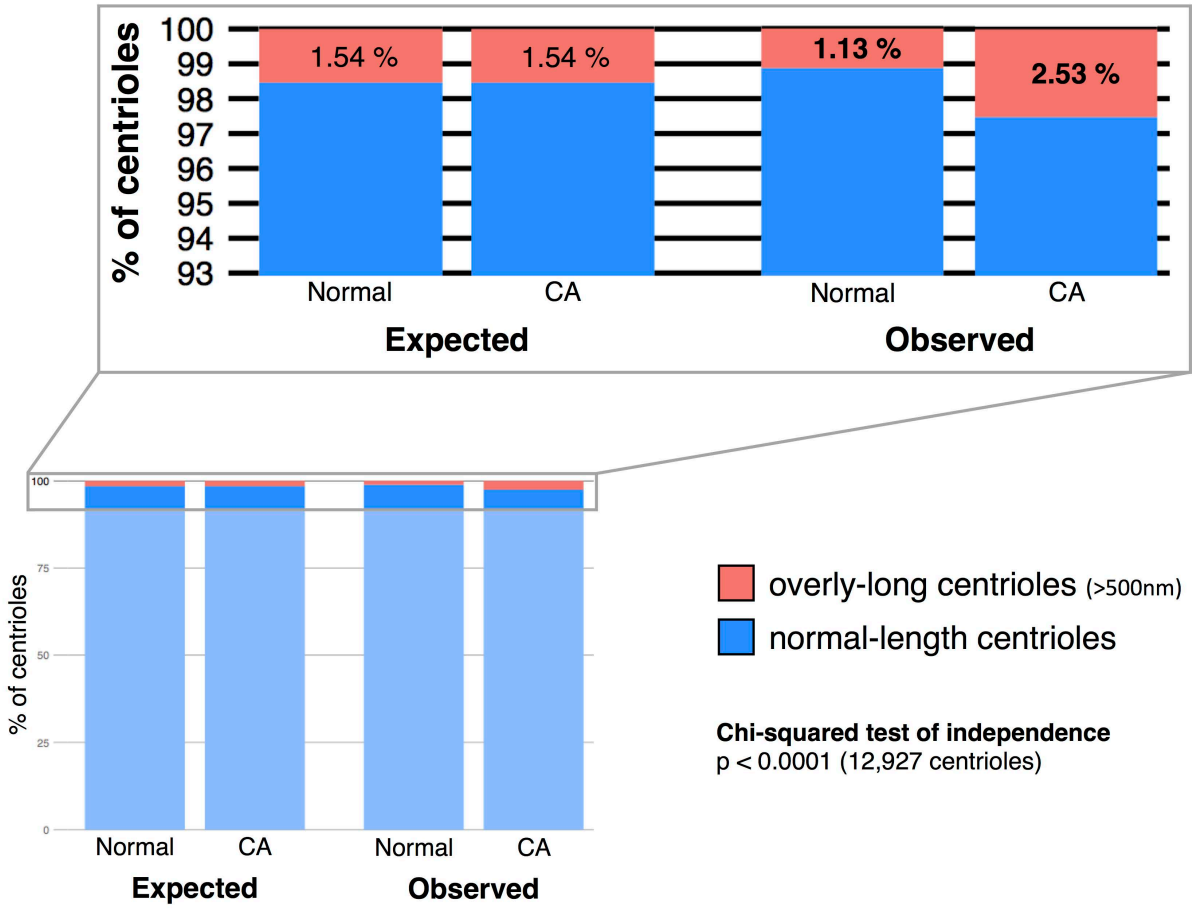


Figure 4.9 Centriole over-elongation and centrosome amplification are not independent. Higher proportion of overly-long centrioles (>500nm) in cells with CA (12,927 total centriole number; $p < 0.0001$, Chi-squared test of independence), compared with the expected proportions under the null hypothesis of independence. The expected and observed percentages of overly-long centrioles in cells with normal centriole number (Normal) and CA are shown, together with a detailed view above, given the low frequency of overly long centrioles.

4.1.4 COE is not a stochastic event in cancer cells

In the present dataset, we observed only 1.54% (199 out of 12,927) of overly-long centrioles, within 5.88% (167 out of 2,842) of all cells (***Figure 4.10***). More precisely, 5.14%, 0.63% and 0.13% of cells (146, 18 and 3 out of 2,842, respectively) had one, two, and more than two overly-long centrioles, respectively. The lower frequency of COE suggests it as a rare phenomenon in cancer.

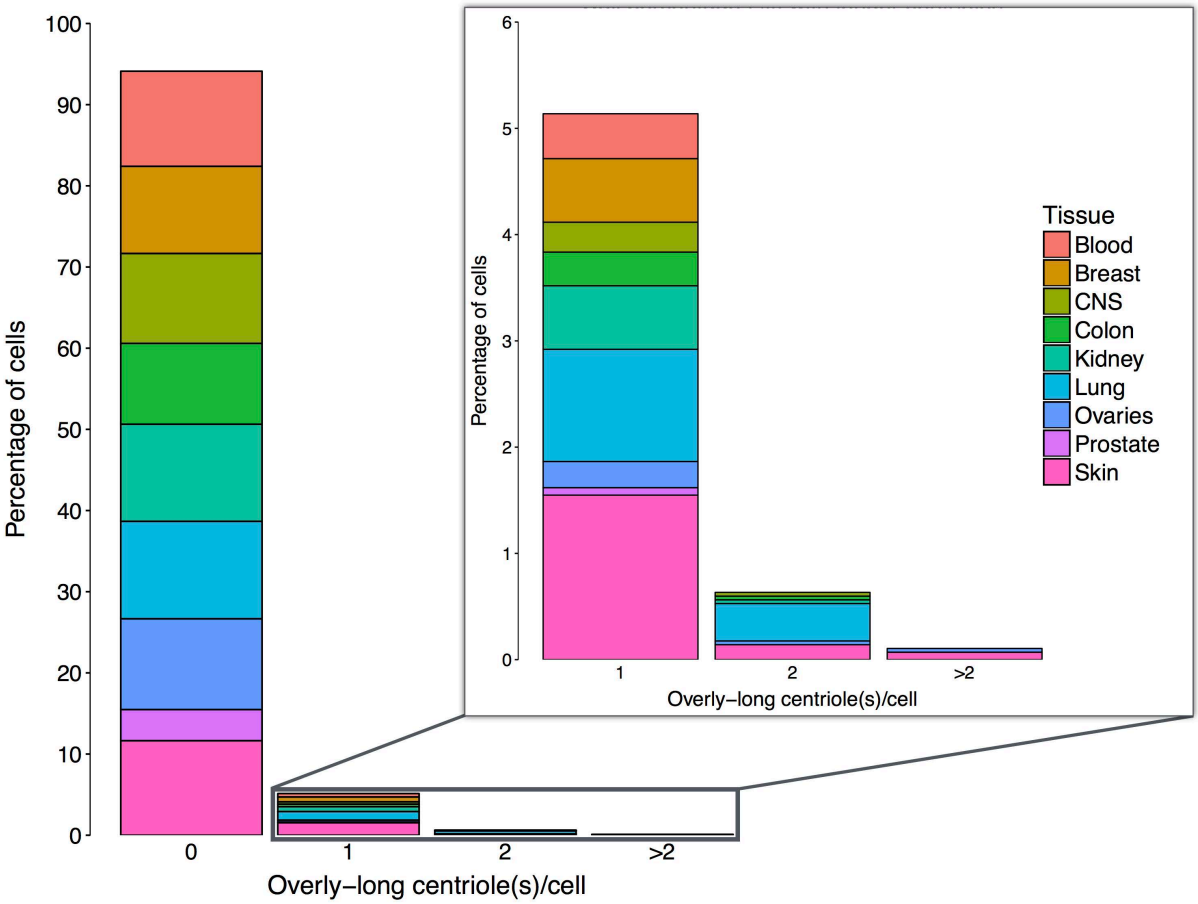


Figure 4.10 Centriole length dysregulation does not affect all centrioles within the same cell. Distribution of the number of overly-long centrioles (0, 1, 2 and more than 2) in each cell from the secondary screening, coloured by tissue of origin. Most cells with centriole length abnormalities (detailed view on box) only have one centriole affected by over-elongation.

To determine whether the COE distribution observed in this cancer panel is similar to a typical one of a rare event, we first generated a random distribution of COE per cell (***Figure 4.11a,b - Expected***). This distribution represented the expected frequencies of over-elongation within cancer cells, as if it was a stochastic event. Then we compared the observed frequencies (***Figure 4.11a,b - Observed***) with the expected ones, under the null hypothesis that both frequencies are equal. We found a significant difference between the observed frequencies of COE and the expected ones (Chi-squared goodness of fit test $p < 0.0001$; ***Figure 4.11***), suggesting that this feature is not a completely stochastic event in cancer cells. Particularly, we observed a higher frequency of cells with more than one overly-long centriole than what is expected by chance (***Figure 4.11b***).

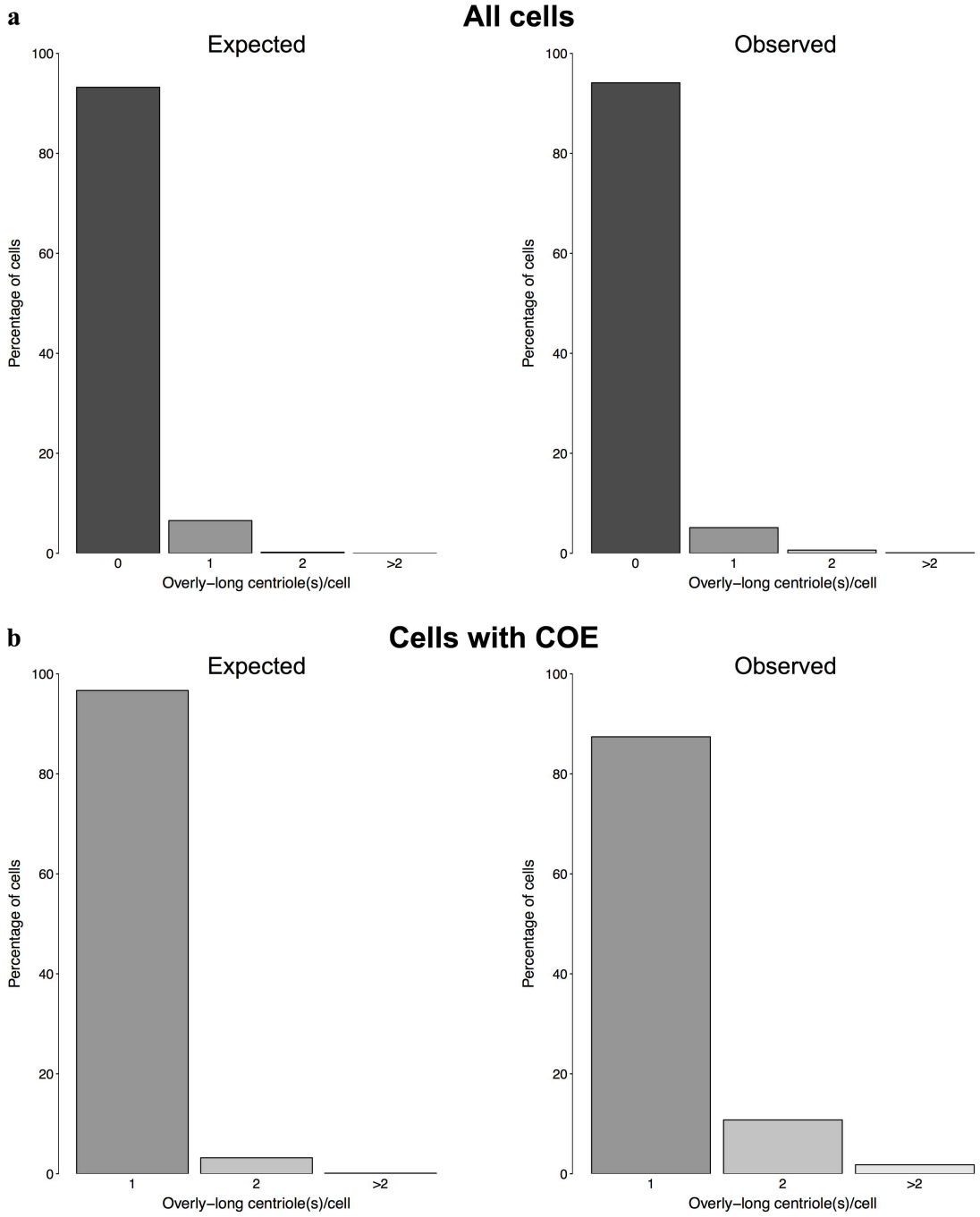


Figure 4.11 Centriole over-elongation is not a stochastic event in cancer cells. Comparison of the expected (calculated from a random distribution) and observed distributions of the number of overly-long centrioles (0, 1, 2 and more than 2) per cell within **(a)** all cells or **(b)** cells with COE. Significant differences were observed: Chi-squared goodness of fit test $p < 0.0001$.

4.1.5 COE is cell state-dependent

Since COE was not deemed to be a completely stochastic event, with the frequency of cells with more than one overly-long centriole being higher than expected by chance (***Figure 4.11b***), we investigated if the cell state (physiological condition of a given cell) could be influencing the frequency of this abnormality. Namely, our hypothesis was that cells with

already one overly-long centriole had a tendency to have another overly-long centriole. To test this hypothesis, we used the binomial test, where we considered a cell with more than one overly-long centriole as a success, within all cells with COE (i.e. at least one overly-long centriole). Then we compared the number of successes observed in the panel with a hypothesised probability of success calculated from the COE random distribution. We observed a significantly higher proportion of successes in the observed distribution (one-tailed Binomial test $p < 0.0001$; ***Figure 4.11b***), meaning that COE occurred more frequently within the same cell and, thus, is a cell state-dependent feature.

However, this observation can differ between cells with four centrioles and cells with CA. To address this question, we took the same approach, but now independently for each of those groups of cells. For each group, we have generated a random distribution of COE per cell and calculated the expected frequencies of cells with one, two, and more than two overly-long centrioles (***Figure 4.12***). Then, we again performed binomial tests. We observed, in both cases (within the 95 cells with four centrioles and COE, and within the 68 cells with CA and COE), a significantly higher proportion of cells with more than one overly-long centriole than expected by chance (one-tailed Binomial test $p < 0.05$ and $p < 0.001$, respectively; ***Figure 4.12***).

Regarding cells with four centrioles and COE, five of the 95 (5.26%) exhibited two overly-long centrioles and no cells were found with more than two. The latter could be explained by the insufficient number of observations, since, by chance, one would expect 0.01% of cells to have more than two overly-long centrioles, i.e. less than one cell out of 95 (frequencies from the random distribution; ***Figure 4.12a***). 16 of the 68 (23.5%) cells with both CA and COE exhibited two or more overly-long centrioles, contrasting with the expected 9.62% (***Figure 4.12b*****).** These observations suggest COE as a widely cell state-dependent feature.

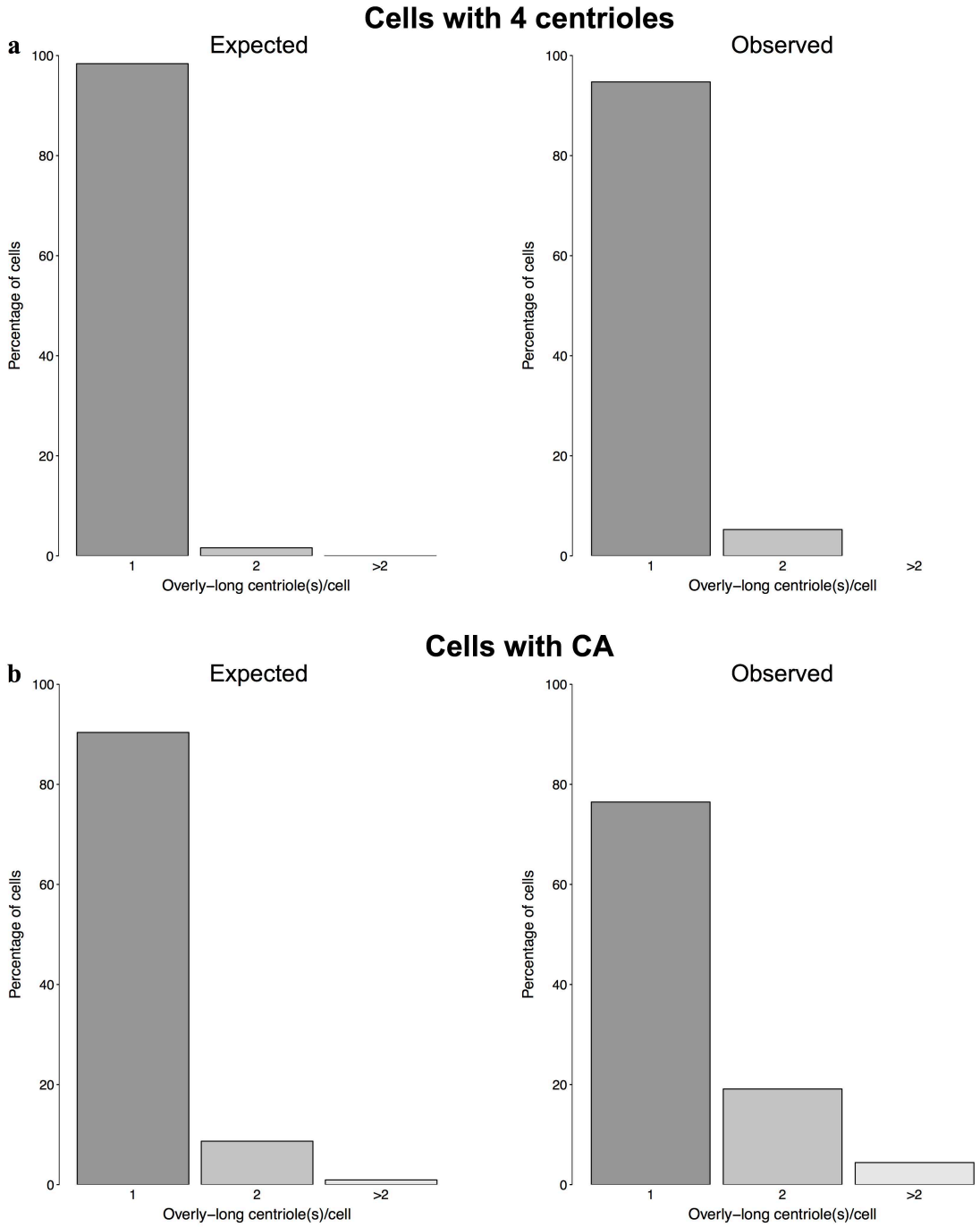


Figure 4.12 Centriole over-elongation is a cell state-dependent feature. Higher proportion of cells (Observed) with more than one overly-long centriole than expected by chance (calculated from a random distribution) within **(a)** cells with four centrioles and COE (one-tailed Binomial test $p < 0.05$) and **(b)** cells with both CA and COE (one-tailed Binomial test $p < 0.01$).

