EXPLORING THE BIOLOGY AND CLINICAL UTILITY OF CIRCULATING TUMOUR DNA

A data-driven investigation of the mechanisms and associates of ctDNA release in lung adenocarcinoma, colorectal cancer and metastatic melanoma cohorts

judit kisistók

PhD Dissertation

Department of Molecular Medicine (MOMA), Aarhus University Hospital
Department of Health, Aarhus University
Bioinformatics Research Centre, Aarhus University

May 2023

Judit Kisistók: *Exploring the biology and clinical utility of circulating tumour DNA,* A data-driven investigation of the mechanisms and associates of ctDNA release in lung adenocarcinoma, colorectal cancer and metastatic melanoma cohorts, © May 2023

AUTHOR
Judit Kisistók
Master of Science in Bioinformatics
Department of Molecular Medicine, Aarhus University Hospital
Department of Health, Aarhus University
Bioinformatics Research Centre, Aarhus University

MEMBERS OF THE ASSESSMENT COMMITTEE
Richard T. Bryan
Professor
Institute of Cancer and Genomic Sciences
Robert Aitken Institute for Clinical Research
College of Medical and Dental Sciences
University of Birmingham, Birmingham, United Kingdom

Mads Thomassen
Professor
Klinisk Institut, Forskningsenhed for Human Genetik, Odense Universitetshospital

Joanna Maria Kalucka
Associate Professor
Department of Biomedicine, Aarhus University

MAIN SUPERVISOR
Nicolai Juul Birkbak
Associate professor
Department of Molecular Medicine, Aarhus University Hospital

CO-SUPERVISORS
Christian Storm Pedersen
Associate Professor
Bioinformatics Research Centre, Aarhus University

Lars Dyrskjøt Andersen
Professor
Department of Molecular Medicine, Aarhus University

Charles Swanton
Professor
The Francis Crick Institute, United Kingdom

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede.
Figures used in the thesis were created in Biorender by Judit Kisistók unless referenced otherwise.

CONTENTS

LIST OF FIGURES

ACRONYMS

ctDNA	circulating tumor DNA
NSCLC	non-small-cell lung cancer
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
COAD	colon adenocarcinoma
CRC	colorectal cancer
MAF	mutant allele frequency
VAF	variant allele frequency
LOD	limit of detection
ITH	intratumor heterogeneity
CIN	chromosomal instability
MSI	microsatellite instability
MMR	mismatch repair
CMS	Consensus Molecular Subtypes
MSigDB	Molecular Signatures Database
EV	extracellular vesicles
CTC	circulating tumor cells
cfDNA	cell-free DNA
PCR	Polymerase Chain Reaction

PCM	Personalized Cancer Monitoring
TRACERx	TRAcking Cancer Evolution through therapy (Rx)
MRD	minimal residual disease
dPCR	digital PCR
ddPCR	Digital-droplet PCR
NGS	next-generation sequencing
SNP	Single-Nucleotide Polymorphisms
WGS	whole genome sequencing
CNA	copy number alteration
SVM	support vector machine
CNV	copy number variants
DAO	Deep alternate observations
UMI	unique molecular identifiers
CCF	cancer cell fraction
CHIP	clonal hematopoiesis of indeterminate potential
wGII	weighted genome integrity index
FLOH	fraction of loss of heterozygosity
WES	whole exome sequencing
GSVA	Gene Set Variation Analysis
TCGA	The Cancer Genome Atlas
WHO	World Health Organization
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
AMP	Anchored Multiplex

PREFACE

This thesis represents my own scientific research efforts and is submitted as part of the Ph.D. degree requirements established by the Graduate School of Health at Aarhus University.

To the best of my knowledge, the materials contained in this submission have not previously been published or written by another person, unless otherwise acknowledged or referenced.

The work has been conducted at the Department of Molecular Medicine (MoMA), Aarhus University Hospital, Denmark, from March 2020 to May 2023; in conjunction with the Department of Clinical Medicine and the Bioinformatics Research Centre, Aarhus University, Denmark.

Judit Kisistok
May 2023

ACKNOWLEDGEMENTS

This work has truly been a team effort, and I am incredibly thankful to everyone who shared this journey with me. I would like to extend my heartfelt gratitude to:

My main supervisor, **Nicolai Juul Birkbak** - you have been the best mentor I could've ever hoped for. Thank you for taking a chance on me and inspiring me to do my best work, even in the toughest of times.

My bioinformatics group mates, **Ditte**, **Johanne**, **Mateo**, **Laura**, and **Asbjørn**, for all the laughs, the scientific discussions, and the very-non-scientific banter. You all made my time at MOMA very memorable - and speaking of memorable, special thanks to Ditte, Johanne, and Mateo for an unforgettable adventure in New Orleans.

The **researchers and staff at the Bioinformatics Research Centre**, for creating a friendly atmosphere and sharing your wealth of knowledge freely.

All my collaborators at **MOMA** and **The Francis Crick Institute** for all your help, feedback, and incredibly high-quality scientific work that inspired me to set new standards for myself.

The team at **Omiics**, especially **Susanne** and **Marni**, for taking me on as an intern and supporting me in my professional development. I'm grateful to have spent my external stay with you.

The members of the assessment committee, **Professor Richard T. Bryan**, **Professor Mads Thomassen**, and **Associate Professor Joanna Maria Kalucka**, for agreeing to review my thesis and participating in my defence.

All of my friends in Denmark and Hungary, especially **Tine**, **Noor**, **Andi**, **Dóri**, and **Noémi** - for the plethora of cherished memories that are way too many to list, and for wanting to hang out with me even when I was a stressed-out mess.

My **Lindy Hop friends**, for all the amazingly fun times we spent together on (and away from) the dancefloor.

Mom, for all your unconditional love, tremendous support, and unwavering belief in me. I wouldn't be anywhere near to where I am today if it hadn't been for you.

Dad, for passing down your stubborn grit to me. I wish you could've been here to see this.

Finally, my boyfriend, **Mathias**, for your relentless love and support. You've always seen the best in me, even at times when I couldn't see it in myself.

MANUSCRIPTS

The following three manuscripts form the core of the thesis:

Manuscript I (Published 2023)
Christopher Abbosh, Alexander M. Frankell, Thomas Harrison, Judit Kisistók, et. al. *Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA.* Nature 616, 553–562 (2023).
doi: 10.1038/s41586-023-05776-4

Manuscript II (Manuscript in preparation, 2023)
Judit Kisistók, Laura Andersen, Tenna Vesterman Henriksen, Jesper Bertram Bramsen, Thomas Reinert, Nadia Øgaard, Trine Block Mattesen, Nicolai Juul Birkbak, Claus Lindbjerg Andersen. *Exploring the biology of ctDNA release in colon cancer.*

Manuscript III (Manuscript accepted in Melanoma Research, 2023)
Judit Kisistók, Ditte Sigaard Christensen, Mads Heilskov Rasmussen, Lone Duval, Ninna Aggerholm-Pedersen, Adam Andrzej Luczak, Boe Sandahl Sorensen, Martin Roelsgaard Jakobsen, Trine Heide Oellegaard, Nicolai Juul Birkbak. *Analysis of circulating tumor DNA during checkpoint-inhibition in metastatic melanoma using a tumor-agnostic panel.*

The following 2 manuscripts are not included in the following dissertation but are products of my Ph.D. study:

Johanne Ahrenfeldt, Ditte Sigaard Christensen, Mateo Sokač, Judit Kisistók, Nicholas McGranahan, Nicolai Juul Birkbak. *Computational Analysis Reveals the Temporal Acquisition of Pathway Alterations during the Evolution of Cancer.* Cancers (Basel). 2022 Nov 25;14(23):5817. doi: 10.3390/cancers14235817. PMID: 36497297; PMCID: PMC9739002.

Ditte Sigaard Christensen, Johanne Ahrenfeldt, Mateo Sokač, Judit Kisistók, Martin K. Thomsen, Lasse Maretty, Nicholas McGranahan, Nicolai Juul Birkbak. *Treatment Represents a Key Driver of Metastatic Cancer Evolution.* Cancer Res. 2022 Aug 16;82(16):2918-2927.
doi: 10.1158/0008-5472.CAN-22-0562. PMID: 35731928.

SUMMARY

Diagnosing, profiling, and monitoring cancer has traditionally been based on the analysis of surgically resected tumor samples. In recent years, liquid biopsies have garnered interest as a minimally invasive supplement or alternative to this medical practice. Investigating the blood, specifically, the tumor-originated DNA fragments in the plasma, termed circulating tumor DNA (ctDNA), enables researchers and oncologists to gain real-time insight into the qualitative composition and the quantity of the malignancy.

Despite considerable research efforts in the field, the exact, cancer type-, histology-, and patient-specific biology behind ctDNA release remains poorly understood, posing a limitation to clinical adoption. The primary aim of this Ph.D. study was, therefore, to elucidate the process of ctDNA shedding by analyzing the biological contributors and clinical associates of ctDNA release across various cancer types and study designs.

In **Manuscript I**, I aimed to investigate the biological drivers of ctDNA release by analyzing multi-region transcriptomic and genomic data collected from a cohort of lung adenocarcinoma patients. I found that ctDNA positive patients carried a distinct, differential phenotype compared to ctDNA negatives, characterized by high proliferation, and associated with aggressive disease.

In **Manuscript II**, I analyzed genomic, transcriptomic, and clinical data to uncover the biology of ctDNA release within the context of a colorectal cancer cohort. I found that ctDNA release in colorectal cancer appears to be driven by a multitude of contributors, most notably, tumor size and proliferative capacity.

In **Manuscript III**, I investigated how genomic alterations may contribute to response to immunotherapy by analyzing longitudinal ctDNA data from a cohort of metastatic melanoma patients. I have found that patients with subpar response to therapy harbored a higher percentage of TERT mutations, indicating that this gene could potentially be used as a marker in the clinic.

DANSK RESUME (DANISH SUMMARY)

Diagnosticering, profilering og overvågning af cancer har traditionelt været baseret på analyse af kirurgisk resekterede tumorprøver. I de senere år har flydende biopsier høstet interesse som et minimalt invasivt supplement eller alternativ til denne medicinske praksis. Undersøgelse af blodet, specifikt de DNA-fragmenter i plasmaet med tumor oprindelse, kaldet cirkulerende tumor-DNA (ctDNA), gør det muligt for forskere og onkologer at få real-time indsigt i den kvalitative sammensætning af tumoren, samt en indikation på størrelsen af tumoren.

På trods af betydelig forskningsindsats på området er den nøjagtige, cancertype-, histologi- og patientspecifikke biologi bag ctDNA-frigivelse stadig dårligt forstået, hvilket udgør en begrænsning for klinisk anvendelse. Det primære formål med dette ph.d.-studie var derfor at belyse processen med ctDNA-udskillelse ved at analysere de biologiske komponenter og kliniske tilstande associeret til ctDNA på tværs af forskellige cancertyper og undersøgelsesdesign.

I **Manuskript I** havde jeg til formål at undersøge de biologiske drivkræfter for ctDNA-frigivelse ved at analysere multiregionale transkriptomiske og genomiske data indsamlet fra en kohorte af lungeadenokarcinom patienter. Jeg fandt ud af, at ctDNA-positive patienter bar en distinkt, differentiel fænotype sammenlignet med ctDNA-negative, karakteriseret ved hastig vækst og forbundet med aggressiv sygdom.

I **Manuskript II** analyserede jeg genomiske, transkriptomiske og kliniske data for at afdække biologien bag ctDNA-frigivelse i en kohorte af kolorektal cancerpatienter. Jeg fandt, at ctDNA-frigivelse i tyktarmskræft ser ud til at være drevet af en lang række faktorer, heriblandt især tumorstørrelse og proliferativ kapacitet.

I **Manuskript III** undersøgte jeg, hvordan genomiske ændringer kan bidrage til respons på immunterapi ved at analysere longitudinelt ctDNA-data fra en kohorte af metastatiske melanompatienter. Jeg fandt, at patienter med subpar respons på terapi havde en højere procentdel af TERT-mutationer, hvilket indikerer, at dette gen potentielt kunne bruges som en klinisk biomarkør.

Part I

INTRODUCTION

1

CANCER BIOLOGY

1.1 CANCER EPIDEMIOLOGY

Cancer is a diverse and devastating disease carrying high social, clinical, and economic burdens [1]. According to the World Health Organization (WHO), it is among the leading causes of death worldwide [2], accounting for an estimated 10 million deaths in 2020 [3]. Cancer incidence and mortality, while high already, are expected to rise by 47% to reach 28.4 million cases in 2040 [3, 4], putting further pressure on affected families and healthcare systems.

1.2 THE DEVELOPMENT AND BIOLOGY OF CANCER

Cancer is a genetic disease that may originate from almost any cell type, giving rise to a remarkably varied set of malignancies. While somatic mutations occur naturally, genetic and environmental risk factors contribute to their accelerated accumulation in the genome, allowing the cell to acquire cancer-like properties. Genetic or endogenous factors include DNA repair abnormalities and cell cycle dysregulation, whereas environmental or exogenous factors include exposure to chemical carcinogens, tobacco smoke, and radiation [5, 6]. Deciphering the relative importance of these factors and their complex interplay is a daunting task, and today, carcinogenesis is known to be a complicated and layered process [5].

In an attempt to conceptualize tumorigenesis across the vast variety of cancer phenotypes and genotypes, Hanahan and Weinberg introduced the Hallmarks of Cancer in 2000. In their work, they summarized the commonalities that cancer cells share as they evolve from normalcy to uncontrolled growth states and continuously attempt to survive and expand [7]. Evading apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential, and tissue invasion and metastasis were named in the original publication. 11 years later, the authors expanded on their previous work and added two emerging hallmark capabilities to the list (deregulating cellular energetics; and avoiding immune destruction), alongside two enabling characteristics (genome instability and mutation; and tumor-promoting inflammation) [8]. In 2022, two additional hallmark capabilities (unlocking phenotypic plasticity; and senescent cells) and two further enabling characteristics (nonmutational epigenetic reprogramming; and polymorphic microbiomes)

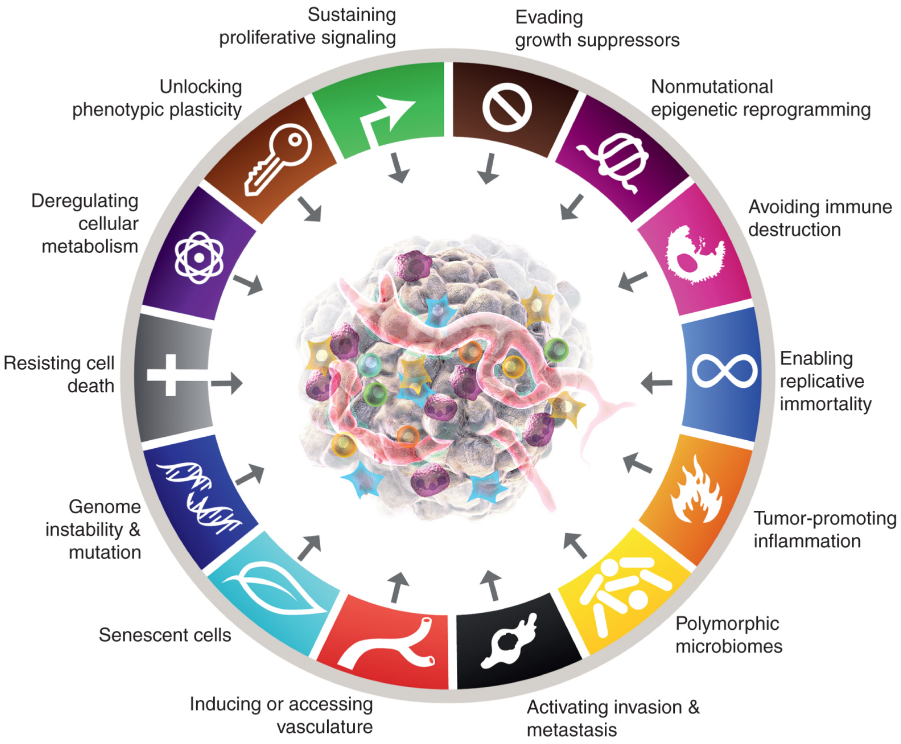

Sustaining proliferative signaling
Evading growth suppressors
Unlocking phenotypic plasticity
Nonmutational epigenetic reprogramming
Deregulating cellular metabolism
Avoiding immune destruction
Resisting cell death
Enabling replicative immortality
Genome instability & mutation
Tumor-promoting inflammation
Senescent cells
Polymorphic microbiomes
Inducing or accessing vasculature
Activating invasion & metastasis

Figure 1: The Hallmarks of Cancer, depicting the most recent version of capabilities and enabling characteristics [9]. Reprinted from Cancer Discovery, 2022, 12 (1): 31–46., Douglas Hanahan; Hallmarks of Cancer: New Dimensions, with permission from AACR.

were added [9] (Figure 1).

Specific mutations playing a role in carcinogenesis are referred to as driver mutations, and the genes in which mutations contribute to the development of cancer are referred to as driver genes. Driver genes can be proto-oncogenes, requiring a gain of function mutation to aid tumorigenesis; or tumor suppressors, requiring loss of function on both alleles to aid cancer growth [10] (Figure 2). Ever since the advent of next-generation sequencing and bioinformatics, considerable research interest has been poured into uncovering a complete list of cancer driver genes, including efforts carried out by the Cancer Gene Census [11], The Cancer Genome Atlas (TCGA) project [12], and the International Cancer Genome Consortium [13]. Famously, aberrations in the tumor suppressor gene TP53 are largely omnipresent across cancer types [14]. Cancer driver genes and mutations, however, are often tissue- and cancer-type-specific, further complicating the characterization, detection, and treatment of the disease. For instance, EGFR activation is a well-known oncogenic driver of non-small-cell lung cancer (NSCLC) [15], APC and KRAS play a role in colorectal cancer (CRC) development [16], and BRAF is a significant contributor to melanoma pathogenesis [17].

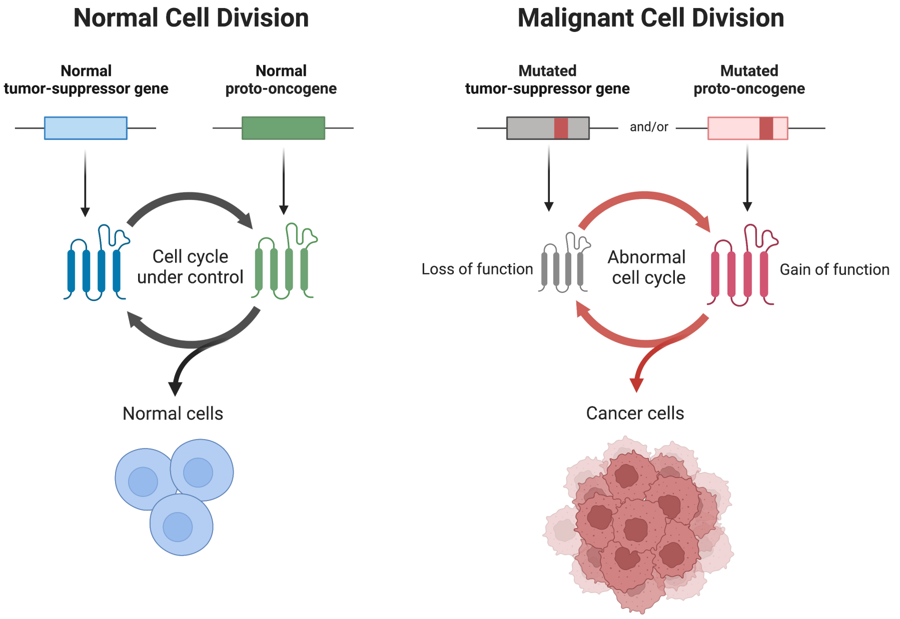


Figure 2: Comparison of normal cell division and malignant cell division. Created with BioRender.com.

1.3 CANCER EVOLUTION

In 1958, Huxley discovered that tumors show genetic diversity[18]. This phenomenon can be attributed to the complex process of cancer evolution, in which cancer cells compete, evolve, and adapt as a response to being exposed to selection pressures, such as therapy or environmental factors [19]. In an attempt to stay viable, driver mutations are selected to increase the reproductive potential of the tumor, for example, by reducing cell death, increasing cell division, or escaping growth suppressors [20]. This competition, often in conjunction with the increased number of mutations, chromosome aberrations, and genome doublings caused by genomic instability, leads to intratumor heterogeneity (ITH). ITH refers to the state where different tumor regions display different mutation profiles depending on the specific conditions they had to adapt to [21–23]. It is as much of an evolutionary puzzle as it is a clinically relevant occurrence. On one hand, the alterations carried by a tumor offer insight into its evolutionary history. In particular, mutations shared by all cancer cells are considered clonal, and alterations that reside only in a subset of the cells are considered subclonal. With this terminology, the common trunk of the tumor's evolutionary tree is composed of the clonal mutations and the branches represent the subclonal variation [24] (Figure 3). On the other hand, ITH poses a significant challenge within the field of precision medicine. Many targeted therapies rely on the presence or absence of certain driver alterations regardless of clonal architecture, potentially leading to the emergence of resistant subclones

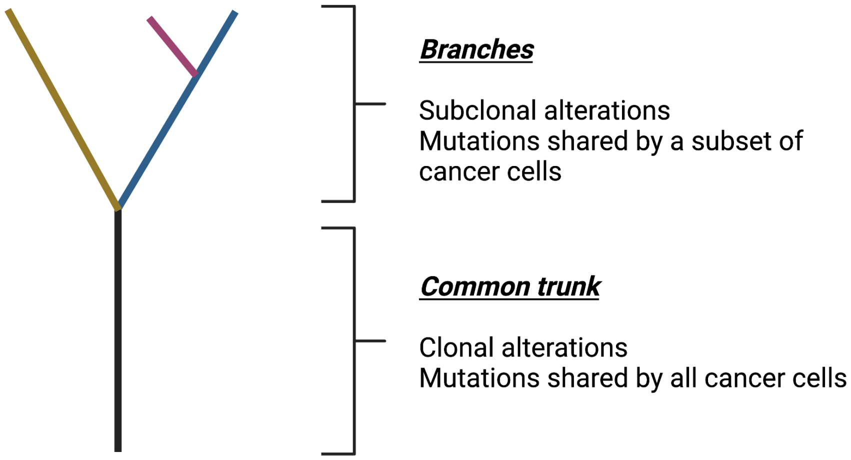

Branches
Subclonal alterations
Mutations shared by a subset of cancer cells
Common trunk
Clonal alterations
Mutations shared by all cancer cells

Figure 3: Evolutionary tree depicting the trunk carrying clonal alterations, and the branches carrying subclonal variation. Created with BioRender.com.

[25, 26].

Reconstructing the clonal structure, however, is not a trivial task. Many bioinformatics solutions have been developed to address this challenge, for example, by focusing solely on point mutations or by combining mutation and copy number data [27–30]. Regardless, the accuracy of these approaches depends heavily on sampling, specifically, the number and purity of the tumor regions being investigated. Through incomplete sampling, a subclonal variant might be falsely classified as clonal, giving rise to the problem termed clonal illusion [24] (Figure 4).

One potential solution to mitigate the effects of clonal illusion is to analyze data from plasma cell-free DNA, since it may offer a more holistic view of the tumor. A novel tool, ECLIPSE, was developed for this purpose as an important part of **Manuscript I**, and it will be described in the *Bioinformatics and ctDNA* section.

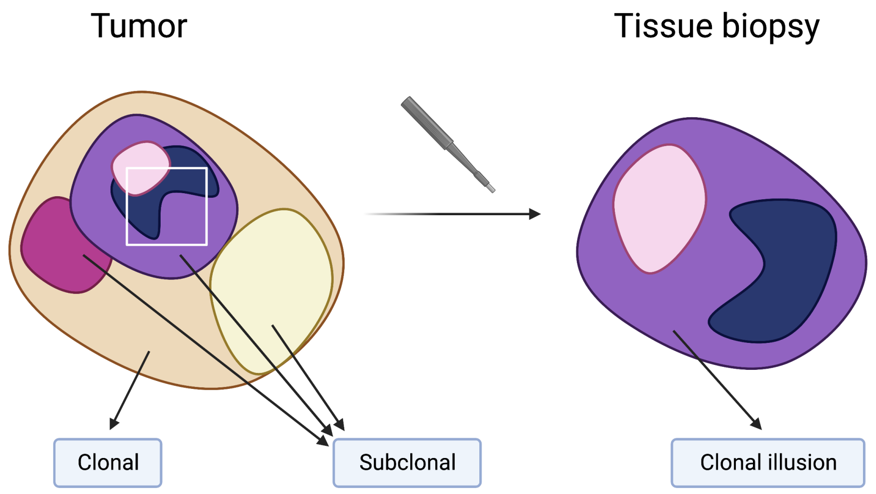


Figure 4: Clonal illusion. The white square depicts the site of a single tissue biopsy; the dark purple subclone appears clonal in the sample. Created with BioRender.com.

In my Ph.D., I worked with non-small-cell lung cancer, colorectal cancer, and metastatic melanoma cohorts. The next section provides a brief introduction to the biology of these distinct cancer types.

1.4 LUNG CANCER

Lung cancer is the leading cause of cancer-related mortality. In 2020, it was responsible for 18% of all cancer deaths, accounting for 1.8 million lost lives worldwide [3]. Lung cancers can be divided into two broad categories with different biology and growth patterns: small-cell lung carcinomas (SCLC) and non-small-cell lung carcinomas (NSCLC) [31]. Its emergence is mainly attributed to long-term tobacco use, however, other environmental factors such as air pollution and exposure to asbestos, as well as genetic predisposition have been proposed as carcinogenic risk factors. Lung cancer patients present with highly diverse symptoms and are often diagnosed at advanced disease stages [31].

1.4.1 *Non-small cell lung cancer*

NSCLC is responsible for 85-90% of all lung cancer incidences. Its most common subtypes are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), arising in the main and peripheral bronchi, respectively [31]. Smoking is the major risk factor in the development of this cancer type, more so in LUSC than LUAD. Never smokers are more likely to develop LUAD compared to LUSC, influ-

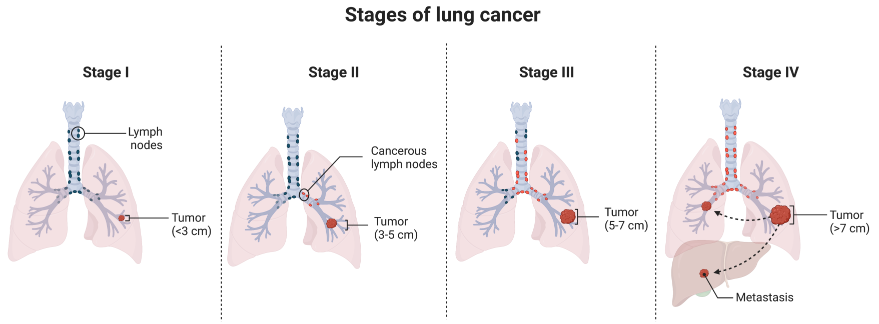


Figure 5: Stages of lung cancer. Created with BioRender.com.

enced by genetic predisposition and environmental factors such as second-hand smoking or air pollution [32]. Five-year survival rates range from 65% to 9% depending on whether the tumor is localized, regional, or distant. These unfavorable outcomes are partly due to rapid proliferation and already present metastatic disease at the time of diagnosis [33]. For Stage I, Stage II, and some Stage III patients, surgical resection is the most effective treatment option, yet a large percentage recurs, with a 5-year survival of 83% for Stage IA to 36% for Stage IIIA disease. Adjuvant cytotoxic therapy is associated with improved survival in these cases. Advanced NSCLC is often treated with platinum-based doublet therapy and unresectable locally advanced NSCLC patients are prescribed a combination of cytotoxic therapy and thoracic radiation. For those who progressed during or after platinum-based therapy, immunotherapy with nivolumab is available [32]. Since research uncovered deeper knowledge about the molecular profiles of the disease and thereby paved the way for personalized medicine, targeted therapies entered the clinic and led to improved outcomes in eligible patients [32]. In particular, the EGFR, PI3K/AKT/mTOR, RAS-MAPK, and NTRK/ROS1 pathways were demonstrated to be targetable in a clinically beneficial manner, leading to the development of drugs that are now used as first-line treatment [34–36]. The introduction of combinational therapies, such as EGFR tyrosine kinase inhibitors for treating both EGFR-sensitive and tyrosine kinase domain mutated tumors, has improved outcomes by targeting the resistance mutations that emerge over the course of and in response to therapy [34, 37].

NSCLC is molecularly heterogeneous and its histologies tend to carry alterations in a distinct set of genes, furthermore, the mutational landscape of smokers is different and richer than that of those who never smoked. The oncogenes KRAS, BRAF, NRAS, PIK3CA, and EGFR, as well as the tumor suppressors TP53, STK11, KEAP1, and NF1 were shown to be commonly mutated in LUADs, and the tumor suppressor genes TP53, PTEN, and CDKN2A tend to be deactivated in LUSCs [32, 38].

1.5 COLORECTAL CANCER

The second largest cause of cancer-related mortalities, colorectal cancer (CRC), was responsible for 9.4% of all cancer deaths in 2020 [3]. It originates from the healthy epithelium, which, through a gradual process, transforms into cancer and becomes malignant as it penetrates the muscularis mucosa [39]. Besides common risk factors such as age and hereditary aspects, lifestyle factors such as obesity, red meat consumption, and low physical activity have been linked to an increased risk of developing the disease [40]. Outcome is highly dependent on the disease stage - while the 5-year survival for early-stage colon and rectal cancers is around 88-92%, it drops to 12-13% for advanced, metastatic cancers [41]. Localized tumors and affected lymph vessels are removed via open surgery or laparoscopic resection. In addition to surgery, patients with a high risk of recurrence are given neoadjuvant radiotherapy. After curative resection, adjuvant therapy with chemotherapy agents like fluorouracil or oxaliplatin is recommended for stage III patients to increase disease-free survival [39].

The onset of colorectal cancer is attributed to a gradual accumulation of mutations over the span of decades. In particular, the tumor suppressor genes APC and TP53, the oncogenes RAS, BRAF, and PIK3CA, and the mismatch repair genes MLH1, MSH2, MSH6, and PMS2 have been associated with the development of CRC [42]. According to the adenoma-carcinoma model, carcinogenesis is initiated by a loss or mutation of APC, leading to the development of a polyp. A polyp is a benign growth, and its resection is standard procedure today. Following the aberration of APC, KRAS, TP53, and DCC mutations accumulate, contributing to the development of the malignant tumor [43] (Figure 6). Tumorigenesis is facilitated by the loss of genomic stability [42, 44]. Approximately 85% of CRCs display chromosomal instability (CIN), defined by the presence of chromosome copy number aberrations; the remaining 15% of CRCs show microsatellite instability (MSI), referring to the presence of genome-wide mono-or dinucleotide repeats. In MSI tumors, DNA mismatch repair (MMR) genes are inactivated, leaving the genome vulnerable to transcription errors [42, 45].

The diversity of the disease brought about the need to devise a method for stratifying patients in a clinically relevant manner. The development of the gene expression-based Consensus Molecular Subtypes (CMS) provided an answer to this problem and accelerated biologically interpretable disease classification. According to this system, CRC tumors can be divided into 4 major subtypes: hypermutated, microsatellite unstable CMS1 tumors with strong immune activation, representing 14% of CRCs; canonical CMS2 tumors charac-

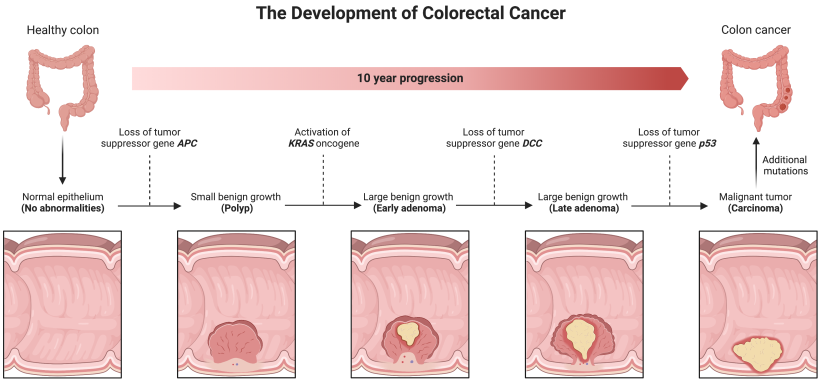

The Development of Colorectal Cancer
Healthy colon
Colon cancer
10 year progression
Loss of tumor suppressor gene *APC*
Activation of *KRAS* oncogene
Loss of tumor suppressor gene *DCC*
Loss of tumor suppressor gene *p53*
Additional mutations
Normal epithelium (No abnormalities)
Small benign growth (Polyp)
Large benign growth (Early adenoma)
Large benign growth (Late adenoma)
Malignant tumor (Carcinoma)

Figure 6: The development of colorectal cancer according to the adenoma-carcinoma model, as defined by Fearon and Vogelstein [43]. Created with BioRender.com.

terized by WNT and MYC signaling activation, representing 37% of CRCs; metabolically dysregulated CMS3 tumors, representing 13% of CRCs; and mesenchymal CMS4 tumors with prominent angiogenesis and stromal invasion, representing 23% of CRCs. The remaining 13% of tumors display mixed features, potentially due to ITH or a transitional phenotype [46]. To expand on molecular stratification, Bramsen and colleagues [47] developed a framework for combining molecular subtypes and subtype-specific prognostic biomarkers. They defined 3 cancer cell subtypes based on epithelial cell and stroma transcripts: secretory (enrichment in KRAS mutation and signaling), serrated (hypermethylated, hypermutated, and immune activated), and adsorptive (low methylation, chromosomal instability, and microsatellite stability). Additionally, they defined 5 tumor subtypes: goblet (characterized by secretory subtype features), stroma (characterized by tumor microenvironment properties, high expression of tumor stroma transcripts), SSC (characterized by serrated subtype features), dARE, and CIN (both characterized by adsorptive subtype features, with dARE being more microbiome-dependent). While cancer cell subtype was not found to be prognostic, they showed that tumor type classification and type-specific biomarker discovery can be used to enhance the prognostication of CRC [47].

1.6 MELANOMA

Malignant melanoma is responsible for approximately 0.6% of cancer mortalities [4], and incidence rates are increasing worldwide. The disease stems from the aberrant transformation of melanocytes that rise from the neural crest. Commonly, it occurs on the skin, yet it may also develop in other areas such as the brain or the gastrointestinal tract. UV-light and sunlight exposure, fair skin, and prior per-

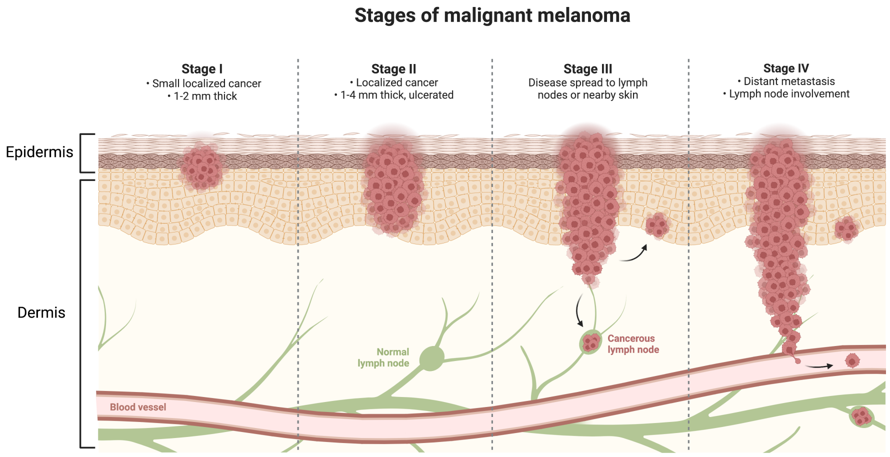


Figure 7: The stages of malignant melanoma. Created with BioRender.com.

sonal or family history are among the well-known risk factors [48]. Survival is highly dependent on the extent of the disease - while early-stage melanoma is highly curable and 97% of patients reach 5 years of survival, the survival rate for metastatic melanoma patients is only 10% [49]. Melanoma cases can be divided into four major disease subtypes based on their development patterns: superficial spreading melanoma (most common, flat and irregular shape), nodular melanoma (quick development, downward growth), lentigo malignant melanoma (slow growth), and acral lentiginous melanoma (rare, mostly found on palms, soles, under the nails or around the big toenail) [48]. Another categorization is based on the tissue of origin. According to this, the major subtypes are cutaneous melanoma (originating from non-glabrous skin), acral melanoma (originating from glabrous skin of palms, soles, and nail beds), mucosal melanoma (originating from the mucosa of internal tissue), and uveal melanoma (originating from the eye's uveal tract) [50]. Clinical intervention is dependent on stage as well as BRAF and PDL-1 status [51]. Early-stage melanomas are primarily treated via surgical removal of the malignant tissue, skin grafting, or tissue transfers, whereas advanced melanoma patients are eligible for chemotherapy or immunotherapy treatment. For instance, targeted therapy with BRAF-inhibitor agents such as vemurafenib and dabrafenib are recommended for BRAF-positive patients, and PD-1/PDL-1 immune checkpoint inhibitors such as pembrolizumab or nivolumab are used in the therapy of patients with high PDL-1 expression [49, 51]. In particular, the introduction of immunotherapy revolutionized metastatic melanoma care; almost 50% of the patients show positive response to immunotherapy treatment and a large fraction of the responders receive lasting clinical benefit [52–54]

Somatic mutation burden has been associated with the site of origin as sites suffering from chronic sun damage carry a higher number of mutations compared to sites that are less exposed to UV radiation. Aberrant activation in the MAP kinase pathway is a common occurrence at the early stages of tumor evolution, most often through BRAF and NRAS mutations [50, 55]. Genomically, melanomas can be classified based on their driver mutations; the most common, clinically distinct subtypes are BRAF-mutant, RAS-mutant, NF1-loss, and triple wild-type melanomas [50]. In particular, approximately 45-50% of cutaneous melanomas carry BRAF-activating mutations, primarily through the V600 codon in exon 15. More than 70% of these are of the V600E type, signaling a valine-glutamic acid substitution [56]. About 30% are of the RAS-mutant, 10-15% are of the NF1-mutant, and 5-10% are of the triple wild type subtypes. Advanced cutaneous melanomas tend to carry mutations in the telomerase gene TERT, which has been linked to poor survival [50, 57].

1.7 CANCER AND DATA

Systematic analysis of sequencing data made it possible to gain a deeper understanding of the signaling pathways and regulatory processes involved in cancer development [58–60]. Today, a wide variety of gene set definitions alongside accompanying data access and analysis tools, such as Gene Ontology (GO) [61], Kyoto Encyclopedia of Genes and Genomes (KEGG) [62], and Reactome [63], allow researchers to summarize genomic information and pursue enrichment analysis on a pathway level.

> In my work, I used the MSigDB Hallmark gene sets, the Reactome pathways, and the Sanchez-Vega pathway definitions.

1.7.1 *MSigDB*

The Molecular Signatures Database (MSigDB) [64] is a comprehensive online database containing more than 10,000 gene sets. Liberzon and colleagues introduced the Hallmark gene set collection in 2016, in an effort to increase the database's utility by reducing the redundancy and heterogeneity brought about by the vast diversity of the included gene sets [65]. They clustered similar sets of genes based on their founder gene set memberships and manually annotated the resulting clusters with associated biological themes. The gene sets were then distilled through transcriptomic data analysis, leaving co-expressed and biologically relevant genes in the set. Through this process, they defined 50 Hallmark gene sets that summarize specific biological states encompassing 8 process categories (cellular component, development, DNA damage, immune, metabolic, pathway, proliferation, and signaling). This refined collection enables sensitive, concise, and highly interpretable analysis of differential gene expression [65].

1.7.2 *Reactome*

The Reactome Knowledgebase is a manually curated pathway and process database with a complementary collection of bioinformatic tools [63]. The pathways were annotated by domain experts based on a reaction-focused, unified data model, and were cross-referenced to a variety of bioinformatic databases [66]. The current version of Reactome includes 10,726 genes grouped into 2546 pathways (such as M phase or DNA repair) and 28 super-pathways (such as metabolism or immune system). ReactomeGSA, their gene set analysis system optimized for different omics approaches, allows the discovery of distinct biological mechanisms, even in a multi-omics and cross-species setting [63].

1.7.3 *Sanchez-Vega pathways*

In 2018, Sanchez-Vega and colleagues curated 10 oncogenic signaling pathways by analyzing genetic alterations in the TCGA database [67]. They included somatic mutation, gene expression, copy number alteration, gene fusion, and DNA methylation data from 9125 tumors and 33 cancer types. Conducting this work on a pan-cancer level was a non-trivial effort given that mutation profiles vary between different tissues and cancer types; nevertheless, they showed that 89% of the tumors carried a mutation in at least one of the resulting pathways. The final set included the RTK/RAS pathway, Nrf2, PI3K, TGFß, Wnt, Myc, p53, cell cycle, Hippo, and Notch pathways [67].

2 LIQUID BIOPSIES

Cancer diagnosis, profiling, and monitoring have traditionally been done via examination of surgically resected tumor samples. This practice has a range of limitations: tumor samples are not always obtainable in sufficient quality and quantity; sampling may be impossible or non-repeatable as the process is highly invasive and may be dangerous to the patient; and a single tissue sample may not give a clear picture of a heterogeneous or metastasized tumor [68, 69]. In recent years, research has been focusing on exploring liquid biopsies as a minimally invasive and highly repeatable alternative to tissue biopsies. Liquid biopsies involve the analysis of bodily fluids, such as cerebrospinal fluid, urine, stool, mucosa, or most commonly, blood [69, 70].

2.1 BLOOD COMPONENTS

Blood contains a variety of cellular elements that differ by size and relative density. Through centrifugation, it can be separated into three layers: plasma, buffy coat, and packed red blood cells [71]. Plasma represents 55% of the blood and contains 91-92% water, electrolytes, proteins, immunoglobulins, coagulants, and notably, analytes that are relevant for cancer research [72]. The buffy coat (less than 1% of the blood) contains platelets and leukocytes and is usually the source of the germline sample in a tumor-normal pair [73–75]. The packed red blood cell layer, corresponding to the remaining 45% of the blood, is composed of erythrocytes [76] (Figure 8). Blood-based liquid biopsies contain a wide range of potentially tumor-originated materials, such as cell-free DNA, cell-free RNA, vesicles, or proteins. In particular, circulating tumor DNA (ctDNA), tumor-derived RNA, tumor-derived extracellular vesicles (EV), and circulating tumor cells (CTC) have captured clinical interest [69, 70].

2.1.1 *Circulating tumor cells*

CTCs are released from primary tumors into the bloodstream and are responsible for the development of metastases or second primaries at distant sites [77]. Their shape depends on the stage and type of the tumor they originated from, and they tend to form cellular aggregates with other cells to protect themselves against the immune system and oxidative stress [70]. As they can be used to follow the tumor's condition in a more real-time manner compared to other blood-based

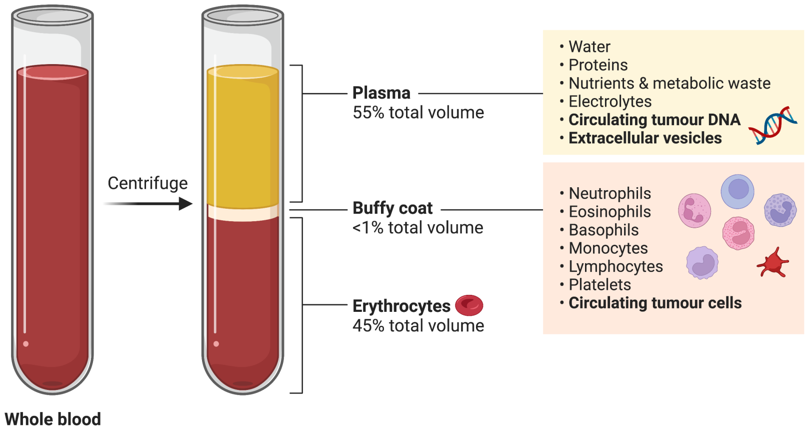


Figure 8: The contents of blood. Created with BioRender.com.

markers, CTCs have garnered significant interest in cancer diagnosis and treatment response monitoring [70, 78, 79]. A significant limitation of using these cells in clinical practice lies in their limited abundance in the blood (~ 1 CTC / 1,000,000 leukocytes), therefore, robust methods for enriching CTC yield need to be developed to increase the marker's applicability [80].

2.1.2 *Extracellular vesicles*

EVs are small, heterogeneous, membrane-bound vesicles secreted by any cell in the body. Originally thought to be cellular garbage bags, they carry a diverse cargo of DNA, RNA, proteins, and other biomolecules; additionally, they have been shown to play a role in cell-to-cell communication and signaling [70, 81]. Tumors, in particular, shed EVs at a remarkable rate, so EV concentration in the plasma is extremely high. Tumor-derived EVs were found to facilitate tumorigenesis via immunity, metastasis, and angiogenesis regulation, and have gained attention as a biomarker for early detection, diagnosis, and treatment monitoring [81–83].

2.1.3 *Cell-free DNA*

In 1948, Mandel and Métais discovered that cells release cell-free nucleic acids into the bloodstream [84]. 41 years later, Stroun and colleagues found evidence that a percentage of the cell-free DNA (cfDNA) is of tumor origin [85]. cfDNA can be found in healthy individuals, where its appearance is mostly attributed to the hematopoietic system [86]. Conditions of various severity, such as physical activity, inflammation, pregnancy, and cancer were shown to lead to elevated levels, therefore, deciphering the origin of cfDNA with sufficiently high specificity can be a challenging task [87–90].

2.1.4 *Circulating tumor DNA*

Circulating tumor DNA (ctDNA) represents <0.1-10% of the total cfDNA volume [91]. It originates from tumor cells, presumably through means of apoptosis, necrosis, or active secretion [92]; and promptly after its release, it gets cleared from the bloodstream via the liver, kidneys, and spleen [69]. It has a short half-life (16 minutes to 2.5 hours) [92] and thus can provide a real-time snapshot into the current state of the malignancy it originated from. This snapshot carries multifaceted information about the state of the disease at the time of sampling; the amount of ctDNA correlates with the tumor burden and the genetic and transcriptomic data derived from the plasma sample provide insight into the qualitative composition of the tumor. This qualitative information can include the presence of point mutations, copy number alterations, methylation profiles, chromosomal aberrations, and fragmentomically derived gene expression data [93–95]. A striking advantage of liquid biopsies and ctDNA analysis over tissue biopsies is therefore not only the relative convenience and higher feasibility due to the minimally invasive sampling but also the timely and versatile nature of the data. Due to these benefits, considerable research efforts have gone into exploring its utility in a wide range of clinical settings - as a prognostic marker, as a proxy for detecting the onset of cancer as well as the occurrence of residual disease, and as a tool to monitor response to therapy [96–100].

