Comprehensive genomic characterization of matched primary-metastatic lung adenocarcinomas using a multiparameter nuclei flow-sorting approach

---

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Thomas Lorber

aus München, Deutschland

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von:

Prof. Dr. Erich Nigg
(Fakultätsverantwortlicher)

Prof. Dr. Lukas Bubendorf
(Dissertationsleiter)

Prof. Dr. Lukas Sommer
(Korreferent)

Basel, den 13. Dezember 2016

Prof. Dr. Jörg Schibler
(Dekan)

I dedicate this work to my dear wife Carolin and my family

"Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts – some of them interesting or curious but making no meaningful picture as a whole."

T. Dobzhansky – "Nothing in Biology Makes Sense Except in the Light of Evolution"
American Biology Teacher, 1973

Table of contents

Acknowledgements

I am deeply grateful to Prof. Dr. Lukas Bubendorf for giving me the opportunity to perform my thesis in his research group and for providing such an important and exciting research topic. The funding of the Swiss National Foundation (SNF) made this project possible.

I am also very grateful to PD Dr. Christian Ruiz for his supervision during my thesis, support and scientific discussions, and his endless motivation.

I would like to thank Prof. Dr. Erich Nigg and Prof. Dr. Lukas Sommer for being member of my thesis committee, for helpful scientific discussions and suggestions.

Special thanks to Prof. Dr. Ruth Chiquet-Ehrismann. She was a valuable member of my thesis committee, but unfortunately passed away way too young on September, 4th 2015. RIP, Ruth.

Warm thanks to all the members of the laboratories for technical help and fruitful scientific discussions, especially Tanja, Valeria, Sabrina, Arthur, Petra, David M., Mariacarla, Luca, Bruno, David J., Sebastian, Gabriel, Cristina, Nadja, Martin, Maarit, Freddy, Nicole, Salvatore, Charlotte, Jasmin H., Katharina, Jasmin M., Alex, Lisa, Anja, Spasenija, Sara, Heike, Matthias, Sandra, Sina, Diren, and Pascal. They generated a friendly, stimulating, and enjoyable environment and helped me throughout these years.

I am grateful for the collaboration with Dr. Noemi Andor. She helped tremendously with the biostatistics in this project. I admire her scientific mind that always came up with very helpful and valuable ideas.

I appreciate the help of my colleague Dr. Michal Kovac. I could always come to him for help. I am also grateful for our friendship and time we spent running in the parks and woods of Basel. He trained with me for my first 42.195 km, which I would not have been able to finish without his help.

I am especially indebted to my dear colleague Dr. Darius Juškevičius for sharing office, laboratories and scientific knowledge with me. Importantly, I am very grateful for our friendship. His support and encouragement was worth much more than I can express on paper and I enjoyed the time we spent in and outside the lab. Particularly, I will remember our endless table tennis games (“SCUB02”).

I would like to thank my graduate school teacher Werner Haag for leading me into the world of science. It was his devotion to biology and teaching that eventually inspired me to study biology.

Personally, I would like to thank my family for moral support throughout the years. I am deeply grateful for unlimited care and generous support of my parents, Waltraud and Hans, and my brother Colin. I have to thank them for who I am today. If character is somehow imprinted in our genes, I inherited the way to critically observe in research as well in life from my paternal genes.

Last but not least, my sincerest and deepest gratitude to my wife, Carolin, for her unwavering love, support, encouragement, patience, devotion and irreplaceable heartiness since 2000.

List of abbreviations

2N	*Diploid population*
aCGH	*Array-comparative genomic hybridization; also Array-CGH*
AEC1	*Type I alveolar epithelial cells*
AEC2	*Type II alveolar epithelial cells*
ALK	*Anaplastic lymphoma kinase*
AN	*Aneuploid population*
ATM	*Ataxia-telangiectasia mutated*
ATP	*Adenosine triphosphate*
BAP	*Benzo[a]pyrene*
CCP	*Comprehensive Cancer Panel*
CGH	*Comparative genomic hybridization*
CIN	*Chromosomal instability*
CNV	*Copy number variation (germline)*
CSC	*Cancer stem cell*
CTC	*Circulating tumor cells*
DAPI	*4´,6-diamidino-2-phenylindole; DNA intercalating dye*
DNA	*Deoxyribonucleic acid*
dsDNA	*Double-stranded DNA*
EDTA	*Ethylenediaminetetraacetic acid*
EGFR	*Epidermal growth factor receptor*
EML4	*Microtubule-associated protein 4 gene*
ETS	*Environmental tobacco smoke*
FCS	*Fetal calf serum*
FFPE	*Formalin-fixed and paraffin-embedded*
GDP	*Guanosine diphosphate*
GTP	*Guanosine-5'-triphosphat*
ICM	*Image-based cytometry*
IfP	*Institute for Pathology of the University Hospital Basel*
IGV	*Integrated Genomics Viewer*
IHC	*Immunohistochemistry*
ITH	*Intratumor heterogeneity*
KRAS	*V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog*
LOH	*Loss of heterozygosity*
LUAD	*Lung adenocarcinoma*
LUSC	*Squamous cell lung carcinoma*
MDM2	*Mouse double minute 2 homolog*
NGS	*Next-generation sequencing*
NIM	*Nuclei isolation medium*
NNK	*Nicotine-derived nitrosamine ketones*
NSCLC	*Non-small cell lung cancer*
PAH	*Polycyclic aromatic hydrocarbons*
PBS	*Phosphate-buffered saline*
RTK	*Receptor tyrosine kinase*
SAC	*Spindle assembly checkpoint*
SCNA	*Somatic copy number aberration*
SNP	*Single nucleotide polymorphism*
SNV	*Single nucleotide variant*
SOX10	*Sry-related HMg-Box gene 10*
TK	*Tyrosine kinase*
TKI	*Tyrosine kinase inhibitor*
TSG	*Tumor suppressor gene*
TTF1	*Thyroid transcription factor 1*
VAF	*Variant allele frequency*
WGA	*Whole genome amplification*
WGD	*Whole genome duplication*

Glossary

Aneuploidy	*Abnormal number of (parts of) chromosomes (Figure 13); absence of euploidy*
Array-comparative genomic hybridization	*A microarray-based high-resolution CGH method*
Chromoplexy	*Rearrangements across several chromosomes*
Chromothripsis	*The shattering of the genome in one catastrophic event*
Comparative genomic hybridization	*Conventional cytogenetic method for analyzing CNVs and SCNAs*
Copy number variation	*Germline (normal) copy number variation in normal cells and tumor cells*
Diploid tumor	*Tumor or tumor population that appears as a diploid peak in a DNA histogram*
Driver mutation	*Mutation that confers a selective advantage*
log2-ratios	*Logarithmic intensity ratio of tumor and control DNA in an aCGH experiment*
Loss of heterozygosity	*Complete loss of one of the two parental alleles with or without (copy-neutral LOH) a copy number change*
Next-generation sequencing	*High-throughput sequencing technologies based on massively parallel DNA amplification and sequencing*
Oncogene	*Gene that confers a selective advantage if hit by a gain-of-function mutation*
P/M divergence	G*enetic divergence between primary-metastatic pairs*
Passenger mutation	*A neutral mutation that does not confers a selective advantage*
Population	*A group of cells (or nuclei) that form a peak in a DNA histogram*
Private mutation	*A SNV or SCNA that is present in only one tumor lesion investigated for a given patient*
Purity	*The relative proportion of tumor vs normal cells in a sample*
Shared mutation	*A SNV or SCNA that is present in more than one but not all tumor lesions investigated for a given patient*
Single nucleotide polymorphism	*Mutation that exists in the human population and is found in both tumor and normal tissue*
Single nucleotide variant	*Mutation that occurred in the tumor, comprising also mutations detected by NGS*
Somatic copy number aberration	*Chromosomal aberration in the cancer genome of tumor cells*
Truncal mutation	*A SNV or SCNA that is present in all tumor cells in a single tumor lesion*
Tumor suppressor gene	*A gene that confers a selective advantage if hit by a loss-of-function mutation*
Tumorigenesis	*The process of tumor development*
Ubiquitous mutation	A *SNV or SCNA that is present in all tumor lesions investigated for a given patient*
Variant allele frequency	*Fraction of alleles (in NGS data, reads) carrying a mutation*

Summary

Lung cancer is the most frequent cancer worldwide and is responsible for more death than any other tumor type. Genomic intratumor heterogeneity (ITH) is thought to be implicated in the resistance to current therapies, but in spite of its high prevalence, only a few studies have investigated genomic ITH in non-small cell lung cancer (NSCLC). Moreover, these studies have mainly focused on primary tumors. Yet, it is the progression to metastatic disease that makes lung cancer so lethal.

We sought to shed light on the evolution of metastasizing lung adenocarcinoma (LUAD) by investigating the extent of genomic ITH in clonally related primary and metastatic sites in a cohort of 16 LUAD patients. We established and validated the use of a refined multiparameter nuclei flow-sorting approach and demonstrated that it substantially increases the purity of tumor DNA for subsequent genomic analyses. We applied this method to sort tumor populations from bulk tissues in order to determine somatic copy number aberrations (SCNAs) by array-comparative genomic hybridization (aCGH) and single nucleotide variants (SNVs) in 409 well-known cancer genes by targeted ultra-deep sequencing.

This comprehensive genomic analysis revealed that 88% of SCNAs and 78% of SNVs were propagated from primary tumors to metastases, which indicated their accumulation in the primary tumors before metastatic dissemination. The fact that SCNAs were not increased at metastatic sites resulted in stable ploidies across our cohort even over long time periods. Despite the fact that the pure presence of SCNAs implies chromosomal instability (CIN) at some point during tumorigenesis, our data suggested that there is no or little increase in CIN during metastatic progression. Nevertheless, the analysis of shared and private alterations demonstrated a broad continuum from linear to parallel progression, indicating a late and early dissemination of metastases from their primary tumors, respectively. Furthermore, we confirmed that both primary tumors and metastases can be the source of further metastases. However, no recurrent alterations enriched in metastases were detected, which suggested that LUADs obtain their metastatic phenotype via additional mechanisms that were not investigated by this work, such as modification on an epigenetic and transcriptomic level in interaction with the tumor microenvironment and the immune system.

Our data might have important clinical implications, because they indicate that most of the SCNAs and driver mutations can be detected in the primary tumor. However, the observation that established metastases spread to other sites in a cascading manner argues in favor of aggressive local treatment of all metastatic sites in patients with oligometastatic disease.

1. Introduction

1.1 Importance of this work

In the last years, genomic intratumor heterogeneity (ITH) in cancer has gained increasing attention, since its presence might challenge therapy decision making based on single biopsies. If a tumor consists of multiple clones that depend on different biological pathways because of individual mutations, resistance against (targeted) therapies might arise early because the treatment is not successful in eliminating all clones within a tumor. Additionally, if a single biopsy does not represent all of the existing clones and their mutations because of ITH, one or the other clone could be easily missed, because, despite being present in the tumor, it might not be present in a specific biopsy sample. This concept of genomic ITH and its impact on personalized medicine has recently been elaborated on by a landmark paper in renal cell carcinomas, which challenges therapy decision-making based on single biopsies.[1] Additionally, most cancer studies focus on genomic ITH in primary tumors only. However, it is its transformation to a metastatic disease that makes cancer so deadly, but studies of matched primary tumors and metastases are sparse.

This is especially true for the data situation in non-small cell lung cancer (NSCLC). Only a small number of studies have investigated ITH between primary tumors and metastatic sites.[2–4] Most of these studies have concentrated on exploring ITH of epidermal growth factor receptor (EGFR) mutations (reviewed in ref[5]) because they have direct clinical consequences. This is, because of the two most common mutated genes in NSCLCs, *KRAS* and *EGFR*, only mutations in *EGFR* are currently druggable. A recent NSCLC study makes aware that the extent to which genomic ITH is detected is attributable to differences in sequencing depth and to the presence of mixed (tumor and non-tumor) cell populations.[6] These limitations have encouraged the design of this study, which aims technologically to establish a method for purification of tumor cell nuclei that are intermixed with normal cells. This study also aims biologically to infer the clonal relationship of primary lung adenocarcinomas (LUADs) and their metastases by investigating ITH of genomic alterations, such as copy number aberrations and cancer gene mutations, which allow quantifying the extent of chromosomal instability (CIN) over the course of metastatic disease.

We believe that this study of temporally separated, matched lesions is better suited than one of unmatched material to studying CIN, because CIN describes a process of change, which has a temporal dimension by definition.

1.2 Lung cancer

Lung cancer has the highest incidence and mortality rates of all cancers worldwide. It has been estimated that 1.8 million new cases were diagnosed and 1.6 million patients died from this disease in 2012 (Figure 1). Despite advances in earlier detection and improved treatment options, lung cancer is still the most frequently diagnosed cancer among males and the third most commonly diagnosed among females (after cancers of the breast and colon).[7] Improving survival therefore remains a major challenge in lung cancer oncology, where 5-year survival rates are 17.7% across all stages and only 1% in metastatic disease (Table 1).[8] One major reason for the high mortality rate of lung cancer is that the majority (70%) of patients are diagnosed with advanced stages or metastatic disease. Additionally, many patients with localized disease are often unsuitable for curative surgical treatments due to concomitant medical illness.[9]

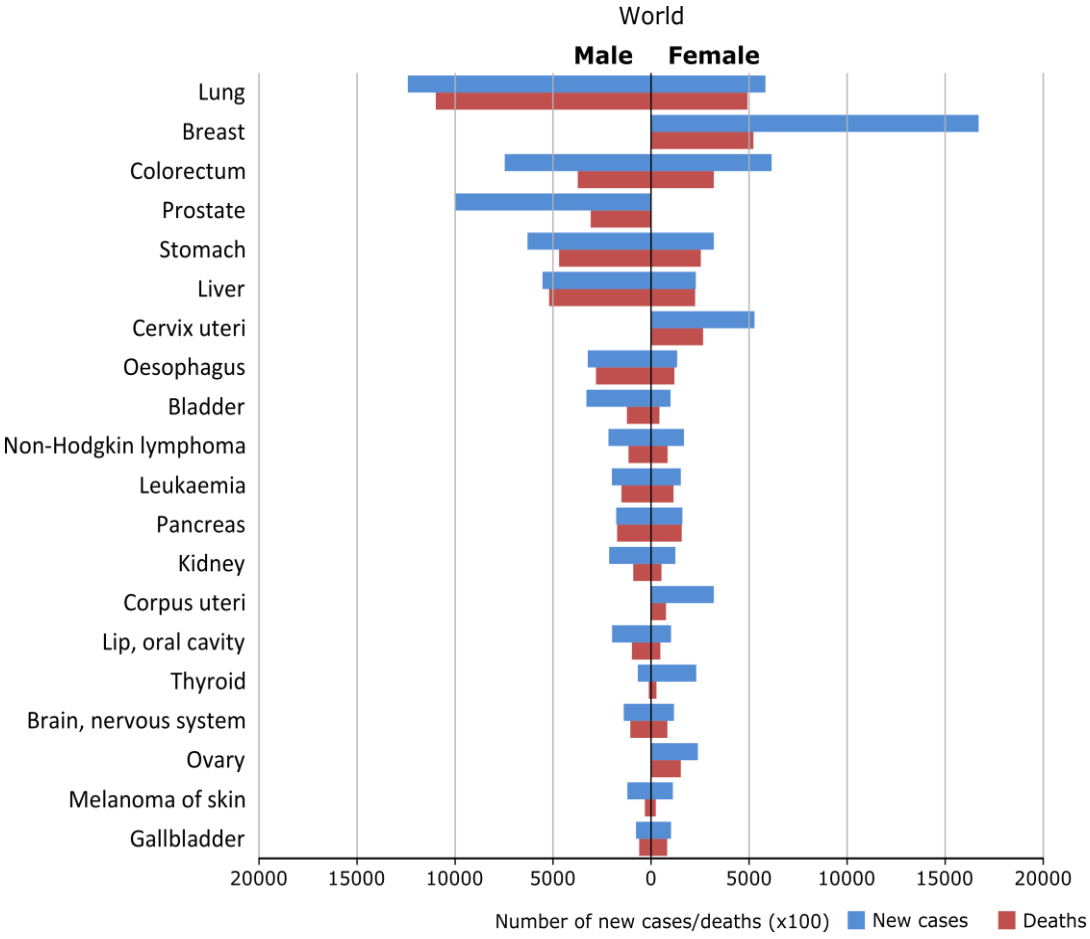


Figure 1 | Worldwide incidence and mortality of the top 20 cancers in 2012. Data retrieved from Globocan 2012 (http://globocan.iarc.fr/).[7] Lung cancer has the highest overall incidence and mortality rates, but in women alone, breast cancer is more frequent than lung cancer.

Table 1 | 5-year survival rates of NSCLC according to stage. These numbers are based on data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database, based on people who were diagnosed with NSCLC between 1998 and 2000. For staging of NSCLC see Figure S1.

Stage	5-year survival rate
IA	49%
IB	45%
IIA	30%
IIB	31%
IIIA	14%
IIIB	5%
IV	1%

1.2.1 Etiology and risk factors

Tobacco smoking

Cigarette smoking is considered by far the most important etiological risk factor for lung cancer.[10] The link between tobacco smoking and lung cancer was established by two landmark papers in 1950.[11,12] Both quantity and duration of smoking are known to correlate with an increased life-time risk for developing lung cancer.

Awareness of the impact of tobacco smoke and improved screening methods have resulted in a decline in the incidence rate since the mid-1980s in men and since the mid-2000s in women in the USA. This gap in time between genders is attributed to historical differences in the timing of uptake and cessation. Death rates began declining in the USA with a delay of some years and decreased from 2007 to 2011 by 2.9% and 1.9% per year in men and women, respectively.[13] In contrast, in other parts of the world, lung cancer-related incidences and deaths are on the rise.[14] Even though pipe and cigar smoking is considered less dangerous by the public, the risk of developing lung cancer is comparable to that of light cigarette smoking.[15,16]

Never smokers

The term "never smoker" defines a person that has smoked less than 100 cigarettes in his or her lifetime.[17] This includes lifetime non-smokers. Globally, it has been estimated that 25% of patients diagnosed with lung cancer are never smokers[18], a value that underlies gender and geographic variations. Overall, 85% of lung cancers in men, but only half of all lung cancers in women, are related to smoking. Additionally, 60-80% of women that develop lung cancer in East and South Asia

are never smokers, compared to 15-20% in Europe and North America.[19] Yet lung cancer in never smokers alone ranks in seventh place of leading causes of cancer death in the world and is responsible for more deaths every year than cervical, pancreatic, or prostate cancer.[20]

Other risk factors

Epidemiological studies have established an association between several environmental, genetic, hormonal and viral factors and the risk of developing lung cancer.[21–31] Given the dominant role of tobacco smoking in lung cancer tumorigenesis, the association of passive smoking, also referred to as environmental tobacco smoke (ETS), with lung cancer risk has been widely studied. Environmental tobacco smoke comprises sidestream (released by cigarettes or other smoking devices) and mainstream (exhaled by smokers) smoke; it contains the same carcinogenic toxins – polycyclic aromatic hydrocarbons (PAHs), benzo[a]pyrene (BAP), and nicotine-derived nitrosamine ketones (NNK) – that are inhaled by smokers, although in lower concentrations.[32] Nevertheless, metabolites of NNK have been detected in the urine of non-smokers exposed to ETS.[33] Therefore, non-smokers exposed to ETS, for instance due to living with a smoker or working in a smoking environment, have a 20-25% higher risk of developing lung cancer in their lifetimes. [32,34,35] Moreover, children raised in a smoker's household have a threefold increased risk of developing lung cancer during adulthood.[36]

Other risk factors, such as asbestos, ionizing radiation, air pollution, exposure to cooking fumes, and diet, all of which are unrelated to smoking, have been implicated.[37–40] However, only weak associations have been established, resulting in the fact that major causes of lung cancer in never smokers have yet to be identified.

Overall, the lifetime risk of developing lung cancer is 10-20 times higher in smokers than in never smokers and cigarette smoking accounts for 85-90% of all lung cancers in the Western world.[41] Consequently, in terms of prevention, the best strategy is to prevent the sequence of events that eventually lead to lung cancer. Therefore, efforts to prevent smoking and the initiation of smoking in children and teenagers in particular are important.[17]

1.3 Non-small cell lung cancer

Non-small cell lung cancer accounts for 80-85% of all lung cancer cases.[8,42] Based on its histology, NSCLC is divided into three major pathologic subtypes: LUAD, squamous cell carcinoma (LUSC), and large cell carcinoma. The most frequent subtype of NSCLC is LUAD, accounting for 40%, while LUSC and large cell carcinoma account for 25% and 15%, respectively.[43] LUAD arises from distal airways, while LUSC is connected to the more proximal epithelium.[44]

1.3.1 Connection to smoking

All subtypes of NSCLC are associated with tobacco smoking, yet the strongest connection exists between smoking and LUSC.[45] Conversely, LUAD is the most common form of lung cancer in never smokers.[46] In addition to the global variation of the histopathological subtypes, there is a global trend of increasing rates of LUAD that co-occur with a decline in LUSC. This dominance of LUAD over LUSC has been hypothesized to be connected to the emergence of filtered cigarettes with low tar- and nicotine content that began to be produced once the detrimental effects of smoking were realized.[47] This change in cigarette design has presumably led smokers to inhale deeper to reach the same nicotine saturation as before.[48] This could then result in a more peripheral distribution of tobacco smoke, prompting a shift from central LUSC to peripheral LUAD.[47] Moreover, the decrease of PAHs, known LUSC inducers, has led to a concurrent increase in NKKs, which have been shown to promote LUAD.[17] Together, this has resulted in a shift of the histopathological subtype rather than a decline in lung cancer rates.

1.3.2 Cells of origin

The airway epithelium consists of various cell types that vary in both their composition and proportion along the proximo-distal axis of the lung. The alveolar epithelium consists of type I and type II alveolar epithelial cells (AEC1 and AEC2, Figure 2).[49] Although functional evidence is still lacking, different cell types are hypothesized to be the origin of the different histological subtypes of lung cancer. For instance, basal cells in the proximal airway are shown to serve as the cell of origin for LUSC.[50] The evidence for LUAD is not that clear. In V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (KRAS)-mutant mouse models, AEC2, clara cells, and putative bronchoalveolar stem cells are able to yield hyperplasia, yet only AEC2s have the capability to advance to malignant adenocarcinomas.[51,52] This suggests that AEC2s might be the initiating cells for LUAD.

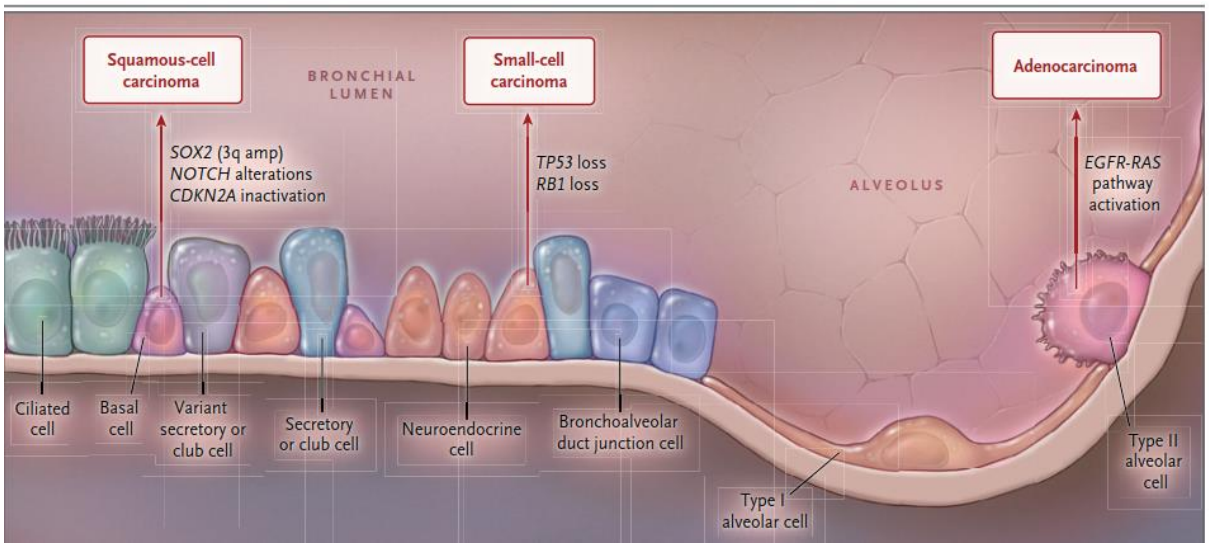


Figure 2 | Cells of origin and characteristic alterations according to histologic subtype. "Reproduced with permission from ref[50], Copyright Massachusetts Medical Society.

1.3.3 Molecular biology

NSCLC is a heterogeneous disease, and despite the fact that some molecular alterations are shared among the histological subtypes, they markedly differ in the genes and therefore pathways affected by structural aberrations and mutations.

Genomic landscape

Whole-exome sequencing of smoking-related lung cancer has revealed a subtype independent mean somatic mutation rate of 8-10 mutations per megabase (10^6 base pairs).[53–55] The mutational burden of the LUAD subtype of never smokers, however, is much lower (0.6-0.9 mutations per megabase).[56] Additionally, tumors from smokers are characterized by a high frequency of cytosine-adenine (C:G/A:T) transversion mutations (purine to pyrimidine or vice versa), a feature that is directly associated with tobacco exposure. On the other hand, cytosine-thymine (C:G/T:A) transition mutations (purine to purine or pyrimidine to pyrimidine) are predominant in lung cancers of never smokers.[56] Further complexity ranges from the large number of somatic copy number aberrations (SCNAs) and gene rearrangements. This makes it extremely difficult to discover new genetic drivers among the background of a high burden of genetic alterations. It is estimated that approximately 3,000 samples are needed to be able to detect genes that are significantly mutated in > 2% of lung cancers in smokers.[57]

Comparison of LUAD and LUSC

Genomic analyses of tumors from NSCLC patients have revealed alterations that are common across different histological subtypes, yet the majority of genomic alterations are specific to one or the other subtype. A recent study, the most comprehensive of its kind, compares somatic genome alterations in LUAD and LUSC.[58] It finds that only six genes – *TP53*, *RB1*, *ARID1A*, *CDKN2A*, *PIK3CA*, and *NF1* – are significantly mutated in both tumor types, and of these, *TP53*, *CDKN2A*, and *PIK3CA* are more frequently mutated in LUSC (Figure 3a). On the level of SCNAs, only 11 focal amplifications are altered in both tumor types (Figure 3b), including amplifications of *TERT*, *MDM2*, *MYC*, *CCND1*, *EGFR* and *ERBB2*. Similarly, only 13 genes are focally deleted in both subtypes, with *CDKN2A* being the most significant one (Figure 3c). Interestingly, a comparison of significantly mutated genes in LUAD and LUSC compared to 19 other tumor types from the TCGA[59] database has revealed that both NSCLC subtypes have a greater overlap with other tumor types than with each other.[58] This indicates that the pathways and biological processes that are altered differ tremendously between these two subtypes.

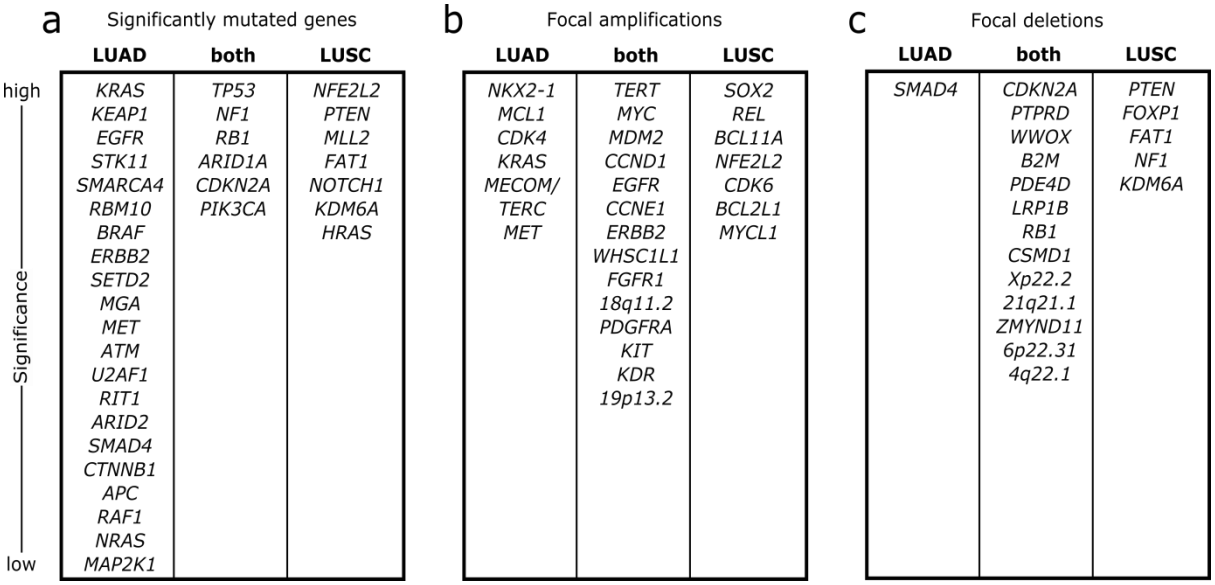


Figure 3 | Common and uncommon somatic alterations in LUAD and LUSC. Only genes that have been previously connected to lung cancer are included. Data were retrieved from ref[58].

Pathway alterations in LUAD

A recent study that integrates data of whole-exome and transcriptome sequencing presents the most comprehensive map of alterations in LUAD. It reveals that in 76% of cases, genes of the RTK/RAS/RAF pathway are either mutated or otherwise altered. Other pathways that are affected are the PI(3)K-mTOR pathway (25%) and the p53 pathway (63%). Further, alterations of cell cycle regulators (64%),

oxidative stress response pathways (22%), and mutations in various chromatin modulators and RNA splicing factors (49%) are frequently affected (Figure 4).[53] Mutations in the oncogenes *EGFR* (11%) and *KRAS* (32%) result in a constitutive activation of the RTK/RAS/RAF pathway[60,61] and are mutually exclusive with each other and with alterations in three other oncogenes of this pathway, namely *ALK*, *BRAF* and *ERBB2*.[62,63]

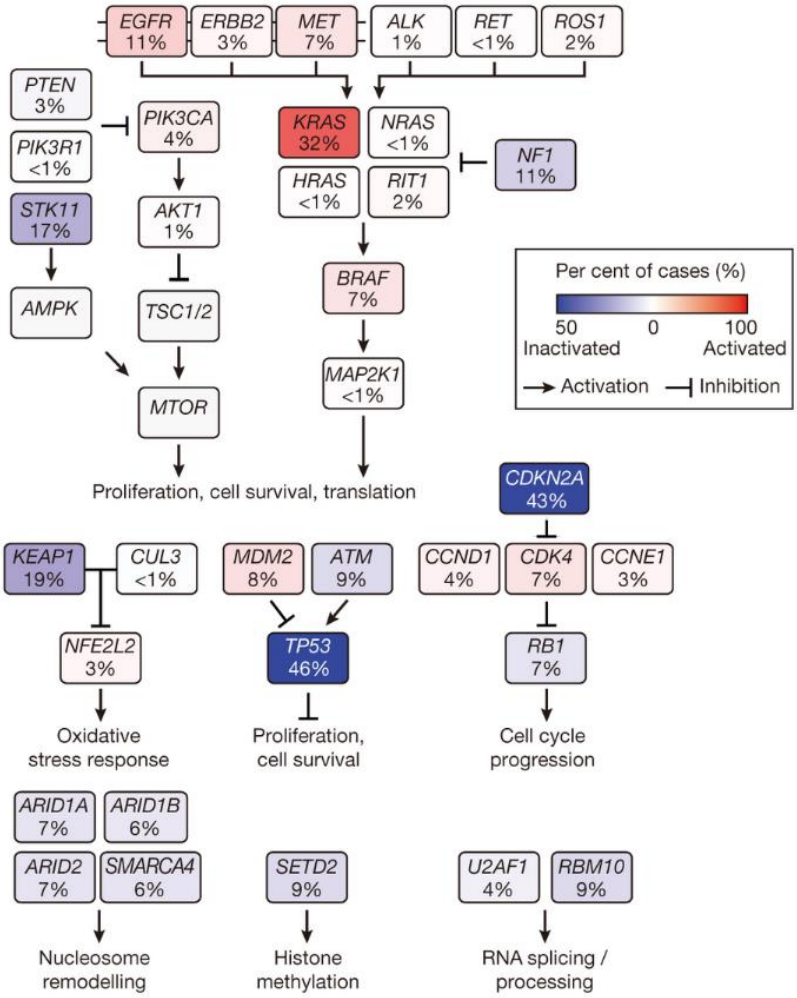


Figure 4 | Pathway alterations in LUAD. Somatic alterations in key pathways that are dysregulated in LUAD are shown. Figure 4a from ref[53]: "Republished under the Creative Commons license (Attribution-Noncommercial) for non-commerical/educational purposes from Nature Publishing Group.

KRAS

KRAS is in fact the most frequently mutated gene in LUAD and is more prevalent in smokers than never smokers.[64] Like *HRAS* and *NRAS*, it encodes for a Guanosine-5'-triphosphatase (GTPase) that functions as a molecular switch regulating cell proliferation and survival. KRAS is normally tightly regulated between an inactive guanosine diphosphate (GDP)-bound state and an active GTP-bound conformation.[65] The majority of mutations occur at codons 12, 13, and 61, so-called "hotspots". Mutations at these conserved sites favor GTP binding and result in a constitutive activation of

mitogenic signaling.[66–68] The most common KRAS mutation in LUAD is the G12C mutation that results from a C:G/A:T transversion mutation and is associated with exposure to tobacco smoke.[69,70] Despite this high frequency of mutations, KRAS is currently not druggable, and strategies, such as blocking C-terminal farnesylation, a post-translational modification that is required for protein activity, have failed in clinical trials to provide a statistically significant survival benefit.[71,72] Other approaches, such as targeting the downstream effectors of KRAS, RAF and MEK, have been unsuccessful as well.

EGFR

The epidermal growth factor receptor (EGFR) is one of the four members of cell-surface receptor tyrosine kinases (RTK) that together constitute the epidermal growth factor (EGF) family. These regulate many developmental and metabolic processes. Like all members of this family, EGFR consists of an extracellular ligand-binding domain, a single hydrophobic transmembrane domain and an intracellular tyrosine kinase (TK) domain. Receptor activation is mediated upon ligand binding, which induces receptor dimerization and subsequent activation of the intrinsic TK.[73] In cancer, TK activity can be activated by *EGFR* mutations, gene copy number amplification, and EGFR overexpression, all of which result in the promotion of cell survival, proliferation, invasion and metastasis. EGFR is overexpressed in many solid cancers, including stomach and colon cancers and in about 50-60% of NSCLCs.[74–76] It is higher in LUSC than in LUAD and correlates with poor prognosis.[77]

In contrast to *KRAS*, *EGFR* mutations are more frequent in LUAD in women, never smokers, and Asians, with widely varying frequencies across these populations.[78–81] This is because the proportion of patients who are lifetime never smokers is higher among females and in Asian ethnicities.[82] All mutations that are known to lead to a ligand-independent, constitutive TK activation are found in the first four exons (18-21) of the TK domain (Figure 5).[83,84] The predominant single-point mutation is the EGFR L858R mutation in exon 21 and accounts for 43% of all EGFR TK mutations. Together with exon 19 deletion mutations, EGFR L858R constitutes more than 90% of all EGFR activating mutations. Although the exact mechanisms are not fully understood, crystallographic analyses have suggested that these mutations confer a 50-fold increased TK activity by destabilizing the autoinhibited conformation that is normally formed in the absence of ligand binding in wild-type EGFR.[84–88] Mutations in *EGFR* have important clinical implications, since they can be exploited by targeted therapies.

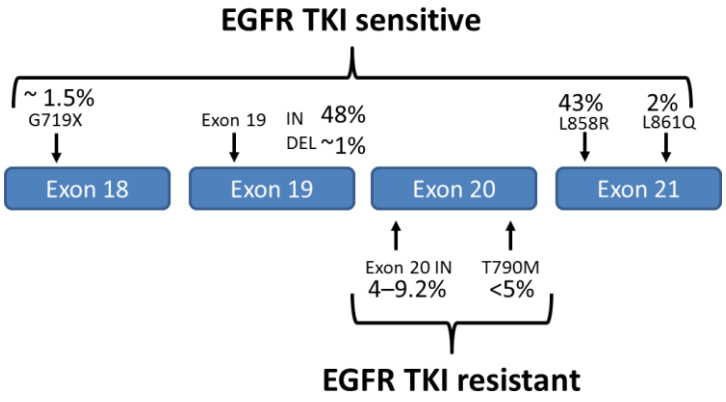


Figure 5 | Schematic of *EGFR* mutations. Mutations in the four exons confer either resistance (exon 20) or sensitivity (exons 18,19, and 21) to small EGFR TKIs. The majority of mutations occur in exons 19 and 21. *Note: the resistance mutation* EGFR T790M *has been implicated in a small proportion of untreated EGFR-mutated tumors.*[89] Adapted from https://www.mycancergenome.org/content/disease/lung-cancer/egfr/

1.4 Therapy of non-small cell lung cancer

Surgery remains the standard treatment for stage I and II NSCLC, whereas chemotherapy is still the standard therapy for patients with advanced lung cancer. Despite the fact that histological heterogeneity in lung cancer was acknowledged as early as the 1950s,[90] subtype-independent treatment decisions were continued until 2004. It was then recognized that specific mutations in *EGFR* are important predictive determinants for mediating sensitivity to gefitinib.[75,91] Gefitinib, an EGFR small tyrosine kinase inhibitor (TKI), was initially approved in Japan and the USA for treatment of NSCLC as a result of the observation that EGFR is more abundantly expressed in tumor cells than in adjacent normal lung tissue.[92] However, a consistent correlation between EGFR expression and response to gefitinib could not be detected.[93] In 2008, it was recognized that tumor histology is an important determinant for response to chemotherapy. The combination of the two chemotherapeutics cisplatin and gemcitabine was found to be superior to cisplatin plus pemetrexed. However, this was true only for LUSC, because the survival among LUAD patients was better under treatment with cisplatin plus pemetrexed.[94] Since then, dramatic improvements have been achieved in the outcomes of NSCLC patients, which can be attributed mainly to improvements in understanding NSCLC as a heterogeneous disease entity and efforts to classify NSCLC into subtypes based on both genotype and histology (Figure 6).

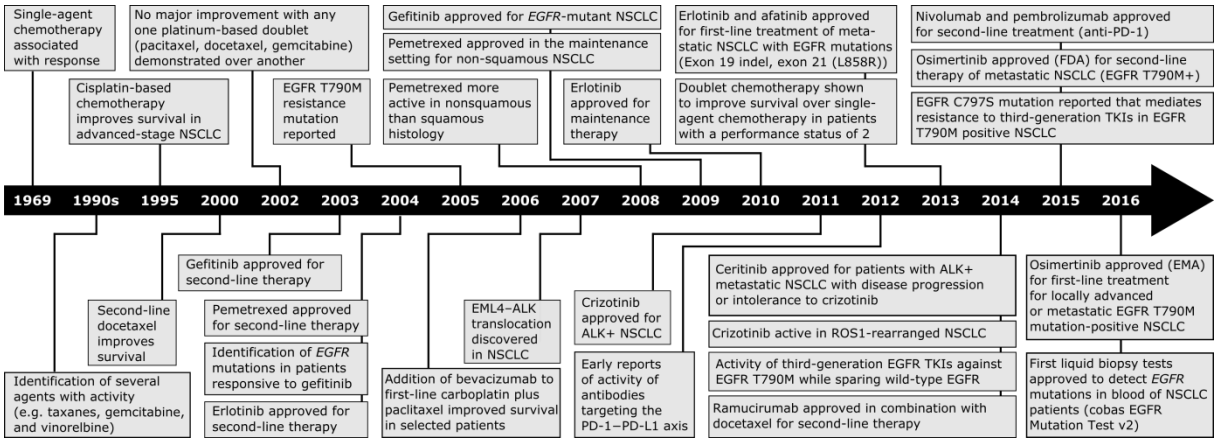


Figure 6 | Timeline depicting the historical milestones in the development of therapies for NSCLC. Until the 1990s, there was considerable pessimism about the treatment of NSCLC, and the benefit of chemotherapy for treating NSCLC was unclear. Treatment stratification by histology was not performed in these times. EGFR TKIs were initially approved as second-line therapy, as it was recognized that EGFR is overexpressed in NSCLC. In 2004, *EGFR* mutations were identified in patients who responded to first-generation EGFR TKIs, the beginning of the molecular era for personalized treatment of NSCLC. ALK-translocations were discovered in 2007, and 4 years later, the first drug was approved for the treatment of ALK-rearranged NSCLC. First reports of the activity of immune-checkpoints inhibitors were reported in 2012 and the first anti-PD-1 antibodies were approved in 2015. The EGFR T790M resistance mutation was reported as early as 2005, and after second-generation EGFR TKIs proved less effective than hoped at blocking T790M-mutant tumors, the first third-generation EGFR TKI against T790M was approved at the end of 2015 for second-line treatment in the USA, followed by EU approval for first-line treatment in early 2016. Recently, a second resistance mutation, EGFR C797S, was discovered, which confers resistance to third-generation EGFR TKIs. The first test to detect *EGFR* mutations in the blood of NSCLC patients was approved in June 2016 with the hope that consistent monitoring results in a benefit for metastatic *EGFR* mutation-positive NSCLC patients. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Clinical Oncology, ref[95], copyright 2015, and updated.

Targeted therapy and resistance

LUAD can be considered a pioneer disease for personalized cancer medicine, because many treatment decisions today are based on the presence of predictive genomic alterations. Patients with *EGFR* mutations in exons 18, 19, and 21, for instance, are treated with small TKIs, such as gefitinib, erlotinib and afatinib. Stratifying patients according to mutations in *EGFR* helped increase response rates from 10% of unselected patients to approximately 75%.[96] The inhibition results from a higher binding affinity of these small TKIs than adenosine triphosphate (ATP) and is so effective, because *EGFR*-mutant tumors depend on their aberrant EGFR signaling for survival.[97,98]

The use of specific TKIs has also been studied for other molecular targets and impressive advances have been achieved. The small TKI crizotinib, for instance, was approved in 2011 for the treatment of anaplastic lymphoma kinase (*ALK)* rearranged NSCLCs. This was just 4 years after the fusion protein EML4-ALK – the result of a translocation between ALK and the echinoderm microtubule-associated protein 4 gene (EML4) – was identified in LUAD in 2007.[99]

Despite initial responses to these targeted therapies, resistance ultimately develops. More than 50% of patients that are treated with EGFR inhibitors, for instance, eventually acquire a resistance via the secondary EGFR mutation T790M.[100–102] It is believed that this mutation confers resistance either

through steric hindrance of EGFR TKIs or because it increases the affinity of EGFR for ATP.[103,104] Drugs are therefore necessary to overcome this resistance mutation. Just recently, osimertinib (AZD9291), a new EGFR TKI, was found to be highly active in EGFR T790M mutated lung cancers.[105] In fact, it received accelerated approval for the treatment of patients with EGFR T790M mutation-positive NSCLC in the USA, EU, and Japan just recently. However, an additional resistance mutation has already been identified. Interestingly, this mutation (EGFR C797S) is mutated at exactly the position, where osimertinib makes a covalent bond with EGFR, the cysteine residue 797.[106]

Resistance mutations are thought to be already present in few cells of the tumor before the application of a targeted therapy and because of their low abundance they cannot be detected by current diagnostic tests. Indeed, a recent study in lung cancer cell lines has estimated that the EGFR T790M resistance mutation had pre-existed in approximately 0.6% of the parental population. The application of an EGFR targeted treatment resulted in the selection and outgrowth of cells with the T790M resistance mutation.[107] Heterogeneity in a tumor due to the presence of different cell clones might be indeed a challenge for overcoming resistance in cancer therapies if clones with alterations that confer resistance to specific therapies exist prior to treatment.

1.5 Intratumor heterogeneity

Heterogeneity can be observed in many different ways both in tumor and normal tissue. This includes heterogeneity of phenotypes and cell morphology, gene expression, metabolism, motility, and angiogenic, immunogenic, and metastatic potential in tumors.[108] The first observations of heterogeneity within cancers occurred as early as the mid-19th century and were of a phenotypic nature. In 1855, German pathologist Rudolf Virchow proposed that cancers arise from cells in mature tissues. He observed that cells in cancers differ in their cellular morphology. Despite now being considered the father of modern pathology, Virchow's work was largely ignored in the 19th century and revived only toward the early and mid-20th century, when studies documented functional and genetic heterogeneity of tumors. It was shown that different cell populations with distinct cytogenetic profiles differ in their tumorigenicity in animal models.[109] Further studies revealed that distinct subpopulations of cancer cells within tumors vary in their abilities to metastasize and in their resistance to treatment.[110–112]

Although tumor development is regarded as a process of Darwinian evolution, with the understanding that selection forces act on populations of cells that differ in their heritable traits, the relative contributions of heritable (genomic) and non-heritable mechanisms are still not fully understood. Two mutually exclusive models have emerged to explain ITH: (i) the cancer stem cell

(CSC) model and (ii) the stochastic model (Figure 7). The CSC model assumes a hierarchical organization in which only a minor fraction of cells in a tumor is responsible for its progression and maintenance in a non-heritable fashion. This is because the ability to self-renew and differentiate into other non-stem cells is unique to CSCs. Therefore, CSCs are biologically distinct in that they have different intrinsic features than the bulk of other non-stem cells. This can be exploited to enrich tumor-initiating activity by sorting CSCs based on these intrinsic characteristics.[113] Conversely, the stochastic model posits that a tumor is biologically homogeneous and random intrinsic or extrinsic factors result in the heterogeneous behaviors of the cancer cells, including tumor initiation capacity. Therefore, cell behavior cannot be predicted by intrinsic characteristics, and tumor-initiating activity cannot be enriched.[114] The essential difference between these models is that phenotypic and genetic heterogeneity are irrelevant for tumor progression under the CSC model, as long as they do not affect the CSCs, because selection only acts on heritable phenotypes of CSCs. In contrast, genetic heterogeneity involves variation, and variation is necessary for selection under the stochastic model.

The existence of CSCs has been demonstrated in hematopoietic malignancies[115] and has been suggested with increasing evidence also in solid tumors.[116] Still, it remains a subject of debate for many reports on solid tumors. This is because the CSC model requires, by definition, experimental evidence, such as tumor initiation at limited dilution. [117] Most of the data supporting the CSC model originate from mouse transplantation assays, which underestimate the number of cells in human tumors with tumorigenic potential due to variations in xenotransplantation conditions.[118] In addition, in situations where many cells in a tumor have the ability to initiate tumors, which are then defined as CSCs, enriching for them becomes meaningless.

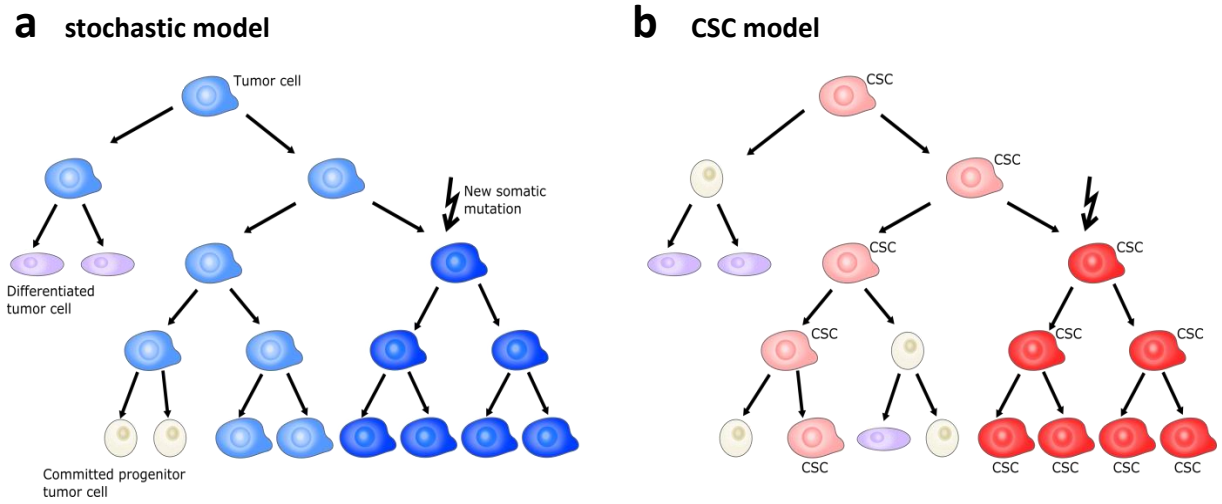


Figure 7 | Two models of tumor growth. The stochastic model (a) results in tumor heterogeneity because all cells can self-renew and differentiate, whereas in the CSC model (b) heterogeneity can only arise from CSCs. In both models, mutations can further contribute to ITH. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer, ref[117], copyright 2013.

1.6 Tumor evolution

Tumor evolution is loosely related to the concepts of evolutionary biology. The traditional so called "Darwinian" model of tumor evolution posits that selection plays a critical role in tumor development and that this results in multiple clonal expansion rounds, each of which is characterized by the acquisition of an additional mutational event that leads to a so-called "selective sweep" (Figure 8a).[119] Selection favors well-adapted clones over less well-adapted clones and thus allows them to produce more offspring. Over time, this results in clonal outgrowth and can ultimately lead to a situation where one clone dominates the entire tumor. Yet, this process can only occur if there is enough time for the selected clone to sweep through the tumor, i.e. enough time needs to pass for the next driver mutation to happen. Consequently, mutations and selection are inextricably linked with each other, but whereas mutations are considered to happen randomly, selection is a non-random process.[120] This depiction is based macroscopically on a single time point and therefore limited in reflecting tumor dynamics over time. Chromosomal instability (CIN) with ongoing alterations and mutations is likely to alter the selective pressures that are experienced by individual populations. Therefore, tumor evolution in the presence of CIN is probably non-linear on a microscopic scale, resulting in genetic heterogeneity due to coexisting tumor cell populations (Figure 8b).