4.1.6 Total centriolar mass per cell is apparently not controlled

Given this phenomenon of centriole length dysregulation in cancer, we then tested if cancer cells still control their centriolar mass (given by the sum of centrioles' length) when centriole number increases. If they control it, one would expect constant centriolar mass in cells with supernumerary centrosomes, i.e no statistical association between the number of

centrioles and the sum of centrioles' length. However, we observed that centriolar mass increased monotonically with the number of centrioles within the cell (Spearman correlation coefficient: 1, $p < 0.0001$; correlation between mean of centriolar mass, per group of cells with the same number of centrioles, and this number of centrioles per cell), suggesting that cancer cells do not control their centriolar mass (***Figure 4.13a***).

Next, we investigated if there was any evidence that CA arose from centriole fragmentation, one of the mechanisms proposed by Marteil and co-workers (Marteil et al., *manuscript in preparation*). If this hypothesis was true, one would expect that cells with CA had shorter centrioles, putatively resulting from fragments, and the mean of centriole length per cell would be negatively associated with its centriole number. We found no association between those variables (Spearman correlation coefficient: -0.01, p: n.s. (not significant); correlation between mean of centriole length, per group of cells with the same number of centrioles, and this number of centrioles per cell; ***Figure 4.13b***). Since the mean of centriole length was maintained in cells with supernumerary centrosomes, there was no evidence, in this panel of cell lines, that CA arose from centriole fragmentation.

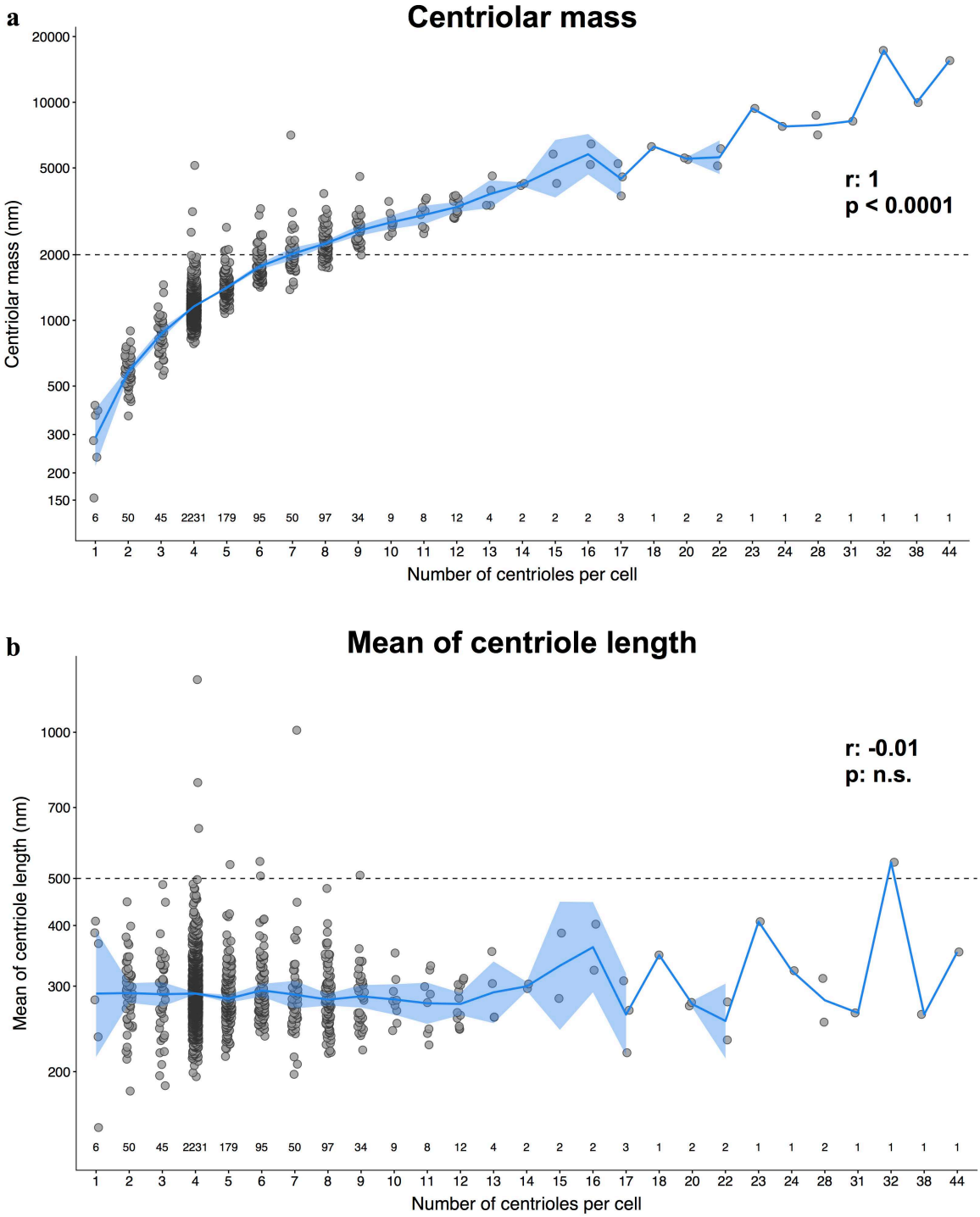


Figure 4.13 Centriolar mass increased in association with the number of centrioles within the cell, while centriole length mean was maintained. Distribution of **(a)** centriolar mass (sum of centrioles' length per individual cell; nm) and **(b)** mean of centriole length (nm) across cells with different number of centrioles (2,842 cells; Spearman correlation coefficient: 1 ($p < 0.0001$) and -0.01 (p: n.s.), respectively). Lines connecting group means (blue shades represent their associated 95% confidence interval) and the numbers of cells per group (at the bottom) are shown. Horizontal dashed lines at **(a)** 2000nm and **(b)** 500nm represent the stipulated maximums of centriolar mass (as in cells with four normal-length centrioles, i.e. 4 times 500nm) and length of a normal centriole, respectively. The Y-axis is in logarithmic scale.

4.2 Discovery of novel molecular origins of centriole abnormalities in cancer

The molecular origins of centriole abnormalities in cancer were studied by integrating their prevalence with publicly available molecular data for the NCI-60 panel. In addition to the percentage of cells with CA and COE metrics, that reflect the prevalence of these abnormalities at the cell line level, we used the single-centriole data to generate two other metrics that characterize the centriolar landscape per cell line: the means of centriole number and length per cell (***Annex 1***). Together, the four metrics complementarily depicted different features of centriolar abnormalities and were used in the following analyses.

4.2.1 Cancer cell lines with overly-long centrioles have more cells in G1

We started by investigating a putative relationship between centriole abnormalities and cell proliferation rates, taking advantage of the publicly available data on NCI-60 cell line doubling times (inversely proportional to proliferation rates). Through correlation analyses, we found no association between doubling time and centriole number (Spearman correlation coefficient: 0.18 (p: n.s.) and 0.2 (p: n.s.) for percentage of cells with CA and mean of centriole number, respectively; ***Figure 4.14a***), but a positive correlation with both the percentage of cells with COE and the mean of centriole length (Spearman correlation coefficient: 0.28 ($p < 0.05$) and 0.4 ($p < 0.01$), respectively; ***Figure 4.14b***). In summary, cancer cell lines with overly-long centrioles showed lower proliferation rates.

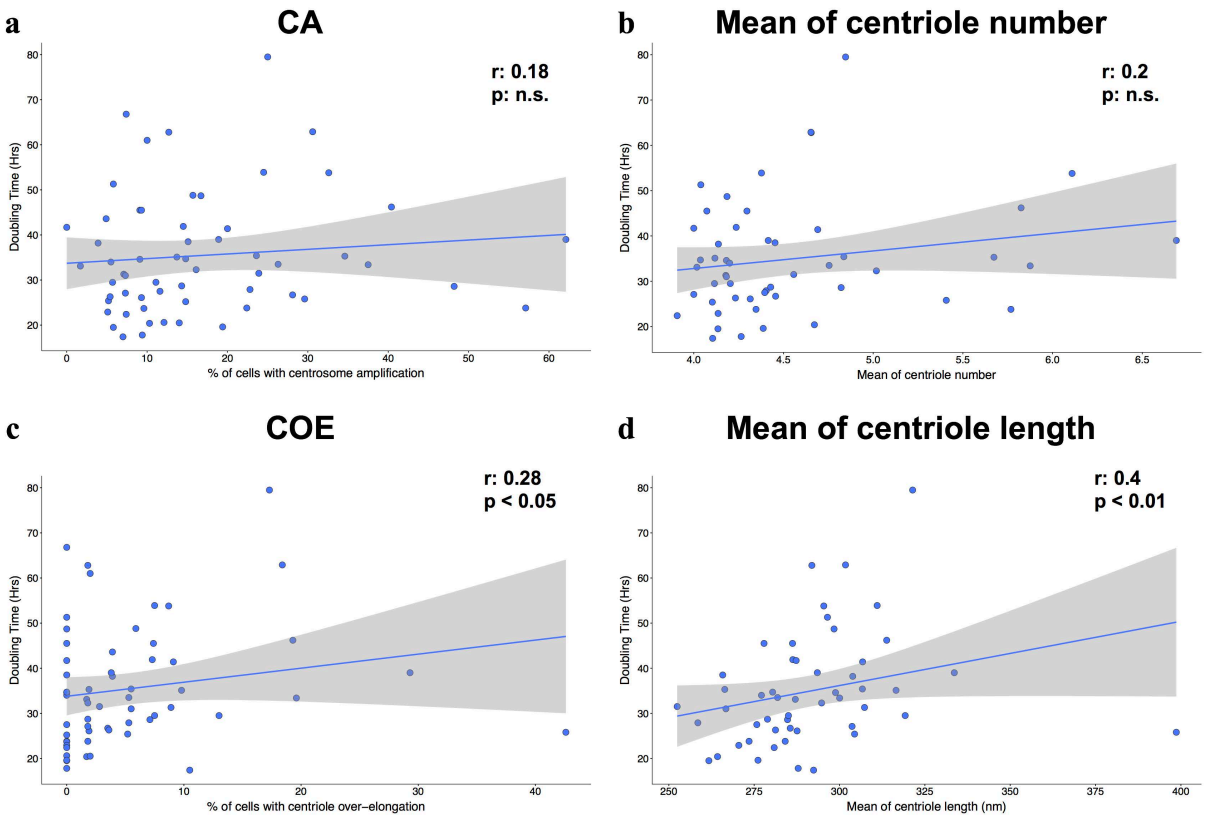


Figure 4.14 Cancer cell lines with overly-long centrioles had lower proliferation rates. Relation between cell line doubling time (hours) and **(a)** percentage of cells with CA, **(b)** mean of centriole number, **(c)** percentage of cells with COE or **(d)** mean of centriole length (nm), for NCI-60 cancer cell lines (Spearman correlation coefficient: 0.18 (p: n.s.), 0.2 (p: n.s.), 0.28 ($p < 0.05$) and 0.4 ($p < 0.01$), respectively). Grey shades around linear regression lines as in *Figure 4.7a*. Higher doubling time means lower proliferation rate.

With this result in mind, we wondered how is the cell cycle time associated with centriole length? Is this association specific of a particular cell division phase? To answer these questions, we used the FACS data provided by *O'Connor et al., 1997*, where, for each cell line, the percentages of cells in each of the G1, S and G2/M phases were estimated. We then analysed the correlation between those cell proportions and both the centriole length metrics. Overall, we observed a positive correlation with the percentage of cells in G1 phase (Spearman correlation coefficient: 0.28 ($p < 0.05$) and 0.26 (p: n.s.), for percentage of cells with COE and mean of centriole length, respectively; ***Figure 4.15a,b***), whereas no association was found for the other phases (***Figure 4.15c-f***). Thus, COE was associated with a longer G1.

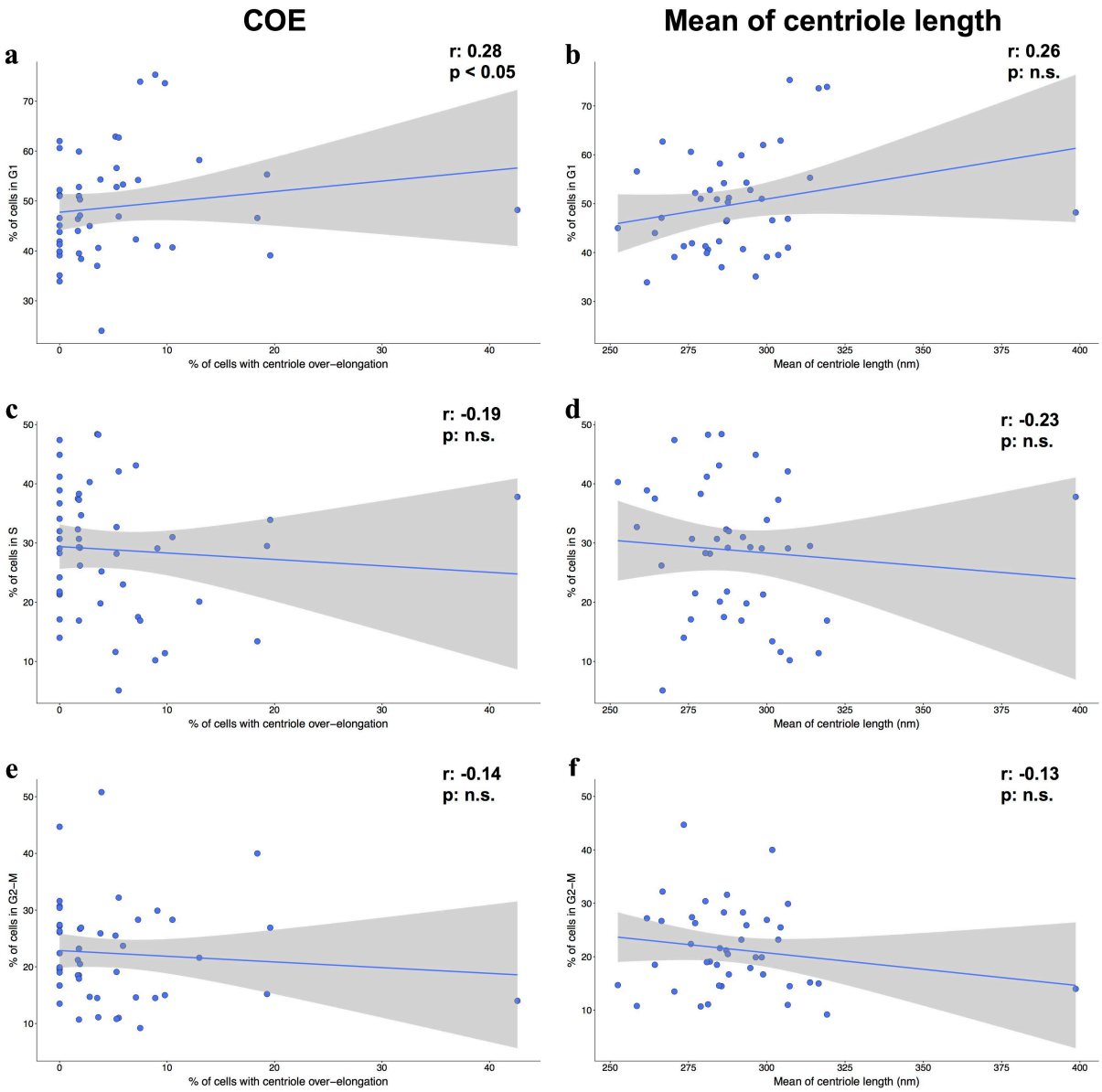


Figure 4.15 Cancer cell lines with overly-long centrioles have more cells in G1. Relation between percentage of cells in **(a,b)** G1, **(c,d)** S or **(e,f)** G2/M cell cycle phases and **(a,c,e)** percentage of cells with COE or **(b,d,f)** mean of centriole length (nm). Spearman correlation coefficient: **(a)** 0.28 ($p < 0.05$), **(b)** 0.26 (p: n.s.), **(c)** -0.19 (p: n.s.), **(d)** -0.23 (p: n.s.), **(e)** -0.14 (p: n.s.) and **(f)** -0.13 (p: n.s.). Grey shades around linear regression lines as in *Figure 4.7a*.

Correlation does not mean causation, so we then hypothesized that centriole length defects may cause a cell cycle delay in G1, which would be expected to be stronger for cells that still have an intact p53 response. To test this hypothesis, we compared the correlation between centriole length metrics and the percentage of cells in G1 across cell lines with wild-type (WT) and mutated (MT) *TP53*. However, we did not find statistically significant different associations between *TP53* statuses (ANCOVA p: n.s. for both metrics; ***Figure 4.16***).

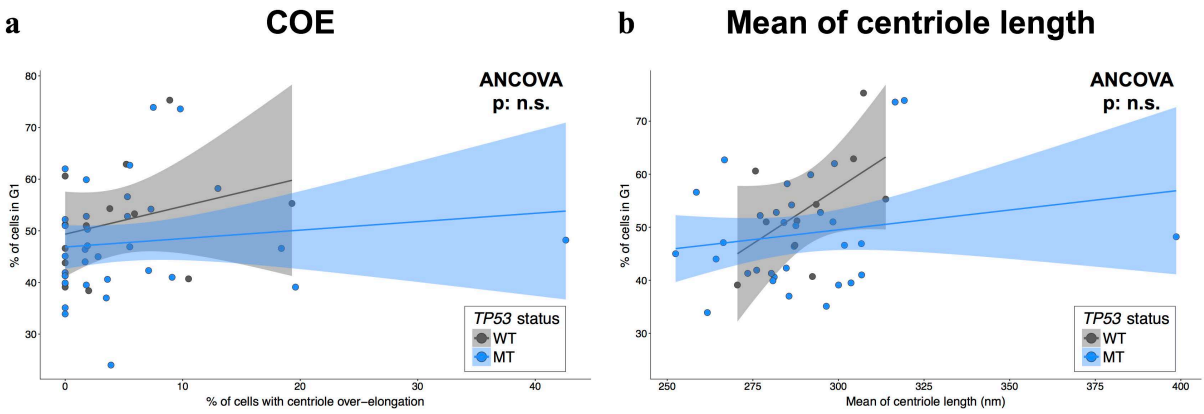


Figure 4.16 *TP53* status was not associated with the relation between centriole length defects and cell cycle delay in G1. Relation between percentage of cells in G1 phase and **(a)** percentage of cells with COE and **(b)** mean of centriole length, according to *TP53* status (WT: wild-type, MT: mutated; ANCOVA test p: n.s. for both). The grey/blue shades around the grey/blue linear regression lines represent their 95% confidence intervals for WT/MT.

4.2.2 PRKACA is a putative promoter of COE in cancer

In order to identify putative centriole length regulators, we performed correlation analyses between gene and protein expression levels of centrosomal genes, retrieved from the CentrosomeDB - Centrosomal Proteins Database (Alves-Cruzeiro et al., 2014; Nogales-Cadenas et al., 2009), and the two centriole length metrics, in the NCI-60 panel. From the 870 centrosomal genes tested, we identified only one gene significantly correlated with the mean of centriole length, whereas none was associated with the percentage of cells with COE (FDR-adjusted p-value lower than 0.05). Regarding protein expression, we did not find any protein significantly correlated with any of the centriole length metrics (FDR-adjusted p-value lower than 0.05). However, we also did not find any significant association between centriole length and *CPAP* gene and protein levels, or *CP110* gene expression levels (no protein data available), two known centriole length regulators (Schmidt et al., 2009), suggesting that centriole length dysregulation in cancer could be associated with different mechanisms.