2.2 THE UTILITY OF CTDNA

The presence of ctDNA has been shown to have prognostic relevance, as ctDNA positive patients faced worse outcomes compared to their ctDNA negative counterparts in many studies across a variety of cancer types [101–104]. It also holds the potential to enable early cancer detection, which, paired with intervention, could lead to improved survival [105–107]. As cfDNA has been found to carry tissue-specific and cell-specific information, it may also be used for localizing the disease and identifying the tissue of origin; for example, in the case of cancers of unknown primary, distant metastases, or in a screening setting [108, 109]. Due to the highly repeatable sampling and the marker's snapshot-like nature, it has garnered interest for monitoring disease progression and response to therapy. It has been shown that ctDNA dynamics associate with response, and it may be used to detect response or relapse earlier than imaging-based follow-up approaches [98, 110, 111]. Recurrence monitoring and minimal residual disease detection are of particular importance as they allow clinicians

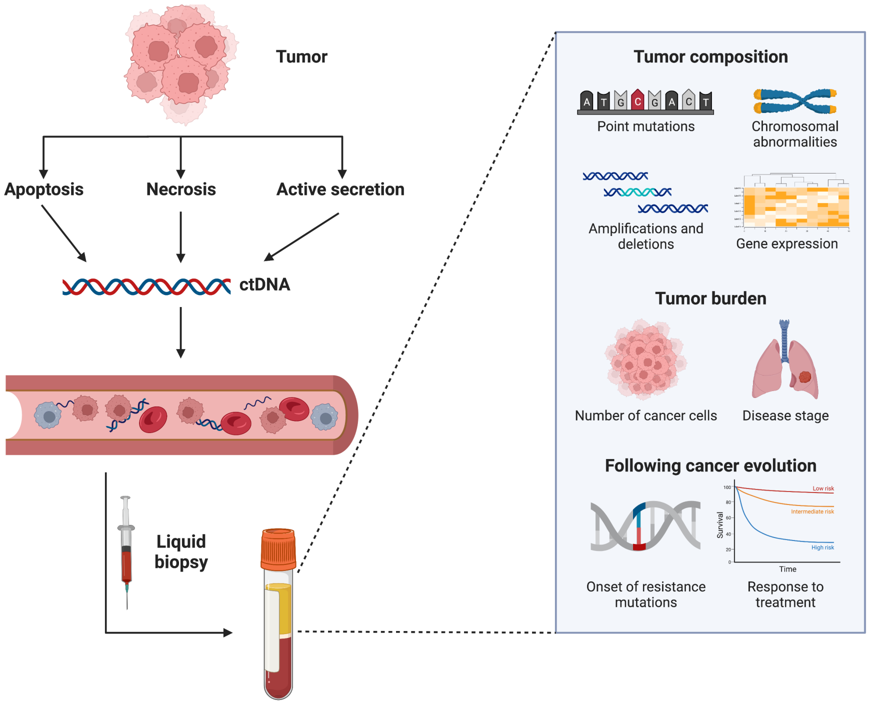


Figure 9: The origin and utilization of circulating tumor DNA. Created with BioRender.com.

to identify high-risk patients who are most likely to benefit from adjuvant treatment, based on their post-surgery ctDNA status [112, 113].

ctDNA can also provide insight into the composition of the tumor in a more holistic manner compared to a single tissue biopsy. It is now well-known that tumors can be a composite of distinct tumor regions with different mutation profiles [114], and, as those tumor regions shed ctDNA, this heterogeneity is reflected in the blood [115, 116]. By following the rise and disappearance of certain mutations over time, it may also be possible to monitor cancer evolution and identify variants that could have an effect on treatment response [117–119]. This opens the door to adaptive therapy, where the treatment regime is adjusted based on the appearance of resistance mutations in the patient's blood [92].

2.3 CTDNA DETECTION METHODS

Since ctDNA represents a small portion of the cfDNA volume, sensitive assay technologies are needed for reliable detection. Two main approaches, tumor-informed and tumor-agnostic, are available for the assessment of ctDNA content. In a tumor-informed setting, the genomic information obtained from the patient's primary tumor sample is used to guide the assay design, whereas, in a tumor-agnostic setting, panel design is more uniformized and is guided by domain knowledge [120]. A tumor-informed assay offers higher sensitivity and specificity compared to a tumor-agnostic approach and allows the detection of low-volume disease [121]. However, these benefits come at a cost - on one hand, the technology is expensive, and on the other hand, it requires a tumor sample from the patient which is invasive and at times unfeasible. In contrast, tumor-agnostic approaches tend to be more affordable, yet less sensitive; and they require rigorous data analysis to filter out false positives, germline variants, and sequencing errors.

ctDNA analysis methods can also be classified according to the genomic scale. In increasing order of cost and coverage, these range from single-locus or multiplexed assays, through targeted sequencing techniques, to whole-genome sequencing approaches. Single-locus or multiplexed assays include allele-specific or mutant allele-enriched Polymerase Chain Reaction (PCR) technologies, such as digital PCR, BEAMing, COLD-PCR, and SCODA; and can achieve a limit of detection (LOD) of 0.001%-0.01% variant allele frequency (VAF). Amplicon-based (eg. TAm-Seq, Safe-SeqS) or hybrid-capture (CAPP-Seq, digital sequencing) targeted sequencing approaches can reach <0.0.1% - 5% VAF LOD, depending on whether the assay is custom-built, off-the-shelf, or utilizes exome sequencing. Shallow whole genome sequencing (sWGS, eg. PARE) and amplicon-based (eg. FAST-SeqS) genome-wide approaches offer a sensitivity of 5%-10% VAF LOD [92].

In my projects, ctDNA detection was performed using three methods: Invitae's Personalized Cancer Monitoring with Anchored Multiplex PCR, Natera's Signatera assay, an in-house digital-droplet PCR method, and a custom Qiagen QIAseq targeted DNA panel. A brief introduction to these techniques is given below.

2.3.1 *Personalized Cancer Monitoring with Anchored Multiplex PCR*

The Personalized Cancer Monitoring (PCM) platform was created by the San Francisco-based company Invitae. This proprietary, pan-cancer, and tumor-informed technology was developed to detect ctDNA with high sensitivity and specificity, thus enabling the early detection of minimal residual disease (MRD) and post-treatment relapse. First, the patient's tumor and blood samples are sequenced using whole-exome sequencing, where the tumor sample is used to discover the patient- and cancer-specific variants and the paired blood sample is used to identify germline mutations. Based on this data, the company creates patient-specific assays with ~50 tumor-specific variants that can be used to monitor disease progression over time. Anchored Multiplex (AMP) PCR chemistry is used in the platform to enable the error-corrected detection of tumor DNA fragments in the plasma [122].

The assay has been shown to achieve >99% sensitivity for 0.005% VAF [123]. It has been validated and used in a number of publications involving the TRAcking Cancer Evolution through therapy (Rx) (TRACERx) multi-center NSCLC study [123, 124] and is continuing to undergo clinical validation.

The AMP PCR method was co-developed in collaboration with the TRACERx study. This assay was used in **Manuscript I** for tracking a median of 200 variants that were chosen based on tissue multiregion exome analysis. These variants included clonal and subclonal Single Nucleotide Variants (SNVs) and neoantigens, in other words, cancer-specific antigens originated from tumor-specific alterations [125].

2.3.2 *Signatera*

Signatera, developed by the Austin-based company Natera, is an FDA-approved technology built for personalized MRD surveillance [126]. Similar to Invitae's assay, it is a tumor-informed method relying on whole-exome sequencing of the patient's primary tumor [127]. After obtaining the patient-specific mutations from the tumor tissue and the buffy coat, the top 16 clonal somatic variants are used to design individualized primers for multiplex PCR. Blood samples are then collected longitudinally at predefined intervals, allowing response or recurrence monitoring over time. The extracted cfDNA is assayed with the 16-plex PCR, and after PCR amplification, ultra-deep next-generation sequencing is performed to detect the presence of ctDNA. The technology promises reliable detection of variants at 0.01%-0.1% VAF levels with >99.5% specificity [128]. The prototype was first used in an NSCLC context as part of a TRACERx study [129], and it has since been used in multiple publications across a variety of cancer types [130–132].

2.3.3 *Digital-droplet PCR*

PCR is an enzymatic assay that can be used to amplify a certain DNA fragment of interest [133]. Since its discovery in 1990 [134], it has gone through three iterations: the first version of PCR was purely qualitative; the second version, named real-time quantitative PCR, allowed quantitative analysis; and the third version, termed digital PCR (dPCR), provided the sensitivity and accuracy needed for detecting rare variants or trace amounts of DNA. Digital-droplet PCR (ddPCR) is a specific dPCR technique where the enzymatic amplification reaction is carried out in water-in-oil droplets [135]. It is easier and faster compared to next-generation sequencing (NGS) techniques, however, it cannot uncover novel targets and it is limited to testing single mutations [136]. As it offers a favorable LOD range of 0.001%-0.1%, it is considered to be one of the most accurate methods for detecting and quantifying ctDNA today [137].

2.3.4 *QIAseq targeted DNA panel*

The QIAseq technology and gene panel design service is offered by Qiagen, a biotechnology company based in Hilden, Germany. In contrast with the Invitae and Natera assays, these panels are not intended for clinical use. They offer a uniform, tumor-agnostic solution for low-frequency variant detection and promise >90% sensitivity for VAF of 1% [138]. It was developed for medium- and high-throughput NGS sequencers and it uses a target enrichment technology to allow the sequencing of specific genomic regions of interest. It implements an optimized reaction chemistry through the combination of targeted and universal PCR cycles, supported by the use of molecular barcoding to reduce false positives, PCR artifacts, and library bias [139]. Likely due to its affordability and universality, it has been a popular choice among researchers and has been featured in a wide array of publications [140–143].

2.4 BIOINFORMATICS AND CTDNA

The abundance of generated sequencing data necessitated the development of effective bioinformatics algorithms. Somatic variant callers like Mutect2 [144, 145], Strelka2 [146], and Shearwater [147] have been applied to a wide range of NGS data, whereas tools like MRDetect [148], INVAR [149], MRD-EDGE [150], DREAMS [151], ichorCNA [152], and MRD caller - ECLIPSE [153] were specifically developed for ctDNA detection and analysis. A striking advantage of the ctDNA-optimized tools lies in their enhanced sensitivity compared to the general variant callers, however, they are often more limited in scope and may require certain lab protocols or a pre-existing fragment database [151].

> In my work, I implemented a pipeline using Shearwater and I worked with data obtained from Mutect2 and MRD caller - ECLIPSE. These technologies are detailed below, and the rest of the tools are briefly introduced.

2.4.1 *Mutect2*

Mutect2 is a somatic variant caller developed and maintained by the Broad Institute. It is based on the germline variant caller Haplotype-

Caller - while detecting germline variants is relatively straightforward as they only need to be compared against the reference genome, calling somatic variants include determining the variation between two samples in contrast to the reference. For best results, Mutect2 should be used with a tumor sample and a matched normal, however, tumor-only mode is available yet its usage is not advisable since it produces a high number of false positives [154]. First, the tool determines the read assembly intervals by identifying active regions where variations are likely to occur. This is done by calculating a per-position activity profile via a simplified somatic genotyping model, denoting the probability that the position carries a variant. Subsequently, the positions that surpass a predefined activity threshold are identified. For each active region, a local assembly graph is generated from the reference sequence, and the reads are compared to the different segments of the graph. As mismatches between the reads and the reference graph are identified, the graph is updated with new nodes to account for the variation. The graph is scored using a version of the active region likelihood model. The resulting graph is then pruned by a two-pass adaptive pruning process, where first a global, empirical error rate is calculated based on the non-branching, unlikely subgraphs in the original graph, then the resulting error rate is used to recalculate likelihoods and prune the subgraphs. Candidate haplotype sequences, in other words, the statistically associated set of Single-Nucleotide Polymorphisms (SNP) are assembled by traversing the pruned graph, and the haplotypes with the highest scores are retained. Haplotype scoring is done at traversal. For each edge, the transition probability is calculated as the number of reads supporting the edge divided by the sum of read support for other edges originating from the same vertex. The haplotype's score is then the product of all the transition probabilities belonging to its edges. The retained haplotypes are aligned to the reference [155]. For evaluating the haplotypes, Mutect2 uses the Pair-HMM model, which is a Hidden Markov Model adapted for pairwise sequence alignment. Each read is aligned against each haplotype, including the reference, and a score denoting the likelihood of having been sequenced from a given haplotype is calculated. Based on the per-read haplotype likelihoods, read-allele matrices are calculated. Evidence for individual alleles is obtained from a Bayesian model, where the model evidence with the allele excluded from the set is divided by the model evidence supporting the full allele set [145, 156]. Filtering can be done with a companion tool, FilterMutectCalls, which filters the resulting variants based on their Mutect2 annotations. It implements a combination of hard filters (eg. poor mapping quality, panel of normals blacklist) and probabilistic error models (eg. germline model, normal artifact model), and it calculates the probability for each call to be a real somatic variant versus a germline variant, sequencing error, contamination, or other

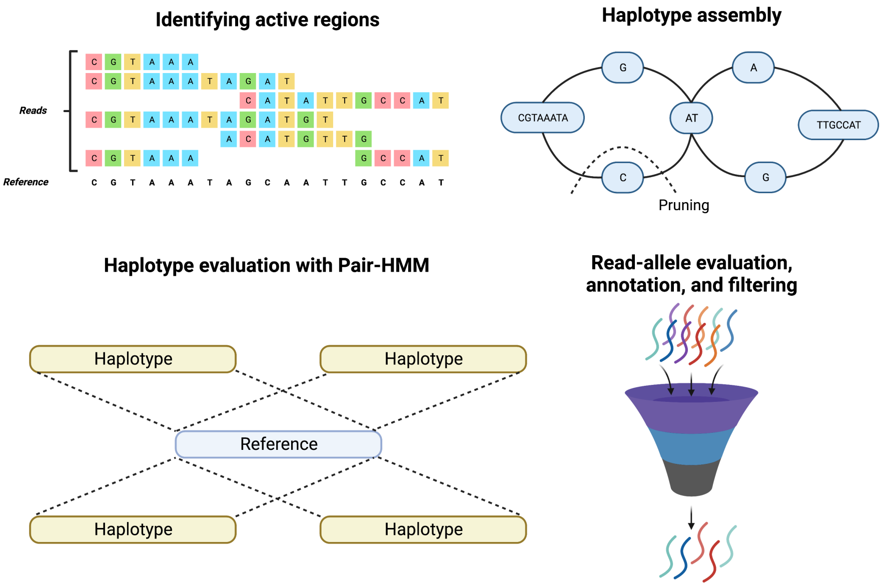

Identifying active regions
Reads
Reference
C G T A A A T A G C A A T T G C C A T
Haplotype assembly
CGTAAATA
G
C
AT
A
G
TTGCCAT
Pruning
Haplotype evaluation with Pair-HMM
Haplotype
Haplotype
Reference
Haplotype
Haplotype
Read-allele evaluation, annotation, and filtering

Figure 10: The Mutect2 pipeline. Created with BioRender.com.

artifacts [145, 157]. Mutect2 is a widely used tool providing stable and accurate results in a tissue analysis setting, however, it is computationally intensive and as a general somatic variant caller tool, it may not be sensitive enough for robust ctDNA analysis and cancer detection [151, 158].

2.4.2 *Strelka2*

Strelka2 is a germline and somatic variant calling tool developed by Illumina. In principle, it works similarly to Mutect2 as it involves read realignment and statistical modeling to evaluate the support for a given variant. The main differences lie in the haplotype modeling and variant evaluation steps. Strelka2 implements a tiered haplotype model, where the software uses a simple read alignment-based model or a more complex local assembly approach depending on the locus being investigated. Additionally, instead of a pair-HMM or other complete solution examining all possible pairwise alignments, it uses a small set of candidate alignments to approximate read likelihoods. Finally, a pre-trained random forest model is used to evaluate the variants, taking into account factors like genome probability, mapping quality, and read support. These considerations have been successful in decreasing computational burden without compromising scientific performance, as the authors showed that it could outperform Mutect2 and other variant callers in terms of precision as well as runtime. However, it was not specifically tailored for ctDNA detection, and it has been demonstrated that its performance decreases greatly at low allele frequencies [146, 158].

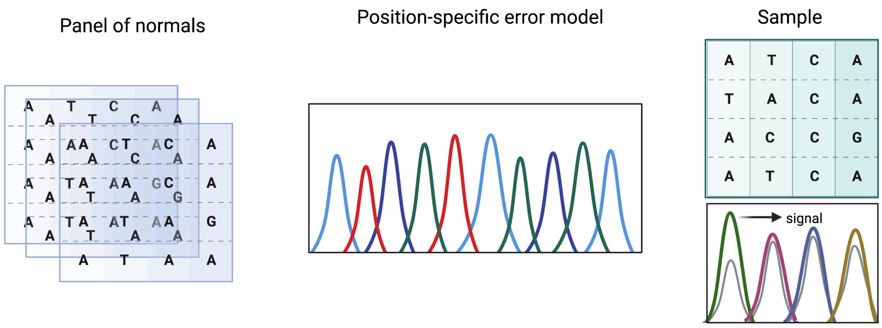


Figure 11: The Shearwater pipeline. Created with BioRender.com.

2.4.3 *Shearwater*

Shearwater, implemented as part of the deepSNV R package, is an algorithm designed for detecting low-frequency somatic variants in deep sequencing data. Based on a panel of normals, it builds a position-specific error model that captures the likelihood of variation in the background samples. The investigated sample's count distribution is then compared against the control distribution and a mutation is called if its signal surpasses the expected value. In this framework, nucleotide counts are modeled by the beta-binomial distribution, which differs from the binomial distribution in that the probability of success is not fixed. A true variant is understood as a variant that is present on both the forward and backward strands with high frequencies. According to the null hypothesis, the distribution of counts in the sample is not significantly different from the control count distribution; whereas the alternative hypothesis claims that the variant counts are significantly larger than the control counts. By default, the algorithm calculates Bayes factors using a likelihood ratio test; however, p-values can be obtained using its maximum likelihood adaptation, ShearwaterML [159]. The default method may be supplemented by a prior to account for previous knowledge regarding mutation frequency, for example, by using cancer type- or tissue-specific priors. While this tool can be used with a wide range of sequencing data, it is better suited for ctDNA analysis compared to Mutect2 or Strelka2 due to its optimization for low-frequency variants [147, 160].

In **Manuscript III**, I adapted Shearwater to the longitudinal setting, where, in the absence of control samples, the panel of normals was composed of the collection of samples independent of the patient whose sample was being analyzed.

2.4.4 *MRDetect*

MRDetect is a tumor-informed approach developed for monitoring low-burden disease in liquid biopsies. The main challenge surrounding ctDNA detection lies in its low concentration in the blood, which this tool aims to overcome by analyzing whole genome sequencing (WGS) data. MRDetect works by contrasting tumor and buffy coat samples, thereby identifying the tumor-derived SNV and copy number alteration (CNA) profiles. The primary tumor is used as a prior, and the buffy coat is used to filter out germline variation. A support vector machine (SVM) algorithm, trained on a pre-defined, curated truth set of alterations and errors, is then applied to the reads to perform noise reduction and classification in a sensitive manner. The tool is further enriched by the capability of integrating CNA data, which, combined with the orthogonal SNV information, can be used to strengthen the detection signal. This architecture allows effective ctDNA detection at low frequencies, however, it shows limited sensitivity for identifying individual mutations [161].

2.4.5 *MRD-EDGE*

MRD-EDGE is a WGS-based method for detecting SNVs and copy number variants (CNV) in ctDNA. It was created by the authors of MRDetect, and it is considered to be a more sensitive and widely applicable successor of the original tool. SNV detection is performed using an ensemble convolutional neural network / multilayer perceptron model that, instead of germline samples, is trained on high tumor fraction plasma samples, thereby learning a ctDNA-specific feature space. CNV analysis is assisted by a noise reduction method based on robust principal component analysis [162]. CNVs are evaluated by pooling information from minor allele frequency and fragmentomic analyses. The authors showed that this method can outperform MRDetect in the tumor-informed setting, and may also be used to call de novo mutations without a matched tumor sample due to its enhanced signal enrichment capabilities [150, 162].

2.4.6 *ichorCNA*

ichorCNA, developed by researchers at the Broad Institute, uses ultra-low-pass WGS data to evaluate the tumor fraction of the cfDNA sample. CNA signal is modeled as a tumor-derived and non-tumor-derived fragment admixture on the basis of the tumor proportion, a specific alteration's copy number, and in the subclonal CNA case, the tumor proportion from which the given alteration is missing. Then, a hidden Markov model is implemented for simultaneous genome segmentation, CNA prediction, and tumor fraction estimation. The tool can be used without a matched control, however, building and utilizing a panel of normals for the purposes of noise reduction and accuracy enhancement is also possible. Since ultra-low-pass sequencing is cost-effective, ichorCNA has the potential to aid ctDNA detection-based patient stratification at scale or identify cases where deeper sequencing is needed [152].

2.4.7 *INVAR*

INVAR is a method developed for monitoring ctDNA primarily in a patient-specific setting. The error model used for noise reduction is based on trinucleotide contexts, and the statistical model in the pipeline compares the distribution of fragment lengths to distinguish between ctDNA and cfDNA. To optimize the ctDNA detection threshold to high sensitivity and specificity, the method uses independent patients' samples as negative controls. Even though INVAR was developed as a patient-specific method, the authors show that it can be applied to general WGS data as long as a curated tumor mutation list is available, however, the error suppression framework is less effective in this case [149].

2.4.8 *DREAMS*

At its core, DREAMS is a read-level error modeling approach enabling position-specific error rate estimation. A neural network model is trained to predict the error rates, using a combination of local sequence context features such as trinucleotide context or GC content, and read-level features such as fragment lengths or strand as input. Using the resulting error model, two statistical methods, DREAMS-vc and DREAMS-cc were developed for calling variants and detecting cancer from ctDNA, respectively. The method does not have strict requirements for the model training data since it does not necessarily need to be of ctDNA origin, rather, independent normal samples, mutation-filtered tumor samples, or matched control samples may be used. These tools can only be used for detecting SNVs, however, it promises good accuracy for this use case as authors showed that it can outperform Shearwater and Mutect2 [151].

2.4.9 *MRD caller - ECLIPSE*

MRD caller and ECLIPSE are two algorithms that represent the core computational methodology behind **Manuscript I**. First, the MRD caller estimates trinucleotide-specific, intralibrary sequencing error rates based on the reads aligned against the reference genome. The reads are supported by unique molecular identifiers (UMI) to increase accuracy. The consensus sequence is then filtered for noise and artifacts, such as strand bias, background error, and variant allele frequency outliers; and variation at tumor-informed positions is evaluated. Deep alternate observations (DAO) are calculated as the number of UMI-supported alterations for each trinucleotide context within the assay's capture regions. ctDNA detection is determined by conducting a Poisson-test comparing the number of observed DAOs across the tumor-specific positions to the expected DAOs obtained from the background error model. Individual mutations are called in a similar manner, assessing whether the DAOs were significantly higher than the background error for a given alteration. Subsequently, ECLIPSE utilizes the resulting variant data, clonal/subclonal mutation status, background noise estimates, and copy number information derived from the primary tumor to resolve clonal architecture, in particular, to calculate the cancer cell fraction (CCF) of the clones. Tumor purity, or the percentage of tumor-originated cells from which DNA is derived, is estimated using the tumor and normal copy numbers, the VAF, and multiplicity, referring to the number of mutant DNA copies residing in mutant cells. The CCF is then computed on the basis of tumor purity, tumor and normal copy numbers, VAF, and multiplicity. CCFs range from 0 to 1, where a CCF of 1 means that 100% of cancer cells carry a specific event, a CCF of 0.5 means that 50% of cancer cells carry the event, and a CCF of 0 means that the event is absent from all of the cancer cells. Additionally, ECLIPSE can identify clonal sweeps, a phenomenon where a previously subclonal mutation reaches a CCF of 1, in other words, is now present in every cancer cell. The presence of a clonal sweep is detected by conducting a Wilcoxon test to compare the CCFs of the subclones to the CCFs of clonal mutations within a given sample. Using the CCF information over time, metastatic dissemination patterns can be identified. Monoclonal relapse means that all subclones found in the primary tumor were also present with CCF = 1 in the postoperative ctDNA, while in polyclonal relapse, some subclones were found to be present in significantly less than 100% of the cells. Polyclonal relapse can be divided into polyclonal monophyletic and polyclonal polyphyletic relapse, depending on whether the subclones were direct descendants or branched into different lineages in the evolutionary tree, respectively. Besides enabling the discovery of complex evolutionary information, a significant strength of the method is that it was optimized to work with ctDNA levels as low as <1% [153].

2.5 THE ADVANTAGES AND CHALLENGES OF CTDNA ANALYSIS

Liquid biopsies, and in particular, ctDNA analysis shows considerable potential at different stages of the cancer care lifecycle. It is a versatile marker - its presence carries prognostic value and can indicate the onset or recurrence of the disease, and its quantity and quality

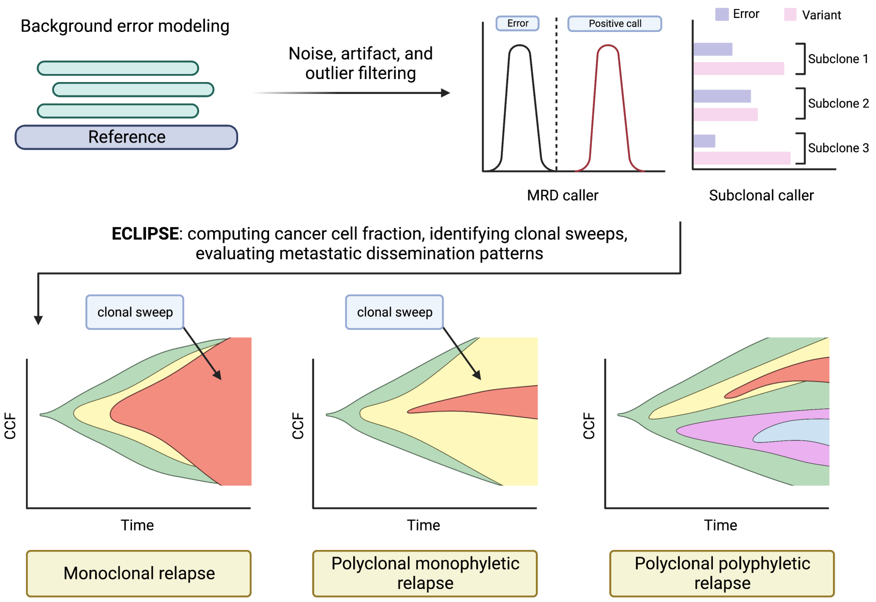

Background error modeling
Reference
Noise, artifact, and outlier filtering
Error
Positive call
MRD caller
Error
Variant
Subclone 1
Subclone 2
Subclone 3
Subclonal caller
ECLIPSE: computing cancer cell fraction, identifying clonal sweeps, evaluating metastatic dissemination patterns
clonal sweep
clonal sweep
CCF
CCF
CCF
Time
Time
Time
Monoclonal relapse
Polyclonal monophyletic relapse
Polyclonal polyphyletic relapse

Figure 12: The MRD caller - ECLIPSE pipeline. This figure was inspired by Abbosh et al. 2023 [153] and redrawn in BioRender.com by Judit Kisistók.

provide a window into the quantity and quality of the tumor of origin at the time of sampling. Blood samples are less invasive to obtain compared to tissue biopsies, and blood tests can be repeated more often than imaging-based follow-up methods such as CT scanning. When implemented in the clinic, ctDNA holds immense promise to provide clinicians with timely and actionable data.

This promise, however, is not completely fulfilled yet. Despite the striking advantages, the widespread adoption of ctDNA analysis faces a set of biological and technical challenges.

From a technical point of view, the challenge lies in the low concentration of ctDNA available in a standard blood test, a problem even more pronounced in the case of early detection since small tumors shed less ctDNA. It is, therefore, probabilistically possible that a given blood sample simply does not contain any ctDNA fragments or a specific mutation of interest. This highlights the need for sensitive, multi-target assays, as each target increases the chance to identify the presence of a tumor-originated variant [92, 110]. Conducting plasmapheresis to collect a larger plasma volume would be another solution, however, it is less convenient and more time-consuming compared to a routine blood test.

From a biological perspective, challenges stem from the questions surrounding the cause and associates of ctDNA release. First, liquid biopsies contain biological noise, primarily due to clonal hematopoiesis of indeterminate potential (CHIP). CHIP refers to the normal, non-tumor-origin accumulation of somatic variants. These variants are rarely present below the age of 40, but they increase in frequency with age (ranging from 9.5%-18.4% of people between ages 70-108) and can cause false positives during analysis. Sequencing control samples, such as the buffy coat, can be used to filter out CHIP variants [163, 164]. Secondly, ctDNA needs to be sufficiently distinguished from cfDNA. Since it has been shown that they display different fragmentation patterns where cfDNA shows a maximum peak at 167 base pairs (range of 120-220 base pairs) and ctDNA shows a maximum peak around 145 base pairs (range of 50-150 base pairs), size-optimized extraction methods can help identify ctDNA content [165–167]. Thirdly, ctDNA levels vary between cancer types, histologies, stages, and even individual patients. While some of the variation can be explained by tumor size [104], it has been shown that tumor burden alone does not account for the full spectrum of ctDNA shedding [103, 129]. Avanzini et al. [168] note that apart from tumor volume, other factors like clinicopathological features, histology, and disease stage may also influence shedding rate, and variance across tumor types as well as across individual patients complicate the matter further. Bladder, colorectal, and gastroesophageal cancers are among the ubiquitous ctDNA-shedder cancer types, whereas glioma, thyroid, renal cell carcinoma, and prostate cancers were found to sparsely release ctDNA in detectable quantities [103]. Granular, histology-specific investigation is necessary, however, as it has been demonstrated in non-small-cell lung cancer that while lung squamous cell carcinomas are predominantly ctDNA positives, lung adenocarcinomas do not shed ctDNA with the same near-uniform frequency [129].

In summary, liquid biopsies and ctDNA analysis hold tremendous potential in enabling oncologists to follow the effect of cancer treatment in real time by looking at the amount of ctDNA as a proxy for the number of cancer cells; and to observe the rise of potentially resistant subclones harboring acquired resistance mutations. It is also evident, however, that a deeper understanding of ctDNA biology is needed to fulfill this potential. This matter is not yet well-understood and in particular, the observation that some tumors shed more ctDNA than others remains largely unexplained. Early disease detection is particularly demanding as the technical challenges associated with the limit of detection become a concern at low ctDNA fractions.

Part II

RESULTS

3 OVERVIEW OF THE TOPICS OF THIS THESIS

Liquid biopsies offer a promising and minimally invasive alternative to tissue biopsies. While sample collection is relatively easy and the obtained information provides a multi-faceted and real-time snapshot into the quality and quantity of the tumor, the method is not without challenges. In particular, cancer type-, histology-, and patient-level variations in ctDNA shedding are not well-understood and thus limit the marker's utility in the clinic.

Despite considerable interest and efforts in the field, the exact biology of ctDNA release remains elusive. The primary aim of this Ph.D. study was, therefore, to elucidate the biological associates and clinical utility of ctDNA in a variety of cancer types and study settings, broadening the understanding of the mechanisms of ctDNA release, and, hopefully, contributing to better implementation of this marker in a clinical setting. My work encompasses the following two themes:

3.1 EXPLORING CTDNA BIOLOGY

In **Manuscript I**, I investigated the differential biology behind ctDNA shedding in a cohort of LUAD patients, utilizing multi-region transcriptomic, genomic, and chromosomal instability data.

In **Manuscript II**, I built on the findings of **Manuscript I**. I investigated the biology of ctDNA release in a CRC cohort, and I performed a comparative analysis of NSCLC and CRC biology as it relates to ctDNA shedding.

3.2 LONGITUDINAL DISEASE TRACKING USING CTDNA

In **Manuscript III**, I analyzed longitudinal ctDNA data collected from a small cohort of metastatic melanoma (skin cutaneous melanoma, SKCM) patients, and aimed to investigate genomic alterations that may provide an early insight into response to therapy.

4 RESULTS

4.1 MANUSCRIPT I

Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA

Nature, 2023 [153]

Minimal Residual Disease (MRD) refers to the number of cancer cells present in the body following treatment, and is an indicator of long-term outcome. Current clinical practice is cautiously adopting ctDNA as a tool to detect and profile MRD, and to potentially guide adjuvant treatment [169]; however, a more comprehensive understanding of this marker's biology is needed to aid its utility and clinical implementation.

In this study, I aimed to investigate the associates of preoperative ctDNA detection in a LUAD cohort, collected as part of the TRACERx study [170]. Additionally, I sought to explore whether a well-defined, differential ctDNA shedding phenotype exists in this cancer type.

I analyzed the transcriptomic landscape of 34 ctDNA positive and 28 ctDNA low-shedder LUADs. Differential gene expression analysis revealed a striking difference between these subgroups; in particular, genes associated with proliferation, cell cycle, and DNA repair (eg. *AURKB, POLQ, CDK4, BUB1B*) were upregulated in ctDNA shedders. By exploring enrichment within the context of the Hallmark gene sets, I strengthened these findings further as pathways related to M-phase, cell cycle, and proliferation (eg. *E2F targets, G2M checkpoint, MYC targets V1, MYC targets V2*) were significantly enriched in patients who exhibited preoperative ctDNA release. When I tested the association between shedding status and ORACLE [171], a transcriptomic-based, prognostic biomarker, I found that ctDNA positives harbored significantly higher risk scores, suggesting worse clinical outcomes compared to their ctDNA negative counterparts (Figure 13).

I compared the genomic profiles of these tumors and I observed no significant differences in clonal or subclonal mutation frequency, neither on an individual nor per-pathway [67] level. I found that ctDNA positive LUADs showed significantly higher levels of chromosomal instability, quantified by the weighted genome integrity index (wGII) and fraction of loss of heterozygosity (FLOH); and displayed

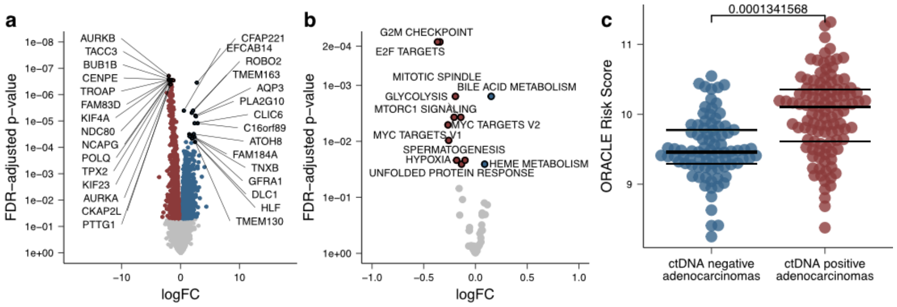


Figure 13: A-B) Gene-level and pathway-level transcriptomic analyses of the LUAD cohort. C) Association between ctDNA shedding and the ORACLE risk score.

a higher prevalence of whole-genome doubling events compared to ctDNA negatives. By performing copy number analysis on our genomic dataset, I, with the help of collaborators, identified 20 amplified cytobands in the ctDNA shedder group. Of the 966 genes located on these cytobands, 21 were previously identified as cancer genes by the COSMIC cancer gene census [172], including genes that play a role in proliferative processes (eg. *CCND1*, *CDK4*, *MDM2*).

In conclusion, I demonstrated that ctDNA shedding in LUAD associates with an aggressive, highly proliferative disease phenotype, and similarly, less aggressive disease is linked to a phenotype that does not shed ctDNA in detectable quantities.

4.2 MANUSCRIPT II

Exploring the biology of ctDNA release in colon cancer

ready for submission

Colorectal cancer is one of the leading causes of cancer-related deaths worldwide. Clinically relevant patient stratification, early detection of the emergence of the cancer as well as detection of post-treatment MRD may have a significant positive impact on patients suffering from this heterogeneous disease. While circulating tumor DNA offers a promising solution to these challenges, the biology and mechanisms of ctDNA shedding in this cancer type and its distinct subtypes have not been thoroughly explored so far.

In this study, I analyzed transcriptomic, clinical, and whole exome sequencing (WES) data collected from a cohort of Stage I-IV CRC patients treated at Aarhus University Hospital.

I conducted pairwise statistical analyses comparing the clinical characteristics of the two subgroups. Notably, I observed that tumors of ctDNA positive patients were significantly larger than those of ctDNA

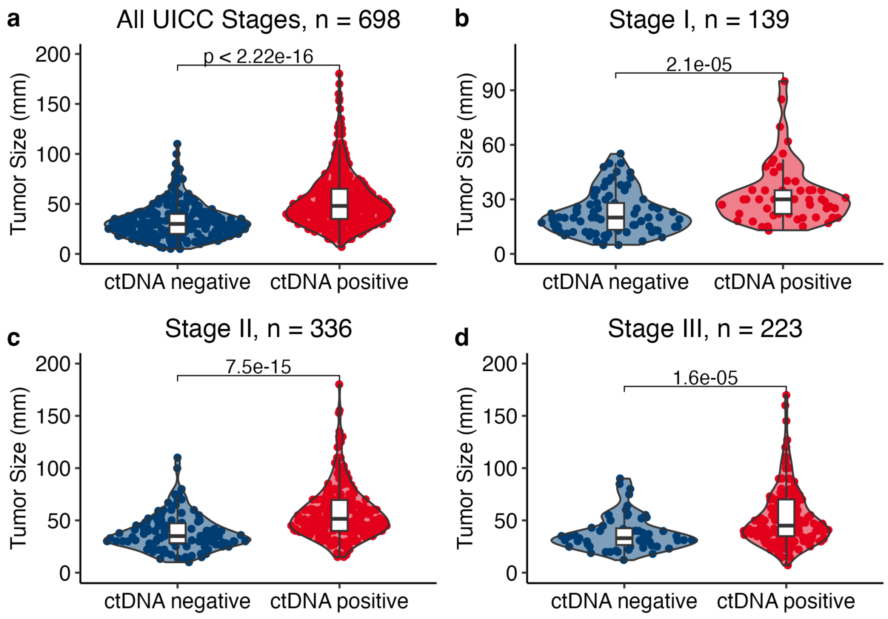


Figure 14: Association between ctDNA shedding in CRC and the tumor's largest diameter, stratified according to disease stage.

negatives, regardless of disease stage (Figure 14). In addition to tumor size, I identified significant associations between ctDNA shedding and molecular subtypes, MMR status, recurrence, and tumor location. In particular, I observed that secretory and CMS3 tumors, as well as MMR proficient tumors, shed lower amounts of ctDNA than their goblet and adsorptive, CMS1, CMS2, and CMS4, and MMR deficient counterparts, respectively. Additionally, ctDNA positive patients tended to suffer from recurrence and harbor tumors located in the left colon or rectum with higher frequency compared to ctDNA negatives.

When comparing the mutational landscape of ctDNA positive and negative tumors, I have found that the distribution of mutations affecting genes or pathways did not show an association with shedding status. Comparing primary tumor transcriptomic data from 86 ctDNA positives and 15 ctDNA negatives did not reveal any differentially expressed genes, however, I did note a moderate proliferative signal on the pathway level. When visualizing the significant hallmark Gene Set Variation Analysis (GSVA) scores on a per-patient basis, I noted a distinct proliferation-driven clustering of individuals, characterized by the enrichment of *MYC targets V1 and V2, G2M checkpoint, E2F targets, unfolded protein response, MTORC1 signaling*, and *DNA repair* pathways. Involving these pathways, I derived patient categories by classifying ctDNA positives as low-or high-proliferation shedders based on their mean GSVA enrichment scores. I observed that high-proliferation shedders release higher amounts of ctDNA and tend to

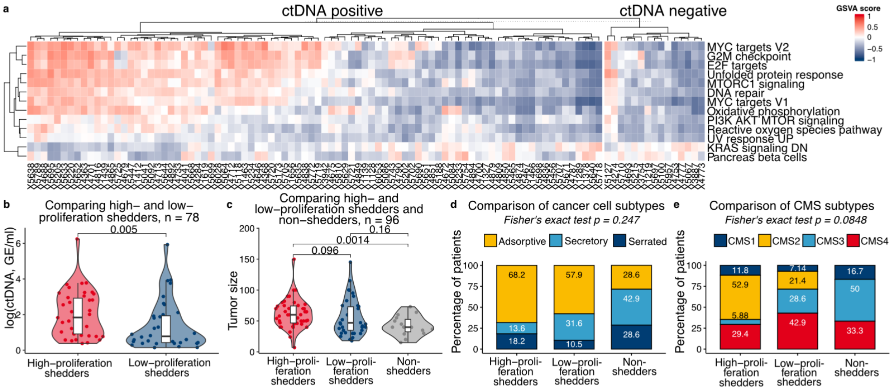


Figure 15: CRC proliferation categories. A) Per-patient GSVA scores. B-E) Association between proliferation and ctDNA concentration, tumor size, cancer cell subtypes, and CMS subtypes.

be larger than low-proliferation shedders and nonshedders. Additionally, low-proliferation shedders and nonshedders showed an overrepresentation of the secretory, CMS3, and CMS4 subtypes, compared to the high-proliferation shedders (Figure 15).

Interestingly, when a collaborator and I compared the transcriptomic profiles of 228 colon adenocarcinoma (COAD), 486 LUAD, and 475 LUSC tumor samples using the TCGA dataset, we found that both LUSC and COAD tumors exhibit a highly proliferative phenotype, whereas, in line with the findings of Manuscript I, LUADs display a bimodal proliferation distribution. These results were reproduced when we compared our CRC dataset with a previously published NSCLC cohort of 58 LUAD and 31 LUSC samples.

In conclusion, I proposed that ctDNA shedding in CRCs is a complex process mainly driven by proliferation and tumor size. Additionally, I demonstrated that COAD and CRC biology shows a resemblance with LUSC biology, suggesting that, similarly to the high-shedder LUSCs, CRC tumors also release ctDNA ubiquitously.

4.3 MANUSCRIPT III

Analysis of circulating tumor DNA during checkpoint-inhibition in metastatic melanoma using a tumor-agnostic panel

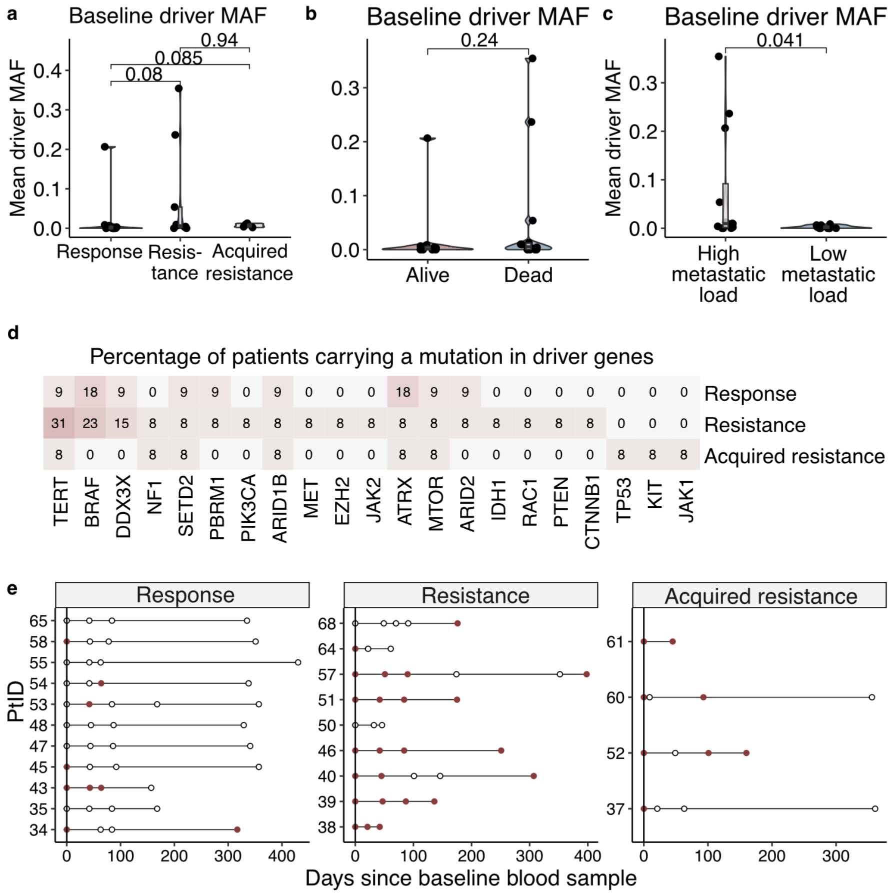
accepted in Melanoma Research, 2023

Metastatic melanoma is an aggressive disease with poor prognosis. Even though the introduction of immunotherapy treatment brought about a significant positive change in clinical outcomes, less than 50% of the patients receive lasting benefits [52–54]. Since immunotherapy is associated with persistent and severe side effects [173], identifying cases where the drawbacks outweigh the benefits is crucial in guiding or stopping treatment. Currently, disease progression is being monitored through CT-scans taken at regular intervals, however, a less invasive and more frequently repeatable alternative to this practice would be favorable. In this study, I investigated whether tumor-agnostic ctDNA analysis would be a feasible tool for following cancer evolution and gaining insight into therapy response.

I analyzed longitudinal ctDNA samples from 24 patients undergoing checkpoint-inhibition therapy. Out of these 24 patients, 11 were categorized as responders, 9 were resistant to therapy, and 4 showed initial response yet acquired resistance over time. The samples were collected at four different time points, one at baseline and three during the course of treatment. Plasma was sequenced using a custom panel of 40 well-established, melanoma-specific genes, with the purpose of tracking cancer progression and identifying the rise of mutant clones that might be associated with the onset of acquired resistance. Mutation calling was performed by an in-house bioinformatic pipeline.

In this limited cohort, I observed that baseline mutant allele frequency (MAF) was not significantly associated with response or survival. I found an association with metastatic burden, where ctDNA baseline MAF was significantly higher for patients with high metastatic load. Interestingly, the frequency of baseline ctDNA detection did not associate with metastatic load or survival, yet it varied with response category, where resistant and acquired resistant patients were ctDNA positive at baseline more often than responders. Of the 40 genes included in our panel, genomic analysis revealed that *TERT*, a gene whose aberrant activation is associated with poor outcome, was the most mutated gene in the nonresponder, high metastatic load, and deceased subgroups. ctDNA detection was possible in 12/13 resistant and acquired resistant patients, indicating that our tumor-agnostic panel was able to detect cancer in 92.3% of patients. When following the emergence and disappearance of specific alterations over the course of treatment, I did not observe a pattern in ctDNA dynamics that may be used as an indicator for immunotherapy response (Figure 16).

In conclusion, I demonstrated the utility of genomic data analysis within the context of understanding immunotherapy response in metastatic melanoma patients. I identified *TERT* as a potential predictor for poor response to therapy, and I showed that a tumor-agnostic ctDNA panel may be used to gain information about cancer progression.