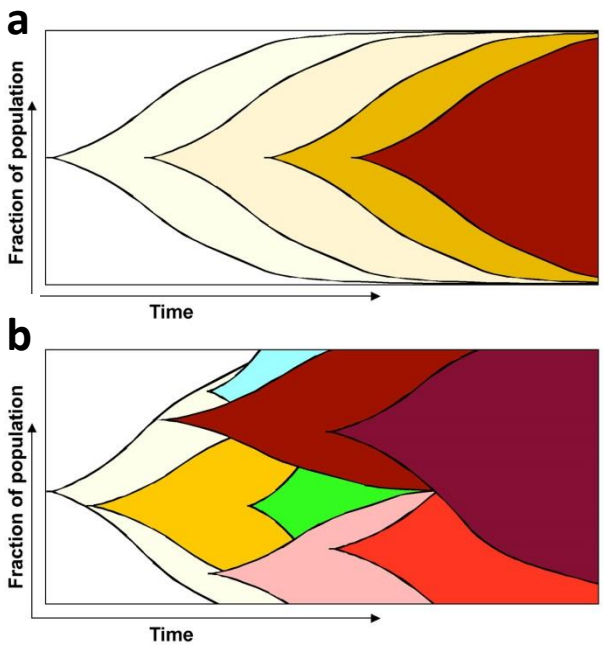


Figure 8 | Schematic view of monoclonal and multiclonal models of tumor progression (a) Linear model of tumor progression where multiple clonal expansions driven by mutations in oncogenes and tumor suppressor genes eventually result in one dominant clone due to concomitant selection of the newly expanded clone. (b) Multiclonal model of tumor progression. Despite the fact that tumors are the product of one cell that initially experienced mutations, inherent chromosomal instability can allow for clonal diversity. Clone sizes can change during tumor evolution due to selection of clones with benefical mutations. This can lead to the extinction of clones, but also to the emergence of new clones over time. Note: this process of clonal diversity does not necessarily imply gradual processes, but can also result from punctuated events (see Figure 10). "Reprinted from ref[113], copyright 2010, with permission of Elsevier".

A central question in tumor evolution is how fast mutations appear that eventually lead to ITH with the presence of multiple clones within a neoplasm. Two theories have been proposed for this context and are valid only under the assumption that selection drives cancer evolution: gradualism and punctuation (Figure 10). Gradualism posits that tumor evolution derives from a constant accumulation of mutations with subsequent selective clonal outgrowth, whereas the punctuated theory proposes that tumor evolution happens in short time intervals of change followed by longer time periods of stasis.

The punctuation theory of cancer evolution would hereby be analogous to the saltation theory of evolutionary biology. Saltation theory suggests that large phenotypic changes from one generation to the next are the result of sudden large-scale mutations, so-called "macromutations".[121] In other words, the genetic events themselves and not the phenotypic changes they cause are punctuated. This should not be confused with the evolutionary theory of punctuated equilibrium, which proposes that large phenotypic change results from rapid, but gradual genetic evolution in an isolated population (Figure 9a).[122] Both saltation theory and punctuated equilibrium theory are in fact in concordance with Darwin's theory of evolution by natural selection. Richard Goldschmidt postulated his theory about macromutations in 1940 and was the first scientist to use the term "hopeful monsters", which are the result of single macromutational events (Figure 9b).[123]

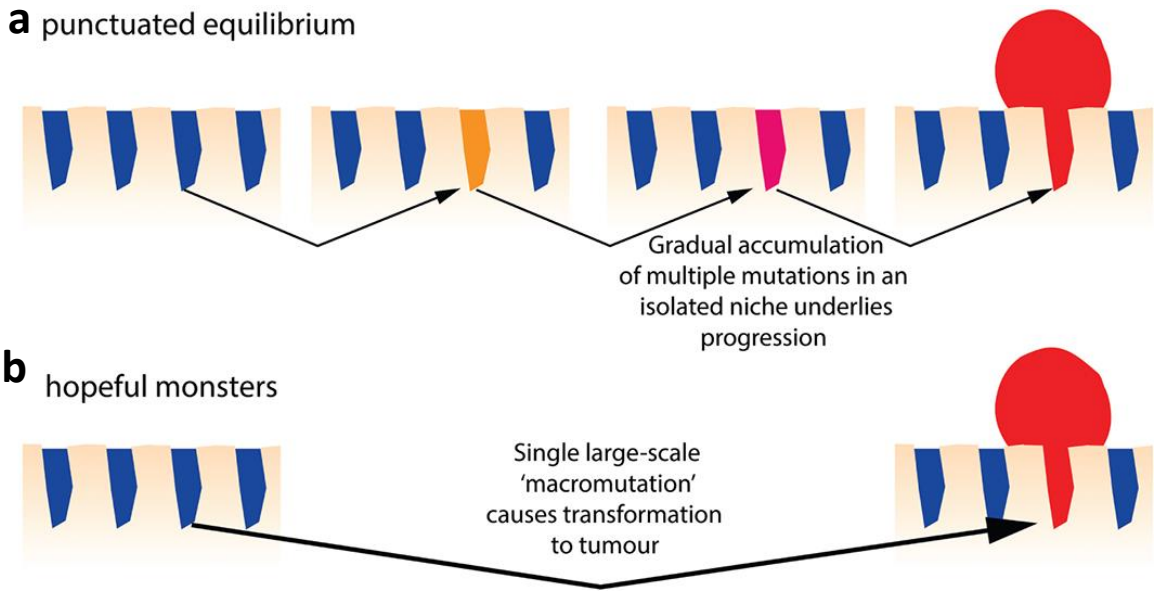


Figure 9 | Punctuated equilibrium and hopeful monsters. (a) Gradual accumulation of mutations in small spatially isolated niches (here, an intestinal crypt). Only after a sufficient number of driver mutational events (orange, pink) are acquired a clonal expansion takes place (red). Despite the fact that mutations are acquired gradually, the evolution of the tumor appears punctuated from a macroscopic perspective. (b) A hopeful monster with a large change in phenotype is generated in a single cell division. Therefore, tumor formation is the result of a single catastrophic event. Reprinted by permission from John Wiley and Sons: Journal of Pathology, ref[122], copyright 2016.

Recently, macromutations, punctuated changes in genotypes, have been assumed to occur during cancer evolution. Chromothripsis, for instance, describes the shattering and reassembly of focal regions deriving from one catastrophic event, and has been reported in many cancer types.[124] Chromoplexy on the other hand refers to the interweaving of multiple larger chromosomal regions involving multiple chromosomes, and was first identified in prostate cancer.[125] Genome doubling, also referred to as "whole genome duplication (WGD)", affects approximately 40-50% of solid cancers and can be viewed as another saltation theory, because it is believed to happen as a failure of a single mitotic event.[126]. A recent study has provided insights into the evolution of SCNAs during tumorigenesis by quantifying SCNAs of single cells in breast cancer.[127] This study has assumed that changes of copy number alterations occur as punctuated events during breast cancer evolution, because, despite the fact that ITH of SCNAs was omnipresent, single cells with intermediate patterns of these SCNAs have not been detected.

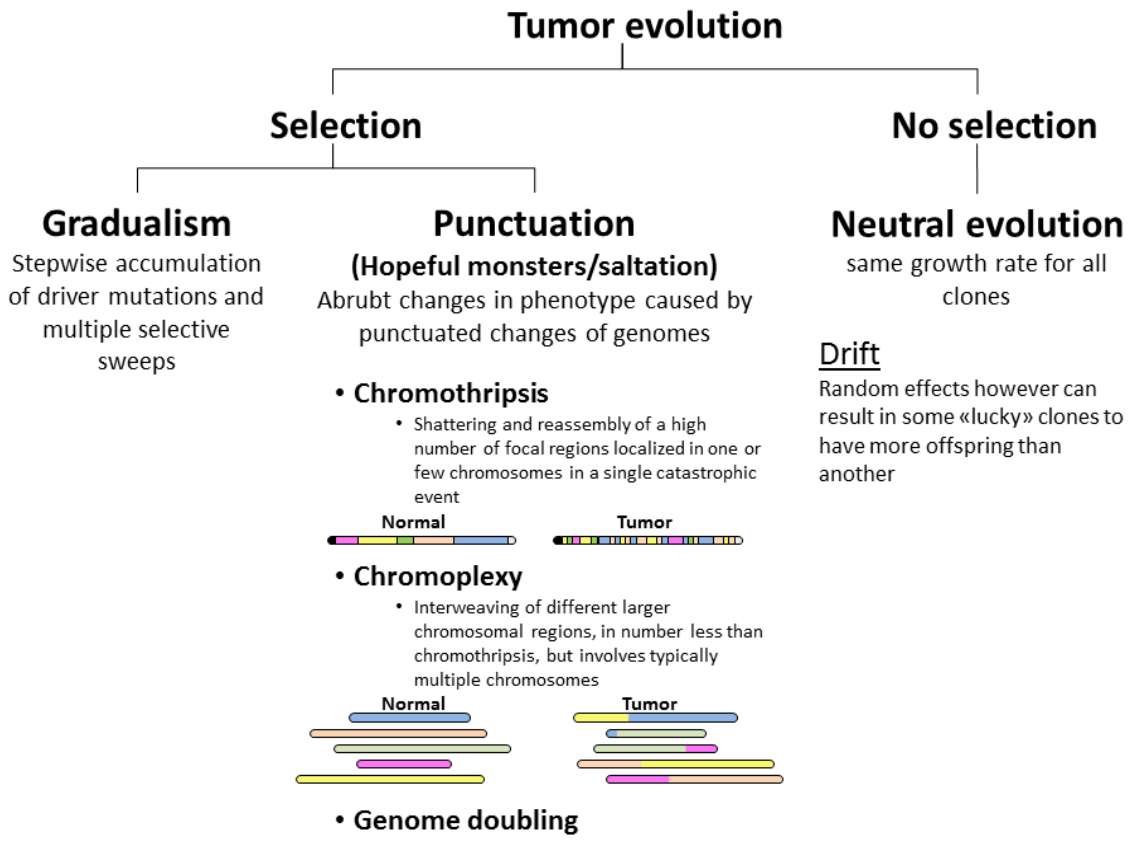


Figure 10 | Theories of tumor evolution implying the presence or absence of selection. Under selection, the aberrations that are present in a tumor at a single point in time can be the result of gradual changes or of single events that are not repeated. In contrast, neutral evolution occurs in the absence of selection and all cells grow at equal rates. Yet, in a drifting population, random effects can result in an increase or decrease of clone sizes.

Recent studies challenge, at least in part, that selection is a ubiquitous force during tumor evolution. In some tumors, the genetic diversity is orders of magnitudes higher than expected from a selection process as anticipated by the Darwinian model of tumor evolution.[128] Likewise, a study of colorectal cancer provides evidence that tumors grow as a single clonal expansion, a so-called "Big Bang", in which the clonal distribution is not determined by selection but by the point in time when the clone is generated, followed by neutral evolution.[129]

These observations have resulted in the theory of neutral evolution in cancer development. Neutral evolution is defined as "the absence of clonal selection"[130] and proposes that all cells grow at the same rate as one another. In a recent study across 14 different solid cancers[131], 30% of tumors seemed to have evolved by neutral evolution, assuming that selection might be less widespread than perhaps expected from a gradualist evolutionary perspective. In this context, it is the timing of the occurrence of a new mutation rather than strong clonal selection for this mutation that predicts subclonal dynamics of a tumor. However, it has been noted that a neutrally evolved tumor dominated by a "lucky" clone, which just happens to grow faster due to random effects, is indistinguishable from a tumor that is the result of clonal selection.[122]

1.7 Metastasis

Historically, metastases were thought to be secondary neoplasms that arise independently from primary tumors and are therefore unrelated to them. Virchow himself proposed the idea that primary tumors infect the blood with so-called "tumor poisons" that eventually elicit metastatic cell growth at distant sites, a hypothesis that excluded the involvement of cells that disseminate from the primary tumor and travel to other sites of the body.[132] We know now that metastases are the end result of a multistage process during which malignant cells spread from the primary tumor to other sites. This biological process is termed the "invasion-metastasis cascade" and comprises the following steps: cancer cell migration, local invasion, intravasation, survival in circulation, arrest at secondary sites, extravasation, micrometastasis formation, and metastatic colonization.[133] So-called "circulating tumor cells" (CTCs) can be detected in the blood of the majority of carcinoma patients, including those who might never develop metastases.[134] This obscures the fact, that the last steps of the metastatic cascade are extraordinarily inefficient (Figure 11). Overall, it has been estimated that fewer than 0.01% of all tumor cells that manage to reach the bloodstream eventually become visible as macroscopic metastases.[135] However, assuming that a tumor of 1 cm^3 in size consists of 10^9 cells, a therapy that destroys 99.9% still leaves 10^6 cells to proliferate and these provide a basis for cellular diversity.

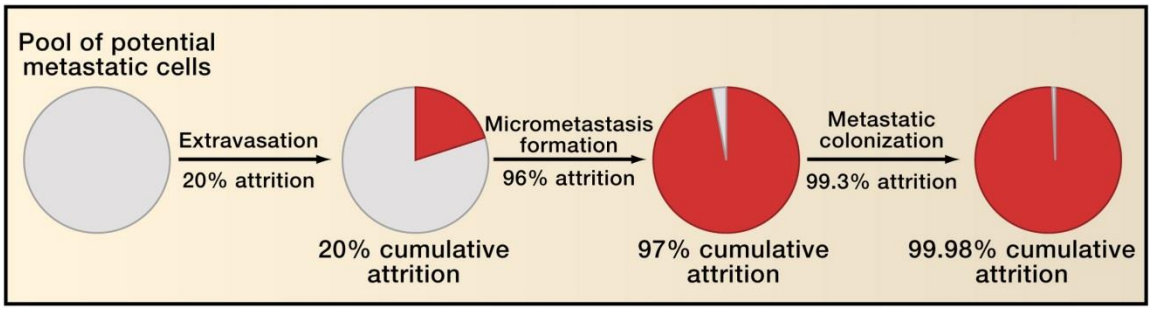


Figure 11 | Inefficiency of the invasion-metastasis cascade. "Reprinted from ref[133], copyright 2011, with permission of Elsevier".

It remains controversial whether all metastases originate from (different) clones of the primary tumor or if metastases have the potential to spread to other distant sites. Further, it is currently unknown at which moment during tumorigenesis a metastatic founder cell or clone disseminates from the primary tumor. The only time that can be measured is the relative time point of clinical detection compared to the primary tumor. Therefore, it is important to study tumors in context of their metastatic biology, because it is the metastatic disease that makes cancer so deadly. In fact, metastases account for > 90% of cancer mortality, while primary tumors can often be treated well with surgical resection or radiotherapy.[136,137]

1.7.1 Two fundamental models of metastatic progression

Metastasis is the absolute characteristic that distinguishes benign from malignant tumors.[138] However, a better understanding of the biological principles of when and how metastases arise is crucial to treat tumors that have spread to other sites of the body. Metastases can be detected either at the same time as – or at least shortly after – the primary tumor (synchronous) or several months or even years after the primary tumor (metachronous). Currently, there are no reliable methods to predict whether and when a specific tumor will relapse or seed metastases. In colorectal cancer, for instance, primary tumor characteristics such as size, stage, or grade are not significantly different between patients with synchronous or metachronous metastases.[139] And metastasis is not a self-contained disease. It involves a variety of interactions with other cell types and host factors, and its fate is subject to the interaction of stromal and immunological components of the local and systemic environment.[138,140]

Several conceptual frameworks have been proposed that aim to explain the behavior of a metastatic cancer during development. This includes the two major models of metastatic progression: (i) the linear and (ii) the parallel progression models.[141] Both assume that the metastases are clonally related to the primary tumor, that is, they derive from a common ancestral cell. For a clonal

relationship to be detectable, some mutational events such as single nucleotide variants (SNVs) or SCNAs are required to be present in both the primary tumor and the metastases.

The linear progression model

The linear progression model (Figure 12a) reflects the traditional paradigm of how metastatic evolution occurs. It is called "linear" because it assumes that the metastatic capacity of cells from the primary tumor involves the accumulation of unidirectional events.[142,143] Its central assumption is that cancer cells acquire mutations such as SNVs and SCNAs sequentially and that this eventually leads to genetically better-adapted cancer cells. This step-wise acquisition of mutations is accompanied by multiple clonal expansions that eventually lead to a clone capable of colonizing other organs, as postulated by the invasion-metastasis cascade. Because not all mutations are beneficial or provide a fitness advantage, the acquisition of mutations necessary for the metastasizing phenotype is considered to underlie a random process. Therefore, the chance of such mutations occurring increases with the number of cell divisions. According to the linear progression model, dissemination typically happens shortly before the metastasis is clinically detectable, and thus the genetic distance between the primary tumor and the metastasis is small. This means that a large number of mutations is shared between the primary tumor and the metastasis and that they were already present in their most recent common ancestor.[144]

Connected to the linear progression model is the concept of "metastatic cascades", which describes a situation in which metastases give rise to other metastases in a cascading manner (Figure 12a).[141] In many solid tumors, metastases occur 2-3 years after the primary tumor has been detected, and because the linear progression model assumes short time intervals between metastatic dissemination and detection, the theory of metastatic cascades assumes a very high turnover of cells at the metastatic sites. Consequently, the genetic distance among metastases is lower than to their primary tumor, meaning that metastases are more closely related to each other than to the primary tumor.

The parallel progression model

In comparison to the linear progression model, the parallel progression model (Figure 12b) posits that the metastatic clone disseminates early during tumorigenesis.[145] The primary tumor and the metastasis evolve in parallel, providing both enough time to accumulate substantial private mutations and alterations. The result is a high genetic divergence between the primary tumor and the metastases. In a situation with multiple metastases, this model assumes that all of the metastatic founder clones disseminate early during primary tumor growth, leading to large genetic differences among the metastases. This is in contrast to the model of metastatic cascades, which postulates a series of metastatic expansions that are seeded from other metastases in a relatively short time.

Genomic profiling of the primary tumor with the aim of providing targets for therapeutic interventions of metastatic sites is presumed to be largely ineffective under the parallel progression model, because the genetic alterations that are required for successful colonization at metastatic sites are thought to happen outside of the primary tumor.[145]

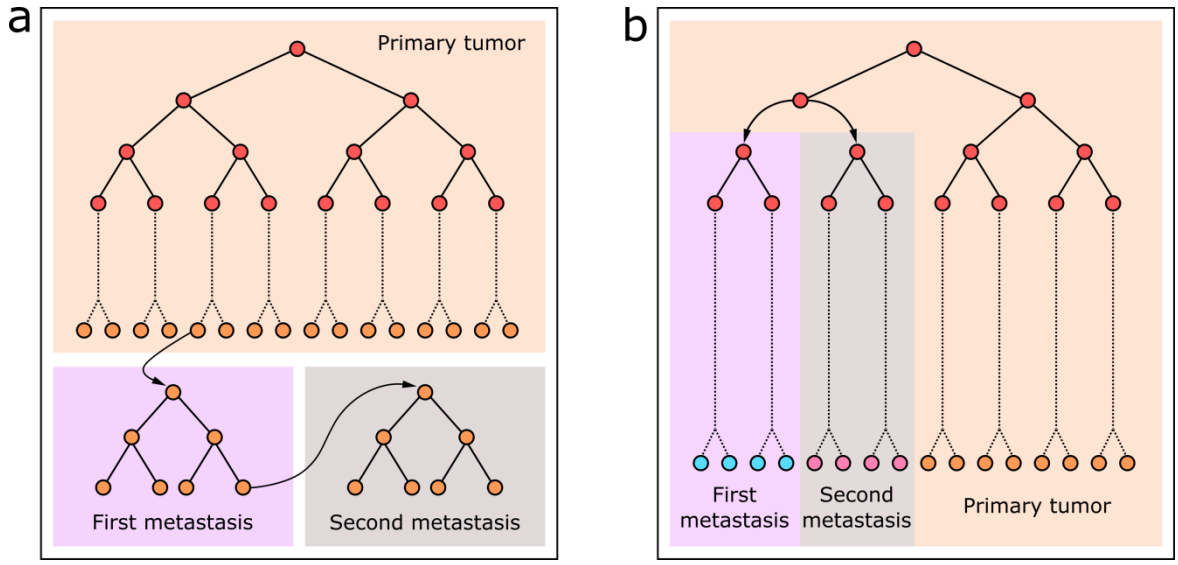


Figure 12 | Overview of models of metastatic progression in human cancer. (a) The linear progression model and metastatic cascade. (b) The parallel progression model. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Clinical Oncology, ref[141], copyright 2015.

1.8 Aneuploidy and chromosomal instability

1.8.1 Definition of ploidy

DNA-ploidy refers to the DNA quantity in cells. Historically, ploidy was estimated by DNA cytometry methods such as flow cytometry and image-based cytometry (ICM). While ICM is considered superior to flow cytometry when working with formalin-fixed tissue samples[146–148], flow cytometry remains the standard for high-throughput and sensitive multiparametric cell assessment of fresh or fresh-frozen tissues. The possibility to couple flow cytometry with cell sorting allows for the investigation of specific cells of interest, which is particularly useful in the research setting. These methods quantify the DNA content on a single cell level based on a stoichiometric DNA staining, e.g. with the DNA-intercalating dye 4´,6-diamidino-2-phenylindole (DAPI), and report differences in the quantity of DNA with a DNA histogram. Cell populations possessing a characteristic DNA content become evident as peaks on this histogram.

Tumors are classified as diploid, tetraploid or aneuploid. This has historical roots from a time when tumors were distinguished solely according to where their cell populations appeared on these DNA-

ploidy histograms, irrespective of the presence or absence of structural or numerical chromosome aberrations or translocations (Figure 14). Biologically, however, the term "diploidy" describes the state where a nonreplicating cell nucleus contains a homologous set of chromosomes that is in number an integral multiple of the normal haploid number of a given species, without the presence of chromosomal aberrations, a state also referred to as euploidy.[149] Aneuploidy therefore is the absence of euploidy (Figure 13). Yet, to be in line with the current scientific usage of the term "diploid tumor," this thesis considers a tumor population to be "diploid" if it appears as a diploid peak in a DNA-ploidy histogram (Figure 14a).

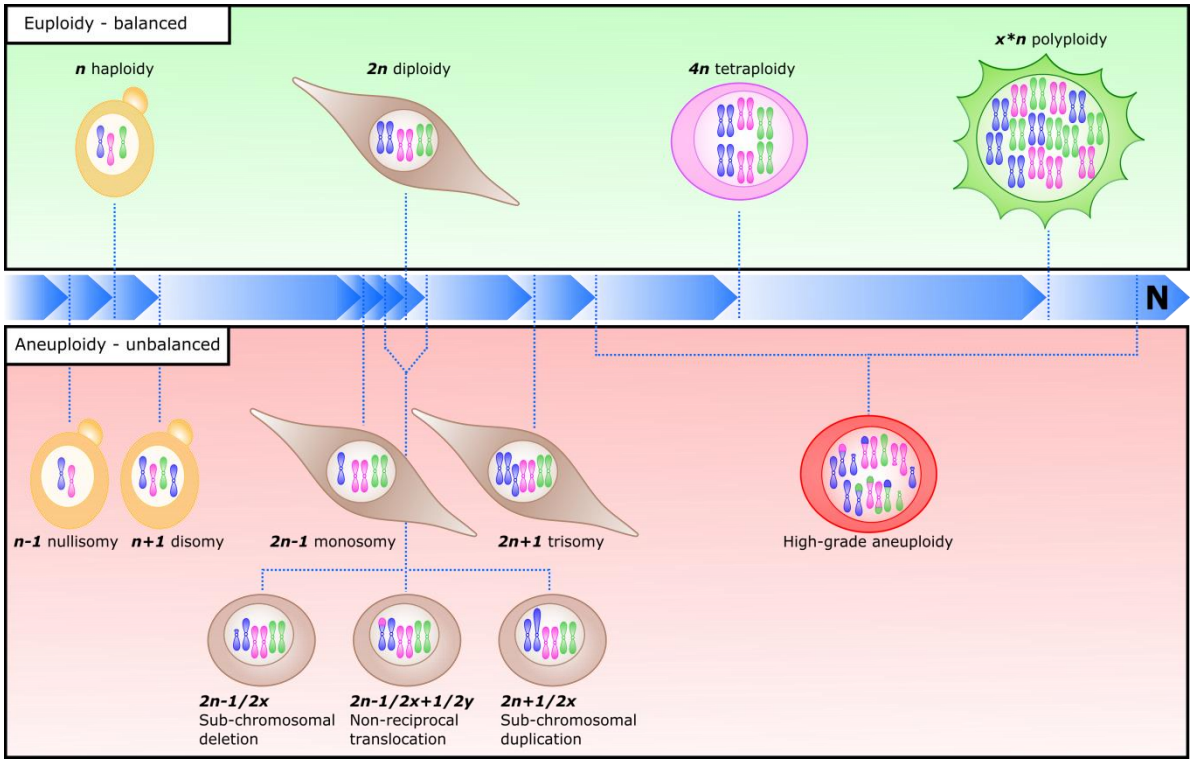


Figure 13 | Definition of euploidy and aneuploidy. Depending on the species and cell type, euploidy refers to a haploid (e.g. yeast), diploid (e.g. humans) or polyploid (e.g. plants) karyotype. They are balanced genomic states of euploid karyotypes. In contrast, aneuploidy is defined as the absence of euploidy and describes an unbalanced genomic state. Whole chromosomes can be lost (nullisomy or monosomy) or gained (disomy or trisomy). In addition, only parts of chromosomes can be lost, translocated or gained. High-grade aneuploidy describes a complex situation that often occurs in cancer, where whole chromosomes are deleted or amplified, as well as the presence of chromosomal rearrangements. N, ploidy increasing from left to right. Adapted by permission from John Wiley and Sons: EMBO Reports, ref[149], copyright 2012.

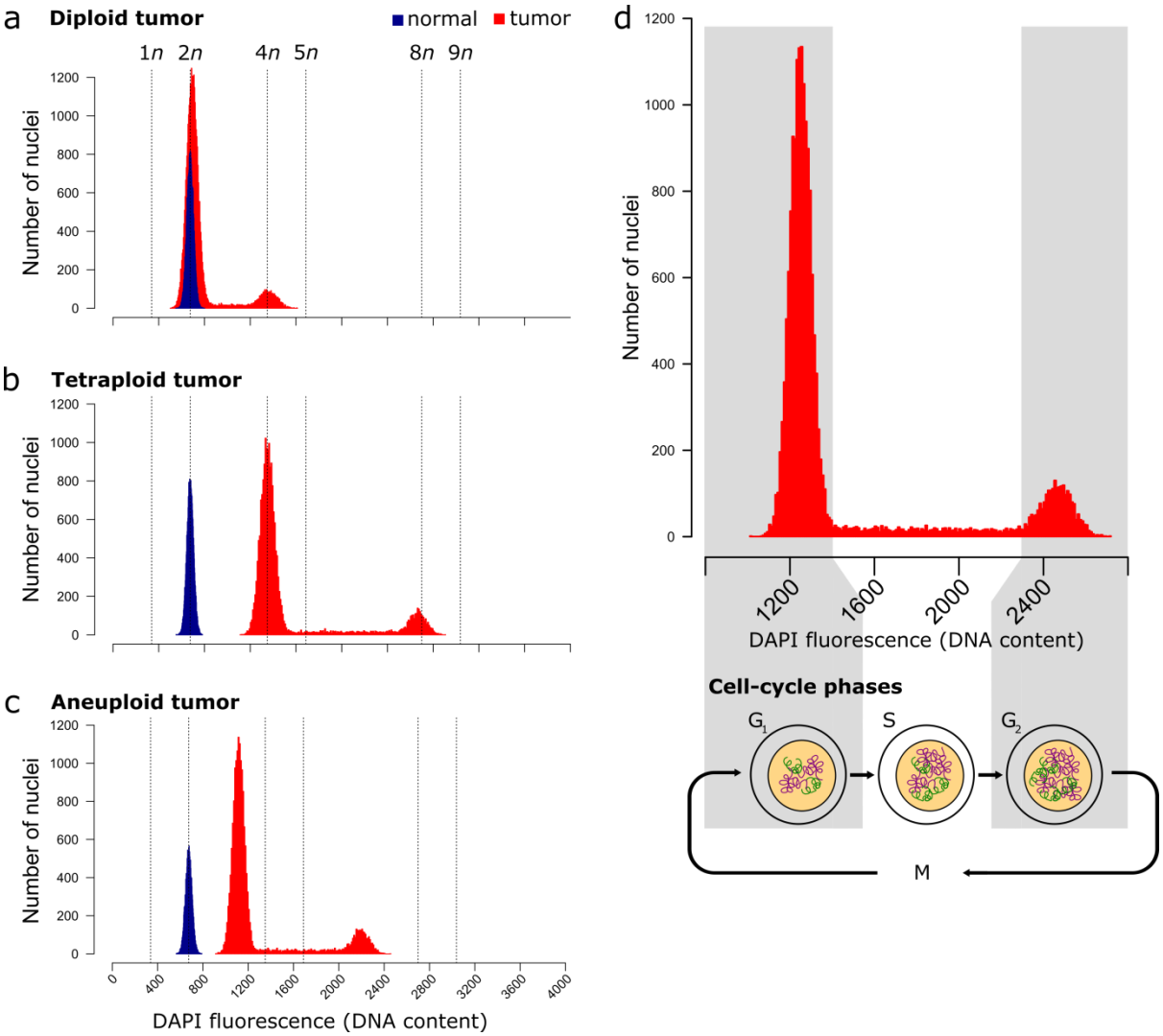


Figure 14 | DNA-ploidy status assessed by DNA cytometry. Schematic examples of DNA-ploidy histograms obtained from tumors that would be reported as diploid (a), tetraploid (b), and aneuploid (c). The x-axis in each graph represents the DNA quantity that can be measured with DNA-intercalating dyes such as DAPI. The vertical lines show the scale of DNA-ploidy values (2*n*, 4*n*, and so on), where 2*n* represents the DNA-ploidy for nuclei of diploid cells in the G_0/G_1 phases of the cell cycle. Diploid, normal cells (blue) that are usually intermixed with the tumor cells (red) such as fibroblasts, lymphocytes or epithelial cells are used as a reference for the 2*n* value. (d) An example of a DNA content histogram in order to show the connection between DNA-ploidy and the different stages of the cell-cycle. Cells in the S-phase lie along a continuum between cells in G_1-phase and cells in G_2-phase. Cells in G_2-phase have already completed DNA replication, but not mitosis (M). These DNA content histograms classify tumors as diploid, tetraploid or aneuploid solely based on their DNA content as compared to normal, diploid cells (2*n*). The presence of chromosomal alterations is only evident for an aneuploid DNA-histogram. Yet, the terms “diploid” or “tetraploid” tumor do not imply the absence of such alterations, and are used as a consequence of historical tumor-ploidy estimations by DNA cytometry. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Clinical Oncology, ref[150], copyright 2015.

1.8.2 Aneuploidy

Aneuploidy was first mentioned in 1922 due to the lack of a term that describes "a chromosome number that is not a multiple of the base chromosome number….Aneuploidy refers to hyper and hypodiploid chromosome numbers."[151,152] Aneuploidy therefore describes a situation when a cell, often tumor cell, harbors an incorrect number of chromosomes.[150] "Correct" in this context refers to the aforementioned state of euploidy. In diploid cells, tetraploidy occurs transiently in later stages of the mitotic cycle (G_2- and M-phase), at a moment when the entire genome has been replicated but before the cell has divided. In humans with a diploid genome, persistent tetraploidy is embryonically lethal.[153] However, polyploidy (integer multiples of 2*n*) occurs in a few cell types naturally, such as megakaryocytes, hepatocytes, and placental trophoblasts.[150]

Normally, cellular DNA content is tightly controlled by the cell-cycle associated apparatus to prevent genomic aberrations, including polyploidy and aneuploidy.[154] Aneuploidy eventually results from chromosome missegregation during mitosis. It has been proposed since the 1990s that aneuploidy leads to a destabilization of the genome, which results in an imbalance of the genes required for mitosis. The dysregulations of two surveillance mechanisms are important in this context. First, deficits in the spindle assembly checkpoint (SAC), the main regulator of chromosome segregation during mitosis, are known to result in numerical aneuploidy.[155] This mitotic checkpoint tightly regulates the correct separation of sister chromatids during mitosis. Any defects caused by reduction or overexpression of proteins that are implicated in this regulation can lead to numerical aneuploidy. Second, inappropriate connections between kinetochores and spindle microtubules can result in lagging chromosomes and eventually in chromosome missegregation.[156]

1.8.3 Chromosomal instability

Chromosomal instability eventually results in aneuploid progeny.[157] However, the presence of aneuploidy at a given moment does not necessarily indicate an ongoing contribution of CIN; aneuploidy can simply be the result of a single event that is not repeated. Therefore, CIN is better understood as an ongoing process of change. This is an important distinction, as highly aneuploid tumors do not inevitably have to be chromosomally unstable. Indeed, half of all human tumors have been shown to be karyotypically stable.[158,159] This means that in order to identify CIN, it is necessary to study spatially or temporally separated tumors, either within the same tumor mass or by comparing primary tumors to their metastases. Both aneuploidy and CIN have been implicated as principal contributors to genetic heterogeneity between cancers but also within single tumor lesions.

1.8.4 Aneuploidy, chromosomal instability and cancer

Aneuploidy has been associated with cancer for more than 100 years. German zoologist Theodor Boveri was the first to propose that abnormal chromosome numbers are implicated in cancer development, based on his observations that sea urchin eggs undergo abnormal mitotic divisions.[160] Since then, recurring non-random patterns of aneuploidy in cancer have been recognized, yet no general theory has emerged to explain them. That cellular aneuploidy indeed could be a driving force in tumorigenesis had only become evident in 2013, when it was shown that chromosomal imbalances due to aneuploidy likely result in haploinsufficiency of tumor suppressor genes (TSGs).[161] This is in contrast to the common "two-hit hypothesis," which states that TSGs require two hits, first a mutation and then a loss of heterozygosity (LOH) due to loss of the wild-type allele in order to result in a phenotype with increased tumor incidence.[162] Conversely, only one hit is needed under the assumption of haploinsufficiency, because wild-type protein levels might be not sufficient in executing their TSG function.

Owing to the typically late-stage disease of lung cancer patients, aneuploidy is omnipresent in lung cancers. In addition, aneuploidy induced by CIN in a mutant KRAS-driven lung cancer mouse model has been recently implicated in relapse and recurrence. In this model, CIN is induced by the overexpression of the SAC protein MAD2, and the tumors that overexpress MAD2 and KRAS are larger, more aggressive than tumors that only overexpress KRAS, and relapse quickly.[163,164] This indicates a role of CIN and/or aneuploidy in lung tumorigenesis that might by independent of KRAS, potentially due to an increased heterogeneity of karyotype formation. Although increased CIN can facilitate the emergence of advantageous karyotypes, it more likely has deleterious effects. A recent study has illustrated this bivalent nature of CIN: Interference with the SAC in a mouse model with increased incidence of lung and spleen tumors has resulted in increasing cell death. This indicates that for tumor formation and maintenance, the rate of CIN needs to be high enough to promote tumorigenesis, but not so high that it leads to cell death or tumor suppression.[165,166] Therefore, depending on the dosage, CIN can both promote and inhibit tumorigenesis.

1.8.5 Aneuploidy and patient prognosis

Since it has been recognized that most advanced-stage cancers are aneuploid, efforts have been made to link aneuploidy to patient prognosis. Early attempts in the 1970s relied primarily on the quantification of DNA content of nuclei isolated from solid human tumors by flow cytometry.[157] In fact, this was one of the first large-scale applications of flow cytometry.[167] A correlation between the DNA-ploidy status of tumors and the clinical outcome for the patients was only obtained in large studies that allowed for stratification according to tumor stage.[150] Since then, it has been shown in

breast[168], colorectal carcinoma[169], and other solid cancers[150] that patients with diploid tumors have a general better prognosis than patients with aneuploid tumors. Similarly, NSCLC patients with diploid tumors benefit from a substantially lower risk of NSCLC-related death compared to patients with aneuploid tumors, as indicated by a meta-analysis of 35 DNA-ploidy studies in lung cancer that where published before the year 2000 and included a total of more than 4,000 patients.[170] Nonetheless, ploidy measurements are not performed routinely anymore, because of overall discordant results due to the lack of stratification of mixed patient cohorts with tumors of all clinical stages.

2. Aims

This thesis is divided into two parts. The first, Part A, aims to develop and establish a technology that allows for the enrichment and genetic characterization of tumor cell nuclei from bulk tumor tissues. The second, Part B, aims to apply this technology to the genomic characterization of spatial and temporal genomic ITH of LUADs.

Part A: Develop and establish a multiparameter nuclei flow-sorting technology

Part A establishes a technique that aims to separate nuclei of tumor cells from nuclei of normal cells, in order to increase the precision of genomic analysis techniques for tumor populations. Therefore, a previously applied nuclei flow-sorting technique is refined. This includes the addition of an immunostaining for a tumor marker in order to allow for the isolation of tumor nuclei from bulk tissues. This is especially important for diploid tumor nuclei that cannot be distinguished from normal nuclei by DNA-ploidy histograms alone. Validation is performed by array-comparative genomic hybridization (aCGH) and targeted next-generation sequencing (NGS). Array-CGH is used to quantify numerical and structural copy number alterations and NGS is performed to detect somatic mutations. This is demonstrated in both fresh-frozen tissues and formalin-fixed and paraffin-embedded (FFPE) tissues of malignant melanomas and LUADs.

Part B: Decipher the genomic intratumor heterogeneity between primary LUADs and matched metastases

Part B aims to apply the technique established in Part A in order to investigate ITH between matched, longitudinal primary-metastatic tumors of a cohort of LUADs. This helps decipher the clonal relationship between primary tumors and metastases and therefore infer their evolution over time and space. From a technological viewpoint, it comprises a comprehensive genomic characterization of sorted tumor populations through genome-wide SCNA analysis and mutational analysis of 409 cancer census genes. Array-CGH and targeted ultra-deep sequencing are applied, respectively.

3. Materials and Methods

Patients and tissue samples

Fresh-frozen tissues were obtained from the biobank and FFPE tissues from the archive of the Institute for Pathology of the University Hospital Basel (IfP). Multiple frozen and FFPE test tissues were obtained for establishing the multiparameter flow-sorting approach, including LUADs and malignant melanomas. For Part B of this thesis, patients were chosen based on the following inclusion criteria: i) histologically diagnosed LUAD, ii) availability of two or more fresh-frozen biopsies, and iii) tissue samples differing in time point or site. In addition, FFPE tissues of different sites were used if available. In total, 35 fresh-frozen and two FFPE tissues were investigated. All tumors were evaluated histopathologically by two experienced lung pathologists from the IfP (Prof. Lukas Bubendorf and Dr. Sasenija Savic Prince). This study was approved by the Ethics Committee of Northwestern and Central Switzerland with the approval number EKBB 31/12.

Cell lines and cytospins

The cell lines A375 (malignant melanoma), HCC78 (NSCLC, LUAD), and H522 (NSCLC, LUAD) were obtained as a courtesy of Rosemarie Chaffard from the IfP. All cell lines were cultured in standard RPMI medium (Corning GmbH HQ, Wiesbaden, Germany) supplemented with 10% fetal bovine serum and grown in a humidified incubator with 5% CO_2 at 37°C. Cells were split by trypsinization every 3-4 days when they reached a confluency of 80-90%. Cytospins were made by spinning 50µL cell suspension in a Cytospin™ 4 Cytocentrifuge (Thermo Fisher Scientific, MA, USA) at 300 rpm for 2 min, followed by immediate fixation of the cells in Delaunay fixative (50% [v/v] ethanol, 50% [v/v] acetone, 2mM trichloroacetic acid).

Flow-sorting experiments

Isolation of nuclei

Nuclei were isolated according to published protocols.[171,172] Briefly, fresh-frozen tumors were minced in ice-cold nuclei isolation medium (NIM) buffer (10 mM Tris-HCl [pH 7.4] buffer containing 146 mM NaCl, 22 mM $MgCl_2$, 2 mM $CaCl_2$, 10% (v/v) DMSO, 0.05% (w/v) BSA, and 0.1% (v/v) IGEPAL CA-630). The nuclei were pelleted in a cooled centrifuge (300 g, 3 min, 4°C). The nuclei pellet was washed three times and finally resuspended in ice-cold NIM buffer and stored on ice for the subsequent flow cytometric analysis up to 8 h or at -20°C for longer storage up to 6 months.

For FFPE tissues, the isolation of nuclei was performed as previously described[173,174], with minor modifications. Histological sections of 55 μm were deparaffinized in xylene and rehydrated in

subsequent washings of 100%, 95%, 70%, 50%, and 30% (v/v) ethanol. Antigen retrievals for both the thyroid transcription factor 1 (TTF1) and the Sry-related HMg-Box gene 10 (SOX10) were performed in citrate buffer (pH 6.0) in a heat block at 95°C for 30 min. Nuclei were obtained after digestion with Collagenase III (final: 50 units/mL in 0.1 mM $CaCl_2$/Phosphate-buffered saline [PBS]) in a shaking heat block (16 h, 37°C, 1,000 rpm). After the nuclei were pelleted at 500 g for 5 min, they were resuspended in FACS buffer (PBS, pH 7.4 containing 1 mM EDTA, 1% (v/v) fetal calf serum [FCS], 0.1% (v/v) IGEPAL CA-630) to stop the digestion. After subsequent washings, isolated nuclei were stored in FACS buffer at 4°C for up to 1 week or at -20°C for up to 6 months.

TTF1 and SOX10 as markers in flow cytometry

SOX10 and TTF1 were used in combination with DAPI for setting up the staining protocols and procedures. SOX10 is a marker of neural crest origin and involved in the development of melanocytes. It is expressed in all stages, from neural crest stem cells to terminally differentiated melanocytes.[175–177] Nuclear SOX10-expression is found in practically all melanomas[175,178] and SOX10 was shown to have the highest overall detection rate when compared to the microphthalmia-associated transcription factor (MITF) and other pan-melanoma cocktails.[179] TTF1 is a protein that is expressed in the thyroid, lung, and ventral forebrain.[180] It is a marker of cellular lineage of the terminal respiratory unit and is routinely applied to differentiate LUAD from LUSC: 72% of all LUADs are TTF1-positive.[181]

Multiparameter flow sorting and DNA-ploidy analysis

Nuclei were stained with DAPI for DNA quantification and ploidy analysis. In addition to DAPI, SOX10 was also used in test tissues of malignant melanoma to set up the protocol and optimize the conditions for nuclei staining and sorting. Uniform TTF1 or SOX10 expression was ascertained by immunohistochemistry on corresponding FFPE tissue sections and evaluated by experienced lung (Prof. Lukas Bubendorf, Dr. Spasenija Savic Prince) or melanoma (Prof. Kathrin Glatz) pathologists from the IfP. In the LUAD cohort (Part B), 10 of 16 patients (62.5%) had TTF1-positive tumors. TTF1 was therefore used as a second parameter in 25 tumor manifestations of these 10 patients to distinguish tumor nuclei from stromal components.

All centrifugation steps were performed at 300 g for 3 min in a cooled centrifuge (4°C). Stainings were executed in nuclear NIM buffer for nuclei from fresh-frozen tissues or in FACS buffer for nuclei derived from FFPE sections (for composition of buffers see the “Isolation of nuclei” section). Nuclei were stained with an anti-TTF1 antibody (clone SPT24, Novocastra™, Product: NCL-L-TTF1) or anti-SOX10 antibody (Human/Rat SOX10 Antibody, Monoclonal Mouse IgG_1 Clone# 20B7, R&D Systems,

Product: MAB2864) on ice for 4 h, washed twice and incubated for 30 min with an anti-mouse IgG-Alexa 647 secondary antibody (F(ab')2-Goat anti-Mouse IgG (H+L) Secondary Antibody, Alexa Fluor® 647 conjugate, Thermo Fisher, Product: A-21237). To verify the TTF1- and SOX10-specific stainings, an IgG_1 isotype control was used as a primary antibody control (Mouse IgG_1 Isotype Control, R&D Systems, Product: MAB002). Concentrations of primary and secondary antibodies were 5 μg/mL and 2 μg/mL, respectively. Stained nuclei were filtered through a 40 μm mesh, and DAPI was added to a final concentration of 5 μg/mL at least 30 min before flow sorting was performed on a BD Influx™ cell sorter (Becton-Dickinson, San Jose, CA, USA). The instrument setup is displayed in Figure S2. A specific gating strategy was developed and applied as explained in the "Gating strategy of the multiparameter flow-sorting approach" section. Ploidy was calculated from the ratio of the geometric means of the DAPI intensity signal in the tumor population versus the respective non-tumor population (diploid without SCNAs, or "flat genomes"). Flow cytometric analysis and visualisation was performed with FlowJo X version 10.0.7 (FlowJo LLC., Ashland, OR, USA).

Gating strategy of the multiparameter flow-sorting approach

Additionally, a specific gating strategy was applied for the enrichment of single (Figure 15b), intact nuclei (Figure 15c). In all cases, an isotype control antibody was used to control for the sensitivity and specificity of the primary antibody and to quantify the background fluorescence that results from non-specific cellular protein interactions (Figure 15e).

In a representative example of a SOX10-positive melanoma in Figure 15 three peaks of different ploidy are visible from staining with DAPI alone (Figure 15c). DNA peaks can overlap, which results in cross-contamination if gating is performed in DNA-ploidy histograms because they are one-parametric visualizations of the DNA content only. To minimize this contamination here, each peak was gated separately in a two-parameter scatter plot (Figure 15e). Additionally, this helped set the threshold for positive staining, as the background fluorescence differs among different ploidy peaks (Figure 15f-h). Compared to the control (Figure 15f), staining with the specific antibody (here: anti-SOX10) was able to separate two populations from the diploid peak (Figure 15k), one positive and one negative for SOX10. The other two aneuploidy peaks were both positive for SOX10 (Figure 15l,m).

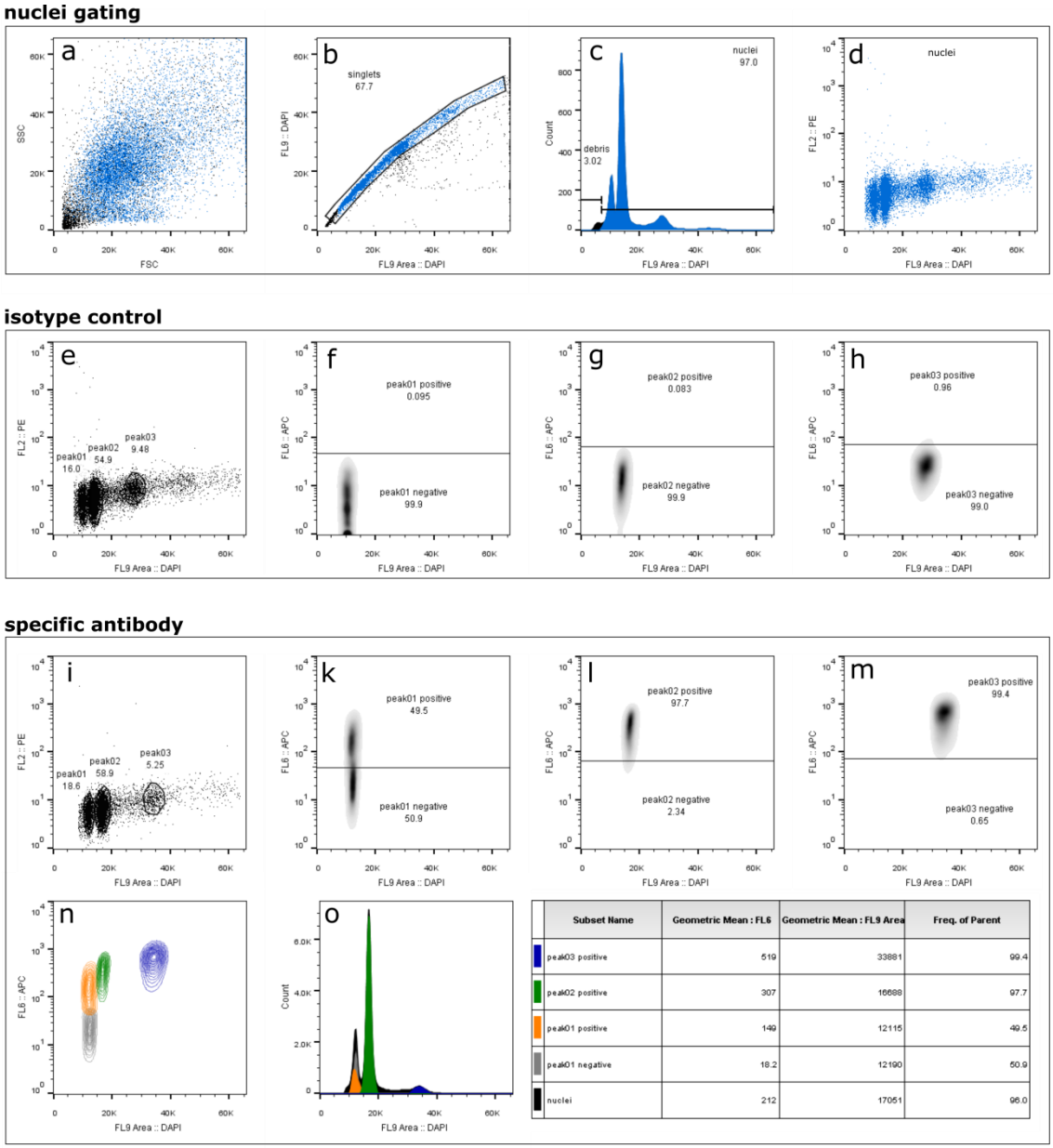


Figure 15 | Gating strategy of the multiparameter nuclei flow-sorting approach. A multiploid SOX10-positive melanoma with a diploid and aneuploid tumor population is shown as a general example. For all cases, the initial population (a) was stained for DNA content with DAPI and gated for single events (b). Only intact nuclei (c) were further included. (d) The final nuclei population is displayed in a two-parameter scatter plot. This procedure was done for an isotype control (e-h) in order to evaluate background staining intensities of different ploidies (f-h). The same was performed for nuclei stained with an antibody against SOX10 (i-m). Due to different background intensities of the different ploidies, each peak was gated separately (i) to separate positive from negative events. The ploidy of a population was calculated from the ratio of the geometric means of the DAPI signal (x-axis, FL9 Area DAPI) in each gated population versus the diploid, marker negative population. This gating strategy and analysis was applied on all TTF1-positive LUADs and all SOX10-positive melanomas with either an anti-TTF1 antibody or an anti-SOX10 antibody, respectively.