The candidate identified in these analyses as being associated with centriole length in cancer was *PRKACA*. This gene encodes one of the catalytic subunits of protein kinase A (PKA), a family of kinases that phosphorylate many substrates in the cytoplasm and the nucleus and whose activity is dependent on cellular levels of cyclic AMP (cAMP; Skålhegg and Tasken, 2000). PKA was found to localize at the centrosome (Nigg et al., 1985) and its higher activation is associated with increased primary cilium length in mammalian cells (Besschetnova et al., 2010). However, there is no previous information about a possible relationship between this kinase, or even this specific gene, and centriole length regulation. In the NCI-60 panel, *PRKACA* was positively correlated with the mean of centriole length

(Spearman correlation coefficient: 0.58, p < 0.0001; ***Figure 4.17a***), suggesting that higher expression of this gene is also associated with increased centriole length in cancer. Regarding protein expression, correlation with centriole length was not so significant (Spearman correlation coefficient: 0.27, p: 0.07; ***Figure 4.17b***).

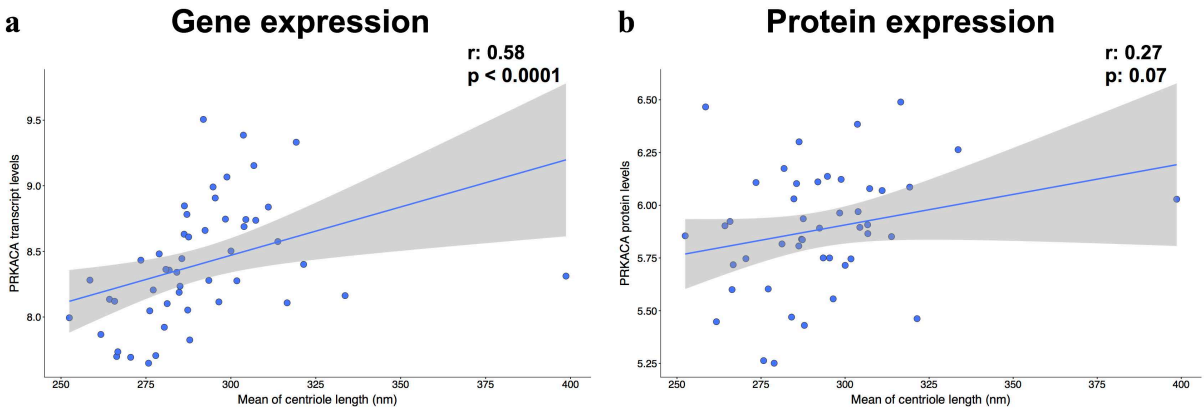


Figure 4.17 *PRKACA* gene expression levels were positively correlated with centriole length. Relation between *PRKACA* **(a)** gene or **(b)** protein expression levels and mean of centriole length (nm; Spearman correlation coefficient: 0.58 (p < 0.0001) and 0.27 (p: 0.07), respectively). Grey shades around linear regression lines as in *Figure 4.7a*.

To further explore the potential role of *PRKACA* in centriole length, we took advantage of an independent screen for putative centriole length regulators, done in Dr Mónica Bettencourt-Dias' lab (*unpublished data*), where this gene was already tested. Here, *PRKACA* was knocked down using two different siRNAs and then CEP135 and centrin protein intensity levels were measured as readouts for centriole length. Indeed, in one siRNA experiment, *PRKACA* knock down was found to significantly decrease centrin protein levels (-2.27 robust z-score against the control; t-test p < 0.05), while no significant change was observed in CEP135 intensity (siRNA-1 on ***Figure 4.18***), highlighting *PRKACA* as a putative promoter of COE in cancer. However, the second experiment (siRNA-2 on ***Figure 4.18***) did not show any significant changes in both markers.

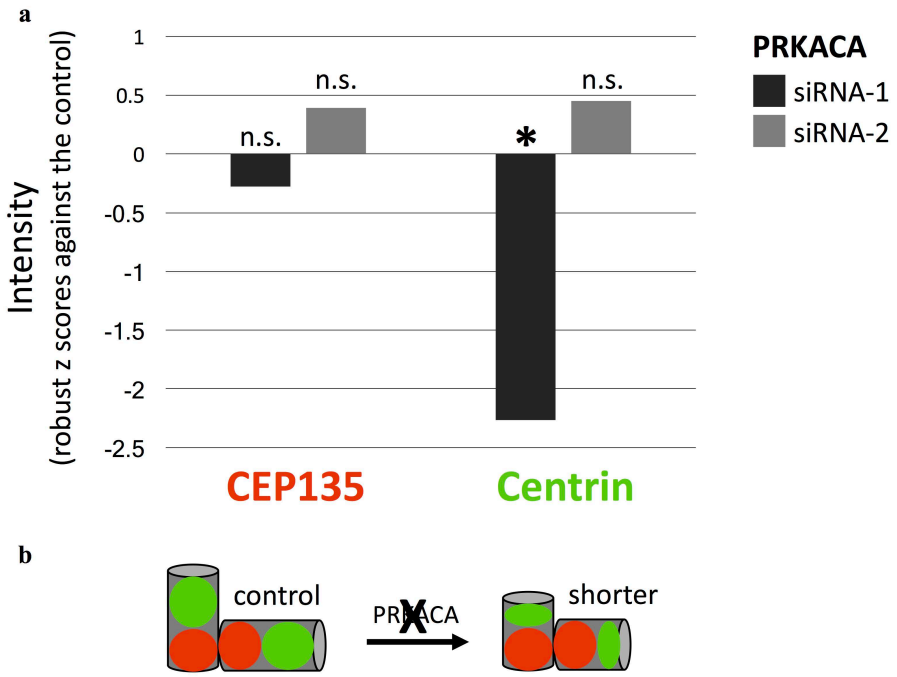


Figure 4.18 *PRKACA* knock down decreased centriole length. a) *PRKACA* expression levels were knocked down using two different siRNAs (siRNA-1 and siRNA-2), followed by CEP135 (red) and centrin (green) protein intensity levels measurement. siRNA-1 (black) was found to significantly decrease centrin intensity levels (-2.27 robust z-score against the control; t-test $p < 0.05$), but not CEP135 intensity (t-test p: n.s.). In turn, siRNA-2 (grey) did not affect centriole length (t-test p: n.s. for both CEP135 and centrin changes). **b)** Depletion of *PRKACA* led to shorter centrioles in the siRNA-1 experiment.

4.2.2.1 Increased centriole length is associated with higher interaction with ECM and lower efficiency in DNA repair

Since we did not observe strong transcriptional changes associated with centriole length abnormalities at the level of individual genes, we performed GSEA to identify gene sets putatively associated with centriole length abnormalities in cancer.

In the present work, GSEA results on the transcriptomic changes associated with both centriole length metrics could be confounded with the cell line doubling time associated ones, since they are positively correlated across the NCI-60 cell lines (***Figure 4.14c,d***). Therefore, for each length metric, we used multiple linear regression models to decouple the effects of length and doubling time (explanatory variables) on gene expression and determine which genes are independently associated with each length metric. Afterwards, the new lists of genes associated with those two metrics (ranked by the moderated t-statistic for differential expression) were used in GSEA (***Figure 4.19***).

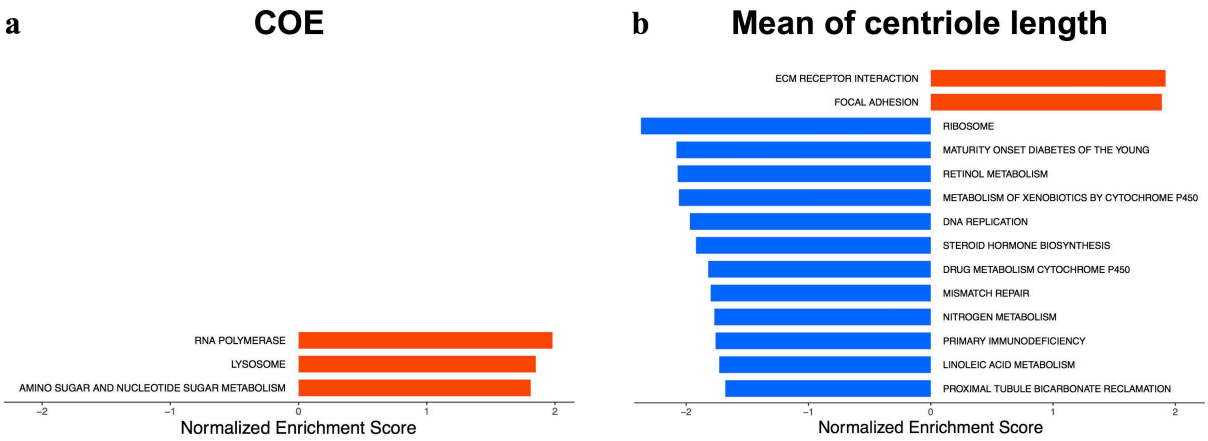


Figure 4.19 KEGG pathways associated with centriole length abnormalities in cancer. GSEA results on the transcriptomic changes associated with **(a)** percentage of cells with COE and **(b)** mean of centriole length. Pathways positively (red) and negatively (blue) enriched, with a FDR lower than 5%, are shown.

GSEA identified three KEGG pathways enriched in genes positively correlated with the percentage of cells with COE levels: amino sugar and nucleotide sugar metabolism, lysosome and RNA polymerase (***Figure 4.19a***). Furthermore, GSEA on the transcriptomic changes associated with the mean of centriole length (***Figure 4.19b***) identified ECM (extracellular matrix) receptor interaction and focal adhesion (a type of adhesive contact between the cell and ECM) pathways (***Figure 4.20a***) as positively associated with centriole length, whereas multiple pathways related with the ribosome, cell metabolism (retinol, xenobiotics, drugs, nitrogen and linoleic) and DNA repair (DNA replication and mismatch repair; ***Figure 4.20b***) were found enriched in genes negatively correlated with centriole length.

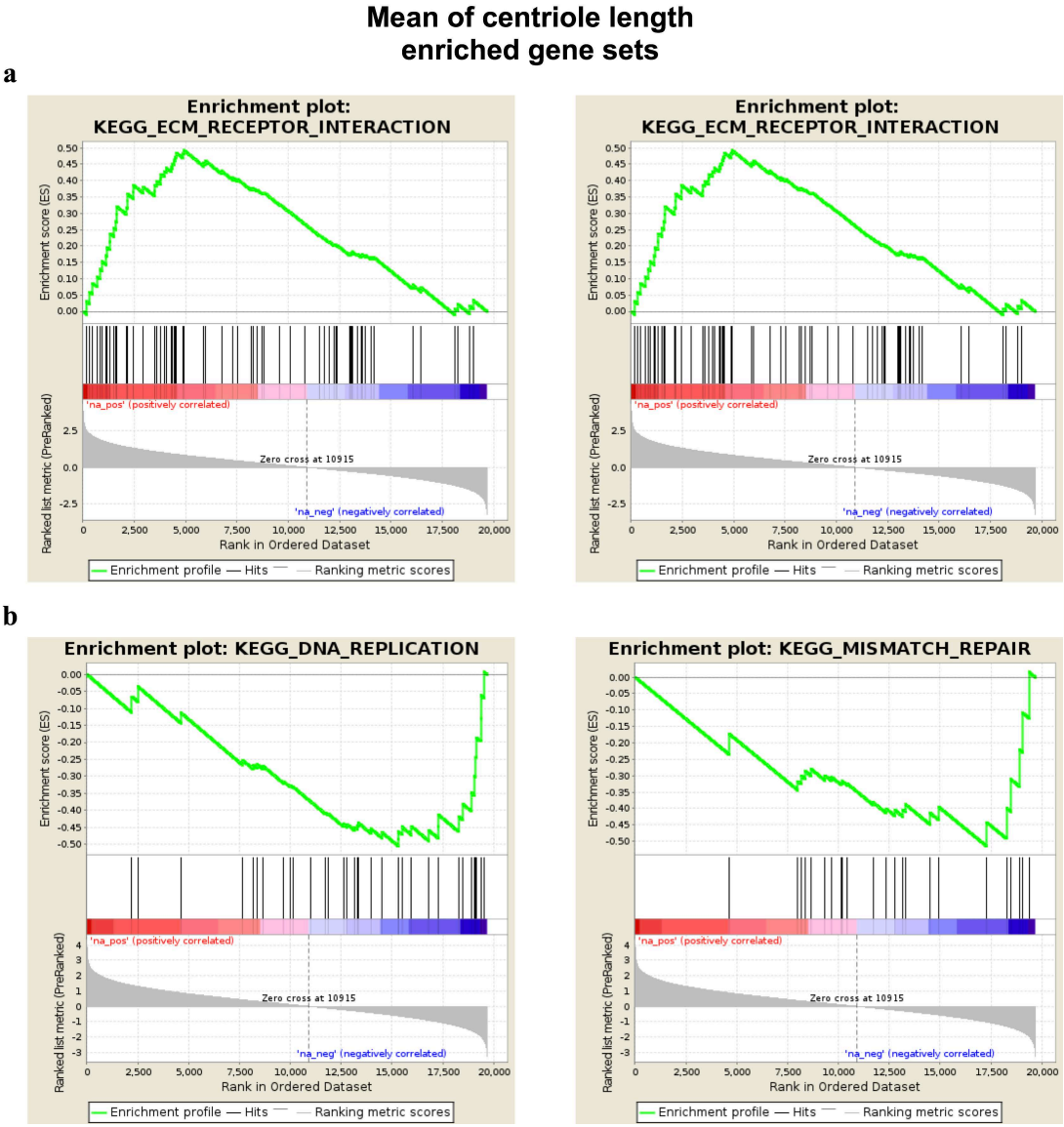


Figure 4.20 Examples of KEGG pathways associated with the mean of centriole length in cancer. Enrichment plots of pre-selected pathways **(a)** positively (ECM receptor interaction (NES: 1.92, FDR < 0.05) and focal adhesion (NES: 1.89, FDR < 0.05)) and **(b)** negatively (DNA replication (NES: -1.97, FDR < 0.01) and mismatch repair (NES: -1.8, FDR < 0.05)) enriched in genes overexpressed in cell lines with overly-long centrioles (GSEA enrichment plot explained in Materials and Methods, section 3.2.2).

4.2.3 Cancer cells with less proteasome activity are more susceptible to CA

Similarly to centriole length analyses, we performed correlation analyses between centriole number abnormality metrics and the 870 centrosomal genes, at both gene and protein level. From these analyses, we did not find any gene significantly correlated with any of the centriole number metrics (FDR-adjusted p-value lower than 0.05). At the protein expression level, one was significantly correlated with the mean of centriole number (FDR-adjusted p-value lower than 0.05). Yet, we also did not find any significant association with *PLK4* (protein levels were not available), *STIL* and *SAS*-6 gene and protein levels in this panel.

The identified candidate was protein PSMD1, whose expression levels were negatively correlated with centriole number (Spearman correlation coefficient: -0.53, $p < 0.0001$; ***Figure 4.21***). This result suggests this protein as putatively protecting cells from CA. Since we did not find this gene's transcript levels to vary with centriole number, PSMD1 regulation in cancer is proposed to be post-transcriptional. This protein is part of the proteasome complex and is responsible for substrate recognition and binding (Coux et al., 1996).

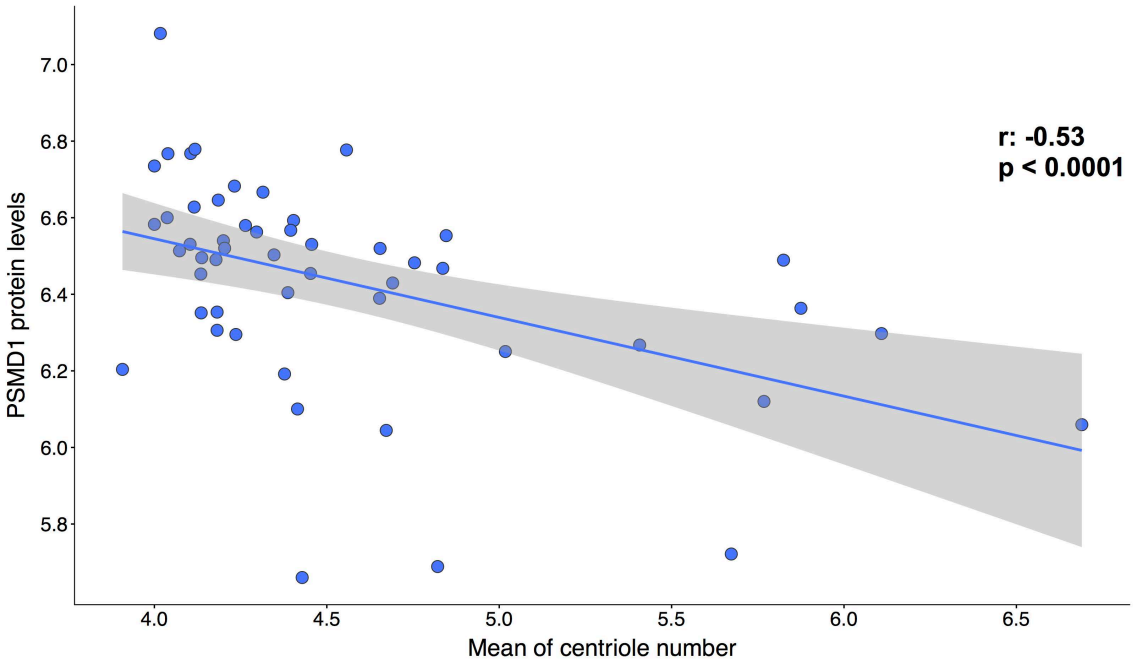


Figure 4.21 PSMD1 protein levels were negatively correlated with centriole number (Spearman correlation coefficient: -0.53, $p < 0.0001$). Grey shades around linear regression lines as in *Figure 4.7a*.

We then performed genome-wide correlation analyses between gene expression and centriole number metrics, followed by GSEA (using the Spearman correlation coefficient to rank genes), to identify pathways putatively associated with centriole number abnormalities in cancer. We did not identify any pathway significantly associated with the percentage of cells with CA at a FDR lower than 5%.

Interestingly, beyond the previous identification of the PSMD1 proteasome component, GSEA also identified the proteasome pathway as enriched in genes negatively correlated with the mean of centriole number (NES: -2.25, FDR < 0.001; ***Figure 4.22a***). Proteasomes are large protein complexes involved in many essential cellular functions that degrade ubiquitinated proteins by an ATP-dependent mechanism. The most common form used in mammals is the cytosolic 26S proteasome that contains one 20S protein subunit and two 19S regulatory cap

subunits (Coux et al., 1996). Most of the 26S component genes contributed to the proteasome negative enrichment result (***Figure 4.22b***).