	TERT	BRAF	DDX3X	NF1	SETD2	PBRM1	PIK3CA	ARID1B	MET	EZH2	JAK2	ATRX	MTOR	ARID2	IDH1	RAC1	PTEN	CTNNB1	TP53	KIT	JAK1
Response	9	18	9	0	9	9	0	9	0	0	0	18	9	9	0	0	0	0	0	0	0
Resistance	31	23	15	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	0	0	0
Acquired resistance	8	0	0	8	8	0	0	8	0	0	0	8	8	0	0	0	0	0	8	8	8

Figure 16: ctDNA detection and genomic analyses in a metastatic melanoma cohort. A-C) Comparison of baseline MAFs in response, survival, and metastatic load categories. D) Percentage of patients carrying a mutation in the investigated genes, per response category. E) Longitudinal ctDNA detection, per response category.

5 DISCUSSION

5.1 EXPLORING CTDNA BIOLOGY

Circulating tumor DNA has been the subject of active research interest in recent years, due to its potential of offering a real-time snapshot into the malignancy. It is hypothesized to be released into the bloodstream through apoptosis, necrosis, and active secretion [92], and the process has been previously linked to various factors such as tumor size, rate of cell birth and cell death, chromosomal instability, and clearance [168, 174].

It has been demonstrated that different cancer types and even different histologies within cancer types may display different shedding behavior [103, 129]. In **Manuscript I** and **Manuscript II**, I aimed to investigate this phenomenon in a lung adenocarcinoma and a colorectal cancer cohort.

Interestingly, I found that LUADs carry two distinct biological phenotypes. When comparing the transcriptomics of ctDNA positives and ctDNA negatives, I found that genes and pathways involved in proliferative processes were overexpressed and enriched in shedders. The copy number analysis supported this result, as I identified 21 known cancer genes, including proliferation-associated ones, located on amplified cytobands that were enriched in ctDNA shedders. High proliferation is a hallmark of malignant, uncontrolled tumor growth [8] and has been linked to poor prognosis in a variety of cancer types [175–178]. Apart from high proliferation, high CIN is also regarded as a hallmark of aggressive disease [179]. Indeed, I found evidence that ctDNA shedder LUADs exhibit higher levels of CIN compared to ctDNA low-shedders, indicating potentially unfavorable outcomes for these patients. When I tested ctDNA shedding against ORACLE, a proliferation-associated prognostic biomarker where higher levels associate with poor survival in LUAD, I found that ctDNA positives harbored significantly higher risk scores.

Tumor purity and tumor size represent potential confounders to this analysis. On one hand, higher-purity tumor samples may represent a stronger signal, thereby disproportionately inflating my results. To address this, I compared the purity measures of our samples and I found no significant differences between ctDNA positives and ctDNA negatives. On the other hand, not only does tumor size correlate with ctDNA shedding [129], but larger tumor sizes also imply higher levels of proliferation, thus posing the question of whether tumor size, instead of differential biology, drives my findings. To investigate this, I excluded the smallest 25% of low-shedder LUAD tumors and repeated the analysis. My results largely reproduced, indicating that biology may indeed be the main driving force behind my results. Additionally, to minimize the effect of technical ctDNA detection limitations, I, with the help of a collaborator, used a regression model to identify ctDNA negatives that may be non-shedders only due to their tumor size, then removed these patients from the biological analyses.

Synthesizing all these observations, I established that in LUADs, a highly proliferative, aggressive disease phenotype exists, in contrast with a more indolent disease type harboring phenotypes that do not shed ctDNA in detectable quantities. Integrating this finding as a form of ctDNA-based patient stratification into clinical practice opens the possibility of escalating therapy for those who might be facing a poor disease prognosis, thus improving patient outcomes while reducing the risk of overtreating those that are less likely to benefit.

In comparison, CRC tumors display strikingly different behavior. While LUADs are bimodal in terms of their shedding phenotype, CRCs appear to be uniformly high-proliferative, ubiquitous shedders. I observed that in this cancer type, tumor size shows the strongest association with ctDNA shedding, and the process might be further supported by enriched proliferative capacity. This is in stark contrast with the biologically distinct LUADs, where tumor size alone, albeit important, does not fully explain shedding behavior [129]. Additionally, I found that CMS3 and secretory subtypes shed lower amounts of ctDNA. These subtypes are associated with less aggressive disease, which is in line with current findings linking ctDNA detection to poor outcomes [129, 180, 181]. Additionally, they tend to fall into the lower end of the proliferation distribution and carry an enrichment in KRAS mutations, suggesting that ctDNA may be promptly cleared from the bloodstream due to enhanced metabolic adaptation [46, 47].

Since I did not identify further significant differences in clinical factors or genomic alterations in cancer driver genes or pathways when comparing ctDNA positive and ctDNA negative tumors, I hypothesize that ctDNA shedding in CRC is not driven by phenotypic separation. This finding was strengthened by a comparative analysis of our in-house CRC cohort and a previously published NSCLC dataset [129]. In contrast with the high-and low-proliferative LUADs, I observed that both CRC and LUSC tumors displayed uniformly high levels of proliferation. In order to confirm that this result is not driven by batch effects or artifacts arising from small sample sizes, a collaborator and I compared COAD, LUAD, and LUSC transcriptomics using the TCGA dataset, where I observed similar trends. A limitation to this analysis, however, is that TCGA does not contain ctDNA data, so I can not reliably extrapolate from this biology to shedding behavior. A future large-scale study containing transcriptomic, genomic, and ctDNA data would be helpful for confirming the association and would greatly enrich the field of ctDNA research.

The size measurement I had access to during the analysis, the tumor's largest diameter, poses a limitation to the study as it might be of limited accuracy for non-spherical tumors. In a future study, it would be desirable to cross-reference this measure with imaging data in order to obtain more accurate tumor size measurements; nevertheless, the association I see between shedding behavior and size is significant. Furthermore, I hypothesize that the ctDNA negatives in our cohort might release ctDNA in quantities that fail to reach the limit of detection due to their small size and insufficient proliferative capacity. Summarizing the results from this study, I propose that CRC tumors, similarly to LUSCs, are ubiquitous ctDNA shedders; some of them might just appear ctDNA negative due to the tumor's insufficient shedding rate compounded by technological sensitivity challenges.

In conclusion, I analyzed the biology of ctDNA shedding in two distinctly different cancer types. As it is apparent from these studies, the contributors of ctDNA shedding are complex and vary on a per-cancer type, per-histology, or even per-patient basis, thus further research is needed to understand the extent of these mechanisms.

5.2 LONGITUDINAL DISEASE TRACKING USING CTDNA

The progression of metastatic melanoma is currently followed and evaluated via regularly scheduled CT-scans. This procedure is expensive, cumbersome, and exposes the patient to radiation; thus, it is performed relatively sparsely, often with months between scans. This delay in follow-up is detrimental to patient outcomes, given that immunotherapy benefits less than 50% of the patients yet is associated with severe side effects. In contrast, liquid biopsies offer a cost-effective and more timely addition to radiologic surveillance.

In **Manuscript III**, I analyzed longitudinal ctDNA data within the context of treatment response and aimed to decipher a mutational signal that may be used to identify patients who do not benefit from checkpoint-inhibition treatment. The plasma samples were sequenced using a tumor-agnostic, custom gene panel composed of 40 melanoma-specific genes. The benefit of this approach is two-fold: first, it allowed us to sequence our targeted set of genes at a high (15,000x) depth without the high price tag of whole exome or whole genome sequencing, and secondly, its tumor-agnostic nature allowed us to base our analysis on the blood, without needing to have access to tissue samples from the localized tumor. On the other hand, by using this highly curated approach, I cannot disregard the possibility that some genes of clinical interest were simply not included in the panel and thus been missed by our study; furthermore, sensitivity issues may stem from the lack of paired germline control or tumor biopsy samples. In the absence of controls, my in-house bioinformatic pipeline relied on independent samples as background for variant calling and utilized known SNP databases combined with a strict MAF cutoff threshold of 0.4 to filter out normal variants. I expect that, as a result of this setup, some relevant alterations may have remained undetected or were erroneously filtered out. In silico benchmarking revealed that the sensitivity threshold of my method, quantified as the median MAF of a significant variant, is 6.2%. This is high compared to the 1-5% LOD promised by the state-of-the-art targeted sequencing approaches [92], however, it could likely be improved by sequencing and analyzing the buffy coat in conjunction with the patient's plasma sample.

I found evidence that mutations in the *TERT* telomerase gene occur with higher frequency in patients who do not respond to immunotherapy treatment and those who carry a high metastatic load. This finding is in line with recent studies that implicate *TERT* with poor outcome [57]. I thus hypothesize that *TERT* might be a good candidate for a biomarker identifying aggressive disease and subpar response to therapy. Due to the limitations outlined above, as well as the limitations stemming from a small cohort size, this finding would need to be validated in a larger patient group. While ctDNA levels have been associated with disease burden in metastatic melanoma patients [182], I did not observe a correlation between baseline ctDNA MAF and therapy response in our study. I did find a significant association between baseline ctDNA MAF and metastatic load, indicating that these patients shed higher amounts of ctDNA into the bloodstream, an observation in line with previous work associating ctDNA levels with cancer cell burden [168]. Additionally, I found significant differences in baseline ctDNA detection between the three response groups, suggesting that resistant and acquired resistant patients shed ctDNA in detectable quantities more often than their responder counterparts. Taken together, these findings suggest that ctDNA may be a valuable tool for monitoring disease burden in the clinic, however, more comprehensive studies need to be conducted to confirm the extent of its utility. My longitudinal variant analysis did not yield any patterns or specific alterations that may provide insight into the emergence of acquired resistance. I hypothesize that the limitations of our study may have hindered my ability to identify a sensitive signal in this case.

In conclusion, I demonstrated that a tumor-agnostic ctDNA panel and corresponding data analysis may be used as a supplement to current standard-of-care methods in order to monitor disease evolution and stratify patients according to response. This observation is in line with previous published work demonstrating that serial ctDNA sampling and analysis, even in a tumor-agnostic manner, offers a strong prognostic and clinical value in a variety of cancer types [183–185]. In order to aid clinical adoption, it would be necessary to gain a deeper understanding of ctDNA dynamics in metastatic melanoma as it relates to outcomes and immunotherapy response; and verify my findings by utilizing a more sensitive ctDNA detection and variant calling method in a larger cohort.

6 CONCLUSIONS AND FUTURE PERSPECTIVES

This thesis investigated the biology and clinical utility of ctDNA release across a variety of cancer types and study designs.

First, I explored the biology behind ctDNA shedding in a lung adenocarcinoma and a colorectal cancer cohort. Here, I demonstrated that while LUADs exhibit a distinct, differential phenotype conducive to ctDNA shedding, ctDNA release in CRC appears to be a mechanistic process primarily governed by tumor size. This finding carries clinical relevance since ctDNA status in LUAD patients may necessitate urgent action such as change of treatment, whereas, in CRC patients, the presence of shedding alone may be a less informative marker. Nevertheless, the contrast between these two cancer types elucidates the need for gaining a deeper understanding of ctDNA biology, as it appears to be clear that the origins and associates of ctDNA release show cancer type-, histology-, and potentially, patient-level variations.

Additionally, I explored the utility of ctDNA for tracking cancer progression in a metastatic melanoma cohort. In this small and limited study, I demonstrated that ctDNA may be a valuable tool for following the evolution of the disease in relation to treatment response. In order to adopt liquid biopsies in the clinic for this purpose, more comprehensive, sensitive studies would need to be conducted to draw solid lines between longitudinal genomics and patient outcomes.

Overall, ctDNA holds immense promise for improving patient outcomes in the clinic, and particularly, for the future of personalized medicine. My work expands on the current understanding of ctDNA biology and offers a glimpse into the variations occurring between and within cancer types. ctDNA shedding is a multi-factor, complex process and there is still much left to explore - and I hope that my studies presented here may serve as a starting point for future work.

BIBLIOGRAPHY

[1] Camilla Mattiuzzi and Giuseppe Lippi. "Current Cancer Epidemiology." en. In: *J. Epidemiol. Glob. Health* 9.4 (Dec. 2019), pp. 217–222.

[2] *Global health estimates: Leading causes of death*. en. `https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death`. Accessed: 2023-5-9.

[3] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." en. In: *CA Cancer J. Clin.* 71.3 (May 2021), pp. 209–249.

[4] Valentin Titus Grigorean and Daniel Alin Cristian. "Cancer-Yesterday, Today, Tomorrow." en. In: *Medicina* 59.1 (Dec. 2022).

[5] Jeffrey M Peters and Frank J Gonzalez. "The Evolution of Carcinogenesis." en. In: *Toxicol. Sci.* 165.2 (Oct. 2018), pp. 272–276.

[6] Hiroshi Saeki and Keizo Sugimachi. "Carcinogenic risk factors." In: *Japan Med. Assoc. J.* 44.6 (2001), pp. 245–249.

[7] D Hanahan and R A Weinberg. "The hallmarks of cancer." en. In: *Cell* 100.1 (Jan. 2000), pp. 57–70.

[8] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation." en. In: *Cell* 144.5 (Mar. 2011), pp. 646–674.

[9] Douglas Hanahan. "Hallmarks of Cancer: New Dimensions." en. In: *Cancer Discov.* 12.1 (Jan. 2022), pp. 31–46.

[10] Shruti Morjaria. "Driver mutations in oncogenesis." en. In: *Int. J. Mol. Immuno Oncol.* 6.100 (May 2021), pp. 100–102.

[11] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. "A census of human cancer genes." en. In: *Nat. Rev. Cancer* 4.3 (Mar. 2004), pp. 177–183.

[12] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. "The Cancer Genome Atlas Pan-Cancer analysis project." en. In: *Nat. Genet.* 45.10 (Oct. 2013), pp. 1113–1120.

[13] International Cancer Genome Consortium et al. "International network of cancer genome projects." en. In: *Nature* 464.7291 (Apr. 2010), pp. 993–998.

[14] A J Levine, J Momand, and C A Finlay. "The p53 tumour suppressor gene." en. In: *Nature* 351.6326 (June 1991), pp. 453–456.

[15] Mathieu Chevallier, Maxime Borgeaud, Alfredo Addeo, and Alex Friedlaender. "Oncogenic driver mutations in non-small cell lung cancer: Past, present and future." en. In: *World J. Clin. Oncol.* 12.4 (Apr. 2021), pp. 217–237.

[16] Dongdong Huang et al. "Mutations of key driver genes in colorectal cancer progression and metastasis." en. In: *Cancer Metastasis Rev.* 37.1 (Mar. 2018), pp. 173–187.

[17] Eran Hodis et al. "A landscape of driver mutations in melanoma." en. In: *Cell* 150.2 (July 2012), pp. 251–263.

[18] Julian Huxley. "BIOLOGICAL ASPECTS OF CANCER." In: *Am. J. Med. Sci.* 238.2 (Aug. 1959), p. 256.

[19] Chris Bailey, James R M Black, James L Reading, Kevin Litchfield, Samra Turajlic, Nicholas McGranahan, Mariam Jamal-Hanjani, and Charles Swanton. "Tracking Cancer Evolution through the Disease Course." en. In: *Cancer Discov.* 11.4 (Apr. 2021), pp. 916–932.

[20] Johannes G Reiter, Ivana Bozic, Benjamin Allen, Krishnendu Chatterjee, and Martin A Nowak. "The effect of one additional driver mutation on tumor progression." en. In: *Evol. Appl.* 6.1 (Jan. 2013), pp. 34–45.

[21] Brett E Johnson et al. "Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma." en. In: *Science* 343.6167 (Jan. 2014), pp. 189–193.

[22] Roberto Vendramin, Kevin Litchfield, and Charles Swanton. "Cancer evolution: Darwin and beyond." en. In: *EMBO J.* 40.18 (Sept. 2021), e108389.

[23] Franck Raynaud, Marco Mina, Daniele Tavernari, and Giovanni Ciriello. "Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability." en. In: *PLoS Genet.* 14.9 (Sept. 2018), e1007669.

[24] Nicholas McGranahan and Charles Swanton. "Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future." en. In: *Cell* 168.4 (Feb. 2017), pp. 613–628.

[25] Andriy Marusyk, Michalina Janiszewska, and Kornelia Polyak. "Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance." en. In: *Cancer Cell* 37.4 (Apr. 2020), pp. 471–484.

[26] Nicholas McGranahan and Charles Swanton. “Biological and therapeutic impact of intratumor heterogeneity in cancer evolution.” en. In: *Cancer Cell* 27.1 (Jan. 2015), pp. 15–26.

[27] Scott L Carter et al. “Absolute quantification of somatic DNA alterations in human cancer.” en. In: *Nat. Biotechnol.* 30.5 (May 2012), pp. 413–421.

[28] Andrej Fischer, Ignacio Vázquez-García, Christopher J R Illingworth, and Ville Mustonen. “High-definition reconstruction of clonal composition in cancer.” en. In: *Cell Rep.* 7.5 (June 2014), pp. 1740–1752.

[29] Yuchao Jiang, Yu Qiu, Andy J Minn, and Nancy R Zhang. “Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.37 (Sept. 2016), E5528–37.

[30] Christopher A Miller et al. “SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution.” en. In: *PLoS Comput. Biol.* 10.8 (Aug. 2014), e1003665.

[31] Hassan Lemjabbar-Alaoui, Omer Ui Hassan, Yi-Wei Yang, and Petra Buchanan. “Lung cancer: Biology and treatment options.” en. In: *Biochim. Biophys. Acta* 1856.2 (Dec. 2015), pp. 189–210.

[32] Roy S Herbst, Daniel Morgensztern, and Chris Boshoff. “The biology and management of non-small cell lung cancer.” en. In: *Nature* 553.7689 (Jan. 2018), pp. 446–454.

[33] *Lung cancer survival rates*. en. https://www.cancer.org/cancer/types/lung-cancer/detection-diagnosis-staging/survival-rates.html. Accessed: 2023-5-12.

[34] Min Yuan, Li-Li Huang, Jian-Hua Chen, Jie Wu, and Qing Xu. “The emerging treatment landscape of targeted therapy in non-small-cell lung cancer.” en. In: *Signal Transduct Target Ther* 4 (Dec. 2019), p. 61.

[35] Sitanshu S Singh, Achyut Dahal, Leeza Shrestha, and Seetharama D Jois. “Genotype Driven Therapy for Non-Small Cell Lung Cancer: Resistance, Pan Inhibitors and Immunotherapy.” en. In: *Curr. Med. Chem.* 27.32 (2020), pp. 5274–5316.

[36] Nissim Hay and Nahum Sonenberg. “Upstream and downstream of mTOR.” en. In: *Genes Dev.* 18.16 (Aug. 2004), pp. 1926–1945.

[37] Hua Shen et al. “Alteration in Mir-21/PTEN expression modulates gefitinib resistance in non-small cell lung cancer.” en. In: *PLoS One* 9.7 (July 2014), e103305.

[38] Jill E Larsen and John D Minna. "Molecular biology of lung cancer: clinical implications." en. In: *Clin. Chest Med.* 32.4 (Dec. 2011), pp. 703–740.

[39] Hermann Brenner, Matthias Kloor, and Christian Peter Pox. "Colorectal cancer." en. In: *Lancet* 383.9927 (Apr. 2014), pp. 1490–1502.

[40] Anna Lewandowska, Grzegorz Rudzki, Tomasz Lewandowski, Aleksandra Stryjkowska-Góra, and Sławomir Rudzki. "Title: Risk Factors for the Diagnosis of Colorectal Cancer." en. In: *Cancer Control* 29 (2022), p. 10732748211056692.

[41] Prashanth Rawla, Tagore Sunkara, and Adam Barsouk. "Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors." en. In: *Prz Gastroenterol* 14.2 (Jan. 2019), pp. 89–103.

[42] Eric R Fearon. "Molecular genetics of colorectal cancer." en. In: *Annu. Rev. Pathol.* 6 (2011), pp. 479–507.

[43] E R Fearon and B Vogelstein. "A genetic model for colorectal tumorigenesis." en. In: *Cell* 61.5 (June 1990), pp. 759–767.

[44] K W Kinzler and B Vogelstein. "Lessons from hereditary colorectal cancer." en. In: *Cell* 87.2 (Oct. 1996), pp. 159–170.

[45] Kai Li, Haiqing Luo, Lianfang Huang, Hui Luo, and Xiao Zhu. "Microsatellite instability: a review of what the oncologist should know." en. In: *Cancer Cell Int.* 20 (Jan. 2020), p. 16.

[46] Justin Guinney et al. "The consensus molecular subtypes of colorectal cancer." en. In: *Nat. Med.* 21.11 (Nov. 2015), pp. 1350–1356.

[47] Jesper Bertram Bramsen et al. "Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer." en. In: *Cell Rep.* 19.6 (May 2017), pp. 1268–1280.

[48] Z Kozovska, V Gabrisova, and L Kucerova. "Malignant melanoma: diagnosis, treatment and cancer stem cells." en. In: *Neoplasma* 63.4 (2016), pp. 510–517.

[49] Jonathan B Heistein, Utkarsh Acharya, and Shiva Kumar R Mukkamalla. *Malignant Melanoma*. StatPearls Publishing, Jan. 2023.

[50] Roy Rabbie, Peter Ferguson, Christian Molina-Aguilar, David J Adams, and Carla D Robles-Espinoza. "Melanoma subtypes: genomic profiles, prognostic molecular markers and therapeutic possibilities." en. In: *J. Pathol.* 247.5 (Apr. 2019), pp. 539–551.

[51] Ye Liu, Xilan Zhang, Guoying Wang, and Xinchang Cui. "Triple Combination Therapy With PD-1/PD-L1, BRAF, and MEK Inhibitor for Stage III-IV Melanoma: A Systematic Review and Meta-Analysis." en. In: *Front. Oncol.* 11 (June 2021), p. 693655.

[52] Sarah A Weiss, Jedd D Wolchok, and Mario Sznol. "Immunotherapy of melanoma: Facts and hopes." en. In: *Clin. Cancer Res.* 25.17 (Sept. 2019), pp. 5191–5201.

[53] Maartje W Rohaan et al. "Tumor-Infiltrating Lymphocyte Therapy or Ipilimumab in Advanced Melanoma." In: *N. Engl. J. Med.* 387.23 (Dec. 2022), pp. 2113–2125.

[54] Anne Vest Soerensen, Eva Ellebaek, Lars Bastholt, Henrik Schmidt, Marco Donia, and Inge Marie Svane. "Improved Progression-Free Long-Term Survival of a Nation-Wide Patient Population with Metastatic Melanoma." en. In: *Cancers* 12.9 (Sept. 2020), p. 2591.

[55] Elin S Gray et al. "Circulating tumor DNA to monitor treatment response and detect acquired resistance in patients with metastatic melanoma." en. In: *Oncotarget* 6.39 (Dec. 2015), pp. 42008–42018.

[56] Eric Loo, Parisa Khalili, Karen Beuhler, Imran Siddiqi, and Mohammad A Vasef. "BRAF V600E Mutation Across Multiple Tumor Types: Correlation Between DNA-based Sequencing and Mutation-specific Immunohistochemistry." en. In: *Appl. Immunohistochem. Mol. Morphol.* 26.10 (2018), pp. 709–713.

[57] Sara Gandini et al. "TERT promoter mutations and melanoma survival: A comprehensive literature review and meta-analysis." en. In: *Crit. Rev. Oncol. Hematol.* 160 (Apr. 2021), p. 103288.

[58] Bert Vogelstein and Kenneth W Kinzler. "Cancer genes and the pathways they control." en. In: *Nat. Med.* 10.8 (Aug. 2004), pp. 789–799.

[59] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz Jr, and Kenneth W Kinzler. "Cancer genome landscapes." en. In: *Science* 339.6127 (Mar. 2013), pp. 1546–1558.

[60] Levi A Garraway and Eric S Lander. "Lessons from the cancer genome." en. In: *Cell* 153.1 (Mar. 2013), pp. 17–37.

[61] M A Harris et al. "The Gene Ontology (GO) database and informatics resource." en. In: *Nucleic Acids Res.* 32.Database issue (Jan. 2004), pp. D258–61.

[62] M Kanehisa and S Goto. "KEGG: kyoto encyclopedia of genes and genomes." en. In: *Nucleic Acids Res.* 28.1 (Jan. 2000), pp. 27–30.

[63] Marc Gillespie et al. “The reactome pathway knowledgebase 2022.” en. In: *Nucleic Acids Res.* 50.D1 (Jan. 2022), pp. D687–D692.

[64] Aravind Subramanian et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (Oct. 2005), pp. 15545–15550.

[65] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. “The Molecular Signatures Database (MSigDB) hallmark gene set collection.” en. In: *Cell Syst* 1.6 (Dec. 2015), pp. 417–425.

[66] David Croft et al. “Reactome: a database of reactions, pathways and biological processes.” en. In: *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D691–7.

[67] Francisco Sanchez-Vega et al. “Oncogenic Signaling Pathways in The Cancer Genome Atlas.” en. In: *Cell* 173.2 (Apr. 2018), 321–337.e10.

[68] Emily Crowley, Federica Di Nicolantonio, Fotios Loupakis, and Alberto Bardelli. “Liquid biopsy: monitoring cancer-genetics in the blood.” en. In: *Nat. Rev. Clin. Oncol.* 10.8 (Aug. 2013), pp. 472–484.

[69] Giulia Siravegna, Silvia Marsoni, Salvatore Siena, and Alberto Bardelli. “Integrating liquid biopsies into the management of cancer.” en. In: *Nat. Rev. Clin. Oncol.* 14.9 (Sept. 2017), pp. 531–548.

[70] Saife N Lone et al. “Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments.” en. In: *Mol. Cancer* 21.1 (Mar. 2022), p. 79.

[71] Debdatta Basu and Rajendra Kulkarni. “Overview of blood components and their preparation.” en. In: *Indian J. Anaesth.* 58.5 (Sept. 2014), pp. 529–537.

[72] Joscilin Mathew, Parvathy Sankar, and Matthew Varacallo. *Physiology, Blood Plasma*. StatPearls Publishing, Apr. 2022.

[73] Jin H Bae et al. “Single duplex DNA sequencing with CODEC detects mutations with high sensitivity.” en. In: *Nat. Genet.* 55.5 (May 2023), pp. 871–879.

[74] W Teetson, C Cartwright, B J Dreiling, and M H Steinberg. “The leukocyte composition of peripheral blood buffy coat.” en. In: *Am. J. Clin. Pathol.* 79.4 (Apr. 1983), pp. 500–501.

[75] M Böck, S Rahrig, D Kunz, G Lutze, and M U Heim. “Platelet concentrates derived from buffy coat and apheresis: biochemical and functional differences.” en. In: *Transfus. Med.* 12.5 (Oct. 2002), pp. 317–324.

[76] Laura Dean. *Blood and the cells it contains*. National Center for Biotechnology Information (US), 2005.

[77] David R Parkinson et al. "Considerations in the development of circulating tumor cell technology for clinical use." en. In: *J. Transl. Med.* 10 (July 2012), p. 138.

[78] Johann S de Bono, Howard I Scher, R Bruce Montgomery, Christopher Parker, M Craig Miller, Henk Tissing, Gerald V Doyle, Leon W W M Terstappen, Kenneth J Pienta, and Derek Raghavan. "Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer." en. In: *Clin. Cancer Res.* 14.19 (Oct. 2008), pp. 6302–6309.

[79] Jeffrey B Smerage et al. "Circulating tumor cells and response to chemotherapy in metastatic breast cancer: SWOG S0500." en. In: *J. Clin. Oncol.* 32.31 (Nov. 2014), pp. 3483–3489.

[80] Menno Tamminga, Sanne de Wit, T Jeroen N Hiltermann, Wim Timens, Ed Schuuring, Leon W M M Terstappen, and Harry J M Groen. "Circulating tumor cells in advanced non-small cell lung cancer patients are associated with worse tumor response to checkpoint inhibitors." en. In: *J Immunother Cancer* 7.1 (July 2019), p. 173.

[81] Raghu Kalluri. "The biology and function of exosomes in cancer." en. In: *J. Clin. Invest.* 126.4 (Apr. 2016), pp. 1208–1215.

[82] Gabriella Dobra et al. "Small Extracellular Vesicles Isolated from Serum May Serve as Signal-Enhancers for the Monitoring of CNS Tumors." en. In: *Int. J. Mol. Sci.* 21.15 (July 2020).

[83] Khalid Al-Nedawi, Brian Meehan, Johann Micallef, Vladimir Lhotak, Linda May, Abhijit Guha, and Janusz Rak. "Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells." en. In: *Nat. Cell Biol.* 10.5 (May 2008), pp. 619–624.

[84] P Mandel and P Metais. "Nuclear acids in human blood plasma." fr. In: *C. R. Seances Soc. Biol. Fil.* 142.3-4 (Feb. 1948), pp. 241–243.

[85] M Stroun, P Anker, P Maurice, J Lyautey, C Lederrey, and M Beljanski. "Neoplastic characteristics of the DNA found in the plasma of cancer patients." en. In: *Oncology* 46.5 (1989), pp. 318–322.

[86] Yanni Y N Lui, Ki-Wai Chik, Rossa W K Chiu, Cheong-Yip Ho, Christopher W K Lam, and Y M Dennis Lo. "Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation." en. In: *Clin. Chem.* 48.3 (Mar. 2002), pp. 421–427.

[87] Yan-Yan Yan, Qiao-Ru Guo, Feng-Hua Wang, Rameshwar Adhikari, Zhuang-Yan Zhu, Hai-Yan Zhang, Wen-Min Zhou, Hua Yu, Jing-Quan Li, and Jian-Ye Zhang. "Cell-Free DNA: Hope and Potential Application in Cancer." en. In: *Front Cell Dev Biol* 9 (Feb. 2021), p. 639233.

[88] S A Leon, B Shapiro, D M Sklaroff, and M J Yaros. "Free DNA in the serum of cancer patients and the effect of therapy." en. In: *Cancer Res.* 37.3 (Mar. 1977), pp. 646–650.

[89] David Sidransky. "Emerging molecular markers of cancer." en. In: *Nat. Rev. Cancer* 2.3 (Mar. 2002), pp. 210–219.

[90] Anne Jan van der Meer, Anna Kroeze, Arie J Hoogendijk, Aicha Ait Soussan, C Ellen van der Schoot, Walter A Wuillemin, Carlijn Voermans, Tom van der Poll, and Sacha Zeerleder. "Systemic inflammation induces release of cell-free DNA from hematopoietic and parenchymal cells in mice and humans." en. In: *Blood Adv* 3.5 (Mar. 2019), pp. 724–728.

[91] Ilaria Alborelli et al. "Cell-free DNA analysis in healthy individuals by next-generation sequencing: a proof of concept and technical validation study." en. In: *Cell Death Dis.* 10.7 (July 2019), p. 534.

[92] Jonathan C M Wan, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. "Liquid biopsies come of age: towards implementation of circulating tumour DNA." en. In: *Nat. Rev. Cancer* 17.4 (Feb. 2017), pp. 223–238.

[93] Mohammad Shahrokh Esfahani et al. "Inferring gene expression from cell-free DNA fragmentation profiles." en. In: *Nat. Biotechnol.* 40.4 (Apr. 2022), pp. 585–597.

[94] Laura Keller, Yassine Belloum, Harriet Wikman, and Klaus Pantel. "Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond." en. In: *Br. J. Cancer* 124.2 (Jan. 2021), pp. 345–358.

[95] Bhuvan Molparia, Eshaan Nichani, and Ali Torkamani. "Assessment of circulating copy number variant detection for cancer screening." en. In: *PLoS One* 12.7 (July 2017), e0180647.

[96] Elizabeth M Swisher, Melissa Wollan, Sarita M Mahtani, Julia B Willner, Rochelle Garcia, Barbara A Goff, and Mary-Claire King. "Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer." en. In: *Am. J. Obstet. Gynecol.* 193.3 Pt 1 (Sept. 2005), pp. 662–667.

[97] Frank Diehl et al. "Circulating mutant DNA to assess tumor dynamics." en. In: *Nat. Med.* 14.9 (Sept. 2008), pp. 985–990.

[98] Sarah-Jane Dawson et al. "Analysis of circulating tumor DNA to monitor metastatic breast cancer." en. In: *N. Engl. J. Med.* 368.13 (Mar. 2013), pp. 1199–1209.

[99] Hideharu Kimura, Kazuo Kasahara, Makoto Kawaishi, Hideo Kunitoh, Tomohide Tamura, Brian Holloway, and Kazuto Nishio. "Detection of Epidermal Growth Factor Receptor Mutations in Serum as a Predictor of the Response to Gefitinib in Patients with Non–Small-Cell Lung Cancer." en. In: *Clin. Cancer Res.* 12.13 (July 2006), pp. 3915–3921.

[100] Yanan Kuang, Andrew Rogers, Beow Y Yeap, Lilin Wang, Mike Makrigiorgos, Kristi Vetrand, Sara Thiede, Robert J Distel, and Pasi A Jänne. "Noninvasive Detection of EGFR T790M in Gefitinib or Erlotinib Resistant Non–Small Cell Lung Cancer." en. In: *Clin. Cancer Res.* 15.8 (Apr. 2009), pp. 2630–2636.

[101] Evan J Lipson, Victor E Velculescu, Theresa S Pritchard, Mark Sausen, Drew M Pardoll, Suzanne L Topalian, and Luis A Diaz. "Circulating tumor DNA analysis as a real-time method for monitoring tumor burden in melanoma patients undergoing treatment with immune checkpoint blockade." In: *Journal for ImmunoTherapy of Cancer* 2.1 (Dec. 2014), p. 42.

[102] O Gautschi et al. "Origin and prognostic value of circulating KRAS mutations in lung cancer patients." en. In: *Cancer Lett.* 254.2 (Sept. 2007), pp. 265–273.

[103] Chetan Bettegowda et al. "Detection of circulating tumor DNA in early- and late-stage human malignancies." en. In: *Sci. Transl. Med.* 6.224 (Feb. 2014), 224ra24.

[104] Christine A Parkinson et al. "Exploratory Analysis of TP53 Mutations in Circulating Tumour DNA as Biomarkers of Treatment Response for Patients with Relapsed High-Grade Serous Ovarian Carcinoma: A Retrospective Study." en. In: *PLoS Med.* 13.12 (Dec. 2016), e1002198.

[105] L Mao, R H Hruban, J O Boyle, M Tockman, and D Sidransky. "Detection of oncogene mutations in sputum precedes diagnosis of lung cancer." en. In: *Cancer Res.* 54.7 (Apr. 1994), pp. 1634–1637.

[106] Great Britain. Office for National Statistics, John Broggio, and Neil Bannister. *Cancer Survival by Stage at Diagnosis for England (experimental Statistics): Adults Diagnosed 2012, 2013 and 2014 and Followed Up to 2015 : Statistical Bulletin*. en. Office for National Statistics, 2016.

[107] Emmanuelle Gormally et al. "TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study." en. In: *Cancer Res.* 66.13 (July 2006), pp. 6871–6876.

[108] Nitzan Rosenfeld et al. "MicroRNAs accurately identify cancer tissue origin." en. In: *Nat. Biotechnol.* 26.4 (Apr. 2008), pp. 462–469.

[109] John A Olsen, Lauren A Kenna, Regine C Tipon, Michael G Spelios, Mark M Stecker, and Eitan M Akirav. "A Minimally-invasive Blood-derived Biomarker of Oligodendrocyte Cell-loss in Multiple Sclerosis." en. In: *EBioMedicine* 10 (Aug. 2016), pp. 227–235.

[110] Tim Forshew et al. "Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA." en. In: *Sci. Transl. Med.* 4.136 (May 2012), 136ra68.

[111] Antonio Marchetti et al. "Early Prediction of Response to Tyrosine Kinase Inhibitors by Quantification of EGFR Mutations in Plasma of NSCLC Patients." en. In: *J. Thorac. Oncol.* 10.10 (Oct. 2015), pp. 1437–1443.

[112] Jeanne Tie et al. "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer." en. In: *Sci. Transl. Med.* 8.346 (July 2016), 346ra92.

[113] Isaac Garcia-Murillas et al. "Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer." en. In: *Sci. Transl. Med.* 7.302 (Aug. 2015), 302ra133.

[114] Marco Gerlinger et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing." en. In: *N. Engl. J. Med.* 366.10 (Mar. 2012), pp. 883–892.

[115] M Jamal-Hanjani et al. "Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer." en. In: *Ann. Oncol.* 27.5 (May 2016), pp. 862–867.

[116] L De Mattos-Arruda et al. "Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle." en. In: *Ann. Oncol.* 25.9 (Sept. 2014), pp. 1729–1735.

[117] Sandra Misale et al. "Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer." en. In: *Nature* 486.7404 (June 2012), pp. 532–536.

[118] Kenneth S Thress et al. "Acquired EGFR C797S mutation mediates resistance to AZD9291 in non–small cell lung cancer harboring EGFR T790M." en. In: *Nat. Med.* 21.6 (May 2015), pp. 560–562.

[119] J S Frenel, S Carreira, J Goodall, D Roda, and others. "Serial Next-Generation Sequencing of Circulating Cell-Free DNA Evaluating Tumor Clone Response To Molecularly Targeted Drug AdministrationLiquid Biopsy and ... " In: *Clin. Cancer Res.* (2015).

[120] Pashtoon Murtaza Kasi et al. "Tumor-informed assessment of molecular residual disease and its incorporation into practice for patients with early and advanced-stage colorectal cancer (CRC-MRD Consortia)." In: *J. Clin. Orthod.* 38.15_suppl (May 2020), pp. 4108–4108.

[121] Shannon Zhang, Danielle Brazel, Priyanka Kumar, Liudmila N Schafer, Benjamin Eidenschink, Maheswari Senthil, and Farshid Dayyani. "Utility of tumor-informed circulating tumor DNA in the clinical management of gastrointestinal malignancies." en. In: *J. Gastrointest. Oncol.* 12.6 (Dec. 2021), pp. 2643–2652.

[122] *Invitae opens early access to liquid biopsy-based Personalized Cancer Monitoring as a central laboratory service*. en. https://ir.invitae.com/news-and-events/press-releases/press-release-details/2021/Invitae-opens-early-access-to-liquid-biopsy-based-Personalized-Cancer-Monitoring-as-a-central-laboratory-service/default.aspx. Accessed: 2023-5-17.

[123] Chris Abbosh et al. "Abstract CT023: Phylogenetic tracking and minimal residual disease detection using ctDNA in early-stage NSCLC: A lung TRACERx study." en. In: *Cancer Res.* 80.16_Supplement (Aug. 2020), CT023–CT023.

[124] Jo Cavallo. *CtDNA may be a prognostic biomarker of disease recurrence in patients with lung cancer - the ASCO post*. en. https://ascopost.com/news/april-2020/ctdna-may-be-a-prognostic-biomarker-of-disease-recurrence-in-patients-with-lung-cancer/. Accessed: 2023-5-17.

[125] Uttara Saran and Chendil Damodaran. "The application of RNA sequencing in precision cancer medicine." In: *Reference Module in Biomedical Sciences*. Elsevier, Jan. 2023.

[126] *Signatera – circulating tumor DNA blood test*. en. https://www.natera.com/oncology/signatera-advanced-cancer-detection/. Accessed: 2023-5-17. Dec. 2020.

[127] Alan P Venook. "Colorectal Cancer Surveillance With Circulating Tumor DNA Assay." en. In: *JAMA Netw Open* 5.3 (Mar. 2022), e221100.

[128] *Natera: A global leader in cell-free DNA testing*. en. http://natera.com. Accessed: 2023-5-18. May 2020.

[129] Christopher Abbosh et al. "Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution." en. In: *Nature* 545.7655 (Apr. 2017), pp. 446–451.

[130] Thomas Reinert et al. "Analysis of Plasma Cell-Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer." en. In: *JAMA Oncol* 5.8 (Aug. 2019), pp. 1124–1131.

[131] R C Coombes, K Page, R Salari, R K Hastings, and others. "Personalized Detection of Circulating Tumor DNA Antedates Breast Cancer Metastatic RecurrencePersonalized ctDNA Detection of Breast Cancer Recurrence." In: *Clin. Cancer Res.* (2019).

[132] Emil Christensen et al. "Abstract 913: Early detection of metastatic relapse and monitoring of therapeutic efficacy by ultra-deep sequencing of plasma cell-free DNA in patients with urothelial bladder carcinoma." en. In: *Cancer Res.* 79.13_Supplement (July 2019), pp. 913–913.

[133] Lilit Garibyan and Nidhi Avashia. "Polymerase chain reaction." en. In: *J. Invest. Dermatol.* 133.3 (Mar. 2013), pp. 1–4.

[134] K B Mullis. "The unusual origin of the polymerase chain reaction." en. In: *Sci. Am.* 262.4 (Apr. 1990), pp. 56–61, 64–5.

[135] Xiaodong Mao, Chao Liu, Hua Tong, Yajun Chen, and Kangsheng Liu. "Principles of digital PCR and its applications in current obstetrical and gynecological diseases." en. In: *Am. J. Transl. Res.* 11.12 (Dec. 2019), pp. 7209–7222.

[136] Vicky Rowlands, Andrzej J Rutkowski, Elena Meuser, T Hedley Carr, Elizabeth A Harrington, and J Carl Barrett. "Optimisation of robust singleplex and multiplex droplet digital PCR assays for high confidence mutation detection in circulating tumour DNA." en. In: *Sci. Rep.* 9.1 (Sept. 2019), p. 12620.

[137] Marisol Huerta et al. "Circulating Tumor DNA Detection by Digital-Droplet PCR in Pancreatic Ductal Adenocarcinoma: A Systematic Review." en. In: *Cancers* 13.5 (Feb. 2021).

[138] *QIAseq targeted DNA panels*. en. https://www.qiagen.com/us/products/discovery-and-translational-research/next-generation-sequencing/dna-sequencing/somatic-panels/qiaseq-targeted-dna-panels. Accessed: 2023-5-18.

[139] QIAGEN. *QIAseq Targeted DNA Panel Handbook*. https://www.qiagen.com/us/resources/resourcedetail?id=b95f7e26-90db-4565-ae50-15b64066d1a8&lang=en. Accessed: 2023-5-18.

[140] Wey Cheng Sim, Chet Hong Loh, Grace Li-Xian Toh, Chia Wei Lim, Akhil Chopra, Alex Yuan Chi Chang, and Liuh Ling Goh. "Non-invasive detection of actionable mutations in advanced non-small-cell lung cancer using targeted sequencing of circulating tumor DNA." en. In: *Lung Cancer* 124 (Oct. 2018), pp. 154–159.

[141] Ya-Sian Chang, Hsin-Yuan Fang, Yao-Ching Hung, Tao-Wei Ke, Chieh-Min Chang, Ting-Yuan Liu, Yu-Chia Chen, Dy-San Chao, Hsi-Yuan Huang, and Jan-Gowth Chang. "Correlation of genomic alterations between tumor tissue and circulating

tumor DNA by next-generation sequencing." en. In: *J. Cancer Res. Clin. Oncol.* 144.11 (Nov. 2018), pp. 2167–2175.

[142] Chang Xu, Mohammad R Nezami Ranjbar, Zhong Wu, John DiCarlo, and Yexun Wang. "Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller." en. In: *BMC Genomics* 18.1 (Jan. 2017), p. 5.

[143] N Guibert, Y Hu, N Feeney, Y Kuang, V Plagnol, G Jones, K Howarth, J F Beeler, C P Paweletz, and G R Oxnard. "Amplicon-based next-generation sequencing of plasma cell-free DNA for detection of driver and resistance mutations in advanced non-small cell lung cancer." en. In: *Ann. Oncol.* 29.4 (Apr. 2018), pp. 1049–1055.

[144] *Mutect2*. en. https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2. Accessed: 2023-5-27.

[145] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. "Calling Somatic SNVs and Indels with Mutect2." en. Dec. 2019.

[146] Sangtae Kim et al. "Strelka2: fast and accurate calling of germline and somatic variants." en. In: *Nat. Methods* 15.8 (Aug. 2018), pp. 591–594.

[147] Moritz Gerstung, Elli Papaemmanuil, and Peter J Campbell. "Subclonal variant calling with multiple samples and prior knowledge." en. In: *Bioinformatics* 30.9 (May 2014), pp. 1198–1204.

[148] Elizabeth D Lightbody, Ankit K Dutta, and Irene M Ghobrial. "MRDetect: An Ultrasensitive Solution to Monitor Low Tumor Burden in Liquid Biopsies." en. In: *The Hematologist* 17.5 (Aug. 2020).

[149] Jonathan C M Wan et al. "ctDNA monitoring using patient-specific sequencing and integration of variant reads." en. In: *Sci. Transl. Med.* 12.548 (June 2020).

[150] Adam J Widman et al. "Machine learning guided signal enrichment for ultrasensitive plasma tumor burden monitoring." en. Jan. 2022.

[151] Mikkel H Christensen et al. "DREAMS: deep read-level error model for sequencing data applied to low-frequency variant calling and circulating tumor DNA detection." en. In: *Genome Biol.* 24.1 (Apr. 2023), p. 99.

[152] Viktor A Adalsteinsson et al. "Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors." en. In: *Nat. Commun.* 8.1 (Nov. 2017), p. 1324.

[153] Christopher Abbosh et al. “Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA.” en. In: *Nature* (Apr. 2023), pp. 1–10.

[154] *Somatic calling is NOT simply a difference between two callsets*. en. https://gatk.broadinstitute.org/hc/en-us/articles/360035890491. Accessed: 2023-5-28.

[155] *Local re-assembly and haplotype determination (HaplotypeCaller and Mutect2)*. en. https://gatk.broadinstitute.org/hc/en-us/articles/360036227612-Local-re-assembly-and-haplotype-determination-HaplotypeCaller-and-Mutect2-. Accessed: 2023-5-28.

[156] *Evaluating the evidence for haplotypes and variant alleles (HaplotypeCaller and Mutect2)*. en. https://gatk.broadinstitute.org/hc/en-us/articles/360036227632-Evaluating-the-evidence-for-haplotypes-and-variant-alleles-HaplotypeCaller-and-Mutect2-. Accessed: 2023-5-28.

[157] *Somatic short variant discovery (SNVs + Indels)*. en. https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-. Accessed: 2023-5-28.

[158] Zixi Chen, Yuchen Yuan, Xiaoshi Chen, Jiayun Chen, Shudai Lin, Xingsong Li, and Hongli Du. “Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency.” en. In: *Sci. Rep.* 10.1 (Feb. 2020), p. 3501.

[159] Iñigo Martincorena et al. “Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin.” en. In: *Science* 348.6237 (May 2015), pp. 880–886.

[160] *DeepSNV*. en. https://bioconductor.org/packages/release/bioc/html/deepSNV.html. Accessed: 2023-5-28.

[161] Asaf Zviran et al. “Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring.” en. In: *Nat. Med.* 26.7 (July 2020), pp. 1114–1124.

[162] Aditya Deshpande, Trent Walradt, Ya Hu, Amnon Koren, and Marcin Imielinski. “Robust foreground detection in somatic copy number data.” en. Nov. 2019.

[163] Siddhartha Jaiswal et al. “Age-related clonal hematopoiesis associated with adverse outcomes.” en. In: *N. Engl. J. Med.* 371.26 (Dec. 2014), pp. 2488–2498.

[164] J Liu et al. “Biological background of the genomic variations of cf-DNA in healthy individuals.” en. In: *Ann. Oncol.* 30.3 (Mar. 2019), pp. 464–470.

[165] Miguel Alcaide et al. “Evaluating the quantity, quality and size distribution of cell-free DNA by multiplex droplet digital PCR.” en. In: *Sci. Rep.* 10.1 (July 2020), p. 12564.