DNA extraction and quantification

Genomic DNA was extracted with the Maxwell® 16 FFPE Plus LEV DNA Purification Kit (Promega, AS1135, Madison, WI, USA) according to the manufacturer's protocol. Briefly, after sorting, the nuclei were pelleted in a cooled centrifuge (16.100 g, 30 min, 4°C). The nuclei pellet was resuspended in 180 μL Incubation buffer and 20 μL Proteinase K and digested in a shaking heat block (56°C, 16 hours, 500 rpm). After digestion, 400 μL Lysis buffer was added and heated to 80°C for 10 min. The Maxwell® 16 Instrument was used to extract DNA in 50-100 μL nuclease-free water. Double-stranded DNA (dsDNA) was quantified with the Qubit® dsDNA HS (High Sensitivity) Assay Kit (Invitrogen™), which was used according to the manufacturer's instructions. Typical DNA extractions from sorted nuclei reached quantities of 100-200 ng/100,000 nuclei, depending on the ploidy.

Whole genome amplification

DNA of sorted populations from fresh-frozen tissues was amplified with the illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences,PA, USA). This whole genome amplification (WGA) is based on an isothermal multiple displacement DNA amplification with the Φ29 enzyme, a high-fidelity DNA polymerase with proofreading (3'-5' exonuclease) activity. Amplified DNA was used for both aCGH and NGS. DNA from sorted populations from FFPE tissues was not amplified, in order to reduce sequencing biases. Notably, three criteria were applied to optimize the quality of amplified DNA and to reduce potential amplification biases. First, amounts of 50-100 ng were used whenever possible and second, equal amounts of test and reference DNA (46XX, Promega, Product: G1521) were amplified as suggested previously.[182] Last, amplification was performed only for 1 h, because shorter reaction times with Phi29 were shown to significantly decrease amplification biases.[183]

The DNA from the sorted populations was SpeedVac™-ed to 1 μL and mixed with 9 μL Sample Buffer. The mixture was denatured at 95°C for 3 min and immediately cooled on ice for at least 5 min. A mix of 9 μL Reaction buffer and 1 μL of the enzyme was added to the sample. Amplification was performed at 30°C for 1 h in a thermocycler, followed by inactivation of the Φ29 enzyme at 75°C for 10 min. High molecular weight DNA (average ≥ 10 kb) was assured by gel electrophoresis with the FlashGel™ DNA System (Lonza, Basel, Switzerland). Nuclease-free water was used as a negative control. Amplified DNA was diluted 1/10 in nuclease-free water and quantified with the Qubit® dsDNA HS Assay Kit (Invitrogen™). Amplifications resulted in a total DNA quantity of around 4.5-5.5 μg (225-275ng/uL) and no DNA product in the negative water control.

Array-comparative genomic hybridization

Note: The DNA of the sorted populations from FFPE tissues (n=2) was not amplified and not digested. For FFPE-DNA, 600 ng of unamplified reference DNA was digested as described below. Labelling, filtering and hybridization procedures were not different between DNA from fresh-frozen and FFPE tissues.

Digestion of DNA from fresh-frozen tissues

The amplified DNA of the sorted populations from fresh-frozen tissues was digested with DNaseI to reach a DNA smear from 100 to 1,000 bp. Therefore, 1 μg of both the amplified sample and reference DNA was diluted with nuclease-free water to a total volume of 7 μL and mixed with 1 μL of 10x Reaction buffer (200mM TrisHCl [pH 7.5], 100 mM $MgCl_2$, 10 mM $CaCl_2$). A DNAseI dilution was prepared by combining 1.25 μL DNAseI (5,000 units/mL, Thermo Fisher, Cat#89835) and 500 μL 1x Reaction buffer. Two microliters of this DNAseI dilution were added to the amplified DNA to reach a total volume of 10 μL. DNA was digested in a thermo cycler for 15 to 18 min at 30°C, followed by DNAseI inactivation at 75°C for 10 min. Degraded DNA with a molecular weight of 100-1,000 bp was assured by gel electrophoresis with the FlashGel™ DNA System (Lonza, Basel, Switzerland).

Labelling and filtering

Digested test and reference DNA were labeled with Cy3-dUTP and Cy5-dUTP, respectively, using the BioPrime® Array-CGH Genomic Labeling System (Thermo Fisher Scientific, MA, USA). Therefore, degraded DNA was incubated with 2.5x Random Primers in a total volume of 40 μL at 95°C for 5 min, followed by the addition of 9 μL Labeling Master Mix (5 μL 10x dUTP, 1 μL Exo-Klenow, 3 μL Cy3-UTP or Cy5-dUTP). Labeling was performed in a thermo-cycler at 37°C and stopped after 2 h by the addition of 5 μL EDTA. Labeled DNA was filtered with Amicon Ultra 30K Filter Units (Millipore, Cat# UFC503096). To do so, labeled DNA was mixed with 450 μL TE-buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0.) and centrifuged at 8,000g for 10 min, followed by one wash with 500 μL TE-buffer. Finally, the DNA was eluted, and successful labeling was verified by calculating the specific activity (>30) of each sample (pmol dye/μg DNA) using a Nanodrop for evaluation.

Hybridization, washing and scanning

Labeled test and reference DNA were pooled and adjusted to a volume of 39 μL with TE-buffer. 71 μL Hybridization Master Mix (5 μL Cot-1 DNA, 11 μL Agilent 10x Blocking Agent, 55 μL 2x Agilent Hybridization Buffer) were added to each test/reference sample and incubated at 98°C for 3 min, followed by another incubation at 37°C for 30 min. Subsequently, each sample was hybridized on

180K SurePrint G3 Human CGH Microarrays (Agilent Technologies, CA, USA) for 24 h at 67°C in a rotating hybridization chamber. All microarray slides were washed according to the manufacturer's instructions and scanned with the Agilent 2565C DNA scanner. Images were analyzed with Agilent's Feature Extraction v10.7 using default settings.

Detection of somatic copy number aberrations

Feature-extracted aCGH data were evaluated using Agilent's CytoGenomics v3.0.1.1 software. Recentralization was performed by setting ploidy within +/- 0.5 of the prior ploidy estimate, as calculated from DNA-ploidy flow cytometry. Aberrations were called with the aberration detection algorithm ADM2 set to a threshold of 12.0, with Fuzzy Zero and GC-content (window size: 2 kb) correction. All aberrations were manually inspected and corrected before the aberration data table and raw probe log2-ratios were exported and further processed by a custom workflow programmed in the statistical computing software R version 3.2.3. This included the Bioconductor package "copynumber" version 3.4[184] for visualization of the aCGH data as produced in the frequency plot in Figure 29 and the sample specific copy number plot in Figure 33.

It is important to note that diploid populations with "flat genomes," that is, with no aberrations present except of common copy number variants (CNVs), were considered to be of non-tumor origin (see 4.1.3). They were used as germline controls to identify both SCNAs in aCGH data and SNVs in sequencing analysis (see the "Next-generation sequencing" section in "Materials and Methods"). In cohort-wide analyses, each SCNA was considered only once per patient to exclude overrepresentation of ubiquitous or shared SCNAs present in patients with more than two tumors.

Next-generation sequencing with the Ion Ampliseq™ Comprehensive Cancer Panel

Library preparation and sequencing

Sequencing was performed on all sorted populations, both tumor and non-tumor populations. The non-tumor populations were used as germline controls. For three patients, non-tumor tissue was available and utilized to validate the usage of the sorted non-tumor populations as germline controls in NGS analysis (see 4.1.3). Library preparation was carried out according to the manufacturer's instructions[185]. The Ion Ampliseq™ Comprehensive Cancer Panel (CCP), which targets the exons of 409 genes frequently mutated in cancer, was used for both fresh-frozen and FFPE populations. Most importantly, the CCP includes all genes that are well known to be mutated in LUADs (a list of the genes can be found in Table S1). Quality checks of the libraries were performed for both the expected library size (using Agilent's High Sensitivity DNA Analysis Kit on a Bioanalyzer) and library

concentration (quantitative real-time PCR with the Ion Library Quantitation Kits). Sequencing was performed with the Ion PI™ Chips on the Ion Proton™ Sequencer with a mean coverage of 965x.

Variant calling and filtering

Sequence alignment to target regions from the hg19 genome was performed with the IonTorrent TorrentSuite™ software. Variant calling was performed with the Torrent Variant Caller 5.0 plugin from the TorrentSuite (Thermo Fisher Scientific, MA, USA) using low-stringency settings (Somatic – Proton – Low Stringency), as suggested by the manufacturer . Variants with a Phred-scored quality of ≥ 50 and a strand bias < 0.95 were considered, but had to meet the following thresholds: minimum coverage ≥10, minimum variant allele coverage ≥ 5, minimum variant allele frequency (VAF) ≥ 4%, homopolymer ≤ 10, and common_signal_shift ≤ 0.2. A variant was then considered present, if the ratio $\frac{VAF(tumor)}{VAF(germline\ control)} \geq 4$ or if $VAF(germline\ control) = 0$. We applied this last filtering step, due to the use of sorted non-tumor populations (diploid populations without SCNAs, or "flat genomes") as germline controls. Minute amounts of tumor DNA are still present in these populations, which is inevitable in the flow-sorting process. Known hotspot mutations of *KRAS* and *EGFR* were always considered somatic. All variants were annotated using ANNOVAR[186] and curated by manual inspection using the Integrated Genomics Viewer (IGV)[187].

Whole-exome sequencing with the IonAmpliseq™ Exome RDY Kit

Library preparation and sequencing

Whole-exome sequencing was performed for the four tumor populations of patient 42 (three tumor population in the primary tumor and one in the metastasis). The Ion Ampliseq™ Exome RDY Kit was used to prepare libraries following the manufacturer's protocol.

Quality checks were performed as described in the "Library preparation and sequencing" subsection of the "Next-generation sequencing" section of this chapter. Template preparation and chip loading was done with the Ion Torrent™ Ion Chef™. Sequencing was performed with the Ion PI™ Chips on the Ion Proton™ Sequencer with a mean coverage of 85x (range: 68.3 - 100.5x).

Variant calling and filtering of whole exome sequenced tumors from patient 42

We performed very stringent variant calling and filtering for the whole-exome sequencing data for patient 42, as our focus was to investigate the genomic relationships between the tumor populations. Therefore, a smaller number of high-confident and validated mutations were prioritized over a large number of mutations. Variant calling was performed with both the Torrent Variant Caller

5.0 plugin from the TorrentSuite and the "somatic"-tools from VarScan2 v2.3.9, as described below. To achieve a high stringency, the intersection (variants called by both algorithms) was used for further ultra-deep re-sequencing with an Ion Ampliseq™ custom validation panel. The intersection, rather than the union, of the two variant calling algorithms was used, as high-confidence variants were favored over the number of variants in respect to evolutionary analysis.

Variant calling with the TorrentSuite

Sequence alignment to target regions from the hg19 genome was performed with the Ion Torrent TorrentSuite™ software. Variant calling was performed with the Torrent Variant Caller 5.0 plugin from the TorrentSuite using low-stringency settings (Somatic – Proton – Low Stringency). The following thresholds were applied for filtering the variants identified by the Ion Torrent Variant Caller: Phred-scored based quality ≥ 50, strand bias < 0.90, coverage ≥ 10, variant allele coverage ≥ 5, VAF ≥ 4%, and common signal shift ≤ 0.2. Variants, which were not filtered out, were considered present if the ratio of $\frac{VAF(tumor)}{VAF(germline\ control)} \geq 4$ or if $VAF(germline\ control) = 0$.

Variant calling with VarScan2

A second, separate variant calling was performed between tumor and matched germline using the "somatic" command from VarScan2 v2.3.9.[188] The input for VarScan2 was the SAMtools[189] mpileup output from the combined tumor/normal samples. The "somatic" command from VarScan2 was applied with default settings, except for the following: minimum coverage (--min-coverage-normal/tumor 10) was set to 10 for both the tumor and the normal sample; the minimum variant frequency (--min-var-freq 0.04) was set to 4%; both tumor and normal purity (--normal/tumor-purity 0.85) were applied with a value of 85%; and variants with >90% support on one strand only (--strand-filter 1) and fewer than five supporting reads (--min-reads2 5) were filtered out.

Finally, for the purpose of this analysis, small insertions and deletions were not evaluated. Variants that were called by both approaches (intersection of both Variant Caller and VarScan2) were further curated by manual inspection using the IGV[187]. The variants (n=112) meeting all these criteria were eventually used for further validation by ultra-deep re-sequencing (mean coverage 5,864x) with an Ion Ampliseq™ custom validation panel. All variants were annotated using ANNOVAR[186].

Ion Ampliseq™ custom validation panel for patient 42

A total of 112 mutations were detected in the four tumor populations of patient 42 by whole-exome sequencing (see "Whole-exome sequencing with the IonAmpliseq™ Exome RDY Kit"). The presence or absence of all mutations was validated with an Ion Ampliseq™ custom panel. This panel was

designed using the online designer (www.ampliseq.com). Multiplex PCRs were performed according to the manufacturer's protocol. Barcoded libraries (read length: 400 bp) were constructed, and their quality was examined with both a Bioanalyzer for the expected size range and a qPCR to assess quantity. Template preparation and chip loading was done with the Ion Torrent™ Ion Chef™. Ultra-deep sequencing was performed with Ion 316™ Chips v2 on the Ion Torrent™ Ion S5™ system with a mean depth of 5,864x (range 5,203-6,118x).

In total, 16 mutations were either absent in all tumor regions or identified as germline variants (overall validation rate 85.7%). The mean sample specific validation rate was 90.4% (range 87.5%-93.8%). The presence or absence of the mutations was ensured by two additional quality check: (i) sorting replicates (which comprise resorting of the primary tumor and metastasis, DNA extraction, and WGA) were subjected to re-sequencing with this custom panel and (ii) sequencing replicates (in duplicates), starting from the library preparation step with individual barcoding. Both were performed for further certainty of the VAFs of the detected mutations. The correlation of the VAFs of both the sorting and the sequencing replicates were assessed by linear regression analyses. Here, the mean squared correlation coefficient R^2 was 0.93 (range 0.91-0.96, Figure S13) for **inter**-sort comparison and 0.93 (range 0.90-0.95, Figure S14) for **intra**-sort comparison, including all tumor populations from the primary tumor (n=3) and the metastasis (n=1).

Categorization of mutations and copy number aberrations based on their presence in primary tumors and matched metastases

Mutational events, both SCNAs and SNVs, were categorized as ubiquitous, shared, and private , according to their presence in the primary tumor and corresponding metastases. An SCNA or SNV was defined as "ubiquitous", when it was present in all tumors investigated for a given patient, or as "private" otherwise. An exception was made for patients with more than one metastasis and for patient 42 with a multiploid primary tumor. Here, the definition "shared" was used if an SCNA or SNV was shared between at least two, but not all, sorted tumor populations. Therefore, "private" was used in these patients to describe the presence in only one tumor population. Non-synonymous SNVs and splice-site- or regulatory SNVs (generally named nonsilent) were distinguished from synonymous and intronic SNVs (also referred to as silent). A functional effect prediction was further applied to nonsilent mutations as described below.

Functional effect prediction and classification of mutations

Combined scoring of five functional protein prediction algorithms was applied to nonsilent SNVs in order to distinguish pathogenic SNVs from non-pathogenic SNVs. All identified nonsilent mutations

were classified into five categories, based on the functional predictors SIFT[190], Polyphen2[191], MutationTaster[192], FATHMM[193] and Provean[194]. Binary scoring of these predictors was performed, where 1 is a "deleterious" and 0 is a "non-deleterious" mutation.

In order to evaluate the level of confidence (C) of calling a mutation "pathogenic," the predictions of all five predictor algorithms were weighted as follows:

$$C = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ X_i = \{0,1\}$$

where n is the number of predictors with available information for a specific mutation, and X_i is the binary prediction of a predictor algorithm with 0 being a "benign", and 1 being a "pathogenic" mutation. The impact of the mutation on the protein function was than divided into five categories, depending on the value of C: Category 1 "confidently deleterious" (C = 1), Category 2 "deleterious" (0.66 ≤ C < 1), Category 3 "possibly deleterious" (0.5 ≤ C < 0.66), Category 4 "possibly non-deleterious" (0.2 ≤ C < 0.5) and Category 5 "confidently non-deleterious" (C < 0.2).

Mutations in Categories 1-3 as well as nonsense (stopgain/truncating), splice-site, and frameshift mutations were considered "pathogenic". Additionally, mutations in Categories 4 and 5 were considered "pathogenic" if these were predicted to be "driver" alterations by CHASM[195] (Lung-Adenocarcinoma; available online under www.cravat.us/CRAVAT/) or considered "passengers" otherwise. A default threshold of p ≤ 0.05 (p-value [missense]) was set for CHASM cancer driver mutations.

Mean pairwise genetic divergence

The genetic divergence of primary tumors and metastases for each patient was the number of SCNAs and SNVs differing between two samples divided by the number of SCNAs and SNVs that were present in both, the latter of which were likely present in their most recent common ancestor. The mean pairwise divergence (D) is defined as

$$D = \frac{2}{n(n-1)}\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{X_{i,j}}{Y_{i,j}}$$

where n is the number of samples in a patient, $X_{i,j}$ is the number of SNVs and SCNAs private to either sample i or j but not present in both, and $Y_{i,j}$ is the number of SNVs and SCNAs present in both samples i or j. The mean pairwise divergence was calculated by comparing the primary tumor with the first biopsied metastasis only.

Purity estimation and characterization of phylogenetic relationship for tumor populations of patient 42

Purity estimation

We used EXPANDS[196] version 1.7.2 with default parameters to infer the purity of 39 tumor populations from primary and metastatic LUAD samples from 16 patients. The size of the largest clone detected is a direct indicator for purity.[197] SNVs that could not be explained by a clone present in 10% or more of the sample, at a ploidy of six or less, were excluded. The remaining detected SNVs and copy number segments were used to predict the number of clones that coexisted in each tumor sample, clone size, clone specific SNVs/SCNAs, and phylogenetic relations between clones. Yet, due to the overall low number of somatic SNVs in most cases, we used EXPANDS for purity estimations and not for subclonal detection. Purity estimations were available for 35 tumor populations, but unsuccessful for four samples due to insufficient numbers of SCNAs (n=2) or SNVs (n=2).

Phylogenetic relationship analysis for tumor populations of patient 42

Phylogenetic relationships among tumor populations in patient 42 were calculated using the neighbor-joining algorithm provided by the R-package "ape". Hereby, pairwise distances between clones were calculated as the number of copy number segments for which both clones had the same copy number, divided by the total number of copy number segments for which both clones had available copy number information. EXPANDS analysis was performed by Karen Pereira and Stephanie Greer (Division of Oncology, Department of Medicine, Stanford University School of Medicine, USA). Phylogenetic analysis was performed by computational biologist Dr. Noemi Andor (Division of Oncology, Department of Medicine, Stanford University School of Medicine, USA).

Immunohistochemistry

Immunohistochemistry (IHC) was performed on FFPE tissue sections using the automated immunostainer Benchmark XT (Ventana, Tucson, AZ, USA) following the manufacturer's guidelines. The IHC for TTF1 was performed to validate uniform staining in TTF1-positive tumors (10/16 patients, 62.5%) before multiparameter flow sorting. TTF1 can be expressed by normal lung alveolar epithelial cells or in thyroid follicular cells, as was the case in the thyroid gland metastasis of patient 12 (Figure S3). This information was necessary to validate the non-tumor origin of those cells in sorts with a TTF1-positive population that carried no SCNAs ("flat genome") as detected by aCGH. The IHC for p16, a protein encoded by *CDKN2A*, was performed on the primary tumor and metastasis of patient 42. SOX10-IHC was performed to validate uniform staining in SOX10-positive malignant melanomas.

4. Results

The results of this thesis are divided into two parts. The first part (Part A) introduces a refined nuclei sorting method that includes tumor marker-assisted flow cytometry to enrich tumor nuclei from bulk tissues, most importantly from diploid tumors. To establish the methodology, cell cultures and human tissues of two cancer types were used: malignant melanoma and LUAD. The second part (Part B) of this thesis presents data that were obtained by applying the method described in Part A to a comprehensive genomic characterization, comprising genome-wide copy number aberrations and deep-sequenced cancer gene mutations, of matched primary-metastatic LUADs from 16 patients.

4.1 Results Part A: Establish a multiparameter flow-sorting approach for genomic characterization of tumor nuclei

4.1.1 Cell lines staining experiment

Cell lines were used to test the possibility of nuclei being immunostained for transcription factors in addition to DAPI, which is used for the quantification of DNA. Two transcription factors were investigated: SOX10, because it is expressed in virtually all melanomas, and TTF1, because 70% of LUADs are TTF1-positive. Furthermore, TTF1 is used to differentiate LUAD (TTF1-positive) from LUSC (TTF1-negative) in diagnostic routine. For this purpose, three cell lines were used: HCC78, an NSCLC (subtype: LUAD) cell line, that was previously tested to express TTF1[198], A375, a malignant melanoma cell line that was used in other studies as a positive control for SOX10[199]; and H522, an NSCLC (subtype: LUAD) cell line that does not express TTF1[200].

In line with these data, we detected nuclear expression of TTF1 in HCC78 by IHC (Figure 16a). Similarly, the malignant melanoma cell line A375 expressed the transcription factor SOX10 in the nucleus (Figure 17a). H522 was negative for both TTF1 and SOX10 by IHC (Figure 16a and Figure 17a) and therefore used as a negative control for flow cytometry analyses. In agreement with IHC, nuclei extracted from HCC78 were positive for TTF1 in flow cytometry, whereas no TTF1 signal above the isotype control could be detected in H522 nuclei (Figure 16b), which is therefore considered TTF1-negative. While isotype controls are used widely in flow cytometry to measure the maximum background noise, the best control is to mix positive and negative samples before staining with an antibody for a protein of interest. By doing so, we were able to separate TTF1-positive HCC78 nuclei from TTF1-negative H522 nuclei (Figure 16c). Similar results were obtained with A375 and H522 for SOX10. In agreement with IHC, A375 nuclei were SOX10 positive, whereas H522 nuclei were negative for SOX10 by flow cytometry (Figure 17c). Mixing nuclei of both cell lines before addition of the anti-

SOX10 antibody resulted in the separation of SOX10-positive A375 nuclei and SOX10-negative H522 nuclei (Figure 17d).

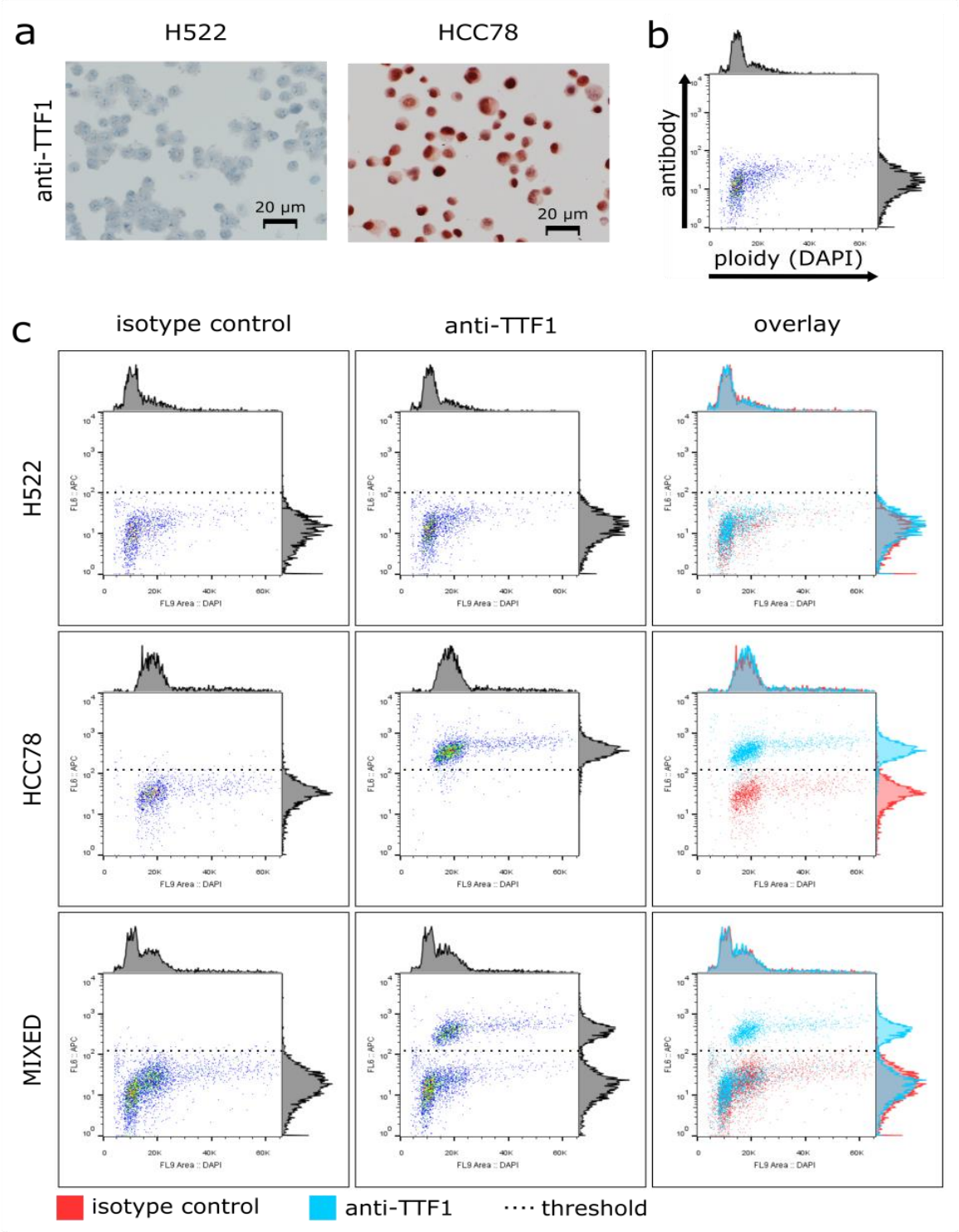


Figure 16 | Multiparameter nuclei flow cytometry with TTF1 in cell lines HCC78 and H522. (a) IHC for TTF1 on cytospins from the cell lines H522 and HCC78, respectively. (b,c) Multiparameter flow cytometry of nuclei extracted from both cell lines and stained for DNA content with DAPI (x-axis) and TTF1 or control antibody (y-axis). Cell lines were either stained in separate tubes or mixed before the addition of the anti-TTF1 antibody.

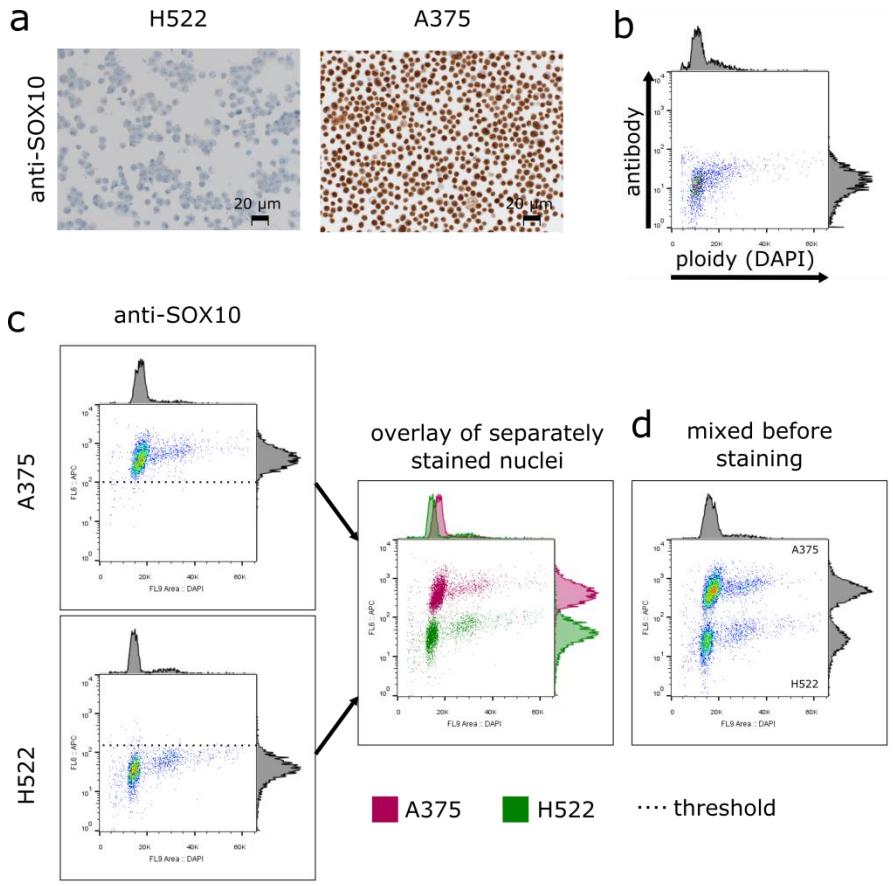


Figure 17 | Multiparameter nuclei flow cytometry with SOX10 in cell lines A375 and H522. (a) IHC for SOX10 on cytospins from the cell lines H522 and A375, respectively. (b,c) Multiparameter flow cytometry of nuclei extracted from both cell lines and stained for DNA content with DAPI (x-axis) and SOX10 or control antibody (y-axis). Cell lines were either stained in separate tubes (c) or mixed (d) before the addition of the anti-SOX10 antibody.

4.1.2 Multiparameter flow sorting enriches for tumor DNA of diploid tumors

Detection of SCNAs and inference of absolute copy numbers is difficult because of the issues that arise from samples with low to medium tumor purity and assumptions of ploidy for computational deconvolution. This is especially true for diploid tumors, where tumor cells cannot be separated from diploid normal cells by ploidy alone. Therefore, we applied this multiparameter flow-sorting approach to nuclei derived from diploid tumors followed by detection of SCNAs by aCGH. As a proof of concept, four diploid tumors, two malignant melanomas, and two LUADs were used. The melanomas were stained for SOX10 and the LUADs for TTF1.

In all four cases, a diploid, marker-positive (SOX10 or TTF1) population (2N+) and a diploid, marker-negative (2N-) population was detected (Figure 18a,c,e,g). Array-CGH of the sorted populations revealed the presence of SCNAs in all 2N+ populations (Figure 18b,d,f,h). The second, smaller peaks were considered the proliferating fraction (cells in G_2-phase) of the diploid tumors, based on three observations: (i) Their ploidy was approximately two times the ploidy of the 2N+ populations, (ii) they were in number much less than the 2N+ populations, and (iii) their aCGH-profiles were identical to

those of the 2N+ populations, without the presence of any additional private SCNAs. In contrast to the 2N+ populations, no SCNAs were detected in the 2N- populations, except of common CNVs and aberrations due to gender differences (chromosomes X, Y). These data suggest that the 2N- populations consist of nuclei of normal cells that are also present in tumors such as fibroblasts, lymphocytes and/or epithelial cells.

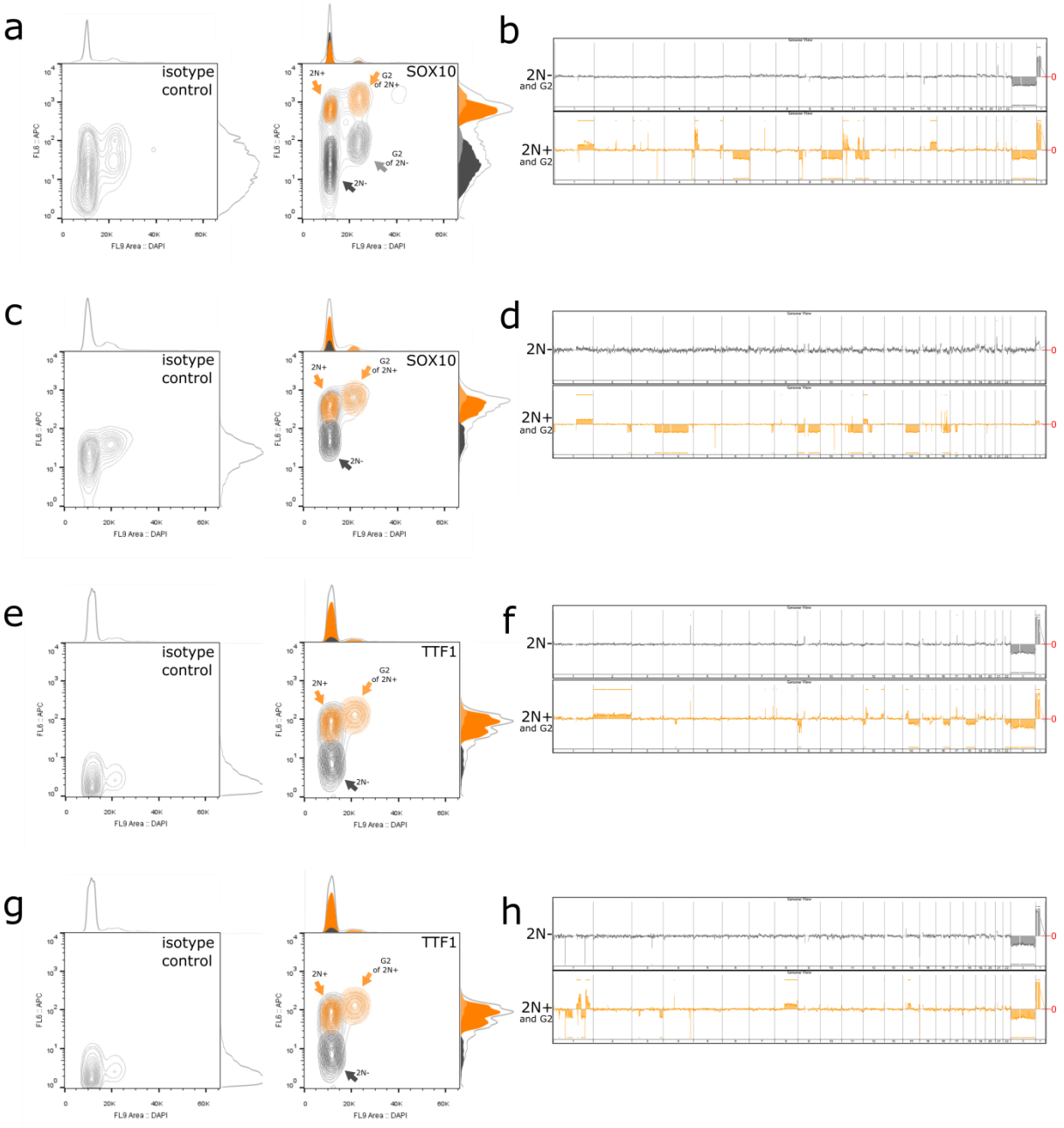


Figure 18 | Application of the multiparameter nuclei flow-sorting approach on diploid tumors. Sorts of two malignant melanomas (a,c) with subsequent genome-wide copy number aberration analysis (b,d) of both the diploid, SOX10-positive and diploid, SOX10-negative sorted population. Sorts of two LUADs (e,g) and the aCGH profiles of diploid, TTF1-positive and diploid, TTF1-negative populations (f,h). In all cases, isotype controls were used to detect antibody-specific staining. Orange: diploid, marker-positive populations; grey: diploid, marker-negative populations. The red 0 next to the aCGH profiles indicates a log2-ratio of zero, which is identical to a copy number of two. Therefore, any aberration above this line is a chromosomal gain or amplification and any aberration below is a loss or deletion. *Note: The populations indicated as G_2 consist of nuclei from proliferating tumor cells (light orange) or proliferating normal cells (light grey). Their aCGH profiles were identical to their non-proliferation counterparts and are therefore not shown to exclude redundancy.*

4.1.3 Diploid populations with flat genomes are of non-tumor origin and can serve as germline controls in sequencing studies

The presence of SCNAs in all the 2N+ populations in the previous experiment suggests a tumor origin. However, one must be cautious, if the protein of interest is not expressed by the tumor alone, but also by cells of non-tumor origin. TTF1, for instance, is a lineage marker of the terminal respirator unit[181] and is therefore always expressed by normal AEC2 cells in the lung and by normal thyroid follicular cells.[180] Hence, both tumor cells and normal cells might be TTF1-positive in a given LUAD or in a thyroid metastasis.

To genomically validate such circumstances, we selected two cases. The first case was a TTF1-positive primary LUAD with co-occurring TTF1-positive, normal alveolar epithelial cells (Figure 19a). In agreement with IHC, the multiparameter flow sorting by DNA content and TTF1 demonstrated the presence of both diploid, TTF1-positive (2N TTF1+) and diploid, TTF1-negative (2N TTF1-) populations in the primary tumor (Figure 19b). The matched adrenal gland metastasis contained only TTF1-positive tumor nuclei but no TTF1-positive normal nuclei as verified by both IHC and flow cytometry (Figure 19c,d). As expected from IHC analysis for both the primary tumor and the metastasis, no SCNAs were found in either the diploid, TTF1-positive or TTF1-negative populations (Figure 19e). In contrast, a large amount of copy number alterations were detected in the aneuploid TTF1-positive tumor populations (AN TTF1+) of both the primary tumor and the adrenal gland metastasis (Figure 19e). A similar situation was found in a thyroid gland metastasis of a patient with a TTF1-negative LUAD (Figure S3). Here, a 2N TTF1+ population was also visible in the flow cytometry analysis (Figure S3b). Yet, these were normal, diploid thyroid follicular cells, as ascertained by IHC (Figure S3a) and the absence of SCNAs (Figure S3c).

To assess whether these sorted diploid populations without SCNAs (2N SCNA-negative) can be used as substitutes for germline DNA as controls in NGS studies, we submitted them to sequencing of 409 well-known cancer genes for three cases. DNA from non-tumor material for each patient served as germline control. As in the aCGH analysis for SCNAs, no SNVs were detected in the 2N SCNA-negative populations for all three patients (Figure 21, Figure S4, and Figure S5). Conversely, all aneuploid tumor populations had somatic mutations, some of them with VAFs close to 100%. Likewise, heterozygous single nucleotide polymorphisms (SNPs: germline mutations) were presented with VAFs around 50% in the 2N SCNA-negative populations and reached VAFs of 100% or were lost in the tumor populations, indicating of LOH due to the loss of the wild-type or mutant allele, respectively.

To further assure the non-tumor origin of the 2N SCNA-negative populations, we analyzed their distribution pattern of VAFs of heterozygous SNPs. The VAFs of these SNPs clustered around 50% in the 2N SCNA-negative populations similarly to in the germline controls, whereas the tumor

populations were characterized by skewed VAF distributions, attributable to SCNAs in the respective loci. In addition, similar distribution patterns were obtained for the tumor populations no matter if the tumor-free control DNA or the DNA extracted from the sorted 2N SCNA-negative populations (representative case in Figure 20, other two patients not shown) was used. These data suggest not only that the 2N SCNA-negative populations are of non-tumor origin, but also that DNA from these populations can be used as germline controls in sequencing experiments.

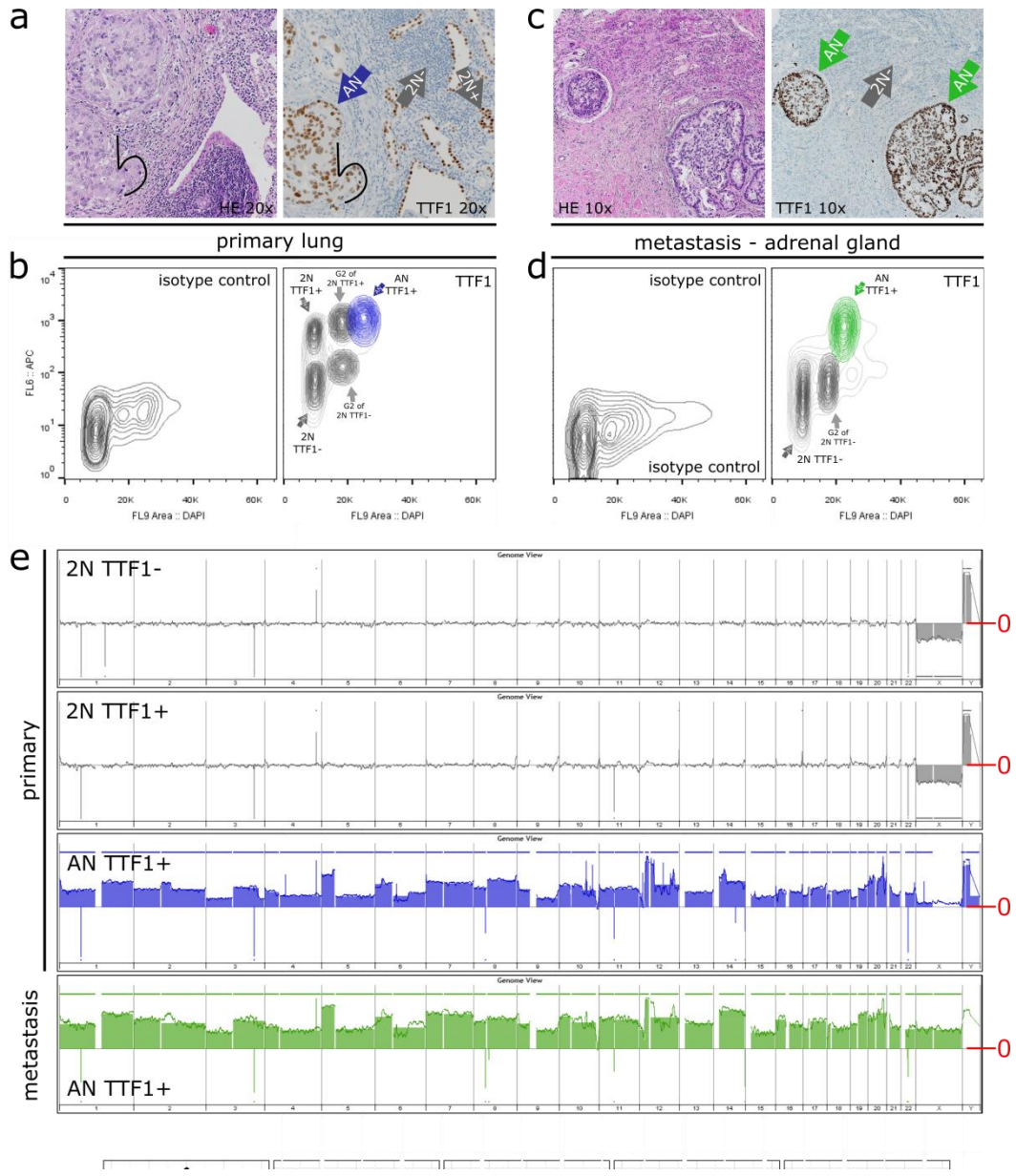


Figure 19 | Example of diploid, TTF1-positive normal cells. Nuclei flow sorting and aCGH analysis of an aneuploid, TTF1-positive LUAD with diploid, TTF1-positive alveolar epithelial cells. IHC for TTF1 of the primary tumor (a) and metastasis (c). Multiparameter flow sorting of the primary tumor (b) and metastasis (d). (e) Genome-wide copy number plots from aCGH analysis of sorted populations. Grey, normal cells without any SCNAs; blue, aneuploid, TTF1-positive tumor population of the primary tumor; green, aneuploid, TTF1-positive tumor population of the metastasis. The red 0 next to the aCGH profiles indicates a log2-ratio of zero, which is equal to a copy number of two. *Note: This is patient 44 from the LUAD cohort from Part B of this thesis.*

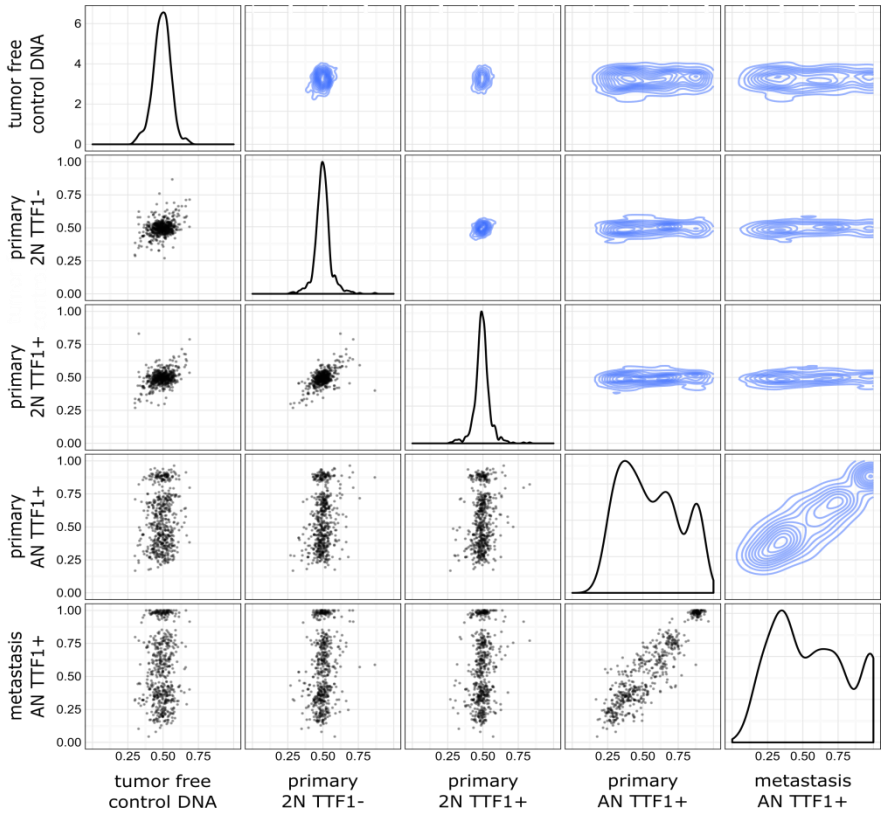


Figure 20 | Distribution of VAFs of heterozygous germline SNPs. The VAFs of diploid, SCNA-negative populations from Figure 19 (primary 2N TTF1- and primary 2N TTF1+) are compared to a tumor-free control from the same patient. The VAFs of the aneuploid tumor populations of the primary tumor and the metastasis are included as well to demonstrate the skewed VAF distributions of a tumor (as a result of various copy number states of the mutated loci) in comparison to the VAFs of a non-tumor population. VAFs > 80% in the tumor populations indicate the presence of LOH. *Note: These are data from patient 44 from the LUAD cohort from Part B of this thesis.*

					Tumor free	Primary			Metastasis			
gene	Chrom	pos	ref	var	control	2N TTF1-	2N TTF1+	AN TTF1+	AN TTF1+	Type	effect	Protein change
ADAMTS20	chr12	43826576	C	G	0.00	0.00	0.00	0.00	0.04	somatic	missense	p.Cys920Ser
PIK3CG	chr7	106509630	C	T	0.00	0.00	0.00	0.00	0.21		missense	p.Arg542Trp
LRP1B	chr2	141458130	T	C	0.00	0.00	0.00	0.00	0.23		missense	p.Asn2163Ser
PDE4DIP	chr1	144874726	C	T	0.00	0.00	0.00	0.07	0.00		missense	p.Glu1628Lys
FH	chr1	241663767	T	C	0.00	0.00	0.01	0.10	0.00		missense	p.Met454Val
MARK4	chr19	45801160	C	T	0.00	0.00	0.01	0.20	0.40		missense	p.Pro609Ser
MTOR	chr1	11317061	C	A	0.00	0.00	0.01	0.20	0.17		missense	p.Val145Leu
SMO	chr7	128845139	G	T	0.00	0.00	0.01	0.21	0.36		missense	p.Glu211Asp
PTPRD	chr9	8521226	C	A	0.00	0.00	0.00	0.21	0.40		intronic	
MTR	chr1	237057688	A	T	0.00	0.01	0.01	0.24	0.23		missense	p.Tyr1079Phe
EPHA7	chr6	93965562	G	T	0.00	0.00	0.00	0.30	0.32		missense	p.Ala789Asp
CRKL	chr22	21272375	G	T	0.00	0.00	0.00	0.36	0.36		synonymous	p.(=)
ATM	chr11	108115675	T	C	0.00	0.00	0.00	0.38	0.23		synonymous	p.(=)
CSMD3	chr8	113277736	T	C	0.00	0.01	0.02	0.47	0.53		missense	p.Met3198Val
AR	chrX	66765577	C	T	0.00	0.01	0.01	0.72	0.93	somatic LOH	nonsense	p.Gln197Ter
HOOK3	chr8	42798388	G	C	0.00	0.02	0.01	0.77	1.00		intronic	
FGFR4	chr5	176524693	G	T	0.00	0.01	0.02	0.80	0.95		utr_3	
TP53	chr17	7578406	C	T	0.00	0.01	0.02	0.80	0.97		missense	p.Arg175His
PTPRT	chr20	40743829	A	G	0.44	0.58	0.53	0.94	0.99	germline LOH	intronic	
ASXL1	chr20	31024274	T	C	0.50	0.58	0.57	0.94	1.00		synonymous	p.(=)
AKT1	chr14	105238820	G	C	0.56	0.49	0.53	0.93	0.99		intronic	
AURKA	chr20	54961541	A	T	0.41	0.45	0.46	0.06	0.00	germline LOH	missense	p.Phe31Ile
NUP214	chr9	134072992	G	C	0.54	0.47	0.40	0.07	0.01		missense	p.Gly1371Arg
LRP1B	chr2	141771058	T	A	0.51	0.57	0.48	0.19	0.04		intronic	

Figure 21 | Comparison of mutations in DNA of sorted 2N SCNA-negative populations, tumor-free control DNA and tumor populations. SNVs and selected SNPs are included to strengthen the fact that both SNPs and SNVs with VAF > 80% can be detected in sorted tumor populations.