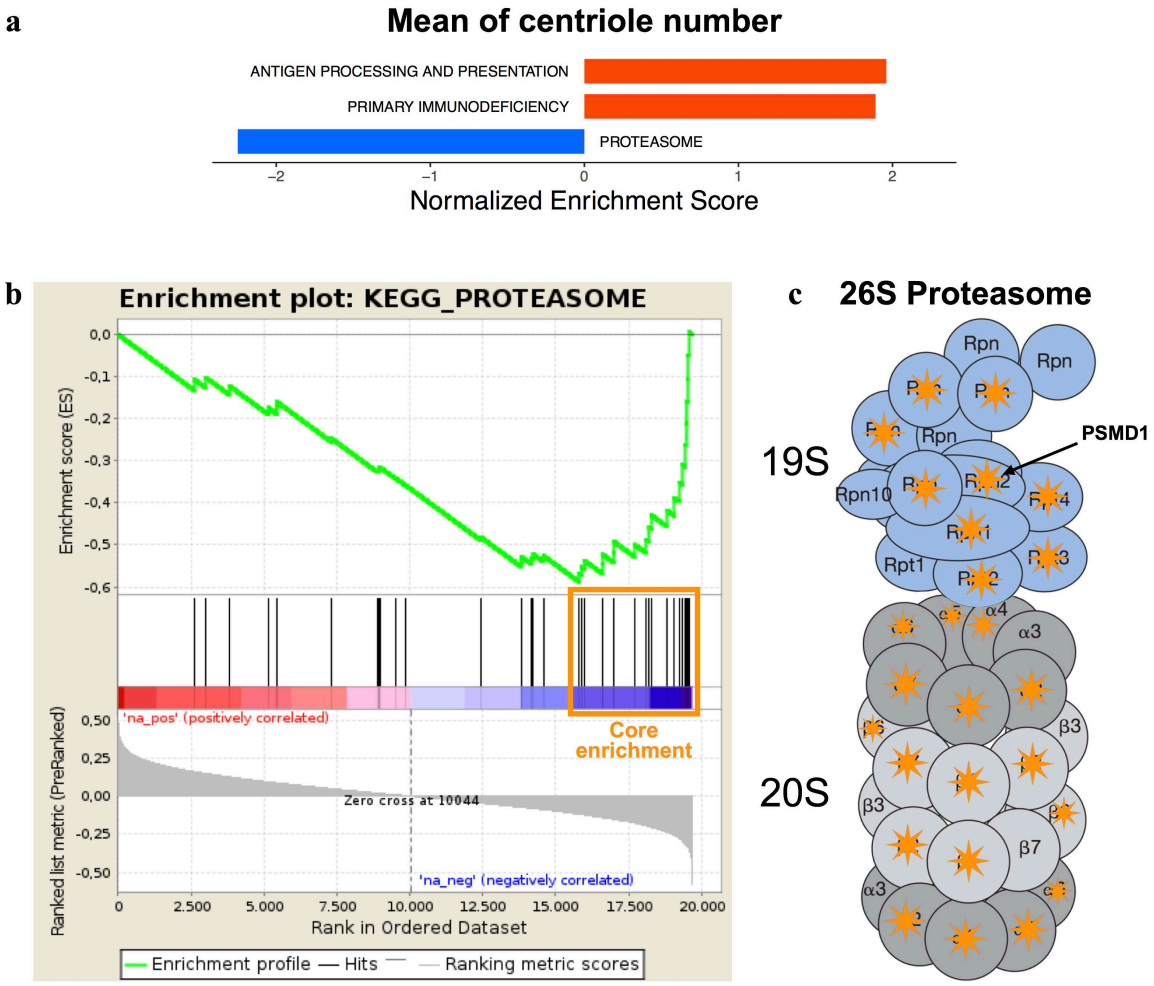


Figure 4.22 Lower expression of proteasome components was associated with increased centriole number levels. a) GSEA results on the mean of centriole number-associated transcriptomic alterations in the NCI-60 panel. KEGG pathways positively (red) and negatively (blue) enriched respectively in genes over- and underexpressed in cell lines with higher centriole numbers, with a FDR lower than 5%, are shown. **b)** The proteasome pathway was enriched in the negatively associated genes (NES: -2.25, FDR < 0.001). Genes that contributed the most for that enrichment are highlighted (orange box). GSEA enrichment plot explained in Materials and Methods, section 3.2.2. **c)** Representation of the 26S proteasome structure, as well as its 19S and 20S subunits (adapted from Dahlmann, 2005). 26S component genes that contributed to the proteasome negative enrichment result are highlighted (orange star), including the PSMD1 component (arrow).

GSEA also identified blood-related pathways (primary immunodeficiency and antigen processing and presentation) as being associated with increased centriole number (***Figure 4.22a***), most likely due to the observed higher CA levels in blood cell lines (***Figure 4.1a***).

4.3 Identification of new compounds that target CA

To explore the CA therapeutic potential in cancer, we took advantage of the compound activity data available in the NCI-60 panel. We combined these data with the CA profiles uncovered for the panel to both understand the general role of CA in cell lines' compound sensitivity and identify new compounds that putatively kill cancer cells through CA. The last aim was approached in two different ways: correlation analyses between compound activity and both CA levels and expression levels of CA-associated genes.

4.3.1 CA is associated with higher sensitivity to compound activity

CA is a common feature in cancer but it is still unknown if it generally confers sensitivity or resistance to anti-cancer therapies. To address this question, we performed global Spearman correlation between the activity of each of 14,005 compounds and the prevalence of CA (given by the percentage of cells with CA) in NCI-60 cell lines. If there was no particular association, we expected most correlations to be around zero and similar proportions between positive and negative correlations (***Figure 4.23 - Expected***). However, we observed a statistically significant higher proportion of positive correlations (60%, Chi-squared goodness of fit test $p < 0.0001$; ***Figure 4.23 - Observed***), meaning that cell lines with higher CA levels likely tend to have higher sensitivity to compound activity. Wherever this is a causal relationship, i.e. CA confers higher sensitivity to drug activity, or CA just emerges as a surrogate for other factors needs to be experimentally tested.

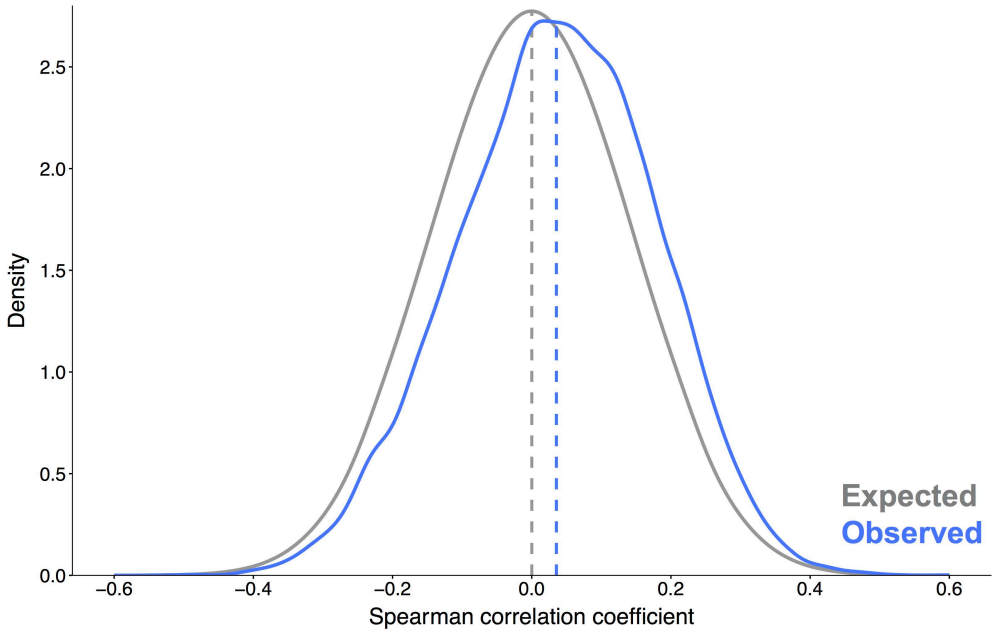


Figure 4.23 Enrichment of positive correlations between compound activity and centrosome amplification prevalence. Expected (grey, resulting from 1000 drug activity permutations) and observed (blue) density distributions of Spearman correlation coefficients between activity of each of 14,005 compounds and the percentage of cells with CA across the NCI-60 panel of cell lines. According with the expected distribution, positive and negative correlations should present similar proportions: 50% (Chi-squared goodness of fit test p: n.s.). However, we observed a significant higher proportion of positive correlations: 60% (Chi-squared goodness of fit test $p < 0.0001$). Vertical dashed lines represent the mean of each Spearman correlation coefficient distribution.

4.3.2 Compounds that selectively kill cancer cells with CA

Since we observed that CA was associated with higher sensitivity to compound activity in the panel, we exploited the above-mentioned global Spearman correlation analyses to prioritize candidate compounds that selectively kill cancer cells with higher incidence of this abnormality. Unfortunately, we did not find any compound positively correlated with the percentage of cells with CA at a FDR lower than 5%, i.e. with statistically significant higher activity in cell lines that have higher incidence of this abnormality. Nevertheless, the ten compounds with the most significant positive correlations (***Table 4.1***) were selected for further *in vitro* experimental tests, in order to test their CA-selectivity. Compound NSC 633109 showed the strongest association between its activity and the percentage of cells with CA (Spearman correlation coefficient: 0.6, $p < 0.001$; ***Figure 4.24***).

Table 4.1 Compounds that have higher activity in cell lines with higher incidence of centrosome amplification. Table with the ten compounds with the most significant positive correlations between their activity and percentage of cells with CA across NCI-60 cell lines, ordered by nominal p-value. NSC ID is the US National Cancer Institute internal ID number and PubChem SID is the PubChem substance identifier. The compound with the strongest positive correlation is highlighted (blue box) and depicted in *Figure 4.24*.

Drug	NSC ID	Spearman correlation coefficient	P.value	FDR-corrected p.value	PubChem SID
NA	633109	0.60	1.49E-04	0.78	496250
NA	734057	0.45	4.79E-04	0.84	48430269
s-(n-hydroxy n-methylcarbamoyl)-glutathione	647036	0.48	5.56E-04	0.84	503341
NA	732188	0.47	6.27E-04	0.84	48429416
NA	654104	0.48	6.93E-04	0.84	506238
NA	656084	0.52	7.40E-04	0.84	507009
polyozellin	670123	0.44	8.79E-04	0.84	513587
NA	647589	0.42	1.12E-03	0.84	503644
NA	641822	0.42	1.27E-03	0.84	500894
arylpurines	737433	0.42	1.40E-03	0.84	48431476

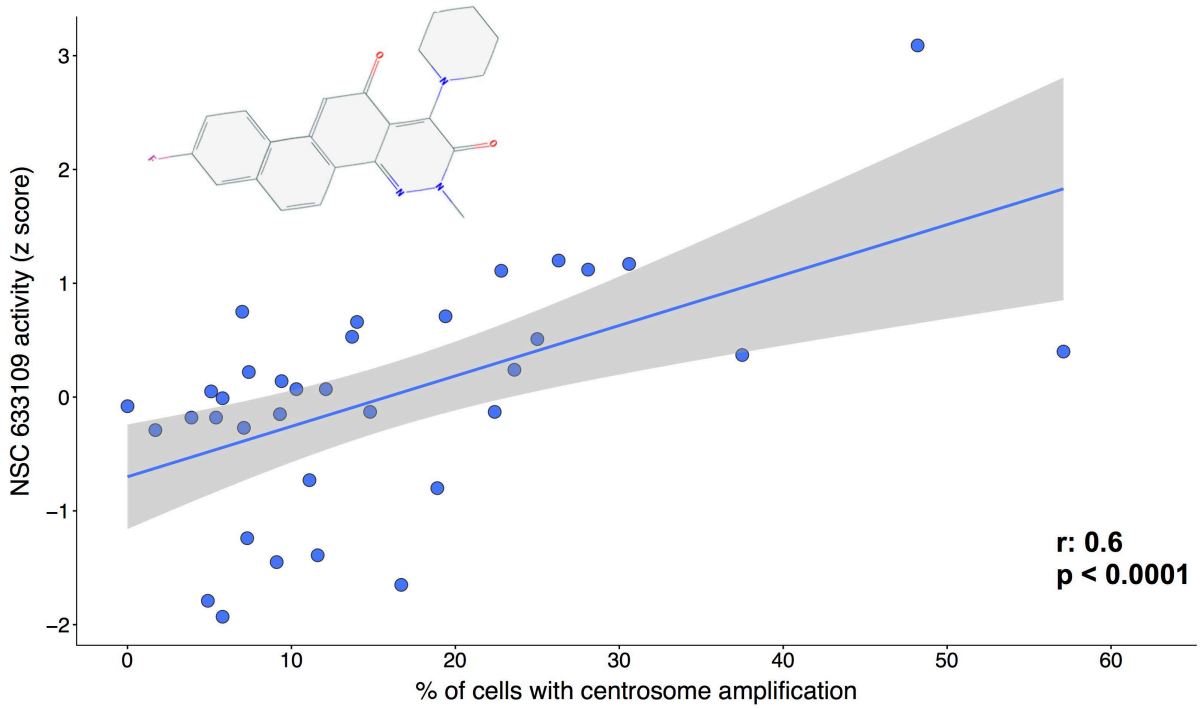


Figure 4.24 Correlation between activity of top compound NSC 633109 (z score) and percentage of cells with centrosome amplification (Spearman correlation coefficient: 0.6, $p < 0.001$). Grey shades around linear regression lines as in *Figure 4.7a*. The compound's chemical structure is shown above.

4.3.3 Compounds that target CA-associated proteins

An additional way of exploiting CA in cancer therapy would be through the selective targeting of proteins that promote supernumerary centrosomes (as discussed in Holland and Cleveland, 2014, regarding the targeting of PLK4). To explore this approach, we performed correlation analyses between cell line sensitivity for each of the 14,405 compounds and expression levels of the gene associated with the main CA-associated protein – PLK4 – across NCI-60 cell lines. We used gene expression because protein levels were not available for this kinase. The rationale was that compounds inhibiting a specific protein are expected to have higher activity in cell lines highly expressing that protein. We did not identify any compound with activity positively correlated with *PLK4* expression, taking a FDR-corrected p-value lower than 0.05. Still, these analyses were used to prioritize the ten compounds with the most significant positive correlations (***Table 4.2***), whose CA-selectivity will be tested experimentally using *in vitro* assays. Compound NSC 658364 showed the strongest association between its activity and *PLK4* expression levels (Spearman correlation coefficient: 0.62, $p < 0.0001$; ***Figure 4.25***). This analysis represents the proof of concept for this approach, which could be applied with hereafter identified CA-associated proteins and different datasets containing both compound activity and gene and protein expression data.

Table 4.2 Compounds that target the centrosome amplification-associated protein PLK4. Table with the ten compounds with the most significant positive correlations between their activity and *PLK4* expression across NCI-60 cell lines, ordered by nominal p-value. NSC ID and PubChem SID explained in *Table 4.1*. The compound with the strongest positive correlation is highlighted (blue box) and illustrated in *Figure 4.25*.

Drug	NSC ID	Spearman correlation coefficient	P.value	FDR-corrected p.value	PubChem SID
NA	658364	0.62	1.26E-05	0.16	508234
6-selenoguanosine	137679	0.51	4.59E-05	0.22	425064
NA	658368	0.51	5.24E-05	0.22	508238
NA	709319	0.48	1.34E-04	0.31	582484
azaserin	758188	0.48	1.60E-04	0.31	*NA*
duramycin	71935	0.47	1.62E-04	0.31	*NA*
(z)-5-chloro-3-((1-(3,5-dimethoxyphenyl)-9h...	761834	0.49	1.68E-04	0.31	*NA*
NA	635525	0.47	2.18E-04	0.35	497704
NA	704319	0.46	2.84E-04	0.35	528327
NA	683833	0.46	3.16E-04	0.35	520009

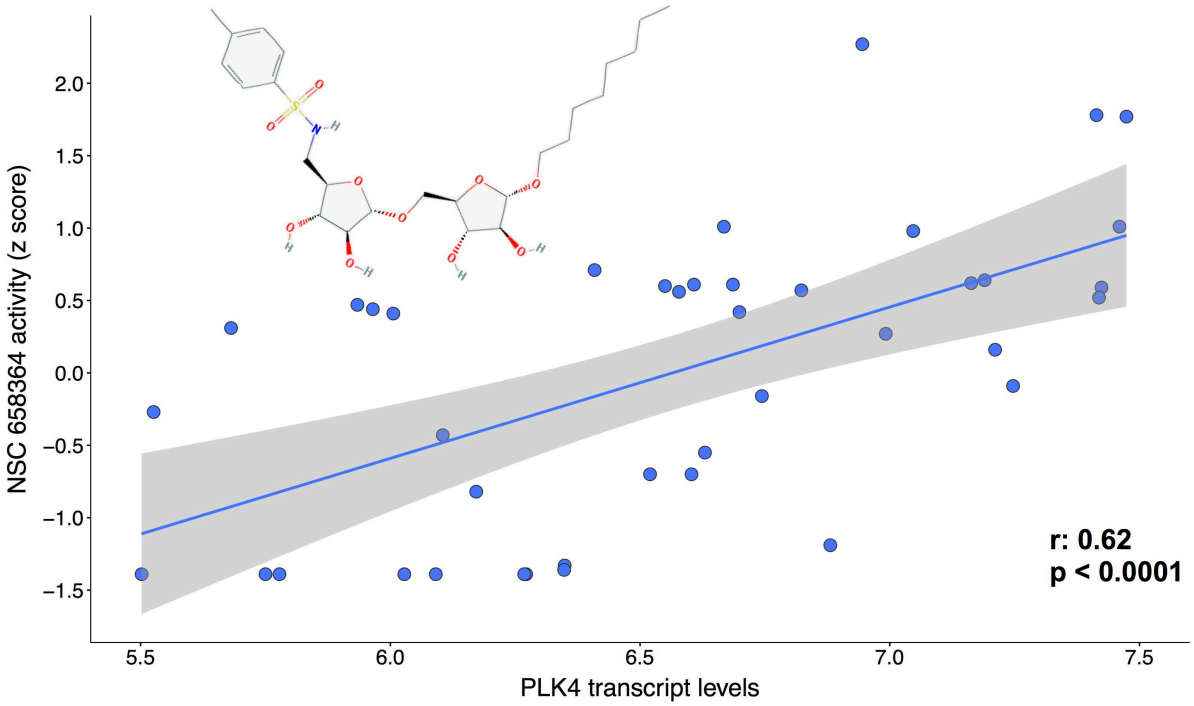


Figure 4.25 Correlation between activity of top compound NSC 658364 (z score) and *PLK4* transcript levels (Spearman correlation coefficient: 0.62, $p < 0.0001$). Grey shades around linear regression lines as in *Figure 4.7a*. The compound's chemical structure is shown above.

CHAPTER 5 – DISCUSSION

Previous studies have demonstrated the importance of centriole number dysregulation in cancer, both as a promotor of tumourigenesis (Basto et al., 2008; Coelho et al., 2015; Levine et al., 2017; Serçin et al., 2015) and an appealing target for anti-cancer therapies (Mason et al., 2014; Watts et al., 2013). Moreover, the Bettencourt-Dias Lab have identified COE as a recurrent feature of cancer cells that also promotes CA (Marteil et al., *manuscript in preparation*). However, the prevalence and the molecular mechanisms underlying these centriole abnormalities in cancer, as well as their therapeutic value, remain poorly understood, preventing their use in the clinic.

We present an innovative and multidisciplinary approach for studying centriole abnormalities in cancer that has provided further insights into their associated molecular mechanisms and the development of clinical applications based on selectively targeting these Achilles' heels of cancer cells.