[166] Florent Mouliere, Bruno Robert, Erika Arnau Peyrotte, Maguy Del Rio, Marc Ychou, Franck Molina, Celine Gongora, and Alain R Thierry. “High fragmentation characterizes tumour-derived circulating DNA.” en. In: *PLoS One* 6.9 (Sept. 2011), e23418.

[167] Havell Markus, Dineika Chandrananda, Elizabeth Moore, Florent Mouliere, James Morris, James D Brenton, Christopher G Smith, and Nitzan Rosenfeld. “Refined characterization of circulating tumor DNA through biological feature integration.” en. In: *Sci. Rep.* 12.1 (Feb. 2022), p. 1928.

[168] Stefano Avanzini, David M Kurtz, Jacob J Chabon, Everett J Moding, Sharon Seiko Hori, Sanjiv Sam Gambhir, Ash A Alizadeh, Maximilian Diehn, and Johannes G Reiter. “A mathematical model of ctDNA shedding predicts tumor detection size.” en. In: *Science Advances* 6.50 (Dec. 2020), eabc4308.

[169] Daniele Frisone, Alex Friedlaender, and Alfredo Addeo. “The Role and Impact of Minimal Residual Disease in NSCLC.” en. In: *Curr. Oncol. Rep.* 23.12 (Nov. 2021), p. 136.

[170] Mariam Jamal-Hanjani et al. “Tracking the Evolution of Non-Small-Cell Lung Cancer.” en. In: *N. Engl. J. Med.* 376.22 (June 2017), pp. 2109–2121.

[171] Dhruva Biswas et al. “A clonal expression biomarker associates with lung cancer mortality.” en. In: *Nat. Med.* 25.10 (Oct. 2019), pp. 1540–1548.

[172] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.” en. In: *Nat. Rev. Cancer* 18.11 (Nov. 2018), pp. 696–705.

[173] Filipe Martins et al. “Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance.” en. In: *Nat. Rev. Clin. Oncol.* 16.9 (May 2019), pp. 563–580.

[174] Jonas Kabel et al. “Impact of Whole Genome Doubling on Detection of Circulating Tumor DNA in Colorectal Cancer.” en. In: *Cancers* 15.4 (Feb. 2023).

[175] A Warth et al. “Tumour cell proliferation (Ki-67) in non-small cell lung cancer: a critical reappraisal of its prognostic role.” en. In: *Br. J. Cancer* 111.6 (Sept. 2014), pp. 1222–1229.

[176] Hongyue Dai et al. “A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients.” en. In: *Cancer Res.* 65.10 (May 2005), pp. 4059–4066.

[177] Jiawen Teng, Xingfeng Guo, and He Wang. "CCEPR is a novel clinical biomarker for prognosis and regulates cell proliferation through PCNA in osteosarcoma." en. In: *J. Cell. Biochem.* 120.8 (Aug. 2019), pp. 12796–12802.

[178] Fumihiro Tanaka, Kazuhiro Yanagihara, Yosuke Otake, Yozo Kawano, Ryo Miyahara, Kazumasa Takenaka, Hiromichi Katakura, Shinya Ishikawa, Harumi Ito, and Hiromi Wada. "Prognostic factors in resected pathologic (p-) stage IIIA-N2, non-small-cell lung cancer." en. In: *Ann. Surg. Oncol.* 11.6 (June 2004), pp. 612–618.

[179] Samuel F Bakhoum and Lewis C Cantley. "The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment." en. In: *Cell* 174.6 (Sept. 2018), pp. 1347–1360.

[180] Midhun Malla, Jonathan M Loree, Pashtoon Murtaza Kasi, and Aparna Raj Parikh. "Using Circulating Tumor DNA in Colorectal Cancer: Current and Evolving Practices." en. In: *J. Clin. Oncol.* 40.24 (Aug. 2022), pp. 2846–2857.

[181] Gabriella Taques Marczynski, Ana Carolina Laus, Mariana Bisarro Dos Reis, Rui Manuel Reis, and Vinicius de Lima Vazquez. "Circulating tumor DNA (ctDNA) detection is associated with shorter progression-free survival in advanced melanoma patients." en. In: *Sci. Rep.* 10.1 (Oct. 2020), p. 18682.

[182] Ashleigh C McEvoy et al. "Correlation between circulating tumour DNA and metabolic tumour burden in metastatic melanoma patients." en. In: *BMC Cancer* 18.1 (July 2018), p. 726.

[183] Nadia Øgaard, Thomas Reinert, Tenna V Henriksen, Amanda Frydendahl, Emilie Aagaard, Mai-Britt W Ørntoft, Marie Ø Larsen, Anders R Knudsen, Frank V Mortensen, and Claus L Andersen. "Tumour-agnostic circulating tumour DNA analysis for improved recurrence surveillance after resection of colorectal liver metastases: A prospective cohort study." In: *Eur. J. Cancer* 163 (Mar. 2022), pp. 163–176.

[184] Mingchao Xie et al. "586 Durvalumab (D) ± tremelimumab (T) + platinum-etoposide (EP) in extensive-stage small-cell lung cancer (ES–SCLC): RNA sequencing analysis to explore canonical markers of IO activity in CASPIAN." en. In: *J Immunother Cancer* 10.Suppl 2 (Nov. 2022).

[185] Hiu Ting Chan, Satoshi Nagayama, Masumi Otaki, Yoon Ming Chin, Yosuke Fukunaga, Masashi Ueno, Yusuke Nakamura, and Siew-Kee Low. "Tumor-informed or tumor-agnostic circulating tumor DNA as a biomarker for risk of recurrence in resected colorectal cancer patients." en. In: *Front. Oncol.* 12 (2022), p. 1055968.

Part III

MANUSCRIPTS

Part IV

MANUSCRIPT I

Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA

Article

Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA

https://doi.org/10.1038/s41586-023-05776-4

Received: 6 April 2022

Accepted: 30 January 2023

Published online: 13 April 2023

Check for updates

Christopher Abbosh[1,101]✉, Alexander M. Frankell[1,2,101], Thomas Harrison[3,101], Judit Kisistok[4,5,6,101], Aaron Garnett[3,101], Laura Johnson[3], Selvaraju Veeriah[1], Mike Moreau[3], Adrian Chesh[3], Tafadzwa L. Chaunzwa[7,8], Jakob Weiss[7,8,9], Morgan R. Schroeder[3], Sophia Ward[1,2,10], Kristiana Grigoriadis[1,2,11], Aamir Shahpurwalla[3], Kevin Litchfield[1,12], Clare Puttick[1,2,11], Dhruva Biswas[1,2,13], Takahiro Karasaki[1,2,14], James R. M. Black[1,11], Carlos Martínez-Ruiz[1,11], Maise Al Bakir[1,2], Oriol Pich[2], Thomas B. K. Watkins[2], Emilia L. Lim[1,2], Ariana Huebner[1,2,11], David A. Moore[1,2,15], Nadia Godin-Heymann[16], Anne L'Hernault[16], Hannah Bye[16], Aaron Odell[3], Paula Roberts[3], Fabio Gomes[17], Akshay J. Patel[18], Elizabeth Manzano[1], Crispin T. Hiley[1,2], Nicolas Carey[19], Joan Riley[19], Daniel E. Cook[2], Darren Hodgson[16], Daniel Stetson[20], J. Carl Barrett[20], Roderik M. Kortlever[21], Gerard I. Evan[21], Allan Hackshaw[22], Robert D. Daber[3], Jacqui A. Shaw[19], Hugo J. W. L. Aerts[7,8,23], Abel Licon[3], Josh Stahl[3], Mariam Jamal-Hanjani[1,14,24], TRACERx Consortium*, Nicolai J. Birkbak[1,2,4,5,6,102], Nicholas McGranahan[1,11,102]✉ & Charles Swanton[1,2,24,102]✉

Circulating tumour DNA (ctDNA) can be used to detect and profile residual tumour cells persisting after curative intent therapy[1]. The study of large patient cohorts incorporating longitudinal plasma sampling and extended follow-up is required to determine the role of ctDNA as a phylogenetic biomarker of relapse in early-stage non-small-cell lung cancer (NSCLC). Here we developed ctDNA methods tracking a median of 200 mutations identified in resected NSCLC tissue across 1,069 plasma samples collected from 197 patients enrolled in the TRACERx study[2]. A lack of preoperative ctDNA detection distinguished biologically indolent lung adenocarcinoma with good clinical outcome. Postoperative plasma analyses were interpreted within the context of standard-of-care radiological surveillance and administration of cytotoxic adjuvant therapy. Landmark analyses of plasma samples collected within 120 days after surgery revealed ctDNA detection in 25% of patients, including 49% of all patients who experienced clinical relapse; 3 to 6 monthly ctDNA surveillance identified impending disease relapse in an additional 20% of landmark-negative patients. We developed a bioinformatic tool (ECLIPSE) for non-invasive tracking of subclonal architecture at low ctDNA levels. ECLIPSE identified patients with polyclonal metastatic dissemination, which was associated with a poor clinical outcome. By measuring subclone cancer cell fractions in preoperative plasma, we found that subclones seeding future metastases were significantly more expanded compared with non-metastatic subclones. Our findings will support (neo)adjuvant trial advances and provide insights into the process of metastatic dissemination using low-ctDNA-level liquid biopsy.

ctDNA is a multifaceted biomarker; presurgical ctDNA levels reflect relapse risk in early-stage NSCLC[3–5] and postoperative ctDNA detection highlights impending NSCLC recurrence[4–9]. Potential exists for postoperative ctDNA to guide the administration of adjuvant therapy[10,11]. Longitudinal measurements of clonal composition across metastatic sites can also be captured using ctDNA[7,12–14]. Within the TRACERx study[2], patients undergoing resection of NSCLC are managed according to National Institute of Clinical Excellence approved care pathways[15] and are followed for 5 years after surgery. Plasma is collected preoperatively and at three-monthly postoperative intervals during the first two years, followed by six-monthly intervals between years three and five. Previously, we demonstrated that 13 out of 14 patients with NSCLC recurrence had detectable postoperative ctDNA, which could provide insights into the clonal structure of residual disease[7]. Here we analysed 1,069 plasma samples from 197 patients with a median follow-up of 4.6 years in event-free patients. We implemented phylogenetic tracking technologies, including patient-specific anchored-multiplex PCR (AMP)[16] and cell-free DNA (cfDNA) enrichment tracking a median of 200 tumour mutations, combined with an informatic tool (ECLIPSE) to extract clonal composition in the context of the low ctDNA levels (<1%) encountered in early-stage NSCLC[17]. We address the prognostic implications of preoperative ctDNA detection, alongside postoperative

A list of affiliations appears at the end of the paper.

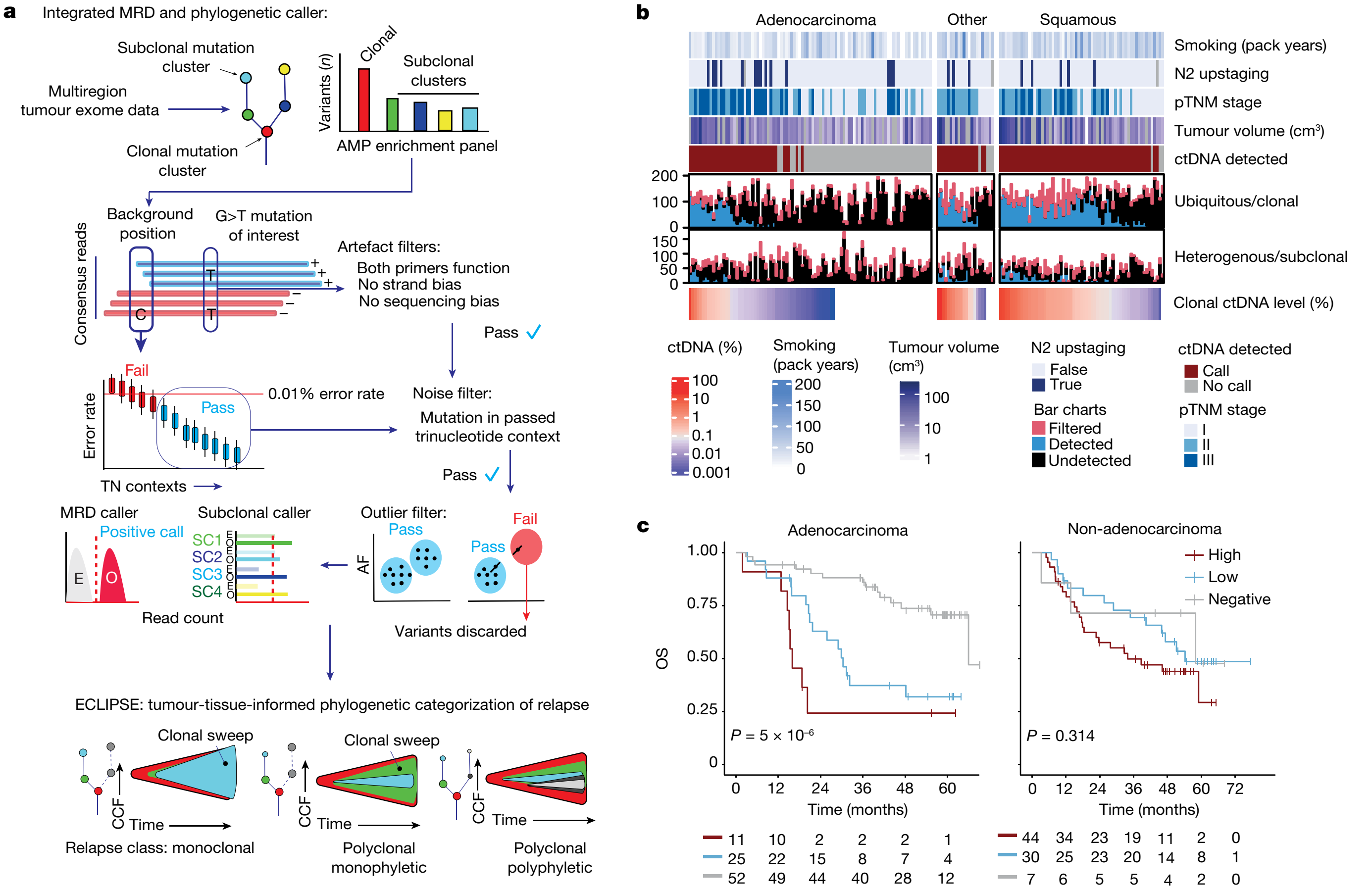


Fig. 1 | Overview of the cohort and ctDNA calling. a, The ctDNA detection method estimates intralibrary, trinucleotide-specific sequencing error rates. For calling ctDNA, the number of consensus reads at all positions targeted by a PSP that pass the described filters are compared with the expected error rates. To detect subclones, ECLIPSE evaluates the collective signal across all mutations in each subclone and integrates this with primary-tumour-derived copy-number information to estimate plasma CCFs (the percentage of tumour cells corresponding to clone), clonal sweeps (where a subclone reaches 100% CCF) and metastatic dissemination patterns. The ctDNA analysis approach is described further in the Supplementary Note. AF, allele frequency; E, expected signal; O, observed signal; SC, subclone. **b**, The clinical features associated with preoperative ctDNA analyses in patients in the non-pilot TRACERx study (with non-synchronous primary tumours). N2 upstaging row: patients clinically staged with N0/1 lymph node involvement upstaged to N2 disease by pathology. Grey, no mediastinal lymph nodes sampled at surgery. pTNM-stage row: pathological tumour node metastasis (v7) stage. Volumetrics row: $\log_{10}$-transformed tumour volume (in cm^3) measured using CT. Grey, unevaluable. The bar charts show mutations tracked by a patient's PSP categorized by clonality. Black, mutation undetected (per-variant one-sided Poisson $P > 0.01$; Methods); red, mutation filtered by MRD caller; blue, mutation detected. Clonal ctDNA level: the $\log_{10}$-transformed mean percentage of mutant consensus reads across all clonally mutated positions tracked by a PSP (Methods); patients with 0% level are given a white colour, a non-zero clonal ctDNA level can occur in ctDNA-negative patients for whom the signal was insufficient to result in confident detection of ctDNA. **c**, Kaplan–Meier curves demonstrating the overall survival outcomes in ctDNA-high (dark red), ctDNA-low (blue) and ctDNA-negative (grey) patients with non-synchronous adenocarcinoma (left) and non-synchronous non-adenocarcinoma (right). ctDNA high and low was categorized on the basis of median clonal ctDNA levels across all ctDNA-positive NSCLCs (0.16%). *P* values were calculated using log-rank tests.

ctDNA detection as an indicator of both impending disease relapse and phylogenetic pattern of metastatic dissemination.

ctDNA detection using AMP

AMP patient-specific cfDNA enrichment panels (PSPs) targeted a median of 200 mutations preidentified in multiregion exome analyses of early-stage NSCLC surgical resection samples (range, 72–201). A median of 126 clonal mutations were tracked, enabling sensitive identification of ctDNA; a median of 64 subclonal mutations (representing a median of 88% of subclones sampled in surgical tissue) were tracked to infer subclonal evolution at relapse (Fig. 1a, Extended Data Fig. 1a,b and Supplementary Table 1). The median cfDNA input into the AMP assay was 23 ng (interquartile range, 15–37 ng; Extended Data Fig. 1c and Supplementary Table 2). A molecular residual disease (MRD) detection algorithm evaluated background (non-variant) sequencing positions to estimate library error rates to enable ctDNA detection (Methods, Fig. 1a, Extended Data Fig. 1d–h and Supplementary Note). An MRD algorithm *P* value of 0.01 was determined to be optimal through analyses of a ten-patient pilot cohort (Supplementary Note and Extended Data Fig. 2a–d). The patients in the pilot cohort were excluded from subsequent ctDNA analyses, apart from ECLIPSE examination of subclonal kinetics (Methods).

Analytical and orthogonal validation of variant DNA detection using the locked-assay was performed (Supplementary Note). A total of 659 spike-in samples was analysed at assay DNA inputs of 2 ng to 80 ng and variant DNA levels of 0.003% to 0.1% (Methods). Sensitivity for variant DNA detection using a 50-variant PSP at 0.01% variant DNA level (representative of ctDNA levels encountered after resection of NSCLC, using current MRD assays[8]) was higher than 90% at DNA inputs of 20 ng and above. Below 20 ng input, sensitivity for 0.01% variant DNA declined. Specificity was 100% in analyses of 48 samples from healthy participants (Extended Data Fig. 2e–h and Supplementary Table 3). Orthogonal validation of preoperative ctDNA-positive calls was performed using digital droplet PCR (Extended Data Fig. 2i,j and Supplementary Table 4). Tracking more than 50 mutations improved the assay sensitivity at lower DNA inputs (Extended Data Fig. 2k and Supplementary Tables 5 and 6).

Features of preoperative ctDNA detection

Preoperative cfDNA was analysed across 187 patients in the TRACERx study (Supplementary Tables 7 and 8 and Extended Data Fig. 3a). A total of 178 patients had a single primary NSCLC (Fig. 1b) and 9 patients had synchronous primary NSCLCs at diagnosis (Extended Data Fig. 3b and Supplementary Note). In agreement with previous findings[3,7], higher rates of preoperative ctDNA detection in non-adenocarcinoma histologies compared with lung adenocarcinoma were observed (39 out of 93 lung adenocarcinomas were ctDNA positive versus 78 out of 85 non-adenocarcinomas; Fig. 1b). Patients exhibiting preoperative ctDNA detection had a higher smoking pack-year history (Wilcoxon rank sum test, $P = 0.023$; Fig. 1b and Extended Data Fig. 3c). Preoperative ctDNA detection was associated with clinically occult mediastinal lymph node disease in patients with adenocarcinoma. In total, 81 adenocarcinomas were clinical N0/1 stage and, after pathological nodal staging performed in 80 out of 81 patients, 14 out of 80 were upstaged to pN2 status. A total of 11 out of 14 (79%) patients who were pN2 upstaged were ctDNA positive versus 19 out of 66 (29%) patients who were not upstaged (χ^2 test, $P = 0.001$; Fig. 1b). Thus, preoperative ctDNA detection could guide mediastinal resection strategies in adenocarcinoma.

Preoperative ctDNA and clinical outcome

Given the variation in ctDNA detection across NSCLC subtypes, we assessed preoperative ctDNA status (negative (absent detection) or low or high (classified on the basis of clonal ctDNA level, the mean percentage of mutant consensus reads across clonally mutated positions tracked by a PSP)) as a prognostic biomarker separately in patients with single (non-synchronous) adenocarcinomas ($n = 88$) and single non-adenocarcinomas ($n = 81$) evaluable for survival analyses (Methods). In patients with adenocarcinoma, ctDNA status was associated with overall survival (OS) (log-rank test, $P = 5 \times 10^{-6}$; Fig. 1c). The 52 out of 88 (59%) patients with adenocarcinoma who were preoperative ctDNA negative had superior OS outcomes (90% 2 year OS, 95% confidence interval (CI) = 82–99%) compared with ctDNA low (63% 2 year OS, 95% CI = 46–85%, $n = 25$) or high adenocarcinoma (24% 2 year OS, 95% CI = 8–74%, $n = 11$). In non-adenocarcinoma, 7 out of 81 patients negative for ctDNA had OS outcomes that were indistinguishable from patients with low or high ctDNA and ctDNA status was not strongly prognostic (log-rank test, $P = 0.314$; Fig. 1c; when the seven patients negative for ctDNA were excluded, log-rank test, $P = 0.2$). Similar findings were observed in freedom from recurrence (FFR) analyses (Extended Data Fig. 3d). In multivariable survival analyses including pathological TNM (pTNM) stage, adjuvant therapy status, age and unique sequencing coverage, preoperative ctDNA status was associated with FFR and OS in adenocarcinoma but not in non-adenocarcinoma (Extended Data Fig. 3e). This supports prior findings that preoperative ctDNA status is a robust prognostic factor for RFS in adenocarcinoma, but not squamous cell carcinoma[5]. In patients with adenocarcinoma, preoperative ctDNA detection was associated with extrathoracic metastasis and poor post-recurrence outcomes. A total of 18 out of 20 (90%) recurrences involving extrathoracic sites occurred in patients who were preoperative ctDNA positive, compared with 8 out of 18 (44%) intrathoracic-only recurrences (χ^2 test, $P = 0.008$); post-recurrence survival was shorter in those who were preoperative ctDNA positive relative to those who were preoperative ctDNA negative (log-rank test, $P = 0.003$, Extended Data Fig. 3f,g).

Biology of ctDNA detection

Computed tomography (CT) volumetric data were available for 150 out of 178 patients with non-synchronous NSCLC (Extended Data Fig. 4a and Supplementary Table 8). In NSCLC, 10 cm^3 tumour volume has been associated with ctDNA levels of around 0.1% (ref. 7,18; a level detectable by AMP; Extended Data Fig. 2). A total of 17 out of 42 (41%) patients with adenocarcinoma and tumour volumes of ≥10 cm^3 were preoperative ctDNA negative, compared with only 2 out of 50 (4%) patients with non-adenocarcinoma (χ^2 test, $P < 0.001$; Extended Data Fig. 4b). The relative absence of ctDNA detection in higher-volume adenocarcinomas suggested a low-ctDNA shedding phenotype. We developed a regression model in 96 preoperative ctDNA positive cases to estimate clonal ctDNA levels on the basis of tumour histology and volume (Methods and Extended Data Fig. 4c). We then estimated clonal ctDNA levels in the 47 ctDNA-negative adenocarcinomas, categorizing these tumours as low-shedders (ctDNA detection expected on the basis of tumour volume, but not observed (31 out of 47 cases)) or technical negatives (tumour volume predicted for ctDNA levels below the sample limit of detection (16 out of 47 cases); Methods and Extended Data Fig. 4d). The latter group was excluded from analyses of ctDNA detection and tumour biology.

Available multiregion transcriptomic data enabled the comparison of 34 ctDNA-positive adenocarcinomas with 28 low-shedder adenocarcinomas (Fig. 2a and Supplementary Tables 9 and 10). Genes upregulated in ctDNA-positive adenocarcinomas included those associated with M phase, cell cycle and DNA repair (Supplementary table 11); and gene set variation analysis (GSVA[19]) using the Hallmark gene sets (which summarize 50 biological states[20]) revealed upregulation of proliferation and cell-cycle-associated gene sets (Fig. 2b–d). We evaluated our published prognostic biomarker associated with outcomes in lung adenocarcinoma (ORACLE[21]). Preoperative ctDNA-positive adenocarcinomas demonstrated higher ORACLE scores relative to negative adenocarcinomas ($P = 0.000134$; Fig. 2e). We observed no difference between ctDNA-positive adenocarcinoma and low-shedders when we analysed tumour purity and subclonal and clonal somatic driver mutations, individually and summarized to pathways (Extended Data Fig. 4e–g). We observed that ctDNA-positive adenocarcinomas showed increased levels of both weighted genome integrity index (wGII)[22] and fraction of loss of heterozygosity (FLOH)[23] relative to low-shedders ($P = 0.0286$ and $P = 0.00443$) and an increased percentage of ctDNA-positive adenocarcinomas were subject to whole-genome doubling (WGD; any WGD compared to none, 86% versus 61%, $P = 0.0400$; Fig. 2f,g and Extended Data Fig. 4h). We used GISTIC2.0[24] to assess whether the increased levels of chromosomal alterations in ctDNA-positive tumours were linked to the observed increase in cell proliferation (Methods). We identified 20 amplified cytobands enriched in ctDNA shedders (false-discovery rate (FDR) $q < 0.05$) with a GISTIC score difference (GSD) of at least 0.5 (Fig. 2h,i), a previously defined threshold for comparing two sample sets[25]. A total of 966 genes are located within these cytobands, of which 21 are listed in the COSMIC cancer gene census[26] as cancer genes (Supplementary table 12), including proliferation-associated genes *CCND1* (11q13.3), *CDK4* (12q14.1), *MDM2* (12q15) and *CCNE1* (19q12). These results were largely recapitulated when excluding the bottom quartile of tumour volumes from low-shedding adenocarcinomas (Supplementary Note and Extended Data Fig. 4i–k), indicating that tumour biology is probably the main driver behind our observations.

Postoperative ctDNA detection without relapse

Postoperative cfDNA samples from 42 recurrence-free patients and 19 patients who subsequently developed new primary cancers during follow-up (on the basis of histological or clinical findings) were analysed to assess AMP clinical specificity (PSPs are specific to the excised NSCLC and are not expected to detect new primary cancers; Fig. 3a,b and Supplementary Table 13). In total, 10 out of 426 (2%) postoperative

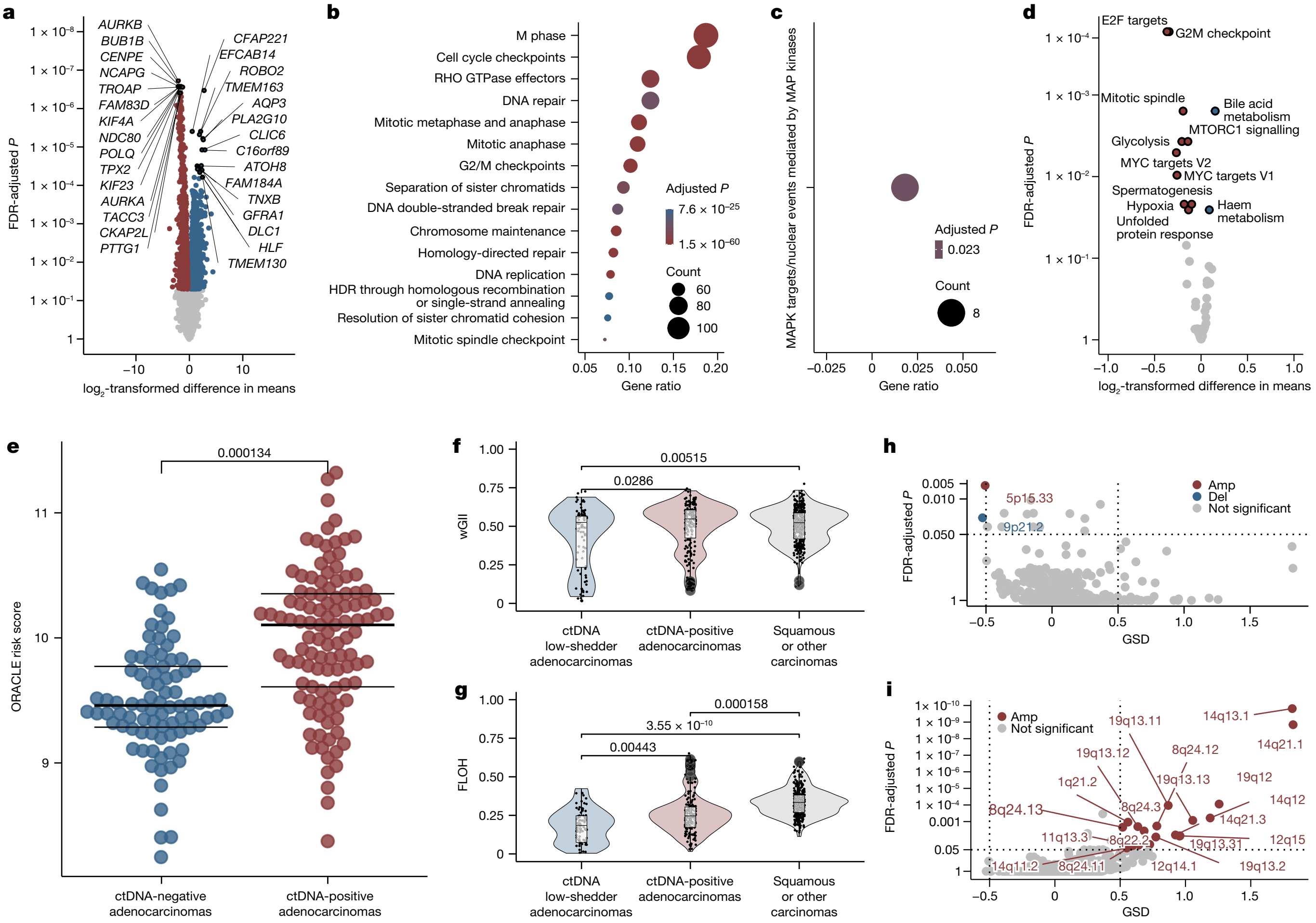


Fig. 2 | Genomic and transcriptomic predictors of ctDNA detection in early-stage NSCLC. a, Differential gene expression analysis comparing 34 ctDNA-positive adenocarcinomas (101 regions) with 28 ctDNA low-shedder adenocarcinomas (62 regions). The *x* axis shows the $\log_2$-transformed difference in mean value; the *y* axis shows two-sided FDR-adjusted *P* values. Statistical testing was performed by computing moderated *t*-statistics from a linear model fit to the transformed expression data (Methods). Genes significantly overexpressed in ctDNA-positive adenocarcinomas and ctDNA low-shedder adenocarcinomas (technical non-shedders excluded) are shown in red and blue, respectively. The top 15 genes are labelled per detection category. **b**,**c**, Reactome pathway enrichment analysis on the basis of the 1,759 significant genes found in **a**. The *y* axis lists pathways and the *x* axis shows the proportion of genes involved. **b**, The top 15 pathways in ctDNA-positive adenocarcinomas. HDR, homology-directed repair. **c**, The only significantly enriched pathway in ctDNA low-shedders. Size indicates gene count; colour indicates the one-sided hypergeometric *P* value. **d**, Differential enrichment analysis based on the Hallmark gene sets. Samples, axes and colours are as described in **a**. **e**, ORACLE gene expression scores in ctDNA-positive (35 patients, 109 regions) versus ctDNA-negative (42 patients, 87 regions) adenocarcinomas. The centre lines show the median values. Colours are as described in **a**. **f**,**g**, wGII (**f**) and FLOH (**g**) levels of ctDNA-positive adenocarcinomas (35 patients, 166 regions), ctDNA low-shedder adenocarcinomas (28 patients, 79 regions) and squamous or other carcinomas (74 patients, 303 regions). The hinges correspond to first and third quartiles, the whiskers extend to the largest/smallest value no further than 1.5× the interquartile range, and the centre lines represent the median values. **h**,**i**, GISTIC score analysis comparing 35 ctDNA-positive adenocarcinomas (166 regions; **i**) and 28 ctDNA low-shedders (79 regions; **h**). Red, amplifications (amp); blue, deletions (del); grey, non-significant values. The *y* axis shows the one-sided *P* values computed using GISTIC 2.0 permutation-based statistical methods, and the *x* axis shows the GSD. The dotted lines show the *G*-score and significance cut-offs. Pairwise comparisons were performed using linear mixed-effects models; *P* values are two-sided.

samples from 3 out of 61 (5%) of patients exhibited ctDNA detection (Fig. 3a,b). Patient CRUK0086 was ctDNA positive before radiation therapy, CRUK0269 was ctDNA positive after surgery and developed a new primary NSCLC, and CRUK0498 had false-positive ctDNA detection at 7 out of 8 postoperative timepoints, probably due to PSP mistargeting of somatic mutations associated with a lymphoid aggregate present in primary tumour tissue (Supplementary Note and Extended Data Fig. 5a–e).

Postoperative ctDNA detection and relapse

In total, 365 postoperative plasma samples were analysed from 70 patients with either recurrence of their NSCLC (*n* = 66) or incomplete resection (macroscopic residual disease, *n* = 4; Fig. 3c–e and Supplementary Table 13). ctDNA was detected postoperatively (before or after relapse) in 59 out of 70 (84%) of these patients. A total of 3 out of 11 patients relapsing without postoperative ctDNA detection lacked plasma sampling within 100 days of clinical relapse (CRUK0303, CRUK0495 and CRUK0570). In those with plasma sampled close to relapse, 2 out of 11 patients had unresected hilar or mediastinal lymph node metastases on postoperative imaging (CRUK0230 and CRUK0234), 4 out of 11 had intracranial recurrence (CRUK0331, CRUK0407, CRUK0567 and CRUK0736) and 2 out of 11 had intrathoracic recurrence (CRUK0329 and CRUK0490; Fig. 3c–e and Supplementary Table 14). Intracranial recurrence has previously been associated with absent postoperative ctDNA detection[9,27]. Here,

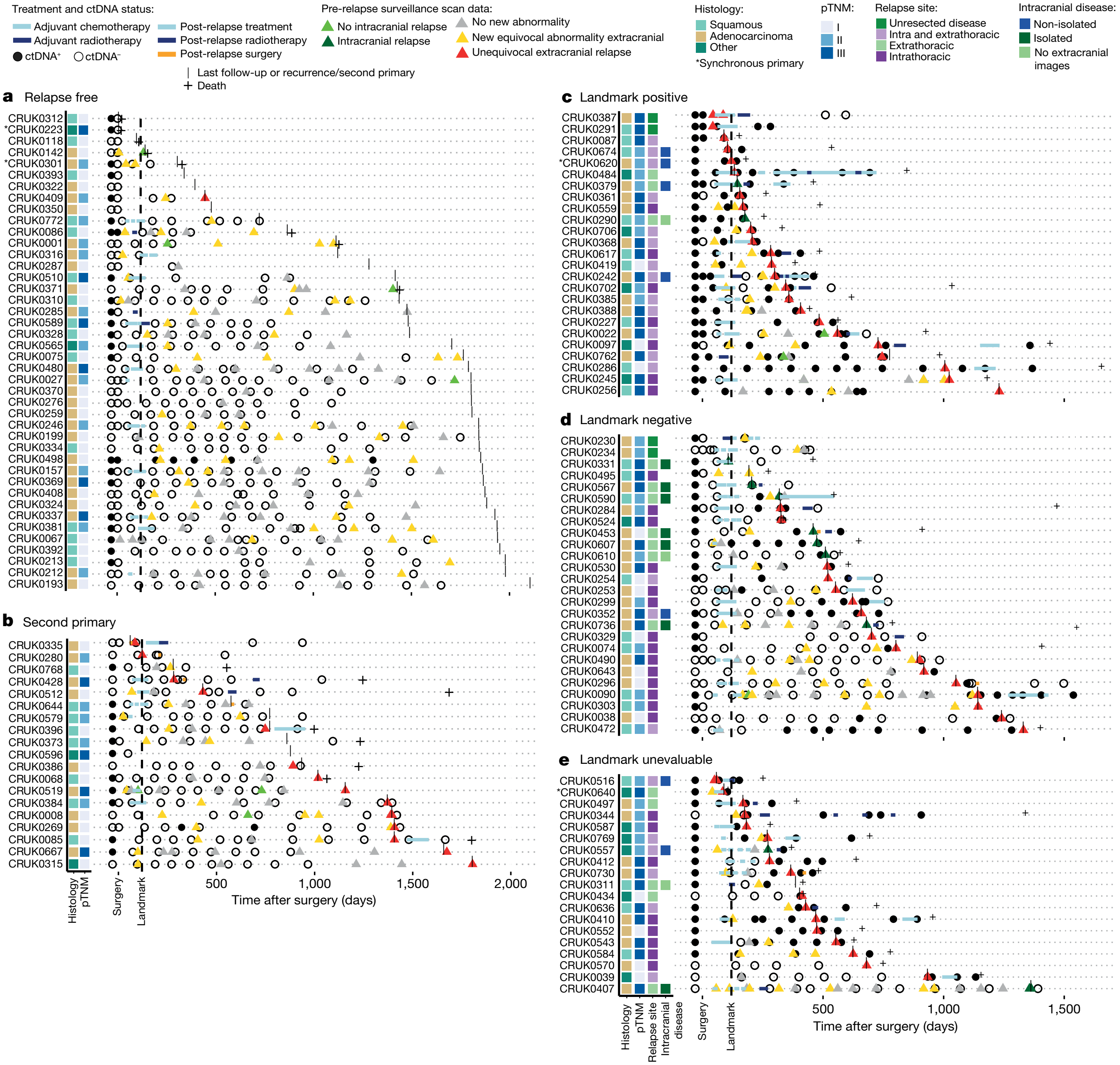


Fig. 3 | Postoperative minimal residual disease detection in early-stage NSCLC. a–e, Longitudinal ctDNA data from non-pilot patients with no evidence of NSCLC recurrence (**a**; $n = 42$); development of a second primary cancer (**b**; $n = 19$); recurrence of NSCLC in landmark-positive patients (**c**; $n = 25$ patients); recurrence of NSCLC in landmark-negative patients (**d**; $n = 26$ patients); and recurrence of NSCLC in landmark-unevaluable patients (**e**; $n = 19$ patients). For **a–e**, each circle represents a cfDNA sampling timepoint. Circles to the left of the surgical day are preoperative timepoints, circles to the right of the surgical day are postoperative timepoints. Black filled circle, positive ctDNA detection; light blue rectangles, chemotherapy; dark blue rectangles, radiotherapy; orange rectangles, patient received post-recurrence surgery. The triangles represent standard of care postoperative CT, PET or MRI surveillance imaging (imaging up until first relapse or last follow-up displayed on plot). Imaging was classified as no disease (grey), equivocal images (yellow) or unequivocal imaging evidence of extracranial relapse (red). Light green triangles, no evidence of intracranial relapse; dark green triangles, intracranial relapse. The vertical black lines show the event date for a patient (if death, second primary, NSCLC recurrence occurred); otherwise, the vertical line represents last TRACERx follow-up. Crosses indicate patient death events. The annotation plots on the left highlight histology, pTNM (pathological TNM) status, relapse site, and details regarding whether an intracranial relapse was isolated (brain-only) or non-isolated (brain and extracranial site) or occurred without extracranial imaging to confirm solitary status. For the relapse site and intracranial disease annotations, anatomical sites of disease were identified within a 180 day post-recurrence period.

17 patients experienced brain metastases within 180 days of relapse and 14 out of 17 patients also had extracranial imaging at relapse. Of these 14 patients, 3 out of 7 patients with isolated (brain-only) intracranial relapse versus 7 out of 7 with non-isolated intracranial relapse exhibited postoperative ctDNA detection (Fig. 3c–e and Extended Data Fig. 6a).

Landmark MRD analysis

We explored postoperative ctDNA detection within a landmark analysis framework, where the landmark timepoint refers to establishing the ctDNA status of a patient within an 120 day period following completion of surgery[1,6] (Fig. 3a–d). Here, 108 out of 131 patients with postoperative plasma sampling performed were evaluable for landmark analysis based on ≥1 plasma sample obtained within 120 days of surgery, before adjuvant therapy or relapse (Supplementary Table 7). A total of 51 out of 108 patients relapsed, with disease recurrence ($n = 47$) or incompletely resected disease detected during follow-up ($n = 4$). At landmark, 27 out of 108 patients (25%) exhibited 1 or more positive ctDNA calls and 25 out of 27 of these patients relapsed (positive predictive value of landmark for relapse 93%, negative predictive value 68%, sensitivity of landmark for relapse 49%). Landmark-positive status was associated with higher pTNM stage (5 out of 41 (12%) patients with stage I, 8 out of 35 (23%) patients with stage II and 14 out of 32 (44%) patients with stage III were landmark positive, χ^2 test, $P = 0.008$). In total, 15 out of 21 (71%) relapse events occurring within 1 year of surgery were landmark-positive versus 8 out of 26 (31%) events occurring later than 1 year (4 patients with incomplete resections excluded, χ^2 test, $P = 0.01$). The median clonal ctDNA level at MRD detection in landmark-positive patients who relapsed was 0.08% (range, 0.002–2.41%, $n = 25$; Extended Data Fig. 6b).

Twelve patients were landmark positive before adjuvant therapy (Supplementary Table 15 and Fig. 3a–d). A pre-adjuvant patient positive for ctDNA (CRUK0086) had undetectable ctDNA after adjuvant radiotherapy and was disease-free until non-cancer associated death; the remaining 11 out of 12 patients eventually clinically relapsed despite 5 out of 11 patients exhibiting undetectable ctDNA after adjuvant therapy, indicating that ctDNA clearance in this setting may not always predict a positive outcome (Extended Data Fig. 6c).

In 102 out of 108 patients evaluable for survival analyses, landmark-positive patients exhibited a hazard ratio (HR) of 5.3 (95% CI = 2.9–9.7, log-rank test, $P = 1 \times 10^{-9}$) for OS and an HR of 6.8 for FFR (95% CI = 3.7–12.3, log-rank test, $P = 6 \times 10^{-13}$) relative to landmark-negative patients (Methods and Extended Data Fig. 6d,e).

In total, 16 out of 81 (20%) landmark-negative patients emerged to be ctDNA positive during ctDNA surveillance before, or at, clinical relapse; this occurred a median of 359 days postoperatively (range, 120–929 days), after a median of 3 negative postoperative plasma samples (range, 1–9) at a median clonal ctDNA level of 0.02% (range, 0.003–6.67%) (Fig. 3a,b,d and Extended Data Fig. 6b).

ctDNA lead times

The overall median lead time encountered in the cohort was 119 days (range, 0–1,137 days, $n = 63$; Methods). Lead times were associated with landmark status (Kruskal–Wallis test, $P = 0.006$); landmark-positive patients had the longest lead times (median, 228 days; range, 0–1,137 days; $n = 23$) relative to landmark-negative patients (median 76 days; range, 0–980 days, $n = 24$, Wilcoxon rank sum test, $P = 0.010$) and landmark-unevaluable patients (median, 56 days; range, 0–477 days, $n = 16$, Wilcoxon rank sum test, $P = 0.005$; Extended Data Fig. 6f).

Imaging and ctDNA

We assessed postoperative ctDNA detection in the context of standard-of-care extracranial CT, magnetic resonance imaging or positron emission tomography imaging surveillance in the adjuvant setting (Methods, Fig. 3 and Supplementary Table 16). In patients who eventually experienced relapse, we identified 44 surveillance scans from 23 patients that showed no new abnormalities compared with previous imaging; 22 out of 23 patients had plasma sampling performed before these scans (Fig. 3c–e). In total, 9 out of 22 patients were ctDNA positive before the scan and 8 out of 9 patients positive for ctDNA had eventual recurrence at sites covered by the extracranial scans (CRUK0590 experienced intracranial recurrence; Extended Data Fig. 6g). Thus, in some cases, positive postoperative ctDNA status preceded new abnormalities on surveillance imaging. Postoperative ctDNA detection before equivocal abnormalities occurred in 23 patients—20 out of 23 had subsequent NSCLC recurrence (Fig. 3 and Extended Data Fig. 6g,h). Before surveillance imaging showing new equivocal lymphadenopathy, 14 patients were ctDNA positive and 20 patients were ctDNA negative. A total of 11 out of 14 (79%) patients positive for ctDNA subsequently relapsed with lymph node involvement at the equivocal site versus 6 out of 20 (30%) patients negative for ctDNA before the scan (Fisher's test, $P = 0.013$; Extended Data Fig. 6i). Establishing ctDNA status may facilitate definitive therapeutic intervention at equivocal radiological sites, supporting previous findings from a cohort predominantly consisting of locally advanced NSCLC treated with chemo-radiation therapy[6].

ctDNA-based measurement of clonal architecture

To estimate tumour subclonal composition from deep targeted sequencing of plasma cfDNA, we developed ECLIPSE. ECLIPSE leverages background noise estimates and tumour-tissue-derived copy-number information to assess the presence or absence of specific tumour subclones and calculate their respective cancer cell fractions (CCFs) from low-tumour-content cfDNA data (Methods, Extended Data Fig. 7 and Supplementary Note). Plasma samples with clonal ctDNA levels of ≥0.1% (64% of ctDNA-positive samples) had an estimated minimally detectable CCF of ≥20% for a representative subclone (Methods, Supplementary Note and Extended Data Fig. 8a–d). Using 76,263 subclones constructed in silico from the AMP analytical validation spike-in data, we estimated a detection sensitivity of 94% for 20% CCF subclones in 0.1% clonal-ctDNA-level plasma with 4 tracked mutations and 10 ng DNA input (Extended Data Fig. 8e and Supplementary Note). We observed a decline in detection rates below 10 ng DNA input and therefore considered samples with ≥0.1% clonal ctDNA level and ≥10 ng cfDNA input as 'high-subclone-sensitivity', and analysed their clonal composition with ECLIPSE.

ECLIPSE measures of subclonal CCF from preoperative plasma samples were proportional to tumour exome multiregion sequencing measures of subclonal CCF sampled at surgery (Pearson $R = 0.78$, m (gradient) = 1, median clonal ctDNA level = 0.9%; Extended Data Fig. 9a,b and Supplementary Note). Subclone detection rates in preoperative plasma increased with subclone size (CCF) in the primary tumour (Extended Data Fig. 9c). Using plasma-based CCFs, we found evidence of sampling bias in measurements of tissue CCF for subclones unique to a single tumour region (Extended Data Fig. 9d–g and Supplementary Note).

Refining heterogeneity estimates using ctDNA

In the TRACERx 421 cohort[28] a median of 12% of mutations were determined to be present in all cancer cells of at least one resected tumour region but were absent from other regions of the tumour, therefore exhibiting a clonal illusion (Fig. 4a). ctDNA may be released from several regions of the tumour and resolve the true subclonal nature of mutations displaying a clonal illusion. In 71 TRACERx patients with high-subclone-sensitivity plasma samples available preoperatively, plasma-based CCFs were lower for clonal illusion mutations compared with mutations ubiquitous across all resected tumour regions (Wilcoxon test, $P < 0.001$; Fig. 4a), and plasma CCFs could predict clonal illusion with an area under the curve of 0.81 (95% CI = 0.79–0.82; Extended Data Fig. 9h). This suggests that collection of plasma alongside a single tumour biopsy can overcome tissue sampling bias, potentially increasing the accuracy of future heterogeneity-based clinical biomarkers[29,30].