4.1.4 Diploid and aneuploid cells coexist in multiploid tumors

To establish the multiparameter flow-sorting technology, we screened multiple human tumor samples, and in most cases the tumors consisted of only one tumor population. These were either solely diploid or aneuploid. However, in one LUAD (see Part B: 4.2.8) and three malignant melanomas, we found 2N marker-positive populations in otherwise aneuploid tumors (Figure 22a,b and Figure 23a). In contrast to the case described in 4.1.3, where the 2N marker-positive population was of non-tumor origin, all diploid, SOX10-positive (2N+) populations of these three multiploid malignant melanomas were characterized by multiple SCNAs (Figure 22c,d and Figure 23d).

In all three cases, shared SCNAs provided evidence for a clonal relationship between the 2N+ and the aneuploid tumor (AN+) populations. Indeed, in the first case, we did not detect a single breakpoint that was private to either the 2N+ or the AN+ population. Nonetheless, all tumor populations of this tumor were characterized by the same pattern of multiple focal gains and losses on chromosome 9p, comprising a homozygous deletion of *CDKN2A* (Figure 22e).

In contrast to this first case, private and shared SCNAs were present in the second and third tumor. This included a breakpoint on 12q21.1 that existed only in the 2N+ population (Figure 22f) of the second case. Additionally, two other SCNAs were either shared by the 2N+ and AN2+ populations (11q22.1) or by the two aneuploid tumor populations (5q21.1). Similarly, the 2N+ population of the third case was characterized by two SCNAs on 5q11 and 11q22 that were not present in the coexisting aneuploid tumor population (Figure 23e).

4.1.5 Clonal evolution deciphered by multiparameter flow sorting

The malignant melanoma of the above described third case was a relapse at the primary site (local relapse). We thought to investigate the clonal relationship of the diploid (2N+) and aneuploid (AN+) tumor populations in this relapse with two additional lesions that were available for this patient; a lymph node metastasis and a skin metastasis. In contrast to the local relapse, the metastatic lesions consisted of single tumor populations, and these were purely aneuploid, not diploid (Figure 23b,c). Array-CGH analysis revealed a relationship between the 2N+ population of the relapse and the lymph node metastasis: two SCNAs on 5q11 and 11q22 that were private to the 2N+ population of the local relapse were also present in the lymph node metastasis (Figure 23d,e). In addition, a heterozygous CNV on 11q12.1 remained heterozygous for the diploid tumor population of the relapse and the lymph node metastasis, but was completely lost (homozygous deletion) in both the AN+ population of the relapse and the skin metastasis (Figure 23e). Conversely, the SCNA on chr4q was absent in the lymph node metastasis, but present in all the other tumor populations (Figure 23e). The lymph node

metastasis carried in fact multiple focal amplifications scattered around chr4 (Figure 23e). To further investigate the relationship between all populations and to infer the clonal evolution in this tumor, we sequenced 28 genes (full list in Table S2) that are commonly mutated in melanoma, including *BRAF*, *NRAS*, *TP53*, and the recently reported *ARID2*[201], *GRIN2A*[202], *PREX2*[203], and *RAC1*[204].

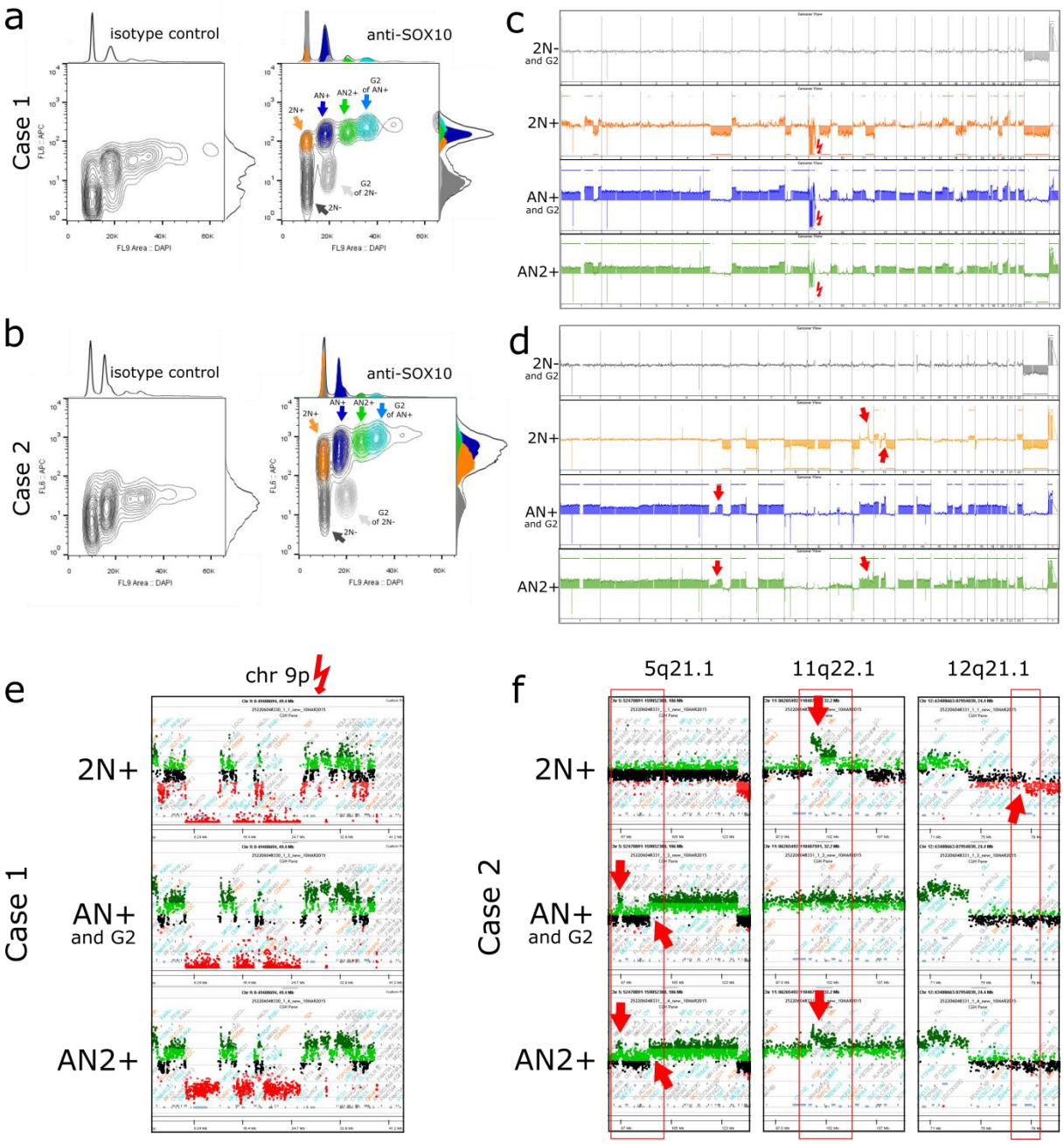


Figure 22 | Multiparameter nuclei flow sorting of multiploid malignant melanomas. Flow cytometry (a,b) and genome-wide copy number plots (c,d) of two cases. Selected chromosomal regions that revealed a relationship between the populations are enlarged for case 1 (e) or for case 2 (f). Red arrows hint towards genomic breakpoints that lead to SCNAs. The red lightning bolt (e) indicates the abundance of multiple deletions clustered on chr9p in case 1. Tumor populations designated as G2 (cell in G_2-phase) had the same aCGH profiles as their counterparts in G_1-phase and only one profile is therefore plotted.

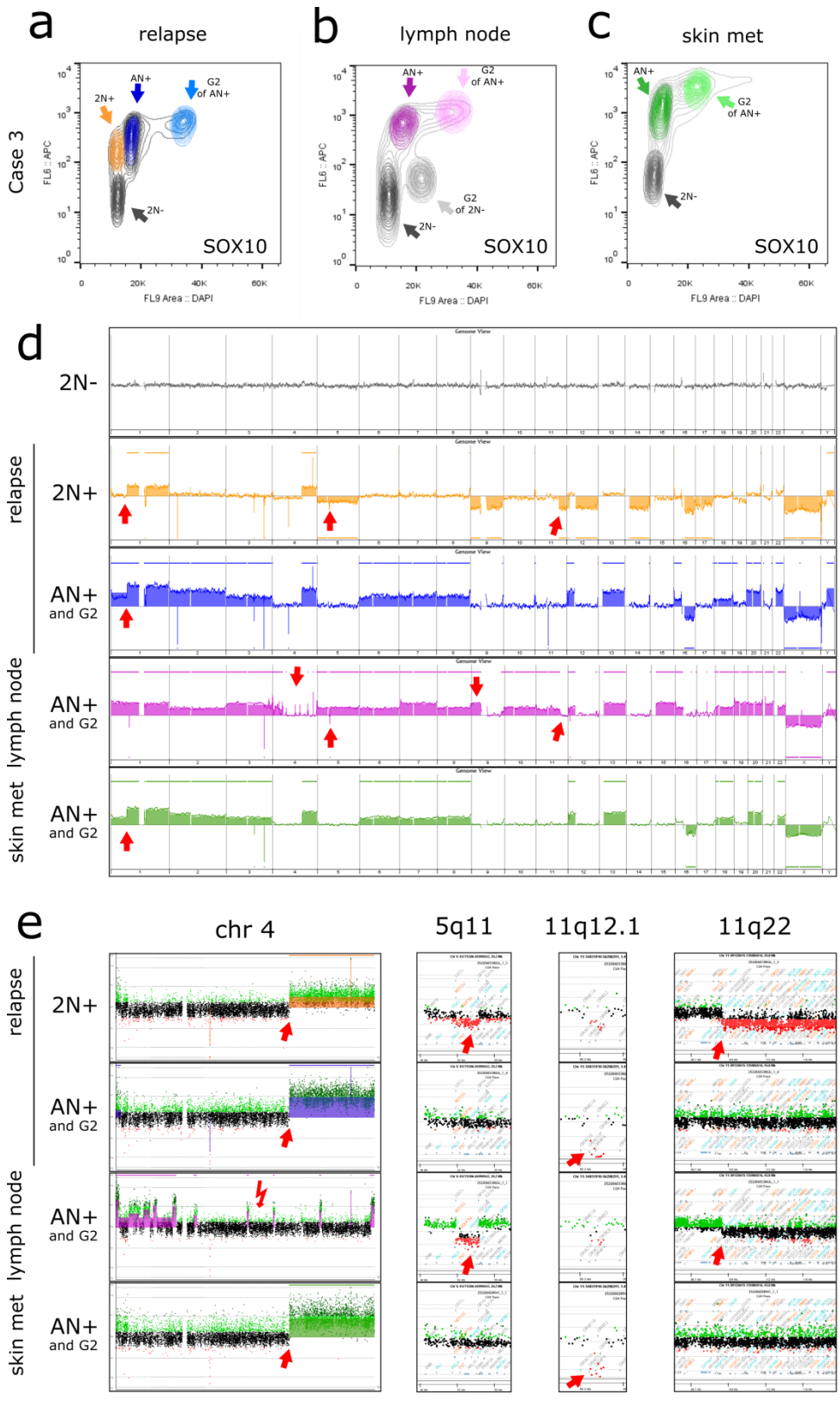


Figure 23 | Multiparameter nuclei flow sorting of three tumor lesions of a melanoma patient. Flow sorting of the local relapse (a), the lymph node metastasis (b), and the skin metastasis (c). Genome-wide copy number plots of all sorted tumor populations (d) and enlarged regions of selection aberrations (e) that show private but also shared copy number aberrations. Red arrows hint towards genomic breakpoints that lead to copy number aberrations. The red lightning bolt (e) indicates the abundance of multiple amplifications scattered across chr4 in the lymph node metastasis. Tumor populations designated as G2 (cell in G_2-phase) had the same aCGH profiles as their counterparts in G_1-phase and only one profile is therefore plotted.

The diploid, SOX10-negative population (2N-) from the relapse did not have any SCNAs and was therefore used as the germline control for the detection of SNVs and as the reference for SNPs. Sixteen SNVs were found, of which 10 were missense or nonsense mutations (Figure 24a). This included ubiquitous mutations of *TP53* (VAF: 93-100%) and the common melanoma driver mutation NRAS Q61R[70] (VAF: 32-73%). Interestingly, the VAF of the *NRAS* mutation was the highest in the lymph node metastasis. NRAS is located on chr1 and copy number analysis of the *NRAS* locus did not detect the chromosomal break on chr1 for the lymph node metastasis as compared to the other three tumor populations of the local relapse and the skin metastasis. This indicates a different evolutionary event on chr1 for the lymph node metastasis which caused the observed increase of *NRAS*-mutant alleles.

Another informative locus was chr9p. The 2N+ population of the local relapse had only one copy of this chromosomal region, whereas the aneuploid population of the relapse and the skin metastasis had two copies. The lymph node metastasis even possessed four copies of this locus (Figure 23e). This is of interest because this region contains the gene *PTPRD,* and one SNV and multiple heterozygous SNPs of *PTPRD* were detected in this tumor. *PTPRD* mutations had a VAF of 100% in all tumor populations but the lymph node metastasis. Here, the VAFs were around 50% for all the SNPs and the SNV (Figure 24a). Taken together, we proposed an evolution where most of the mutations, including *PTPRD*, happened early and some accumulated at a VAF of 100% due to LOH, as seen for *ARID2* on chr12 and *TP53* on chr17. Furthermore, exemplified by the *PTPRD* mutation and the CNV on 11q21.1, the combined copy number and mutation data suggest that a common precursor contained one mutated *PTPRD* allele and the heterozygous loss of 11q22-11q25. Following additional losses of whole chromosomes or chromosome parts, the aneuploid tumor populations might have eventually resulted from a WGD of either a cell from the diploid tumor population of the local relapse or a common diploid progenitor cell (Figure 24b), a hypothetical model that is in agreement with a branched evolution of the lymph node and the skin metastasis.

a

					control	local relapse		lymph node	skin met			
gene	chrom	pos	ref	var	2N SOX10-	2N SOX10+	AN SOX10+	AN SOX10+	AN SOX10+	Type	effect	Protein change
ARID2	chr12	46230691	C	T	0.05	0.92	0.99	0.98	1.00	somatic LOH	missense	p.Arg314Cys
TP53	chr17	7578201	CA	AT	0.05	0.92	0.99	0.95	1.00		missense	p.Val216Asp
TP53	chr17	7578212	G	A	0.05	0.93	0.99	0.96	1.00		nonsense	p.Arg213Ter
PTPRD	chr9	8486022	G	A	0.07	0.95	1.00	0.51	0.99	somatic LOH/ somatic	missense	p.Ser932Leu
ERBB4	chr2	212483938	C	T	0.03	0.54	0.51	0.44	0.59	somatic	synonymous	p.(=)
ERBB4	chr2	212252636	C	T	0.06	0.52	0.49	0.68	0.44		intronic	
ERBB4	chr2	212530199	G	A	0.04	0.51	0.51	0.35	0.53		missense	p.Pro574Ser
KIT	chr4	55597561	G	A	0.03	0.47	0.54	0.54	0.50		missense	p.Asp737Asn
ERBB4	chr2	212488624	G	A	0.04	0.47	0.48	0.39	0.47		intronic	
PREX2	chr8	68934283	C	T	0.02	0.44	0.66	0.40	0.67		missense	p.Arg117Cys
GRIN2A	chr16	9943693	G	A	0.04	0.43	0.35	0.57	0.32		synonymous	p.(=)
ROS1	chr6	117710780	C	T	0.06	0.42	0.29	0.36	0.35		missense	p.Asp498Asn
SNX31	chr8	101620805	G	T	0.04	0.41	0.34	0.56	0.29		intronic	
ROS1	chr6	117639344	G	A	0.03	0.39	0.31	0.28	0.35		intronic	
NRAS	chr1	115256529	T	A	0.05	0.32	0.42	0.73	0.41		missense	p.Gln61Leu
PREX2	chr8	69129915	C	T	0.04	0.43	0.64	0.00	0.58		missense	p.Arg1557Cys
BRAF	chr7	140476936	G	A	0.47	0.39	0.63	0.45	0.46	germline	intronic	
PREX2	chr8	69136871	T	A	0.52	0.44	0.28	0.58	0.36		intronic	
TACC1	chr8	38677932	A	G	0.50	0.45	0.31	0.55	0.37		synonymous	p.(=)
TACC1	chr8	38684853	T	C	0.54	0.47	0.33	0.59	0.30		synonymous	p.(=)
BRAF	chr7	140449150	T	C	0.48	0.50	0.64	0.46	0.61		synonymous	p.(=)
PREX2	chr8	69143589	A	G	0.50	0.50	0.32	0.58	0.34		synonymous	p.(=)
BRAF	chr7	140449071	C	G	0.46	0.50	0.65	0.51	0.64		intronic	
SNX31	chr8	101661634	C	T	0.57	0.52	0.30	0.50	0.33		intronic	
BRAF	chr7	140500135	C	T	0.49	0.54	0.64	0.55	0.65		intronic	
BRAF	chr7	140477756	C	G	0.53	0.57	0.68	0.47	0.69		intronic	
PTPRD	chr9	8436703	C	G	0.56	0.97	0.99	0.44	1.00	germline LOH/ germline	intronic	
PTPRD	chr9	8465598	A	G	0.60	0.97	0.99	0.48	1.00		synonymous	p.(=)
PTPRD	chr9	8454566	T	C	0.56	0.98	1.00	0.48	0.99		intronic	
PTPRD	chr9	8733894	G	A	0.49	1.00	1.00	0.48	1.00		utr_5	
PTPRD	chr9	8389364	C	G	0.50	0.03	0.01	0.48	0.00	germline LOH/ germline	synonymous	p.(=)
DKN2A	chr9	21968199	C	G	1.00	0.99	0.99	0.97	1.00	germline homozygous	utr_3	
ROS1	chr6	117678083	A	G	1.00	1.00	1.00	1.00	1.00		intronic	
SNX31	chr8	101661603	T	G	1.00	1.00	1.00	1.00	1.00		intronic	

b

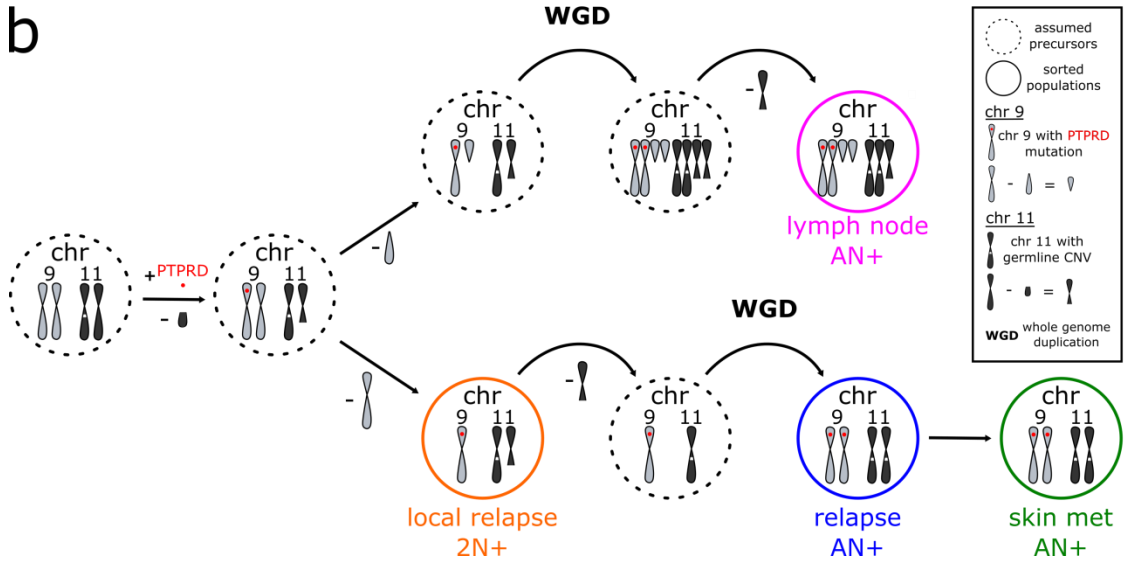


Figure 24 | Clonal evolution of a malignant melanoma. (a) List of mutations detected in all populations with a targeted melanoma panel. Note that all mutations except the *PREX2* mutation are ubiquitous. This includes 13 SNVs with VAFs between 30% and 70% and four SNVs with VAFs close to 100%, which indicates LOH at the loci of these SNVs. In addition all detected germline SNPs are shown. Of interest in this context is the locus of *PTPRD* (bold) with one SNV and five SNPs. (b) Hypothetical model of the clonal evolution of this tumor exemplified based on the somatic *PTPRD* mutation and the LOH of chr11q (Figure 23), as revealed from a CNV at 11q12.1 and a SCNA at 11q22. Data are in agreement with the assumption that all aneuploid tumor populations evolved from diploid precursors due to WGDs.

4.2 Results Part B: Genome-wide copy number and mutational analysis in longitudinal biopsies of matched primary and metastatic lung adenocarcinomas using a multiparameter flow-sorting approach

4.2.1 Overview of the project

This project aimed to investigate temporal and spatial ITH between primary tumors and matched metastases to study clonal tumor evolution in LUAD. Patients were included if at least two longitudinal fresh-frozen LUAD biopsies were available in the biobank of the IfP, and one of these was the primary tumor. Nineteen patients met these inclusion criteria (Figure 25).

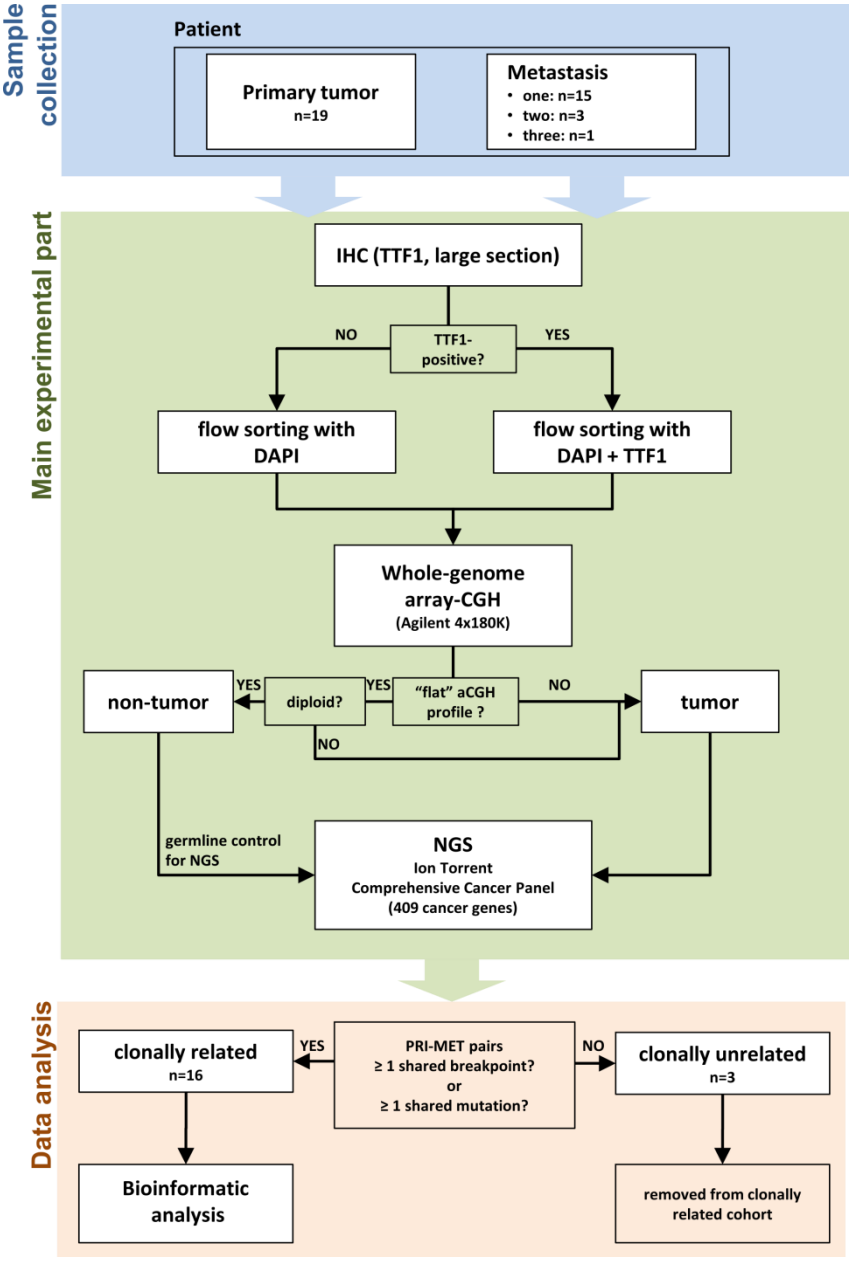


Figure 25 | Overview of the project

Additional metastases were included, if available. TTF1 expression was validated on large sections of each tumor. TTF1-positive tumors (n=25, 58%) were detected in 11 patients (58%) and subjected to multiparameter flow sorting with DAPI for DNA content (ploidy) and TTF1 to enrich for tumor cells. Tumors without TTF1 expression (n=18, 42%) were found in eight patients (42%) and sorted with DAPI for ploidy only. The tumors of 3 of the 19 patients that were initially included did not share a single breakpoint or a single SNV in aCGH and NGS analysis, respectively (Figure S6). The suspected metastases of these three patients were therefore considered second primary tumors and were therefore removed from this study of clonally related primary-metastatic LUADs. The final cohort included 12 primary-metastasis duos, three primary-metastasis trios, and one patient with a primary tumor and three metastases. This entailed patients with brain (n=4), intrapulmonary (n=4), adrenal gland (n=3), kidney (n=1), and thyroid gland (n=1) metastases, and with pleural effusions (n=3). A graphical summary of the localization of the primary tumors and their clonally related metastases is provided in Figure S7. The detailed composition of the cohort, including clinical annotation, is summarized in Table 2.

Table 2 | Characteristics of 16 patients with clonally related LUADs

Patient	Sex	Age at diagnosis (years)	Smoker at diagnosis	smoking pack-years	Tumor Type	Radio-therapy	Chemo-therapy	Delta time PRI-MET (months)	Metastatic site(s)	Ploidy(N) PRI	Ploidy(N) MET
12	M	58	NA	NA	LLL	NA	NA	2	thyroid gland	3.3	3.2
20	F	40	active	25	LUL	No	No	41 41	kidney, left kidney right	2.6	2.6 2.6
22	F	59	NA	NA	LUL	No	No	7 11	brain, right (parietal) Locoregional: LLL	3.4	3.4 3.5
31	F	64	NA	NA	RUL	No	No	64 64	Locoregional: LLL Locoregional: RLL	3.0	3.1 3.3
34	M	60	former	45	RUL	No	No	27	Locoregional: RLL	2.0	2.0
35	F	52	NA	NA	RLL	No	No	10	brain, left (frontal)	3.3	3.1
36	M	51	NA	NA	LUL	No	No	11 28 28	Locoregional: RLL Locoregional: LLL Locoregional: RLL	2.1	2.0 2.0 2.1
38	M	66	former	45	LUL	Yes	Cisplatin, Taxotere, Erbitux	14	adrenal gland, left	3.2	3.1
39	M	78	former	30	RLL	No	No	9	Locoregional: RUL	3.2	3.2
41	F	64	active	40	LUL	No	No	0	brain, left (parietal)	2.9	3.0
42	F	81	former	1-2	RLL	No	No	7	pleura, right	2.1 3.4 4.7	3.8
44	M	51	active	35	LUL	No	Cisplatin, Alimta	1	adrenal gland, left	5.4	5.0
46	F	74	NA	NA	RUL	No	Carboplatin, Taxotere	44	pleura, right	3.2	2.8
48	M	67	active	50	LUL	No	No	0	adrenal gland, left	4.1	4.7
49	M	76	former	20	RUL	No	No	152	pleura, left	2.0	2.0
50	M	57	never	0	RUL	No	Cisplatin, Taxotere	22	brain, left (parietal)	3.2	3.0

Abbreviations: LLL, left lower lobe; LUL, left upper lobe; RLL, right lower lobe; RUL, right upper lobe; NA, unknown

4.2.2 Multiparameter flow sorting results in tumor DNA of high purity

As recently demonstrated, the availability of high purity tumor samples is fundamental for not overestimating ITH in genomic studies.[6] All tumor specimens (n=37) from the 16 patients of the final cohort were therefore subjected to flow sorting with DAPI for DNA content (n=12 tumors, 6 patients) or additionally with the lineage marker TTF1 (n=25 tumors, 10 patients, Figure 26). The resulting 39 sorted tumor populations were genomically profiled by aCGH to determine genome-wide SCNAs and by deep sequencing of 409 cancer-relevant genes to detect SNVs (mean coverage 965x). As expected, diploid, TTF1-negative populations carried no SCNAs (Figure 28a). Similarly, diploid populations of tumors that were sorted by DNA content only did not have any SCNAs (data not shown). We confirmed the absence of SNVs in these diploid populations in three patients with matched non-tumor tissue (see Results Part A 4.1.3 and Figure S5). Consequently, diploid populations without any SCNAs were considered to be of non-tumor origin and served as germline controls for the NGS analyses.

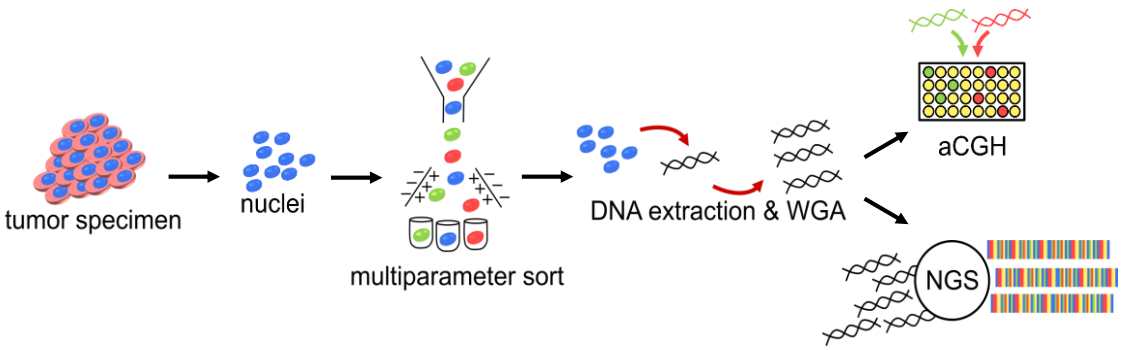


Figure 26 | Workflow of the multiparameter flow-sorting approach and genomic characterization of the sorted tumor cell populations. This approach comprises the isolation of tumor nuclei from both fresh-frozen and FFPE tissues, followed by a multiparameter sort with DAPI for DNA content and TTF1 for tumor cells. Subsequently, DNA of sorted populations was extracted and subjected to WGA with Phi29, followed by detection of SCNAs and SNVs with aCGH and NGS, respectively.

In four patients (25%), TTF1-positive tumor populations of diploid DNA content (2N±0.2) were detected. The tumors of all other patients were aneuploid (>2.2N, Table 2). Almost all tumors (97%) consisted of tumor cells of one single ploidy. Only the primary tumor of patient 42 showed more than one tumor population in the multiparameter flow cytometry analysis (Figure 38a). The application of this flow-sorting approach increased the mean tumor purity from 54% (range: 7-89%) of unsorted material to 92% (range: 79-99%) as evaluated by EXPANDS[205] (Figure 27). This is of particular importance for the genomic characterization of biopsies with low tumor cell proportions and of diploid tumors (Figure 28). Indeed, by using this multiparameter sorting approach, we achieved a tumor cell content of up to 90% in these sorted diploid tumor populations.

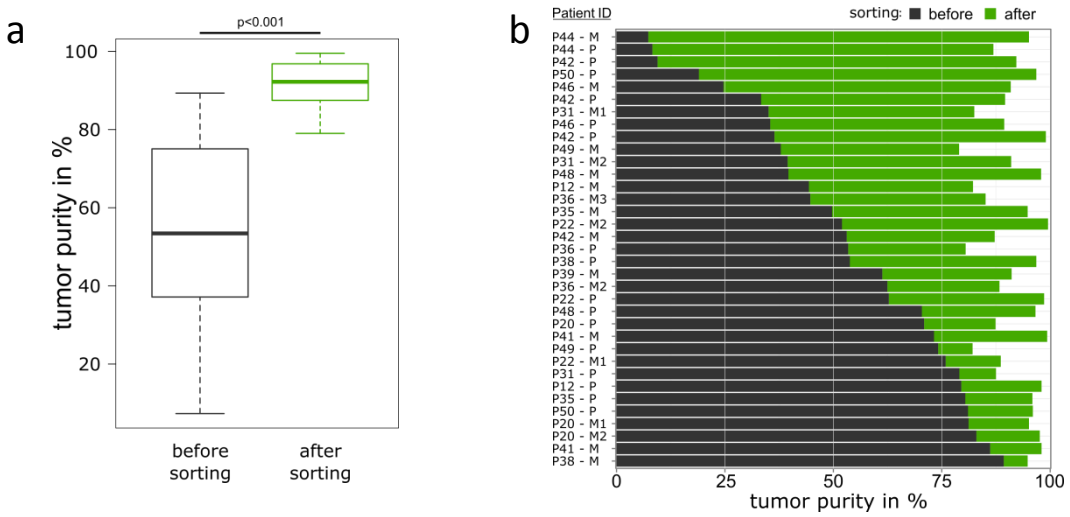


Figure 27 | Purity estimation before and after sorting. (a) Box plots show the median, first quartile, third quartile, minimum, and maximum of purity before and after sorting. (b) The bar plot illustrates the increase in purity because of sorting per sample (n=35). *Note: Four out of the 39 sorted tumor populations were excluded because EXPANDS could not predict purity due to a low number of SCNAs and SNVs.*

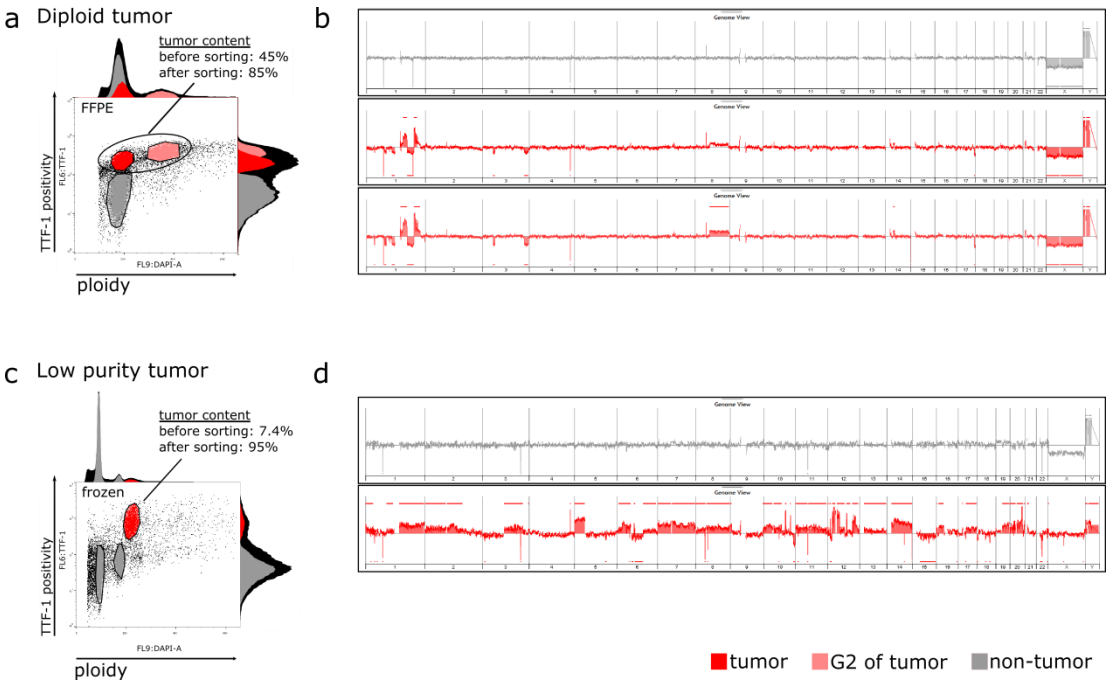


Figure 28 | Multiparameter nuclei flow sorting of diploid tumor or low purity tumors. The diploid tumor (a) was separated from diploid, normal nuclei by the addition of TTF1 (y-axis) as a second parameter extra to DAPI (DNA content, x-axis). Second, a tumor with low tumor cell content (c) was sorted and purity was increased from 7.4% to 95%. Diploid normal nuclei (grey) displayed no SCNAs except of gender-specific differences of X and Y chromosomes (b,d). The TTF1-positive populations carried distinct SCNAs (red).

4.2.3 Recurrent mutations and copy number aberrations

Targeted sequencing 39 tumor populations identified 318 pathogenic mutations affecting 96 of the 409 investigated genes. An overview of genes recurrently mutated or affected by SCNAs is displayed in Figure 30. The most common ubiquitous mutations were found in genes previously known to be mutated in LUAD.[53,206] These includes *KRAS* in six patients (37.5%); *TP53* in five patients (31.3%); *EGFR, ATM,* and *STK11* in four patients (31.3%); and *CSMD3, EPHA7, KEAP1, NF1,* and *NOTCH4* in three patients (18.8%). As expected, mutations with LOH were mainly detected in TSGs such as *ATM, KEAP1, NF1, PTPRD, STK11,* and *TP53.* As a matter of fact, ubiquitous mutations of *KEAP1, STK11* and *TP53* always occurred in the context of LOH.

Recurrent copy number gains involved the chromosomal regions 1q, 5p, 7p, 8q, and both arms of chromosomes 12 and 14 (Figure 29). This was in line with previous data[207] that mapped these regions to genes known to be amplified in LUAD, such as *ARNT* (1q21), *TERT* (5p15), *EGFR* (7p11), *MYC* (8q24), *KRAS* (12p12), *NKX2-1, and FOXA1* (both 14q13). Chromosome 9p21 was the only recurrently deleted region, with both homozygous (7 patients|43.8%) and heterozygous deletions (4|25%). Homozygous deletions are of general interest, because they cause an irreversible loss of genes, therefore indicating TSG function. In addition to the genes located on chr9p21, we detected only seven additional genes with homozygous deletions: *TP53*, *RB1, PTEN, DCN, THSD4, HS3ST4,* and *WWOX*. Concordantly, all of them have previously reported tumor suppressor activity. [208–211]

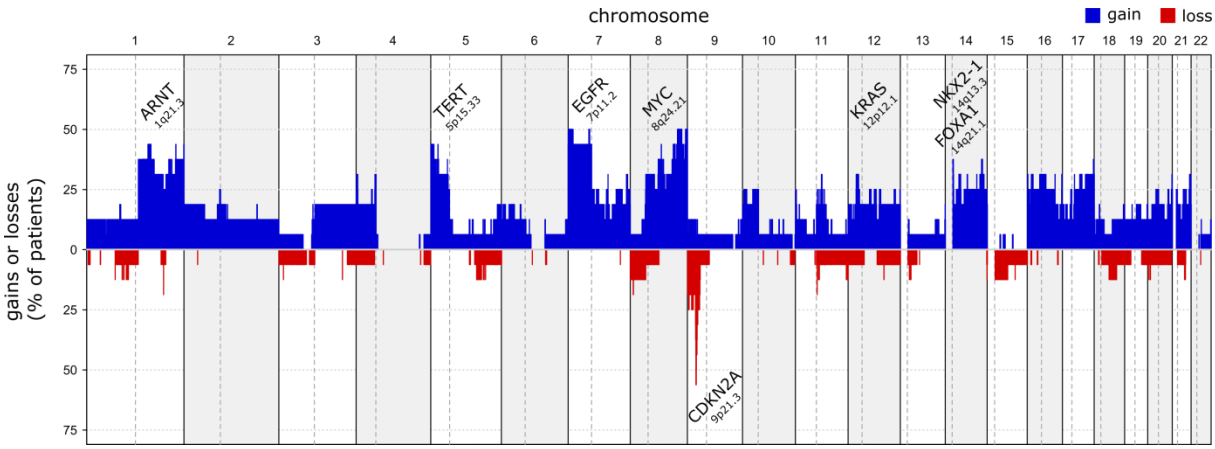


Figure 29 | Overview of recurrent amplifications and deletions. (a) Six regions were amplified in >30% of the patients (1q, 5p, 7p, 8q, chr12, chr14). Only one region (9p21) was recurrently (>30% of patients) deleted in this cohort. Gains, blue; losses, red. *Note: To exclude over- or underrepresentation, because of different numbers of samples per patient, each aberration was counted only once per patient.*

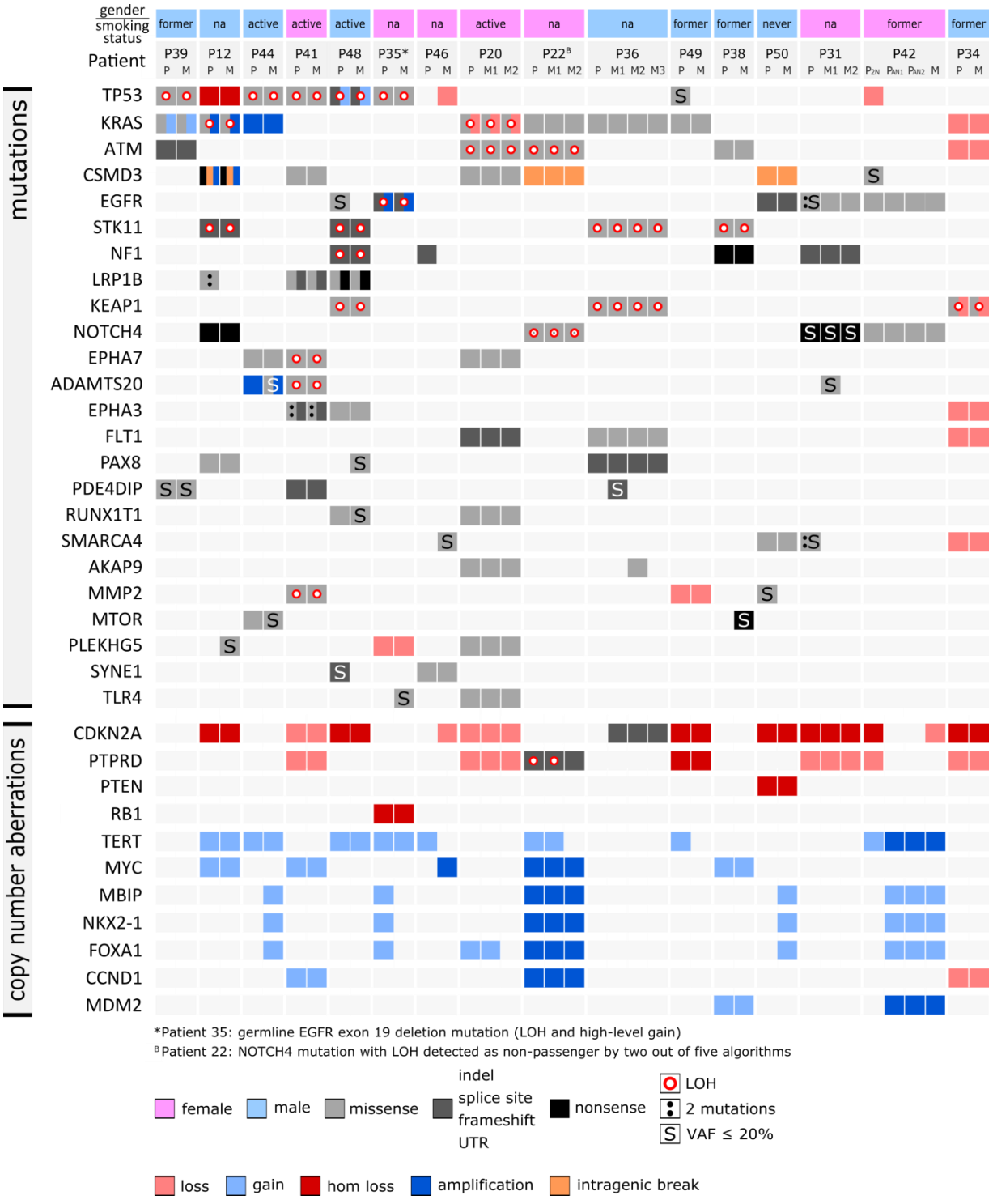


Figure 30 | Overview of genes recurrently affected by SNVs and SCNAs. Overall, a high proportion of SNVs and SCNAs were ubiquitously affected in all samples of individual patients. Many TSGs *(TP53, ATM, STK11, PTPRD)* were mutated with the complete loss of the wild-type allele. Of note, *KEAP1* was mutated in three patients, was always ubiquitous, and occurred in the context of LOH in all samples.

4.2.4 Stable ploidy and high concordance of copy number aberrations and cancer gene mutations

Flow cytometry revealed that the ploidy of individual tumors was stable over time regardless of the time interval between the primary tumor and occurrence of the metastasis (Figure 31, p=0.89, Wilcoxon signed-rank test). Likewise, the metastasis of patient 49 remained diploid, despite the fact that it emerged 12.7 years after the detection of the diploid primary tumor. A high ploidy was associated with a short time until emergence of the metastatic tumors (p<0.05, univariate cox analysis, data not shown) and correlated with the total amount of SCNAs (r=0.79, p<0.001, Figure 32a). Moreover, tumors with biallelic inactivation of *TP53* or *ATM* had a higher burden of SCNAs (p=0.003, Wilcoxon signed-rank test) and higher ploidy (p=0.001, Wilcoxon signed-rank test) than the rest of the cohort. Of note, all diploid tumors were negative for *TP53* or *ATM* mutations and also had the lowest numbers of SCNAs.

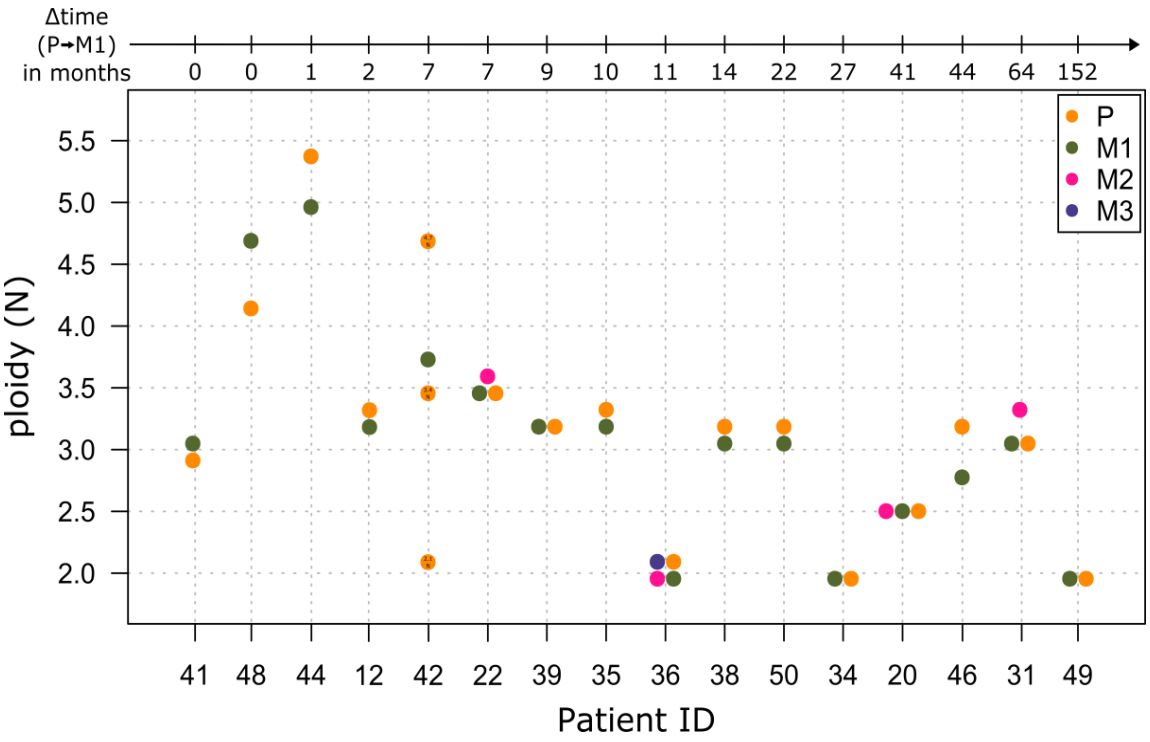


Figure 31 | Ploidy of tumors per patient. Ploidy of individual tumors did not change significantly (p=0.89, Wilcoxon signed-rank test), over time as calculated for the time interval between primary tumor detection and the emergence of the first metastasis (Δtime P -> M1). P, primary tumor; M1-3, first, second, third metastasis. *Note: Patient 42 had a multiploid primary tumor and is excluded from this analysis.*

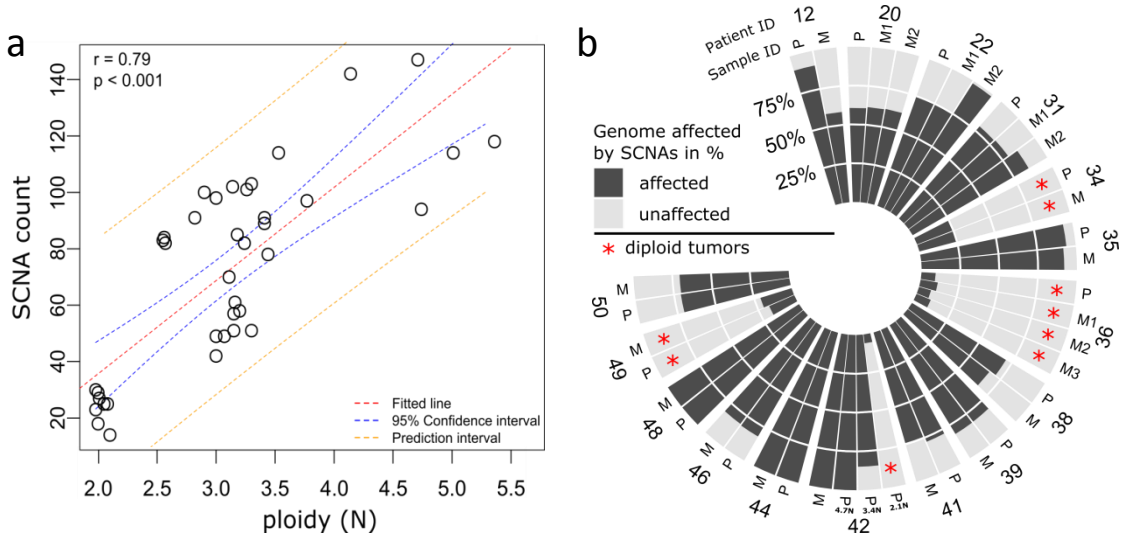


Figure 32 | Scatter plot of ploidy and SCNA count (a) and Circos plot of the percentage of genome affected by SCNAs (b). (a) Ploidy correlated significantly (linear regression, r=0.79, p<0.001) with the burden of SCNAs. The number of SCNAs therefore increases with the ploidy of the tumor. (b) Circos plot showing the relative amount of the genome that is affected by SCNAs per sample.