5.1 NCI-60 profile of centriole abnormalities

Our analyses at the cell population level confirmed CA and COE as widespread phenomena among different cancer cell lines and cancer types, where dysregulation in number was more common than that in length. However, we recognize two fundamental problems with this centriole screen that prevent reaching stronger conclusions about the prevalence of centriole abnormalities in cancer. First, the reduced number of profiled cells for each cell line does not allow us to accurately estimate their abnormality levels. For instance, with a minimum sample of 50 cells and CA observed in eight of them (close to the panel mean of 17%), the 95% confidence interval for the CA frequency estimate ranges from 7.6% to 29.7%, which means a great deal of uncertainty in our estimates of CA prevalence in many cell lines. Second, the lack of negative controls does not allow us to define the variability of centriole number and length in non-cancerous cell lines, thus limiting the comparison between cancerous and non-cancerous centriole abnormality profiles. To overcome these issues, the Bettencourt-Dias Lab are now increasing the sample size per cell line, as well as performing additional tests in non-cancerous cell lines.

We took advantage of the single-centriole resolution of the NCI-60 screen to perform the first profiling of abnormalities in individual centrioles in cancer. We observed centriole number variability across cell lines from different tissues of origin, with colon and lung being those with most heterogeneous ones. Surprisingly, all tissues presented more cells with eight than

with seven centrioles. Three possible explanations for this observation: i) cancer cells live more stably with eight centrioles, twice the normal number; ii) most cancer cells with eight centrioles are a result of cell cycle defects (e.g. cytokinesis failure, mitotic slippage and cell–cell fusion) and have accumulated the centrioles from the two expected daughter cells, resulting in twice the number expected for an individual cell; iii) some cancer cells with eight centrioles are indeed two cells but were detected as being one due to technical issues. Although all three causes seem plausible, the technical one is unlikely, given that all microscopy pictures analysed in the screen were manually curated, and cell cycle defects are most likely, since they were already observed to be a cause of CA (Godinho and Pellman, 2014).

Similarly, we observed variability at centriole length level across cell lines from different cancer tissues of origin, where lung and skin constituted those from which cell lines with higher heterogeneity were derived. Interestingly, even within these two tissues of origin there were big discrepancies in length variability across cell lines. Particularly, cell line MDA-MB-435 showed significantly higher length heterogeneity, when compared with other skin cancer cell lines, that could be explained by the curious origin of this cell line: it was originally described as a human breast cancer cell line, since it was derived from the pleural effusion of a patient with breast cancer, but its subsequent molecular characterization (transcript and protein expression and drug sensitivity) was consistent with a melanoma origin (Ross et al., 2000; Scherf et al., 2000). Furthermore, single nucleotide polymorphism analysis revealed that MDA-MB-435 is actually derived from the same individual as the melanoma cell line M14 (Garraway et al., 2005; Rae et al., 2007). The panel designation for this cell line is still a topic of discussion (Chambers, 2009). Nevertheless, given the controversial origin of MDA-MB-435, it is interesting that it stood out in the NCI-60 panel as that with the highest penetrance of COE in the panel.

CA was found to be associated with particularly aggressive breast (basal) and colon (CIN) cancer molecular subtypes in the NCI-60 panel, both characterized by poor prognosis (Kocarnik et al., 2015; Parker et al., 2009; Phipps et al., 2015; Sørlie et al., 2001). Breast cancer findings were further validated in human tissue samples, in Dr Joana Paredes' lab (Marteil et al., *manuscript in preparation*). The higher prevalence of CA in these molecular subtypes is likely explained by their specific molecular features, such as increased incidence of *BRCA1* mutations in basal-like breast carcinomas (Foulkes et al., 2003), shown to induce centrosome overduplication (Ko et al., 2006; Starita et al., 2004), and observed amplification of the Aurora-A gene in CIN colon cancer (Grady, 2004), whose overexpression leads to supernumerary

centrosomes (Ghadimi et al., 2000; Lentini et al., 2007; Meraldi et al., 2002). Altogether, these results support the tumourigenic potential of CA in cancer (Chan, 2011).

We validated the association between COE and CA in the panel both at the cell population level, where COE was positively correlated with CA, and at the single-centriole level, where we observed higher proportion of overly-long centrioles in cells with CA. However, we were not able to establish a causal association and corroborate the previous finding of COE as promoter of CA *in vitro* (Marteil et al., *manuscript in preparation*). Actually, the existence of cell lines with high levels of CA but reduced COE, contrasting with the absence of cell lines with high COE but low CA, suggests that CA is necessary for cell lines to develop COE. Nevertheless, both hypotheses are compatible: CA can prompt COE, which in turn can promote CA by both centriole fragmentation and ectopic procentriole formation (***Figure 5.1***). This positive feedback hypothesis should be tested in the future, in order to better understand the association between both types of centriole abnormality.

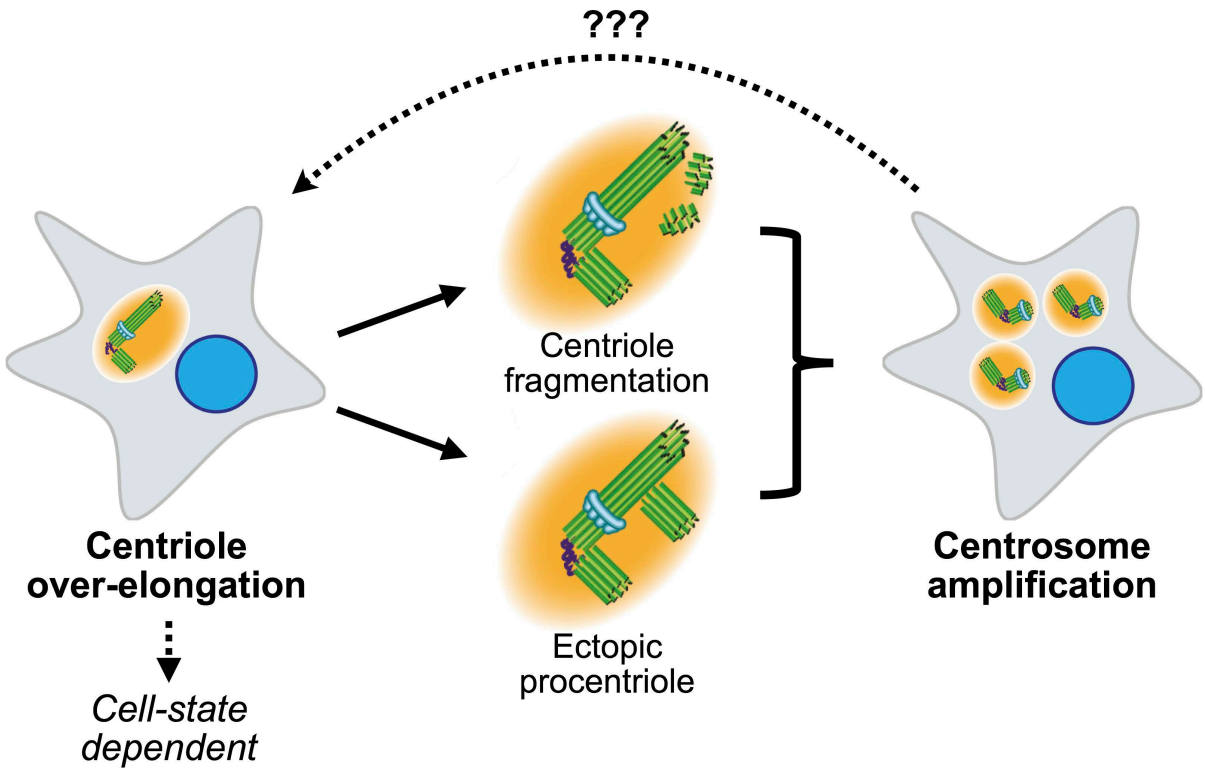


Figure 5.1 Hypothesized portrait of the COE-CA association in cancer. COE was shown to promote CA both by centriole fragmentation and ectopic procentriole formation. However, our results in the NCI-60 panel were also compatible with a positive feedback hypothesis where CA also promotes COE. Nevertheless, this hypothesis requires further experimental validation, given that there is no previous evidence suggesting CA as a driver of COE. Moreover, our analyses suggest COE as a cell-state dependent feature, i.e. specific physiological conditions may be enhancing COE. Hypothesized associations are shown in dashed lines (adapted from Marteil et al., *manuscript in preparation*).

Although COE had low frequency in the NCI-60 panel, our results showed that it was not a purely stochastic event in cancer cells and it occurred more frequently within the same cell than expected by chance, both in cells with or without CA, thus suggesting COE as a widely cell state-dependent feature (***Figure 5.1***), i.e. specific physiological conditions may be enhancing COE. Better understanding of the underlying mechanisms would provide novel insights in centriolar biology.

Beyond centriole number and length dysregulation, we found that cancer cells apparently do not control their centriolar mass, since the average total centriolar mass increased monotonically with the average number of centrioles per cell. In addition, we observed that the average centriole length was maintained in cells with supernumerary centrosomes. Thus, there was no evidence that any CA arose from centriole fragmentation in the NCI-60 panel. However, there are two limitations that prevent us from dismissing this hypothesis: i) microscopy resolution was not enough for the detection of centriole shortening; ii) short centriole fragments then generally grow until they reach normal length. Nevertheless, we must consider the possibility that CA did not arise from centriole fragmentation and other molecular mechanisms may be causing number dysregulation in this panel.

5.2 Novel molecular mechanisms underlying centriole abnormalities in cancer

The second aim of this project was to explore the molecular origins of both CA and COE in cancer by integrating the quantitative centriole profiles along NCI-60 cell lines with the publicly available molecular data for the panel.

Regarding centriole length dysregulation, we identified for the first time an association between centriole length and cell cycle progression. We found that cancer cell lines with overly-long centrioles proliferated slower and showed an accumulation of cells in G1 phase, suggesting that centriole length defects could lead to a cell cycle delay in G1. If this was the case, one would expect a stronger association for cells that still have an intact p53 response than for those with mutated *TP53*. However, we did not find a significant difference in the relation between centriole length defects and cell cycle delay in G1 between *TP53* statuses. Although we did not validate this p53-dependent hypothesis in the NCI-60 panel, it would be worth to experimentally test the association between centriole length, cell cycle progression and the accumulation of cells in G1 phase, given that other p53-independent G1 checkpoint pathways could be involved in this process.

Our transcriptomic analysis identified *PRKACA* as a putative promoter of COE in cancer (***Figure 5.2***). Among the 870 centrosomal genes tested, only *PRKACA* expression levels were

significantly correlated with centriole length mean. Neither *CPAP* nor *CP110*, two known centriole length regulators (Schmidt et al., 2009), had their expression significantly correlated with centriole length in the panel, suggesting that centriole length dysregulation could be associated with different molecular mechanisms.

PRKACA encodes one of the catalytic subunits of PKA, a family of cAMP-dependent serine/threonine kinases that in their inactive form consist of two catalytic (C) and two regulatory (R) subunits. Upon binding of cAMP to R subunits, C subunits are released and phosphorylate many substrates in the cytoplasm and the nucleus (Skålhegg and Tasken, 2000). PKA is mainly docked at the centrosome (Nigg et al., 1985), a process dependent on the A-kinase anchoring proteins pericentrin and AKAP9 (Grady, 2004), and it is involved in cell growth, differentiation, and proliferation, as well as in cell cycle control (Duncan et al., 2006; Eyers et al., 2005; Kovo et al., 2006). Particularly, higher PKA activation was associated with increased primary cilium length in mammalian cells (Besschetnova et al., 2010). However, to our knowledge there are no reports on a possible relationship between that kinase and centriole length regulation.

PRKACA was previously tested in an independent screen for putative centriole length regulators, where its depletion led to shorter centrioles in one of two siRNA experiments. The other experiment did not yield any difference in centriole length. The screen involved a large-scale analysis in which the efficiency of depletion was not checked and thus the false negative could not be controlled for, which could explain the discrepancy between the outcomes of the two siRNA experiments. Therefore, the putative role of *PRKACA* in the regulation of centriole length requires further experimental validation. To validate *PRKACA* as a centriole length regulator, we are planning to first validate *PRKACA* upregulation in cell lines with overly-long centrioles, and then determine the effects of *PRKACA* up- and downregulation on centriole length in cell lines. These analyses will be performed in close collaboration with the Bettencourt-Dias Lab.

We then performed GSEA on the genome-wide COE-associated transcriptomic changes to explore the molecular mechanisms associated with centriole length abnormalities in cancer. Strikingly, we identified pathways involved in the interaction between the cell and the ECM (ECM receptor interaction and focal adhesion) positively associated with COE. Marteil and co-workers have shown that overly-long centrioles generate over-active centrosomes that nucleate more microtubules (Marteil et al., *manuscript in preparation*), a known cause of invasiveness, but this relationship between overly-long centrioles and cell invasion is not established yet. Here we have shown that cell lines with overly-long centrioles presented higher

expression of genes involved in the interaction with the ECM, thus providing some new insights into a possible effect of overly-long centrioles in cancer cell migration (***Figure 5.2***). The identification of genes involved in DNA repair mechanisms (DNA replication and mismatch repair pathways) as negatively associated with centriole length also suggests that a drop in efficiency of those mechanisms may allow the presence of overly-long centrioles in the NCI-60 cell lines (***Figure 5.2***). The pathways identified in GSEA need to be properly tested in order to establish a causal link with COE, as well as to identify its underlying molecular mechanisms.

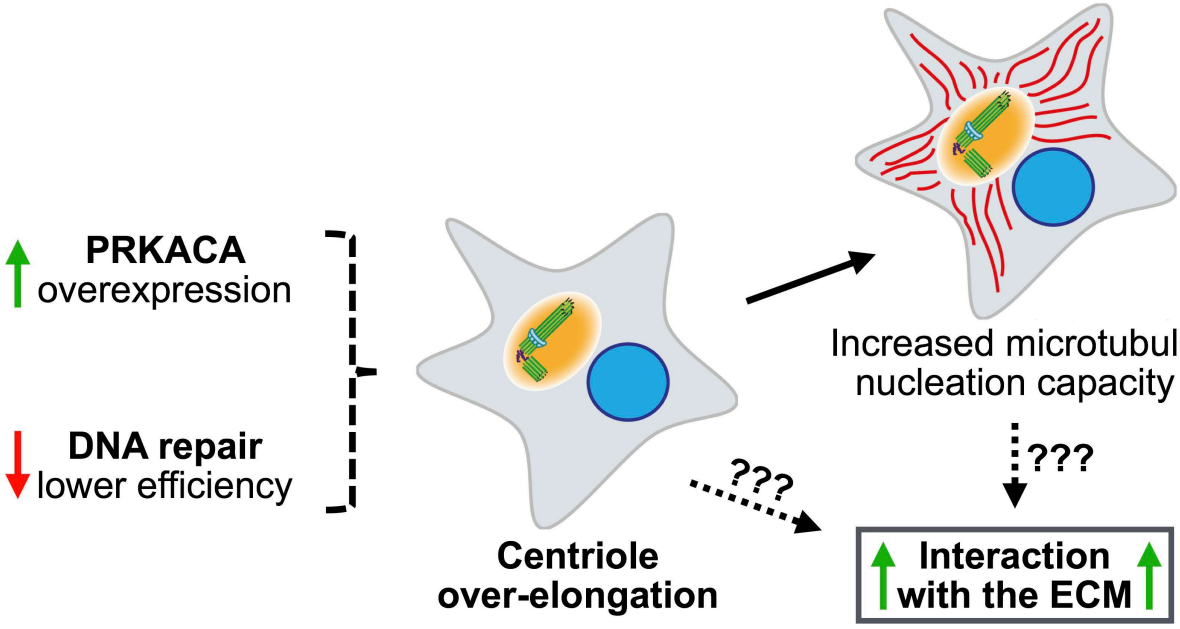


Figure 5.2 Hypothesis of centriole over-elongation origins and consequences in the NCI-60 panel. We identified *PRKACA* and lower expression of DNA repair genes as putative promoters of COE in cancer. On the other hand, Marteil and co-workers have shown that overly-long centrioles have increased microtubule nucleation capacity. Our GSEA results showed an association between COE and higher interaction with the ECM, that we hypothesize to be an indirect consequence through higher microtubule nucleation capacity. However, a microtubule-independent hypothesis should also be considered. Hypothesized associations are shown in dashed lines (adapted from Marteil et al., *manuscript in preparation*).

In similar genome-wide analyses, we did not identify any strong individual association of centriole number with known centriole biogenesis regulators: *PLK4* (gene expression level), *STIL* (gene and protein) and *SASS6* (gene and protein). This could suggest that centriole number dysregulation in cancer is associated with different molecular mechanisms that need to be explored. Indeed, we found a strong negative association between the proteasome and CA, an unprecedented observation in cancer research. Several genes that encode for proteasome components were negatively associated with centriole number, namely *PSMD1*.

The 26S proteasome, the form most commonly used in mammals, is a large multisubunit enzyme complex present both in the nucleus and cytoplasm of all eukaryotic cells that primarily degrades proteins in a ATP/ubiquitin-dependent process (Adams, 2003). It is composed by one 20S protein subunit and two 19S regulatory cap subunits (Coux et al., 1996). *PSMD1* encodes for the largest non-ATPase subunit of the 19S regulator proteasome lid, responsible for substrate recognition and binding, and its individual protein levels were negatively correlated with centriole number.

Proteasomes regulate the turnover of numerous cellular proteins involved in essential cellular processes, such as cell-cycle progression and apoptosis (Adams, 2003; Livneh et al., 2016). Remarkably, proteasomes have been shown to localize to centrosomes (Wigley et al., 1999), where they play a critical role in regulating centrosomal proteins, hence maintaining proper centrosome function (Didier et al., 2008). For instance, proteasome inhibition resulted in a significant increase in centrosome size (Adams, 2003), putatively due to an accumulation of different centrosome proteins (Didier et al., 2008). Moreover, all the members of the PLK4–STIL–SAS-6 module were shown to be regulated through proteasome-mediated degradation (Arquint and Nigg, 2014; Cunha-Ferreira et al., 2009; Strnad et al., 2007) and proteasome inhibition was shown to induce CA in *Drosophila* (Cunha-Ferreira et al., 2009), highlighting the relevance of proteasome in preventing centriole number abnormalities.

Although a negative association between the proteasome and CA has been previously established in *Drosophila*, there is no evidence that it can happen in cancer *per se*. Therefore, the proteasome-CA negative association observed in the NCI-60 panel provides a remarkable insight into the putative molecular mechanisms driving CA in cancer. Together with previous knowledge from literature, this result led us to hypothesize that NCI-60 cell lines that present lower expression of proteasome machinery genes, i.e. putative lower proteasome activity, can also have higher levels of PLK4–STIL–SAS-6 module proteins (and possibly other centrosomal proteins), due to lower protein degradation, driving the observed higher CA levels (***Figure 5.3***). Thus, proteasome is proposed to protect cells from CA, meaning that cells with less proteasome activity would be more susceptible to exhibit supernumerary centrosomes susceptible of conferring advantage and consequent higher tumourigenic potential to those cells.