Clonal expansions forecast metastasis

Predicting the subclonal composition of the subsequent metastatic recurrence at the time of surgery could inform precision adjuvant therapies against subclone(s) driving disease relapse. Primary tumour subclones (subclones detected in primary tumour tissue, excluding subclones unique to lymph node or ipsilateral pulmonary metastases resected at initial surgery; Methods) detected in postoperative cfDNA

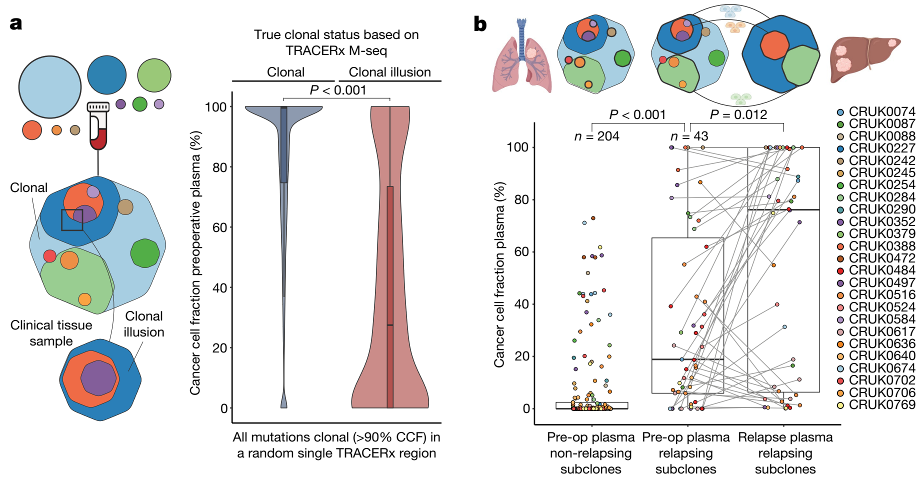


Fig. 4 | Clonality measurements in preoperative plasma overcome sampling bias from a single tissue sample and predict metastatic seeding potential. **a**, Depiction of a clonal illusion in which a dark blue subclone is found in 100% of cells in a single clinical tissue sample. Such clonal illusion mutations may be detected in the clinical setting using ctDNA derived from many different tumour regions to increase the accuracy of intratumour heterogeneity (ITH) measurements in the clinic. Mutations that were clonal (CCF > 90%) in a single, randomly selected tumour region were compared using plasma-based preoperative CCFs splitting by those truly clonal across all tumour regions in TRACERx (clonal) and those that, although they were clonal in the randomly selected region, were absent from other tumour regions (clonal illusion). Only data from a single randomly selected region were used by ECLIPSE to generate these CCFs. The distribution of plasma CCFs in each case is represented by a violin plot and a box and whisker plot. A Wilcoxon rank sum test was used to compare groups. Only preoperative samples with at least a 0.1% clonal ctDNA level (high-subclone-sensitivity samples, 71 samples from 71 patients) were included in this analysis (Supplementary Note for analysis of lower ctDNA levels). M-seq, multiregional sequencing. **b**, Preoperative (pre-op) plasma primary tumour subclone CCFs split by whether a given subclone was found to be present or absent in cfDNA samples at relapse and postoperative plasma CCFs for relapse subclones at the last high-subclone-sensitivity timepoint. Only tumours with at least one sample with >0.1% clonal ctDNA level (high subclone sensitivity) both preoperatively and postoperatively were included. $n = 26$ tumours with CCFs from 247 subclones included. Two-sided Wilcoxon tests were used to compare groups. The schematic in **b** was created using BioRender.

displayed larger CCFs in plasma samples taken before surgery relative to subclones that were not detectable postoperatively (Wilcoxon test, $P < 0.001$) and these metastatic subclones tended to expand further at relapse (Wilcoxon test, $P = 0.027$; Fig. 4b). This result indicates that primary tumour subclonal expansion measured non-invasively using ctDNA is associated with metastatic potential. In our companion papers, we demonstrate a similar effect using metastasis tissue sampling[31] and describe increased proliferative transcriptional signatures associated with metastasis-seeding primary tumour subclones[32].

Metastatic dissemination patterns in ctDNA

Comprehensive tissue sampling is challenging in the early-relapse setting. A total of 44% of patients with relapse had a tissue sample obtained at relapse, yet ease of plasma sampling enabled us to obtain high-subclone-sensitivity postoperative plasma samples in 61% of patients with relapse (mean 2 samples per patient). In total, 38% of patients with relapse had high-subclone-sensitivity postoperative plasma samples, but lacked a relapse tissue sample (Extended Data Fig. 10a). In 26 patients with both high-subclone-sensitivity postoperative plasma and recurrence tissue, we found a high concordance between subclones detected in recurrence tissue and postoperative ctDNA (98% sensitivity with 50 out of 51 relapse tissue subclones detected that were tracked by PSPs; Extended Data Fig. 10b,c). Additional subclones detected in postoperative ctDNA but absent from relapse tissue were found in 6 out of 26 patients (20 subclones). These subclones may have evaded tumour biopsy detection due to undersampling of metastatic sites at relapse (Supplementary Note). This is consistent with our companion article[31], which suggests that a single metastatic biopsy is not sufficient to confidently capture all metastatic dissemination events.

ECLIPSE-mediated calculation of subclone CCFs coupled with PSP targeting of the majority of sampled subclones in NSCLC resections (Extended Data Fig. 1b) facilitated the estimation of dissemination patterns from the primary tumour to relapse using ctDNA (Supplementary Note). Tumours were categorized by the number of relapse-seeding primary tumour subclones (monoclonal = 1, polyclonal ≥ 1) and relapse-seeding primary tumour phylogenetic tree branches (monophyletic = 1, polyphyletic ≥ 1; Fig. 1a and Methods). Longitudinal plasma- and tissue-based clonal composition estimates from surgery to relapse are presented for 44 patients with high-subclone-sensitivity postoperative plasma (Methods, Fig. 5a and Supplementary Fig. 1). We found an increased frequency of polyclonal metastatic dissemination at relapse when using ctDNA compared with recurrence biopsy tissue, driven by detection of ctDNA-unique subclones (10% polyclonal dissemination using tissue versus 24% polyclonal dissemination using ctDNA in matched cases; Extended Data Fig. 10d). Overall, 32 out of 44 recurrent tumours were defined as monoclonal dissemination and 12 out of 44 as polyclonal dissemination (3 polyclonal monophyletic and 9 polyclonal polyphyletic). Shorter OS from study registration and from the first ctDNA positive timepoint was observed in patients exhibiting polyclonal dissemination versus monoclonal dissemination (Fig. 5b and Extended Data Fig. 10e; post-registration OS: HR = 3.49, 95% CI = 1.57–7.77, $P = 0.001$, log-rank test, $n = 44$). OS from first postoperative ctDNA detection remained significant after adjustment for maximum postoperative clonal ctDNA level, assay DNA input amount, pTNM, preoperative ctDNA positivity, ctDNA detection in the first postoperative plasma sample and histology in multivariable analysis (Extended Data Fig. 10f).

Longitudinal tracking of clonal evolution

We addressed whether phylogenetic tracking could detect changes in subclonal composition, which may represent therapy-induced shifts in selection pressure. In 18 out of 42 (43%) patients with a high-subclone-sensitivity postoperative plasma sample available, we estimated that subclones tracked from the surgically resected tumour had undergone a complete clonal sweep at recurrence, whereby a subpopulation of cells expands to become clonal across all tumour sites (Methods, Extended Data Fig. 10g–i and Supplementary Note).

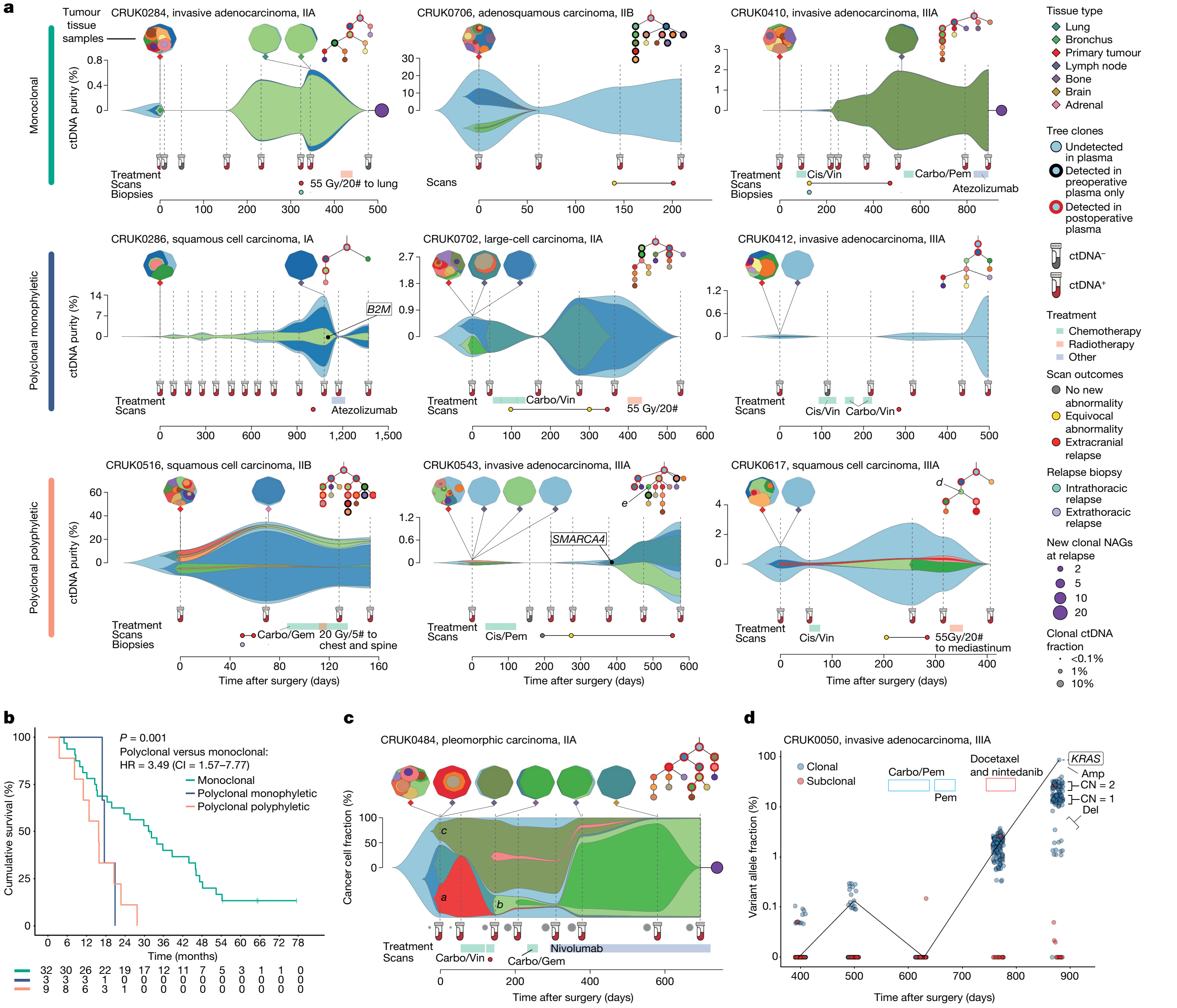


Fig. 5 | Longitudinal measurements of clonal evolution in the plasma from surgery to therapy and recurrence. a–**d**, The ctDNA purity for each clone was calculated by multiplying the clone CCF by the ctDNA purity of the plasma sample (Methods) and represents the fraction of all cells from which cfDNA was derived that harbour a given tumour clone at each timepoint. The *y*-axes indicate the total tumour purity at each vertically symmetrical point on the plots. The area of each subclone then represents the total proportion of the tumour in which each subclone is present at each timepoint. Clonal nesting is based on the phylogenetic tree for each tumour. Data from all ctDNA-positive plasma samples are shown, including results from ECLIPSE of samples with less than 0.1% clonal ctDNA level. Clone maps for each tumour tissue mass are depicted above the ctDNA-based clonal structure with the phylogenetic tree. Metastatic dissemination class was defined using primary tumour subclones, excluding metastatic unique clones in surgically excised lymph nodes or intrapulmonary metastases (Methods). Both CRUK0617 subclone D and CRUK0543 subclone E were not detected in ctDNA but their presence was inferred by detection of its daughter subclones (Supplementary Note). **a**, Depictions of longitudinal tumour evolution for examples of monoclonal, polyclonal monophyletic and polyclonal polyphyletic metastatic dissemination patterns. **b**, A Kaplan–Meier plot depicting differences in the overall survival between metastatic dissemination classes (*n* = 44 tumours, which had at least 1 high subclone sensitivity postoperative sample). A log-rank test was used to compare survival in the two groups. **c**, CCFs depicted through time and therapy for patient CRUK0484, who experienced a polyclonal polyphyletic relapse. **d**, Variant allele fractions for mutations tracked in patient CRUK0050 at recurrence. Carbo, carboplatin; Cis, cisplatin; CN, copy number; Gem, gemcitabine; Gy, Gray; NAG, neoantigen, Pem, pemetrexed; Vin, vinorelbine; # indicates the number of radiotherapy fractions.

We observed shifts in clonal composition in patient CRUK0484 concurrent with treatment (Fig. 5c and Supplementary Note), including extinguishing of a subclone present in more than half of tumour cells after surgery (clone *a*) during adjuvant chemotherapy, and expansion of a minor subclonal lineage (clone *b*) during post-recurrence immunotherapy treatment, which eventually outcompeted a parallel lineage (clone *c*). Despite three relapse tissue biopsies at different timepoints and metastatic sites, the dominant clone *c* was not detected in post-surgical tissue samples but only in a surgically excised lymph node. In patient CRUK0050, we observed a rapid increase in clonal ctDNA levels at day 876, after treatment of recurrent lung disease with cytotoxic chemotherapy (Fig. 5d). The multimodal distribution of clonal variant allele frequencies (VAFs) observed in the plasma suggested that 59 out of 130 clonal mutations had altered their copy-number state compared with samples taken at surgery (Methods), including evidence for amplification of an oncogenic *KRAS*^{G12R} mutation (84% VAF). This implies the expansion of a new subclone during treatment with significant chromosomal instability, not directly tracked by the PSP.

Discussion

In summary, we have demonstrated that preoperative ctDNA detection is prognostic in early-stage adenocarcinoma and implicated chromosomal instability as a predictor of ctDNA detection in this NSCLC subtype. These findings suggest that management of early-stage adenocarcinomas that are deemed to be high risk on the basis of preoperative ctDNA detection is inadequate, with innovation urgently needed.

Postoperative ctDNA detection forecasted impending NSCLC relapse, agreeing with previous findings[5–9,33–35]. Here, 25% of patients were landmark MRD positive and 93% of these patients relapsed a median of 228 days after ctDNA detection. Assessment of early treatment escalation in this high-risk population is required. ctDNA surveillance identified impending relapse in 20% of landmark-negative patients—emergence of ctDNA during surveillance may reflect low-burden metastatic disease initially shedding ctDNA quantities below assay limit of detection (around 95% sensitivity at 0.008% ctDNA level in ≥30 ng DNA input samples). Landmark MRD detection rates could increase with next-generation assays with improved ctDNA limits of detection[36–38].

Previous publications used high-tumour-fraction ctDNA samples (>10%) to calculate subclonal CCFs[12–14]. However, such samples are rare[17], comprising only 9% of ctDNA-positive samples in this study from 14 out of 145 (10%) patients in which ctDNA was ever detected. ECLIPSE, combined with AMP PSPs, enabled an estimated 94% detection sensitivity for 20% CCF subclones in plasma samples with 0.1% tumour content and could accurately estimate CCFs using such samples. We demonstrated that ctDNA can sample clonal structure from multiple different surgically excised tissue sites and capture additional heterogeneity at relapse when compared to analysis of relapse tissue samples. Despite this, two-thirds of patients with disease recurrence still had only one ctDNA-detectable metastasizing primary tumour subclone (monoclonal dissemination). However, low ctDNA levels and incomplete primary tumour sampling may limit the detection of additional disseminating primary tumour subclones. We observed a more aggressive disease course in patients with multiple metastatic dissemination events (polyclonal dissemination), suggesting that heterogeneity in the seeding population may provide fuel for Darwinian adaptation to different metastatic niches. However, the requirement to perform multiregional primary tumour sequencing currently limits the feasibility of determining metastatic dissemination patterns in the clinic.

ctDNA is poised to change (neo)adjuvant trial designs. Measurements of subclonal expansion in the plasma before surgery may enable the prediction of future metastatic subclones, offering the possibility for early intervention and suggesting new routes for biomarker development to target and eradicate such clones months or even years before relapse.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-05776-4.

1. Moding, E. J., Nabet, B. Y., Alizadeh, A. A. & Diehn, M. Detecting liquid remnants of solid tumors: circulating tumor DNA minimal residual disease. *Cancer Discov.* **11**, 2968–2986 (2021).
2. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
3. Chabon, J. J. et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245–251 (2020).
4. Peng, M. et al. Circulating tumor DNA as a prognostic biomarker in localized non-small cell lung cancer. *Front. Oncol.* **10**, 561598 (2020).
5. Xia, L. et al. Perioperative ctDNA-based molecular residual disease detection for non-small cell lung cancer: a prospective multicenter cohort study (LUNGCA-1). *Clin. Cancer Res.* **28**, 3308–3317 (2021).
6. Chaudhuri, A. A. et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discov.* **7**, 1394–1403 (2017).
7. Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).
8. Gale, D. et al. Residual ctDNA after treatment predicts early relapse in patients with early-stage non-small cell lung cancer. *Ann. Oncol.* **33**, 500–510 (2022).
9. Zhang, J.-T. et al. Longitudinal undetectable molecular residual disease defines potentially cured population in localized non-small cell lung cancer. *Cancer Discov.* **12**, 1690–1701 (2022).
10. Powles, T. et al. ctDNA guiding adjuvant immunotherapy in urothelial carcinoma. *Nature* **595**, 432–437 (2021).
11. Tie, J. et al. Circulating tumor DNA analysis guiding adjuvant therapy in stage II colon cancer. *N. Engl. J. Med.* **386**, 2261–2272 (2022).
12. Parikh, A. R. et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat. Med.* **25**, 1415–1421 (2019).
13. Murtaza, M. et al. Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.* **6**, 8760 (2015).
14. Herberts, C. et al. Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. *Nature* **608**, 199–208 (2022).
15. *Lung Cancer: Diagnosis and Management NICE Guideline* NG122 (NICE, 2019).
16. Zheng, Z. et al. Anchored multiplex PCR for targeted next-generation sequencing. *Nat. Med.* **20**, 1479–1484 (2014).
17. Abbosh, C., Birkbak, N. J. & Swanton, C. Early stage NSCLC—challenges to implementing ctDNA-based screening and MRD detection. *Nat. Rev. Clin. Oncol.* **15**, 577–586 (2018).
18. Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
19. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7 (2013).
20. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
21. Biswas, D. et al. A clonal expression biomarker associates with lung cancer mortality. *Nat. Med.* **25**, 1540–1548 (2019).
22. Burrell, R. A. et al. Replication stress links structural and numerical cancer chromosomal instability. *Nature* **494**, 492–496 (2013).
23. Wang, Z. C. et al. Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome. *Clin. Cancer Res.* **18**, 5806–5815 (2012).
24. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
25. Shih, D. J. H. et al. Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. *Nat. Genet.* **52**, 371–377 (2020).
26. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
27. Garcia-Murillas, I. et al. Assessment of molecular relapse detection in early-stage breast cancer. *JAMA Oncol.* **5**, 1473–1478 (2019).
28. Frankell, A. M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* https://doi.org/10.1038/s41586-023-05783-5 (2023).
29. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614 (2021).
30. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
31. Al Bakir, M. et al. The evolution of non-small lung cancer metastases in TRACERx. *Nature* https://doi.org/10.1038/s41586-023-05729-x (2023).
32. Martínez-Ruiz, C. et al. Genomic–transcriptomic evolution in lung cancer and metastasis. *Nature* https://doi.org/10.1038/s41586-023-05706-4 (2023).
33. Moding, E. J. et al. Circulating tumor DNA dynamics predict benefit from consolidation immunotherapy in locally advanced non-small cell lung cancer. *Nat. Cancer* **1**, 176–183 (2020).
34. Chen, K. et al. Perioperative dynamic changes in circulating tumor DNA in patients with lung cancer (DYNAMIC). *Clin. Cancer Res.* **25**, 7058–7067 (2019).
35. Li, N. et al. Perioperative circulating tumor DNA as a potential prognostic marker for operable stage I to IIIA non–small cell lung cancer. *Cancer* **128**, 708–718 (2021).
36. Kurtz, D. M. et al. Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA. *Nat. Biotechnol.* **39**, 1537–1547 (2021).
37. Cohen, J. D. et al. Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. *Nat. Biotechnol.* **39**, 1220–1227 (2021).
38. Gydush, G. et al. Massively parallel enrichment of low-frequency alleles enables duplex sequencing at low depth. *Nat. Biomed. Eng.* **6**, 257–266 (2022).

[1]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [2]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK. [3]Invitae, San Francisco, CA, USA. [4]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [5]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [6]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [7]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. [8]Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [9]Department of Radiology, Freiburg University Hospital, Freiburg, Germany. [10]Advanced Sequencing Facility, The Francis Crick Institute, London, UK. [11]Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [12]Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK. [13]Bill Lyons Informatics Centre, University College London Cancer Institute, London, UK. [14]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK. [15]Department of Cellular Pathology, University College London Hospitals, London, UK. [16]AstraZeneca, Cambridge, UK. [17]The Christie NHS Foundation Trust, Manchester, UK. [18]University Hospital Birmingham NHS Foundation Trust, Birmingham, UK. [19]Cancer Research Centre, University of Leicester, Leicester, UK. [20]AstraZeneca, Waltham, MA, USA. [21]Department of Biochemistry, University of Cambridge, Cambridge, UK. [22]Cancer Research UK & UCL Cancer Trials Centre, London, UK. [23]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands. [24]Department of Oncology, University College London Hospitals, London, UK. [101]These authors contributed equally: Christopher Abbosh, Alexander M. Frankell, Thomas Harrison, Judit Kisistok, Aaron Garnett. [102]These authors jointly supervised this work: Nicolai J Birkbak, Nicholas McGranahan, Charles Swanton. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: c.abbosh@ucl.ac.uk; nicholas.mcgranahan.10@ucl.ac.uk; Charles.Swanton@crick.ac.uk

TRACERx Consortium

Charles Swanton[1,2,24,102], Nicholas McGranahan[1,11,102], Christopher Abbosh[1,101], Alexander M. Frankell[1,2,101], Judit Kisistok[4,5,6,101], Selvaraju Veeriah[1], Sophia Ward[1,2,10], Kristiana Grigoriadis[1,2,11], Kevin Litchfield[1,12], Clare Puttick[1,2,11], Dhruva Biswas[1,2,13], Takahiro Karasaki[1,2,14], James R. M. Black[1,11], Carlos Martínez-Ruiz[1,11], Maise Al Bakir[1,2], Oriol Pich[2], Thomas B. K. Watkins[2], Emilia L. Lim[1,2], Ariana Huebner[1,2,11], David A. Moore[1,2,15], Akshay J. Patel[18], Crispin T. Hiley[1,2], Joan Riley[19], Allan Hackshaw[22], Jacqui A. Shaw[19], Mariam Jamal-Hanjani[1,14,24], Jason F. Lester[25], Amrita Bajaj[26], Apostolos Nakas[26], Azmina Sodha-Ramdeen[26], Keng Ang[26], Mohamad Tufail[26], Mohammed Fiyaz Chowdhry[26], Molly Scotland[26], Rebecca Boyles[26], Sridhar Rathinam[26], Claire Wilson[27], Domenic Marrone[27], Sean Dulloo[27], Dean A. Fennell[26,27], Gurdeep Matharu[19], Lindsay Primrose[19], Ekaterini Boleti[28], Heather Cheyne[29], Mohammed Khalil[29], Shirley Richardson[29], Tracey Cruickshank[29], Gillian Price[30,31], Keith M. Kerr[31,32], Sarah Benafif[24], Kayleigh Gilbert[33], Babu Naidu[34], Aya Osman[18], Christer Lacson[18], Gerald Langman[18], Helen Shackleford[18], Madava Djearaman[18], Salma Kadiri[18], Gary Middleton[18,35], Angela Leek[36], Jack Davies Hodgkinson[36], Nicola Totten[36], Angeles Montero[37], Elaine Smith[37], Eustace Fontaine[37], Felice Granato[37], Helen Doran[37], Juliette Novasio[37], Kendadai Rammohan[37], Leena Joseph[37], Paul Bishop[37], Rajesh Shah[37], Stuart Moss[37], Vijay Joshi[37], Philip Crosbie[37,38,39], Fabio Gomes[17], Kate Brown[17], Mathew Carter[17], Anshuman Chaturvedi[17,39], Lynsey Priest[17,39], Pedro Oliveira[17,39], Colin R. Lindsay[40], Fiona H. Blackhall[40], Matthew G. Krebs[40], Yvonne Summers[40], Alexandra Clipson[39,41], Jonathan Tugwood[39,41], Alastair Kerr[39,41], Dominic G. Rothwell[39,41], Elaine Kilgour[39,41], Caroline Dive[39,41], Hugo J. W. L. Aerts[7,8,23], Roland F. Schwarz[42,43], Tom L. Kaufmann[43,44], Gareth A. Wilson[2], Rachel Rosenthal[2], Peter Van Loo[45,46,47], Zoltan Szallasi[48,49,50], Mateo Sokac[4,5,6], Roberto Salgado[51,52], Miklos Diossy[48,49,53], Jonas Demeulemeester[47,54,55], Abigail Bunkum[1,14,56], Aengus Stewart[57], Alastair Magness[57], Andrew Rowan[2], Angeliki Karamani[58], Antonia Toncheva[1], Benny Chain[58], Brittany B. Campbell[2], Carla Castignani[47,59], Chris Bailey[2], Clare E. Weeden[57], Claudia Lee[2], Corentin Richard[1], Cristina Naceur-Lombardelli[1], David R. Pearce[58], Despoina Karagianni[58], Dina Levi[57], Elena Hoxha[58], Elizabeth Larose Cadieux[47,59], Emma Colliver[2], Emma Nye[60], Eva Grönroos[57], Felip Gálvez-Cancino[58], Foteini Athanasopoulou[1,2,10], Francisco Gimeno-Valiente[1], George Kassiotis[61,62], Georgia Stavrou[58], Gerasimos Mastrokalos[58], Haoran Zhai[1,2], Helen L. Lowe[58], Ignacio Matos[58], Jacki Goldman[57], James L. Reading[58], Javier Herrero[13], Jayant K. Rane[2,58], Jerome Nicod[10], Jie Min Lam[1,14,24], John A. Hartley[58], Karl S. Peggs[63,64], Katey S. S. Enfield[2], Kayalvizhi Selvaraju[58], Kerstin Thol[1,11], Kevin W. Ng[61], Kezhong Chen[58], Krijn Dijkstra[65,66], Krupa Thakkar[1], Leah Ensell[58], Mansi Shah[58], Marcos Vasquez[58], Maria Litovchenko[58], Mariana Werner Sunderland[1], Mark S. Hill[2], Michelle Dietzen[1,2,11], Michelle Leung[1,2,11], Mickael Escudero[57], Mihaela Angelova[2], Miljana Tanić[59,67], Monica Sivakumar[1], Nnennaya Kanu[1], Olga Chervova[58], Olivia Lucas[1,2,24,56], Othman Al-Sawaf[1,2,14], Paulina Prymas[1], Philip Hobson[57], Piotr Pawlik[58], Richard Kevin Stone[60], Robert Bentham[1,11], Robert E. Hynds[58], Roberto Vendramin[57], Sadegh Saghafinia[1], Saioa López[58], Samuel Gamble[58], Seng Kuong Anakin Ung[58], Sergio A. Quezada[1,68], Sharon Vanloo[1], Simone Zaccaria[1,56], Sonya Hessey[1,14,56], Stefan Boeing[57], Stephan Beck[59], Supreet Kaur Bola[58], Tamara Denner[57], Teresa Marafioti[15], Thanos P. Mourikis[58], Victoria Spanswick[58], Vittorio Barbè[57], Wei-Ting Lu[57], William Hill[57], Wing Kin Liu[1,14], Yin Wu[58], Yutaka Naito[57], Zoe Ramsden[57], Catarina Veiga[69], Gary Royle[70], Charles-Antoine Collins-Fekete[71], Francesco Fraioli[72], Paul Ashford[73], Tristan Clark[74], Martin D. Forster[1,24], Siow Ming Lee[1,24], Elaine Borg[15], Mary Falzon[15], Dionysis Papadatos-Pastos[24], James Wilson[24], Tanya Ahmad[24], Alexander James Procter[75], Asia Ahmed[75], Magali N. Taylor[75], Arjun Nair[75,76], David Lawrence[77], Davide Patrini[77], Neal Navani[78,79], Ricky M. Thakrar[78,79], Sam M. Janes[78], Emilie Martinoni Hoogenboom[80], Fleur Monk[80], James W. Holding[80], Junaid Choudhary[80], Kunal Bhakhri[80], Marco Scarci[80], Martin Hayward[80], Nikolaos Panagiotopoulos[80], Pat Gorman[80], Reena Khiroya[15], Robert CM. Stephens[80], Yien Ning Sophia Wong[80], Steve Bandula[80], Abigail Sharp[22], Sean Smith[22], Nicole Gower[22], Harjot Kaur Dhanda[22], Kitty Chan[22], Camilla Pilotti[22], Rachel Leslie[22], Anca Grapa[81], Hanyun Zhang[81], Khalid AbdulJabbar[81], Xiaoxi Pan[81], Yinyin Yuan[82], David Chuter[83], Mairead MacKenzie[83], Serena Chee[84], Aiman Alzetani[84], Judith Cave[85], Lydia Scarlett[84], Jennifer Richards[84], Papawadee Ingram[84], Silvia Austin[84], Eric Lim[86,87], Paulo De Sousa[87], Simon Jordan[87], Alexandra Rice[87], Hilgardt Raubenheimer[87], Harshil Bhayani[87], Lyn Ambrose[87], Anand Devaraj[87], Hema Chavan[87], Sofina Begum[87], Silviu I. Buderi[87], Daniel Kaniu[87], Mpho Malima[87], Sarah Booth[87], Andrew G. Nicholson[88,89], Nadia Fernandes[87], Pratibha Shah[87], Chiara Proli[87], Madeleine Hewish[90,91], Sarah Danson[92], Michael J. Shackcloth[93], Lily Robinson[94], Peter Russell[94], Kevin G. Blyth[95,96,97], Craig Dick[98], John Le Quesne[95,96,99], Alan Kirk[100], Mo Asif[100], Rocco Bilancia[100], Nikos Kostoulas[100], Mathew Thomas[100] & Nicolai J. Birkbak[1,2,4,5,6,102]

[25]Singleton Hospital, Swansea Bay University Health Board, Swansea, UK. [26]University Hospitals of Leicester NHS Trust, Leicester, UK. [27]University of Leicester, Leicester, UK. [28]Royal Free Hospital, Royal Free London NHS Foundation Trust, London, UK. [29]Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [30]Department of Medical Oncology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [31]University of Aberdeen, Aberdeen, UK. [32]Department of Pathology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [33]The Whittington Hospital NHS Trust, London, UK. [34]Birmingham Acute Care Research Group, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. [35]Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. [36]Manchester Cancer Research Centre Biobank, Manchester, UK. [37]Wythenshawe Hospital, Manchester University NHS Foundation Trust, Wythenshawe, UK. [38]Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK. [39]Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK. [40]Division of Cancer Sciences, The University of Manchester and The Christie NHS Foundation Trust, Manchester, UK. [41]Cancer Research UK Manchester Institute Cancer Biomarker Centre, University of Manchester, Manchester, UK. [42]Institute for Computational Cancer Biology, Center for Integrated Oncology (CIO), Cancer Research Center Cologne Essen (CCCE), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [43]Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany. [44]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [45]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [46]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [47]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. [48]Danish Cancer Society Research Center, Copenhagen, Denmark. [49]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [50]Department of Bioinformatics, Semmelweis University, Budapest, Hungary. [51]Department of Pathology, ZAS Hospitals, Antwerp, Belgium. [52]Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [53]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. [54]Integrative Cancer Genomics Laboratory, Department of Oncology, KU Leuven, Leuven, Belgium. [55]VIB–KU Leuven Center for Cancer Biology, Leuven, Belgium. [56]Computational Cancer Genomics Research Group, University College London Cancer Institute, London, UK. [57]The Francis Crick Institute, London, UK. [58]University College London Cancer Institute, London, UK. [59]Medical Genomics, University College London Cancer Institute, London, UK. [60]Experimental Histopathology, The Francis Crick Institute, London, UK. [61]Retroviral Immunology Group, The Francis Crick Institute, London, UK. [62]Department of Infectious Disease, Faculty of Medicine, Imperial College London, London, UK. [63]Department of Haematology, University College London Hospitals, London, UK. [64]Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [65]Department of Molecular Oncology and Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [66]Oncode Institute, Utrecht, The Netherlands. [67]Experimental Oncology, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia. [68]Immune Regulation and Tumour Immunotherapy Group, Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [69]Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [70]Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, UK. [71]Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [72]Institute of Nuclear Medicine, Division of Medicine, University College London, London, UK. [73]Institute of Structural and Molecular Biology, University College London, London, UK. [74]University College London, London, UK. [75]Department of Radiology, University College London Hospitals, London, UK. [76]UCL Respiratory, Department of Medicine, University College London, London, UK. [77]Department of Thoracic Surgery, University College London Hospital NHS Trust, London, UK. [78]Lungs for Living Research Centre, UCL Respiratory, University College London, London, UK. [79]Department of Thoracic Medicine, University College London Hospitals, London, UK. [80]University College London Hospitals, London, UK. [81]The Institute of Cancer Research, London, UK. [82]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [83]Independent Cancer Patients' Voice, London, UK. [84]University Hospital Southampton NHS Foundation Trust, Southampton, UK. [85]Department of Oncology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [86]Academic Division of Thoracic Surgery, Imperial College London, London, UK. [87]Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [88]Department of Histopathology, Royal Brompton and Harefield Hospitals, Guy's and St Thomas' NHS Foundation Trust, London, UK. [89]National Heart and Lung Institute, Imperial College London, London, UK. [90]Royal Surrey Hospital, Royal Surrey Hospitals NHS Foundation Trust, Guilford, UK. [91]University of Surrey, Guilford, UK. [92]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [93]Liverpool Heart and Chest Hospital, Liverpool, UK. [94]Princess Alexandra Hospital, The Princess Alexandra Hospital NHS Trust, Harlow, UK. [95]School of Cancer Sciences, University of Glasgow, Glasgow, UK. [96]Cancer Research UK Beatson Institute, Glasgow, UK. [97]Queen Elizabeth University Hospital, Glasgow, UK. [98]NHS Greater Glasgow and Clyde, Glasgow, UK. [99]Pathology Department, Queen Elizabeth University Hospital, NHS Greater Glasgow and Clyde, Glasgow, UK. [100]Golden Jubilee National Hospital, Clydebank, UK.

Methods

Patients and tissue samples

The TRACERx study (ClinicalTrials.gov: NCT01888601) is a prospective observational cohort study that aims to transform our understanding of NSCLC, and the design of which has been approved by an independent research ethics committee (NRES Committee London, REC ref:13/LO/1546). Informed consent for entry into the TRACERx study was mandatory and obtained from every patient. All patients were assigned a study identity number that was known to the patient. These were subsequently converted to linked study identities such that the patients could not identify themselves in study publications. All human samples (tissue and blood) were linked to the study identity number and barcoded such that they were anonymized and tracked on a centralized database, which was overseen by the study sponsor only. The ctDNA cohort represents 188 TRACERx 421 cohort eligible patients and 9 additional patients (the following 9 patients were excluded from the final TRACERx T421 cohort (after ctDNA analyses were performed) and were analysed in this manuscript: CRUK0230, 0234, 0291, 0335, 0387, 0480, 0490, 0498, 0622). The reasons for exclusion from final T421 cohort were as follows: CRUK0480 and 0490: C>A artefact uncovered in exome data (excluded from ECLIPSE analyses); CRUK0291, 0234, 0230, 0387 and 0622: incomplete resection of NSCLC; CRUK0335: concurrent oesophageal primary present at diagnosis; CRUK0498: 1 of 2 tumour regions contained lymphoid associated variants. The remaining preoperative plasma from 19 patients published previously[7] was also analysed in this paper; these patients can be identified by CRUK IDs shared between the papers. Extended Data Fig. 3a describes the structure of the patient cohort analysed, patients analysed in the Extended Data Fig. 2 pilot cohort to assess optimum ctDNA detection thresholds were excluded from clinical analyses associating pre- and postoperative ctDNA detection with patient characteristics and survival outcomes (Figs. 1 and 3) and biological analyses of ctDNA detection in lung adenocarcinoma (Fig. 2). However, these patients were included in ECLIPSE clonality analyses (Figs. 4 and 5). Multiregion tumour sampling was performed as previously described[2]. Relapse tissue samples, excess to diagnostic requirements, were also acquired. Sample extraction from tissue and whole blood was performed according to the protocol in the TRACERx 100 cohort and exome sequencing was performed as previously described[2].

Analyses of adjuvant surveillance and relapse scan reports

Relapse site data were collected from anonymized standard of care imaging scan reports that occurred within 180 days of confirmed clinical relapse (Supplementary Table 14). Each report was reviewed by two clinicians and sites of disease documented. Two patients lacked available scan reports (CRUK0311 and 0452); for these two patients, data were gathered from TRACERx case report forms. Where an anatomical site was not covered by a recurrence scan, this was marked as not evaluable. Anonymized surveillance (pre-relapse or relapse) scan reports were reviewed from 121 out of 131 non-pilot patients who had donated longitudinal plasma samples (321 CT scans, 7 magnetic resonance imaging scans and 36 whole-body positron emission tomography scans). Surveillance scan reports were not available in 10 out of 131 non-pilot patients. These reports were categorized as showing no new abnormality compared with previous imaging, new equivocal abnormality (an equivocal abnormality was defined as any new change compared to a previous scan, equivocal changes were categorized as being related to new lung tissue abnormality including nodules, enlarging lymph-nodes, pleural abnormality or pleural effusion, lung atelectasis or collapse or other changes) or new unequivocal abnormality (scans showing a change that was viewed as definitive malignancy and resulted in a change in clinical management; Supplementary Table 16). This central review of reports was performed blinded to a patient's disease and death status. In cases in which questions regarding interpreting the report arose, there was a dialogue with the cancer centre to establish an agreed assessment.

Plasma samples

Blood samples were collected and processed to plasma as previously described[7]. Up to 4 ml of plasma per case was evaluated for the study (range, 0.5–4 ml; median, 4 ml; Supplementary Table 2). For 1,074 out of 1,095 samples, circulating cfDNA was purified from plasma using the MagMAX Cell-Free DNA Isolation Kit in conjunction with the KingFisher Flex Purification System (Thermo Fisher Scientific). KingFisher 24-deepwell processing plates were prepared according to the manufacturer's instructions (plate setup option for KingFisher Flex Magnetic Particle Processor 24DW, 4 ml of plasma, 75 µl elution volume). Automated cfDNA isolation was performed on the KingFisher Flex system. For the remaining 21 samples, cfDNA was extracted as previously described[7]. Eluted cfDNA samples were quantified on the Qubit 3.0 Fluorometer using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. The single-nucleotide polymorphism (SNP) profile of cfDNA from a patient was matched back to normal exome data and samples exhibiting discordant SNP profiles were excluded as sample swaps ($n = 26$ out of 1,095 plasma samples analysed).

Volumetric analyses

Tumour volume was determined on the basis of pretreatment (PET-) CT scans using 3D Slicer. Contours of the primary tumour were manually segmented on each axial CT slice. Window settings were adjusted if necessary to exclude vessels, lymph nodes or adjacent mediastinal tissue. If no accurate delineation of the primary tumour was possible (for example, large cavity, pleural effusion or atelectasis), the patient was excluded from volume analysis (Extended Data Fig. 4a); patients with minor cavities within tumours were included. These steps were performed by a trained resident and all contours were confirmed and edited where necessary by an experienced radiologist. Relevant clinical demographics including gender and tumour location were cross checked with imaging appearances for each scan analysed. Volumetric data are provided in Supplementary Table 8.

Library preparation using AMP

AMP is a nested multiplex–PCR enrichment chemistry that incorporates strand-specific priming and the incorporation of unique molecular identifiers (UMIs) into sequenced reads[16]. cfDNA, fragmented peripheral blood mononuclear cell (PBMC) DNA (60 ng) or fragmented normal tissue DNA (60 ng) was end-repaired phosphorylated and A-tailed. An adapter containing a universal priming site, the indexes for multiplexing and a UMI are then ligated onto DNA. One round of target-specific PCR was performed with a gene-specific primer 1 (GSP1) which amplifies against the P5 primer in the adapter, and a further round of PCR was then performed with a second nested gene-specific primer (GSP2) and a primer that incorporates a second primer containing a P7 index. Strand-specific priming was performed in both rounds of amplification facilitating the identification of positive- and negative-strand input DNA molecules during informatic analyses.

For cfDNA libraries, indexed libraries were quantified on either the ViiA 7 Real-Time PCR System or QuantStudio Dx Real-Time PCR Instrument (Thermo Fisher Scientific) using the KAPA Library Quantification Kit (Roche). Libraries were individually normalized on the Fluent 1080 Automated Workstation (Tecan), then symmetrically pooled and adjusted to a final concentration of 2 nM or 1.25 nM for standard or Xp NovaSeq loading workflows, respectively. Library pools were prepared and sequenced on the NovaSeq 6000 System (Illumina) according to the manufacturer's protocol. We aimed to sequence each library to about 10 million reads. The on-target deduplication ratio of the library, which describes the ratio of raw on-target reads to UMI-supported on-target reads (UMI-supported reads contained five or more supporting raw reads with a matched molecular index) was then evaluated. In samples for which initial sequencing depth resulted in an on-target de-duplication ratio of less than 10:1, additional sequencing was performed; this quality-control step was introduced to maximize the recovery of UMI families (which require at least five UMI-supported reads) from high-complexity samples to ensure recoverable information from these samples, thereby reducing bias (given that only UMI families are considered in our analyses). This quality-control step resulted in the majority of cfDNA libraries (1,052 out of 1,069) having median de-duplication ratios of more than 5 (Extended Data Fig. 1f). PBMC and normal tissue libraries were sequenced on the NovaSeq 6000 system (Illumina) or the NextSeq system (Illumina).

MRD calling algorithm

We generated an MRD caller (v.0.1) that investigated background sequencing noise on an intralibrary basis (Fig. 1a and Supplementary Note). The MRD caller used the Archer informatic pipeline to clean input reads and generate deduplicated UMI-supported reads. The cleaned, deduplicated and error-corrected UMI-supported reads were aligned to the hg19 genome and used to evaluate alternate observations at predefined positions where tumour-specific variants were present in the patient's tumour (tumour-informed positions). Only 'deep' consensus reads supported by five or more PCR duplicates (UMI corrected) were used to infer expected sequencing noise as well as calculate signal for the MRD calling algorithm.

Alternate bases at tumour-informed positions were subject to a strict set of quality filters consisting of an off target filter, a read strand bias filter, a sequencing strand bias filter, background error rate filter and variant AF outlier filter to remove artefactual signals. The variant AF outlier filter functioned by performing partitioning around medoids (PAM) clustering of the VAFs of the tumour-informed positions that passed previously described filters. *K* was set to 2 in the clustering algorithm, therefore yielding a high-VAF group and a low-VAF group. If one of the two clusters had significantly higher VAFs (as indicated by non-overlapping confidence intervals of the highest VAF of the low VAF cluster and the lowest VAF of the higher VAF cluster) and contained three or fewer tumour-specific variants, those variants were removed from consideration downstream in the algorithm.

Next, intralibrary background error rates (ERs) were calculated. ERs were used to establish the level of noise present in each library that had to be confidently exceeded to allow an MRD call to be made. To calculate background library ERs, the number of UMI-supported alternate observations (deep alternate observations (DAOs)) were tallied across the assay's region of interest for each trinucleotide context (TNC) and for each possible alternate position based on the plus strand of the reference sequence. The ER corresponding to each TNC alternate was calculated as DAO/DDP (where DDP is the deep UMI-corrected depth across a TNC alternate). To measure only PCR and sequencing error, a position in the region of interest was not included in the TNC ER calculation if the VAF at that position for a particular alternate was >1% (on the basis that this could represent a clonal haematopoiesis-associated mutation or a SNP).

A mapping of tumour-observed variants and their accompanied TNC ERs was generated. Any tumour-observed variant with a corresponding TNC ER upper confidence interval that was above 0.01% was filtered from the MRD calling algorithm. PAM clustering was used to generate four 'D-groups' of TNC ERs from qualified TNCs. The population-weighted average TNC ER was calculated for each of the four D-groups based on the product of the TNC ERs included in each D-group cluster and the total DDP for each TNC. The generation of four D-groups ensured that there was sufficient intralibrary DDP coverage of each D-group to make precise estimations regarding ERs for variants within each group.

To determine whether ctDNA was present in the sample, the total observed DAOs summed across tumour specific positions remaining after filters were compared to the number of DAOs that were expected due to background ERs as dictated by the D-groups. A one-tailed exact Poisson test was applied where the total remaining observed DAOs served as the value being tested and the expected number of DAOs due to error served as the lambda of the Poisson distribution. If the resulting *P* value of the test was below a prespecified alpha threshold set to 0.01, then the sample was classified as MRD positive. The Supplementary Note contains details regarding how the prespecified alpha threshold of 0.01 used in these analyses was generated.

To investigate whether a single mutation targeted by a panel was present, we used the specific trinucleotide ER corresponding to the mutation of interest and a one-tail Poisson test to assess whether the number of DAOs across the mutation of interest was above expected background ER. If the number of DAOs was higher than expected background error using an alpha threshold of 0.01, then a variant was deemed to be confidently detected. Supplementary Tables 13 and 17 contain sample- and variant-level outputs of the MRD caller pipeline.

Estimating the effect of panel size on MDAF

We estimated the minimally detectable allele fraction (MDAF) for total ctDNA to estimate our ctDNA sensitivity in each TRACERx plasma sample. We estimated the number of observed consensus mutant reads that would be required to produce a ctDNA-positive call at a threshold of $P < 0.01$, given the total background noise estimated across all mutations considered. To assess the effect of the number of mutations tracked on our ctDNA sensitivity, we randomly subsampled 1, 2, 5, 10, 20, 50, 75, 100 and 150 mutations for each of our 200 mutation panels and assessed the MDAF. The median MDAF for samples with 20 ng to 30 ng using 50 mutations (0.008%) was very similar to the sensitivity estimated using our in vitro validation data (>90% sensitivity at 0.01% allele fraction).