An overview of the SCNAs on a per-sample basis is depicted in Figure 33. Tumors from one patient were more similar to each other than to the tumors of other patients, as evaluated by hierarchical clustering of Euclidean distances based on SCNAs (Figure S8). One exception was the diploid primary tumor population (2.1N) of patient 42, which clustered together with the diploid tumor populations of the other three patients with diploid tumors probably due to the overall large number of copy number neutral regions (percentage of genome unaffected by SCNAs) of all diploid tumors in this cohort (Figure 32b).

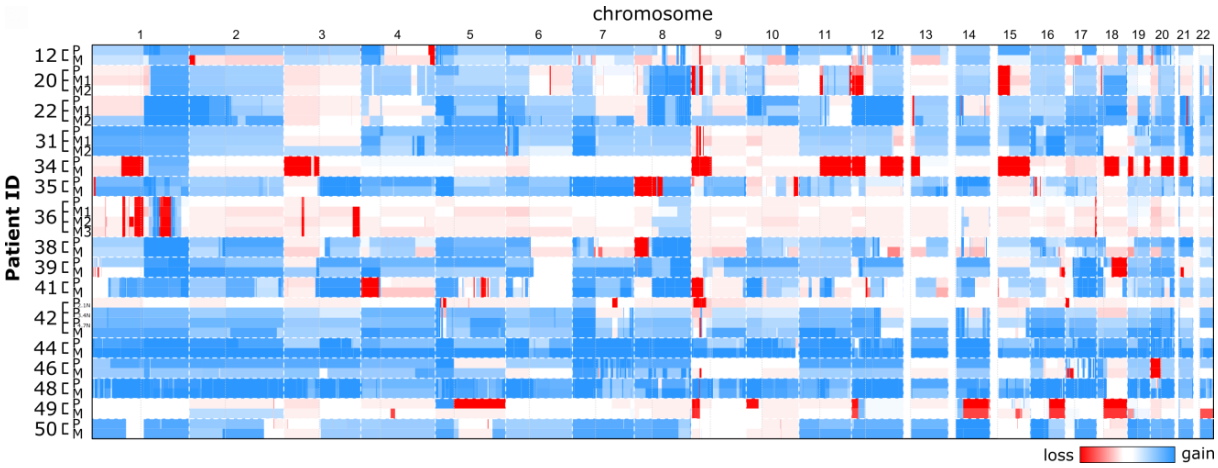


Figure 33 | Overview of genome-wide SCNAs per tumor population. Sex chromosomes were excluded. Red, loss; blue, gain; white, copy number neutral. P: primary tumor, M1-3: metastases.

No significant difference in the number of additional non-ubiquitous (private and shared) mutational events (SCNAs and SNVs) were detected in the primary tumors or the metastases (median 12.8 vs. 13.6 non-ubiquitous events per tumor, respectively; p=0.83, Wilcoxon signed-rank test), regardless that the time between primary and metastatic tumor detection was different in each patient (Figure 34a). Most importantly, 80% of all detected SCNAs were ubiquitously present in the primary tumor and metastases (Figure 34b). Moreover, 88% (range: 68%-100%) of SCNAs were passed from the primary tumor on to the metastases. Likewise, 87% (range: 57%-100%) of SCNAs detected in the metastases were shared with the primary tumor.

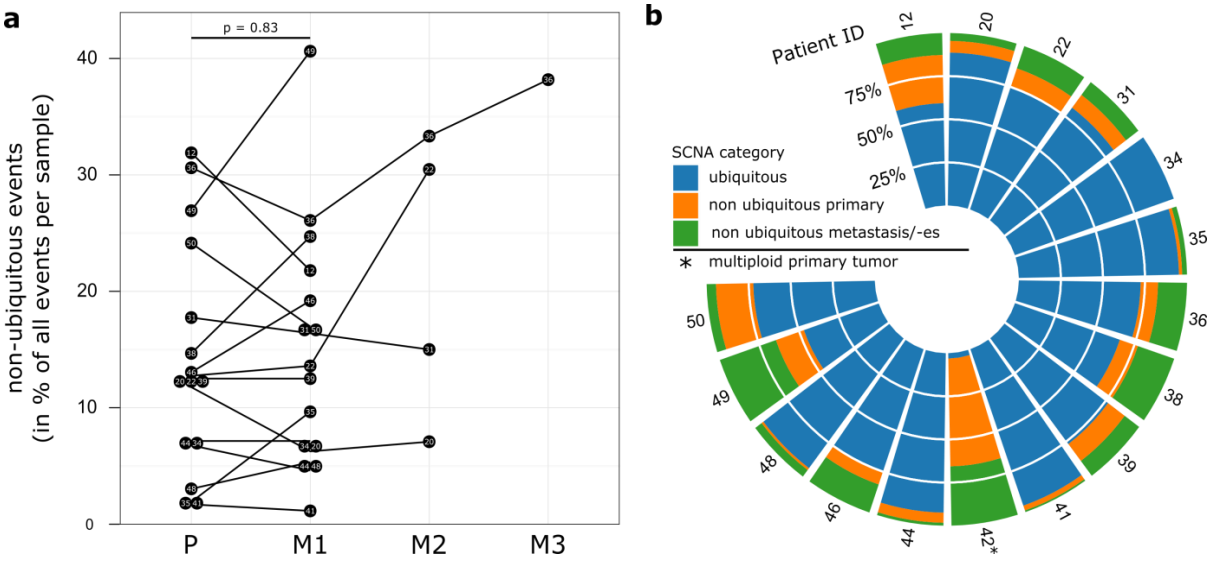


Figure 34 | Relative proportion of SCNAs and SNVs per sample and patient. (a) Proportion of non-ubiquitous (private and shared SNVs and SCNAs) per sample and patient. Metastases did not have higher amounts (p=0.83, Wilcoxon signed-rank test) of non-ubiquitous SCNAs and SNVs than primary tumors. (b) Circos plot showing the relative proportion of SCNAs per category (ubiquitous, primary only, metastasis/es only) per patient. P, primary tumor; M1-3, metastases.

Similar concordance rates were found for SNVs. Here, 78% (range: 50%-100%) of mutations in the primary tumor were propagated to the metastases, and 73% (range: 33%-100%) of SNVs in the metastases were shared with the primary tumor. The VAFs of these ubiquitous mutations were significantly higher than of mutations that were private to only the primary or metastatic tumor (Figure S9; Figure S12). Most importantly, 86.3% of all pathogenic mutations in the top 11 mutated genes *(TP53, KRAS, ATM, CSMD3, EGFR, STK11, NF1, LRP1B, KEAP1, NOTCH4 and EPHA7)* were ubiquitous. This value was increased to 93.3%, if only mutations with VAF > 20% were considered. Yet, one mutation in *TP53* (patient 49, VAF: 14%) and *NF1* (patient 46, VAF: 49%), and two mutations in *LRP1B* (both patient 12, both VAFs of 33%), each private to the primary tumors were not propagated to their metastases (Figure 30). For *EGFR*, we detected two mutations that occurred in primary tumors only. One was an exon 21 mutation (H850Y) with a VAF of 4% in patient 31 who

already had the common EGFR driver mutation L858R (VAF: 63%). A double-mutation of EGFR H850Y with another exon 21 mutation has been recently reported[212], but its effect on EGFR protein function is unknown. Another private *EGFR* mutation occurred in patient 48. This mutation (EGFR D379N, VAF: 6%) occurred in exon 10 and is of unknown significance.

4.2.5 High concordance of absolute copy numbers results in stable VAFs of ubiquitous mutations

Next, the distribution of the VAFs and their change over time in primary tumors and their first biopsied metastases was investigated. Interestingly, VAFs of ubiquitous mutations remained largely unchanged (r=0.93, p<0.001, linear regression) between primary tumors and metastases (Figure 35a), irrespective of the differences in time intervals between the tissue resections per patient (Figure S10). The same was true for absolute copy numbers (Figure S11). Furthermore, this resulted in a density distribution pattern of VAFs and copy numbers that indicated a truncal origin for ubiquitous mutations (Figure 35b; Figure S12). For instance, the majority of mutations with a copy number of three clustered around VAFs of 33% or 66%, indicating the presence of one or two mutated alleles, respectively. Similarly, mutations with a copy number of four displayed a high density at VAFs around 25% and 50%. Overall, 94% of mutations with a VAF ≤ 10%, a value indicating subclonality, were not spread to the metastases. Conversely, 92% of mutations with VAF ≥ 50% in the primary tumors and 97% of mutations with VAF ≥ 50% in the metastases were ubiquitously present in all biopsies, further supporting their truncal nature.

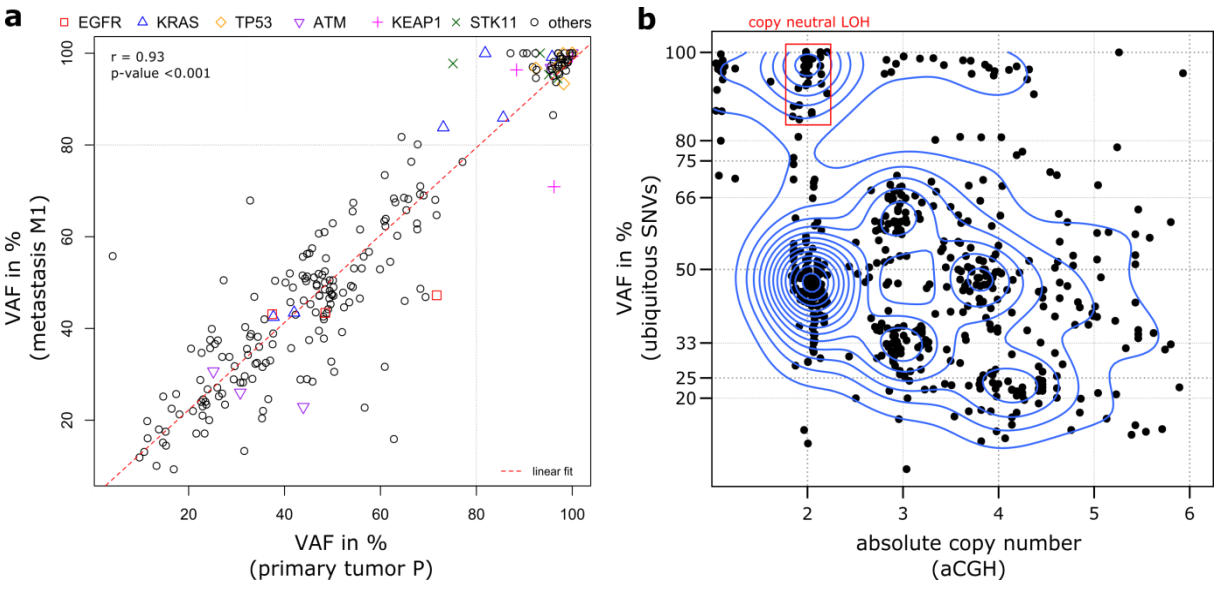


Figure 35 | VAFs and copy numbers of ubiquitous mutations. (a) Scatter plot of the VAFs of ubiquitous mutations in primary tumors (P, x-axis) versus the first biopsied metastasis (M1, y-axis) revealed a significant correlation (linear regression, r=0.93, p<0.001). (b) Density plot of the absolute copy number (x-axis) and VAF (y-axis) of ubiquitous mutations, indicating their truncal origin.

4.2.6 Genetic divergence between primary tumors and metastases reveals two patterns of evolution

Metastatic dissemination follows the two general models of "linear progression" or "parallel progression", which are based on the relative timing of emergence and genetic divergence between the primary tumor and its metastases.[144] The mean pairwise divergence (D)[213] is a measurement of the genetic divergence between primary/metastatic (termed P/M divergence) pairs. A broad continuum of these two progression models was detected in this cohort (Figure 36). The P/M divergence ranged from 0.03-0.51. At the lower end of this distribution was the rather linear progressing tumor of patient 41. Here, 97% of all SCNAs and SNVs were present in both the primary tumor and the brain metastasis. The highest P/M divergence was detected in patient 49 with a contralateral pleural effusion that had emerged 12.7 years after the detection of the primary tumor. Both shared only 49% of all SCNAs and SNVs, indicating a parallel progression. The median time interval between primary tumor detection and the first biopsied metastasis was 10.7 months. Removing patient 49 as an outlier from the analysis lead to the result that the genetic P/M divergence did not correlate with the relative timing to metastasis (r=0.2, p=0.42). Notably, tumors with *TP53* or *ATM* mutations displayed a low P/M divergence and a tendency toward early emergence of metastases. However, these were only significant (p=0.042 for P/M divergence and p=0.006 for time to first metastasis, Wilcoxon signed-rank tests) for tumors with biallelic inactivation of *TP53* or *ATM*.

4.2.7 Two patterns of metastatic spread

Multiple, spatially separated metastases were available for four patients (patients 20, 22, 31 and 36). To investigate the directionality of metastatic spread in such oligometastatic settings, hierarchical clustering of Euclidean distances between primary tumors and their metastases was performed in these patients. The results are supportive for two patterns of metastatic evolution in LUAD (Figure 37): (i) polyclonal spread from the primary tumor and (ii) monoclonal metastasis-to-metastasis spread.

The primary tumor of patient 22 was more closely related to the brain metastasis than to the intrapulmonary (of a different lobe in the ipsilateral lung) metastasis. Besides the fact that *KRAS*, *ATM,* and *PTPRD* were mutated in all three tumor manifestations, biallelic inactivation of PTPRD, a TSG that is frequently inactivated in glioblastoma multiforme and malignant melanoma[214], was detected only in the primary tumor and the brain metastasis. Conversely, a mutation in VEGFR3 (gene: *FLT4*), a TK that is implicated in both the RTK/RAS/RAF and PI(3)K-mTOR pathways,

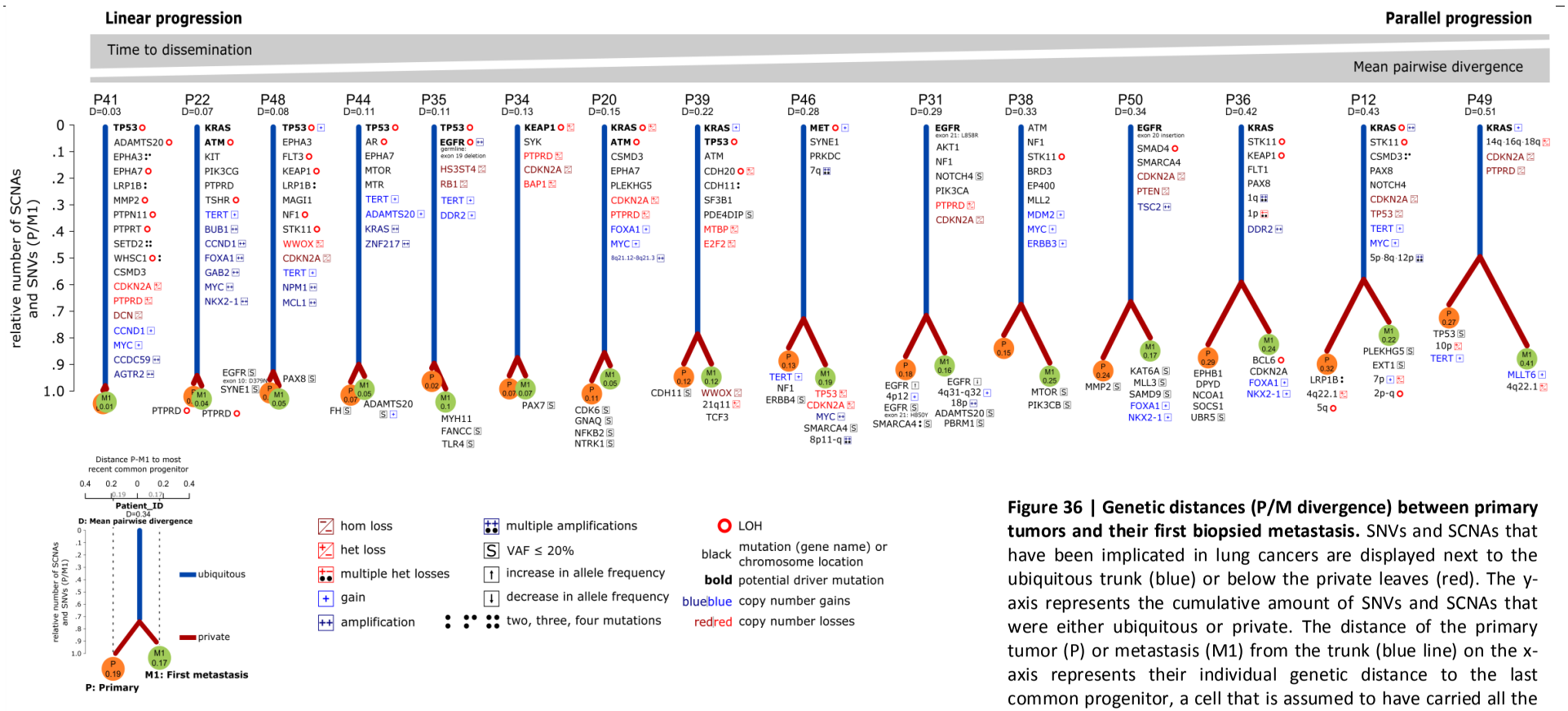


Figure 36 | Genetic distances (P/M divergence) between primary tumors and their first biopsied metastasis. SNVs and SCNAs that have been implicated in lung cancers are displayed next to the ubiquitous trunk (blue) or below the private leaves (red). The y-axis represents the cumulative amount of SNVs and SCNAs that were either ubiquitous or private. The distance of the primary tumor (P) or metastasis (M1) from the trunk (blue line) on the x-axis represents their individual genetic distance to the last common progenitor, a cell that is assumed to have carried all the ubiquitous alterations before the metastatic founder cell or clone disseminated from the primary tumor. A broad continuum of linear to parallel progression is present in this cohort. *Note: Patient 42 with a multiploid primary tumor is excluded from this analysis. Interestingly, despite the highest P/M divergence in this cohort, the primary tumor of patient 49 was diploid, and the metastasis remained diploid when it emerged 12.7 years after the primary tumor.*

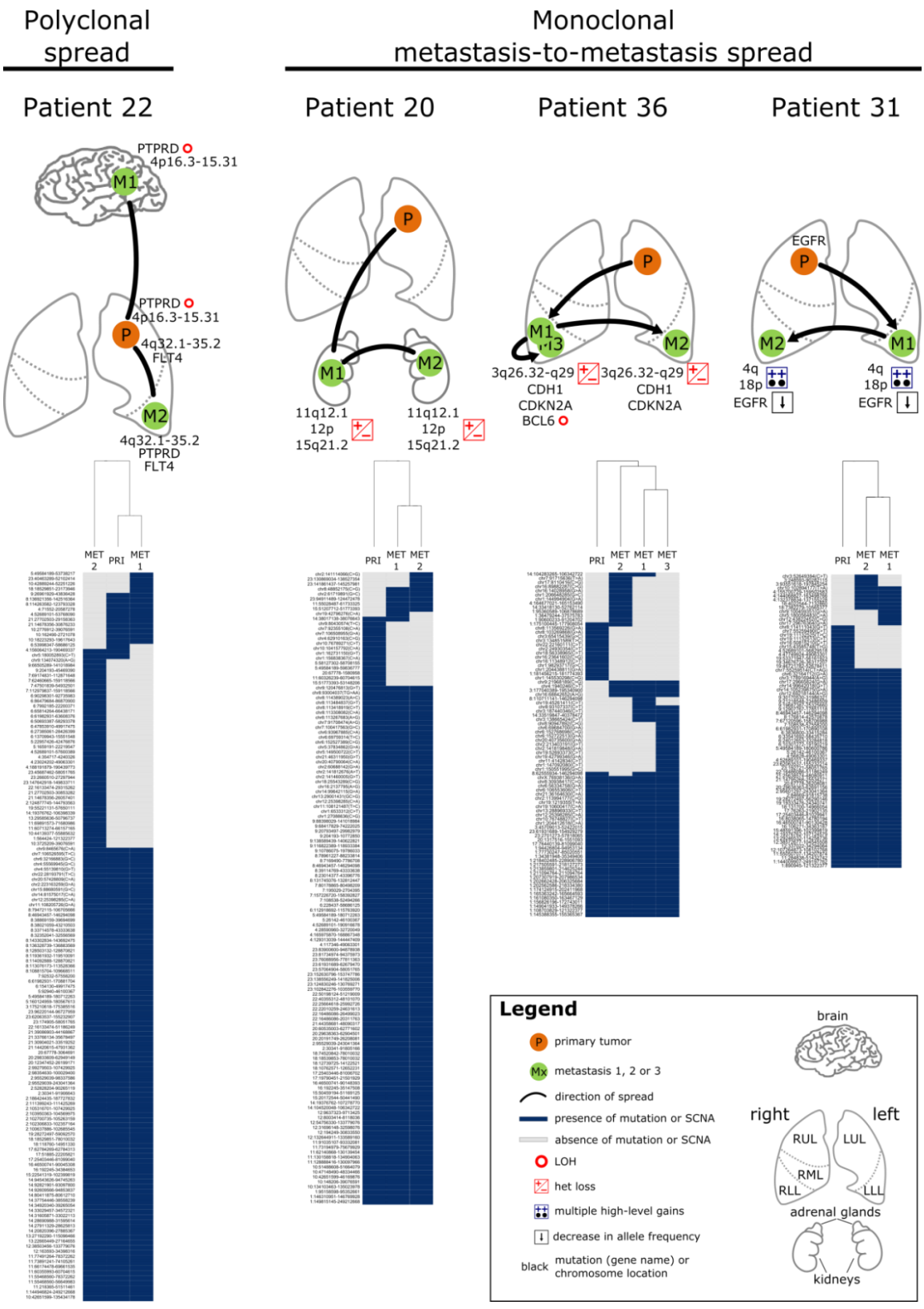


Figure 37 | Two patterns of metastatic spread. Hierarchical clustering of SNVs (including silent mutations) and SCNAs of sorted tumor populations of four patients with ≥2 metastases.

was shared between the primary tumor and the intrapulmonary metastasis. Additionally, of the two SCNAs on chromosome 4 that were both present in the primary tumor, a different one was detected in the brain and intrapulmonary metastasis, respectively (Appendix I). These data support the idea that two clones disseminated from the primary tumor independently and colonized different parts of the body ("polyclonal spread").

In contrast to patient 22, the metastases of patients 20, 31, and 36 were more closely related to each other than to their primary tumors. Known drivers (*KRAS*: patient 20, 36; *EGFR*: patient 31) and a large number of SNVs and SCNAs in patients 20 and 31, respectively, were ubiquitously present in all tumors and can therefore be considered early events in the evolution of these tumors. The tumor patient 36, however, was characterized by a substantial number of private events. The phylogenetic relationship analysis for all three patients suggests a metastatic cascade, in which one metastasis continued to seed other metastases ("metastasis-to-metastasis spread"). This comprised contralateral spreads that were either intrapulmonary (patients 31 and 36) or from one kidney to the other (patient 20) with durations from 11 months (patient 36) and up to five years (patient 31) after the detection of the primary tumors. Notably, besides few additional SCNAs, the metastases and not the primary tumor of patient 36 had a mutation in the 3'untranslated region (3'UTR) of the *CDKN2A* gene. Mutations in the 3'UTR of *CDKN2A* have been previously associated with tumor development.[215] In addition, patient 31 presented with a decrease of the mutant allele of the EGFR L858R mutation during the course of the metastatic disease (VAF: 63%, primary tumor; VAF: 39% and 33% in the first and second metastasis, respectively).

4.2.8 Multiparameter flow sorting reveals a complex situation with substantial ITH

In contrast to all other tumors in this cohort, the primary tumor of patient 42 showed more than one tumor population in the multiparameter flow cytometry analysis. It consisted of three TTF1-positive tumor populations. Of these, one diploid 2.1N population accounted for 33.3% of all cells in the tumor, and two aneuploid populations, 3.4N and 4.7N, had shares of 9.5% and 36.4%, respectively, of the tumor mass (Figure 38a). A diploid, TTF1-negative population made up the rest (20.8%) of the tumor and because no SCNAs or SNVs were detected (Figure 38c, Figure S5), it was considered to comprise the normal tumor stroma. Conversely, all three tumor populations were characterized by distinct SCNAs (Figure 38c) and a previously reported *EGFR* exon 20 insertion mutation (p.N771delinsGY)[216] that is known to confer resistance to EGFR TKIs was ubiquitously present.

In order to study the relationship among the three tumor populations of this multiploid primary tumor, we complemented aCGH analysis with whole-exome sequencing (mean coverage 85x) of all sorted populations. Sorting and ultra-deep targeted resequencing (mean coverage 5,864x) with a

custom panel of both the primary tumor and the metastasis were performed in replicates to validate the detected SNVs and their VAFs. A high correlation of the VAFs was detected for all tumor populations (Figure S13 and Figure S14).

The aCGH analysis demonstrated that all tumor populations were of clonal origin (Figure 38b). This was evident from the common aberration pattern on chromosome 5p, which entailed an amplification of *TERT* (Figure 38b, panel 1). We also detected private and shared SCNAs, at least two of them of substantial interest for defining the evolution of this tumor. First, the homozygous deletion of the gene *CDKN2A* (9p21), which encodes for the tumor suppressor p16, was exclusive to the diploid 2.1N population (Figure 38b, panel 2). Second, a high amplification of the mouse double minute 2 homolog (*MDM2)* oncogene (chr 12) was shared among the 3.4N and 4.7N populations of the primary tumor and the metastasis but was absent in the 2.1N population of the primary tumor (Figure 38b, panel 3). We confirmed the presence of distinct tumor populations by using p16 IHC on large sections of the primary tumor (Figure 39a); only the diploid tumor cells with the homozygous deletion of *CDKN2A* were characterized by the absence of the p16 protein expression. In addition, the breakpoint of the *CDKN2A* aberration (Figure 38b, panel 2) and an SCNA breakpoint on chr7q (Figure 38c), both absent in the 3.4N population, indicated a relationship between the 2.1N population, the 4.7N population, and the metastasis.

Mutational analysis validated a total of 96 SNVs and revealed a rather complex picture. Only 27% of the mutations were ubiquitous, yet 47% of all SNVs were shared by one or another tumor population (Figure 39c). Indeed, the phylogenetic relationship on metapopulation resolution, as analyzed by EXPANDS[205], indicated a relationship between the 2.1N population and both the 3.4N and 4.7N populations (Figure 39b). However, the 4.7N population was most closely related to the metastasis. We further investigated regions of LOH to determine if they were responsible for the absence of mutations. However, only 5 of 31 SNVs absent in the 3.4N population but shared between other tumor populations could be explained by losses of the mutant alleles due to LOH. In contrast, LOH was not responsible for the absence of any SNVs in the 2.1N population (Figure 39c). Interestingly, two regions affected by LOH indicated a relationship between the 2.1N and the 3.4N population and were absent in the 4.7N population and the metastasis. However, a closer analysis of these two regions revealed the other homologous chromosome was lost in the 2.1N and the 3.4N population, respectively (Figure S15). Conversely, LOH events that were ubiquitous or shared among the 3.4N, 4.7N and metastatic populations always affected the same chromosome. Notably, the 4.7N population and the metastasis lacked most of the LOH events that were found in either the 2.1N or the 3.4N population, including LOH of the two aforementioned regions present in the 2.1N and 3.4N populations (Figure S15).

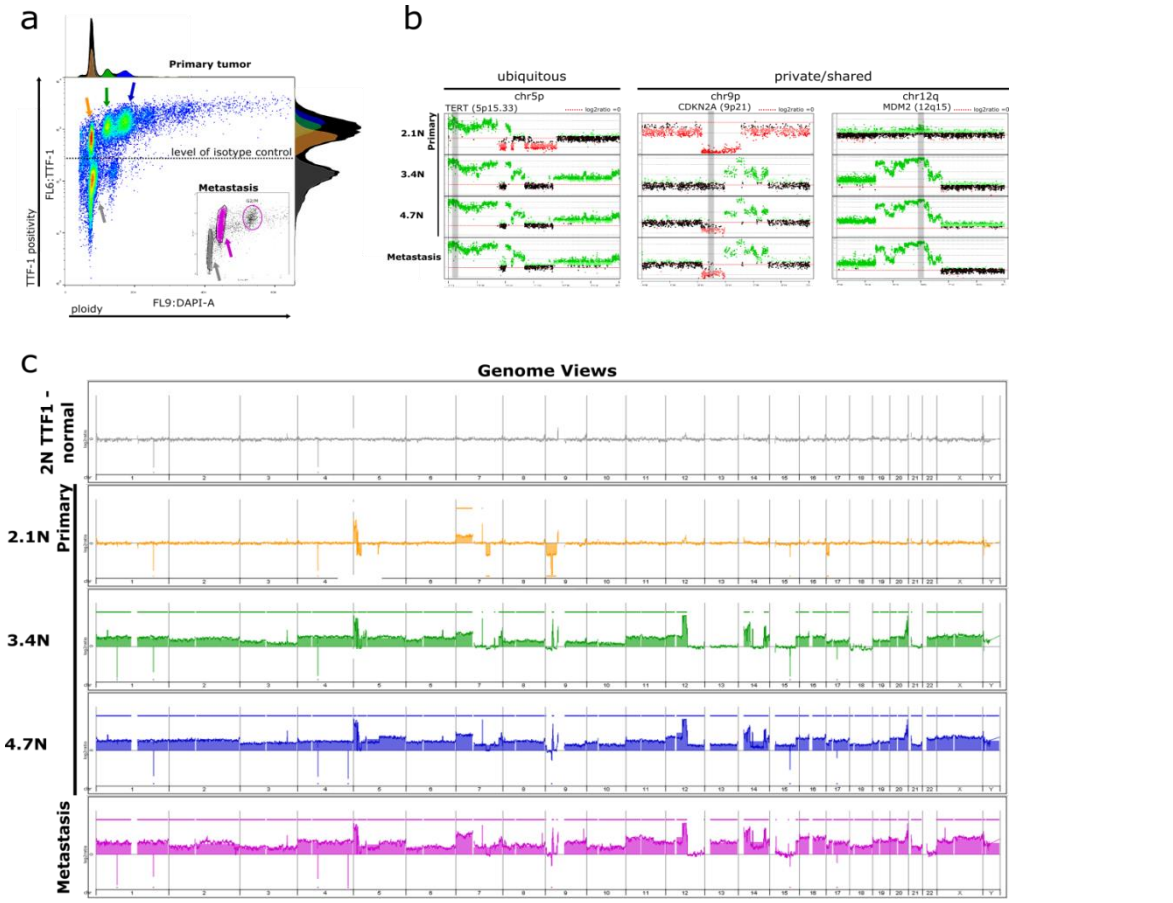


Figure 38 | Genomic ITH in the multiploid primary tumor of patient 42. The primary tumor of patient 42 consisted of three TTF1-positive tumor populations (2.1N: orange, 3.4N: green and 4.7N: blue), whereas only one tumor population (purple) was present in the metastasis (a). All tumor populations carried SCNAs (c) that were either ubiquitous (b: chr5p, *TERT*), shared between the aneuploid tumor populations (b: chr12, *MDM2*), or private to the 2.1N population (b: homozygous *CDKN2A* deletion, chr9p).

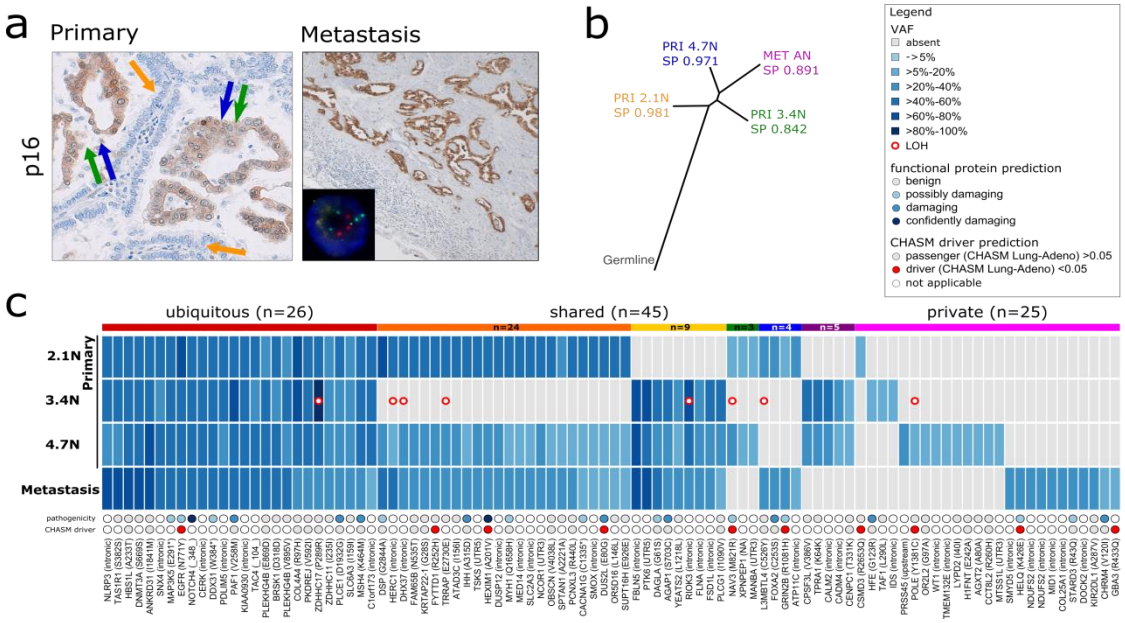


Figure 39 | Clonal tumor evolution in patient 42. IHC for p16 (a) validated the loss of p16 expression in the 2.1N population of the primary tumor with the private homozygous deletion of *CDKN2A*. The phylogenetic relationship on metapopulation resolution based on SCNAs as calculated by EXPANDS (b). Overview of 96 high confidence SNVs with their VAFs (c). Image b was produced by Dr. Noemi Andor (Stanford).

5. Discussion

Many computational methods have been developed to discover copy number aberrations directly from DNA microarrays [217–220] or sequencing data[221,222], including aCGH. Discovering copy numbers and VAFs in an absolute scale is biologically more relevant[223], yet more challenging. This is because the detection of absolute values in a tumor can be affected by two confounding factors: (i) tumor purity, i.e. the fraction of tumor cells within a heterogeneous cancer sample, and (ii) tumor ploidy, the baseline copy number of genomic segments or entire chromosomes[223,224], both of which are unknown and themselves need to be estimated. One promising approach to exclude such prior estimations of ploidy and purity is flow cytometry. It can increase purity by sorting only the cells of interest, while quantifying the ploidy of a tumor at the same.

Previous studies have applied such a sorting approach on nuclei extracted from tumor tissues in order to genomically characterize them.[225–227] These studies have sorted nuclei solely based on their DNA content. However, around 50% of human tumors are diploid (Figure 40) and can therefore not be distinguished from normal, diploid cells based on their ploidy alone. Currently, there is no method that allows for the flow cytometric enrichment of tumor nuclei from diploid tumors. To overcome these limitations, we aimed to improve the nuclei flow-sorting method from a simple ploidy-sorting approach to a tumor marker-assisted multiparameter flow-sorting technology. This allows for the genomic characterization, including the detection of SCNAs and SNVs, of diploid tumors to an unprecedented purity.

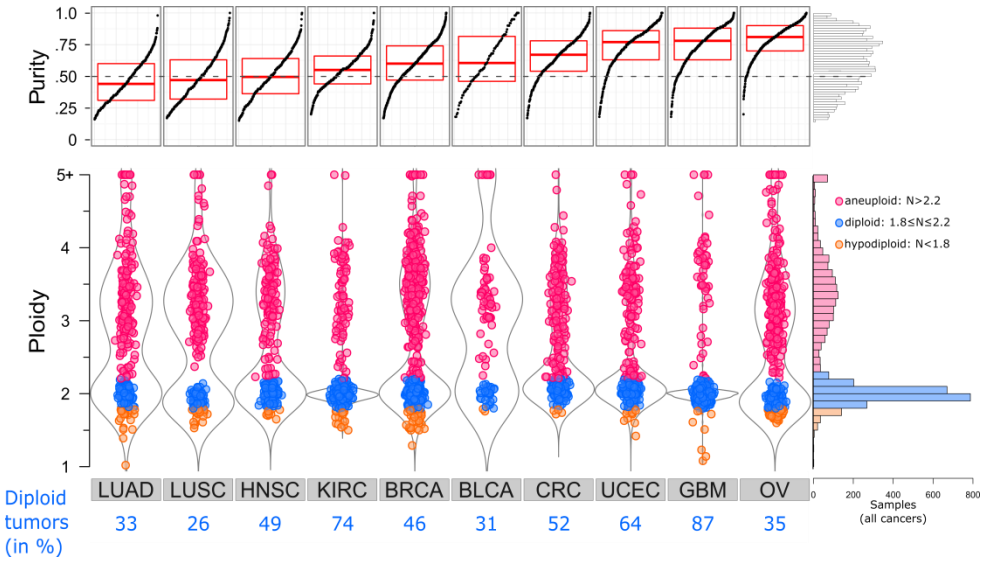


Figure 40 | Distribution of ploidy and purity across 10 cancer types (data retrieved from https://www.synapse.org/#!Synapse:syn1710466.2, version 2 (547.909 KB, modified on 2013-03-20), ref[126]); tumor purity (top) and ploidy (pirate plot, bottom). LUAD, lung adenocarcinoma; LUSC, lung squamous cell; HNSC, head and neck squamous cell; KIRC, kidney renal cell; BRCA, breast; BLCA, bladder; CRC, colorectal; UCEC, uterine cervix; GBM, glioblastoma multiforme; OV, ovary. Box plots show the median, first quartile and third quartile of purity for each cancer type. Diploid samples are designated in blue, aneuploidy tumors in pink, and hypodiploid tumors in orange. *Note: Ploidy axis is limited to a ploidy of five for visualization purposes. Tumors with a ploidy>5 are plotted under 5+.*

Part A

Part A of this thesis demonstrated the feasibility of this approach as a proof-of-concept for two markers: TTF1 and SOX10. They were chosen (i) because they are lineage markers for LUAD and malignant melanoma, respectively, (ii) because they are expressed by a large number of tumors (LUAD: 72% TTF1-positive; melanoma: 100% SOX10-positive), and (iii) because these two proteins are transcription factors and therefore localized in the nucleus. The latter point is important because the cellular integrity of cells is disrupted when frozen tissues are thawed or when FFPE tissues are digested. Consequently, only nuclei and not cells can be obtained from such archived materials.

The cell culture experiments provide evidence that both the anti-TTF1 antibody and the anti-SOX10 antibody can be used to detect TTF1 and SOX10, respectively, in nuclei extracted with our protocols and under our experimental conditions. While isotype controls are often used to quantify unspecific background staining in flow cytometry, we think that the best control to measure specificity of immunostainings is to mix positive and negative controls before the addition of the antibody. By doing so, we showed that both antibodies are sensitive and specific in separating TTF1- or SOX10-positive from TTF1- or SOX10-negative nuclei.

When we started to apply this method to human tumors, we initially focused on samples that were purely diploid – as indicated by the absence of a prominent aneuploid peak – and positive for the markers TTF1 or SOX10. We identified four samples that matched these criteria: two diploid, TTF1-positive LUADs and two diploid, SOX10-positive malignant melanomas. As expected, the diploid peaks split into a positive and a negative population according to the antibody used. We sorted each of these populations and performed copy number analysis by aCGH. Distinct SCNAs in the diploid, marker-positive fraction revealed their tumor origin, whereas no SCNAs could be detected in the diploid, marker-negative populations. We hypothesized that diploid populations without any SCNAs were of non-tumor origin. This was supported by the observation that we did not detect any SNVs in 409 cancer genes in these populations as compared to non-tumor control DNA for three patients. In addition, the VAF distribution pattern was identical to those of the tumor-free controls. Many tumors are highly intermixed with stromal and immune cells, and it is likely that they accounted for the diploid, marker-negative populations in our experiments. Based on these results, we propose the use of DNA from these populations as a surrogate for germline DNA in situations where normal germline DNA of the patients is not available. This is often the case in retrospective studies, where most of the patients have already died from their disease at the time of investigation. Matched normal DNA is important to increase sensitivity and specificity of SNV detection in tumors, as demonstrated by a recent study that has showed that 31% (targeted) or 65% (exome) of mutations have been falsely considered true positive SNVs when matched normal DNA has not been used for the analysis.[228]

The nuclei sorting technique was initially refined with the intention of sorting diploid tumor nuclei from bulk tissues of diploid tumors only. However, during screening and flow cytometric analysis of multiple tumors, we found diploid, marker-positive populations in four otherwise aneuploid tumors: three malignant melanomas and one LUAD. In contrast to the LUAD, where the diploid TTF1-positive population was of non-tumor origin, all three diploid, SOX10-positive populations in the melanomas were of tumor origin, because they had distinct SCNAs. These diploid tumor populations were clonally related to their aneuploid counterparts as indicated by a large number shared SCNAs. This – along with the fact that the ploidy of the aneuploid populations was approximately twice the ploidy of the diploid tumor populations – suggests that the aneuploid tumor populations have emerged from a cell of the diploid tumor population or from a common diploid precursor cell by a WGD event.

It has been recently demonstrated that 40% of tumors across 10 different solid cancer types are the consequence of WGDs.[126] To the best of my knowledge, the co-existence of diploid and aneuploid tumor populations in single tumor lesions has not been reported yet in malignant melanomas. In fact, this situation is limited overall to two case reports of prostate carcinoma.[225,229] Another study in head and neck squamous cell carcinomas has reported the presence of diploid tumor populations together with aneuploid populations[230], but lacks further genomic validation that the diploid population is of tumor origin, such as the presence of SCNAs or SNVs in these cells. That study has used cytokeratin as a marker of epithelial origin and has concluded that cytokeratin-positive, diploid cells are tumor cells. However, as we have learned from a similar case of a TTF1-positive LUAD in Part A of this thesis, diploid TTF1-positive populations are no necessarily tumor because TTF1 can be expressed by normal alveolar cells similarly to cytokeratin, which is also expressed by normal cells of epithelial origin. Validation is therefore necessary in suspected diploid tumor populations, either by IHC or by detection of multiple SCNAs or SNVs.

The fact that we have detected three cases with such diploid/aneuploid coexistence, suggests that this is not an uncommon state in tumors, or at least in melanomas. This is not surprising, given that all aneuploid tumors eventually originate from a normal diploid cell that became malignant during the initiation of tumorigenesis. Two possibilities could explain this co-existence. First, we could just have sampled at the right moment, meaning that the tumor biopsy was taken shortly after this WGD event, where the time interval was too short for one of the populations to sweep over the entire tumor and to eradicate the other tumor population. This implies however, that one population is better adapted than the other and that this adaption goes hand in hand with a survival and growth benefit. A second explanation, however, is that both populations are present because they cooperate and depend on each other. This theory of “cooperation among tumor cells” is not new in cancer research, and it is thought to be linked to at least three cancer hallmarks: (i) angiogenesis, which results in additional blood supply, the products of which, oxygen and nutrients, help all of the nearby

cells and not just the cell that secreted the vascular endothelial growth factor (VEGF); (ii) self-sufficiency of growth signals, where cancer cells produce several growth factors that act in a paracrine fashion on other cancer cells or stromal cells and activate them to release other growth factors; and (iii) tissue invasion and metastasis, in which proteases expressed from cancer-associated stromal cells and cancer cells themselves contribute to neoplastic progression by degradation of the extracellular matrix, allowing further proliferation, tissue invasion, and eventually metastasis of other cells.[231] Indeed, the hypothesis of tumor cell cooperation may explain some of the observations about solid cancers, such as the non-uniform abundance of certain proteins, including growth factors in IHC.[232,233] However, a more comprehensive transcriptomic and proteomic analysis would be needed to clarify the situation in our samples. Nevertheless, this co-existence emphasizes that diploid tumor cells can persist after the development of an aneuploid clone and suggests that they do so for a deeper biological reason, which probably contributes to further local tumor progression.

In one of the three above-discussed patients with melanoma, tissue of two additional metastatic lesions was available. In contrast to the local relapse, both the lymph node and the skin metastases were entirely aneuploid. We took a deeper look at the tumor evolution in this patient, by performing aCGH and targeted sequencing of frequently mutated melanoma genes from the local relapse and both metastatic sites. Overall, the majority of SCNAs and SNVs were ubiquitously present in all tumor populations, yet some private events suggest that the lymph node metastasis derived from the diploid rather than from the aneuploid tumor population of the local relapse. Conversely, the skin metastasis was a direct outgrowth of the aneuploid tumor population of the relapse, indicated by the observation that they shared all SNVs and SCNAs. The data of this patient are consistent with an early acquisition of mutations. These were probably induced by ultraviolet (UV)-light exposure, because, despite the fact that we detected only 16 with our targeted panel, 15 were C>T:G>A transition mutations, which is a signature of melanomas from sun-exposed skin.[234] Our data indicate that a malignant, diploid precursor had all these SNVs and had experienced losses of chr9, chr12 and chr17, which resulted in mutations of *PTPRD* (chr9), *ARID2* (chr12), and *TP53* (chr17) with VAFs of 100% due to the loss of their wild-type alleles. Different chromosomes or chromosome parts were then further lost prior to or during a WGD. The aneuploid tumor population of the relapse could therefore be the result of a WGD of its diploid counterpart, whereas the lymph node metastasis probably emerged from a different WGD of another precursor cell (Figure 24b). However, it cannot be determined whether this WGD event happened before metastatic dissemination, during the migration or after colonization at the lymph node itself. The fact that each population acquired additional private aberrations supports a model of parallel evolution. These private alterations also included a focal aberration pattern on chr4 in the lymph node metastasis (Figure 23e) that was similar to another patient with multiple focal deletions that were scattered across chr9p (Figure 22e). In both cases,

these focal SCNAs culminated in aberration patterns consistent with chromothripsis[235], a phenomenon that has been reported in melanoma just recently.[236]

In summary, we have provided a tool that allows the enrichment of nuclei from bulk tumor tissues based on their expression of specific proteins. The fact that we detected mutations with VAFs of 100% in all sequenced samples proves, without the necessity of estimation or computational deconvolution, (i) that all cells within the cancer harbor this mutation, (ii) that all alleles are affected, and (iii) how pure the DNA is that is obtained from sorting tumor populations using this approach. It therefore provides a possibility to quantify absolute copy numbers and absolute VAFs in diploid tumors and allows studying tumor evolution in tumors with coexisting diploid/aneuploid populations, a phenomenon that would be missed by bulk analysis. These data suggest that the majority of SCNAs in melanomas are early events or happen at least before WGD because they are shared in large numbers between diploid and aneuploid populations. This is known for SNVs in malignant melanomas because UV-exposure is the main driver of this disease and leads to a steady accumulation of many mutations[237], yet that the majority of SCNAs are present before aneuploidization has not been reported previously.

This approach is therefore applicable for the study of diploid tumors in malignant melanoma and LUAD, but can be extended to other cancers. Another good candidate is prostate cancer, because 65% of all prostate cancers are diploid[223] and 40-60% harbour the *TMPRSS2–ERG* gene fusion that results from a chromosomal rearrangement.[238,239] This gene fusion leads to an androgen-dependent transcription and nuclear expression of the transcription factor ERG of affected prostate cancer cells.[240] This could be exploited for ERG-assisted ploidy flow sorting of prostate cancer nuclei. Furthermore, as long as cells or nuclei of interest express a specific protein, multiparameter flow cytometry is able to sort these and provide them for subsequent analyses.