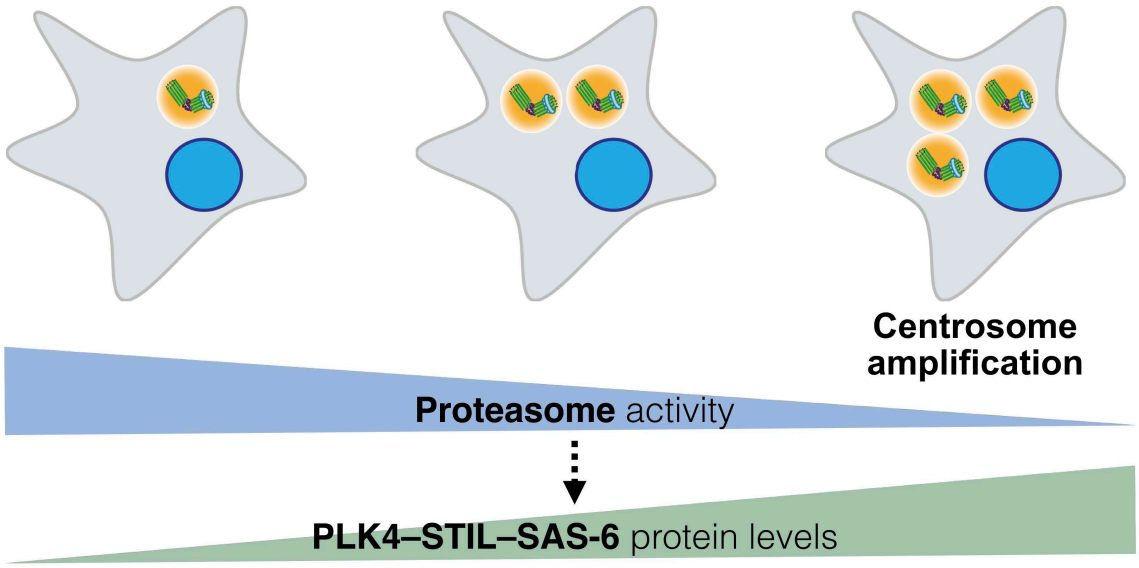


Figure 5.3 Hypothesis of centrosome amplification origin in the NCI-60 panel. Our results identify the proteasome as a putative protector of cells from CA. Since the proteasome is known to regulate several centrosomal proteins at the centrosome, including the PLK4–STIL–SAS-6 module components crucial in centriole biogenesis, we hypothesize that NCI-60 cancer cells present CA levels due to reduced proteasome activity and consequent PLK4, STIL and SAS-6 accumulation. Hypothesized associations are shown in dashed lines (adapted from Marteil et al., *manuscript in preparation*).

Although our correlation analyses in the NCI-60 panel did not identify any correlation between centriole number and individual STIL or SAS-6 protein levels, we were not able to test the association with all the PLK4–STIL–SAS-6 module components, since there were no PLK4 protein levels available for this panel, PLK4 being the master regulator of centriole biogenesis. Therefore, this hypothesis needs to be properly experimentally tested. Indeed, we are already planning those experiments, in collaboration with the Bettencourt-Dias Lab, testing if proteasome inhibition induces CA in cancer cell lines, where we will also assess the respective PLK4, STIL and SAS-6 protein levels.

5.3 Therapeutic value of CA in cancer

One main advantage of the NCI-60 panel is its large database of anticancer drug activity. We have used these data on cell line sensitivity in order to both investigate the CA therapeutic potential in cancer and identify novel compounds that target this abnormality.

Our results disclosed CA as generally associated with higher sensitivity to compound activity. However, since CA could have emerged as a surrogate for other factors, e.g. genomic instability, this direct association needs to be experimentally tested. The validation of this hypothesis will highlight the CA therapeutic potential and intensify the search for novel targeted compounds.

Correlation analyses between compound activity and both CA and *PLK4* gene expression levels did not uncover significantly associated compounds. Still, these analyses were used to prioritize the ten compounds (number of compounds selected based on budget limitations) with the most significant positive correlations in both analyses. Their activity will be experimentally validated using *in vitro* models, particularly in NCI-60 cell lines with different CA levels, together with non-cancerous cell lines, to test compound CA-selectivity. Compounds that show specific targeting of abnormal cell lines, without affecting normal ones, will be selected for subsequent studies to validate their centrosome-directed mechanism of action. Furthermore, we will apply chemical informatics approaches to pinpoint the common chemical properties among the prioritized compounds because we believe this approach will provide new insights into the chemical properties that confer their ability to specifically target cells with CA and thereby become a powerful tool in drug design and modulation. Moreover, by identifying compounds that target CA, we will be able to then explore the molecular mechanisms associated with their activity, thus providing novel insights into the CA-associated mechanisms in cancer.

Both chemical informatics approaches and experimental validation of promising compounds will be performed in close collaboration with Dr Gonçalo Bernardes' group (Instituto de Medicina Molecular), taking advantage of their strong expertise in chemistry and targeted cancer therapeutics.

Given the cancer-specificity of CA, the compounds identified will be the basis for the development of drugs to selectively target cancer cells, hence promising a shift in the way we treat this group of diseases.

CHAPTER 6 – CONCLUSION

Altogether, this project has covered different aspects of centriole abnormalities in cancer, taking advantage of a unique resource created by the rigorous quantification of centrosome structure in a cell line panel also well characterized at the molecular level. Moreover, the created resource allows the investigation of additional relevant questions on centriole abnormalities in cancer. Indeed, we have already quantified the CC levels in the NCI-60 panel, only possible because the centriole screen was performed in mitotic cells, that will now be integrated with the publicly available molecular and drug-sensitivity quantitative profiles for that panel, to uncover the prevalence, origins and therapeutic value of CC in cancer.

The first aim of profiling the centriole abnormalities along the panel has provided a first glimpse of the landscape of such abnormalities in cancer, described in a manuscript (Marteil et al., *manuscript in preparation*), to be submitted to Nature Communications soon, that also includes the NCI-60 centriole screen and the identification of COE as a widespread phenomenon in cancer. The second and third aims have raised interesting and putatively relevant hypothesis in the centrosome-cancer field that will now be experimentally validated in order to better characterize the associated candidate molecular mechanisms and compounds.

This project was presented by this thesis' author in two international meetings: poster presentation at the 3rd EACR Conference in Cancer Genomics, Cambridge, UK; oral presentation at the 12th Young European Scientists meeting, Porto, Portugal, whose abstract was published in Porto Biomedical Journal (de Almeida et al., 2017; *10.1016/j.pbj.2017.07.019*).

To conclude, this work presents the first single-centriole-level portrait of centriole abnormalities in cancer and contributes to a better understanding of their origins, namely by revealing novel molecular mechanisms in cell cycle biology. We ultimately expect the results of our pioneering multidisciplinary approaches to provide novel targeted cancer therapeutic options and inspire a new way of studying and handling cancer.

BIBLIOGRAPHY

Adams, J. (2003). The proteasome: structure, function, and role in the cell. Cancer Treat. Rev. *29*, 3–9.

Ahmad, A.S., Ormiston-Smith, N., and Sasieni, P.D. (2015). Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960. Br. J. Cancer *112*, 943–947.

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J., and Clarke, M.F. (2003). Prospective identification of tumorigenic breast cancer cells. Proc. Natl. Acad. Sci. *100*, 3983–3988.

de Almeida, B.P., Marteil, G., Bettencourt-Dias, M., and Barbosa-Morais, N.L. (2017). Discovery of novel mechanisms of centrosome amplification and their therapeutic value in cancer. Porto Biomed. J. *2*, 182.

Alves-Cruzeiro, J.M.D.C., Nogales-Cadenas, R., and Pascual-Montano, A.D. (2014). CentrosomeDB: A new generation of the centrosomal proteins database for Human and Drosophila melanogaster. Nucleic Acids Res. *42*, 430–436.

American Cancer Society (2016). Cancer Facts & Figures 2016.

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

Anderson, C.T., and Stearns, T. (2009). Centriole age underlies asynchronous primary cilium growth in mammalian cells. Curr. Biol. *19*, 1498–1502.

Arquint, C., and Nigg, E.A. (2014). STIL microcephaly mutations interfere with APC/C-mediated degradation and cause centriole amplification. Curr. Biol. *24*, 351–360.

Arquint, C., and Nigg, E.A. (2016). The PLK4–STIL–SAS-6 module at the core of centriole duplication. Biochem. Soc. Trans. *44*, 1253–1263.

Badano, J.L., Teslovich, T.M., and Katsanis, N. (2005). The centrosome in human genetic disease. Nat. Rev. Genet. *6*, 194–205.

Balestra, F., Strnad, P., Flückiger, I., and Gönczy, P. (2013). Discovering regulators of centriole biogenesis through siRNA-based functional genomics in human cells. Dev. Cell *25*, 555–571.

Basto, R., Brunk, K., Vinadogrova, T., Peel, N., Franz, A., Khodjakov, A., and Raff, J.W. (2008). Centrosome amplification can initiate tumorigenesis in flies. Cell *133*, 1032–1042.

Bauer, M., Cubizolles, F., Schmidt, A., and Nigg, E.A. (2016). Quantitative analysis of human centrosome architecture by targeted proteomics and fluorescence imaging. EMBO J. *35*, 2152–2166.

Van Beneden, E., and Neyt, A. (1887). Nouvelle recherches sur la fécondation et la division mitosique chez l'Ascaride mégalocéphale. Bull. Acad. R. Belgique 3éme Sér. *14*, 215–295.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Besschetnova, T.Y., Kolpakova-hart, E., Guan, Y., Zhou, J., Olsen, B.R., and Shah, J. V. (2010). Identification of signaling pathways regulating primary cilium length and flow-

mediated adaptation. Curr. Biol. *20*, 182–187.

Bettencourt-Dias, M. (2013). Q&A: Who needs a centrosome? BMC Biol. *2*, 11–28.

Bettencourt-Dias, M., and Glover, D.M. (2007). Centrosome biogenesis and function: centrosomics brings new understanding. Nat. Rev. Mol. Cell Biol. *8*, 451–463.

Bettencourt-Dias, M., Rodrigues-Martins, A., Carpenter, L., Riparbelli, M., Lehmann, L., Gatt, M.K., Carmo, N., Balloux, F., Callaini, G., and Glover, D.M. (2005). SAK/PLK4 is required for centriole duplication and flagella development. Curr. Biol. *15*, 2199–2207.

Blows, F.M., Driver, K.E., Schmidt, M.K., Broeks, A., Leeuwen, F.E. Van, Wesseling, J., Cheang, M.C., Gelmon, K., Nielsen, T.O., Blomqvist, C., et al. (2010). Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med. *7*, e1000279.

Borm, G.F., Fransen, J., and Lemmens, W.A.J.G. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. J. Clin. Epidemiol. *60*, 1234–1238.

Boveri, T. (1887). Ueber den antheil des spermatozoon an der teilung des eies. Sitzungsber. Ges. Morph. Physiol. München *3*, 151–164.

Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. J. Cell Sci. *121*, 1–84.

Brinkley, B.R. (2001). Managing the centrosome numbers game: from chaos to stability in cancer cell division. Trends Cell Biol. *11*, 18–21.

Campbell, P.J. (2016). Somatic mutation in cancer and normal cells. Science. *349*, 961–968.

Cancer Research UK (2016). Cancer incidence by age. Retrieved February 24, 2017, from http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age#ref-0

Carvalho-Santos, Z., Azimzadeh, J., Pereira-Leal, J.B., and Bettencourt-Dias, M. (2011). Tracing the origins of centrioles, cilia, and flagella. J. Cell Biol. *194*, 165–175.

Chambers, A.F. (2009). MDA-MB-435 and M14 Cell Lines: identical but not M14 melanoma? Cancer Res. *69*, 5292–5293.

Chan, J.Y. (2011). A clinical overview of centrosome amplification in human cancers. Int. J. Biol. Sci. *7*, 1122–1144.

Chernoff, H., and Lehmann, E.L. (1954). The use of maximum likelihood estimates in $\chi 2$ tests for goodness of fit. Ann. Math. Stat. *25*, 579–586.

Coelho, P.A., Bury, L., Shahbazi, M.N., Liakath-ali, K., Tate, P.H., Wormald, S., Hindley, C.J., Huch, M., Archer, J., Skarnes, W.C., et al. (2015). Over-expression of Plk4 induces centrosome amplification, loss of primary cilia and associated tissue hyperplasia in the mouse. Open Biol. *5*, 150209.

Conover, W.J. (1971). Practical nonparametric statistics (John Wiley & Sons).

Conover, W.J., Johnson, M.E., and Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics *23*, 351–361.

Cooper, G.M. (2000). The Development and Causes of Cancer. In The Cell: A Molecular Approach.

Coux, O., Tanaka, K., and Goldberg, A.L. (1996). Structure and functions of the 20s and 26s

proteasomes. Annu. Rev. Biochem. *65*, 801–847.

Cunha-Ferreira, I., Rodrigues-Martins, A., Bento, I., Riparbelli, M., Zhang, W., Laue, E., Callaini, G., Glover, D.M., and Bettencourt-Dias, M. (2009). The SCF/Slimb ubiquitin ligase limits centrosome amplification through degradation of SAK/PLK4. Curr. Biol. *19*, 43–49.

Dahlmann, B. (2005). Proteasomes. Essays Biochem. *41*, 31–48.

Dammermann, A., and Merdes, A. (2002). Assembly of centrosomal proteins and microtubule organization depends on PCM-1. J. Cell Biol. *159*, 255–266.

Dick, J.E. (2008). Stem cell concepts renew cancer research. Blood *112*, 4793–4807.

Didier, C., Merdes, A., Gairin, J., and Jabrane-ferrat, N. (2008). Inhibition of proteasome activity impairs centrosome-dependent microtubule nucleation and organization. Mol. Biol. Cell *19*, 1220–1229.

Duncan, F.E., Moss, S.B., and Williams, C.J. (2006). Knockdown of the cAMP-Dependent Protein Kinase (PKA) Type Ia Regulatory Subunit in Mouse Oocytes Disrupts Meiotic Arrest and Results in Meiotic Spindle Defects. Dev. Dyn. *235*, 2961–2968.

Everitt, B. (1974). Cluster Analyses (London: Heinemann Educ. Books.).

Eyers, P.A., Liu, J., Hayashi, N.R., Lewellyn, A.L., Gautier, J., and Maller, J.L. (2005). Regulation of the G2/M Transition in Xenopus Oocytes by the cAMP-dependent Protein Kinase. J. Biol. Chem. *280*, 24339–24346.

Fava, L.L., Schuler, F., Sladky, V., Haschka, M.D., Soratroi, C., Eiterer, L., Demetz, E., Weiss, G., Geley, S., Nigg, E.A., et al. (2017). The PIDDosome activates p53 in response to supernumerary centrosomes. Genes Dev. *31*, 34–45.

Ferlay, J., Steliarova-foucher, E., Lortet-tieulent, J., Rosso, S., Coebergh, J.W.W., Combere, H., Formana, D., and Bray, F. (2013a). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. Eur. J. Cancer *49*, 1374–1403.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., and Bray, F. (2013b). GLOBOCAN 2012 v1.0, Cancer incidence and mortality worldwide: IARC CancerBase No. 11.

Ferlay, J., Soerjomataram, I.I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer *136*, E359–E386.

Foulkes, W.D., Ingunn, M., Chappuis, P.O., Be, L.R., Goffin, J.R., Wong, N., Trudel, M., and Akslen, L.A. (2003). Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. J. Natl. Cancer Inst. *95*, 1482–1485.

Fu, J., Lipinszki, Z., Rangone, H., Min, M., Mykura, C., Chao-Chu, J., Schneider, S., Dzhindzhev, N.S., Gottardo, M., Riparbelli, M.G., et al. (2016). Conserved molecular interactions in centriole-to-centrosome conversion. Nat. Cell Biol. *18*, 87–99.

Fuchs, E.J. (2011). Comparing two regression slopes by means of an ANCOVA. Retrieved August 16, 2017, from http://r-eco-evo.blogspot.pt/2011/08/comparing-two-regression-slopes-by.html

Fukasawa, K., Choi, T., Kuriyama, R., Rulong, S., and Woude, G.F. Vande (1996). Abnormal centrosome amplification in the absence of p53. Science (80-.). *271*, 1744–1747.

Gaillard, H., García-muse, T., and Aguilera, A. (2015). Replication stress and cancer. Nat. Rev. Cancer *15*, 276–289.

Ganem, N.J., Godinho, S.A., and Pellman, D. (2009). A mechanism linking extra centrosomes to chromosomal instability. Nature *460*, 278–282.

Garraway, L. a., and Lander, E.S. (2013). Lessons from the cancer genome. Cell *153*, 17–37.

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

Ghadimi, B.M., Sackett, D.L., Difilippantonio, M.J., Schröck, E., Neumann, T., Jauho, A., Auer, G., and Ried, T. (2000). Centrosome amplification and instability occurs exclusively in aneuploid, but not in diploid colorectal cancer cell Lines, and correlates with numerical chromosomal aberrations. Genes Chromosom. Cancer *27*, 183–190.

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

Giehl, M., Fabarius, A., Frank, O., Hochhaus, A., Hafner, M., Hehlmann, R., and Seifarth, W. (2005). Centrosome aberrations in chronic myeloid leukemia correlate with stage of disease and chromosomal instability. Leukemia *19*, 1192–1197.

Global Burden of Disease Cancer Collaboration (2015). The global burden of cancer 2013. JAMA Oncol. *1*, 505–527.

Godinho, S.A. (2015). Centrosome amplification and cancer: branching out. Mol. Cell. Oncol. *2*, e993252.

Godinho, S.A., and Pellman, D. (2014). Causes and consequences of centrosome abnormalities in cancer. Philos. Trans. R. Soc. oB *369*, 20130467.

Godinho, S.A., Kwon, M., and Pellman, D. (2009). Centrosomes and cancer: how cancer cells divide with too many centrosomes. Cancer Metastasis Rev *28*, 85–98.

Godinho, S.A., Picone, R., Burute, M., Dagher, R., Su, Y., Leung, C.T., Polyak, K., Brugge, J.S., Théry, M., and Pellman, D. (2014). Oncogene-like induction of cellular invasion from centrosome amplification. Nature *510*, 167–171.

Gonczy, P. (2015). Centrosomes and cancer: revisiting a long-standing relationship. Nat Rev Cancer *15*, 639–652.

Gould, R.R., and Borisy, G.G. (1977). The pericentriolar material in Chinese hamster ovary cells nucleates microtubule formation. J. Cell Biol. *73*, 601–615.

Grady, W.M. (2004). Genomic instability and colon cancer. Cancer Metastasis Rev. *23*, 11–27.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., and Altman, D.G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur. J. Epidemiol. *31*, 337–350.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Habedanck, R., Stierhof, Y.-D., Wilkinson, C.J., and Nigg, E. a (2005). The Polo kinase Plk4 functions in centriole duplication. Nat. Cell Biol. *7*, 1140–1146.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 319–326.

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

Holland, A.J., and Cleveland, D.W. (2009). Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat. Rev. Mol. Cell Biol. *10*, 478–487.

Holland, A.J., and Cleveland, D.W. (2014). Polo-like kinase 4 inhibition: a strategy for cancer therapy? Cancer Cell *26*, 151–153.

Hoyer-Fender, S. (2010). Centriole maturation and transformation to basal body. Semin. Cell Dev. Biol. *21*, 142–147.