Data inputs for ECLIPSE

For each mutation, ECLIPSE requires mutation identifiers (chromosome, position, reference allele, alternative allele), a sample identifier, the number of supporting reads, sequencing depth, estimated background ER, clone identifier, a binary call for whether the mutation is clonal or subclonal, mutation multiplicity, total copy number at the mutated locus in tumour cells, total copy number at the mutated locus in non-tumour cells (default = 2). ECLIPSE also takes several optional inputs, including variants to be filtered for clone and tumour presence calls due to high background error, variants that should be filtered from all analysis for a specific sample and a measurement for the maximally expected normalized standard deviation of CCF in high confidence clones used to identify clones with incoherent CCF distributions that may represent mutation clusters that are not true clones. The background ER is the probability, for any given read, to observe the specified mutation due to sequencing error. For application of ECLIPSE to our TRACERx data, we estimate this using TNC-specific ERs at non-mutated loci in the deep targeted sequencing data (see the 'MRD calling algorithm' section). The clone identifier, clonal versus subclonal status, mutation multiplicity and total copy number in tumour cells can be calculated using standard copy-number extraction and clonal deconvolution methods (ASCAT[39], Battenberg[40], Pyclone[41], DpCLust[40]) used for high-tumour-purity (>10%) samples—for example, from tissue samples—and these can then be used as estimates for these variables at the time of ctDNA sampling. Clonal status can be more accurately and comprehensively extracted from the sequencing of multiple high-purity samples from the same patient, as is performed in TRACERx, but is not essential. See the 'Application of ECLIPSE to the TRACERx cfDNA data' section for further details.

Stepwise description of ECLIPSE

VAF denoising. VAFs are denoised by subtracting the estimated background error, provided to ECLIPSE for each variant. For a description of estimating background error in this dataset see the 'MRD calling algorithm' section. Variants in each clone are grouped into clusters (via *k*-means clustering) with similar background error profiles, where the number of clustered groups is determined by the sum of the error estimated across all variants, so that if equally dividing the total error from all variants of a clone, each group would have a combined error of at least one mutant read. Therefore, if a clone has a total combined error of less than two mutant reads, only one group will be used. A maximum number of clusters is set to four as the default value (which was used for application to the TRACERx plasma sequencing data). The average background error of each group per variant is subtracted from the number of supporting reads for all variants in each group and divided by the sequencing depth to calculate denoised VAFs.

ctDNA tumour purity calculation. Denoised VAFs are used with mutation multiplicities, total copy number at the mutated locus and clonal versus subclonal mutation status for each mutation provided to ECLIPSE to calculate an estimate of ctDNA tumour purity using the equation shown in Extended Data Fig. 7b for each clonal mutation. The equation shown in Extended Data Fig. 7c is a rearrangement of that shown in Extended Data Fig. 7b for clonal mutations where CCF = 1. We take the mean of these values to provide a final estimate of ctDNA tumour purity per sample.

CCF calculation per mutation and subclone. For all mutations, the sample's ctDNA tumour purity, denoised VAF, multiplicity and total copy number at the mutated locus are used in the equation shown in Extended Data Fig. 7c to calculate an estimate of CCF for each mutation in a given plasma sample. The clone identities for each mutation are provided to ECLIPSE and should be calculated independently using standard methods, which leverage SNP coverage applicable to high purity samples[39–41]. The mean per-mutation CCF is used as a CCF estimate for each clone. Any CCF estimates > 1, presumed to represent noise, are limited to 1.

Poor-quality clone identification. Mutation clustering using standard methodologies is imperfect and will be fitted to the samples of higher purity used for cluster identification (usually matched tissue samples), excluding lower-purity samples that ECLIPSE is able to analyse using deep targeted sequencing. Erroneous clusters may not continue to track at similar CCFs in data from new samples. To identify such clusters, the distribution of ECLIPSE-calculated CCFs in each clone in a ctDNA sample are quantified using normalized s.d. values. The s.d. values can then be compared to the expected CCF distributions of high confidence clones, for example, clonal clusters in higher-purity plasma samples. In our data, we quantified the normalized s.d. of all clonal clusters in samples of greater than 5% purity and took the upper 95% confidence interval for this data calculated at 0.56. Subclonal clusters with normalized s.d. values for CCFs of >0.56 were considered to be of poor quality and were not considered for analysis. This identified 2.6% of clones in the TRACERx data as of poor quality. Expected CCF distributions will be highly dependent on the input data for ECLIPSE and should therefore be benchmarked on each dataset. A function in the ECLIPSE R package is provided to calculate an upper 95% CI of normalized s.d. values for CCFs in clonal clusters in high-purity samples, as was performed for this dataset.

Clone present calling. To determine whether each clone is present or absent from each sample (see the 'High-specificity subclone detection' section), the sum of expected background error is compared with the sum of the observed signal across all variants in the subclone with a one-sided Poisson test. Mutations with high error that should be excluded from these calculations can be specified.

Tumour present calling. To determine whether any tumour cells are present in each sample, the summed expected background error is compared with the summed observed signal across all variants tracked in the sample with a one-sided Poisson test. Mutations with high noise that should be excluded from these calculations can be specified.

Minimal detectable CCF estimation for each subclone. Determination of the CCF equivalent to the minimal number of supporting reads across all variants in a subclone that would be required for a significant call to be made (Poisson test, $P < 0.01$, see the 'High-specificity subclone detection' section).

Minimal detectable CCF estimation for an average subclone for each sample. Determination of the CCF equivalent to the minimal number of supporting reads across all variants in a representative subclone that would be required for a significant clone to be called as present (Poisson test, $P < 0.01$; see the 'High-specificity subclone detection' section). The background is taken as an average of the background error in all subclonal mutations tracked in a given sample and is representative for a subclone tracked by four mutations as default, the average number tracked in this dataset. This value enables comparisons of minimally detectable CCF limits across samples.

Minimal detectable purity estimation for each sample. Determination of the purity equivalent to the minimal number of supporting reads across all tracked clonal variants that would be required for a significant tumour to be called as present (Poisson test, $P < 0.01$).

Testing for the absence of a complete clonal sweep for each subclone. A subclone that is detected in high-purity samples used for mutation clustering may expand through a full clonal sweep later in the disease course. We would therefore expect to observe CCFs of 100%, indistinguishable from CCFs of clonal mutations after such an event. For each subclone in each sample, a Wilcoxon test is performed to compare the CCFs of each subclone to the CCFs of clonal mutations in the same sample. The resulting *P* value indicates whether there is significant evidence that the subclone is significantly below 100% CCF and is therefore present in only the minority of tumour cells, without a full clonal sweep.

Minimal detectable CCF estimates for each subclone

To quantify our limits of detection of CCF in each sample and subclone, ECLIPSE calculates the number of supporting reads for all mutations in each subclone that would be required for a positive clone detected call ($P < 0.01$ threshold) based on the number of expected background error reads using the qpois function in R (stats package, v.4.1.2). This value is then divided by the mean depth of all variants in a subclone to simulate a representative minimal detectable VAF for mutations in a given subclone and these values are input into the equation shown in Extended Data Fig. 7c to calculate the equivalent CCF, using an average of the mutation multiplicity and total copy number across all mutations in the given subclone and the ctDNA purity of the sample (see the 'Determination of 'tumour purity' in plasma' section, Supplementary Note). These minimally detectable CCF thresholds are highly dependent on the number of variants tracked in each subclone; therefore, to provide a single representative and comparable value for each plasma sample, we also simulated the minimal detectable CCF for a subclone containing four mutations, which is the median number of mutations tracked in each subclone in this study but can be altered as an argument to ECLIPSE. The minimal detectable number of supporting reads in these four mutations was estimated using the average background error profile of all subclonal mutations in a given sample.

High-specificity subclone detection

A similar approach to that for high-specificity MRD detection in ctDNA was undertaken for detection of subclones in this study, by estimating the background sequencing error in a TNC-specific manner leveraging non-mutated positions in the target regions of the sequencing library (see the 'MRD calling' section). These background error estimates were then provided to ECLIPSE. These background noise rates were multiplied by depth to calculate the expected number of background reads alternate at each mutated position. These expected background read counts were then summed for all variants in a clone and used as the background lambda for a Poisson test comparing the sum of the observed number of reads across the same mutations. A *P*-value threshold of 0.01 was chosen to call a clone present to match the threshold determined for MRD calling with in vitro spike-in experiments and the pilot cohort of patients comparing post-surgery samples to relapse status.

Application of ECLIPSE to the TRACERx cfDNA data

Inputs to ECLIPSE were prepared from the TRACERx 421 cfDNA and exome sequencing data as follows for all analyses unless otherwise specified. For inputs extracted from matched tissue exome sequencing data, all available samples were used, including from relapse tissue where possible, although all mutations in PSP panels were derived from surgical excised tissue samples only. Clonal versus subclonal status, cluster identities and multiplicity status were extracted using presence and absence informed clustering as previously described[31], which builds on the PyClone algorithm[41]. Total copy number in each tumour sample at each mutated locus was extracted as previously described[31]. Normal copy number was presumed to be two across the genome. For metrics calculated per sample, purity-adjusted averages (which were computed as the sum of the metric per sample, multiplied by the sample purity and divided by sum of all sample purities) were calculated across the whole tumour for input into ECLIPSE for multiplicity and total tumour copy number. The number of variant-supporting reads and depth in each cfDNA sample were calculated considering only unique reads with at least five supporting duplicates to minimize background error. Trinucleotide-specific error estimates were used as input to the background error per variant. 'Hard filtered' variants (those excluded from all ECLIPSE analyses) were those with 'failed filters' of 'primer_abundance_filter', 'primer_strand_bias', 'sequence_strand_bias', 'dro_cutoff' and 'dao_imbalance'. Moreover, 'MRD filtered' variants were those with 'failed filters' 'tnc_error_rate' where the background error was considered to be too high for inclusion in estimates of MRD (see the 'MRD calling algorithm' section) and were also excluded for estimates for clone presence or absence in ECLIPSE (see the 'Steps of ECLIPSE' section).

Validation of ECLIPSE CCFs versus tissue exome M-seq CCFs

To compare ECLIPSE-estimated CCFs to those estimated using validated methods applied to tissue sequencing data at a matched timepoint, we compared purity adjusted averages (see the 'Application of ECLIPSE to the TRACERx cfDNA data' section) of CCFs calculated using surgically excised tumour tissue for each subclonal cluster[31], a benchmarked variant of PyClone[41] to subclonal CCFs estimated in ECLIPSE (Extended Data Fig. 9a). The analysis was performed on high-subclone-sensitivity preoperative samples, which are defined as those with at least 0.1% clonal ctDNA levels. These were samples with an estimated minimally detectable CCF of at least 20% (see the power analysis in Extended Data Fig. 8a) comprising 61% of ctDNA positive preoperative samples from 67 patients. Although a formal method for CCF estimation in deep targeted sequencing data has not been previously published for comparison, we compared ECLIPSE to a VAF-only method for CCF estimation. In this method, which is naive to copy-number status, the mean VAF of each subclonal cluster is divided by the mean VAF of the clonal cluster in each sample (Extended Data Fig. 9b). This caused a consistent underestimation of CCF relative to estimates from tissue exome sequencing, driven by the higher average multiplicity of clonal mutations compared to subclonal mutations, which more commonly occur before large scale copy-number amplifications (for example, WGD), increasing mutation multiplicities of mutations that have already been accrued.

Validation of subclone detection rates using our data and ECLIPSE

To further investigate the sensitivity of subclone detection at different frequencies using ECLIPSE, we analysed data generated using in vitro spike-in experiments described in Extended Data Fig. 2. To generate these data, different mutation allele fractions were spiked into wild-type DNA and different total DNA amounts were inputted into our AMP PCR NGS assay, including 12 replicates for each spike-in mutation fraction and input amount combination. In total, this comprised 398 spike-in samples, each with 50 spiked in mutations, that were then subject to our AMP PCR NGS pipeline, identical to that applied to our plasma-derived cell free DNA samples. We subsampled mutations from each of these spike-in experiments in silico to represent subclones with 1, 2, 4, 10 and 20 mutations (a median of 4 mutations were tracked per subclone in our TRACERx ctDNA panels). Each of these subclones was combined in silico with data from spike-in mutations at higher mutant allele fractions to represent clonal mutations. This enabled us to construct in silico subclones with various CCFs (determined by the ratio of spiked in mutant allele fraction of the subclonal mutations to the spiked in mutant allele fraction of the clonal mutations), across various clonal ctDNA levels (the spiked in mutant allele fraction of the clonal mutations) across a range of total DNA inputs to the assay. Although these data derive from mixing mutations together from different experiments in silico, the concentrations of DNA are known from ground truth; thus, these mixtures provide a deeper level of validation, controlling for various sources of noise in the assay and providing technical replicates. In total, we constructed 76,263 subclones from these data that varied in CCF, clonal ctDNA level, number of mutations per subclone and assay DNA input amount. We ran these data through ECLIPSE using background noise estimates from the same libraries to determine how the rate of subclone detection varies with these four parameters. We focused on the lower DNA inputs (≤10 ng) as the greatest variety of allele fractions were spiked in for these inputs, enabling construction of a wider range of CCFs, and these samples represented the most challenging scenarios for subclone detection. We calculated the fraction of subclones detected for each experimental replicate at each specified clonal ctDNA level and at each CCF. We then used the resulting distribution of detection rates across experimental replicates, for each clonal ctDNA level and CCF, to calculate 95% CIs.

Clonal illusion analyses

For analysis of clonal illusion, we reran ECLIPSE for each TRACERx patient considering only data from a single randomly selected tumour sample to simulate a clinical biopsy, including multiplicity and total copy-number estimates. The clonal status of each mutation was recalculated using a 90% CCF threshold in the selected region and only mutation-specific, rather than clone-specific, estimates of CCF were analysed, which removed the requirement for mutation clustering and clone identification. To analyse clonal illusion, all mutations that would be considered to be clonal in the randomly selected region were split by their clonal status when considering all TRACERx regions. Such mutations were therefore either truly clonal in all regions (labelled clonal) or were in fact subclonal when other tumour regions were considered and therefore harboured clonal illusion in the randomly selected region. ECLIPSE estimates (using only data from the randomly selected region as described) were then displayed for these two mutation groups in Fig. 4a. To determine sensitivity and specificity using ROC analysis of clonal illusion detection, all apparently clonal mutations (>90% CCF) in the randomly selected region were used with the ROCIT R package (v.2.1.1) with scores inputted as the mutation-specific single-region ECLIPSE CCF estimates and final classes considered as the clonal or clonal illusion status leveraging all tumour regions in TRACERx.

Longitudinal depictions of clonal evolution in cfDNA and tissue

Representations of clonal evolution over time were depicted using the ECLIPSE plasma CCFs per subclone, the subclonal CCFs in matched tissue samples extracted either at surgery and the phylogenetic subclone relationships calculated from tissue multiregional exome sequencing as described previously[31]. ECLIPSE plasma subclone dynamics were plotted using modified code from the fishPlot R package (v.0.5)[42] and clonal structure of tissue samples was plotted using an R package developed in-house called cloneMap (v.1.0)[43] distributed on GitHub (https://github.com/amf71/cloneMap). Only clones with at least one cfDNA-tracked mutation that was not hard filtered (see the 'Application of ECLIPSE to the TRACERx cfDNA data' section) in all samples were shown in the ctDNA and tissue clonality representations and the phylogenetic trees. Clonal dynamics in cfDNA were represented by ctDNA purity for each clone, which was calculated by multiplying the CCF of each clone by the ctDNA tumour purity of each cfDNA sample, therefore presenting the proportion of cfDNA-derived cells (including normal haematopoietic cells) that belong to a specific subclone. In total, 44 patients who relapsed from their disease excised at initial surgery and for whom phylogenetic trees were available from tissue exome sequencing were depicted in Fig. 5 and Supplementary Fig. 1. The CCF of a parent clone was maximally limited to the sum of the CCFs of its daughter subclones. In Fig. 5c, CCFs, rather than ctDNA purities, are plotted for each clone, as the purity/ctDNA fraction in this patient varied over several orders of magnitude. Sample purities are depicted in this case as grey circles below the CCFs.

Definition and detection of clonal sweeps at relapse

Subclones undergoing a clonal sweep were those that expanded after surgery, when they were first detected in tissue WES, increasing to 100% CCF, that is, such previously subclonal mutations were now estimated to be present in every tumour cell and parallel subclonal lineages were estimated to have been extinguished. To call instances of a clonal sweep, ECLIPSE performs a Wilcoxon test comparing the CCF of all mutations in a given subclone to the clonal mutation in each sample. The resulting P value indicates the probability that the subclone has undergone a clonal sweep with a null hypothesis of a clonal sweep being present. We considered a clonal sweep present when this P value was greater than 0.05 and absolute mean subclone CCF was at least 90%. For each patient, the latest possible timepoint with high subclone sensitivity (that is, a clonal ctDNA level of at least 0.1%) was used to determine clonal sweeps at relapse. To estimate how these clonal sweeps at relapse modified the tumour trunk, we added all mutations and neoantigen in relapse clonal sweep subclones (including those clustered together in exome sequencing but not tracked in cfDNA) to the clonal mutations for re-estimation of clonal tumour mutational burden and clonal neoantigen burden at relapse. All subclones tracked by PSPs, including those that may have been specific to surgically excised lymph nodes or ipsilateral intrapulmonary metastases, were included in this analysis.

Determination of phylogenetic metastatic dissemination class at relapse

Phylogenetic metastatic dissemination classes at relapse were determined separately using either relapse tissue or post-operative cfDNA for each patient with relapse in this study, where relapse tissue and/or a high-subclone-sensitivity postoperative sample (>0.1% clonal ctDNA level) was available. Our companion article[31] has focused on metastatic disseminations estimated from primary tumour tissue including disseminations detected at surgery to local lymph nodes (also excised at initial surgery). Metastatic dissemination to excised local lymph nodes cannot be estimated in cfDNA alone, as preoperative ctDNA may derive from either metastatic lymph nodes or the primary tumour. For tissue-based metastatic dissemination calls at relapse, relapse seeding primary tumour subclones from Al Bakir et al.[31] that were tracked by PSPs were considered. These clones were used to determine whether a single clone or multiple primary tumour clones seeded the tissue relapse (monoclonal and polyclonal dissemination, respectively). Using the phylogenetic tree in polyclonal cases, we also determined whether clones were directly descended from one-another in the same clade (polyclonal monophyletic) or whether there is branching between the disseminating clones into different clades (polyclonal polyphyletic). For metastatic dissemination calls at relapse based on post-operative cfDNA, the number of relapse seeding clones was determined de novo without reference to the relapse tissue samples. If all primary tumour subclones detected in postoperative ctDNA were direct descendants in the phylogenetic tree and were present at 100% CCF, the relapse was considered to be monoclonal. If any primary tumour subclone was present at significantly less than 100% (using a Wilcoxon test comparing clonal cluster CCFs with each CCFs in each subclonal cluster, $P < 0.05$, and also requiring a mean subclone CCF < 90%), then the metastatic dissemination at relapse was considered to be polyclonal. In polyclonal cases, if the subclones present at relapse were direct descendants of one-another, the metastatic dissemination at relapse was considered to be polyclonal monophyletic and, if they were branched into separate clades, they were considered to be polyclonal polyphyletic. Metastasis-unique subclones tracked by PSPs in surgically excised lymph nodes or intrapulmonary metastases that were also present at relapse were not considered when defining primary tumour to relapse metastatic dissemination patterns, as they represent metastasis-to-metastasis seeding rather than primary-to-metastasis seeding. For example, CRUK0620 is determined to have a monoclonal metastatic dissemination pattern at relapse, despite having multiple subclones and branches present in post-operative ctDNA, as only one of those subclones (subclone D on the phylogenetic tree) is present in the primary tumour and other ctDNA relapse clones were detected only within surgically excised metastases (an ipsilateral intrapulmonary metastasis and several lymph nodes). This definition of metastatic dissemination as primary to metastasis dissemination, rather than surgically excised tumour to recurrence dissemination, is consistent with our companion article and many analyses in the literature[44]. We did not find a significant difference between the number of tracked mutations in post-operative plasma subclones that were detected compared with those that were undetected (Wilcoxon test $P = 0.13$, median number of variants tracked = 4 in both cases) suggesting power of detection did not strongly influence which clones were detected in relapse cfDNA.

Quantifying chromosomal instability in CRUK0050

At the last plasma sample timepoint in CRUK0050, a multimodal distribution of clonal mutation VAFs was observed (Fig. 5d) where each mode represented a set of mutations with a similar average multiplicity across the tumour. To assign each clonal mutation to a VAF cluster, the mclustBIC and then Mclust functions from the mclust package (v.5.4.7) were used. In this case, four VAF clusters were identified. The mutations in the lowest VAF cluster had an average VAF of 1.2% and mutations in the second lowest VAF cluster had an average VAF of 12.1%. If the lowest cluster represented mutations with a multiplicity of 1, the large majority of mutations in the remaining three clusters would therefore be presented at very high multiplicities (>10) given their >10-fold higher average VAF, which would represent a biologically implausible amount of allele duplication across the genome, equivalent to five compounded WGD events. A more plausible explanation of these data is that the lowest cluster represents mutations with a multiplicity of 0 in a new subclonal population that has expanded at the final timepoint to a CCF > 80%. Consistent with this, the mutations in the lowest VAF cluster were present at very similar VAFs in the previous plasma sample timepoint, consistent with the notion that these mutations remained only in those same tumour cells at the final timepoint, but not in the expanded subpopulation. The second lowest VAF cluster also contained 100% of the mutations that were associated with a multiplicity of 1 in the tumour tissue WES data. We therefore assigned the second lowest VAF cluster a multiplicity of 1, the second highest cluster (average VAF of 25%) a multiplicity of 2 and the highest VAF mutation (KRAS G12 variant, VAF = 84%) a multiplicity > 2. These mutation multiplicities were compared to the integer multiplicity estimates in surgically excised tissue WES to determine which mutations had undergone a change in copy number, which was the case for 59 out of 130 clonal mutations.

Designing AMP–MRD enrichment panels

Tumour-informed personalized AMP–MRD enrichment panels were designed for 197 TRACERx patients. A median of 50 variants per panel (range 0 to 50) were chosen using the ArcherDx panel design algorithm (v.0.1) and a median of 150 variants (range, 34 to 153) were chosen using variants selected from TRACERx multiregion exome sequencing data derived from early-stage NSCLC resections (including primary tumour, lymph node metastases or ipsilateral intrapulmonary metastasis if applicable). Owing to alterations in our TRACERx exome sequencing pipeline between panel design (2019 to 2020) and final analysis, a small fraction of mutations (3%) targeted by PSPs was no longer called with high confidence in tissue exome sequencing data. These mutations were included in MRD analyses (to align with the originally intended analysis approach plus prevent any possible bias conferred by manually removing these variants from consideration by the MRD caller) but were excluded from analyses of clonal structure. For Archer variant selection, WES sequencing data from the highest purity tumour region and from the paired germline DNA were used.

The algorithm then determined which variants can be targeted using an ArcherDX AMP panel and, from this set of variants, the 50 most informative mutations were targeted on the basis of the following criteria: the quality of the primers targeting the variant (to ensure high sequencing coverage of the target variant), predicted ER for the variant in error corrected bins and mappability. The predicted ER for each variant is based on an analysis of AMP cfDNA libraries sequenced on a NovaSeq instrument. This ER analysis was performed by running targeted variant calling on every possible single-nucleotide variant (SNV) in a set of Archer LiquidPlex cfDNA libraries. The TRACERx variants were selected from variants called in surgically excised tumour samples using the TRACERx WES pipeline[2]. SNVs were ranked based on their (1) driver designation, (2) TNC as described above, (3) mean mutation allele count. All SNVs were categorized as neoantigen, clonal or subclonal. Up to 50 variants were picked from each category. Neoantigens were additionally ranked by binding affinity[45]. Subclonal mutations were picked to represent all phylogenetic mutation clusters, picking up an equal number of mutations from each cluster when possible, up to a total of 50 maximum. Finally, 50 clonal variants were picked. If the sum of subclonal and neoantigen mutations was less than 100, the difference was picked from the list of clonal mutations.

Each personalized enrichment panel also contained 90 primers targeting 45 common SNPs. During analyses, the zygosity of these SNPs in a cfDNA library is compared to their zygosity in the whole-exome sequencing data for that patient to confirm that a sample swap did not occur. In addition, the coverage provided by these primers helps in establishing the background PCR and sequencing ER for a library. These 45 SNPs were chosen based on being present in each Gnomad subpopulation at a frequency of 25–75% to maximize use in detecting sample swaps.

ArcherDX variant choosing and panel design deviated from the standard workflow in two cases. In the case of the pilot sample CRUK0297, the tumour and non-tumour samples used in design were not properly matched and rare germline variants appeared to be tumour specific as a result. The ArcherDX variant choices in this panel included many germline variants. For this reason, the cfDNA libraries for CRUK0297 underwent manual blanking of the germline targeted variants to facilitate the use of these samples in the pilot patient analyses. All subsequent ArcherDX panel designs included a quality-control step to confirm that the common population polymorphisms in the tumour and non-tumour samples matched. The second case in which panel design deviated from the standard ArcherDX workflow occurred in the design of CRUK0296. The variant call data for a tumour–normal tissue pair could not be obtained in the standard format for this patient. In this case, the standard variant caller could not produce a result so the variant caller VarDict[46] was used and data were not available for the non-tumour sample in the standard format. As a result, two germline variants (chr6:31118898:A:T and chr16:70928307:C:A) were targeted. These two variants were removed from consideration in making the MRD call automatically by the Outlier Filter in every library prepared with this panel (see the ‘Library-specific MRD calling’ section above) but were kept in all analyses and not manually blanked. Two patients lacked Archer-picked variants (CRUK0157 and CRUK0227)—CRUK0157 as the exome data could not be processed by the Archer variant picking pipeline and CRUK0227 owing to an error during PSP primer ordering.

Neoantigen pipeline

HLAHD was used to determine the patient-specific HLA composition. 9–11mer peptides containing non-synonymous mutations coupled with patient-specific HLA were used as input to NetMHCPan4.1. A Rnk_EL < 0.5 was used to determine strong binder peptides.

Analytical validation experiments

For experiment LOD1, 634 samples of fragmented DNA with a known SNP profile (Genome in a Bottle DNA, NA24385) were added to a background of four other fragmented Genome in a Bottle inputs (NA24149, NA24631, NA24694 and NA24695). Six AMP enrichment panels were generated targeting 50 SNPs that were heterozygous in NA24385 and absent from the other four cell lines. To generate contrived samples, NA24385 DNA was spiked into a background of the other four samples at ratios of 0.006% to 0.2% by mass to target VAFs ranging from 0.003% to 0.1% allele fraction (AF), as heterozygous variants are present at 50% in the neat NA24538. As part of the same dilution series, admixtures with target allele frequencies of 1%, 5% and 10% were made. These mixtures were used as input for AMP library preparation to confirm that mixing based on mass achieved the desired target allele levels. The spike-in variant level was measured in these higher AF libraries by adding the number of deep alternate reads across the targeted SNPs and dividing by the total coverage of all deep reads across targeted SNPs. This analysis confirmed that the spike-ins achieved the targeted AFs. Fragmented DNA inputs from 2 ng to 80 ng were used in the experiment to reflect the range of DNA inputs encountered in a clinical setting. Overall, 559 out of 634 samples were deemed to be evaluable for LOD1 analysis (62 samples failed because of incorrect DNA input used, determined by on-target read per primer per ng input of <30 or >400; 8 samples failed because they had less than 10 million reads; and 5 samples failed due to potential duplicate libraries). Clinical samples were used in validation of AMP MRD (LOD2) and were prepared using a similar method to the Genome in a Bottle mixtures. Whole-exome sequencing data from four patients were used to design patient-specific panels with the ArcherDx panel design algorithm containing 50 SNVs. The panels were used to prepare libraries using cfDNA from each patient and the overall tumour variant AF for each sample was calculated by adding the total number of deep unique reads containing a targeted tumour-specific variant and dividing by the sum of the deep unique coverage across all targeted tumour variants. All four patient cfDNA libraries had a total AF of >1%. A single mixture was made using cfDNA from healthy donors and was used to dilute the patient cfDNA. These dilutions were performed as a serial dilution. First a dilution was made targeting a 1% total AF and libraries were prepared using this mixture. The total AF was measured for this sample and a dilution correction factor was calculated to account for differences in conversion efficiency between the background cfDNA. For example, if a 1% AF was targeted and an AF of 1.3% was observed then this would indicate that the patient cfDNA is more efficiently converted to library than the background and more background DNA would need to be used. Mixtures were then made to achieve AFs of 0.1%, 0.05%, 0.01%, 0.008% and 0.005%. A total of 100 libraries were prepared at 5 AFs and 3 input masses. In total, 48 blank samples (DNA donated from 24 healthy donors) were analysed to assess assay specificity. Panel-observed AF values were calculated by taking the number of deep alternate reads noted across the AMP panel, removing estimated background error and dividing by deep depth across the panel. For experiment LOD3, an AMP PSP was generated targeting 300 heterozygous SNPs in Genome in a Bottle product HG002. HG002 was diluted into a background mixture of HG003, HG005, HG006 and HG007 at multiple dilution levels such that heterozygous variants located in HG002 were present at final AFs of 0%, 0.003%, 0.005%, 0.006%, 0.01%, 0.03% and 0.05% and 0.1%. Using stocks of these contrived input materials, 10 ng was input into library preparation. Two libraries were prepared at AFs from 0% to 0.05% and a single library was prepared at 0.1% AF using the 300-variant AMP panel. In silico subsampling was performed on the 15 libraries. Nine in silico panels were generated for each library (3 targeting 200 variants, 3 targeting 100 variants and 3 targeting 50 variants) and MRD caller results evaluated alongside the 300-variant PSP result (overall 150 results were generated from the 15 libraries). For assay sensitivities at specific spike-in categories, Clopper–Pearson binomial two-sided 95% CIs were calculated in Extended Data Fig. 2e,f using the R package DescTools (v.0.99.44)[47] and the function BinomCI.

Simulation analysis to assess specificity

The trinucleotide context of tumour-specific SNVs within each TRACERx AMP-MRD pilot cohort panel was assessed. On the basis of these data, mock tumour signatures (genomic positions covered by the enrichment primers with positions of similar expected ERs of the targeted SNVs) were generated. A mock variant was added to a mock signature if the following criteria were met: it is bidirectionally covered by primers intended for MRD detection; it contained the same TNC-group ER as the true MRD variant that it is replacing; it was not a known population SNP variant as dictated by Ensemble's Variant Effect Predictor v.94.5; had a error-corrected coverage delta of no more than 2,000 compared with the true MRD variant; and was not used within any other mock tumour signature, including itself. Thus, the resulting mock signatures targeted bases that are not mutated in the primary tumour and any positive MRD call from these mock signatures was by default a false positive. In total, 3,157 mock signatures across 91 pilot cfDNA libraries were examined for MRD-positive calls A simulated ctDNA level was estimated for each sample by taking the number of deep alternate reads noted across the mock signature, removing estimated background error and dividing by deep depth across the mock signature. Data from this simulation are provided in Supplementary Table 18.

Digital droplet PCR orthogonal validation

Digital droplet polymerase chain reaction (ddPCR) orthogonal analyses were performed in 30 preoperative plasma samples from TRACERx patients who also had preoperative plasma analysed by the AMP personalized tumour informed approach and 8 negative controls (preoperative plasma from patients diagnosed postoperatively with non-malignant disease). TRACERx patients were selected as having clonal driver mutations that could be targeted by a single ddPCR assay. Clonal driver mutations targeted included *KRAS G12R*, *G12D*, *G12V*, *G12S*, *G12A*, *G12C* and *EGFR L858R*. The ddPCR assays used were SAGAsafe assays (SAGA diagnostics) and had been designed and developed on the BioRad QX200 Droplet Digital PCR system. ddPCR analyses were performed at SAGA, SAGA received plasma (median 4.8 mls; range, 2.5–5.2 mls). cfDNA was extracted using the QiaAMP MinElute ccfDNA Midi Kit (Qiagen). cfDNA was eluted in 40 µl of buffer EB. The entirety of cfDNA material was input in each case and ddPCR analyses were run in four replicate reaction wells per sample. All eight negative controls (each assay tested once, *KRAS G12A* tested twice) exhibited no mutant droplets detectable in control cfDNA; Supplementary Table 4).

Transcriptional data analyses

Gene-level transcription analysis was performed using edgeR (v.3.36.0)[48] and limma (v.3.50.3)[49]. The analysis included 101 tumour regions sampled from 34 patients positive for ctDNA and 62 tumour regions sampled from 28 patients with the biological ctDNA low-shedder classification. The analysis took into account 18,876 protein-coding genes based on the HGNC database, retrieved on 4 March 2022. Genes with insufficient expression levels (count < 30) were filtered out and effective library sizes were calculated using the trimmed mean of *M* values method. Count data were then transformed to $\log_2$[counts per million] (log[CPM]). Before linear modelling, a weight per observation was calculated on the basis of the association between mean and variance. To take into account the association between tumour regions within patients, a per-patient consensus correlation was computed. On the basis of the logCPM table, the within-patient correlations and the ctDNA detection status, a linear model was fitted. A contrast matrix comparing ctDNA positives and biological ctDNA low-shedders was constructed alongside with the associated coefficients and standard errors, and the empirical Bayes method (eBayes function from limma, v.3.50.3)[49] was used to calculate the moderated *t*-statistics of differential expression. The resulting gene-level, two-tailed *P* values were adjusted for multiple testing using the Benjamini–Hochberg (FDR) method. Genes were noted as significantly differentially enriched if their adjusted *P* value was below 0.05.

The set of significantly overexpressed genes per detection category ($n = 876$ for ctDNA positives, $n = 883$ for biological ctDNA low-shedders) was used to calculate Reactome pathway enrichment (ReactomePA, v.1.38.0)[50]. The resulting *P* values were FDR-corrected and an adjusted *P* value cut-off of 0.05 was used.

Moreover, pathway enrichment with respect to the Hallmark gene sets from the msigDB database was investigated. Pathway enrichment analysis was performed on log[CPM] data including 17,815 protein-coding genes using Gene Set Variation Analysis (GSVA, v.1.42.0)[19]. The fold change of GSVA enrichment scores comparing 101 tumour regions from 34 ctDNA positives and 62 tumour regions from 28 biological ctDNA low-shedders was calculated using the estimated marginal means (rstatix, v.0.7.1)[51] method, using a linear mixed-effects (lmerTest, v.3.1-3)[52] model to take into account the patient–tumour region associations, treating detection status as a fixed effect and patient ID as a random effect. The resulting pathway-level *P* values were FDR-corrected for multiple testing.

Mutation analyses

Driver mutations in 181 genes from 70 patients positive for ctDNA or with the biological ctDNA low-shedder classification (39 ctDNA positive, 31 biological ctDNA low-shedder) were included in the analysis. Clonality was determined based on part of the TRACERx WES pipeline. If a patient carried multiple mutations in the same gene with differing clonality, the clonal state was kept. In the gene-level analysis, the top 14 frequently mutated genes were considered. Genes were assigned to pathways as described previously[53]. A Fisher's exact test was conducted in a two-tailed manner to compare the number of patients positive for ctDNA and patients with the ctDNA low-shedder classification carrying alterations in the frequently mutated genes. The resulting *P* values were corrected using the FDR method.

Chromosomal instability analyses

Copy-number data, including allele-specific copy numbers and purity estimates, were derived from the TRACERx WES pipeline and were available for 245 tumour and lymph node regions collected from 63 patients positive for ctDNA or with the biological ctDNA low-shedder classification (166 regions from 35 ctDNA positive and 79 regions from 28 ctDNA negative). Cytoband analysis was conducted using GISTIC (v.2.0)[24], which takes one sample per patient as input. To investigate genomic regions of recurrent gains and losses, we constructed the single-sample copy-number profile for each tumour by selecting the maximum (for gains) or minimum (for losses) ploidy-corrected total copy number per segment across the genome. A GSD of 0.5 was used as a threshold for significance cut-off. Cytobands were overlapped with output from GISTIC2.0 to get a mean GISTIC score for each cytoband. FLOH and wGII were analysed at the region level (548 tumour and lymph node regions from 137 patients, 166 regions from 35 ctDNA-positive adenocarcinomas, 79 regions from 28 biological low-shedder adenocarcinomas and 303 regions from 74 non-adenocarcinomas). Comparing the chromosomal instability metrics between ctDNA positive and biological ctDNA low-shedder adenocarcinomas and non-adenocarcinomas was performed using a linear mixed model, taking into account the within-sample associations. Pairwise comparisons were made using the estimated marginal means method and *P* values were FDR-adjusted. Tumour regions were considered to be WGD if the fraction of the genome with major allele ≥ copy number 2 was greater than 50%, as described previously[54]. Tumours were considered to have WGD if any single region had a WGD event. WGD data were available for 63 patients with lung adenocarcinoma (28 biological ctDNA low-shedder, 35 ctDNA positive).

Purity analysis

Using 245 regions from 63 patients (166 regions from 35 ctDNA positive and 79 regions from 28 ctDNA negative), we performed an estimated marginal means analysis incorporating a linear mixed model approach to account for the within-sample associations. The analysis compared ctDNA positive with biological ctDNA low-shedder samples.

ORACLE analysis

ORACLE scores were calculated by using a previously described method[21], including 196 tumour and lymph node regions from 77 patients positive for ctDNA or in the ctDNA low-shedder category (109 regions from 35 ctDNA positive, 87 regions from 42 ctDNA low-shedder). Pairwise comparisons between the ctDNA shedder and ctDNA low-shedder samples were made using the estimated marginal means method with a linear mixed-effects model to account for the within-patient associations between tumour regions.

Volume adjustment

Biological low-shedder samples were excluded from the volume-adjusted analysis if their size fell in the lowest quartile size range (<6,042.544 mm^3). Transcriptomic and GISTIC analyses were repeated using the volume-adjusted dataset as described above. Taking into account the significantly overexpressed genes and significant cytobands in both datasets, Venn diagrams were constructed for comparison and the Jaccard Similarity Index was calculated to assess the statistical significance of the overlap. The similarity coefficient calculations were performed using the jaccard R package (v.0.1.0)[55], and the corresponding *P* values were computed using the exact method. Venn diagram visualizations were created using eulerr (v.6.1.1)[56] and ggplotify (v.0.1.0)[57].

Clonal mutation ctDNA levels

Mutations that were defined as clonal, either by PyClone clustering as described in our companion manuscript[28], or (in the absence of PyClone data) that were present in every primary tumour tissue region analysed (ITH state = 1), and that were unfiltered by the MRD caller, were used in clonal mutation ctDNA-level estimations. For each mutation, the MRD caller estimated the trinucleotide ER associated with that mutation and the coverage of that mutation was used to estimate the number of expected error-controlled reads we would observe due to error. Clonal mutation ctDNA level was then summarized as the total number of error-corrected reads across selected mutations, minus the expected error across these positions (rounded down to the nearest whole integer) divided by total clonal deep coverage. If the clonal ctDNA level was <0% (where background error was higher than observed variant DNA signal), it was assigned 0%. In two ctDNA-positive samples, clonal ctDNA levels were measured at 0% due to mutations driving ctDNA-positive status not being assigned a clonal status by the TRACERx pipeline (CRUK0296 sample 144717 and CRUK0039 sample 117025).

Identifying probable technical negative and low-shedding adenocarcinomas

We generated a linear regression model (using the stats R package, function lm) where log$_{10}$-transformed tumour volume and histology was used to predict log$_{10}$-transformed clonal ctDNA level in 96 ctDNA-positive non-pilot NSCLCs analysed in this cohort. We used this model to predict clonal ctDNA levels in 47 evaluable adenocarcinomas negative for ctDNA. We tested the ability of this model to predict clonal mutation levels in eight independent ctDNA positive adenocarcinomas with volume data available analysed in our previous work using a separate assay[7]. In this test set, 6 out of 8 (75%) adenocarcinomas evaluated had mean clonal mutation levels above the lower 95% CI of the model estimation. We calculated the minimal detectable clonal ctDNA level (MDCL) in the 47 ctDNA-negative adenocarcinomas by taking the minimum DAOs needed to make a call in patient cfDNA samples and subtracting the estimated deep alternate reads that would occur due to noise in the panel (rounded down to the nearest whole integer). The resulting number was the number of clonal deep alternate reads needed to make a ctDNA positive call (we conservatively assumed that all real deep alternate reads will be clonal). We divided this number by the clonal deep depth across the panel to calculate the minimum clonal ctDNA level that must be exceeded to make a call and called this value MDCL. Using the above linear model, we classified cases as probable technical negatives if the lower 95% CI for predicted clonal ctDNA level was below MDCL and as probable low-shedders if the lower 95% CI for predicted clonal ctDNA level was above MDCL.

Survival analyses

OS (events were death from any cause, outcome predefined in TRACERx protocol)[2], FFR (events were lung cancer recurrence, patients disease-free or experiencing second-primary or death were right censored at last follow-up) and post-relapse survival (time from recurrence to death from any cause) analyses were performed. In total, 169 out of 187 non-pilot cohort patients were evaluated for survival analyses shown in Fig. 1 and Extended Data Fig. 43 (5 patients were excluded as they died within 30 days of surgery—CRUK0115, 0196, 0312, 0487 and 0681; 4 patients were excluded as they had confirmed unresected disease after surgery—CRUK0230, 0234, 0291 and 0387; and 9 patients with synchronous primaries were excluded given the emphasis on associations with tumour histology). For Extended Data Fig. 6d,e, patients with synchronous primaries were included in landmark survival analyses as tumour histology was not considered in survival analysis. R packages survival (v.3.2-13)[58], survivalAnalysis (v.0.3.0)[59] and survminer (v.0.4.9)[60] were used to generate HRs, forest plots, 2-year survival data and Cox regression models in the paper. Differences in OS between metastatic dissemination classes at relapse were analysed using Cox proportional hazard models from the date of study registration and from the date of MRD detection. A multivariable Cox proportional hazard model, including maximum relapse ctDNA level, which is known to co-correlate with tumour burden and power for subclone detection, was used to account for this confounder relative to OS from the date of study registration.

Lead time analyses

Lead time was defined as the time from first postoperative ctDNA detection to radiologically confirmed clinical relapse. For lead time calculations, we analysed patients with NSCLC relapse and assigned patients without postoperative ctDNA detection or with initial detection after clinical relapse lead times of 0 days. We excluded incompletely resected patients ($n = 4$), patients with no ctDNA sampling before clinical recurrence ($n = 3$, CRUK0516, 0557 and 0640) and pilot-patients ($n = 5$) from these analyses.

Statistical data analysis

No statistical methods were used to predetermine sample size. Analysis was performed in the R statistical environment (v.4.1.2)[61]. For input/output operations and general data manipulations, the R packages tidyverse (v.1.3.2)[62], data.table (v.1.14.6)[63], readxl (v.1.4.1)[64], fst (v.0.9.8)[65] and qusage (v.2.28.0)[66–68] were used. All statistical tests were two-sided unless otherwise stated. For chi-squared analyses the R function chisq.test was used and for Wilcoxon rank sum tests the R function wilcox.test was used. For assay performance analyses, positive predictive value was calculated as all true positive results divided by the sum of true-positive and false-positive results; negative predictive value was calculated as all true-negative results divided by the sum of false-negative plus true-negative results; sensitivity was calculated as true-positive results divided by the sum of true-positive and false-negative results; specificity as true negatives divided by the sum of true negatives and false positives. For generation of heat maps, the R package ComplexHeatmap (v.2.11.1)[69] was used. For general visualization purposes, R packages ggplot2 (v.3.3.5)[70], ggpubr (v.0.4)[71], ggrepel (v.0.9.2)[72], ggbeeswarm (v.0.6.0)[73], scales (v.1.2.1.)[74], ggforce (v.0.4.1)[75] and cowplot (v.1.1.1)[76] were used. For plotting paired data, ggpubr (v.0.4)[71] was used.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The cfDNA sequencing files, RNA-seq data and multiregion tumour exome sequencing data (in each case from the TRACERx study) used or analysed during this study have been deposited at the European Genome–phenome Archive (EGA), hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under accession codes EGAS00001006494, EGAS00001006517 and EGAS00001006494 and is under controlled access owing to the nature of the data and commercial partnership arrangements. Details on how to apply for access are available on the linked page.

Code availability

ECLIPSE is available as an R package to install from github (https://github.com/amf71/ECLIPSE) which is only available for academic non-commercial research purposes. Code used to produce the figures in this paper is available on request.

39. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
40. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
41. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
42. Miller, C. A. et al. Visualizing tumor evolution with the fishplot package for R. *BMC Genom.* **17**, 880 (2016).
43. Frankell, A. M., Colliver, E., Mcgranahan, N. & Swanton, C. cloneMap: a R package to visualise clonal heterogeneity. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.26.501523 (2022).
44. Birkbak, N. J. & Mcgranahan, N. Cancer genome evolutionary trajectories in metastasis. *Cancer Cell* **37**, 8–19 (2020).
45. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
46. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
47. Signorell, A., Aho, K., Alfons, A., Anderegg, N. & Aragon, T. DescTools: tools for descriptive statistics. R package version 0.99 (2023).
48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
49. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
50. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
51. Kassambara, A. rstatix: pipe-friendly framework for basic statistical tests. R package version 0.7.1 (2022).
52. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
53. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
54. Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
55. Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinform.* **20**, 644 (2019).
56. Larsson, J. eulerr: area-proportional Euler and Venn diagrams with ellipses. R package version 7.0.0 (2022).
57. Yu, G. ggplotify: convert plot to 'grob' or 'ggplot' object. R package version 0.1.0 (2021).
58. Therneau, T. M. survival: a package for survival analysis in R. R package version v.3.2-13 https://CRAN.R-project.org/package=survival (2021).
59. Wiesweg, M. survivalAnalysis: high-level interface for survival analysis and associated plots. R package version 0.3.0 https://CRAN.R-project.org/package=survivalAnalysis (2022).
60. Kassambara, A., Kosinski, M. & Biecek, P. survminer: drawing survival curves using 'ggplot2'. R package version 0.4.9 https://CRAN.R-project.org/package=survminer (2021).
61. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, 2021).
62. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
63. Dowle, M. et al. data.table: extension of 'data.frame'. R package version 1.14.6 https://CRAN.R-project.org/package=data.table (2022).
64. Wickham, H. et al. readxl: read excel files. R package version 1.4.1 https://CRAN.R-project.org/package=readxl (2022).
65. Klik, M. fst: lightning fast serialization of data frames. R package version 0.9.8 https://CRAN.R-project.org/package=fst (2022).
66. Yaari, G., Bolen, C. R., Thakar, J. & Kleinstein, S. H. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Res.* **41**, e170 (2013).
67. Turner, J. A., Bolen, C. R. & Blankenship, D. M. Quantitative gene set analysis generalized for repeated measures, confounder adjustment, and continuous covariates. *BMC Bioinform.* **16**, 272 (2015).
68. Meng, H., Yaari, G., Bolen, C. R., Avey, S. & Kleinstein, S. H. Gene set meta-analysis with quantitative set analysis for gene expression (QuSAGE). *PLoS Comput. Biol.* **15**, e1006899 (2019).
69. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
70. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
71. Kassambara, A. ggpubr: 'ggplot2' based publication ready plots. R package version 3.3.5 https://CRAN.R-project.org/package=ggpubr (2020).
72. Slowikowski, K. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. R package version 0.9.2 https://CRAN.R-project.org/package=ggrepel (2022).
73. Clarke, E. ggbeeswarm: categorical scatter (violin point) plots. R package version 0.7.1 https://CRAN.R-project.org/package=ggbeeswarm (2022).
74. Wickham, H. et al. scales: scale functions for visualization. R package version 1.2.1 https://CRAN.R-project.org/package=scales (2022).
75. Pedersen, T. L. ggforce: accelerating 'ggplot2'. R package version 0.4.1 https://CRAN.R-project.org/package=ggforce (2022).
76. Wilke, C. O. cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1 https://CRAN.R-project.org/package=cowplot (2020).
77. Lakatos, E. et al. LiquidCNA: tracking subclonal evolution from longitudinal liquid biopsies using somatic copy number alterations. *iScience* **24**, 102889 (2021).