Part B

In Part B of this thesis we applied the multiparameter flow-sorting approach on a cohort of matched primary/metastatic LUADs. We chose this approach because LUADs have the lowest tumor purity (mean: 46%) among solid cancers[223,126], which is probably one of the reasons why data of primary tumors with matched metastases are lacking in this tumor entity. Previous studies on ITH in LUAD have focused mainly on the two main oncogenic drivers *KRAS* and *EGFR*[241] (reviewed in [242]) or have been performed in single biopsies or unmatched tumors of individual patients.[6,243] Hence, despite the fact that these studies have provided a landscape of genomic alterations in LUAD, it is currently unknown to what extent chromosomal aberrations and mutations in other cancer genes are propagated from the primary tumors to metastases. Furthermore, the role of clonal selection during

this process remains largely unidentified. To address these unresolved questions, we applied a comprehensive genomic analysis of highly purified clonal tumor cell populations from matched primary-metastatic biopsies that allows for inferring the clonal relationship of 16 LUADs across time and space. We demonstrated that this approach works for both fresh-frozen and FFPE tissues and most importantly for samples with a low tumor purity and diploid tumors that cannot be distinguished from diploid normal cells by DNA content analysis alone. To the best of my knowledge, this study represents the largest effort so far to elucidate the extent of inter-lesion heterogeneity of SCNAs and cancer gene mutations in longitudinal biopsies of matched primary and metastatic LUAD.

The prevalence of mutations of known cancer genes that we found is in line with previous studies.[53,58,56] In addition, the high purity of sorted tumor cell nuclei enabled us to study VAFs and absolute copy numbers at a previously unmet resolution. As expected, biallelic inactivation mainly affected TSGs, but surprisingly, the oncogenes *KRAS* and *EGFR* were also mutated with a VAF of 100% in three patients. Most importantly, the VAFs of mutations did not change significantly over the course of metastatic disease. Indeed, the correlation of VAFs with their absolute copy numbers indicates that the majority of ubiquitous SNVs were truncal and therefore early events in tumorigenesis. Conversely, the observation that the VAFs of shared and especially of private mutations were significantly lower, suggests that they were present only in a sub-fraction of cancer cells within the tumor and therefore occurred at later stages of tumorigenesis. The fact that these private SNVs in the primary tumor were – by definition – not propagated to the metastasis calls into question their impact in conferring metastatic risk. However, we might not have detected all metastatic sites in each patient, and they could impact local tumor progression. In addition, they might be selected under specific circumstances such as targeted therapy, which has been reported, for instance, for subclonal EGFR T790M mutations.[244] Nevertheless, their pure presence does not necessarily implicate any evolutionary selection, as subclonal mutations with low VAF were shown to often result from neutral evolution.[245]

CIN and aneuploidy are regarded as hallmarks of cancer[246], and inferences about the timing of metastatic dissemination have to take into consideration the possibility of an inherent CIN of tumors. The fact that SCNAs were detected in all tumor populations indicates a history of CIN throughout the tumors of our cohort. Chromosomally unstable tumors are expected to continuously accumulate large-scale genomic alterations over time, resulting in a higher burden of SCNAs at metastatic sites than in primary tumors. Indeed, studies in other solid cancers, e.g. pancreatic cancer[247], renal cell carcinoma[248], or prostate cancer[249] have found that the number of SCNAs is higher in metastases than in primary tumors, which argues for a continuous impact of CIN on the tumor cells. In contrast to these studies, we detected no increase in the number of SCNAs in metastases compared to primary tumors. In fact, 88% of SCNAs were propagated from the primary tumors to the metastases

and 87% of all SCNAs detected in the metastases were shared with the primary tumor. This is surprising given that the median time interval from primary tumor detection to the occurrence of the first biopsied metastasis was almost one year. Such little ITH between the primary and metastatic sites indicates that most SCNAs and SNVs were early events and manifested in the primary tumors before metastatic dissemination.

This high primary-metastatic concordance might be expected for SNVs in a cancer that accumulates mutations early mainly due to the exposure of carcinogens in tobacco smoke[56], yet this has not been previously reported for SCNAs in LUADs and is in agreement with published data of high primary/metastatic SCNAs concordance in other cancers.[250,251] In fact, a study in lung cancer patients has investigated SCNA-patterns at single cell resolution in CTCs – the tumor cells that circulate in the blood of cancer patients and are thought to be the founders of metastases. This study has detected that CTCs exhibited copy number patterns that are reproducible between individual CTCs from the same patient.[252] Similarly, a recent study of single cells in breast cancer has found that most SCNAs happen in a short period of crisis followed by stasis and expansion of stable clones without much further change, resulting in rather homogeneous copy number patterns.[127] Such a saltation theory of punctuated evolution implies the presence of one dominant large clone, the so-called “hopeful monster”, and is compatible with ongoing selection and gradual, yet slow, changes. According to this theory, additional changes are low, compared to the events that happened in short punctuated bursts at the early stages of tumor evolution and lead to a rather chromosomally stable expansion of a clone that eventually forms the tumor mass. Such a theory could explain the observation in our cohort that not only most of the SCNAs were shared between primary tumors and their metastasis, but also that changes in absolute copy numbers were rarely detected.

The idea that most SCNAs are shared between the primary tumor and the metastasis because they occur in the early steps of tumorigenesis, also assumes that they are truncal and that the primary tumor consists of a rather homogeneous population of cells in respect to SCNAs. Indeed, metastases are seeded by single or a few cells from the primary tumor; hence if the primary tumor was heterogeneous, a reduction in genetic diversity would be seen at metastatic sites, and we would have detected a decrease in the number of SCNAs that are propagated from the primary tumor to the metastases. In addition, cells have to undergo many additional cell divisions at metastatic sites in order to form a clinically detectable tumor mass. If these metastatic founders are chromosomally unstable, a higher discrepancy of SCNAs and their absolute copy numbers between two lesions would be expected as compared to what we have identified in our cohort. Therefore, it is possible that LUADs in this cohort received the majority of their SCNAs through a punctuated evolution followed by a rather stable local and metastatic tumor progression (Figure 41b,c). We note that gradual changes could behave similar, but only under the assumption of very long time periods for

these changes to happen (Figure 41a). Otherwise, it would not be plausible that chromosomally unstable tumor cells stop to "destroy" their genomes. CIN eventually result in aneuploidy, but not all aneuploid cells have to be chromosomally unstable. This phenomenon of aneuploid tumors with a uniform, stable karyotype has been reported, but has received much less attention than CIN.[246,253]

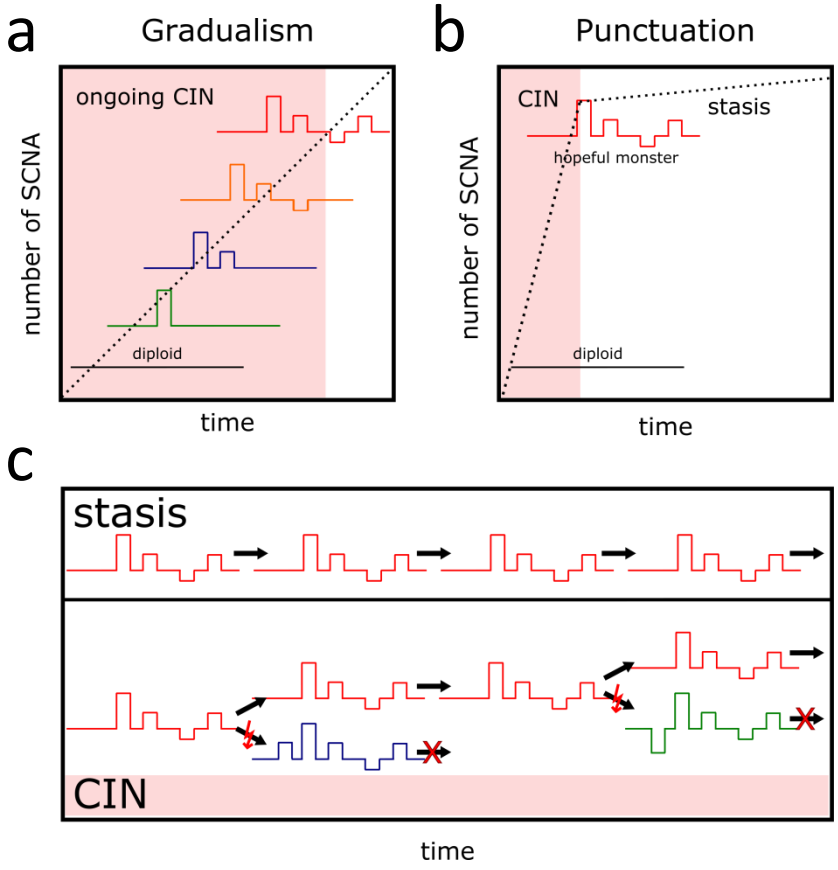


Figure 41 | Models of evolution in the LUAD cohort. Investigating SCNAs at a single time point provides only a snapshot in the evolution of a tumor. With our analysis we cannot distinguish, if the SCNAs in the primary tumor are the consequence of a gradual accumulation of alterations in a chromosomally unstable tumor over time (a), or the consequence of a single, punctuated event that is not repeated (b). However, we analyzed clonally related metastases and found that they share most of their SCNAs with the primary tumors. This can be explained by two models: (i) if the primary tumor is the result of an early punctuated burst of evolution in the early stages of tumor progression (b), is might not exhibit high levels of CIN. Therefore, the primary tumor could be a "hopeful monster" that undergoes stable clonal expansions during metastatic development (c, upper panel); (ii) if the primary tumor has evolved gradually, it is probably chromosomally unstable, and CIN is unlikely to be turned off by the tumor itself. Under this assumption, the little increase of SCNAs at metastatic sites that we have detected might result from a negative selection of new SCNAs that emerge over the course of metastatic progression (c, lower panel).

However, caution should be exercised from which viewpoint the knowledge about CIN is generated. In this LUAD cohort, SCNAs were investigated at the population level – consisting of many thousand cells with the same ploidy – and not at single cell resolution. This is a difference, because what is seen as the result of a population does not have to be true for individual cells. It has been emphasized that stable karyotypes at the population level can persist, despite high rates of ongoing

CIN in single cells.[159] In such steady-state conditions, new SCNAs are established in the population only after long time periods, at least as long as the environment remains constant. In fact, it has been documented that the karyotypes of some cell lines have remained relatively unchanged despite the fact that they have been cultured over years separately in different laboratories.[254,255] This argues for a selection against the emergence of new SCNAs under steady-state conditions. However, if external selection factors such as specific therapies are applied, CIN of individual cells can manifest itself as karyotypic heterogeneity. The cell line MCF-7, for instance, has been found to be karyotypically stable over generations until a selection for a cytotoxic resistance has been applied.[256] Therefore the low ITH that we detected in our cohort of primary-metastatic LUADs on the population level could also be the result of a negative selection of new emerging, deleterious copy number alterations (Figure 41c). To test these hypotheses one would ultimately need to investigate single cells and check for the presence or absence of intermediate genotypes. However, despite advances in single cell technologies, this is a labor- and cost-intensive endeavor that would have gone beyond the scope of this thesis.

CIN and aneuploidy have been previously linked to a dysregulation of the ataxia-telangiectasia mutated (ATM)-p53 pathway.[257] p53 is a tumor suppressor and regarded as "the guardian of the genome"[258] because of its central role in organizing whether a cell responses to various types of stress such as DNA damage with apoptosis, cell cycle arrest, senescence, DNA repair, cell metabolism or autophagy.[259] During normal homeostasis, p53 levels are stabilized by two mechanisms: (i) continuous degradation of p53 in proteasomes mediated by ubiquitination through the E3 ubiquitin ligase Mdm2; and (ii) phosphorylation by ATM, ATR and other kinases stabilizes the p53 protein and promote its DNA binding and subsequent transcription of p53 target genes.[260] Recent studies have shown that p53 is not only activated by DNA damage but also by aneuploidy.[158,257] The activation of p53 in aneuploid cells is mediated through ATM in the presence of elevated levels of reactive oxygen species, which are increased probably due to a higher energy consumption of aneuploid cells.[261,262] Therefore, dysregulation of the ATM-p53 pathway, for instance due to mutations in the TSGs *TP53* and *ATM*, or amplifications of the oncogene *MDM2,* has been implicated with tumorigenesis. In our cohort, biallelic inactivation of *TP53* or *ATM* correlated with a significant high burden of SCNAs. This is in line with previous data that have found the number of SCNAs to be higher in tumors with *TP53* mutations.[126,263,264] In addition, few private SCNAs were evident at metastatic sites in *TP53*-mutated tumor in our cohort, which suggests a late dissemination of the metastatic clone from the primary tumor. In fact, we observed that the time interval between primary and metastatic tumor detection was the shortest in tumors with biallelic inactivation of *TP53* or *ATM*. This is in agreement with previous reports that have shown that loss of p53 accelerates growth in aneuploid tumors[257,265], suggesting that a fast proliferation of *TP53*- or *ATM*-mutated tumors did not allow enough time to

accumulate many additional SCNAs. Of note, all diploid tumor populations were *TP53* and *ATM* wild-type and also had the overall lowest amounts of SCNAs among this cohort, which is in line with previous studies that have found (i) that diploid tumors have fewer genomic aberrations than aneuploid tumors[266], and (ii) that TP53 mutations occur more often in aneuploid than diploid tumors.[126,267] Interestingly, only the aneuploid tumor populations in the multiploid primary tumor of patient 42, and not the diploid tumor population had a high-level amplification of the *MDM2* oncogene, potentially correlating with a loss of p53 function due to overexpression of Mdm2. Inactivation of p53 has been previously associated with tolerance of tetraploidy, the result of a WGD.[265,268–270] The correlation of *TP53* mutations and aneuploidy in our cohort therefore indicates that this association also applies to LUADs.

The early emergence of metastases in tumors with an inactivated ATM-p53 pathway suggested that these tumors follow the linear progression model of tumor evolution. However, our analysis of genetic divergence also discovered some primary-metastatic pairs with clear signs of parallel evolution, which is associated with an early dissemination of the metastatic founder clone from the primary tumor. This is in line with previous data that have found evidence for branched evolution in breast cancer[271], renal cell carcinoma[272] and among other cancer types in a recent pan-cancer analysis.[273] Despite these observations, we were not able to detect a correlation between the type of metastatic progression (linear or parallel) and the relative time interval between primary and metastatic tumor detection. For instance, the P/M divergence was similar in patient 12 and patient 49, yet the time between primary and metastatic tumor detection was 2 months and 12.7 years, respectively. One has to be careful, however, because the tumor of patient 49 was diploid and characterized by only a few SCNAs. In such cases, an absolute small number of additional alterations has a relatively strong effect on the value of the P/M divergence. Nevertheless, genetic distance between two lesions is unlikely to correlate only with the time difference in clinical detection. Pancreatic adenocarcinoma, for instance, was believed to metastasize early because many patients have metastases already at diagnosis. However, this was an oversimplified assumption, as demonstrated by whole-genome sequencing data, which have revealed that late diagnosis, and not early metastasis, is responsible for this observation. In fact, it has been suggested that the metastatic process took decades.[274] This is because other parameters, such as the timing of divergence between the primary tumor and the metastases, the exposure to mutagenic substances, intrinsic genomic instability, and the dynamics of evolution in each lesion unequivocally impact the shaping of cancer genomes.

To study whether metastases are the result of repeated seeding from the primary tumor or arise from a previously established metastasis, multiple tumors of individual patients have to be investigated. Our cohort included four patients with oligometastatic disease. Hierarchical clustering

of SCNAs and SNVs of all tumor lesions indicated that both primary tumors and metastases can spread to further metastatic sites. Patient 22, for instance, showed two separate waves of metastatic spread: from the primary tumor (i) to a brain and (ii) to an ipsilateral lung metastasis. In contrast, in three other patients (patients 20, 31, and 36) the primary tumor was an outgroup to all metastases, i.e. all metastases were more closely related to a common ancestor than to their most recent common ancestor with the primary tumor. Likewise, a recent report has demonstrated that the primary tumor has been an outgroup to all metastases in two LUADs.[273] This relatively high proportion of patients with tumors that evolve through a metastatic cascade might have important clinical implications. It provides a strong argument in favor of aggressive local treatment of all metastatic sites in patients with synchronous or metachronous oligometastatic LUADs.[275,276]

Cancer evolution is a complex process and, despite intensive research, insufficiently understood. In this study an extraordinary example of ITH and cancer evolution was observed in a multiploid primary tumor and its metastasis (patient 42). While the metastasis most likely emerged from the 4.7N population of the primary tumor, analysis of SCNAs and SNVs in the highly purified tumor cell populations of the primary tumor did not reveal a clear picture of either a linear or parallel evolution. Interestingly, the genomic data were more consistent with the 4.7N population being a hybrid of the 2.1N and 3.4N populations. The idea that tumor hybridization contributes to cancer progression was introduced more than 100 years ago.[277–279] Cell fusions – which occur naturally during fertilization, the formation of placenta and muscle fibers, and bone homeostasis[280] – have been described as occurring between cancer cells and other cell types both in *in-vitro* and in *in-vivo* mouse models.[281–283] In humans, however, they are difficult to detect because of the lack of an appropriate genetic marker. Hence, the most convincing result in human cancer derives from a single case report of a woman that developed renal-cell carcinoma after an allogenic bone marrow transplantation from a male donor. Donor Y chromosomes were found in the tumor cells, suggesting that a cell fusion between a donor and a recipient cell had taken place.[284] Despite these limited data regarding human cancers, the concept of cell fusion has been recently revitalized as one of the forces that drive cancer progression.[280] To provide evidence for a tumor-tumor cell fusion in patient 42 of our cohort, the mutational landscape that we found in each sorted tumor cell population would need to be demonstrated at single cell resolution. However, as mentioned earlier, this would go beyond the scope of this work. Even though a definitive answer for the evolution of the multiploid tumor in patient 42 remains outstanding, we provide evidence that cells of the same ploidy share private genomic alterations that would be missed by an analysis of bulk material and demonstrate that cancer evolution is more complex than currently thought.

Nevertheless, this study has certain limitations. Array-CGH, for instance, cannot detect chromosomal gains and losses without copy number changes. Therefore, balanced chromosomal alterations, such

as translocations, inversions, or mosaicism are not identified. Further, the chip size used for our experiments allows the detection of aberrations with a minimum size of 35.2 Kb. Hence, smaller aberrations are not discovered. However, our aCGH data are at high resolution as compared to conventional comparative genomic hybridization (CGH), for which the maximal resolution is around 10 Mb.[285] In addition, aCGH provides copy numbers and aberrations across the entire genome, including non-coding regions, whereas studies that detect copy numbers from whole-exome sequencing are limited to coding regions only. We acknowledge that our focus on the targeted sequencing of cancer genes does not allow for the detection of mutations that occur in other genes previously not associated with cancer. However, our focus is on ITH of important cancer driver gene mutations rather than the discovery of new mutations in other genes. Regardless, it is very likely that many mutations detected in smoking-related LUADs are passenger rather than driver mutations. We are aware that the medium size of our cohort might be too small to find further significant correlations. Yet, sample size is less relevant in qualitative research than in quantitative approaches, and we are convinced we have provided qualitative research by genome-wide detection of SCNAs and ultra-deep sequencing of well-known cancer genes in tumor populations that were sorted to a previously unprecedented purity in LUADs.

The aim of this study was to provide a comparison of matched primary-metastatic LUAD pairs with material collected over two decades at the IfP. In summary, a high concordance of SCNAs and SNVs between primary tumors and metastases was detected that is responsible for the observation of stable ploidies over long periods of time. This overall low ITH between primary and metastatic sites suggests that most SCNAs are early events that accumulate in the primary tumors before metastatic dissemination and that CIN is low on the population level. Similarly, low ITH of 21 cancer gene mutations has been recently revealed in primary LUADs using a multiregion sequencing approach[6]. This observation is hereby extended to include mutations in a large number of other cancer-relevant genes and, to the best of my knowledge, for the first time on the level of copy number aberrations in LUADs.

6. General Conclusion

In this work, we have presented a technology that increases the purity of tumor DNA by sorting nuclei according to their DNA content from bulk tissues. Most importantly, we have demonstrated that the addition of tumor lineage markers allow for the enrichment of tumor nuclei from diploid tumors and tumors of low purity. Multiparameter nuclei flow sorting allowed us to study chromosomal aberrations and their copy numbers as well as mutations and their VAFs to a previously unmet level of precision. This has been demonstrated on three levels: (i) the detection of SNVs with VAFs of 100%; (ii) the identification of multiple tumor populations with distinct private mutations in individual, single tumor lesions; and (iii) the fact that the DNA of the sorted non-tumor populations can serve as a germline control in order to identify the somatic nature of genetic alterations.

We applied this refined nuclei flow-sorting technology to a cohort of 16 patients with primary LUAD and their matched metastases. We validated their clonal relationship by the presence of common SCNAs and mutations and found that the majority of alterations were shared between primary tumors and their metastasis. In fact, all but one tumor consisted of only tumor cells with a single ploidy and the combined analysis of SCNAs and VAFs of SNVs indicated that ubiquitous mutations and aberrations are truncal. This resulted in the initial observation by flow cytometry that tumor ploidy remained unchanged over the course of metastatic disease. This suggests that either the genomes of LUADs are remarkably stable after the primary tumor has been established – despite a history of CIN evident from the presence of SCNAs – or that the majority of new aberrations are deleterious and therefore negatively selected in tumors with ongoing CIN. Nevertheless, the high concordance of alterations between primary tumor and metastatic sites provided evidence that the majority of SCNAs and SNVs were present at the primary site before metastatic dissemination, which suggests that they can already be detected in the primary tumors.

Genomic characterization of high-purity samples at an appropriate depth might therefore be sufficient to identify the majority of SCNAs and SNVs in relevant cancer genes that are propagated from primary tumors to their metastases. Additionally, the observation that metastases are the source of further metastatic spread in three out of four patients has clinical implications; in such a situation, the removal of the primary tumor alone would fail to halt further metastatic progression. In the era of personalized medicine, further approaches in large cohorts are needed to understand how CIN and ITH impact the biology of LUADs that have been treated with targeted therapy. One such approach is the TRACERx (TRAcking Cancer Evolution through therapy [Rx]) lung research project by Cancer Research UK. This project aims to uncover the mechanisms of lung cancer evolution by analyzing ITH using multiregion sequencing of matched primary and metastatic sites to define the relationship between ITH and the clinical outcome of more than 800 NSCLC patients.

References

1. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
2. Pelosi, G. *et al.* Deciphering intra-tumor heterogeneity of lung adenocarcinoma confirms that dominant, branching, and private gene mutations occur within individual tumor nodules. *Virchows Arch.* (2016). doi:10.1007/s00428-016-1931-z
3. Vignot, S. *et al.* Next-Generation Sequencing Reveals High Concordance of Recurrent Somatic Alterations Between Primary Tumor and Metastases From Patients With Non–Small-Cell Lung Cancer. *J. Clin. Oncol.* **31,** 1–7 (2013).
4. Sun, L. *et al.* Comparison of KRAS and EGFR gene status between primary non-small cell lung cancer and local lymph node metastases: implications for clinical practice. *J. Exp. Clin. Cancer Res.* **30,** 30 (2011).
5. Wang, S. & Wang, Z. Meta-analysis of epidermal growth factor receptor and KRAS gene status between primary and corresponding metastatic tumours of non-small cell lung cancer. *Clin. Oncol.* **27,** 30–9 (2015).
6. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346,** 256–259 (2014).
7. Torre, L. A. *et al.* Global Cancer Statistics, 2012. *CA a cancer J. Clin.* **65,** 87–108 (2015).
8. Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, C. K. (eds). SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.
9. Ramalingam, S. S., Owonikoko, T. K. & Khuri, F. R. Lung Cancer : New Biological Insights and Recent Therapeutic Advances. *Cancer* **61,** 91–112 (2011).
10. Prevention, U. S. D. of H. and H. S. P. C. for D. C. and. *The Health Consequences of Smoking: A Report of the Surgeon General*. **7829,** (2004).
11. Wynder, E. L. & Graham, E. A. Etiologic factors in bronchiogenic carcinoma with special reference to industrial exposures; report of eight hundred fifty-seven proved cases. *AMA. Arch. Ind. Hyg. Occup. Med.* **4,** 221–35 (1951).
12. Doll, R. & Hill, A. B. Smoking and carcinoma of the lung; preliminary report. *Br. Med. J.* **2,** 739–48 (1950).
13. American Cancer Society. Cancer facts & figures 2015. *Atlanta Am. Cancer Soc.* (2015).
14. Islami, F., Torre, L. A. & Jemal, A. Global trends of lung cancer mortality and smoking prevalence. *Transl. Lung Cancer Res.* **4,** 327–338 (2015).
15. Shaper, A. G., Wannamethee, S. G. & Walker, M. Pipe and cigar smoking and major cardiovascular events, cancer incidence and all-cause mortality in middle-aged British men. *Int. J. Epidemiol.* **32,** 802–8 (2003).
16. Henley, S. J., Thun, M. J., Chao, A. & Calle, E. E. Association between exclusive pipe smoking and mortality from cancer and other diseases. *J. Natl. Cancer Inst.* **96,** 853–61 (2004).
17. Dela Cruz, C. S., Tanoue, L. T. & Matthay, R. A. Lung Cancer: Epidemiology, Etiology, and Prevention. *Clin. Chest Med.* **32,** 605–644 (2011).
18. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global Cancer Statistics, 2002. *CA. Cancer J. Clin.* **55,** 74–108 (2005).

19. Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers — a different disease. *Nature* **7,** 778–790 (2007).

20. Rudin, C. M., Avila-Tang, E. & Samet, J. M. Lung cancer in never smokers: a call to action. *Clin. Cancer Res.* **15,** 5622–5 (2009).

21. Künzli, N. & Tager, I. B. Air pollution: from lung to heart. *Swiss Med. Wkly.* **135,** 697–702 (2005).

22. Vineis, P. & Husgafvel-Pursiainen, K. Air pollution and cancer: biomarker studies in human populations. *Carcinogenesis* **26,** 1846–55 (2005).

23. Boffetta, P. Epidemiology of environmental and occupational cancer. *Oncogene* **23,** 6392–403 (2004).

24. Hwang, S.-J. *et al.* Lung cancer risk in germline p53 mutation carriers: association between an inherited cancer predisposition, cigarette smoking, and cancer risk. *Hum. Genet.* **113,** 238–43 (2003).

25. Li, X. & Hemminki, K. Inherited predisposition to early onset lung cancer according to histological type. *Int. J. cancer* **112,** 451–7 (2004).

26. Bailey-Wilson, J. E. *et al.* A major lung cancer susceptibility locus maps to chromosome 6q23-25. *Am. J. Hum. Genet.* **75,** 460–74 (2004).

27. Thorgeirsson, T. E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452,** 638–42 (2008).

28. Hung, R. J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452,** 633–7 (2008).

29. Arita, T. *et al.* Bronchogenic carcinoma: incidence of metastases to normal sized lymph nodes. *Thorax* **50,** 1267–9 (1995).

30. Cheng, Y. W. *et al.* The association of human papillomavirus 16/18 infection with lung cancer among nonsmoking Taiwanese women. *Cancer Res.* **61,** 2799–803 (2001).

31. Stabile, L. P. *et al.* Human non-small cell lung tumors and cells derived from normal lung express both estrogen receptor alpha and beta and show biological responses to estrogen. *Cancer Res.* **62,** 2141–50 (2002).

32. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Tobacco smoke and involuntary smoking. *IARC Monogr. Eval. Carcinog. risks to humans* **83,** 1–1438 (2004).

33. Hecht, S. S. *et al.* A tobacco-specific lung carcinogen in the urine of men exposed to cigarette smoke. *N. Engl. J. Med.* **329,** 1543–6 (1993).

34. Vineis, P. *et al.* Tobacco and cancer: recent epidemiological evidence. *J. Natl. Cancer Inst.* **96,** 99–106 (2004).

35. Stayner, L. *et al.* Lung Cancer Risk and Workplace Exposure to Environmental Tobacco Smoke. *Am. J. Public Health* **97,** 545–551 (2007).

36. Vineis, P. *et al.* Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ* **330,** 277 (2005).

37. Alberg, A. J., Brock, M. V & Samet, J. M. Epidemiology of lung cancer: looking to the future. *J. Clin. Oncol.* **23,** 3175–85 (2005).

38. Boffetta, P. Human cancer from environmental pollutants: the epidemiological evidence. *Mutat. Res.* **608,** 157–62 (2006).

39. Subramanian, J. & Govindan, R. Lung Cancer in Never Smokers: A Review. *J. Clin. Oncol.* **25,**

561–570 (2007).

40. Yu, I. T. S., Chiu, Y.-L., Au, J. S. K., Wong, T.-W. & Tang, J.-L. Dose-response relationship between cooking fumes exposures and lung cancer among Chinese nonsmoking women. *Cancer Res.* **66,** 4961–7 (2006).

41. Brownson, R. C., Alavanja, M. C., Caporaso, N., Simoes, E. J. & Chang, J. C. Epidemiology and prevention of lung cancer in nonsmokers. *Epidemiol. Rev.* **20,** 218–36 (1998).

42. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clin. Proc.* **83,** 584–594 (2008).

43. Travis, W. D. Pathology of lung cancer. *Clin. Chest Med.* **23,** 65–81 (2002).

44. Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K.-K. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat. Rev. Cancer* **14,** 535–546 (2014).

45. Khuder, S. A. Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis. *Lung Cancer* **31,** 139–48

46. Toh, C.-K. *et al.* Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity. *J. Clin. Oncol.* **24,** 2245–51 (2006).

47. Lortet-Tieulent, J. *et al.* International trends in lung cancer incidence by histological subtype: adenocarcinoma stabilizing in men but still increasing in women. *Lung Cancer* **84,** 13–22 (2014).

48. Gray, N. The consequences of the unregulated cigarette. *Tob. Control* **15,** 405–408 (2006).

49. Rock, J. R. & Hogan, B. L. M. Epithelial progenitor cells in lung development, maintenance, repair, and disease. *Annu. Rev. Cell Dev. Biol.* **27,** 493–512 (2011).

50. Swanton, C. & Govindan, R. Clinical Implications of Genomic Discoveries in Lung Cancer. *N. Engl. J. Med.* **374,** 1864–1873 (2016).

51. Xu, X. *et al.* Evidence for type II cells as cells of origin of K-Ras-induced distal lung adenocarcinoma. *Proc. Natl. Acad. Sci.* **109,** 4910–4915 (2012).

52. Mainardi, S. *et al.* Identification of cancer initiating cells in K-Ras driven lung adenocarcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 255–60 (2014).

53. Collisson, E. a. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 1–8 (2014). doi:10.1038/nature13385

54. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).

55. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524,** 47–53 (2015).

56. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150,** 1121–1134 (2012).

57. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485,** 502–506 (2012).

58. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48,** 607–616 (2016).

59. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).

60. Pylayeva-Gupta, Y., Grabocka, E. & Bar-Sagi, D. RAS oncogenes: weaving a tumorigenic web. *Nat. Rev. Cancer* **11,** 761–74 (2011).

61. Sordella, R., Bell, D. W., Haber, D. A. & Settleman, J. Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. *Science* **305,** 1163–7 (2004).

62. Gandhi, J. *et al.* Alterations in genes of the EGFR signaling pathway and their relationship to EGFR tyrosine kinase inhibitor sensitivity in lung cancer cell lines. *PLoS One* (2009). doi:10.1371/journal.pone.0004576

63. Ladanyi, M. & Pao, W. Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. *Mod. Pathol.* **21 Suppl 2,** S16-22 (2008).

64. Riely, G. J. *et al.* Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin. Cancer Res.* **14,** 5731–5734 (2008).

65. Coleman, M. L., Marshall, C. J. & Olson, M. F. RAS and RHO GTPases in G1-phase cell-cycle regulation. *Nat. Rev. Mol. Cell Biol.* **5,** 355–366 (2004).

66. Reynolds, S. H. *et al.* Activated protooncogenes in human lung tumors from smokers. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 1085–9 (1991).

67. Barbacid, M. ras genes. *Annu. Rev. Biochem.* **56,** 779–827 (1987).

68. Campbell, S. L., Khosravi-Far, R., Rossman, K. L., Clark, G. J. & Der, C. J. Increasing complexity of Ras signaling. *Oncogene* **17,** 1395–413 (1998).

69. Seo, K. Y., Jelinsky, S. A. & Loechler, E. L. Factors that influence the mutagenic patterns of DNA adducts from chemical carcinogens. *Mutat. Res.* **463,** 215–46 (2000).

70. Prior, I. A., Lewis, P. D. & Mattos, C. A Comprehensive Survey of Ras Mutations in Cancer. *Cancer Res.* **72,** 2457–2467 (2012).

71. Brunner, T. B. *et al.* Farnesyltransferase inhibitors: an overview of the results of preclinical and clinical investigations. *Cancer Res.* **63,** 5656–68 (2003).

72. Whyte, D. B. *et al.* K- and N-Ras are geranylgeranylated in cells treated with farnesyl protein transferase inhibitors. *J. Biol. Chem.* **272,** 14459–64 (1997).

73. Olayioye, M. A., Neve, R. M., Lane, H. A. & Hynes, N. E. The ErbB signaling network: receptor heterodimerization in development and cancer. *EMBO J.* **19,** 3159–67 (2000).

74. Ciardiello, F., De Vita, F., Orditura, M. & Tortora, G. The role of EGFR inhibitors in nonsmall cell lung cancer. *Curr. Opin. Oncol.* **16,** 130–5 (2004).

75. Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350,** 2129–2139 (2004).

76. Al Olayan, A., Al Hussaini, H. & Jazieh, A. R. The roles of epidermal growth factor receptor (EGFR) inhibitors in the management of lung cancer. *J. Infect. Public Health* **5 Suppl 1,** S50-60 (2012).

77. Sharma, S. V, Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7,** 169–181 (2007).

78. Shigematsu, H. *et al.* Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J. Natl. Cancer Inst.* **97,** 339–46 (2005).

79. Kim, H. R. *et al.* Distinct clinical features and outcomes in never-smokers with nonsmall cell lung cancer who harbor EGFR or KRAS mutations or ALK rearrangement. *Cancer* **118,** 729–39 (2012).

80. Usui, K. *et al.* The Frequency of Epidermal Growth Factor Receptor Mutation of Nonsmall Cell Lung Cancer according to the Underlying Pulmonary Diseases. *Pulm. Med.* **2011,** 290132 (2011).

81. Boch, C. *et al.* The frequency of EGFR and KRAS mutations in non-small cell lung cancer (NSCLC): routine screening data for central Europe from a cohort study. *BMJ Open* **3,** e002560 (2013).

82. Subramanian, J. & Govindan, R. Lung cancer in never smokers: a review. *J. Clin. Oncol.* **25,** 561–70 (2007).

83. Shigematsu, H. & Gazdar, A. F. Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers. *Int. J. cancer* **118,** 257–62 (2006).

84. Kumar, A., Petri, E. T., Halmos, B. & Boggon, T. J. Structure and clinical relevance of the epidermal growth factor receptor in human cancer. *J. Clin. Oncol.* **26,** 1742–51 (2008).

85. Zhang, X., Gureasko, J., Shen, K., Cole, P. A. & Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125,** 1137–49 (2006).

86. Yun, C.-H. *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 2070–5 (2008).

87. Carey, K. D. *et al.* Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Res.* **66,** 8163–71 (2006).

88. Yun, C.-H. *et al.* Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **11,** 217–27 (2007).

89. Inukai, M. *et al.* Presence of epidermal growth factor receptor gene T790M mutation as a minor clone in non-small cell lung cancer. *Cancer Res.* **66,** 7854–7858 (2006).

90. Levine, B. & Weisberger, A. S. The response of various types of bronchogenic carcinoma to nitrogen mustard. *Ann. Intern. Med.* **42,** 1089–96 (1955).

91. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304,** 1497–500 (2004).

92. Rusch, V. *et al.* Differential expression of the epidermal growth factor receptor and its ligands in primary non-small cell lung cancers and adjacent benign lung. *Cancer Res* **53,** 2379–2385 (1993).

93. Bailey, R. *et al.* O-242 Gefitinib (‘Iressa,’ ZD1839) monotherapy for pretreated advanced non small-cell lung cancer in IDEAL 1 and 2: tumor response is not clinically relevantly predictable from tumor EGFR membrane staining alone. *Lung cancer* **41,** 71 (2003).

94. Scagliotti, G. V. *et al.* Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J. Clin. Oncol.* **26,** 3543–51 (2008).

95. Thomas, A., Liu, S. V., Subramaniam, D. S. & Giaccone, G. Refining the treatment of NSCLC according to histological and molecular subtypes. *Nat. Rev. Clin. Oncol.* 1–16 (2015). doi:10.1038/nrclinonc.2015.90

96. Eisenhauer, E. A. *et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45,** 228–47 (2009).

97. Mulloy, R. *et al.* Epidermal growth factor receptor mutants from human lung cancers exhibit enhanced catalytic activity and increased sensitivity to gefitinib. *Cancer Res.* **67,** 2325–30 (2007).

98. Godin-Heymann, N. *et al.* Oncogenic Activity of Epidermal Growth Factor Receptor Kinase Mutant Alleles Is Enhanced by the T790M Drug Resistance Mutation. *Cancer Res.* **67,** 7319–

7326 (2007).

99. Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–6 (2007).

100. Kosaka, T. *et al.* Analysis of epidermal growth factor receptor gene mutation in patients with non-small cell lung cancer and acquired resistance to gefitinib. *Clin. Cancer Res.* **12,** 5764–9 (2006).

101. Pao, W. *et al.* Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *PLoS Med.* **2,** e73 (2005).

102. Balak, M. N. *et al.* Novel D761Y and common secondary T790M mutations in epidermal growth factor receptor-mutant lung adenocarcinomas with acquired resistance to kinase inhibitors. *Clin. Cancer Res.* **12,** 6494–501 (2006).

103. Kobayashi, S. *et al.* EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352,** 786–92 (2005).

104. Daub, H., Specht, K. & Ullrich, A. Strategies to overcome resistance to targeted protein kinase inhibitors. *Nat. Rev. Drug Discov.* **3,** 1001–10 (2004).

105. Jänne, P. a *et al.* AZD9291 in EGFR Inhibitor-Resistant Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **372,** 1689–1699 (2015).

106. Cross, D. a E. *et al.* AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discov.* (2014). doi:10.1158/2159-8290.CD-14-0337

107. Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.* **22,** (2016).

108. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev.* **12,** 323–334 (2012).

109. Makino, S. Further evidence favoring the concept of the stem cell in ascites tumors of rats. *Ann. N. Y. Acad. Sci.* **63,** 818–30 (1956).

110. Fidler, I. J. & Kripke, M. L. Metastasis Results From Preexisting Variant Cells Within a Malignant-Tumor. *Science* **197,** 893–895 (1977).

111. Fidler, I. J. Tumor Heterogeneity and the Biology of Cancer Invasion and Metastasis1. *Cancer* 2651–2660 (1978).

112. Heppner, G. H. Tumor heterogeneity. *Cancer Res.* **44,** 2259–2265 (1984).

113. Marusyk, A. & Polyak, K. Tumor heterogeneity: Causes and consequences. *Biochim. Biophys. Acta - Rev. Cancer* **1805,** 105–117 (2010).

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

115. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414,** 105–11 (2001).

116. Visvader, J. E. & Lindeman, G. J. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nat. Rev. Cancer* **8,** 755–68 (2008).

117. Beck, B. & Blanpain, C. Unravelling cancer stem cell potential. *Nat. Rev. Cancer* **13,** 727–738 (2013).

118. Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature* **456,** 593–8 (2008).

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

120. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–8 (1976).

121. Mayr, E. Speciation and Macroevolution. *Evolution (N. Y).* **36,** 1119 (1982).

122. Graham, T. A. & Sottoriva, A. Measuring cancer evolution from the genome. *J. Pathol.* (2016). doi:10.1002/path.4821

123. Dietrich, M. R. Richard Goldschmidt: hopeful monsters and other 'heresies'. *Nat. Rev. Genet.* **4,** 68–74 (2003).

124. Rode, A., Maass, K. K., Willmund, K. V., Lichter, P. & Ernst, A. Chromothripsis in cancer cells: An update. *Int. J. cancer* **138,** 2322–33 (2016).

125. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. **153,** 666–677 (2014).

126. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).

127. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48,** 1–15 (2016).

128. Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Pnas* 1519556112- (2015). doi:10.1073/pnas.1519556112

129. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47,** 209–216 (2015).

130. Kimura, M. *The Neutral Theory of Molecular Evolution*. (1985).

131. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* 1–9 (2016). doi:10.1038/ng.3489

132. Weiss, L. Concepts of metastasis. *Cancer Metastasis Rev.* **19,** 219–234 (2000).

133. Valastyan, S. & Weinberg, R. A. Tumor metastasis: Molecular insights and evolving paradigms. *Cell* **147,** 275–292 (2011).

134. Nagrath, S. *et al.* Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450,** 1235–9 (2007).

135. Chambers, A. F., Groom, A. C. & MacDonald, I. C. Dissemination and growth of cancer cells in metastatic sites. *Nat. Rev. Cancer* **2,** 563–72 (2002).

136. Gupta, G. P. & Massagué, J. Cancer Metastasis: Building a Framework. *Cell* **127,** 679–695 (2006).

137. Steeg, P. S. Targeting metastasis. *Nat. Rev. Cancer* **16,** 201–218 (2016).

138. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–74 (2011).

139. Mekenkamp, L. J. M. *et al.* Clinicopathological features and outcome in advanced colorectal cancer patients with synchronous vs metachronous metastases. *Br. J. Cancer* **103,** 159–164 (2010).

140. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* **9,** 239–52 (2009).

141. Naxerova, K. & Jain, R. K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat. Rev. Clin. Oncol.* **12,** 258–272 (2015).

142. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255,** 197–200 (1975).

143. Foulds, L. The experimental study of tumor progression: a review. *Cancer Res.* **14,** 327–39 (1954).

144. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. *Science* **352,** 169–175 (2016).

145. Klein, C. A. Parallel progression of primary tumours and metastases. *Nat. Rev. Cancer* **9,** 302–12 (2009).

146. Belien, J. A. M. *et al.* Gross genomic damage measured by DNA image cytometry independently predicts gastric cancer patient survival. *Br. J. Cancer* **101,** 1011–1018 (2009).

147. Dunn, J. M. *et al.* Image cytometry accurately detects DNA ploidy abnormalities and predicts late relapse to high-grade dysplasia and adenocarcinoma in Barrett's oesophagus following photodynamic therapy. *Br. J. Cancer* **102,** 1608–1617 (2010).

148. Bol, M. G. W. *et al.* Correlation of Grade of Urothelial Cell Carcinomas and DNA Histogram Features Assessed by Flow Cytometry and Automated Image Cytometry. *Anal. Cell. Pathol.* **25,** 147–153 (2003).

149. Pfau, S. J. & Amon, A. Chromosomal instability and aneuploidy in cancer: from yeast to man. *EMBO Rep.* **13,** 515–527 (2012).

150. Danielsen, H. E., Pradhan, M. & Novelli, M. Revisiting tumour aneuploidy — the place of ploidy assessment in the molecular era. *Nat. Rev. Clin. Oncol.* (2015). doi:10.1038/nrclinonc.2015.208

151. Santaguida, S. & Amon, A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell Biol.* **16,** 473–485 (2015).

152. Täckholm, G. Zytologische Studien über die Gattung Rosa. *Acta Hort. Berg.* 97–381

153. Thorpe, P. H., González-Barrera, S. & Rothstein, R. More is not always better: the genetic constraints of polyploidy. *Trends Genet.* **23,** 263–266 (2007).

154. Nigg, E. A. Mitotic kinases as regulators of cell division and its checkpoints. *Nat. Rev. Mol. Cell Biol.* **2,** 21–32 (2001).

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

156. Cheeseman, I. M. & Desai, A. Molecular architecture of the kinetochore–microtubule interface. *Nat. Rev. Mol. Cell Biol.* **9,** 33–46 (2008).

157. Zasadil, L. M., Britigan, E. M. C. & Weaver, B. A. 2n or not 2n: Aneuploidy, polyploidy and chromosomal instability in primary and tumor cells. *Semin. Cell Dev. Biol.* **24,** 370–379 (2013).

158. Thompson, S. L. & Compton, D. A. Examining the link between chromosomal instability and aneuploidy in human cells. *J. Cell Biol.* **180,** 665–72 (2008).

159. Roschke, A. V, Stover, K., Tonon, G., Schäffer, A. A. & Kirsch, I. R. Stable karyotypes in epithelial cancer cell lines despite high rates of ongoing structural and numerical chromosomal instability. *Neoplasia* **4,** 19–31 (2002).

160. Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **10,** 478–87 (2009).

161. Davoli, T. *et al.* Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell* **155,** 948–962 (2013).

162. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68,** 820–3 (1971).

163. Sotillo, R. *et al.* Mad2 overexpression promotes aneuploidy and tumorigenesis in mice. *Cancer Cell* **11,** 9–23 (2007).

164. Sotillo, R., Schvartzman, J.-M., Socci, N. D. & Benezra, R. Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal. *Nature* **464,** 436–40 (2010).

165. Silk, A. D. *et al.* Chromosome missegregation rate predicts whether aneuploidy will promote or suppress tumors. *Proc. Natl. Acad. Sci.* **110,** E4134–E4141 (2013).

166. Zasadil, L. M. *et al.* High rates of chromosome missegregation suppress tumor progression but do not inhibit tumor initiation. *Mol. Biol. Cell* **27,** 1981–9 (2016).

167. Corver, W. E. & Cornelisse, C. J. Flow cytometry of human solid tumours: clinical and research applications. *Curr. Diagnostic Pathol.* **8,** 249–267 (2002).

168. Li, L. *et al.* Genomic instability and proliferative activity as risk factors for distant metastases in breast cancer. *Br. J. Cancer* **99,** 513–519 (2008).

169. Hveem, T. S. *et al.* Prognostic impact of genomic instability in colorectal cancer. *Br. J. Cancer* **110,** 2159–2164 (2014).

170. Choma, D., Daurès, J. P., Quantin, X. & Pujol, J. L. Aneuploidy and prognosis of non-small-cell lung cancer: a meta-analysis of published data. *Br. J. Cancer* **85,** 14–22 (2001).

171. Summary of Comments on Clonal nature of cancer _ Ruiz _ 2011. 2011–2012

172. Krishan, A. & Dandekar, P. D. DAPI fluorescence in nuclei isolated from tumors. *J. Histochem. Cytochem.* **53,** 1033–6 (2005).

173. Corver, W. E. & ter Haar, N. T. High-resolution multiparameter DNA flow cytometry for the detection and sorting of tumor and stromal subpopulations from paraffin-embedded tissues. *Curr. Protoc. Cytom.* **Chapter 6,** Unit 6.27 (2009).

174. Juskevicius, D. *et al.* Extracavitary primary effusion lymphoma: clinical, morphological, phenotypic and cytogenetic characterization using nuclei enrichment technique. *Histopathology* 1–14 (2014).

175. Shin, J. *et al.* Sox10 is expressed in primary melanocytic neoplasms of various histologies but not in fibrohistiocytic proliferations and histiocytoses. *J Am Acad Dermatol* **67,** 717–726 (2012).

176. Buonaccorsi, J. N., Prieto, V. G., Torres-Cabala, C., Suster, S. & Plaza, J. A. Diagnostic Utility and Comparative Immunohistochemical Analysis of MITF-1 and SOX10 to Distinguish Melanoma In Situ and Actinic Keratosis. *Am. J. Dermatopathol.* **36,** 124–130 (2014).

177. Harris, M. L., Baxter, L. L., Loftus, S. K. & Pavan, W. J. Sox proteins in melanocyte development and melanoma. *Pigment Cell Melanoma Res.* **23,** 496–513 (2010).

178. Mohamed, A. & Gonzalez, R. Tumor Stem Cells (CD271, c-kit, SOX10) in Melanomas: Prognostic and Outcome Implications. *Appl Immunohistochem Mol Morphol* **22,** 142–145 (2014).

179. Clevenger, J., Joseph, C., Dawlett, M., Guo, M. & Gong, Y. Reliability of immunostaining using pan-melanoma cocktail, SOX10, and microphthalmia transcription factor in confirming a diagnosis of melanoma on fine-needle aspiration smears. *Cancer Cytopathol.* **122,** 779–85 (2014).

180. Boggaram, V. Thyroid transcription factor-1 (TTF-1/Nkx2.1/TITF1) gene regulation in the lung. *Clin. Sci.* **116,** 27–35 (2009).

181. Yatabe, Y., Mitsudomi, T. & Takahashi, T. TTF-1 expression in pulmonary adenocarcinomas. *Am. J. Surg. Pathol.* **26,** 767–73 (2002).

182. Arriola, E. *et al.* Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. *Lab. Invest.* **87,** 75–83 (2007).

183. Huang, J., Pang, J., Watanabe, T., Ng, H.-K. & Ohgaki, H. Whole Genome Amplification for Array Comparative Genomic Hybridization Using DNA Extracted from Formalin-Fixed, Paraffin-Embedded Histological Sections. *J. Mol. Diagnostics* **11,** 109–116 (2009).

184. Nilsen, G., Liestol, K. & Lingjaerde, O. copynumber: Segmentation of single- and multi-track copy number data by penalized least squares regression. R package version 3.4.

185. LifeTechnologies. Ion AmpliSeq™ Library Preparation. *Cat. Number* **4472395,** Publication Part Number 4472261 Rev. E

186. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

187. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

188. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

189. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

190. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–4 (2003).

191. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–9 (2010).

192. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7,** 575–6 (2010).

193. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.* **34,** 57–65 (2013).

194. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7,** e46688 (2012).

195. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: Computational prediction of driver missense mutations. *Cancer Res.* **69,** 6660–6667 (2009).

196. Andor, N., Harness, J. V., M??ller, S., Mewes, H. W. & Petritsch, C. Expands: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30,** 50–60 (2014).

197. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22,** 105–113 (2015).

198. Watanabe, H. *et al.* Integrated cistromic and expression analysis of amplified NKX2-1 in lung adenocarcinoma identifies LMO3 as a functional transcriptional target. *Genes Dev.* **27,** 197–210 (2013).

199. www.abcam.com/SOX10-antibody-EPR4007-ab155279.pdf.

200. Harris, T. *et al.* Both Gene Amplification and Allelic Loss Occur at 14q13.3 in Lung Cancer. *Clin. Cancer Res.* **17,** 690–699 (2011).

201. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150,** 251–63 (2012).

202. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.* **43,** 442–446 (2011).

203. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485,** 502–506 (2012).

204. Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat. Genet.* **44,** 1006–1014 (2012).

205. Andor, N., Harness, J. V., Müller, S., Mewes, H. W. & Petritsch, C. Expands: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30,** 50–60 (2014).

206. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,**

1069–1075 (2008).

207. Weir, B. a *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450,** 893–898 (2007).

208. Knudsen, E. S. & Wang, J. Y. J. Targeting the RB-pathway in Cancer Therapy. *Clin. Cancer Res.* **16,** 1094–1099 (2010).

209. Nagayama, K. *et al.* Homozygous deletion scanning of the lung cancer genome at a 100-kb resolution. *Genes, Chromosom. Cancer* **46,** 1000–1010 (2007).

210. Mlakar, V. *et al.* Presence of activating KRAS mutations correlates significantly with expression of tumour suppressor genes DCN and TPM1 in colorectal cancer. *BMC Cancer* **9,** 282 (2009).

211. Paige, A. J. W. *et al.* WWOX: A candidate tumor suppressor gene involved in multiple tumor types. *Proc. Natl. Acad. Sci.* **98,** 11417–11422 (2001).

212. Yang, C.-Y., Lin, M.-W., Chang, Y.-L., Wu, C.-T. & Yang, P.-C. Programmed cell death-ligand 1 expression is associated with a favourable immune microenvironment and better overall survival in stage I pulmonary squamous cell carcinoma. *Eur. J. Cancer* **57,** 91–103 (2016).

213. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38,** 468–73 (2006).

214. Solomon, D. a. *et al.* Mutational inactivation of PTPRD in glioblastoma multiforme and malignant melanoma. *Cancer Res.* **68,** 10300–10306 (2008).

215. Kumar, R. *et al.* A single nucleotide polymorphism in the 3'untranslated region of the CDKN2A gene is common in sporadic primary melanomas but mutations in the CDKN2B, CDKN2C, CDK4 and p53 genes are rare. *Int. J. cancer* **95,** 388–93 (2001).

216. Oxnard, G. R. *et al.* Natural history and molecular characteristics of lung cancers harboring EGFR exon 20 insertions. *J. Thorac. Oncol.* **8,** 179–184 (2013).

217. Bignell, G. R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14,** 287–95 (2004).

218. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.* **18,** 1001–5 (2000).

219. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* **10,** 1126–37 (2000).

220. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20,** 207–11 (1998).

221. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40,** 722–9 (2008).

222. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6,** 99–103 (2009).

223. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30,** 413–421 (2012).

224. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14,** R80 (2013).

225. Ruiz, C. & Lenkiewicz, E. Advancing a clinically relevant perspective of the clonal nature of cancer. *Proc. Natl. Acad. Sci.* **108,** 12054–12059 (2011).

226. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).

227. Holley, T. *et al.* Deep Clonal Profiling of Formalin Fixed Paraffin Embedded Clinical Samples. *PLoS One* **7,** 1–11 (2012).

228. Jones, S. *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7,** 283ra53 (2015).

229. Barrett, M. T. *et al.* Clonal Evolution and Therapeutic Resistance in Solid Tumors. *Front. Pharmacol.* **4,** (2013).

230. Polackova, J. & Hemmer, J. Progression of diploid tumor cells in aneuploid head and neck squamous cell carcinomas. *Int. J. Oncol.* **15,** 315–20 (1999).

231. Axelrod, R., Axelrod, D. E. & Pienta, K. J. Evolution of cooperation among tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 13474–13479 (2006).

232. de Jong, J. S., van Diest, P. J., van der Valk, P. & Baak, J. P. Expression of growth factors, growth-inhibiting factors, and their receptors in invasive breast cancer. II: Correlations with proliferation and angiogenesis. *J. Pathol.* **184,** 53–7 (1998).

233. Royuela, M. *et al.* Immunohistochemical analysis of the IL-6 family of cytokines and their receptors in benign, hyperplasic, and malignant human prostate. *J. Pathol.* **202,** 41–9 (2004).

234. Wangari-Talbot, J. & Chen, S. Genetics of melanoma. *Front. Genet.* **3,** 330 (2012).

235. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144,** 27–40 (2011).

236. Hirsch, D. *et al.* Chromothripsis and focal copy number alterations determine poor outcome in malignant melanoma. *Cancer Res.* **73,** 1454–1460 (2013).

237. Shain, a. H. *et al.* The Genetic Evolution of Melanoma from Precursor Lesions. *N. Engl. J. Med.* **373,** 1926–1936 (2015).

238. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–8 (2005).

239. Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17,** 443–54 (2010).

240. John, J. & Powell, K. TMPRSS2-ERG fusion gene expression in prostate tumor cells and its clinical and biological significance in prostate cancer progression. *J. cancer …* **4,** 94–101 (2012).

241. Yatabe, Y., Matsuo, K. & Mitsudomi, T. Heterogeneous distribution of EGFR mutations is extremely rare in lung adenocarcinoma. *J. Clin. Oncol.* **29,** 2972–2977 (2011).

242. Sherwood, J., Dearden, S., Ratcliffe, M. & Walker, J. Mutation status concordance between primary lesions and metastatic sites of advanced non-small-cell lung cancer and the impact of mutation testing methodologies: a literature review. *J. Exp. Clin. Cancer Res.* **34,** 92 (2015).

243. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346,** 251–256 (2014).

244. Hata, A. N. *et al.* Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat. Med.* (2016). doi:10.1038/nm.4040

245. Sottoriva, A. & Graham, T. A pan-cancer signature of neutral tumor evolution. *bioRxiv* 14894 (2015). doi:10.1101/014894

246. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat. Rev. Genet.* **13,** 189–203 (2012).

247. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–13 (2010).

248. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46,** 225–33 (2014).

249. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **161,** 1215–1228 (2015).

250. Sylvester, B. E. & Vakiani, E. Tumor evolution and intratumor heterogeneity in colorectal carcinoma: insights from comparative genomic profiling of primary tumors and matched metastases. *J. Gastrointest. Oncol.* **6,** 668–75 (2015).

251. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).

252. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Pnas* **110,** 21083–8 (2013).

253. Albertson, D. G., Collins, C., McCormick, F. & Gray, J. W. Chromosome aberrations in solid tumors. *Nat. Genet.* **34,** 369–76 (2003).

254. Chen, T. R., Drabkowski, D., Hay, R. J., Macy, M. & Peterson, W. WiDr is a derivative of another colon adenocarcinoma cell line, HT-29. *Cancer Genet. Cytogenet.* **27,** 125–34 (1987).

255. Macville, M. *et al.* Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer Res.* **59,** 141–50 (1999).

256. Knutsen, T. *et al.* Amplification of 4q21-q22 and the MXR gene in independently derived mitoxantrone-resistant cell lines. *Genes. Chromosomes Cancer* **27,** 110–6 (2000).

257. Li, M. *et al.* The ATM – p53 pathway suppresses aneuploidy- induced tumorigenesis. *Pnas* **107,** 14188–14193 (2010).

258. Lane, D. P. Cancer. p53, guardian of the genome. *Nature* **358,** 15–6 (1992).

259. Kruse, J.-P. & Gu, W. Modes of p53 regulation. *Cell* **137,** 609–22 (2009).

260. Moll, U. M. & Petrenko, O. The MDM2-p53 interaction. *Mol. Cancer Res.* **1,** 1001–8 (2003).

261. Williams, B. R. *et al.* Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science* **322,** 703–9 (2008).

262. Torres, E. M. *et al.* Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317,** 916–24 (2007).

263. Shlien, A. *et al.* Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 11264–9 (2008).

264. Fridlyand, J. *et al.* Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* **6,** 96 (2006).

265. Ho, C. C., Hau, P. M., Marxer, M. & Poon, R. Y. C. The requirement of p53 for maintaining chromosomal stability during tetraploidization. *Oncotarget* **1,** 583–95 (2010).

266. Ghadimi, B. M. *et al.* Centrosome amplification and instability occurs exclusively in aneuploid, but not in diploid colorectal cancer cell lines, and correlates with numerical chromosomal aberrations. *Genes. Chromosomes Cancer* **27,** 183–90 (2000).

267. Campomenosi, P. *et al.* p53 Mutations and DNA Ploidy in Colorectal Adenocarcinomas. *Anal. Cell. Pathol.* **17,** 1–12 (1998).

268. Andreassen, P. R., Lohez, O. D., Lacroix, F. B. & Margolis, R. L. Tetraploid state induces p53-dependent arrest of nontransformed mammalian cells in G1. *Mol. Biol. Cell* **12,** 1315–28 (2001).

269. Dalton, W. B., Yu, B. & Yang, V. W. p53 suppresses structural chromosome instability after

mitotic arrest in human cells. *Oncogene* **29,** 1929–1940 (2010).

270. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148,** 59–71 (2012).

271. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).

272. Brastianos, P. K. *et al.* Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets. *Cancer Discov.* **5,** 1164–1177 (2015).

273. Zhao, Z. *et al.* Early and multiple origins of metastatic lineages within primary tumors. 1–6 (2016). doi:10.1073/pnas.1525677113

274. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–7 (2010).

275. Lanuti, M. Surgical Management of Oligometastatic Non-Small Cell Lung Cancer. *Thorac. Surg. Clin.* **26,** 287–94 (2016).

276. Guerrero, E. & Ahmed, M. The role of stereotactic ablative radiotherapy (SBRT) in the management of oligometastatic non small cell lung cancer. *Lung cancer* **92,** 22–8 (2016).

277. Duelli, D. & Lazebnik, Y. Cell fusion: A hidden enemy? *Cancer Cell* **3,** 445–448 (2003).

278. Pawelek, J. M. & Chakraborty, A. K. Chapter 10 The Cancer Cell-Leukocyte Fusion Theory of Metastasis. *Adv. Cancer Res.* **101,** 397–444 (2008).

279. Pawelek, J. M. & Chakraborty, A. K. Fusion of tumour cells with bone marrow-derived cells: a unifying explanation for metastasis. *Nat. Rev. Cancer* **8,** 377–86 (2008).

280. Lu, X. & Kang, Y. Cell Fusion as a Hidden Force in Tumor Progression. *Cancer Res.* **69,** 8536–8539 (2009).

281. Bjerregaard, B., Holck, S., Christensen, I. J. & Larsson, L. I. Syncytin is involved in breast cancer-endothelial cell fusions. *Cell. Mol. Life Sci.* **63,** 1906–1911 (2006).

282. Yan, B., Wang, J. & Liu, L. Chemotherapy promotes tumour cell hybridization in vivo. *Tumor Biol.* **37,** 5025–5030 (2016).

283. Zhou, X. *et al.* Cell Fusion Connects Oncogenesis with Tumor Evolution. *Am. J. Pathol.* **185,** (2015).

284. Yilmaz, Y., Lazova, R., Qumsiyeh, M., Cooper, D. & Pawelek, J. Donor Y chromosome in renal carcinoma cells of a female BMT recipient: visualization of putative BMT-tumor hybrids by FISH. *Bone Marrow Transpl.* **35,** 1021–1024 (2005).

285. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–21 (1992).

286. UyBico, S. J., Carol Wu, B. C., Robert Suh, B. D. & Nanette, B. H. Lung Cancer Staging Essentials: The New TNM Staging System and Po- tential Imaging Pitfalls. *RadioGraphics* **30,** 1163–1181 (2010).

Remark: This thesis was corrected for spelling and grammar only by a native speaker, as authorized by Prof. Dr. Erich Nigg.

Supplementary Figures

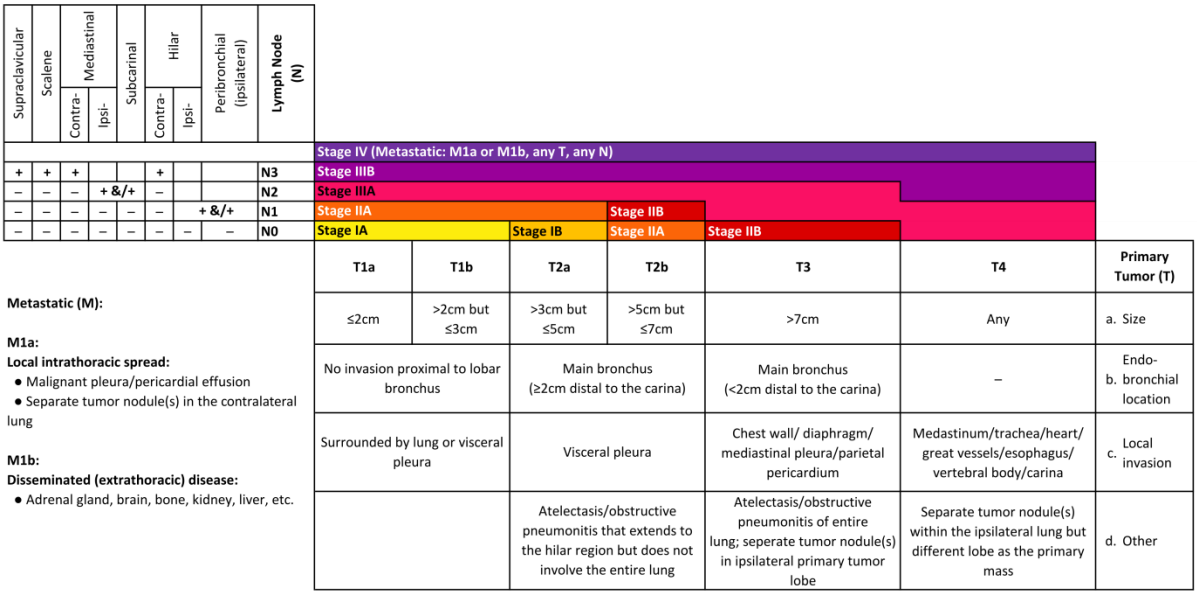


Figure S1 | Characteristics of the seventh edition of the TNM staging system for lung cancer (adapted from ref[286] and modified).

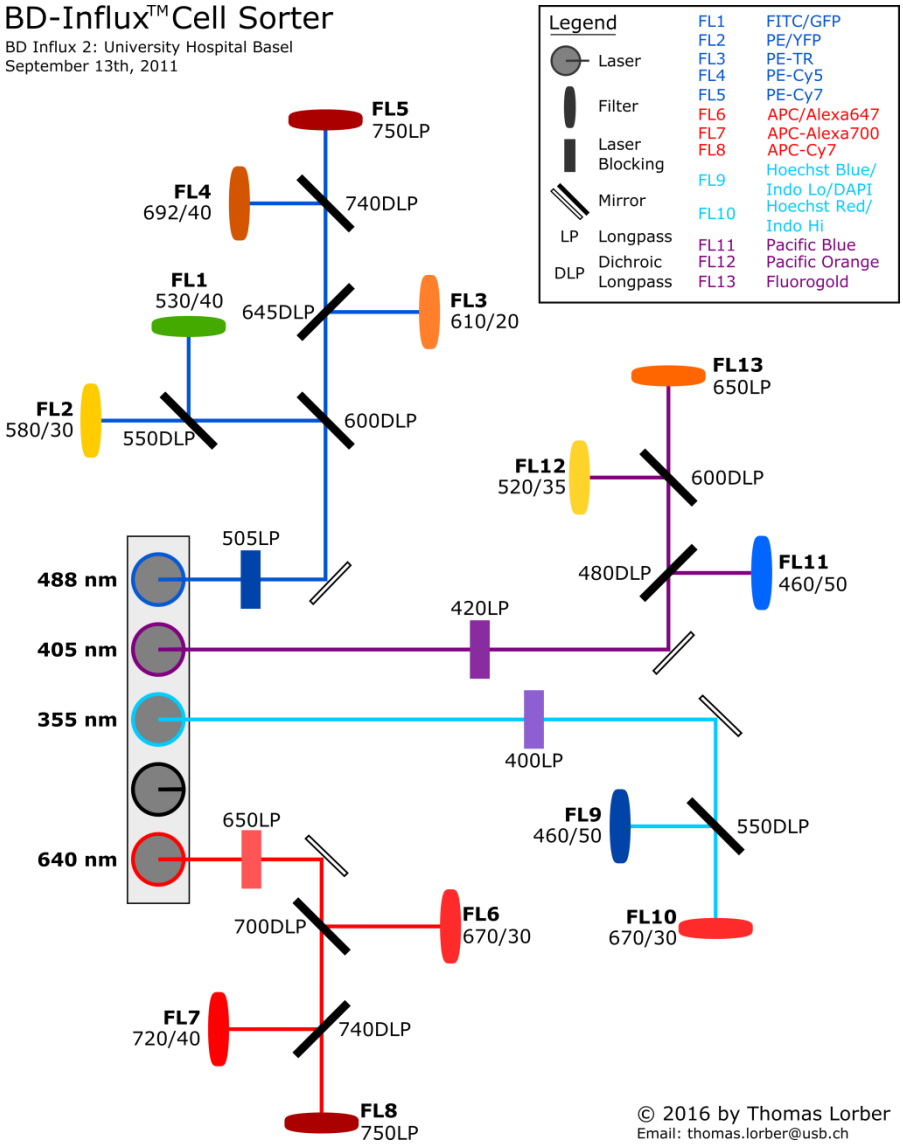


Figure S2 | Filter guide of the sorting device. A BD-Influx™ Cell Sorter (University Hospital Basel Nr.2) was used for all flow-sorting experiments. DAPI (DNA content) intensity was quantified with the FL9 460/50 and TTF1 or SOX10 intensity via the secondary Alexa647-coubled antibody with FL6 670/30.

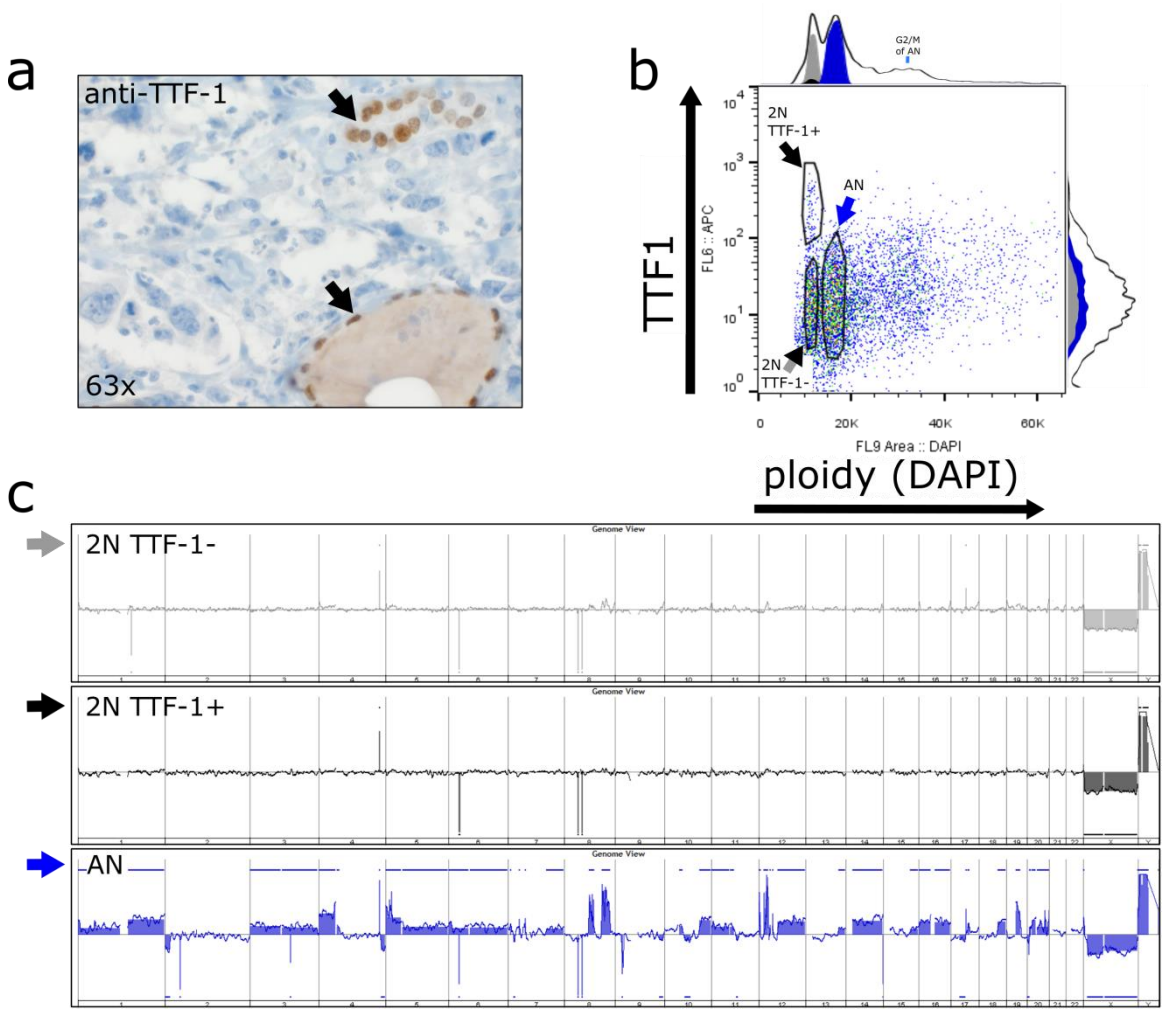


Figure S3 | Thyroid gland metastasis of a TTF1-negative LUAD. (a) IHC for TTF1. TTF1 is expressed by normal thyroid follicular cells. (b) Multiparameter flow sorting detected these cells as a diploid, TTF1-positive (2N TTF1+) population. Consequently, no SCNAs were present in this population (c). In contrast, the aneuploid population (AN) carried many SCNAs. *Note: These are data from patient 12 from the LUAD cohort from Part B of this thesis.*

					Tumor free	Primary		Metastasis			
gene	chrom	pos	ref	var	control	2N	AN	AN	Type	effect	Protein change
TGM7	chr15	43572025	C	T	0.00	0.00	0.00	0.11	somatic	synonymous	p.(=)
PAX8	chr2	113999253	C	G	0.00	0.00	0.02	0.07		missense	p.Gly218Arg|
EGFR	chr7	55224453	G	A	0.00	0.00	0.05	0.00		missense	p.Asp379Asn
SYNE1	chr6	152599416	C	T	0.00	0.00	0.13	0.00		splicesite_5	
TAF1	chrX	70613116	A	G	0.01	0.00	0.20	0.00		intronic	
LRP1B	chr2	141116452	G	T	0.00	0.00	0.20	0.35		nonsense	p.Ser3732Ter
STK36	chr2	219556972	T	A	0.00	0.01	0.23	0.34		missense	p.Leu624His
LRP1B	chr2	141083420	C	A	0.00	0.01	0.26	0.36		missense	p.Arg4084Leu
RUNX1T1	chr8	93088204	T	G	0.00	0.00	0.32	0.13		intronic	
PTPRT	chr20	41385157	C	G	0.00	0.00	0.33	0.12		synonymous	p.(=)
KIT	chr4	55603272	GG	TT	0.00	0.02	0.55	0.65		intronic	
MAGI1	chr3	65361515	T	C	0.00	0.01	0.57	0.55		missense	p.Asn1034Ser
GNAS	chr20	57429939	C	A	0.00	0.01	0.61	0.31		missense	p.Pro540Gln
NOTCH2	chr1	120464266	C	G	0.00	0.02	0.61	0.72		intronic	
EPHA3	chr3	89176389	GG	TT	0.00	0.01	0.69	0.46		missense	p.Gly40Val
PAX8	chr2	113999309	A	C	0.00	0.00	0.70	0.60		intronic	
STK11	chr19	1218499	TATATCC TTTCCGG	T	0.00	0.01	0.93	0.94	somatic LOH	splicesite_3	
TLR4	chr9	120474950	C	A	0.00	0.01	0.94	0.95		missense	p.Leu182Ile
JAK3	chr19	17950389	G	T	0.00	0.01	0.96	0.94		synonymous	p.(=)
RET	chr10	43607633	C	A	0.00	0.01	0.96	0.96		missense	p.Pro537Thr
FLT3	chr13	28608447	C	A	0.00	0.02	0.96	0.98		missense	p.Lys565Asn
TRRAP	chr7	98508084	A	G	0.00	0.01	0.96	0.93		intronic	
NF1	chr17	29663678	TTG	T	0.00	0.03	0.98	0.96		frameshiftDel	p.Asp2059fs
TP53	chr17	7578370	C	A	0.00	0.02	0.98	0.98		splicesite_3	
KEAP1	chr19	10602736	A	T	0.00	0.02	0.99	0.98		missense	p.Leu281Gln
RET	chr10	43610119	G	A	0.44	0.57	0.03	0.02	germline LOH	missense	p.Gly691Ser
NF1	chr17	29679246	G	A	0.48	0.51	0.01	0.01		intronic	
SYNE1	chr6	152694184	T	C	0.46	0.47	0.01	0.01		synonymous	p.(=)
JAK3	chr19	17942412	G	A	0.44	0.53	1.00	0.99	germline LOH	intronic	
PTPRD	chr9	8465598	A	G	0.55	0.65	1.00	0.99		synonymous	
PER1	chr17	8046772	C	G	0.45	0.49	0.99	1.00		missense	p.Ala962Pro

Figure S4 | Comparison of mutations in DNA of sorted diploid, SCNA-negative populations versus tumor-free control DNA, and DNA of sorted tumor populations. Somatic and selected germline variants are included to strengthen the fact that both germline and somatic mutations with VAF > 80% can be detected in sorted tumor populations. *Note: These are data from patient 48 from the LUAD cohort from Part B of this thesis that was sorted with DAPI for ploidy only.*

					tumor free	primary				metastasis		
chrom	posiition	ref	var	gene	control	2N TTF1-	2N TTF1+	AN1 TTF1+	AN2 TTF1+	AN TTF1+	effect	Protein change
chr7	55249013	CA	CGGTT	EGFR	0.00	0.03	0.60	0.49	0.54	0.47	missense	N771delinsGY
chr3	125216154	A	T	SNX4	0.00	0.01	0.58	0.39	0.41	0.58	intronic	
chr1	76343854	A	T	MSH4	0.00	0.03	0.53	0.53	0.62	0.23	missense	K464M
chr5	74364443	T	C	ANKRD31	0.00	0.02	0.53	0.51	0.64	0.61	missense	I1841M
chr1	247587135	T	G	NLRP3	0.00	0.03	0.53	0.49	0.43	0.71	intronic	
chr19	55814161	C	T	BRSK1	0.00	0.02	0.53	0.35	0.40	0.42	synonymous	D318D
chr6	32188296	CG	TC	NOTCH4	0.01	0.02	0.52	0.48	0.54	0.57	nonframeshift	_348_
chr2	227985767	C	T	COL4A4	0.00	0.02	0.52	0.50	0.47	0.37	missense	R97H
chr17	47917211	C	A	TAC4	0.00	0.02	0.51	0.32	0.39	0.45	nonframeshift	_104_
chr22	47108061	T	C	CERK	0.00	0.02	0.50	0.48	0.50	0.48	intronic	
chr6	135318637	C	T	HBS1L	0.00	0.02	0.50	0.50	0.52	0.61	missense	A233T
chr2	25464506	G	A	DNMT3A	0.00	0.02	0.50	0.51	0.45	0.61	synonymous	S669S
chr10	96066356	A	G	PLCE1	0.00	0.02	0.49	0.30	0.26	0.31	missense	D1932G
chr22	45599939	G	A	KIAA0930	0.00	0.02	0.49	0.53	0.51	0.45	intronic	
chr6	137018461	C	A	MAP3K5	0.00	0.02	0.49	0.49	0.52	0.59	nonsense	E291*
chr1	6635338	C	T	TAS1R1	0.01	0.02	0.48	0.48	0.40	0.69	synonymous	S382S
chr9	135523843	C	T	DDX31	0.00	0.03	0.47	0.54	0.54	0.47	nonsense	W384*
chr19	39879285	C	T	PAF1	0.00	0.02	0.47	0.64	0.63	0.45	missense	V258M
chr4	95529130	G	C	PDLIM5	0.00	0.02	0.47	0.48	0.38	0.47	intronic	
chr1	75097620	C	T	C1orf173	0.01	0.02	0.46	0.46	0.54	0.17	intronic	
chr22	46657446	C	T	PKDREJ	0.00	0.02	0.44	0.44	0.47	0.36	missense	V592I
chr12	77216280	C	G	ZDHHC17	0.00	0.02	0.44	0.95	0.65	0.37	missense	P289R
chr5	169653	G	C	PLEKHG4B	0.00	0.06	0.37	0.43	0.41	0.42	missense	E869D
chr5	171181	G	A	PLEKHG4B	0.00	0.06	0.36	0.48	0.46	0.40	synonymous	V895V
chr5	840689	G	A	ZDHHC11	0.00	0.04	0.27	0.39	0.35	0.32	synonymous	I235I
chr5	1432755	G	A	SLC6A3	0.00	0.06	0.23	0.28	0.27	0.24	synonymous	I159I
chr6	7586026	G	C	DSP	0.01	0.02	0.58	0.03	0.23	0.44	missense	G2844A
chr17	48680396	C	A	CACNA1G	0.00	0.02	0.51	0.01	0.17	0.20	nonsense	C1335*
chr19	50266506	G	A	TSKS	0.01	0.01	0.50	0.02	0.20	0.25	intronic	
chr6	24843406	T	G	FAM65B	0.00	0.02	0.49	0.02	0.20	0.32	missense	N535T
chr11	65386152	G	T	PCNXL3	0.00	0.02	0.49	0.02	0.16	0.20	missense	R440L
chr2	219920221	G	T	IHH	0.00	0.01	0.49	0.02	0.14	0.26	missense	A315D
chr1	161720018	G	A	DUSP12	0.00	0.01	0.48	0.02	0.18	0.25	intronic	
chr12	8086395	C	T	SLC2A3	0.00	0.01	0.48	0.02	0.14	0.24	intronic	
chr21	31973521	G	A	KRTAP22-1	0.00	0.02	0.48	0.03	0.24	0.31	missense	G28S
chr17	10399462	C	A	MYH1	0.00	0.02	0.48	0.03	0.19	0.24	missense	Q1658H
chrX	40547811	C	T	MED14	0.00	0.01	0.48	0.02	0.16	0.24	intronic	
chr15	63943359	G	A	HERC1	0.00	0.01	0.47	0.03	0.22	0.44	intronic	
chr1	1391200	C	T	ATAD3C	0.00	0.01	0.47	0.02	0.20	0.26	synonymous	I156I
chr9	131389751	C	T	SPTAN1	0.00	0.02	0.46	0.02	0.23	0.20	synonymous	A2221A
chr20	4167900	C	T	SMOX	0.00	0.01	0.46	0.02	0.14	0.20	intronic	
chr16	68087533	A	G	DUS2L	0.01	0.01	0.46	0.02	0.18	0.18	missense	E80G
chr17	15995167	A	G	NCOR1	0.00	0.01	0.46	0.02	0.20	0.22	intronic	
chr17	43227159	C	T	HEXIM1	0.00	0.01	0.46	0.03	0.20	0.25	missense	A201V
chr1	228480445	G	T	OBSCN	0.00	0.01	0.45	0.02	0.18	0.21	missense	V4038L
chr7	98574357	G	A	TRRAP	0.00	0.01	0.45	0.03	0.22	0.28	synonymous	E2730E
chr3	197505232	G	A	FYTTD1	0.00	0.01	0.45	0.02	0.21	0.30	missense	R252H
chr14	21822582	C	T	SUPT16H	0.00	0.01	0.45	0.02	0.13	0.14	synonymous	E926E
chr11	55606663	C	T	OR5D16	0.00	0.01	0.43	0.02	0.16	0.16	synonymous	L146L
chr12	125448910	C	A	DHX37	0.00	0.01	0.40	0.03	0.23	0.36	intronic	
chr18	21042843	T	A	RIOK3	0.00	0.01	0.01	0.91	0.52	0.30	intronic	
chr14	92413532	G	C	FBLN5	0.00	0.03	0.03	0.74	0.62	0.67	intronic	
chr3	127298649	C	T	TPRA1	0.00	0.01	0.01	0.53	0.31	0.00	synonymous	K64K
chr20	62168684	G	A	PTK6	0.00	0.01	0.02	0.52	0.35	0.66	intronic	
chr1	1248321	G	A	CPSF3L	0.00	0.01	0.01	0.50	0.43	0.00	synonymous	V386V
chr20	39801506	A	G	PLCG1	0.00	0.01	0.01	0.48	0.33	0.18	missense	I1090V
chr9	108226479	A	T	FSD1L	0.00	0.01	0.01	0.48	0.25	0.22	intronic	
chr2	236957919	C	G	AGAP1	0.00	0.01	0.01	0.47	0.35	0.30	missense	S703C
chr11	61488236	G	A	DAGLA	0.01	0.01	0.01	0.46	0.29	0.44	missense	G61S
chrX	153581892	G	A	FLNA	0.00	0.01	0.01	0.43	0.30	0.29	intronic	
chr3	183521844	C	T	YEATS2	0.00	0.00	0.01	0.33	0.18	0.30	synonymous	L1218L
chr10	111640041	A	G	XPNPEP1	0.00	0.01	0.09	0.01	0.13	0.00	intronic	
chr12	78444873	T	G	NAV3	0.01	0.00	0.10	0.02	0.19	0.00	missense	M821R
chr4	103553158	G	A	MANBA	0.00	0.01	0.10	0.01	0.12	0.00	intronic	
chr18	5969456	C	T	L3MBTL4	0.00	0.00	0.22	0.01	0.02	0.30	missense	C526Y
chr12	13716930	C	T	GRIN2B	0.00	0.00	0.06	0.00	0.00	0.21	missense	R1081H
chr20	22563123	A	T	FOXA2	0.00	0.00	0.19	0.01	0.01	0.22	missense	C253S
chrX	138896979	T	A	ATP11C	0.00	0.01	0.28	0.01	0.04	0.20	intronic	
chr8	113318349	C	T	CSMD3	0.00	0.00	0.06	0.00	0.02	0.00	missense	R2653Q
chr4	68380244	G	T	CENPC1	0.00	0.01	0.00	0.05	0.17	0.00	missense	T331K
chr2	47388844	T	G	CALM2	0.00	0.01	0.01	0.45	0.33	0.00	intronic	
chr19	44127609	C	T	CADM4	0.00	0.01	0.00	0.28	0.18	0.00	intronic	
chrX	70597548	G	C	TAF1	0.00	0.00	0.00	0.16	0.01	0.00	synonymous	L290L
chrX	148579485	C	T	IDS	0.00	0.00	0.00	0.04	0.01	0.00	intronic	
chr6	26091568	G	C	HFE	0.00	0.00	0.00	0.05	0.01	0.00	missense	G123R
chr11	32439212	A	G	WT1	0.00	0.00	0.00	0.01	0.13	0.00	intronic	
chr17	32963238	C	A	TMEM132E	0.00	0.00	0.00	0.01	0.10	0.00	intronic	
chr3	46786261	C	A	PRSS45	0.00	0.00	0.02	0.01	0.12	-	intronic	
chr12	133225522	T	C	POLE	0.00	0.00	0.00	0.01	0.12	0.00	missense	Y1381C
chr1	248201859	G	C	OR2L2	0.00	0.00	0.00	0.01	0.18	0.00	missense	G97A
chr16	70697559	A	G	MTSS1L	0.00	0.00	0.00	0.01	0.08	0.00	intronic	
chr8	143832527	G	A	LYPD2	0.00	0.00	0.00	0.01	0.09	0.00	synonymous	I40I
chr12	48723799	A	C	H1FNT	0.00	0.00	0.00	0.01	0.11	0.00	missense	E242A
chr22	17072662	C	T	CCT8L2	0.00	0.00	0.00	0.01	0.08	0.00	missense	R260H
chr5	35039551	T	C	AGXT2	0.00	0.00	0.00	0.01	0.09	0.00	synonymous	A80A
chr17	37809912	G	A	STARD3	0.01	0.00	0.00	0.00	0.00	0.22	missense	R43Q
chr2	73450750	A	T	SMYD5	0.00	0.00	0.00	0.00	0.00	0.29	intronic	
chr1	161180356	C	G	NDUFS2	0.00	0.00	0.01	0.02	0.01	0.27	intronic	
chr1	161180352	G	A	NDUFS2	0.01	0.02	0.02	0.03	0.02	0.25	intronic	
chrX	10422873	T	A	MID1	0.00	0.00	0.00	0.01	0.01	0.24	intronic	
chr19	55294980	C	T	KIR2DL1	0.00	0.00	0.00	0.00	0.00	0.21	missense	A287V
chr4	84368104	T	C	HELQ	0.00	0.01	0.01	0.03	0.03	0.29	missense	K426E
chr4	22820431	G	A	GBA3	0.00	0.00	0.00	0.00	0.00	0.08	missense	R433Q
chr5	169140992	G	T	DOCK2	0.00	0.00	0.00	0.00	0.00	0.22	intronic	
chr4	109863332	C	A	COL25A1	0.00	0.00	0.00	0.00	0.00	0.24	intronic	
chr11	46407750	C	T	CHRM4	0.01	0.00	0.00	0.01	0.01	0.20	missense	V120I

Figure S5 | Comparison of mutations in DNA of sorted diploid, SCNA-negative populations versus tumor-free control DNA and DNA of sorted tumor populations. *Note: These are data from patient 42 with a multiploid primary tumor from the LUAD cohort from Part B of this thesis.*

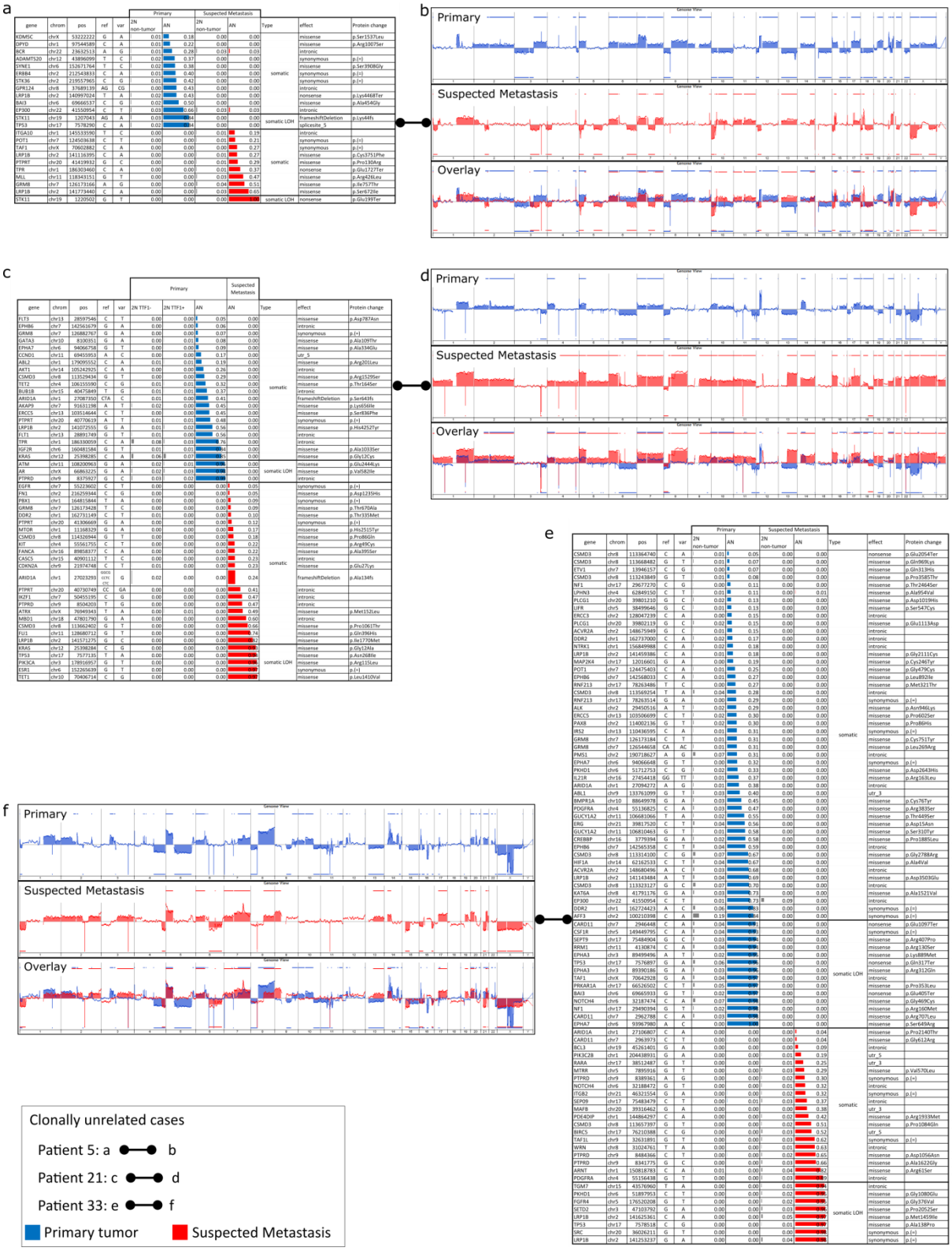


Figure S6 | Clonally unrelated LUAD cases. Three patients were excluded from the cohort of Part B (clonally related LUADs). In all three patients, not a single breakpoint (b,d,f) or SNV (a,c,e) was shared between the two tumors in each patient. Note that whole chromosomes or whole chromosome arms might be gained or lost in both lesions. However, they were not considered evidence for a clonal relationship. A clonal relationship requires the presence of SNVs and intra-chromosomal breaks, and both were absent in these three patients.

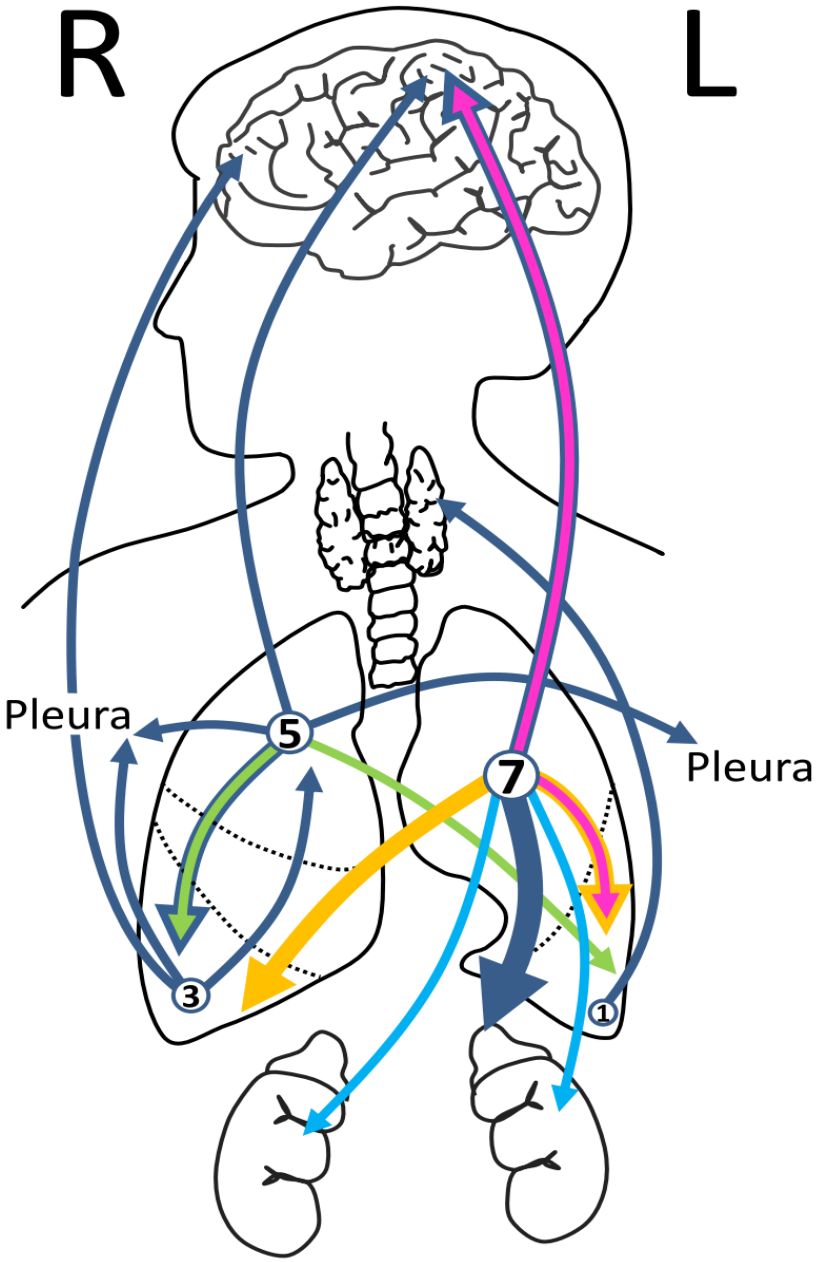


Figure S7 | Localization of primary tumors and their metastases for the 16 patients of the clonally related cohort. The colors indicate the four patients with multiple metastases: patient 20, light blue; patient 22, pink; patient 31, green; patient 36 orange. Note that in these four patients the arrows do not indicate a directionality or relationship of any kind. The numbers correlate with the size of the circles and indicate the number of patients (n = 16) with primary tumors at a specific lobe.

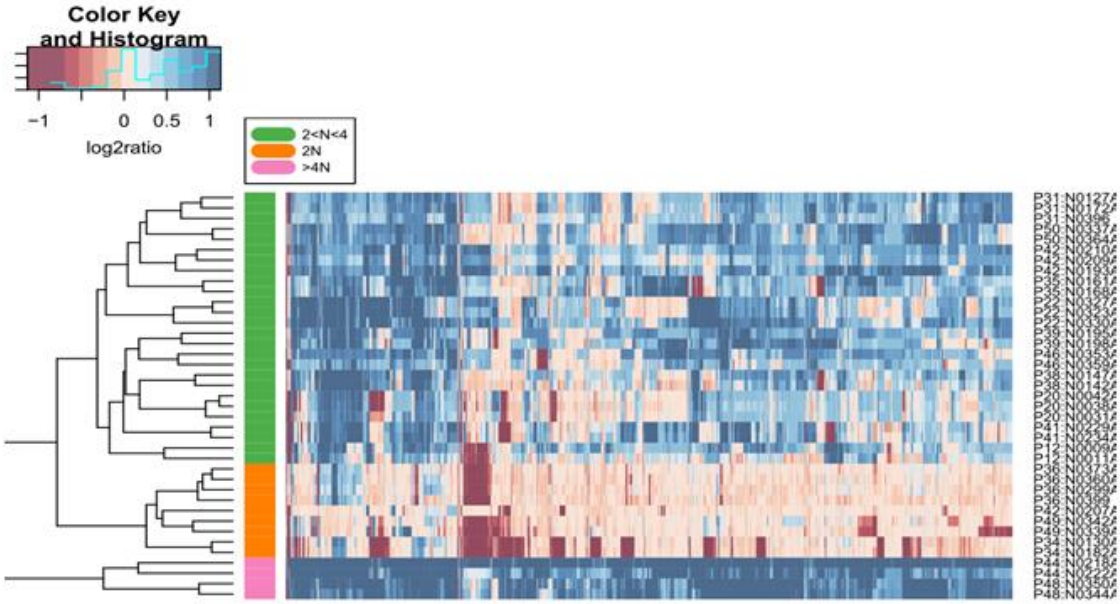


Figure S8| Hierarchical clustering of Euclidean distances based on SCNAs of all sorted tumor populations. Tumors of individual patients (P#) are clustered closer to each other than to tumors of other patients. One exception is the diploid tumor of patient 42 (P42:N0207A), which is clustered with diploid tumors of other patients. Interestingly, this clustering was able to identify three groups, which are represented by their respective ploidies.