Hsu, L., Kapali, M., Deloia, J.A., and Gallion, H.H. (2005). Centrosome abnormalities in ovarian cancer. Int. J. Cancer *113*, 746–751.

Jakobsen, L., Vanselow, K., Skogs, M., Toyoda, Y., Lundberg, E., Poser, I., Falkenby, L.G., Bennetzen, M., Westendorf, J., Nigg, E.A., et al. (2011). Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. EMBO J. *30*, 1520–1535.

Jayat, C., and Ratinaud, M.-H. (1993). Cell cycle analysis by flow cytometry: Principles and applications. In Biology of the Cell, pp. 15–25.

John Quackenbush (2001). Computational analysis of microarray data. Nat. Rev. Genet. *2*, 418–427.

Kammerer, R., Buchner, A., Palluch, P., Pongratz, T., Beyer, W., Johansson, A., Stepp, H., and Baumgartner, R. (2011). Induction of immune mediators in glioma and prostate cancer cells by non-lethal photodynamic therapy. PLoS One *6*, e21834.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *27*, 29–34.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462.

Kawamura, E., Fielding, A.B., Kannan, N., Balgi, A., Eaves, J., Roberge, M., and Dedhar, S. (2013). Identification of novel small molecule inhibitors of centrosome clustering in cancer cells. Oncotarget *4*, 1763–1776.

Kleylein-Sohn, J., Westendorf, J., Le Clech, M., Habedanck, R., Stierhof, Y.D., and Nigg, E.A. (2007). Plk4-Induced centriole Biogenesis in human cells. Dev. Cell *13*, 190–202.

Ko, M.J., Murata, K., Hwang, D., and Parvin, J.D. (2006). Inhibition of BRCA1 in breast cell lines causes the centrosome duplication cycle to be disconnected from the cell cycle. Oncogene *25*, 298–303.

Kocarnik, J.M., Shiovitz, S., and Phipps, A.I. (2015). Molecular phenotypes of colorectal cancer and potential clinical applications. Gastroenterol. Rep. *3*, 269–276.

Kolker, E., Higdon, R., and Hogan, J.M. (2006). Protein identification and expression analysis using mass spectrometry. Trends Microbiol. *14*, 229–235.

Korzeniewski, N., Treat, B., and Duensing, S. (2011). The HPV-16 E7 oncoprotein induces centriole multiplication through deregulation of Polo-like kinase 4 expression. Mol. Cancer *10*, 1–5.

Kovo, M., Kandli-cohen, M., Ben-haim, M., Galiani, D., and Carr, D.W. (2006). An active

protein kinase A (PKA) is involved in meiotic arrest of rat growing oocytes. Reproduction *132*, 33–43.

Krämer, A., Neben, K., and Ho, A.D. (2005). Centrosome aberrations in hematological malignancies. Cell Biol. Int. *29*, 375–383.

Kreso, A., and Dick, J.E. (2014). Evolution of the cancer stem cell model. Cell Stem Cell *14*, 275–291.

Kruskal, W.H., and Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. *47*, 583–621.

Kwon, M., Godinho, S.A., Chandhok, N.S., Ganem, N.J., Azioune, A., Thery, M., and Pellman, D. (2008). Mechanisms to suppress multipolar divisions in cancer cells with extra centrosomes. Genes Dev. *22*, 2189–2203.

Kwon, M., Bagonis, M., Danuser, G., and Pellman, D. (2015). Direct microtubule-binding by myosin-10 orients centrosomes toward retraction fibers and article subcortical actin clouds. Dev. Cell *34*, 323–337.

Lai, T.L., Robbins, H., and Wei, C.Z. (1979). Strong consistency of least squares estimates in multiple regression II. J. Multivar. Anal. *9*, 343–361.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for newcancer-associated genes. Nature *499*, 214–218.

Lazebnik, Y. (2010). What are the hallmarks of cancer? Nat. Rev. Cancer *10*, 232–233.

Leber, B., Maier, B., Fuchs, F., Chi, J., Riffel, P., Anderhub, S., Wagner, L., Ho, A.D., Salisbury, J.L., Boutros, M., et al. (2010). Proteins required for centrosome clustering in cancer cells. Sci. Transl. Med. *2*, 33–38.

Lentini, L., Amato, A., Schillaci, T., and Leonardo, A. Di (2007). Simultaneous Aurora-A/STK15 overexpression and centrosome amplification induce chromosomal instability in tumour cells with a MIN phenotype. BMC Cancer *7*,1–13.

Lerit, D.A., and Poulton, J.S. (2016). Centrosomes are multifunctional regulators of genome stability. Chromosom. Res. *24*, 5–17.

Leroy, B., Girard, L., Hollestelle, A., Minna, J.D., Gazdar, A.F., and Soussi, T. (2014). Analysis of TP53 mutation status in human cancer cell lines: a reassessment Bernard. Hum. Mutat. *35*, 756–765.

Levine, M.S., Bakker, B., Boeckx, B., Moyett, J., Lu, J., Vitre, B., Spierings, D.C., Lansdorp, P.M., Cleveland, D.W., Lambrechts, D., et al. (2017). Centrosome amplification is sufficient to promote spontaneous tumorigenesis in mammals. Dev. Cell *40*, 1–10.

Li, J., Tan, M., Li, L., Pamarthy, D., Lawrence, T.S., and Sun, Y. (2005). SAK, a new Polo-Like Kinase, is transcriptionally repressed by p53 and induces apoptosis upon RNAi silencing. Neoplasia *7*, 312–323.

Lingle, W.L., and Salisbury, J.L. (1999). Altered centrosome structure is associated with abnormal mitoses in human breast tumors. Am. J. Pathol. *155*, 1941–1951.

Lingle, Wi.L., Lutz, W. ard H., Ingle, J.N., Maihle, N.J., and Salisbury, J.L. (1998). Centrosome hypertrophy in human breast tumors: Implications for genomic stability and cell polarity. Proc. Natl. Acad. Sci. *95*, 2950–2955.

Liu, H., Andrade, P.D., Fulmer-smentek, S., Lorenzi, P., Kohn, K.W., Weinstein, J.N.,

Pommier, Y., and Reinhold, W.C. (2010). mRNA and microRNA Expression Profiles of the NCI-60 Integrated with Drug Activities. Mol. Cancer Ther. *9*, 1080–1092.

Livneh, I., Cohen-kaplan, V., Cohen-rosenzweig, C., Avni, N., and Ciechanover, A. (2016). The life cycle of the 26S proteasome: from birth, through regulation and function, and onto its death. Cell Res. *26*, 869–885.

Löffler, H., Fechter, A., Liu, F.Y., Poppelreuther, S., and Krämer, A. (2012). DNA damage-induced centrosome amplification occurs via excessive formation of centriolar satellites. Oncogene *32*, 2963–2972.

Loncarek, J., Hergert, P., Magidson, V., and Khodjakov, A. (2008). Control of daughter centriole formation by the pericentriolar material. Nat. Cell Biol. *10*, 322–328.

Macconaill, L.E., and Garraway, L.A. (2015). Clinical Implications of the cancer genome. J. Clin. Oncol. *28*, 5219–5228.

Macheret, M., and Halazonetis, T.D. (2015). DNA replication stress as a hallmark of cancer. Annu. Rev. Pathol. Mech. Dis. *10*, 425–448.

Maiato, H., and Logarinho, E. (2016). Mitotic spindle multipolarity without centrosome amplification. Nat. Cell Biol. *16*, 386–394.

Van de Mark, D., Kong, D., Loncarek, J., and Stearns, T. (2015). MDM1 is a microtubule-binding protein that negatively regulates centriole duplication. Mol. Biol. Cell *26*, 3788–3802.

Marteil, G., Guerrero, A., Vieira, A., de Almeida, B.P., Machado, P., Mendonça, S., Mesquita, M., Villarreal, B., Fonseca, I., Francia, M., et al. (*manuscript in preparation*). Deregulation of centriole length is widespread in cancer and promotes centriole amplification and chromosome missegregation.

Marthiens, V., Rujano, M. a, Pennetier, C., Tessier, S., Paul-Gilloteaux, P., and Basto, R. (2013). Centrosome amplification causes microcephaly. Nat. Cell Biol. *15*, 731–740.

Mason, J.M., Lin, D.C.-C., Wei, X., Che, Y., Yao, Y., Kiarash, R., Cescon, D.W., Fletcher, G.C., Awrey, D.E., Bray, M.R., et al. (2014). Functional characterization of CFI-400945, a Polo-like Kinase 4 inhibitor, as a potential anticancer agent. Cancer Cell *26*, 163–176.

Mchugh, M.L. (2013). The Chi-square test of independence. Biochem. Medica *23*, 143–149.

Meraldi, P., Honda, R., and Nigg, E.A. (2002). Aurora-A overexpression reveals tetraploidization as a major route to centrosome amplification in p53-/- cells. EMBO J. *21*, 483–492.

Merlo, L.M.F., Pepper, J.W., Reid, B.J., and Maley, C.C. (2006). Cancer as an evolutionary and ecological process. Nat. Rev. Cancer *6*, 924–935.

Miller, G.A., and Chapman, J.P. (2001). Misunderstanding analysis of covariance. J. Abnorm. Psychol. *110*, 40–48.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability - an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. *11*, 220–228.

Nielsen, F.C., Hansen, T.V.O., and Sørensen, C.S. (2016). Hereditary breast and ovarian cancer : new genes in confined pathways. Nat. Rev. Cancer *16*, 599–612.

Nigg, E.A. (2002). Centrosome aberrations: cause or consequence of cancer progression? Nat. Rev. Cancer *2*, 815–825.

Nigg, E.A. (2006). Origins and consequences of centrosome aberrations in human cancers. Int. J. Cancer *119*, 2717–2723.

Nigg, E.A., and Raff, J.W. (2009). Centrioles, centrosomes, and cilia in health and disease. Cell *139*, 663–678.

Nigg, E.A., and Stearns, T. (2011). The centrosome cycle: centriole biogenesis, duplication and inherent asymmetries. Nat. Cell Biol. *13*, 1154–1160.

Nigg, E.A., Schiifer, G., Hilz, H., and Eppenberger, H.M. (1985). Cyclic-AMP-dependent protein kinase type II is associated with the golgi complex and with centrosomes. Cell *41*, 1039–1051.

Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W.C., Waltham, M., Kouros-mehr, H., Bussey, K.J., Lee, J.K., Espina, V., et al. (2003). Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc. Natl. Acad. Sci. *100*, 14229–14234.

Noble, W.S. (2009). How does multiple testing correction work? Nat. Biotechnol. *27*, 1135–1137.

Nogales-Cadenas, R., Abascal, F., Dıez-Perez, J., Carazo, J.M., and Pascual-Montano, A. (2009). CentrosomeDB: a human centrosomal proteins database. Nucleic Acids Res. *37*, 175–180.

Nowak, B.M.A., and Waclaw, B. (2017). Genes, environment, and “bad luck.” Science (80-.). *355*, 1266–1267.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science (80-.). *194*, 23–28.

O’Connell, K.F., Caron, C., Kopish, K.R., Hurd, D.D., Kemphues, K.J., Li, Y., and White, J.G. (2001). The C. elegans zyg-1 gene encodes a regulator of centrosome duplication with distinct maternal and paternal roles in the embryo. Cell *105*, 547–558.

O’Connor, P.M., Jackman, J., Bae, I., Myers, T.G., Fan, S., Mutoh, M., Scudiero, D.A., Monks, A., Sausville, E.A., Weinstein, J.N., et al. (1997). Characterization of the p53 Tumor Suppressor Pathway in Cell Lines of the National Cancer Institute Anticancer Drug Screen and Correlations with the Growth-Inhibitory Potency of 123 Anticancer Agents. Cancer Res. *57*, 4285–4300.

Ogden, A., Rida, P.C.G., and Aneja, R. (2017). Prognostic value of CA20, a score based on centrosome amplification-associated genes, in breast tumors. Sci. Rep. *7*, 262.

Park, E.S., Rabinovsky, R., Carey, M., Hennessy, B.T., Agarwal, R., Liu, W., Ju, Z., Deng, W., Lu, Y., Woo, H.G., et al. (2010). Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. Mol. Cancer Ther. *9*, 257–267.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Peel, N., Stevens, N.R., Basto, R., and Raff, J.W. (2007). Overexpressing centriole-replication proteins In vivo induces centriole overduplication and de novo formation. Curr. Biol. *17*, 834–843.

Phipps, A.I., Limburg, P.J., Baron, J.A., Burnett-hartman, A.N., Weisenberger, D.J., Laird, P.W., Sinicrope, F.A., Rosty, C., Buchanan, D., Potter, J.D., et al. (2015). Association between molecular subtypes of colorectal cancer and patient survival. Gastroenterology *148*, 77–87.

Piel, M., Meyer, P., Khodjakov, A., Rieder, C.L., and Bornens, M. (2000). The respective contributions of the mother and daughter centrioles to centrosome activity and behavior in vertebrate cells. J. Cell Biol. *149*, 317–329.

Pihan, G.A., Purohit, A., Wallace, J., Knecht, H., Woda, B., Quesenberry, P., and Doxsey, S.J. (1998). Centrosome defects and genetic instability in malignant tumors. *58*, 3974–3985.

Pikor, L., Thu, K., Vucic, E., and Lam, W. (2013). The detection and implication of genome instability in cancer. Cancer Metastasis Rev. *32*, 341–352.

Praetorius, H.A., and Spring, K.R. (2005). A Physiological view of the primary cilium. Annu. Rev. Physiol. *67*, 515–529.

Quammen, D. (2008). Contagious cancer: the evolution of a killer. Harpers. (N. Y. N. Y).

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rae, J.M., Creighton, C.J., Meck, J.M., Haddad, B.R., and Johnson, M.D. (2007). MDA-MB-435 cells are derived from M14 Melanoma cells — a loss for breast cancer, but a boon for melanoma research. Breast Cancer Res. Treat. *104*, 13–19.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature *505*, 302–308.

Reinhold, W.C., Sunshine, M., Liu, H., Varma, S., Kohn, K.W., Morris, J., Doroshow, J., and Pommier, Y. (2012). CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. *72*, 3499–3511.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Ring, D., Hubble, R., and Kirschner, M. (1982). Mitosis in a cell with multiple centrioles. J. Cell Biol. *94*, 549–556.

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

Rodrigues-Martins, A., Riparbelli, M., Callaini, G., Glover, D.M., and Bettencourt-Dias, M. (2007). Revisiting the role of the mother centriole in centriole biogenesis. Science (80-.). *316*, 1046–1050.

Rogers, G.C., Rusan, N.M., Roberts, D.M., Peifer, M., and Rogers, S.L. (2009). The SCF/Slimb ubiquitin ligase regulates Plk4/Sak levels to block centriole reduplication. J. Cell Biol. *184*, 225–240.

Roschke, A. V, Tonon, G., Gehlhaus, K.S., Mctyre, N., Bussey, K.J., Lababidi, S., Scudiero, D.A., Weinstein, J.N., and Kirsch, I.R. (2003). Karyotypic Complexity of the NCI-60 Drug-Screening Panel. Cancer Res. *63*, 8634–8647.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Rijn, M. Van De, Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. *24*, 227–235.

Sato, N., Mizumoto, K., Nakamura, M., Nakamura, K., Kusumoto, M., and Niiyama, H.

(1999). Centrosome abnormalities in pancreatic ductal carcinoma. Clin. Cancer Res. *5*, 963–970.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nat. Genet. *24*, 236–244.

Schmidt, T.I., Kleylein-Sohn, J., Westendorf, J., Le Clech, M., Lavoie, S.B., Stierhof, Y.D., and Nigg, E.A. (2009). Control of centriole length by CPAP and CP110. Curr. Biol. *19*, 1005–1011.

Sedgwick, P. (2014). Understanding statistical hypothesis testing. BMJ *348*.

Sedgwick, P. (2015). A comparison of parametric and non-parametric statistical tests. BMJ *350*.

Seo, M.Y., Jang, W., and Rhee, K. (2015). Integrity of the Pericentriolar Material Is Essential for Maintaining Centriole Association during M Phase. PLoS One *10*, e0138905.

Serçin, Ö., Larsimont, J.-C., Karambelas, A.E., Marthiens, V., Moers, V., Boeckx, B., Le Mercier, M., Lambrechts, D., Basto, R., and Blanpain, C. (2015). Transient PLK4 overexpression accelerates tumorigenesis in p53-deficient epidermis. Nat. Cell Biol. *18*, 100–110.

Shackleton, M., Quintana, E., Fearon, E.R., and Morrison, S.J. (2009). Heterogeneity in cancer: cancer stem cells versus clonal evolution. Cell *4*, 822–829.

Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., Major, S., Morita, D., Chary, K.K., Reimers, M.A., Scherf, U., Kahn, A., Dolginow, D., et al. (2007). Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol. Biol. Cell *6*, 820–832.

Shankavaram, U.T., Varma, S., Kane, D., Sunshine, M., Chary, K.K., Reinhold, W.C., Pommier, Y., and Weinstein, J.N. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics *10*.

Shaw, P.A., and Proschan, M.A. (2013). Null but not void: considerations for hypothesis testing. Stat. Med. *32*, 196–205.

Shen, Z. (2011). Genomic instability and cancer. J. Mol. Cell Biol. *3*, 1–3.

Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.

Sieber, O.M., Heinimann, K., and Tomlinson, I.P.M. (2003). Genomic instability - the engine of tumorigenesis? Nat. Rev. Cancer *3*, 701–708.

Skålhegg, B.S., and Tasken, K. (2000). Specificity in the cAMP/PKA signaling pathway. Differential expression, regulation and subcellular localization of subunits of PKA. Front. Biosci. *5*, 678–693.

Sluder, G., and Nordberg, J.J. (2004). The good, the bad and the ugly: the practical consequences of centrosome amplification. Curr. Opin. Cell Biol. *16*, 49–54.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article 3.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., Rijn, M. Van De, Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. *98*, 10869–

10874.

Spearman, C. (1904). The proof and measurement of association between two things. Am. J. Psychol. *15*, 72–101.

Starita, L.M., Machida, Y., Sankaran, S., Elias, J.E., Griffin, K., Schlegel, B.P., Gygi, S.P., and Parvin, J.D. (2004). BRCA1-dependent ubiquitination of γ-tubulin regulates centrosome number. Mol. Cell. Biolology *24*, 8457–8466.

Stevens, N.R., Raposo, A.A.S.F., Basto, R., St Johnston, D., and Raff, J.W. (2007). From stem cell to embryo without centrioles. Curr. Biol. *17*, 1498–1503.

Strachan, T., and Read, A. (1996). Cancer Genetics. In Human Molecular Genetics, (Garland Science), pp. 537–567.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Strnad, P., and Gönczy, P. (2008). Mechanisms of procentriole formation. Trends Cell Biol. *18*, 389–396.

Strnad, P., Leidel, S., Vinogradova, T., Euteneuer, U., and Gönczy, P. (2007). Regulated HsSAS-6 levels ensure formation of a single procentriole per centriole during the centrosome duplication cycle. Dev. Cell *13*, 203–213.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M. a, Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Tomasetti, C., and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science (80-.). *347*, 78–81.

Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., Nichols, M., Nichols, M., Townsend, N., Scarborough, P., Rayner, M., et al. (2016). Cardiovascular disease in Europe: epidemiological update 2016. Eur. Heart J. *37*, 3232–3245.