Acknowledgements The TRACERx study (ClinicalTrials.gov: NCT01888601) is sponsored by University College London (UCL/12/0279) and has been approved by an independent Research Ethics Committee (13/LO/1546). TRACERx is funded by Cancer Research UK (C11496/A17786) and is coordinated through the Cancer Research UK and UCL Cancer Trials Centre, which has a core grant from CRUK (C444/A15953). We thank the patients and relatives who participated in the TRACERx study, and all site personnel, investigators, funders and industry partners who supported the generation of the data within this study. In particular, we acknowledge the support of staff at the Scientific Computing, the Advanced Sequencing Facility and Experimental Histopathology departments at the Francis Crick Institute. We also thank J. Brock for help. This work was supported by the Cancer Research UK Lung Cancer Centre of Excellence and the CRUK City of London Centre Award (C7893/A26233). M.A.B. is supported by Cancer Research UK, the Rosetrees Trust and the Francis Crick Institute. N.J.B. is a fellow of the Lundbeck Foundation (R272-2017-4040) and acknowledges funding from Aarhus University Research Foundation (AUFF-E-2018-7-14) and the Novo Nordisk Foundation (NNF21OC0071483). A. Huebner is supported by Cancer Research UK. D.A.M. is supported by the Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025). T.B.K.W. is supported by the Francis Crick Institute, as well as the Marie Curie ITN Project PLOIDYNET (FP7-PEOPLE-2013, 607722), the Breast Cancer Research Foundation (BCRF), Royal Society Research Professorships Enhancement Award (RP/EA/180007) and the Foulkes Foundation. T.K. is supported by the JSPS Overseas Research Fellowships Program (202060447). C.M.-R. is supported by the Rosetrees (M630) and Wellcome trusts. E.L.L. receives funding from NovoNordisk Foundation (16584). C.T.H. has received funding from NIHR University College London Hospitals

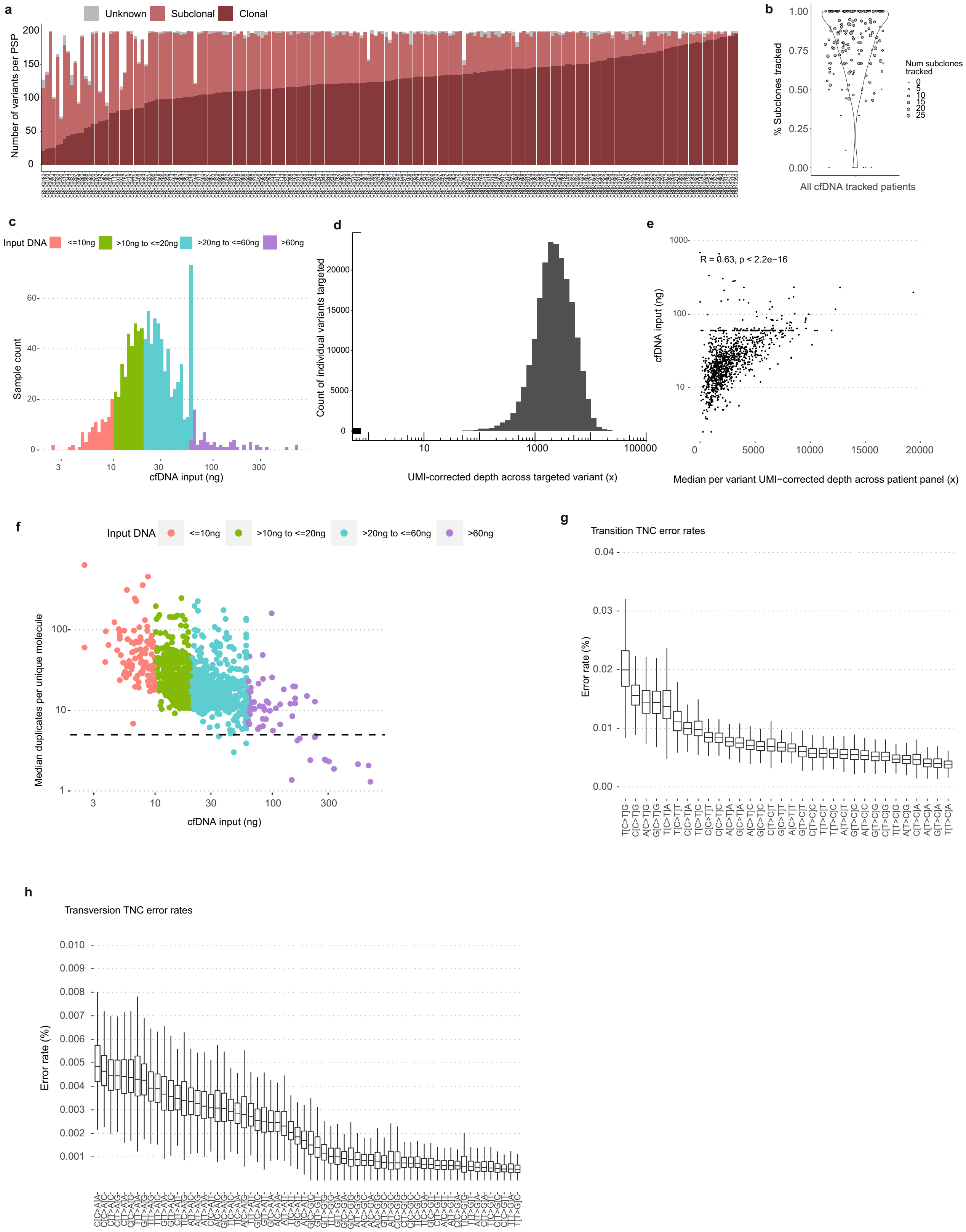
Biomedical Research Centre. M.J.-H. is a CRUK Career Establishment Awardee and has received funding from CRUK, the NIH National Cancer Institute, the IASLC International Lung Cancer Foundation, the Lung Cancer Research Foundation, the Rosetrees Trust, UKI NETs, NIHR and the NIHR UCLH Biomedical Research Centre. N.M. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (grant no. 211179/Z/18/Z) and also receives funding from Cancer Research UK, Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre. T.L.C. acknowledges funding support from the Howard Hughes Medical Institute, and the Radiation Oncology Institute; G.I.E. from Cancer Research UK (A29210) and the European Research Council Advanced Investigator Award (294851); H.J.W.L.A. from the NIH (NIH-USA U24CA194354, NIH-USA U01CA190234, NIH-USA U01CA209414 and NIH-USA R35CA22052) and the European Union–European Research Council (grant agreement no. 866504); and K.L. from the UK Medical Research Council (MR/V033077/1), the Rosetrees Trust and Cotswold Trust (A2437), the Royal Marsden Cancer Charity and Melanoma Research Alliance. BioRender was used in the generation of Fig. 4 and Extended Data Fig. 7. C.S. is a Royal Society Napier Research Professor (RSRP\R\210001); and is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2041), the UK Medical Research Council (CC2041) and the Wellcome Trust (CC2041). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission. C.S. is funded by Cancer Research UK (TRACERx (C11496/A17786), PEACE (C416/A21999) and CRUK Cancer Immunotherapy Catalyst Network); Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield and Stoneygate Trusts; the NovoNordisk Foundation (ID16584); the Royal Society Professorship Enhancement Award (RP/EA/180007); the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre; the Cancer Research UK–University College London Centre; the Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (US) BCRF-22-157; Cancer Research UK Early Detection an Diagnosis Primer Award (Grant EDDPMA-Nov21/100034); and The Mark Foundation for Cancer Research Aspire Award (Grant 21-029-ASP). This work was supported by a Stand Up to Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research grant (grant no. SU2C-AACR-DT23-17 to S. M. Dubinett and A. E. Spira). Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. C.S. is in receipt of an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 835297).

Author contributions C.A., A.M.F., N.J.B., N.M. and C.S. co-wrote the manuscript. C.A., A.M.F., N.J.B., J.S. and C.S. conceived the study design. C.A., A.M.F., J.K., K.G., C.P., D.B., T.L.C., J.W., C.M.-R., M.A.B., O.P., T.B.K.W., E.L.L., A. Huebner, D.A.M., R. Salgado., F.G., A.J.P., E.M., D.E.C., C.T.H., M.J.-H. and N.J.B. integrated clinicopathological data, transcriptomic data, exome data and ctDNA data. C.A., T.H., A.G., A.L., J.S., M.R.S., K.L., L.J., C.P. and C.S. worked to develop and validate the MRD calling algorithm used in this manuscript. A.M.F. developed ECLIPSE and performed analyses of clonal composition used in this manuscript. A.G., M.M., A.C., L.J., P.R. and R.D.D. conducted AMP NGS experimental work for ctDNA data. K.G. performed GISTIC copy-number analysis. S.V., S.W., N.C., J.R., R.D.D., M.M., A.C. and J.A.S. provided oversight of TRACERx patient sample storage and/or DNA extraction and/or sequencing of patient samples. T.L.C., J.W. and H.J.W.L.A. performed radiomic analyses of baseline CT scans. T.H., M.R.S., A.G., A.S., A.O. and A.L. conducted ArcherDx variant selection, PSP design and informatic processing of AMP data. A.M.F., K.G., M.A.B., O.P., T.B.K.W., E.L.L., A. Huebner, D.E.C. and N.M. conducted multiregion sequencing and phylogenetic tree analyses and identified TRACERx variants for PSP design. D.A.M. conducted the pathological review. A. L'Hernault, A.G., L.H., P.R., H.B. and N.G.-H. designed and conducted analytical validation experiments of the AMP MRD assay. C.A. and T.H. designed and conducted in silico specificity experiments for the AMP assay. D.B. and N.J.B. conducted ORACLE analyses. C.A. and T.K. conducted reviews of radiological imaging reports. R.M.K., D.H., D.S., G.I.E. and J.C.B. gave advice on analyses performed in this paper. M.J.-H., J.A.S. and C.S. designed the study protocols. A. Hackshaw gave statistical advice. C.A., N.M., M.J.-H., N.J.B. and C.S. provided overall study oversight. All of the authors approved the final version of the manuscript.

Competing interests C.A. has received speaking honoraria or expenses from AstraZeneca and Bristol-Myers Squibb and reports employment at AstraZeneca. C.A. and C.S. are listed as inventors on a European patent application relating to assay technology to detect tumour recurrence (PCT/GB2017/053289). This patent has been licensed to commercial entities and, under their terms of employment, C.A and C.S are due a revenue share of any revenue generated from such license(s). C.A. and C.S. declare a patent application (PCT/US2017/028013) for methods to detect lung cancer. A.M.F., C.A. and C.S. are named inventors on a patent application to determine methods and systems for tumour monitoring (PCT/EP2022/077987). C.A., C.S., K.L., C.P., T.H., L.J., M.R.S., A.G. and A. Licon are named inventors on a provisional patent protection related to a ctDNA detection algorithm. S.V. is listed as a co-inventor on a patent of methods for detecting molecules in a sample (US patent, 10,578,620). T.H., A.G., M.M., A.C., A.S., A.O., L.J., P.R., M.R.S., R.D.D., A.L. and J.S. are former or current employees of Invitae or ArcherDx and report stock ownership. D.B. reports personal fees from NanoString and AstraZeneca and has a patent (PCT/GB2020/050221) application on methods for cancer prognostication. M.A.B. has consulted for Achilles Therapeutics. D.A.M. reports speaker fees from AstraZeneca, Eli Lilly and Takeda; consultancy fees from AstraZeneca, Thermo Fisher Scientific, Takeda, Amgen, Janssen, MIM Software, Bristol-Myers Squibb and Eli Lilly; and has received educational support from Takeda and Amgen. N.G.-H., A. L'Hernault, H.B., D.H., D.S. and J.C.B. report stock ownership and employment at AstraZeneca. A. Hackshaw has received fees for being a member of independent data monitoring committees for Roche-sponsored clinical trials, and academic projects co-ordinated by Roche. C.T.H. has received speaker fees from AstraZeneca. M.J.-H. has consulted for, and is a member of, the Achilles Therapeutics scientific advisory board and steering committee; has received speaker honoraria from Pfizer, Astex Pharmaceuticals, Oslo Cancer Cluster; and is listed as a co-inventor on a European patent application relating to methods to detect lung cancer (PCT/US2017/028013). This patent has been licensed to commercial entities and, under terms of employment, M.J.-H. is due a share of any revenue generated from such license(s). N.J.B. is listed as a co-inventor on a patent to identify responders to cancer treatment (PCT/GB2018/051912), has a patent application (PCT/GB2020/050221) on methods for cancer prognostication and a patent on methods for predicting anti-cancer response (US14/466,208). H.J.W.L.A. has received personal fees and stock from Onc.AI, Sphera and Love Health, and speaking honoraria from Bristol-Myers Squibb. K.L. has a patent (CA3068366A) on indel burden and CPI response pending and speaker fees from Roche tissue diagnostics and Ellipses Pharmaceuticals, research funding from CRUK TDL/Ono/LifeArc alliance, Genesis Therapeutics and consulting roles with Monopteros Therapeutics and Kynos Therapeutics (all outside of this work). N.M. has received consultancy fees and has stock options in Achilles Therapeutics; and holds European patents relating to targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004) and predicting survival rates of patients with cancer (PCT/GB2020/050221). C.S. acknowledges grant support from AstraZeneca, Boehringer-Ingelheim, Bristol-Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx, collaboration in minimal residual disease sequencing technologies), Ono Pharmaceutical and Personalis; he is an AstraZeneca advisory board member and chief investigator for the AZ MeRmaiD 1 and 2 clinical trials and is also co-chief investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's scientific advisory board. He receives consultant fees from Achilles Therapeutics (also a member of the scientific advisory board), Bicycle Therapeutics (also a member of the scientific advisory board), Genentech, Medicxi, Roche Innovation Centre–Shanghai, Metabomed (until July 2022) and the Sarah Cannon Research Institute; has received honoraria from Amgen, AstraZeneca, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol-Myers Squibb, Illumina and Roche-Ventana; had stock options in Apogen Biotechnologies and GRAIL until June 2021, and currently has stock options in Epic Bioscience, Bicycle Therapeutics, and has stock options and is co-founder of Achilles Therapeutics; and holds additional patent applications related to targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), identifying patients who respond to cancer treatment (PCT/GB2018/051912) and both a European and US patent application related to identifying insertion/deletion mutation targets (PCT/GB2018/051892).

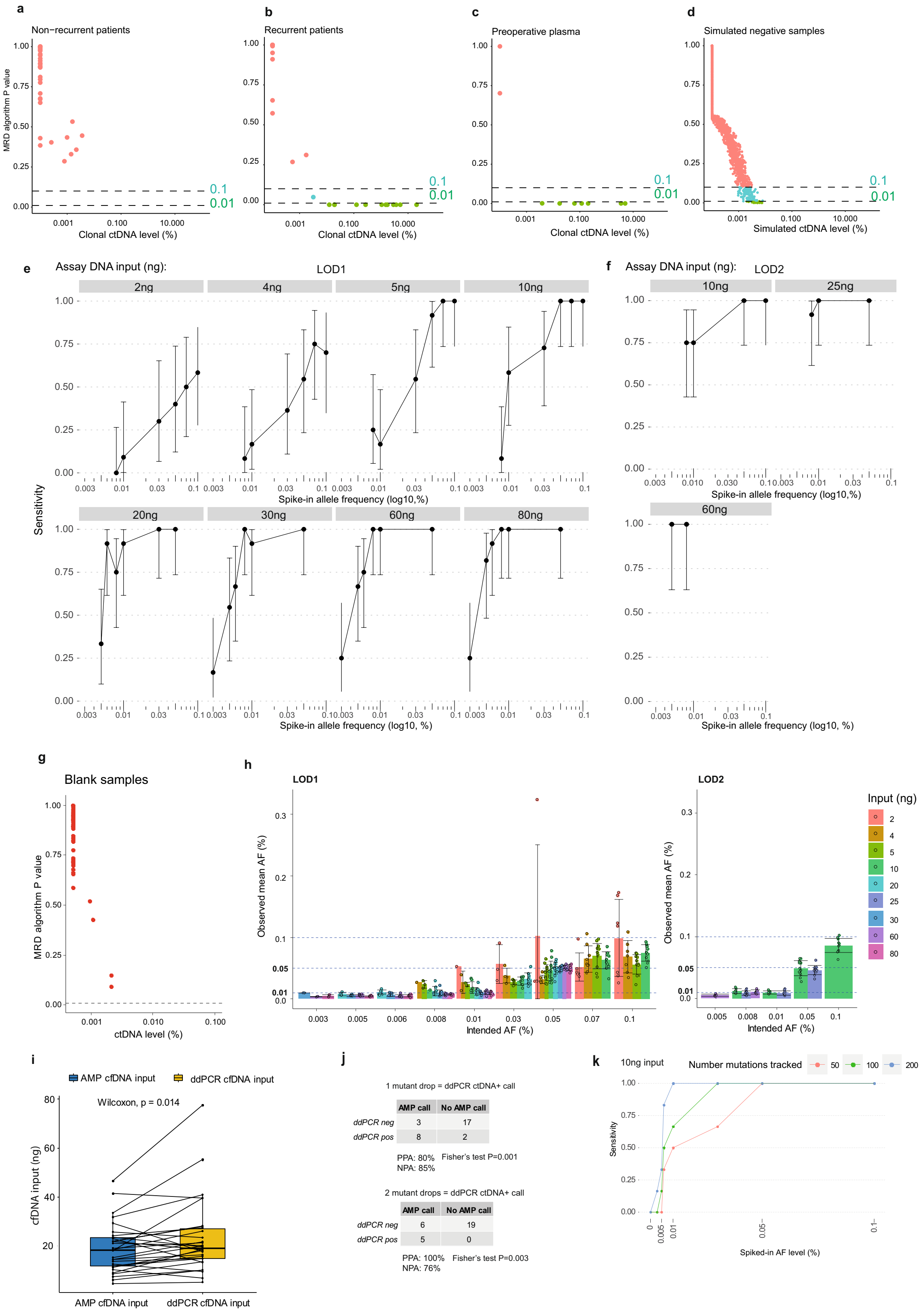
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-05776-4.
Correspondence and requests for materials should be addressed to Christopher Abbosh, Nicholas McGranahan or Charles Swanton.
Peer review information *Nature* thanks Aadel Chaudhuri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



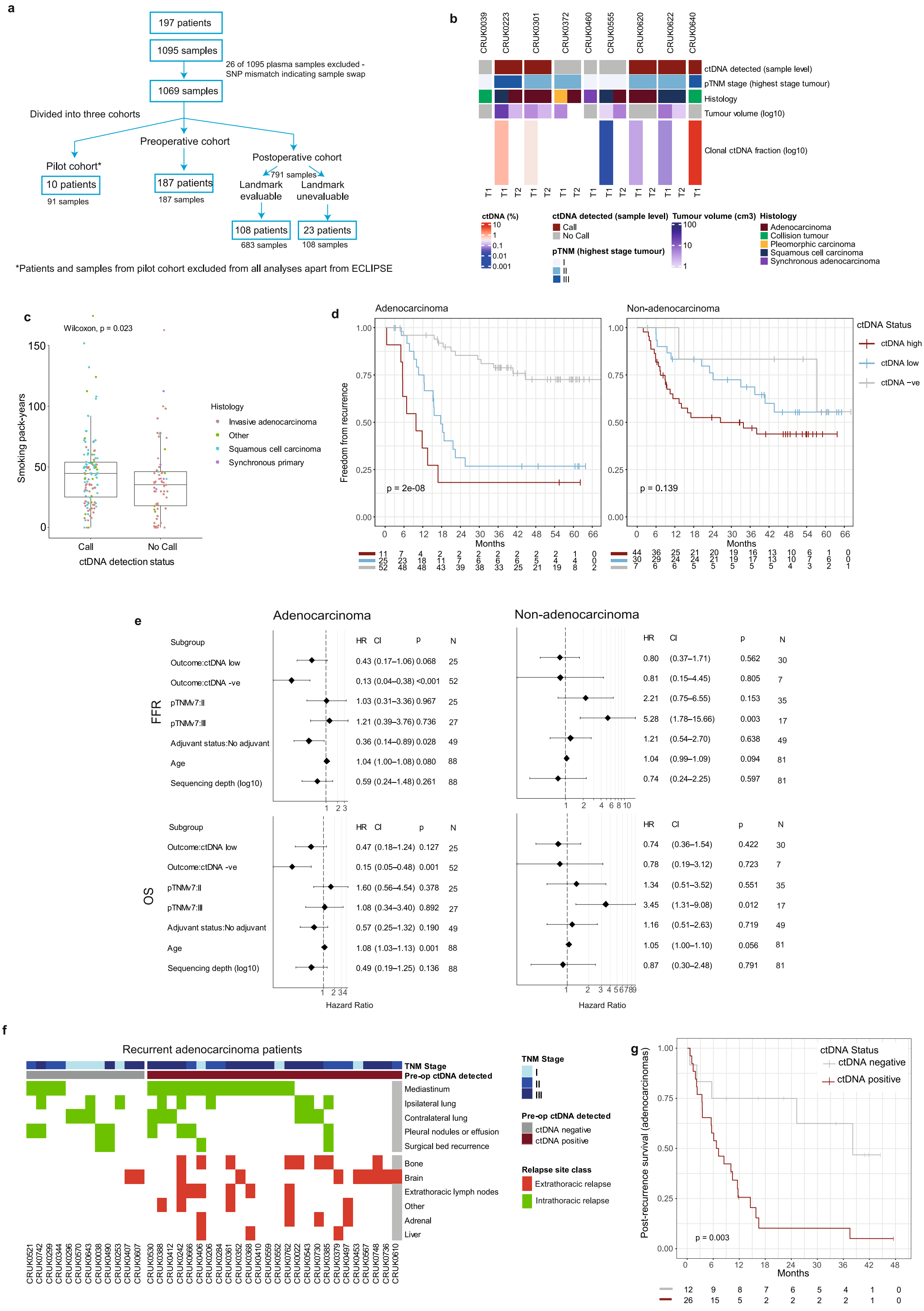
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | TRACERx ctDNA cohort sequencing parameters. **A**. Stacked bar plot of patient specific panels (PSPs) designed from primary tumour sequencing data showing the number of clonal (dark red) and subclonal (light red) variants per panel. Variants lacking clonality information are displayed in grey (median of 3 variants per patient [1-20], these mutations are either no longer called by TRACERx or called by ArcherDx but not TRACERx, see methods). A median of 126 clonal variants (range 21 to 195) and 64 subclonal variants (range 0 to 174) were tracked by the PSPs. Clonality was determined by PyClone analyses of multi-region exome data derived from primary resections of NSCLC (methods), in the absence of PyClone data, variants present in all multi-region sequenced tumour samples were called clonal. **B**. Violin plot demonstrating the % of subclonal clusters derived from multi-region tumour exome data tracked by PSPs on a per patient basis. A median of 88% of the subclonal mutation clusters present in each patient's multi-region exome derived phylogenetic tree were tracked [range 0-100]. 184 tumours with phylogenetic trees were included. **C**. Distribution of cfDNA input values for the cohort, median input of 23 ng, n = 1069 samples. Capping at 60 ng input was performed for some of the cohort explaining the peak at this value; for the remainder of the cohort, all cfDNA extracted was input into the assay (colours represent different cfDNA input categories as indicated). **D**. Histogram demonstrating the distribution of per-variant unique sequencing depth values across the cohort; unique depth refers to error-controlled depth achieved across a position targeted by a PSP (at least 5 unique molecular identifier (UMI) matched reads required to create a consensus error-controlled read, see methods). The median unique depth per-variant tracked by a PSP was 2226x (range 0 to 53789x, n = 201910). **E**. Correlation between cfDNA input (ng, Y axis) into the assay and the median UMI-corrected depth achieved across a PSP across 1069 plasma timepoints (X axis). Spearman's R value = 0.63 and two-sided P value < 2.2e-16. **F**. Association between median deduplication ratio achieved in a sample (Y-axis) and cfDNA input into the assay (ng, X-axis); duplication ratio refers to the median number of duplicate UMI-supported reads within a read family. Resequencing of samples where the median duplication ratio was less than 10 was performed where possible to maximize recoverable information from cfDNA samples, given that 5 UMI-supported reads are required to make a UMI family. 17 of 1069 evaluated cfDNA samples exhibited a final median deduplication ratio less than 5 (corresponds to the horizontal line on the plot). Colours correspond to different cfDNA input categories and match panel c. **G-H**. Boxplots demonstrating the error rates (%, Y axis) per each of 96 mutation trinucleotide contexts (X axis, 192 mutation trinucleotide contexts [TNCs] simplified to 96 reverse-complement identical mutation types), plots divided by transition event (G) and transversion event (H). Background position data from n = 1069 cell-free DNA libraries utilized to generate plots, variants predicted to exhibit low background error rates from pilot data analyses were prioritized for PSP design. Hinges correspond to first and third quartiles, whiskers extend to the largest/smallest value no further than 1.5x the interquartile range. Centre lines represent medians.



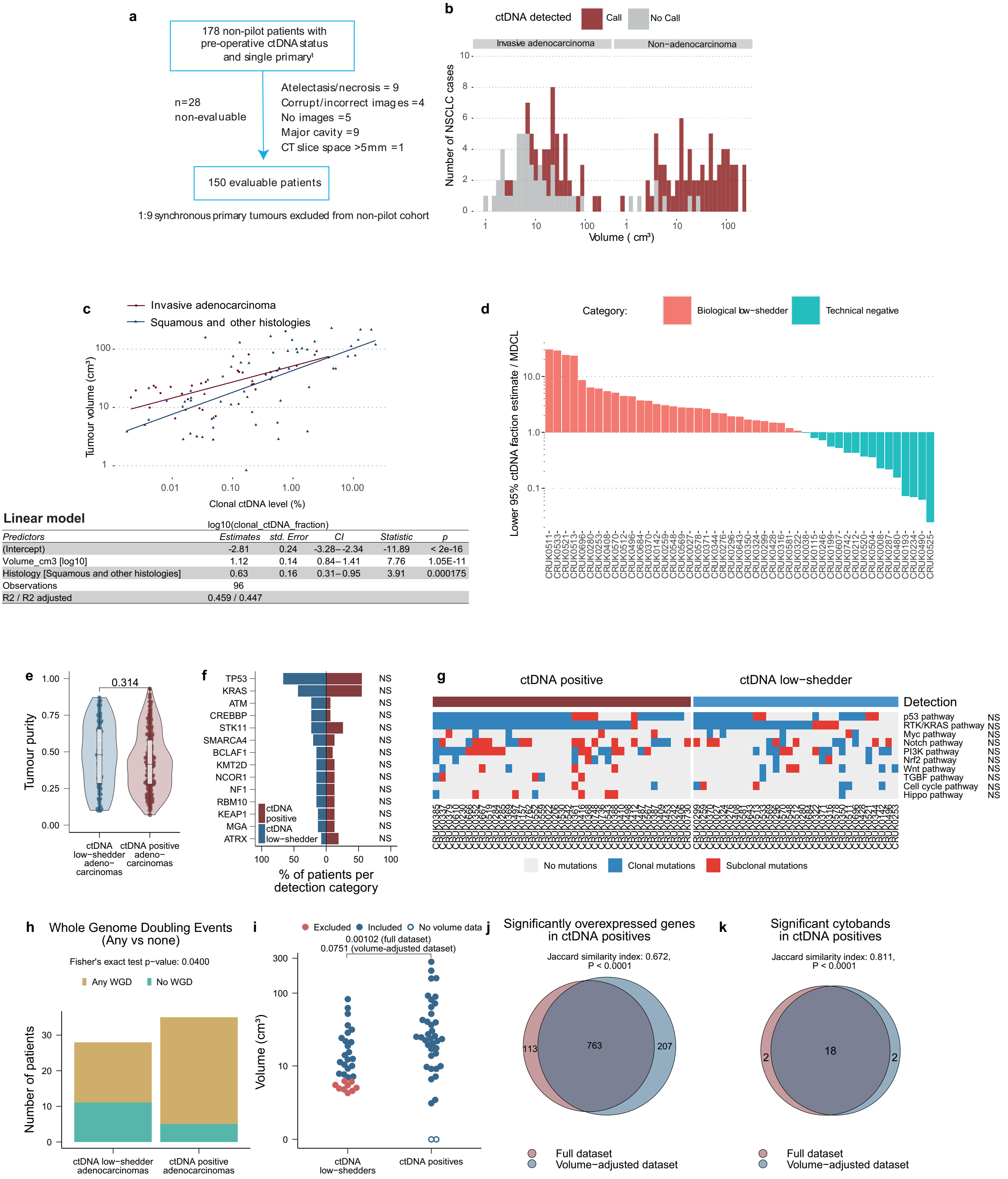
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | MRD calling thresholds and analytical validation. **A-D**. Pre- and postoperative MRD caller P values (Y axis MRD caller P value, one-sided Poisson test, see Methods) observed in pilot-phase of the project. X axis displays clonal ctDNA levels. **A**. Postoperative samples from n = 5 patients who did not have recurrence of their NSCLC; all n = 55 patient samples had caller P values in excess of P > 0.1 threshold meaning that they were deemed negative for ctDNA. **B**. Postoperative caller P values observed in n = 5 patients who had relapse of their NSCLC. 1 of 13 calls was made between caller P values of 0.1 and 0.01, the remaining 12 calls were made at a caller P value less than 0.01. **C**. Preoperative ctDNA calls from pilot cohort; 7 patients had positive ctDNA in plasma prior to surgery, all calls were made at caller P values < 0.01. **D**. In-silico simulation analysis to assess MRD caller specificity. 3157 mock MRD panels were generated within the evaluable pilot patient libraries and MRD caller P values were assessed. At a caller P value < 0.1 threshold, 121/3157 simulated mock panels were ctDNA positive (*in-silico* specificity of 96.2%); at a caller P value threshold < 0.01, 22/3157 simulated mock panels were ctDNA positive (*in-silico* specificity of 99.3%). **E-F**. Analytical validation of 50 variant MRD detection panels. **E**. Fragmented DNA with a known single nucleotide polymorphism (SNP) profile was spiked into a second background of fragmented DNA with a different SNP profile and a patient-specific panel targeted 50 alternate positions present in spiked-in DNA. 559 data points were generated across different DNA input quantities indicated, to establish the limit of detection plots. The Y axis and centre of the error bars demonstrate sensitivity (defined as the proportion of all repeats that resulted in MRD detection using a caller P value of 0.01). The confidence intervals on the plot are Clopper-Pearson confidence intervals (95% CIs). The X axis shows the quantity of variant germline DNA that was spiked into each repeat expressed as a percentage of total DNA in that sample. **F.** Circulating tumour DNA samples with high variant allele fractions were spiked into a different cell-free DNA background. Variant positions in ctDNA were targeted with a 50 variant panel; 100 data points were generated across the DNA input quantities indicated. Axes and error bars are the same as (E). **G**. Data from analyses of 48 blank samples donated by 24 healthy participants, caller P values are displayed. **H**. Barplots demonstrating the intended allele frequencies and the measured allele frequencies in the different spike-ins presented in part (E) and part (F) only data from variant DNA positive samples are presented. The colours of the barplot represent different DNA input masses as shown by the legend. The error bars on the plot represent the mean value of all positive spike-in samples +/− standard deviation of the values. Where the error bar is absent, this is because at this spike-in level and DNA input mass, only one positive sample was observed. Where the error bar led to an observed mean AF less than 0, the error bar was stopped at 0 for visualization purposes (the 0.05% spike-in, 2 ng input mass case). The horizontal dashed lines correspond to 0.1%, 0.05%, and 0.01% spike-in categories. Each data point is represented on the plots by a circle. n = 369 variant DNA positive samples displayed in LOD1 barchart, n = 93 variant DNA positive samples displayed in LOD2 barchart. **I.** Comparison between the content of cell-free DNA input into ddPCR reactions (yellow) and AMP PCR reactions (blue). Hinges correspond to first and third quartiles, whiskers extend to the largest/smallest value no further than 1.5x the interquartile range. Centre lines represent medians. Each dot on the plot represents a data point, lines connect paired samples from the same patient. Significantly more cell-free DNA was input into ddPCR reactions (paired two-sided Wilcoxon-test P = 0.01366). **J**. Orthogonal comparison between ctDNA detection based on AMP panels used in TRACERx and ddPCR against a single clonal variant. ddPCR ctDNA positive call threshold was two mutant droplets (bottom table) and one mutant droplet (top table). Percentage positive agreement (PPA) and percentage negative agreement (NPA) using ddPCR as the comparator is displayed in the table. Two-sided Fisher's test P values are demonstrated under the cross tables. **K**. A 300 mutation patient-specific panel was designed and applied to 10 ng DNA samples containing spike-in variant levels from 0% to 0.1%. *In silico* sub-sampling of the 300 mutations was performed (3 x 200 mutation *in silico* panels, 3x 100 mutation *in silico* panels and 3x 50 mutation *in silico* panels, see methods) and sensitivities are categorized by the number of mutations targeted by the panel.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Preoperative ctDNA detection. A Flow diagram demonstrating different cohorts analysed in this manuscript; the top part of the flow diagram shows the total number of plasma samples that were intended to be analysed (n = 1095 from 197 patients) which reduced to 1069 samples due to single nucleotide polymorphism mismatches between cfDNA and tissue exome data in 26 cases, suggesting sample swap. These samples were analysed in 3 main cohorts, the pilot cohort (left), the preoperative cohort (middle), and the postoperative cohort (right). The postoperative cohort was divided into different categories based on landmark evaluability (relating to samples donated within 120 days of surgery to enable a landmark ctDNA analysis). **B**. Heatmap demonstrating individual tumour-specific clonal ctDNA fractions in patients with synchronous primaries diagnosed at baseline. The annotation rows of the heatmap show the ctDNA call present in that sample across all variants interrogated by the MRD caller, the highest pathological TNM stage, the individual histology, and individual tumour volumes of the two synchronous tumours present at baseline (for this category, grey represents absent data or volume unevaluable). **C**. Boxplot demonstrating the difference in pack-year history across 187 preoperative ctDNA positive NSCLC patients and preoperative ctDNA negative NSCLC patients. Hinges correspond to first and third quartiles, whiskers extend to the largest/smallest value no further than 1.5x the interquartile range. Centre lines represent medians. P value represents a Wilcoxon rank sum test. **D**. Kaplan-Meier curves demonstrating freedom from recurrence outcomes in ctDNA high (dark red), ctDNA low (blue), and ctDNA negative (grey) single primary adenocarcinoma patients (left) and single primary non-adenocarcinoma patients (right). ctDNA high and low were categorized based on median clonal ctDNA levels across ctDNA positive cases and relate to above and below 0.16%. Log-rank P values are displayed on each plot. **E**. Multivariable Cox regression analyses of Overall Survival (OS) and Freedom From Recurrence (FFR, defined as recurrence only) in patients with single (non-synchronous) NSCLC; evaluating ctDNA detection status, pTNM stage (Tumour Node Metastasis pathological stage version 7, categories I, II or III), whether adjuvant therapy was administered, age, and log10-transformed unique sequencing depth as predictors in adenocarcinomas and non-adenocarcinomas separately. Unique sequencing depth was included to adjust for under sequenced samples, representing potential false negatives. n = 88 adenocarcinoma patients and n = 81 non-adenocarcinoma patients were analysed for FFR and OS. On the forest plots, the diamond represents the multivariable Hazard Ratio (HR) with error-bars corresponding to 95% confidence intervals (CI). Multivariable P values (p) are displayed on the plot alongside the number of patients in each category (N). Reference categories were ctDNA positive patients, pTNM stage I patients and patients given adjuvant therapy. The exact Cox regression P value for the Outcome: ctDNA -ve category in the FFR adenocarcinoma plot = 0.00022. **F**. Heatmap showing the site of relapse in recurrent adenocarcinoma cases divided by whether preoperative ctDNA was detected (dark red, right) or undetected (grey, left). Intrathoracic (mediastinum, locoregional, ipsilateral lung, distant lung – green colours) or extrathoracic (bone, brain, liver, adrenal, extrathoracic lymph nodes or other extrathoracic site – red colours) sites of relapse are shown (sites shown are metastatic sites diagnosed within 180 days of clinical relapse). Heatmap is annotated by Tumour Node Metastasis pathological version 7 stage. **G**. Kaplan-Meier curve demonstrating post-relapse survival in recurrent adenocarcinoma patients (n = 38) stratified by preoperative ctDNA positive (red) or preoperative ctDNA negative (grey). Log-rank P value is displayed on the plot.



Linear model

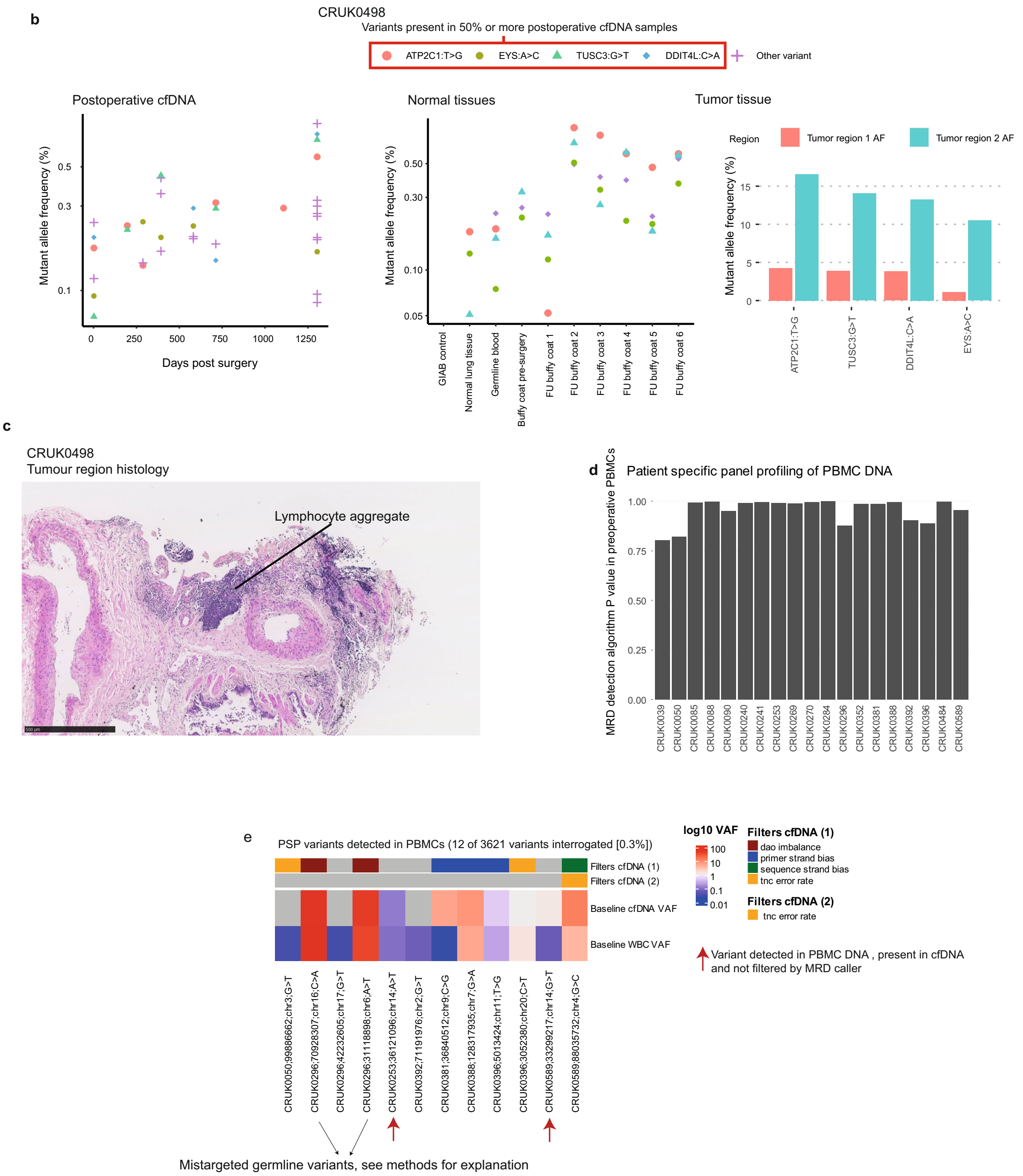
	log10(clonal_ctDNA_fraction)				
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	-2.81	0.24	-3.28 – -2.34	-11.89	< 2e-16
Volume_cm3 [log10]	1.12	0.14	0.84 – 1.41	7.76	1.05E-11
Histology [Squamous and other histologies]	0.63	0.16	0.31 – 0.95	3.91	0.000175
Observations	96				
R2 / R2 adjusted	0.459 / 0.447				

Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Volume and phenotypic analysis of ctDNA positive and ctDNA negative adenocarcinomas. A. Flow chart demonstrating patients available for volumetric analyses and reasons for exclusion. **B**. Histogram showing the number of NSCLC cases by volume, with ctDNA positive samples shown as red bars, and ctDNA negative samples shown as grey bars. n = 150 volume evaluable cases. **C**. Volume versus log10-transformed clonal ctDNA level correlation plot with each individual TRACERx case that was ctDNA positive as a point and coloured by adenocarcinoma status (dark red) and squamous or other histology (dark blue). Fitted line represents a linear model line categorized by tumour histology. Below the correlation plot is a table describing a linear multivariable model based on these data to predict log10-transformed clonal ctDNA levels based on tumour volume and histology (adenocarcinoma and squamous and other categories). P values represent linear model adjusted P values, n = 96 ctDNA positive, volume evaluable NSCLCs analysed. **D**. Based on a multivariable linear regression model fitted to the data in (C), we categorized ctDNA negative adenocarcinomas as biological low-shedders or technical non-shedders (see methods). If a particular tumour volume resulted in an estimated clonal mutation ctDNA level above the clonal ctDNA level a library could detect (95% lower confidence interval for estimated clonal ctDNA level based on tumour volume is above detectable clonal ctDNA level in the preoperative cfDNA library from that patient), then the case was classed as a probable biological low-shedder (red on histogram); otherwise, the case was classed as a probable technical non-shedder (turquoise on histogram). Y axis represents the lower 95% confidence estimate for clonal mutation ctDNA level divided by the minimally detectable clonal mutation ctDNA level (MDCL) for that patient's panel. The X axis is each individual patient analysed. Data from n = 47 ctDNA negative adenocarcinomas presented. **E**. Violin box-plots comparing tumour purity in ctDNA low-shedder adenocarcinomas (blue, n = 79 tumour regions from 28 patients) and ctDNA positive adenocarcinomas (red, n = 166 tumour and lymph node regions from 35 patients). Pairwise comparisons are performed using linear mixed-effects models, P values are two-sided. Boxplot hinges correspond to first and third quartiles, whiskers extend to the largest/smallest value no further than 1.5x the interquartile range and centre lines represent medians. Violins represent the distribution of the underlying data. **F**. Barplots showing gene-level driver alterations between ctDNA positive adenocarcinomas (n = 39 patients) and ctDNA negative low-shedder adenocarcinomas (n = 31 patients). Colours denote ctDNA detection status. Y axis shows the top 14 most frequently altered genes, X axis shows the percentage of patients carrying an alteration in the gene per detection category. NS: Not significant (two-sided Fisher's exact test with FDR P value adjustment). **G**. Pathway-level driver mutations between ctDNA positive adenocarcinomas (n = 39 patients) and ctDNA negative low-shedder adenocarcinomas (n = 31 patients). X axis shows patient IDs, Y axis shows pathways following the Sanchez-Vega definition. Top bar denotes ctDNA detection status (dark red represents ctDNA positives, blue represents biological low-shedders). Heatmap colours display mutations; blue denote clonal mutations and red denote subclonal mutations. No pathway showed significant enrichment in either ctDNA shedder or non-shedder adenocarcinomas (NS: Not significant, using two-sided Fisher's exact test with FDR P value adjustment). **H**. Whole genome doubling status per tumour comparing ctDNA positive adenocarcinomas to ctDNA negative low-shedder adenocarcinomas, using two-tailed Fisher's exact test. Yellow represents the number of tumours subjected to whole genome doubling in at least one region, turquoise represents tumours without any whole genome doublings. **I**. Volume by ctDNA shedding status. Biological non-shedders in red represent the smallest quartile samples. After removal of these from the analysis, no significant difference in tumour volume was found between ctDNA positives and ctDNA low-shedders. Pairwise comparisons are made with two−sided Wilcoxon rank sum tests. **J**. Venn diagram showing the overlap between significantly differentially expressed genes between ctDNA positive and ctDNA low shedder adenocarcinomas obtained from the full dataset, relative to the volume-adjusted dataset. Comparisons are made by computing the Jaccard similarity index and the corresponding two-sided P value using the exact method. **K**. Venn diagram showing the overlap between significantly altered cytobands as called by GISTIC, comparing ctDNA positive to ctDNA low shedder adenocarcinomas obtained from the full dataset, relative to the volume-adjusted dataset. Statistical testing follows (J).