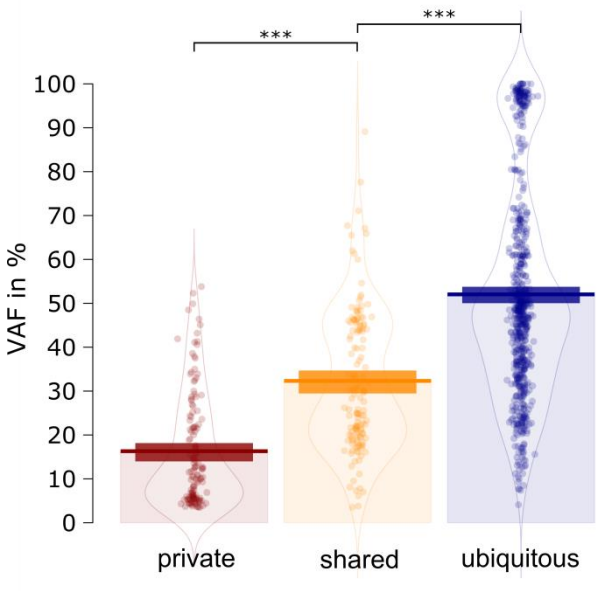


Figure S9| Violin Plots for VAFs. The VAFs of SNVs were plotted according to their presence in all (ubiquitous), more than one but not all (shared, applicable for the four patients with n_{MET} >1), or just one (private) tumor biopsy. The differences are significant (Mann-Whitney-Wilcoxon Test; ***: p-value < 0.001)

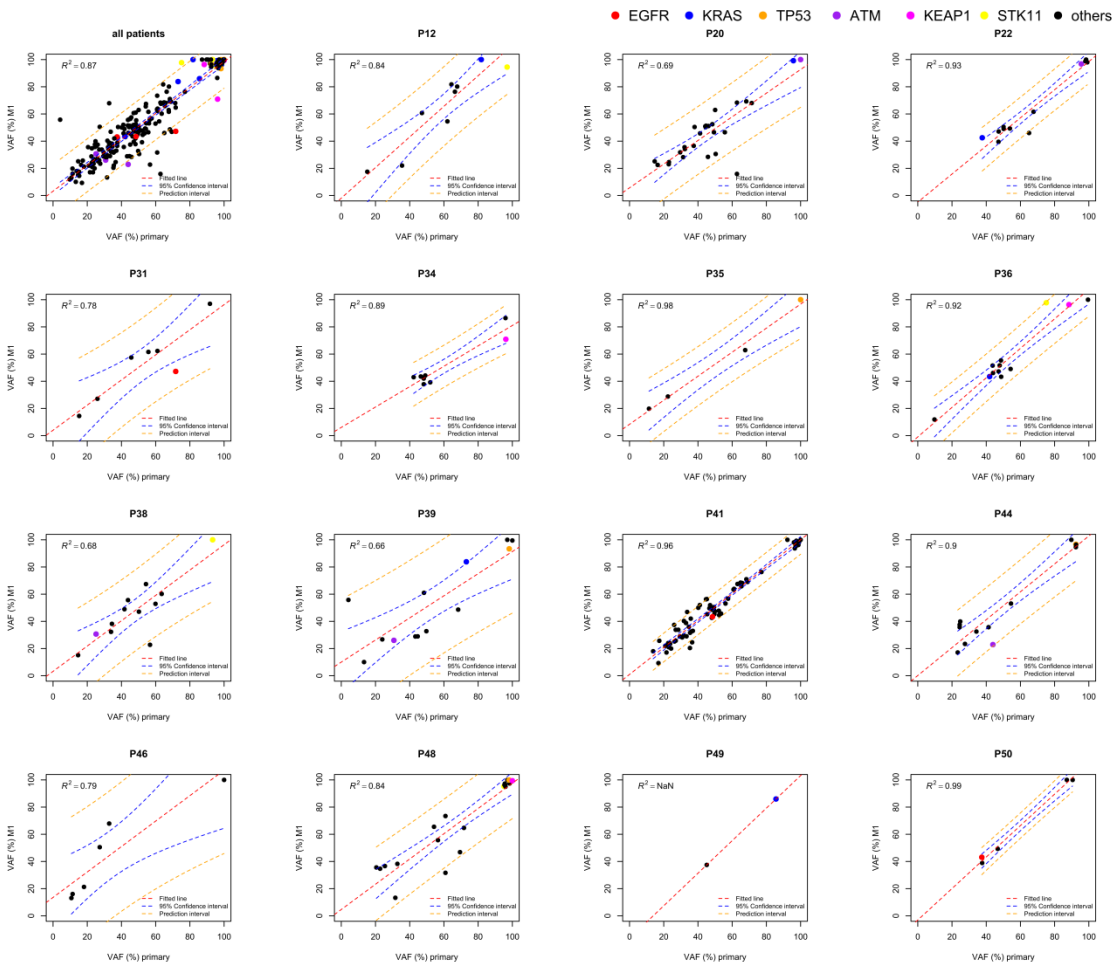


Figure S10 | Scatter plots for VAFs of ubiquitous SNVs per patient. VAFs of mutations shared between the primary tumor (x-axis) and the first biopsied metastasis (y-axis). Patient 42 is excluded, due to the presence of multiple populations in the primary tumor. The correlation among the cohort (top left plot) is linear and significant ($p < 0.001$). P, patient; M1, first biopsied metastasis

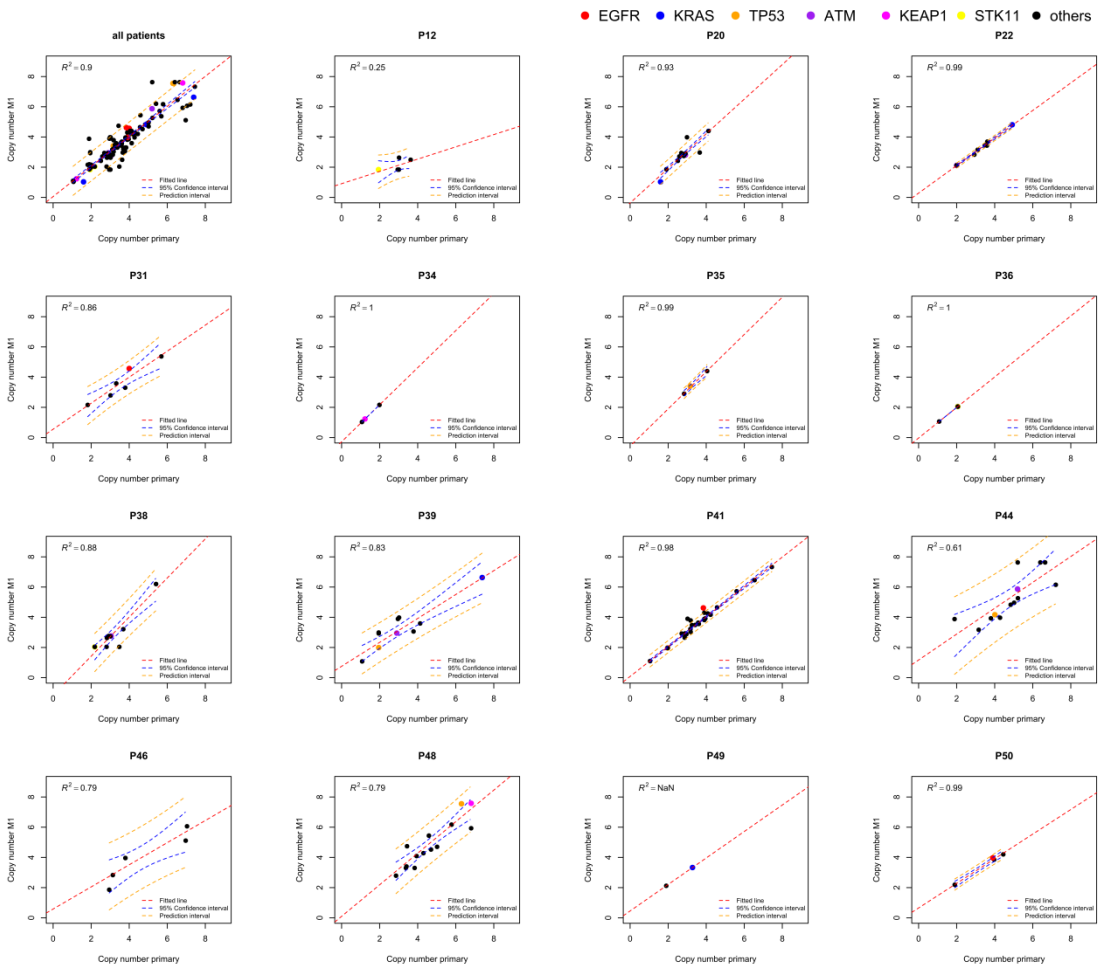


Figure S11 | Scatter plots for copy numbers of ubiquitous SNVs per patient. The copy numbers of mutations shared between the primary tumor (x-axis) and the first biopsied metastasis (y-axis). Patient 42 is excluded, due to the presence of multiple populations in the primary tumor. The correlation among the cohort (top left plot) is linear and significant ($p<0.001$). P, patient; M1, first biopsied metastasis

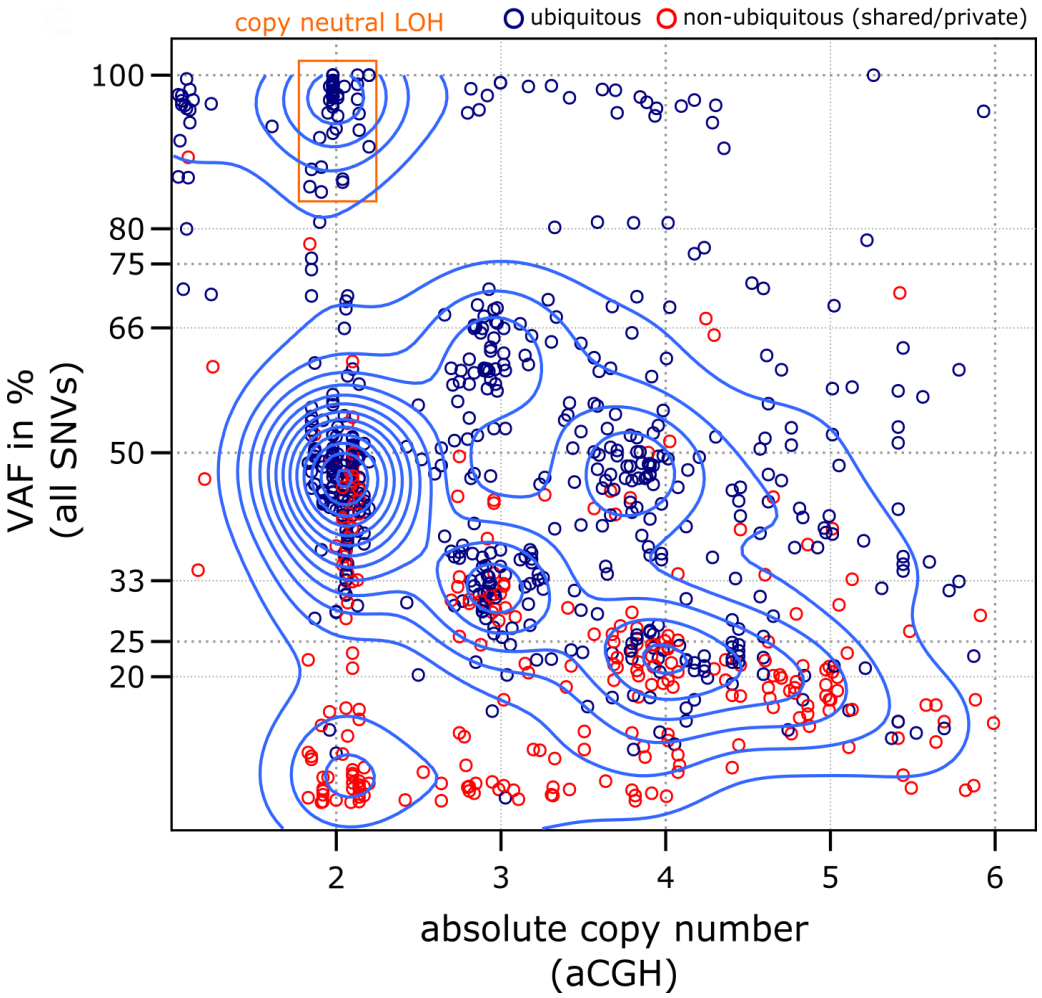


Figure S12 | Variant allele frequencies and copy numbers of all mutations detected in this cohort. Density plot of the absolute copy number (x-axis) and VAF (y-axis) of the mutations. Note that the majority of non-ubiquitous (shared or private) SNVs accumulate at lower VAFs, indicating a non-truncal and therefore subclonal origin.

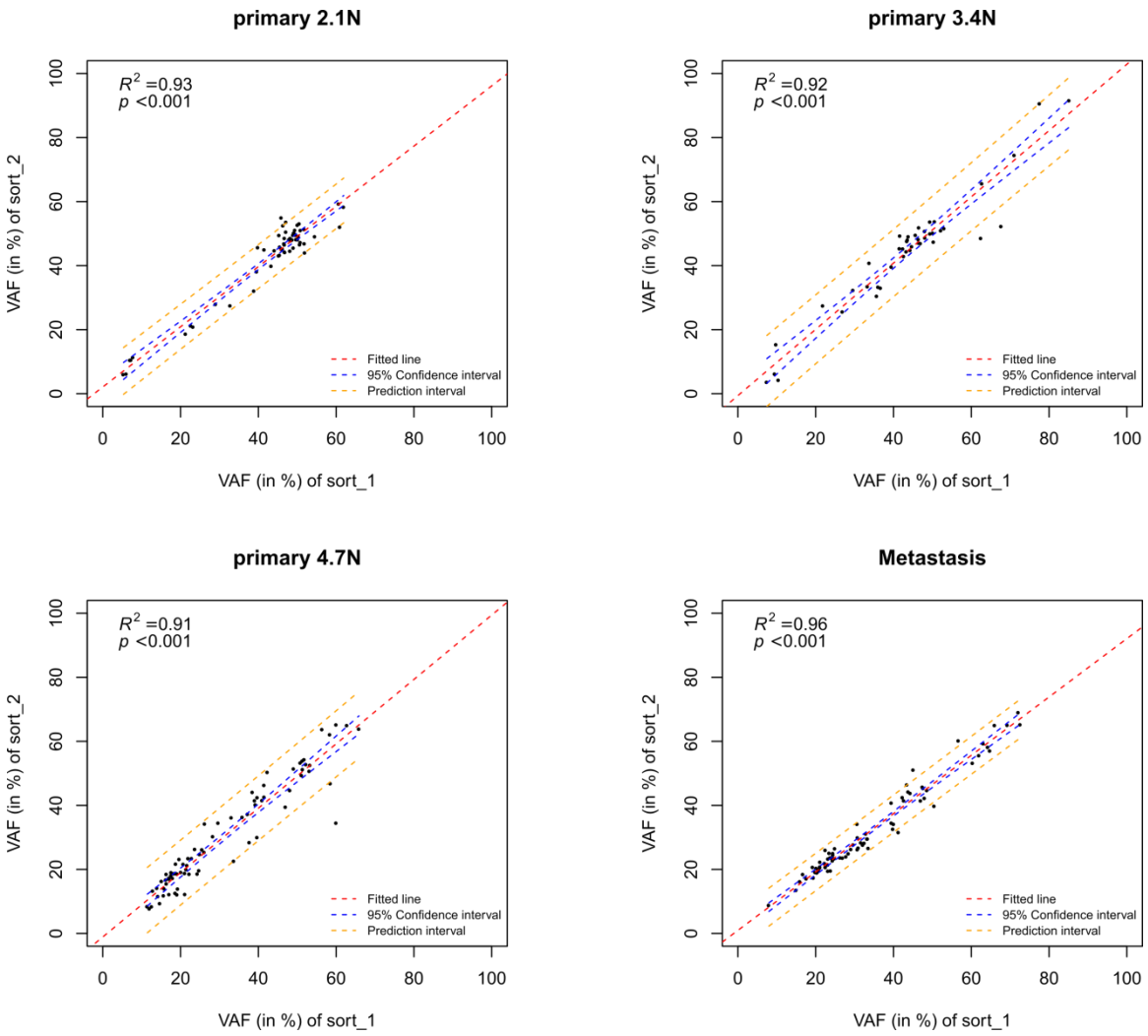


Figure S13 | Scatter plots of VAFs of SNVs from sorting replicates for patient 42. Nuclei were extracted, stained, and sorted on different days. Libraries were constructed and individually sequenced. The VAFs of SNVs from the first sort (x-axis) are plotted versus the VAFs of SNVs from the second sort.

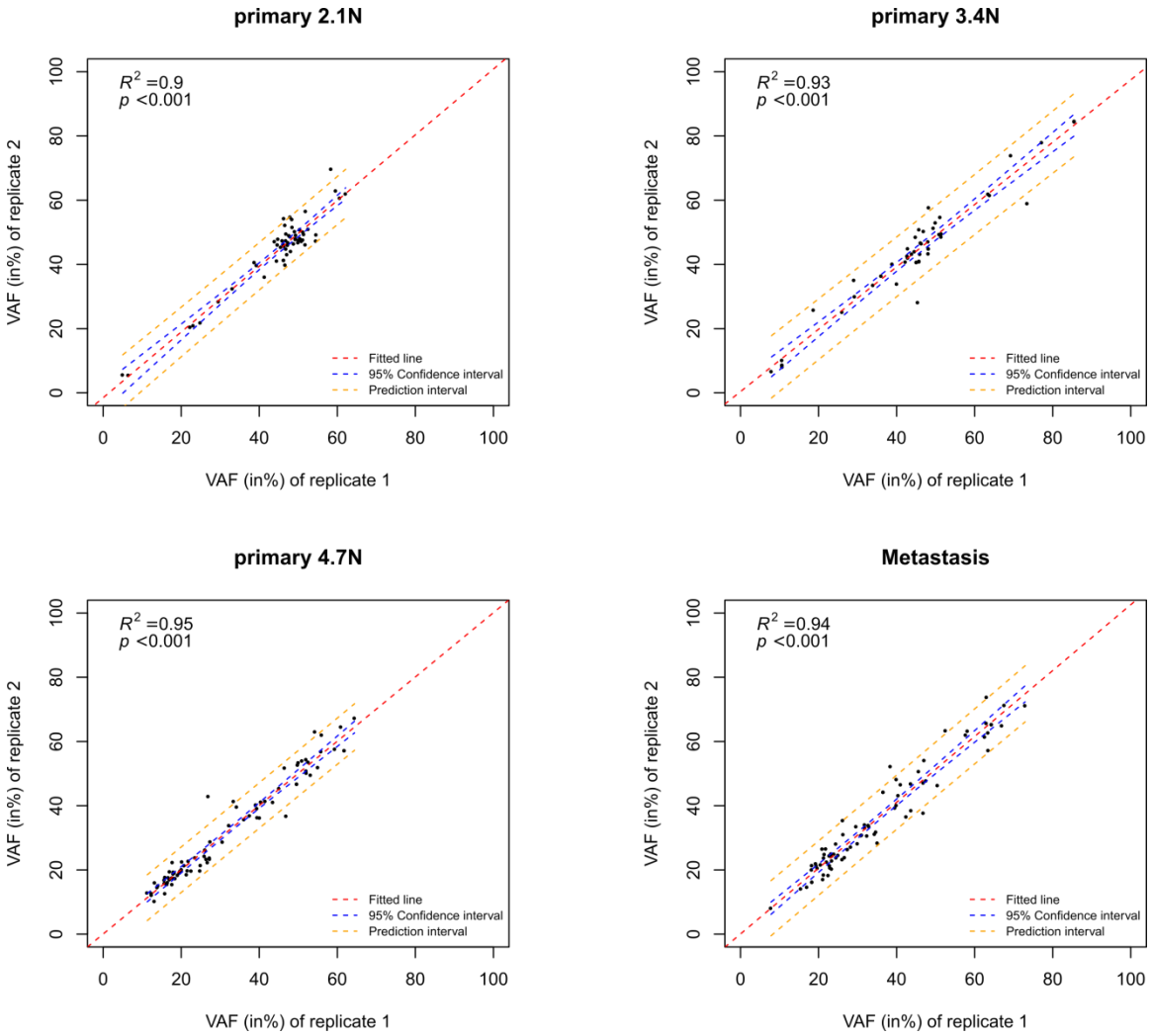


Figure S14 | Scatter plots of VAF of SNVs from <u>sequencing</u> replicates for patient 42. Libraries were constructed in duplicates and individually sequenced. The VAFs of SNVs from the first replicates (x-axis) are plotted versus the VAFs of SNVs from the second replicate.

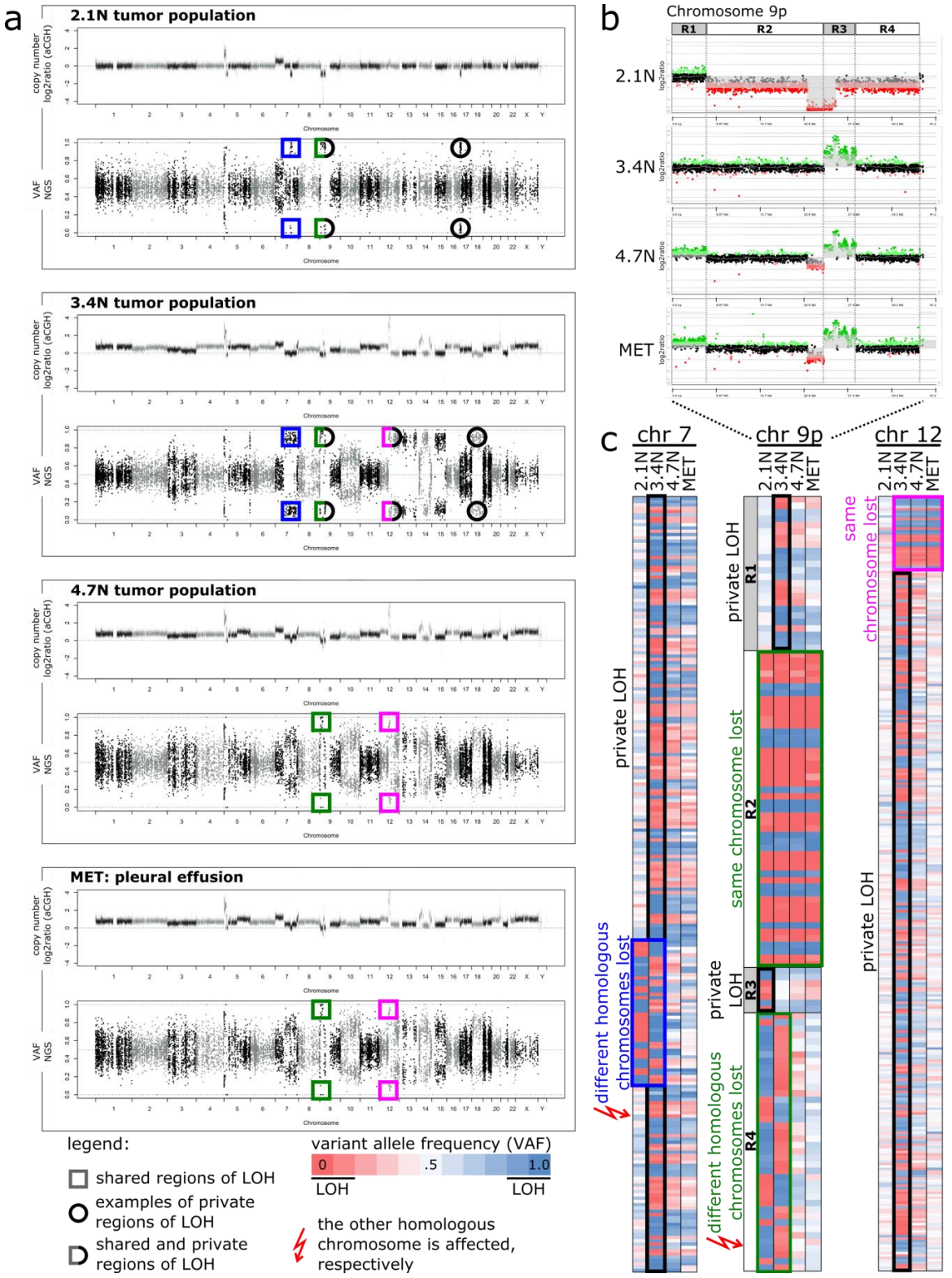


Figure S15 | Regions of LOH in the primary tumor of patient 42. LOH in the three tumor populations from the primary tumor (2.1N, 3.4N, and 4.7N) and the metastasis (a). While the LOH of regions affected in all tumor populations (greenR2 and pink) and LOH events that are shared among the 3.4N population, 4.7N population, and the metastasis affect the same chromosomes, the two LOH events that are shared between the 2.1N and 3.4N populations (blue, green region R4) affect the other homologous chromosome, respectively (c). The 4.7N population and the metastasis have the same breakpoint on chromosome 9p (in region R2) as the 2.1N population covering the *CDKN2A* locus (b), while this is absent in the 3.4N population. Many LOH events in the 3.4N population (chr13, chr14, chr15, chr18) and the LOH in the 2.1N population (chr17) are not present in the 4.7N population or in the metastasis.

Supplementary Tables

Table S1 | Overview of the 409 genes that are included in the Ion Ampliseq™ Comprehensive Cancer Panel. The CCP panel covers all exons of these genes.

The 409 genes in the Ion Ampliseq™ Comprehensive Cancer Panel								
1 AKAP9	51 CTNNB1	101 FANCC	151 RAF1	201 NUMA1	251 TNFAIP3	301 FGFR4	351 MARK4	401 KAT6A
2 BAP1	52 CDK6	102 ETV1	152 NUP214	202 MPL	252 SRC	302 IKBKE	352 NFKB1	402 MBD1
3 BMPR1A	53 CDK4	103 FGFR1	153 PIK3R1	203 NKX2-1	253 SEPT9	303 AKT3	353 RRM1	403 RNF2
4 ATM	54 CCND2	104 ERCC4	154 PDE4DIP	204 PHOX2B	254 WAS	304 CDH2	354 MAP3K7	404 KAT6B
5 BCL3	55 CDKN2B	105 DDB2	155 PTCH1	205 NPM1	255 TSC1	305 CDH5	355 MAPK8	405 CMPK1
6 BLM	56 CCND1	106 FBXW7	156 PALB2	206 NCOA4	256 TFE3	306 GUCY1A2	356 LPHN3	406 TIMP3
7 ABL2	57 HOOK3	107 ETV4	157 PPARG	207 MYCN	257 TP53	307 IRS2	357 SDHA	407 SMUG1
8 BRD3	58 FH	108 DDIT3	158 PAX7	208 PLAG1	258 TSHR	308 SF3B1	358 SGK1	408 RNASEL
9 ALK	59 GPR124	109 FAS	159 PML	209 NRAS	259 TNFRSF14	309 PRDM1	359 PIK3C2B	409 IGF2
10 BCR	60 FLT1	110 ERCC3	160 PDGFRA	210 MYC	260 TAL1	310 RALGDS	360 AURKC	
11 ARNT	61 FLT3	111 FAM123B	161 PTEN	211 SUFU	261 TET2	311 GNA11	361 BIRC5	
12 ASXL1	62 FLI1	112 EXT2	162 PSIP1	212 TAF1L	262 RARA	312 MLL2	362 MAPK1	
13 FANCJ	63 HSP90AA1	113 FANCF	163 POU5F1	213 STK11	263 VHL	313 TCF7L2	363 PIK3CG	
14 BCL11B	64 GNAS	114 MET	164 NUP98	214 SYK	264 BTK	314 ATRX	364 GRM8	
15 CARD11	65 GATA2	115 JAK2	165 PIK3CA	215 STK36	265 AR	315 MLL3	365 PIK3CB	
16 BRAF	66 HNF1A	116 LIFR	166 PAX5	216 RNF213	266 TLX1	316 CTNNA1	366 PIK3R2	
17 BUB1B	67 HSP90AB1	117 JAK3	167 PTPN11	217 RPS6KA2	267 SOX2	317 CDK12	367 ERCC1	
18 AFF1	68 HLF	118 ITGA9	168 PBRM1	218 ROS1	268 SOCS1	318 XPO1	368 WHSC1	
19 APC	69 IDH1	119 KDR	169 MYB	219 SETD2	269 PRKDC	319 ARID2	369 ETS1	
20 AKT1	70 GNAQ	120 KIT	170 NFKB2	220 SDHC	270 PTPRD	320 TRIM24	370 PLEKHG5	
21 ATF1	71 FOXO1	121 KDM6A	171 PDGFRB	221 SMARCA4	271 MAP2K2	321 DAXX	371 CRBN	
22 ARID1A	72 IDH2	122 MDM2	172 MYH11	222 RET	272 PLCG1	322 IL7R	372 BLNK	
23 BIRC3	73 FOXP1	123 MAP2K4	173 NLRP1	223 RUNX1	273 IGF1R	323 MYD88	373 BIRC2	
24 AKT2	74 FLCN	124 MLL	174 MYH9	224 SMO	274 PKHD1	324 ING4	374 PIK3CD	
25 ABL1	75 IKZF1	125 LCK	175 NFE2L2	225 RB1	275 FOXP4	325 IKBKB	375 POT1	
26 BCL9	76 HRAS	126 MAP2K1	176 PIM1	226 SMAD4	276 SMAD2	326 FN1	376 XRCC2	
27 BCL11A	77 FGFR3	127 MITF	177 MUC1	227 RECQL4	277 LRP1B	327 ICK	377 KEAP1	
28 BCL10	78 GATA3	128 MLLT10	178 NF1	228 SMARCB1	278 LTK	328 ITGB3	378 TLR4	
29 AFF3	79 IL21R	129 JAK1	179 NOTCH2	229 XPA	279 EPHB1	329 ITGB2	379 TGM7	
30 BCL2	80 FOXO3	130 MLH1	180 NF2	230 WRN	280 USP9X	330 LAMP1	380 SAMD9	
31 BCL6	81 GATA1	131 ITGA10	181 PBX1	231 TRIP11	281 IGF2R	331 LTF	381 MAGEA1	
32 CDH11	82 FOXL2	132 KRAS	182 MTOR	232 ZNF521	282 PAK3	332 HIF1A	382 CKS1B	
33 CREB1	83 DEK	133 KDM5C	183 MYCL1	233 TSC2	283 TGFBR2	333 HCAR1	383 PTGS2	
34 CASC5	84 ERBB4	134 MEN1	184 NTRK3	234 XPC	284 CCNE1	334 BAI3	384 PTPRT	
35 CREBBP	85 EML4	135 MALT1	185 NBN	235 ZNF384	285 BCL2L2	335 ATR	385 SOX11	
36 CHEK1	86 FANCA	136 LPP	186 NSD1	236 TCF12	286 FLT4	336 AXL	386 SH2D1A	
37 COL1A1	87 EZH2	137 MDM4	187 PMS1	237 RUNX1T1	287 CDH20	337 ACVR2A	387 PGAP3	
38 CDH1	88 ERCC2	138 IL6ST	188 PPP2R1A	238 REL	288 EPHA7	338 ADAMTS20	388 G6PD	
39 CSF1R	89 FANCD2	139 IL2	189 PDGFB	239 TCL1A	289 ERBB3	339 DST	389 DPYD	
40 CEBPA	90 EXT1	140 JUN	190 MRE11A	240 TCF3	290 BCL2L1	340 CSMD3	390 TCF7L1	
41 CBL	91 EGFR	141 IRF4	191 MSH6	241 SSX1	291 EPHB6	341 EP400	391 FZR1	
42 CYLD	92 FGFR2	142 MAML2	192 NCOA1	242 WT1	292 CDK8	342 CYP2C19	392 GDNF	
43 CHEK2	93 FANCG	143 MAF	193 NTRK1	243 RHOH	293 EPHA3	343 DCC	393 SYNE1	
44 CDKN2A	94 DNMT3A	144 KLF6	194 MN1	244 SDHB	294 EPHB4	344 CYP2D6	394 UBR5	
45 CIC	95 ERCC5	145 MAFB	195 PRKAR1A	245 TPR	295 AURKA	345 MMP2	395 TRRAP	
46 CRTC1	96 DICER1	146 PAX3	196 MUTYH	246 SDHD	296 MCL1	346 NOTCH4	396 TAF1	
47 CDC73	97 ERBB2	147 PMS2	197 NOTCH1	247 TRIM33	297 TBX22	347 MTR	397 TNK2	
48 CD79A	98 ERG	148 PAX8	198 MSH2	248 TOP1	298 ESR1	348 PARP1	398 THBS1	
49 CDKN2C	99 DDR2	149 PER1	199 NIN	249 SBDS	299 CRKL	349 MARK1	399 UGT1A1	
50 CD79B	100 EP300	150 RAD50	200 NCOA2	250 TET1	300 AURKB	350 MTRR	400 MAGI1	

Table S2 | Overview of the 28 genes that are included in the Ion Ampliseq™ Custom Melanoma Panel. This panel covers all exons of these genes.

The 28 genes in the Ion Ampliseq™ Custom Melanoma Panel			
1 ARID2	8 FGFR2	15 PIK3CA	22 SNX31
2 BCL2L12	9 GRIN2A	16 PPP6C	23 SOX10
3 BRAF	10 GRM3	17 PREX2	24 STK19
4 CDK4	11 KIT	18 PTEN	25 TACC1
5 CDKN2A	12 MAP2K1	19 PTPRD	26 TERT
6 CTNNB1	13 MITF	20 RAC1	27 TP53
7 ERBB4	14 NRAS	21 ROS1	28 TRRAP

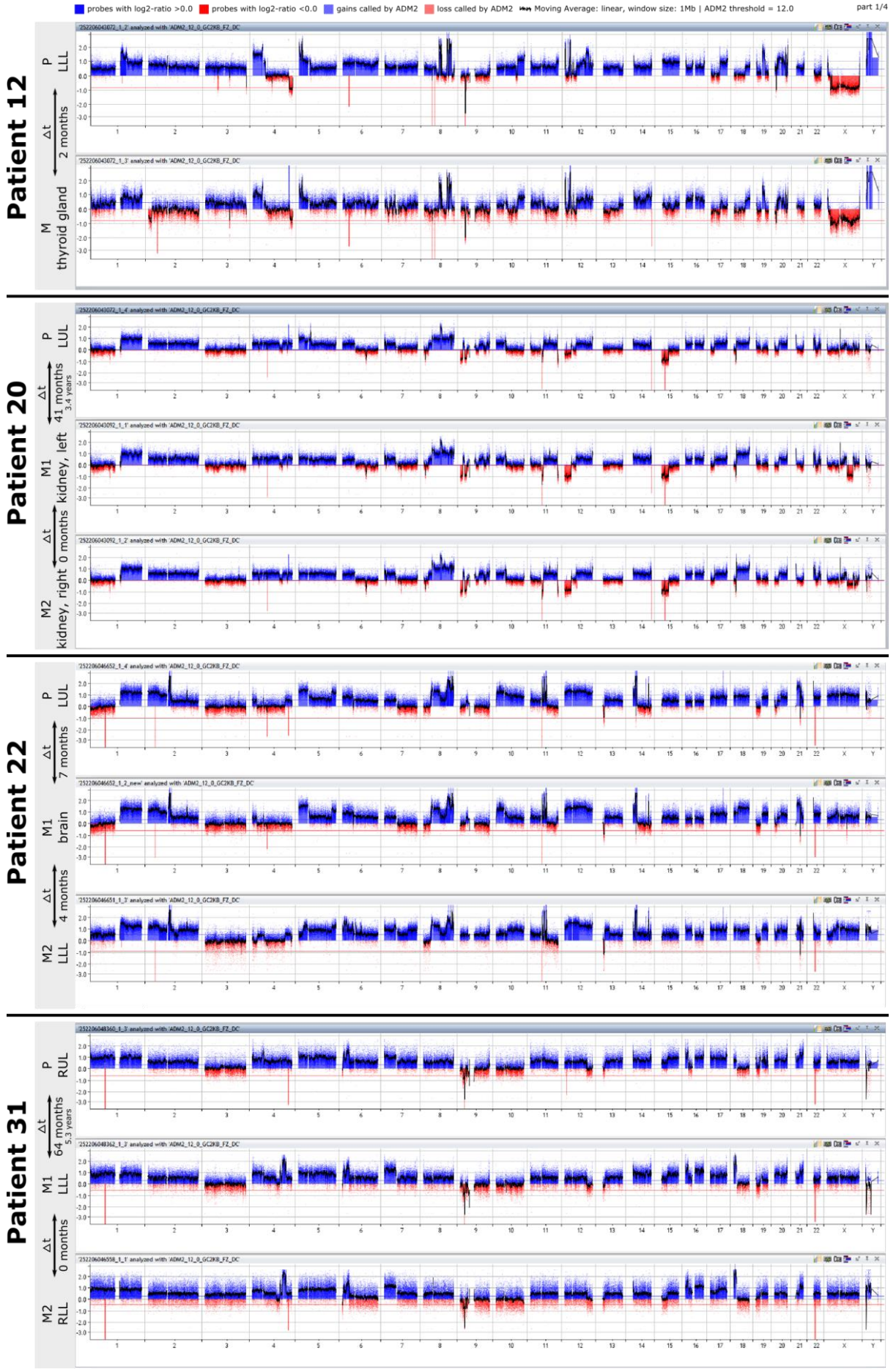
Appendix I – Array-CGH summary Part B

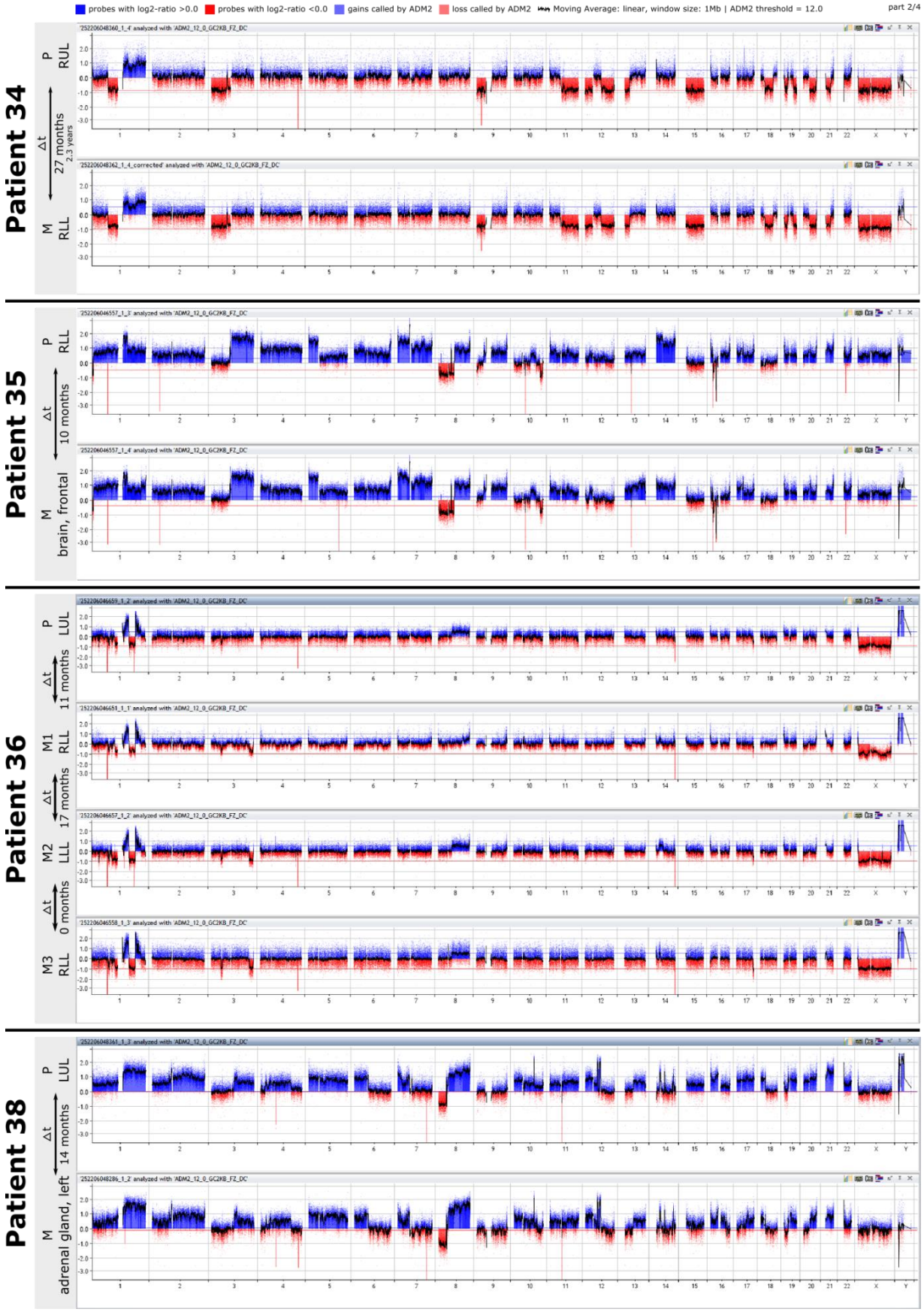
Overview of genome-wide copy number aberrations of all 39 sorted tumors population from 37 tumor samples of 16 patients with matched primary-metastatic LUAD. Shown are chromosome 1-22 and the sex chromosomes X and Y.

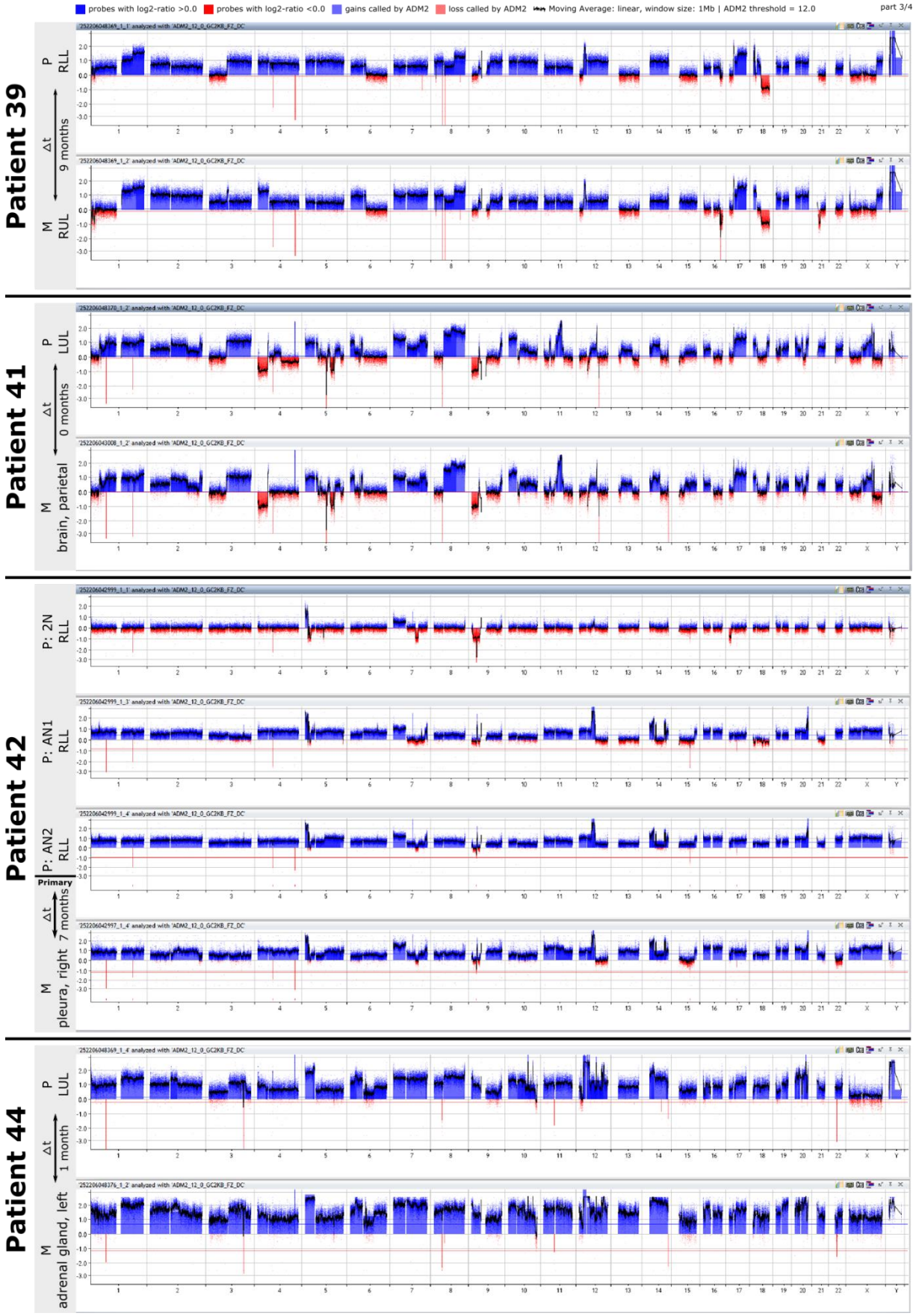
Blue: gain, red: deletion. Abbreviations: P: primary tumor, M: Metastasis, Δt: time between the occurrence of two tumors

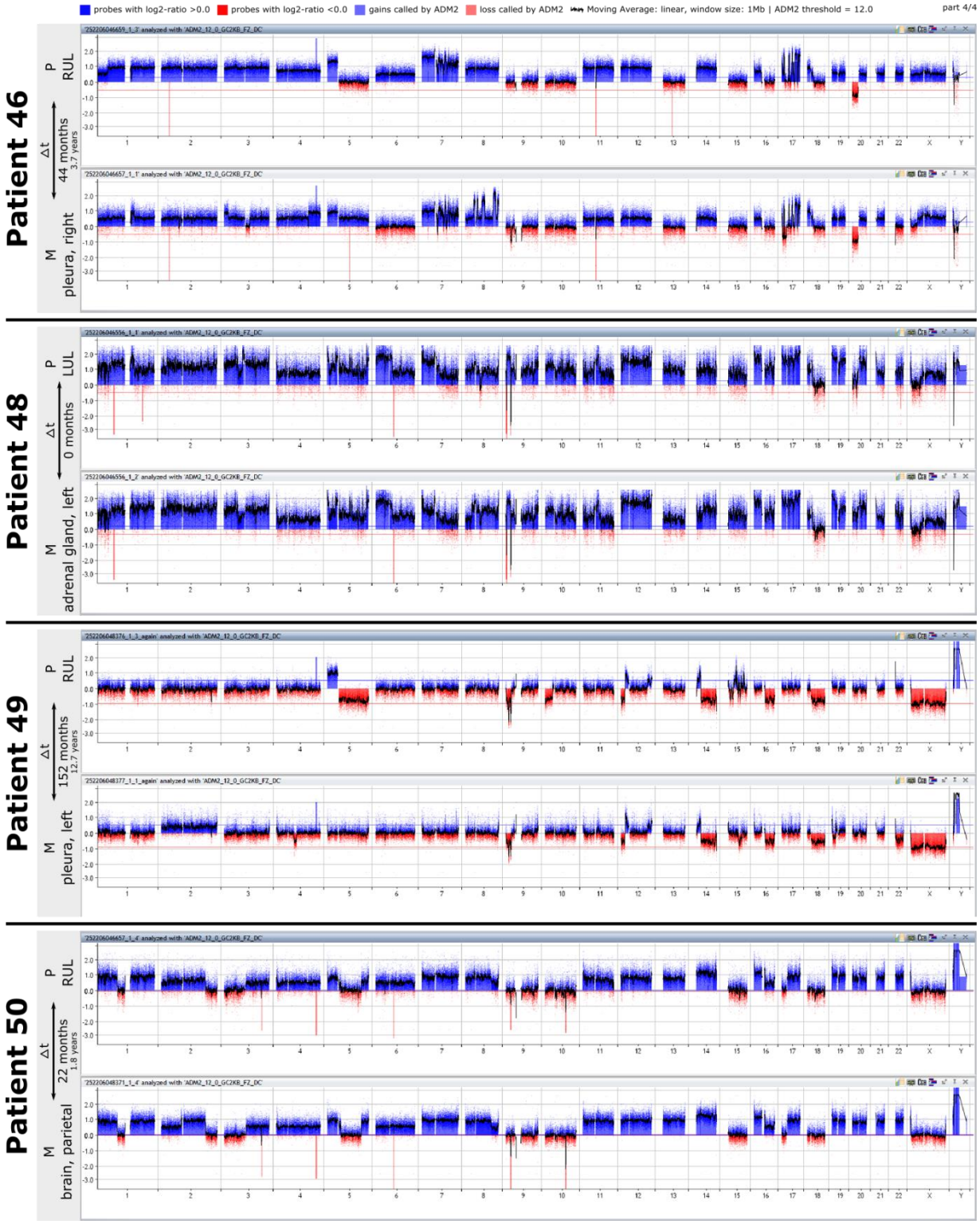
Note: Appendix I is divided into four parts

- Patients 12, 20, 22, 31 | part 1
- Patients 34, 35, 36, 38 | part 2
- Patients 39, 41, 42, 44 | part 3
- Patients 46, 48, 49, 50 | part 4


probes with log2-ratio >0.0
probes with log2-ratio <0.0
gains called by ADM2
loss called by ADM2
Moving Average: linear, window size: 1Mb | ADM2 threshold = 12.0
part 1/4
Patient 12
P LLL
Δt 2 months
M thyroid gland
Patient 20
P LUL
Δt 41 months 3.4 years
M1 kidney, left
Δt 0 months
M2 kidney, right
Patient 22
P LUL
Δt 7 months
M1 brain
Δt 4 months
M2 LLL
Patient 31
P RUL
Δt 64 months 5.3 years
M1 LLL
Δt 0 months
M2 RLL


probes with log2-ratio >0.0
probes with log2-ratio <0.0
gains called by ADM2
loss called by ADM2
Moving Average: linear, window size: 1Mb | ADM2 threshold = 12.0
part 2/4
Patient 34
P RUL
Δt 27 months 2.3 years
M RLL
Patient 35
P RLL
Δt 10 months
M brain, frontal
Patient 36
P LUL
Δt 11 months
M1 RLL
Δt 17 months
M2 LLL
Δt 0 months
M3 RLL
Patient 38
P LUL
Δt 14 months
M adrenal gland, left


probes with log2-ratio >0.0
probes with log2-ratio <0.0
gains called by ADM2
loss called by ADM2
Moving Average: linear, window size: 1Mb | ADM2 threshold = 12.0
part 3/4
Patient 39
P
RLL
Δt
9 months
M
RUL
Patient 41
P
LUL
Δt
0 months
M
brain, parietal
Patient 42
P: 2N
RLL
P: AN1
RLL
P: AN2
RLL
Primary
Δt
7 months
M
pleura, right
Patient 44
P
LUL
Δt
1 month
M
adrenal gland, left


probes with log2-ratio >0.0
probes with log2-ratio <0.0
gains called by ADM2
loss called by ADM2
Moving Average: linear, window size: 1Mb | ADM2 threshold = 12.0
part 4/4
Patient 46
P RUL
Δt 44 months 3.7 years
M pleura, right
Patient 48
P LUL
Δt 0 months
M adrenal gland, left
Patient 49
P RUL
Δt 152 months 12.7 years
M pleura, left
Patient 50
P RUL
Δt 22 months 1.8 years
M brain, parietal