Vinogradova, T., Paul, R., Grimaldi, a. D., Loncarek, J., Miller, P.M., Yampolsky, D., Magidson, V., Khodjakov, A., Mogilner, A., and Kaverina, I. (2012). Concerted effort of centrosomal and Golgi-derived microtubules is required for proper Golgi complex assembly but not for maintenance. Mol. Biol. Cell *23*, 820–833.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz Jr., L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science (80-.). *339*, 1546–1558.

Wang, X., Tsai, J.-W., Imai, J.H., Lian, W.-N., Vallee, R.B., and Shi, S.-H. (2009). Asymmetric centrosome inheritance maintains neural progenitors in the neocortex. Nature *461*, 947–955.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2016). gplots: various R programming tools for plotting data. R package version 3.0.1. https://CRAN.R-project.org/package=gplots.

Watts, C.A., Richards, F.M., Bender, A., Bond, P.J., Korb, O., Kern, O., Riddick, M., Owen, P., Myers, R.M., Raff, J., et al. (2013). Design, synthesis, and biological evaluation of an allosteric inhibitor of HSET that targets cancer cells with supernumerary centrosomes. Chem. Biol. *20*, 1399–1410.

Weaver, B.A.A., Silk, A.D., Montagna, C., Verdier-pinard, P., and Cleveland, D.W. (2007).

Aneuploidy acts both oncogenically and as a tumor suppressor. Cancer Cell *11*, 25–36.

White, R.A., Pan, Z., and Salisbury, J.L. (2000). GFP-centrin as a marker for centriole dynamics in living cells. Microsc. Res. Tech. *49*, 451–457.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis (Springer-Verlag).

Wigley, W.C., Fabunmi, R.P., Lee, M.G., Marino, C.R., Muallem, S., Demartino, G.N., and Thomas, P.J. (1999). Dynamic association of proteasomal machinery with the centrosome. J. Cell Biol. *145*, 481–490.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bull. *1*, 80–83.

Witte, R.S., and Witte, J.S. (2013). Statistics (Wiley).

Wu, J., Mikule, K., Wang, W., Su, N., Petteruti, P., Gharahdaghi, F., Code, E., Zhu, X., Jacques, K., Lai, Z., et al. (2013). Discovery and mechanistic study of a small molecule inhibitor for motor protein KIFC1. ACS Chem. Biol. *8*, 2201–2208.

Yang, B., Lamb, M.L., Zhang, T., Hennessy, E.J., Grewal, G., Sha, L., Zambrowski, M., Block, M.H., Dowling, J.E., Su, N., et al. (2014). Discovery of potent KIFC1 inhibitors using a method of integrated high-throughput synthesis and screening. *57*, 9958−9970.

Yates, F. (2012). Contingency tables involving small numbers and the $\chi 2$ test. Suppl. to J. R. Stat. Soc. *1*, 217–235.

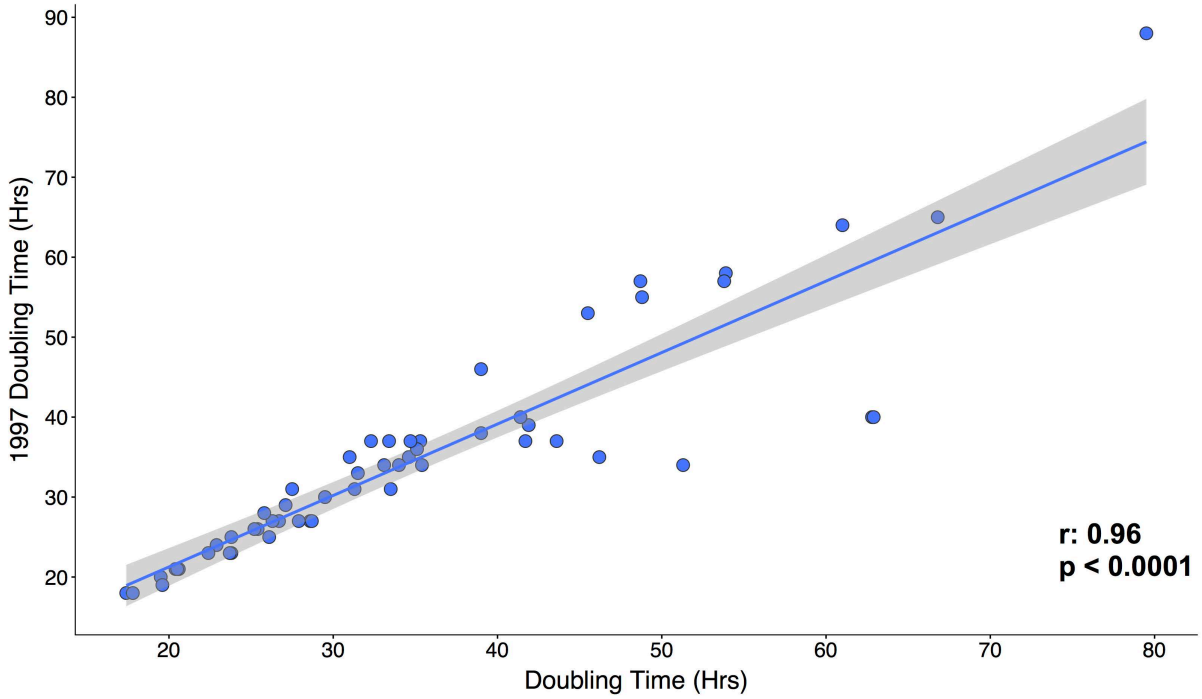
Zyss, D., and Gergely, F. (2009). Centrosome function in cancer: guilty or innocent? Trends Cell Biol. *19*, 334–346.

ANNEXES

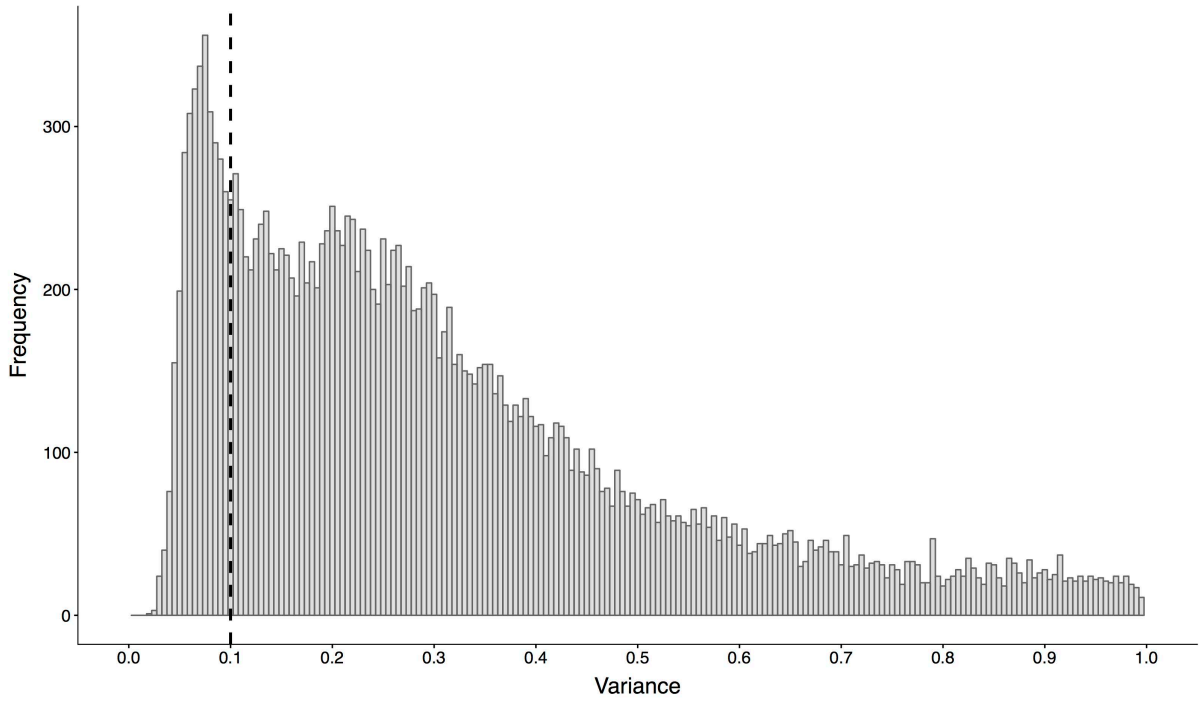
Annex 1 Table with centriole abnormality metrics for each NCI-60 cell line: percentage of cells with CA, mean of centriole number, percentage of cells with COE and mean of centriole length. Cell lines whose metrics were calculated from primary screening data are highlighted (blue).

Cell Line	Tissue	% of cells with CA	Mean of centrioles number	% of cells with COE	Mean of centrioles length (nm)
BT-549	Breast	24.5	4.38	7.5	311.02
HS578T	Breast	32.6	6.11	8.7	295.39
MCF7	Breast	5.2	4.10	5.2	304.41
MDA-MB-231	Breast	14.5	4.24	7.3	286.32
MDA-MB-468	Breast	23.6	4.60	3.6	289.01
T47D	Breast	9.1	4.07	0	286.24
SF-268	CNS	1.7	4.02	1.7	287.06
SF-295	CNS	5.7	4.02	7.5	319.24
SF-539	CNS	23.6	4.84	5.5	306.70
SNB-19	CNS	9.1	4.18	0	298.83
SNB-75	CNS	12.7	4.65	1.8	291.90
U251	CNS	22.4	4.35	0	273.51
COLO205	Colon	57.1	5.77	1.8	284.09
HCC-2998	Colon	23.9	4.49	2.8	252.45
HCT-116	Colon	7	4.11	10.5	292.40
HCT-15	Colon	12.1	*NA*	0	*NA*
HT29	Colon	5.8	4.13	0	261.73
KM12	Colon	9.6	*NA*	0	*NA*
SW-620	Colon	10.3	4.67	1.7	264.27
A549	Lung	5.1	4.14	0	270.50
EKVX	Lung	4.9	*NA*	3.9	*NA*
HOP-62	Lung	62.1	6.69	29.3	333.63
HOP-92	Lung	25	4.85	17.3	321.42
NCI-H226	Lung	10	*NA*	2	*NA*
NCI-H23	Lung	37.5	5.88	19.6	300.03
NCI-H322M	Lung	34.6	5.67	1.9	266.38
NCI-H460	Lung	9.4	4.26	0	287.90
NCI-H522	Lung	3.9	4.14	3.9	303.87
CCRF-CEM	Blood	28.1	4.46	3.5	285.57
HL-60	Blood	48.2	4.82	7.1	284.77
K-562	Blood	19.4	4.39	0	276.12
MOLT-4	Blood	22.8	4.40	5.3	258.51
RPMI-8226	Blood	26.3	4.75	5.3	281.85
SR	Blood	14.3	4.43	1.8	278.90
LOXIMVI	Skin	14	*NA*	2	*NA*
M14	Skin	5.4	4.23	3.6	281.25
MALME-3M	Skin	40.4	5.82	19.3	313.82
MDA-MB-435	Skin	29.6	5.41	42.6	398.65
SK-MEL-2	Skin	9.3	4.30	7.4	277.89
SK-MEL-28	Skin	13.7	4.12	9.8	316.56
SK-MEL-5	Skin	14.8	*NA*	0	*NA*
UACC-257	Skin	15.1	4.45	0	265.78
UACC-62	Skin	7.1	4.18	8.9	307.32
IGROV1	Ovaries	7.3	4.18	5.5	266.72
NCI-ADR-RES	Ovaries	5.5	4.20	0	277.14
OVCAR-3	Ovaries	14.8	4.04	0	280.41
OVCAR-4	Ovaries	20	4.69	9.1	306.77
OVCAR-5	Ovaries	15.7	*NA*	5.9	*NA*
OVCAR-8	Ovaries	9.3	4.31	1.9	287.56
SK-OV-3	Ovaries	16.7	4.19	0	298.40
DU-145	Prostate	16.1	5.02	1.8	294.75
PC-3	Prostate	7.3	4.00	1.8	303.68
786-0	Kidney	7.4	3.91	0	280.86
A498	Kidney	7.4	*NA*	0	*NA*
ACHN	Kidney	11.6	4.40	0	275.77
CAKI-1	Kidney	18.9	4.42	3.8	293.51
RXF-393	Kidney	30.6	4.65	18.4	301.76
SN12C	Kidney	11.1	4.20	13	285.03
TK-10	Kidney	5.8	4.04	0	296.32
UO-31	Kidney	0	4.00	0	287.30

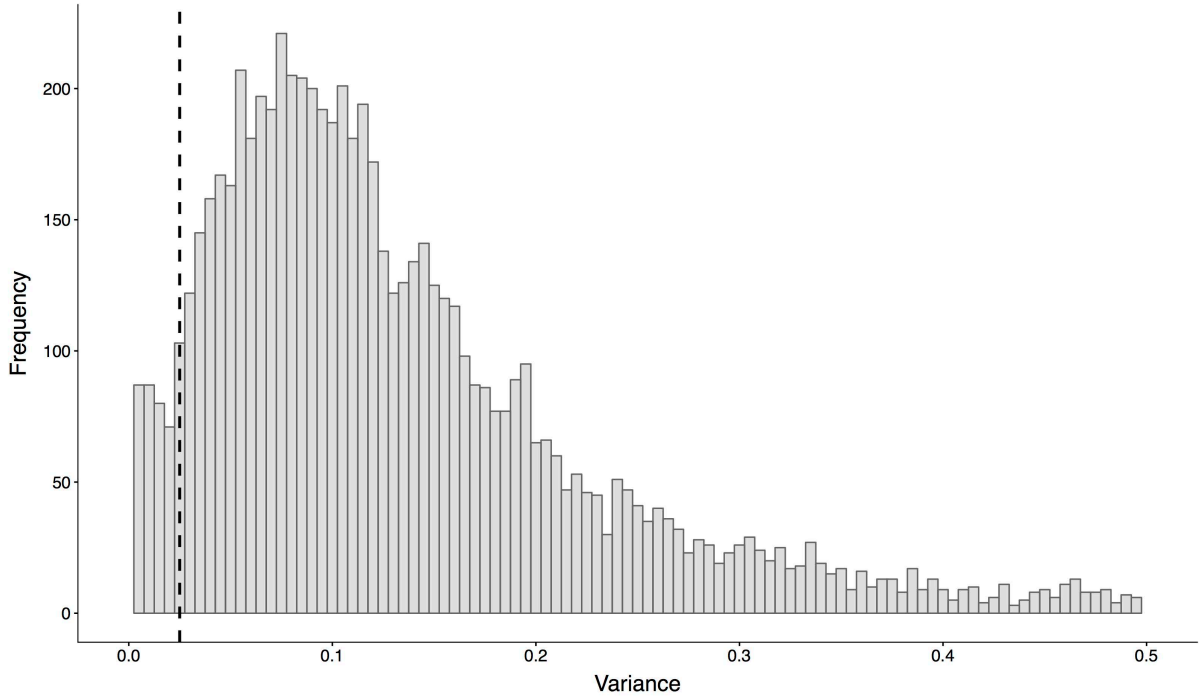
Annex 2 Correlation plot between cell lines' doubling time estimated by the U.S. National Cancer Institute and calculated on O'Connor et al., 1997 (Spearman correlation coefficient: 0.96, $p < 0.0001$). The grey shade around the blue linear regression line represents its 95% confidence interval.



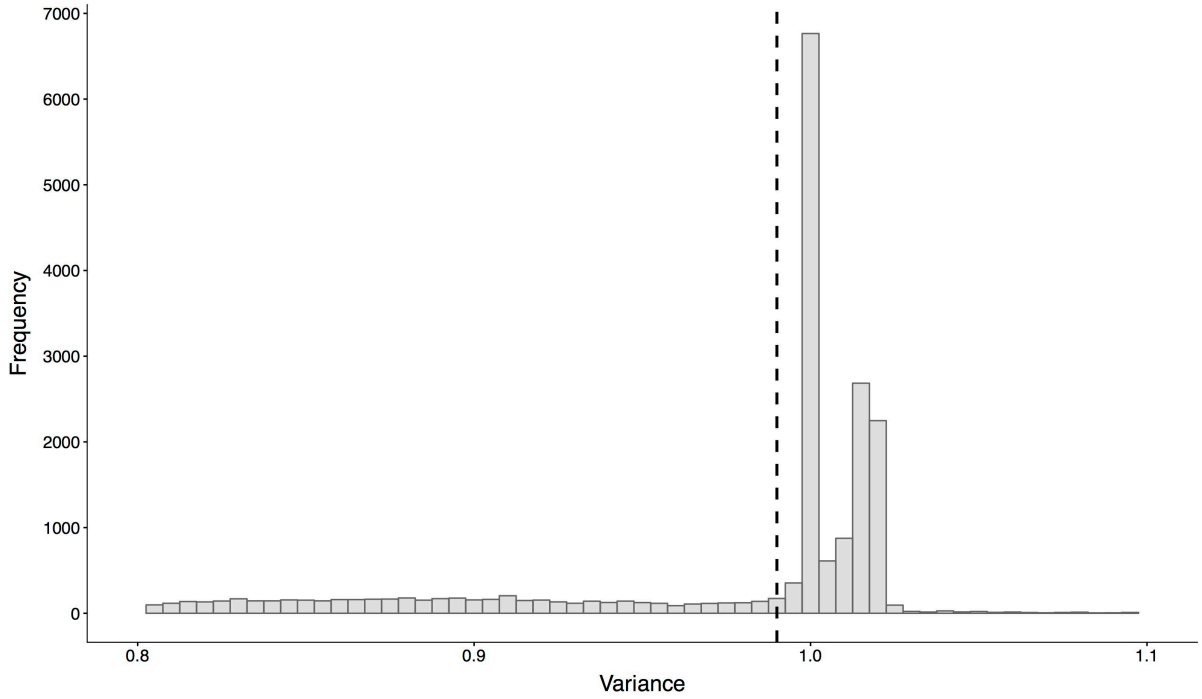
Annex 3 Distribution of gene expression variance across NCI60 samples. Microarray probes with log-intensity variance higher than 1 are not included. Vertical dashed line (log-intensity variance = 0.1) represents the quality control criteria used to remove probes whose expression does not vary across samples.



Annex 4 Distribution of protein expression variance across NCI60 samples. Peptides with log-intensity-based label-free quantification variance higher than 0.5 are not included. Vertical dashed line (log-intensity variance = 0.025) represents the quality control criteria used to remove peptides whose expression does not vary across samples.



Annex 5 Distribution of drug activity variance across NCI60 samples. Compounds with log-intensity variance higher than 1.1 or lower than 0.8 are not included. Vertical dashed line (activity variance = 0.99) represents the quality control criteria used to remove compounds whose activity does not vary across samples. Drug activity measured in z-transformed negative log10 of GI50.



Annex 6 Tissue hierarchical clustering based on centriole heterogeneity. Unsupervised hierarchical clustering using Euclidean distances calculated based on Fligner-Killeen p-values across all combinations of tissue pairs for centriole **(a)** number and **(b)** length heterogeneity. Heatmap colour scale according with -log10 of Fligner-Killeen p-values: darker blue reflects bigger differences across tissue pairs. Main clusters are highlighted with different shades of grey.