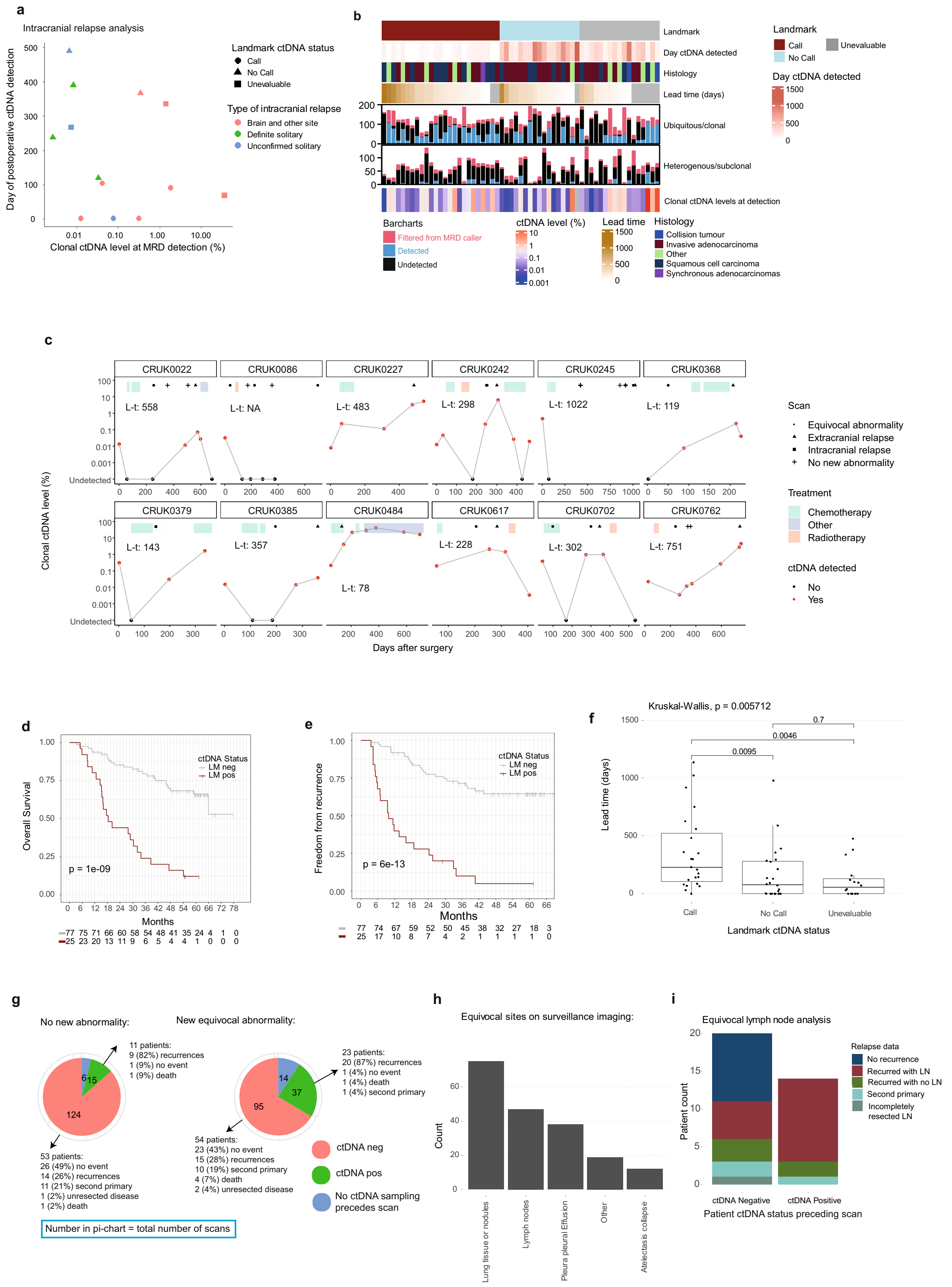
a

Patient ID	Detection timepoint	Clonal ctDNA levels (%)	Clinical context
CRUK0086	Day 2	0.03%	Possible mediastinal disease post surgery (but not clinically confirmed), subsequently treated with radiotherapy and patient remained ctDNA negative at days 130, 193, 284 and 375.
CRUK0498	7 of 8 postoperative timepoints	0.004% to 0.064%	False positive ctDNA positive calls due to mistargeting of normal tissue variants present within an expanded lymphocyte clone within primary tumor tissue. Not subtracted as germline variants due to low mutant allele frequency in peripheral blood.
CRUK0269	Day 693, 322	0.005 and 0.003%	This patient developed a second primary squamous lung cancer at D1407 postsurgery. Lung cancer defined as second primary based on histology (primary was adenocarcinoma). No tissue available for genomic analysis of second primary.CRUK0269's patient specific panel was applied to control genome in a bottle (GIAB), pre operative germline blood, buffy coat from all post operative time points and normal tissue. No variants detected in postoperative cfDNA were present in these samples.



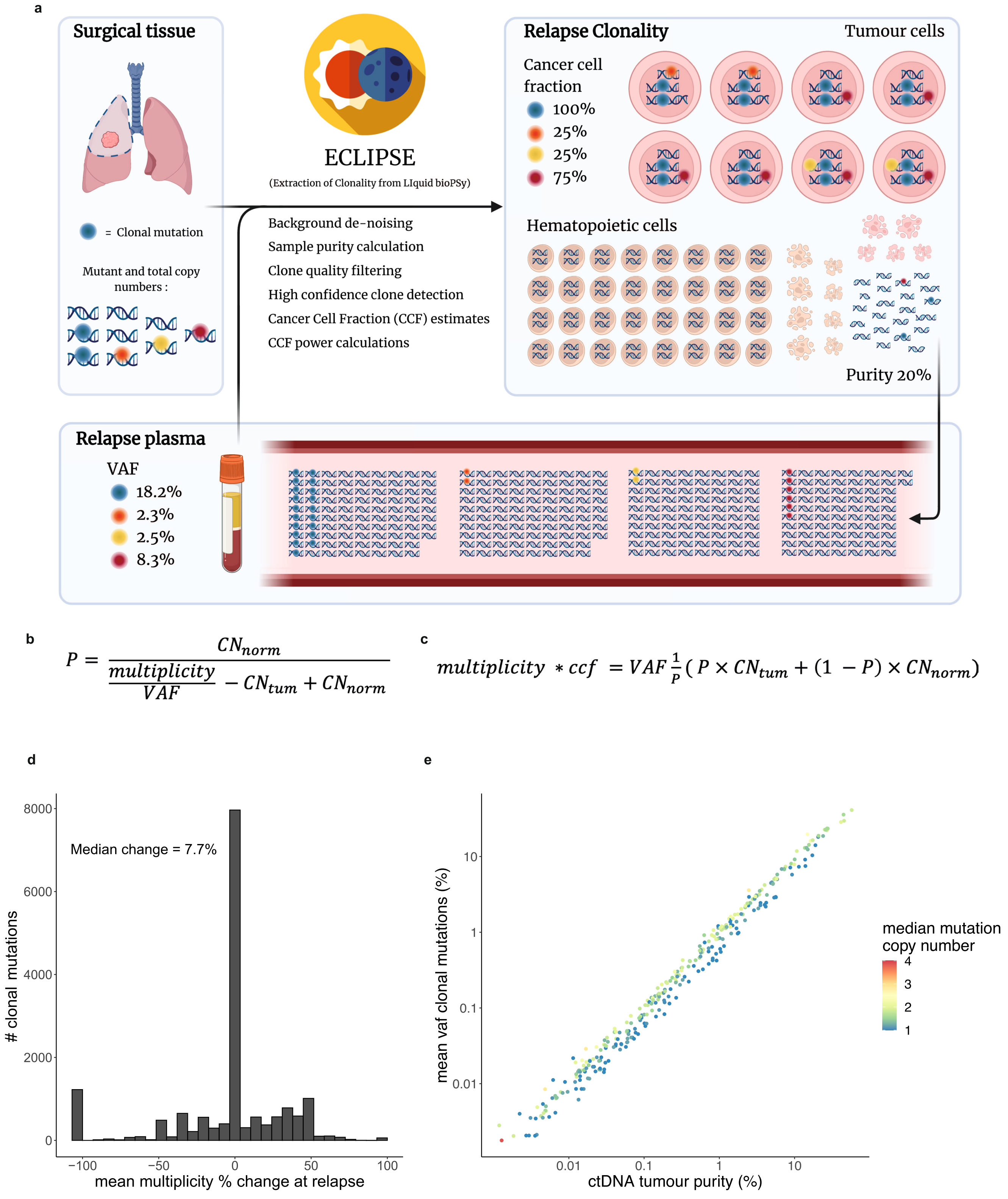
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Exploration of unexpected MRD positive results in non-relapse patients. A. Table demonstrating details of unexpected ctDNA positive results in patients who did not have disease recurrence. **B**. CRUK0498 false positive analysis: Dot-plots represent confidently detected variants at illustrated cfDNA sampling timepoints (left panel), variants confidently detected in normal tissue, control DNA, and peripheral-blood mononuclear cell (PBMC, buffy-coat) DNA based on application of CRUK0498's patient specific panel to these respective samples (middle panel) and the mutant allele frequencies of selected variants in tumour tissue exome data (right panel). The four variants in the legend (variants in genes *ATP2C1*, *DDIT4L*, *EYS*, and *TUSC3*) represent variants confidently called at 50% or more of the timepoints across the cfDNA samples (note that confidently called means an individual variant Poisson one-sided P value of <0.01 [generated by MRD caller, see methods]). **C**. A haematoxylin and eosin image from patient CRUK0498's tumour where exome analysis detected the variants in genes *ATP2C1*, *DDIT4L*, *EYS and TUSC3* at high variant allele-frequencies. This image shows a dense lymphocyte aggregate in this tumour region. Scale bar below image. A single image was analysed. **D**. A further 19 preoperative PBMC samples were analysed from TRACERx patients; no confident panel-wide variant DNA calls were made in these patients' PBMC samples using the MRD calling algorithm. **E**. Variant-level analyses of the preoperative PBMC samples analysed in panel (D) highlighted that 12 of 3621 variants interrogated by the panels were detected (variant level one-sided Poisson P value < 0.01). 8 of 12 detected variants were removed from the MRD caller algorithm in cell-free DNA analyses (cfDNA) due to triggering filters highlighted in the heatmap annotation. Only 2 of the 4 remaining variants carried deep alternate reads in the respective patients' preoperative cfDNA sample (red arrows). The heatmap shows the cfDNA variant allele frequency and the WBC variant allele frequency of the detected variants (grey colour represents no detection of the variant). Two mistargeted germline variants are highlighted by black arrows for patient CRUK0296, variants were targeted in error by the industry panel design pipeline but not by the TRACERx exome pipeline (methods), and were filtered from the MRD calling algorithm due to triggering the outlier filter (dao imbalance filter, dark red).



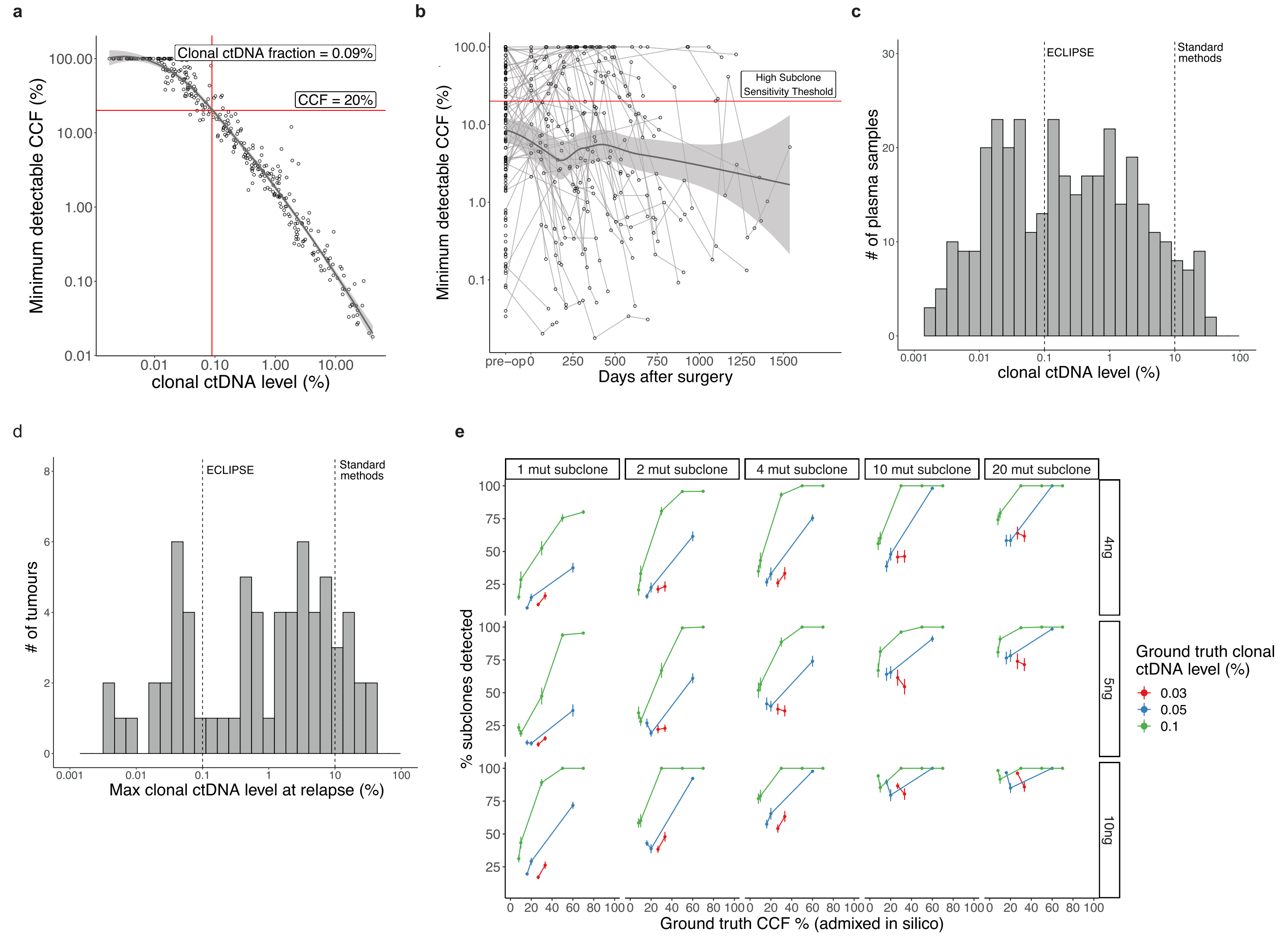
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Expanded postoperative ctDNA and imaging surveillance analysis. A. Analysis of 13 patients who experienced intracranial relapse who were positive for ctDNA in a postoperative blood sample. The X axis shows the clonal ctDNA level at the point of postoperative ctDNA detection and the Y axis shows the day of postoperative ctDNA detection. Points are coloured based on whether the intracranial relapse was solitary (green), accompanied by another extracranial site (red), or unconfirmed solitary (blue, no extracranial imaging performed) and are shaped by landmark ctDNA status. **B**. Heatmap of clonal mutation ctDNA level data at first postoperative ctDNA detection. The annotation rows show the landmark ctDNA status of the patient (landmark positive, ctDNA detected within 120 days postoperatively; landmark negative, ctDNA negative within 120 days postoperatively; unevaluable, landmark status cannot be established), the day ctDNA was detected postoperatively, the histology of the primary tumour, and lead time (days from ctDNA detection to clinical relapse). Where lead time was not applicable (for example incompletely resected disease, ctDNA detected post-relapse, see methods) lead time is coloured grey. The next two rows (bar charts) demonstrate the number of clonal or subclonal mutations tracked by an AMP patient-specific panel (PSP); if the bar is blue, it represents confident detection of an individual variant (based on an individual variant P value of <0.01 [one sided Poisson test based on MRD caller output, see methods]), if the bar is black, it represents absence of confident calling of a variant, if the bar is red, it represents that a variant was filtered by the MRD calling algorithm. The final row represents the mean clonal ctDNA level at the first ctDNA detection time point for a patient. This is on a log-10 scale as displayed in the heatmap legend. For patient CRUK0296, ctDNA detection occurred but clonal ctDNA levels were 0% (grey bar) as the mutation driving ctDNA detection postoperatively did not have a clonal status. **C** Longitudinal per-patient plots in 12 patients who were ctDNA positive prior to adjuvant therapy. Plots are annotated with lead time (L-t), scans performed, and treatment administered (see legend). The Y axis represents clonal ctDNA levels and each circle on the plot represents a blood sampling time point. If the circle is red, it indicates that the blood sample was positive for ctDNA using the MRD caller. The X axis displays days post-surgery. **D-E**. Kaplan-Meier curves in the landmark evaluable population (patients who donated blood within 120 days post-surgery before treatment or clinical recurrence, n = 102/108 landmark evaluable patients were evaluable for survival analysis, see methods for exclusions) showing overall survival (OS,D) or freedom from recurrence (FFR,E) outcomes for landmark positive (dark red) versus landmark negative (grey) patients. Log-rank P values displayed on curves. **F**. Boxplots showing the distribution of lead times (times from ctDNA detection to clinical recurrence) categorized by patient landmark ctDNA status. Hinges correspond to first and third quartiles, whiskers extend to the largest/smallest value no further than 1.5x the interquartile range. Centre lines represent medians. Kruskal-Wallis test P = 0.0057, unadjusted pairwise Wilcoxon-tests compare individual categories, n = 63 patients analysed. **G**. Pie charts demonstrate the number of occurrences of specified ctDNA detection statuses (red – ctDNA negative, green – ctDNA positive, blue – no ctDNA status established), preceding a scan showing no new changes (left) or new equivocal extracranial changes (middle). The ctDNA positive and negative categories are then broken down further into a patient-level analysis showing the outcomes of patients who experienced the occurrence of the specified imaging and ctDNA status event(s). **H**. Barchart showing the count of specific equivocal anatomical sites noted on scans showing new equivocal changes; equivocal lung lesions and lymph nodes were the most common abnormal equivocal findings on NSCLC surveillance imaging. Multiple equivocal sites can be observed on one scan. **I**. Barplot of eventual site of relapse and ctDNA status in 33 patients with ctDNA status established prior to surveillance imaging, showing new equivocal lymph node enlargement. The X axis shows the patient ctDNA detection status preceding surveillance scans. The Y axis shows the patient count. Patient CRUK0090 exhibited occurrences of both negative and positive ctDNA statuses prior to separate equivocal lymphadenopathy scans, so is present in both ctDNA positive and negative categories. Other patients are only included once. Patient CRUK0234 was diagnosed with an unresected lymph node, was ctDNA negative postoperatively and included in the analysis. The barcharts are filled with recurrence status of patients in these categories. Recurred with LN refers to lymph node involvement at relapse (dark red colour). Recurred with no LN refers to recurrence with no lymph node involvement (green colour).

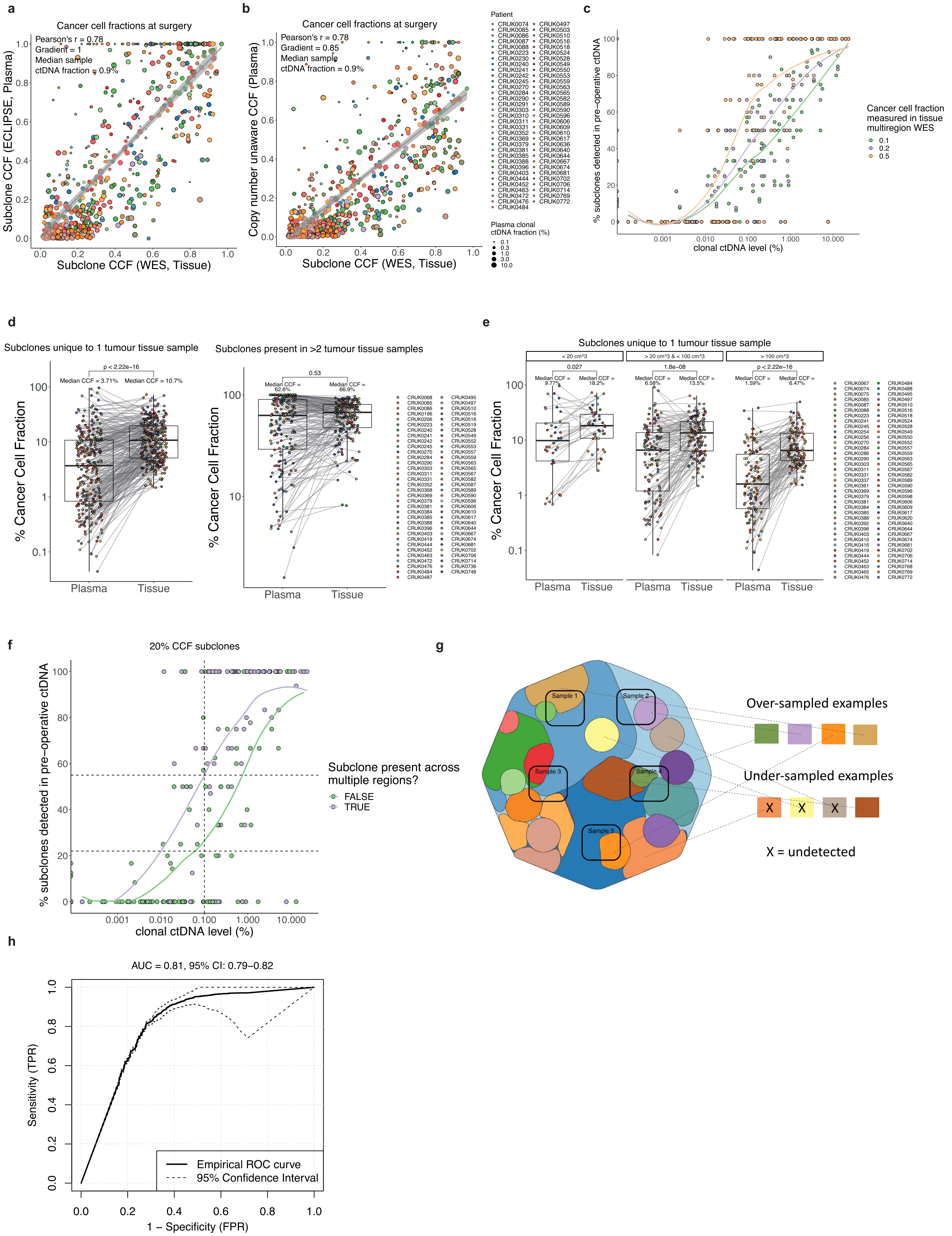


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | ECLIPSE methodology. A. A conceptual overview of the ECLIPSE method and data input types. CCF; cancer cell fraction and VAF; variant allele fraction. The schematic was created using BioRender. **B**. Equation to calculate tumour purity (the % of cells from which the DNA was derived which are tumour cells, see supplementary note 1, also termed 'cellularity' or 'aberrant cell fraction') using clonal mutations. **C**. Equation to calculate cancer cell fraction (CCF). Multiplicity = the number of mutated DNA copies in each mutated cell, CNt = total copy number in the tumour, CNn = total copy number in normal (non-tumour) cells, VAF = variant allele fraction, *P* = tumour purity (the % of cells from which the DNA was derived which are tumour cells, see Supplementary Note 1). **D**. Percentage change in mean multiplicity of clonal mutations comparing measurements in surgical excised tissue samples to tissue samples taken at relapse (46 patients with paired primary and recurrence tissue samples plotted). **E**. A comparison between mean clonal VAF of mutations and ctDNA tumour purity as calculated by ECLIPSE where data points (plasma samples) are coloured by the average copy number of tracked clonal mutations (measured using tissue sequencing). Multi-tumour patients and samples with evidence of copy number of instability at relapse are excluded. A total of 322 samples from 134 patients are plotted.

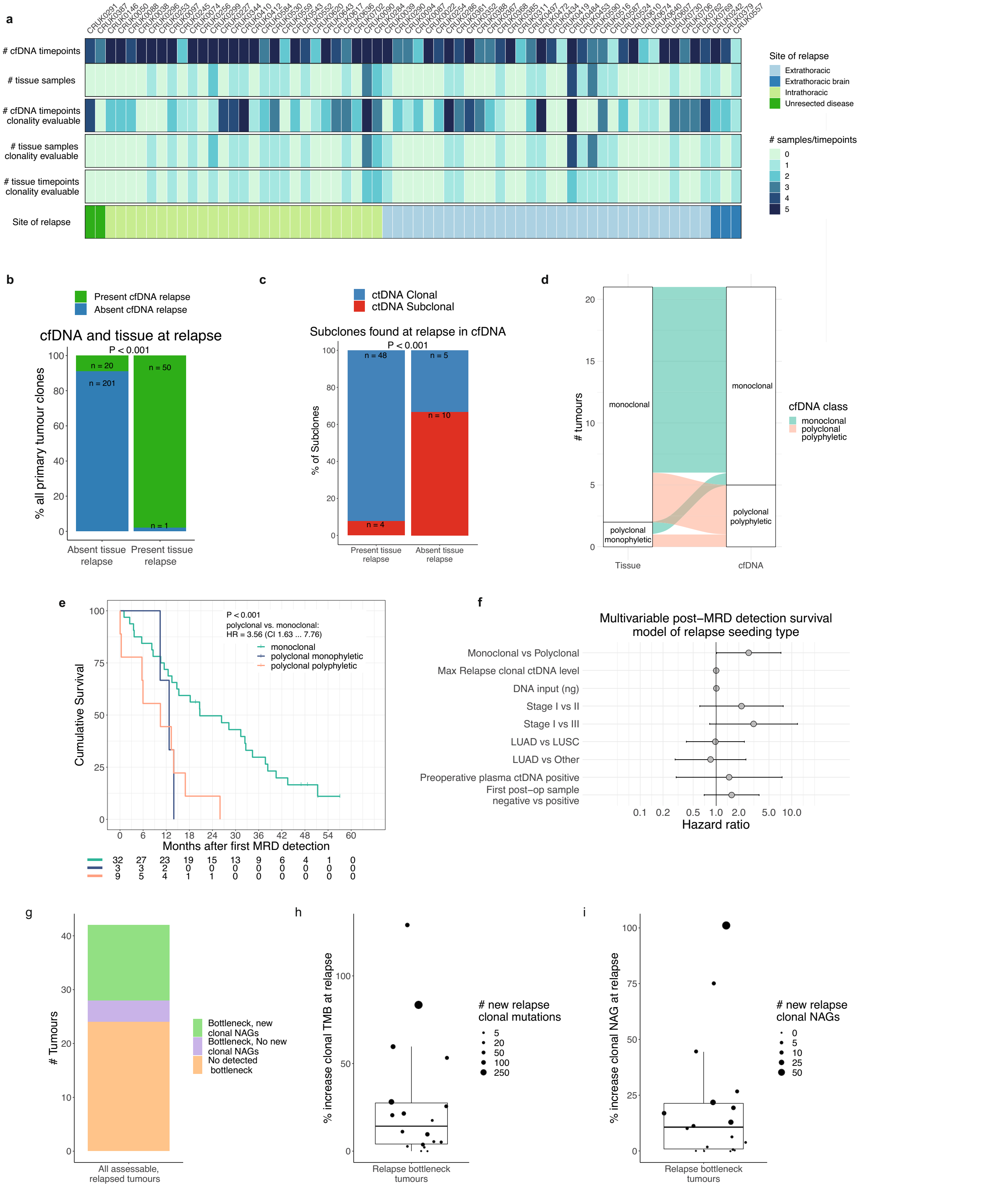


Extended Data Fig. 8 | Subclone detection sensitivity of ECLIPSE. **A**. Minimally detectable CCF for each ctDNA positive sample compared to clonal ctDNA levels for each sample. All ctDNA positive samples included (N = 354). Minimally detectable CCF was calculated using the minimum number of required reads for a positive ($P < 0.01$) clone detection call (methods). **B**. Minimally detectable CCF over time for each patient with a horizontal line indicating the threshold for high subclone sensitivity samples (20% CCF). All ctDNA positive samples included (N = 354). 61% of preoperative MRD positive samples were considered high subclone sensitivity and 66% of postoperative samples were considered of high subclone sensitivity (overall 64% of samples). **C**. A histogram of clonal ctDNA levels for all ctDNA positive samples (N = 354) with vertical lines indicating thresholds for ECLIPSE evaluability and for traditional clonal deconvolution evaluability used for TRACERx tissue samples[28] and previous clonal deconvolution approaches in ctDNA[14,77]. **D**. A histogram of maximum clonal ctDNA levels observed in post-operative samples for each patient with vertical lines indicating thresholds for ECLIPSE evaluability and for traditional clonal deconvolution evaluability (see **C**). This is shown for 66 patients who relapsed with ctDNA positive postoperative plasma . **E**. Validation of ECLIPSE detection rates across varying subclonal mutation number, clonal ctDNA level, subclone cancer cell fraction and DNA input amount into the assay. Subclones were constructed using ground truth in vitro spike-in experiments with 10-12 technical replicates for each input mass-allele fraction combination. These ground truth mutant allele fractions were then mixed in silico to construct 76,263 subclones varying across these parameters. Data from these experimentally derived subclones were then run through ECLIPSE and subclone detection rates across each of these parameters depicted.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Time-matched comparisons between subclonal structure measured in plasma and in tissue at surgery. A. Correlation between cancer cell fractions (CCFs) as measured in preoperative plasma samples with phylogenetic data, >0.1% clonal ctDNA level & >=10 ng DNA input (high subclone sensitivity samples) with ECLIPSE and those measured with multi-region tissue sequencing (M-seq) at surgery (N = 71 patients and 684 subclones included). **B**. Copy number unaware CCFs calculated only using VAFs (methods) compared to tissue CCF from M-seq. All preoperative samples with phylogenetic data, >0.1% clonal ctDNA level & >=10 ng DNA input (high subclone sensitivity samples) were included (N = 71 patients and 684 subclones included). **C**. A scatter plot demonstrating the relationship between clonal ctDNA level and the proportion of multi-region tumour exome (M-seq) defined subclones detected by ECLIPSE based on varying subclonal cancer cell fractions as indicated, loess lines are fitted to the plots, n = 117 ctDNA positive preoperative samples. **D**. A comparison of preoperative plasma CCFs and the average CCFs across all tissue regions sampled at surgery for clones that were unique to one tumour tissue region and for clones that were distributed across more than two tumour tissue regions. N = 71 patients and 684 subclones included. A Wilcoxon-test was used to compare groups. **E**. A comparison of preoperative plasma CCFs and the average CCFs across all tissue regions sampled at surgery for clones that were unique to one tumour tissue region separated between small (<20 cm^3), medium (>20 cm^3 & <100 cm^3), and large (>100 cm^3) tumours as measured on preoperative PET/CT scans. N = 71 patients and 684 subclones included. A Wilcoxon-test was used to compare groups. **F**. A comparison of detection rates in preoperative plasma for 20% CCF subclones across a range of clonal ctDNA levels split by whether the subclones were spread across multiple primary tumour tissue regions or were limited to only a single primary tumour tissue region. 1924 subclones were assessed in 197 preoperative plasma samples. **G**. A map of tumour clones with areas of multi-regional tissue sampling indicated and clones which are over- and undersampled highlighted. Most of the undersampled clones are in fact not in the sampled areas creating a bias towards oversampling in clones which we are able to detect, an effect also called the 'winner's curse'. **H**. A ROC curve describing the sensitivity and specificity of detecting clonal illusion mutations using plasma-based CCFs with 95% confidence intervals generated using bootstrapping across 500-fold cross-validation (N = 71 tumours).



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Clonal composition measurements in ctDNA after surgery. A. An overview of clonal structure evaluability at relapse for TRACERx patients in our cohort (N = 75 tumours) using either cell-free DNA and ECLIPSE or relapse tissue and WES/PyClone. **B**. ctDNA detection status post-operatively of subclones split by detection status in metastatic tissue. Untracked subclones (those without any mutations included in the PSP panels) were excluded (N = 26 tumours). P value indicates the result from Fisher's exact test. **C**. Clonal (estimated as present in 100% of tumour cells) vs subclonal (estimated as present in <100% of cells) status at relapse of primary tumour subclones by whether they were detected in cfDNA and metastatic tissue or cfDNA alone (N = 26 tumours). P value indicates the result from a Fisher's exact test. **D**. Metastatic dissemination class determined by tissue and by cfDNA in 22 cases with a metastatic biopsy, a postoperative high subclone sensitivity plasma sample, and a phylogenetic tree constructed. **E**. Overall survival Kaplan-Meier plot demonstrating time from the first MRD positive timepoint to death stratified by ECLIPSE metastatic dissemination class at relapse (monoclonal: light blue, polyclonal polyphyletic: purple, and polyclonal monophyletic: green). HR: Hazard ratio, CI: confidence interval. 44 patients were included in this analysis. The P value indicates the result of a log-rank test. **F**. A multivariable Cox proportional hazards model to predict overall survival from the time of first MRD detection including the clonality of metastatic dissemination at relapse, stage, maximum postoperative clonal ctDNA level, average DNA assay input, histology, and whether the first plasma sample after surgery was ctDNA positive, including only relapse patients. 44 patients were included in this analysis. Error bars indicate 95% confidence intervals. **G**. The frequency of high confidence subclonal to clonal bottlenecks (methods) at the latest possible plasma sample time point with sufficient clonal ctDNA level (high sensitivity subclone samples, N = 44 tumours) and which of these subclones harbour subclonal neoantigens (NAGs) which therefore become clonal at relapse. **H**. In cases of clonal bottlenecking at relapse, the percentage increase in the number of clonal mutations is shown as a box and whisker plot with the absolute number of new clonal mutations (N = 18 tumours). **I**. In cases of clonal bottlenecking at relapse, the percentage increase in the number of clonal NAGs is shown as a box and whisker plot with the absolute number of new clonal NAGs (N = 18 tumours). NAG = Neoantigen.

nature portfolio

Corresponding author(s):	Charles Swanton Nicholas McGranahan Chris Abbosh
Last updated by author(s):	15-01-2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data
Data analysis	Archer MRD calling algorithm (v0.1) Archer variant selection algorithm (v0.1) R version 4.1.2 (2021-11-01) Copy number analysis: GISTIC2 (version 2.0) Analytical validation: DescTools (version 0.99.44) ECLIPSE (version 1.0.0) R packages: fst (version 0.9.8) survival (version 3.2.13) survivalAnalysis (v0.3.0) survminer (version 0.4.9) ggpubr (version 0.4.0) ggrepel (version 0.9.2) stats (version 4.1.2)

tidyverse (version 1.3.21)
eulerr (version 6.1.1)
jaccard (version 0.1.0)
data.table (version 1.14.6)
readxl (version 1.4.1)
ggplot2 (version 3.3.5)
ggbeeswarm (version 0.6.0)
scales (version 1.2.1)
cowplot (version 1.1.1)
ggforce (version 0.4.1)
ComplexHeatmap (version 2.11.1)
ggplotify (version 0.1.0)
fishPlot (version 0.5)
cloneMap (version 1.0)
edgeR (version 3.36.0)
limma (version 3.50.31)
GSVA (version 1.42.0)
qusage (version 2.28.0)
ReactomePA (version 1.38.0)
rstatix (version 0.7.1)
lmerTest (version 3.1-3)
ROCIT (version 2.1.1)

All code to reproduce the figures will be available on request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The cfDNA sequencing files, RNA-sequencing data and multi-region tumour exome sequencing data (in each case from the TRACERx study), used or analysed during this study have been deposited at the European Genome–phenome Archive (EGA), which is hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under the accession codes (EGAS00001006494, EGAS00001006517, EGAS00001006494) and is under controlled access due to the nature of the data and commercial partnerships arrangements.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

421 patients have been analysed in the longitudinal TRACERx study so far (half-way point of the study) and we previously analysed circulating tumor DNA data from the first 100 TRACERx patients recruited to the study (Abbosh et al., 2017).

Here, we report analyses from 197 TRACERx patients. 197 patients had preoperative plasma samples analysed and 141 of 197 patients had both pre- and postoperative plasma samples analysed including 75 patients who experienced NSCLC relapse. The sample size was chosen based on availability of plasma for circulating tumor DNA analysis. The 197 patients include 169 of the 321 patients subsequently recruited to TRACERx following the first 100 patients, plus an additional 9 patients ineligible for the TRACERx 421 patient cohort, plus 19 patients previously analysed in the TRACERx 100 patient cohort in 2017 (preoperative analyses only). The 9 ineligible patients were excluded from the final TRACERx 421 cohort (for reasons including incompletely resected disease, C>A artefact in tumor exome data, non-lung synchronous primary at diagnosis) but were kept in this analysis since ctDNA analyses were performed prior to the patients being designated ineligible for TRACERx (as described in methods). The breakdown of the cohort is illustrated in Extended Figure 3a. This sample size facilitated assessment of ctDNA utility both in a pre- and postoperative setting across a cohort of patients exhibiting similar demographics to the whole TRACERx 421 cohort (see Supplementary Table 5).

TRACERx is a programme of work of multiple projects built around a single observational cohort study. It is not possible to perform a sample size calculation for each project, especially post hoc. A necessary study size for the complete cohort (n=842 patients) was calculated in relation to tumour heterogeneity and disease free survival:

The sample size is based on demonstrating a relationship between tumours with divergent intratumour heterogeneity index values and clinical outcome. Patients will be split evenly into those with a low and high intratumour heterogeneity index value (and other splits will be considered). Assuming a median Disease Free Survival (DFS) of 30 months and a hazard ratio (HR) of 0.77, with a 2-sided 5% significance level, 90% power, accrual period of 3 years and 5 years follow-up after the end of accrual, the sample size required is almost 400 per group (total of 800 patients). Assuming a 5% dropout rate, a total of 842 patients (421 per group) are required. At 85% power, 705 patients would be required in total, which could be the minimum target. However, we will instead aim for 750 patients and recruitment will continue for the length of time which is funded for accrual in order to get as close as possible to the ideal target of 842 patients. A study size of 842 is also large enough to detect a 10% improvement in a 5 year OS rate from 46% in the high Intratumour Heterogeneity Index (ITB) to 56% in the low Intratumour Heterogeneity Index group (HR=0.75), with 80% power and a 2 sided type I error set at 5% (logrank test). A high/low ITB value will be defined as values above/below the 50th percentile (median ITB). We have a target DFS effect of a 23% reduction in risk (hazard ratio 0.77), which means that our study is powered for an effect at least this large, including a 30% difference (which has been the target for progression-free survival in trials of advanced NSCLC, in relation to expected effects on OS).

Data exclusions

We excluded 26 of 1095 plasma samples analysed due to SNP ID mismatch with exome germline data indicating that these were sample swaps (i.e., were incorrectly assigned to a particular TRACERx ID at clinical sites). We also excluded 44 mutations from patient CRUK0297's patient specific panel as these were germline variants mistargeted by our industry partner in the pilot phase of our project (described in the supplementary note). These mutations were excluded since otherwise this patient could not have contributed to MRD calling threshold optimisation in the pilot phase of the project. Following analyses of 10 pilot patients, these 10 patients were excluded from all analyses evaluating clinical and biological associations with pre- and postoperative ctDNA detection since ctDNA detection thresholds were generated in these patients (CRUK0146,CRUK0050,CRUK0088,CRUK0046,CRUK0270,CRUK0094,CRUK0241,CRUK0297,CRUK0240,CRUK0367). However these patients were included in ECLIPSE analyses of subclonal kinetics.

In the non-pilot patients we excluded 37 of 187 patients from computed tomography (CT) volumetric tumour analyses for the following reasons: lung collapse, necrosis or major cavity within the tumor making volume measurements inaccurate or wide Z spacing (interval between CT slices) making volumetrics inaccurate, patients had synchronous primary cancers (volumetrics inaccurate) or no images available for volumetrics (reasons in non-pilot cohort summarised in extended fig 4a).

We also excluded non-pilot patients from survival analyses due to death within 30 days of surgery (n=5) or incompletely resected disease on postoperative imaging (n=4). Death within 30 days of surgery represents postoperative complication and patients with incompletely resected disease are ineligible for TRACERx survival endpoints, patients with synchronous primary cancers were excluded in Figure 1 preoperative ctDNA detection survival analyses due to emphasis on associations of ctDNA detection, outcome and tumor histology outlined in the manuscript (patients with synchronous primaries cannot be categorised as a single tumor histology), however these patients were included in Landmark MRD survival analyses since tumor histology was not considered.

To maintain high quality estimates of clonal deconvolution we limited our analyses of clonal structure to samples with at least 0.1% clonal ctDNA fraction and 10 ng of DNA input into library preparation for a sample. This threshold was chosen using validation of our subclone detection sensitivity (Supplementary figure 8 ae) using in vitro spike in data arranged in silico into mock subclonal mutation clones which estimated a 94% detection rate of 20% cancer cell fraction subclones (clone present in 1 in 5 cancer cells) at this threshold.

Replication

Both technical and in-silico replicates were used to analytically validate both the ctDNA detection assay and ECLIPSE. These experiments are described in detail in the Supplementary Note and methods. All attempts at technical and in-silico replication of results were successful.

Randomization

No randomization was conducted since this was an observational study, to control for covariates in survival analyses multivariable cox regression was used (including parameters such as pathological TNM stage, adjuvant therapy status, age, sequencing coverage and maximum ctDNA level).

Blinding

Blinding is not relevant as this is an observational study. Patients were not allocated to any intervention and they were followed up and assessed as per routine practice. No biomarker results (tissue and bloods) are reported back to patients, so there is no likelihood of people changing their behaviours based on these findings. The laboratory and radiological report review analyses were all performed without knowing the outcome (DFS or survival) status of the patients, which represents a form of blinding.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
□	☒ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

Baseline demographics describing the 197 patients analysed in this cohort are summarised in Supplementary Table 7.

Please note that the study started recruiting patients in 2016, when TNM version 7 was standard of care. The up-to-date inclusion/exclusion criteria now utilizes TNM version 8.

TRACERx inclusion and exclusion criteria

Inclusion Criteria:
_Written Informed consent
_Patients ≥18 years of age, with early stage I-IIIB disease (according to TNM 8th edition) who are eligible for primary surgery.
_Histopathologically confirmed NSCLC, or a strong suspicion of cancer on lung imaging necessitating surgery (e.g. diagnosis determined from frozen section in theatre)
_Primary surgery in keeping with NICE guidelines planned
_Agreement to be followed up at a TRACERx site
_Performance status 0 or 1
_Minimum tumor diameter at least 15mm to allow for sampling of at least two tumour regions (if 15mm, a high likelihood of nodal involvement on pre-operative imaging required to meet eligibility according to stage, i.e. T1N1-3)

Exclusion Criteria:
_Any other* malignancy diagnosed or relapsed at any time, which is currently being treated (including by hormonal therapy).
_Any other* current malignancy or malignancy diagnosed or relapsed within the past 3 years**.
*Exceptions are: non-melanomatous skin cancer, stage 0 melanoma in situ, and in situ cervical cancer
**An exception will be made for malignancies diagnosed or relapsed more than 2, but less than 3, years ago only if a pre-operative biopsy of the lung lesion has confirmed a diagnosis of NSCLC.
_Psychological condition that would preclude informed consent
_Treatment with neo-adjuvant therapy for current lung malignancy deemed necessary
_Post-surgery stage IV
_Known Human Immunodeficiency Virus (HIV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV) or syphilis infection.
_Sufficient tissue, i.e. a minimum of two tumor regions, is unlikely to be obtained for the study based on pre-operative imaging

Patient ineligibility following registration
_There is insufficient tissue
_The patient is unable to comply with protocol requirements
_There is a change in histology from NSCLC following surgery, or NSCLC is not confirmed during or after surgery.
_Change in staging to IIIC or IV following surgery
_The operative criteria are not met (e.g. incomplete resection with macroscopic residual tumors (R2)). Patients with microscopic residual tumors (R1) are eligible and should remain in the study
_Adjuvant therapy other than platinum-based chemotherapy and/or radiotherapy is administered.

Recruitment

When patients are initially diagnosed with stage I-III lung cancer and then referred for surgical resection, a research nurse identifies them on a clinic/operating list. The patient has an initial eligibility assessment and then provided with written information about the TRACERx study and he/she can ask the research nurse any questions.

Patients have to agree to provide serial blood samples whenever they attend clinic for routine blood sampling, so this represents the only main potential self-selecting bias (i.e. only patients willing to do this would participate). However, it is unclear how this would affect the biomarker analyses. Also, the gender and ethnicity characteristics are in line with patients seen in routine practice.

Inclusion and exclusion criteria are summarised above.

Ethics oversight

The study was approved by the NRES Committee London with the following details:
Study title: TRAcking non small cell lung Cancer Evolution through therapy (Rx)
REC reference: 13/LO/1546
Protocol number: UCL/12/0279
IRAS project ID: 138871

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies
All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

TRACERx Lung https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent Research Ethics Committee, 13/LO/1546

Study protocol

https://clinicaltrials.gov/ct2/show/NCT01888601

Data collection

Clinical and pathological data is collected from patients during study follow up - this period is a minimum of five years. Data collection is overseen by the sponsor of the study (Cancer Research UK & UCL Cancer Trials Centre) and takes place in hospitals across the United Kingdom. A centralised database called MACRO is used for this purpose. Recruitment to TRACERx started in April 2014 and is still ongoing in London and Manchester.

Outcomes

The pre-defined clinical outcome analysed in this manuscript is overall survival (OS): measured from the time of study registration to date of death from any cause. This outcome was pre-defined in the TRACERx protocol (described in Jamal-Hanjani et al., 2017 NEJM) and is described in methods section survival analyses. Additional survival outcomes analysed in this manuscript are: freedom from recurrence (FFR, events were lung cancer recurrence, patients disease-free or experiencing second-primary or death were right censored at last follow-up) and post-relapse survival (time from recurrence to death from any cause).

Part V

MANUSCRIPT II

Exploring the biology of ctDNA release in colon cancer

Exploring the biology of ctDNA release in colon cancer

*Judit Kisistók**, Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital; Bioinformatics Research Centre, Aarhus University

*Laura Andersen**, Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital; Bioinformatics Research Centre, Aarhus University

*Tenna Vesterman Henriksen**, Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital

*Jesper Bertram Bramsen**, Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital

Thomas Reinert, Department of Clinical Medicine - Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital

Nadia Øgaard; Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital

Trine Block Mattesen, Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital

Nicolai Juul Birkbak#, Department of Clinical Medicine - Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital; Bioinformatics Research Centre, Aarhus University

Claus Lindbjerg Andersen#, Department of Clinical Medicine, Aarhus University; Department of Molecular Medicine, Aarhus University Hospital

*These authors contributed equally to this work.

#These authors jointly supervised this work.

INTRODUCTION

Colorectal cancer (CRC) is one of the leading causes of cancer-related mortality, accounting for >900,000 deaths each year worldwide [1]. Generally, CRC develops through a gradual accumulation of mutations, transforming the healthy bowel epithelium to cancer [2,3]. These growths are benign until they penetrate the muscularis mucosa, after which they are classified according to the Union for International Cancer Control (UICC) staging system (stage I-IV). The patient's prognosis is highly correlated to the UICC stages, with decreasing survival with increasing stage [4]. Additionally, CRC prognoses vary by tumor location and histological type. A large proportion of CRC tumors are characterized by chromosomal instability (CIN), while a smaller proportion (~15%) are characterized by microsatellite instability (MSI) [5]. Due to the heterogeneity of the disease, research has been done to devise a clinically relevant method for patient stratification. Gene expression-based subtypes have been widely explored, leading to the development of Consensus Molecular Subtypes (CMS), accelerating disease classification in a biologically interpretable manner [6].

In addition to clinically relevant patient stratification, early cancer detection and detection of minimal residual disease (MRD) has the potential to improve patient outcomes. Circulating tumor DNA (ctDNA) has garnered interest as an efficient and minimally invasive tool for these purposes [7–9]. Most studies to date have focused on the clinical application of ctDNA. However, little is currently known about the biology behind ctDNA release and varying levels of ctDNA shedding across cancer types and subtypes.

It is hypothesized that ctDNA is released into the bloodstream through apoptosis, necrosis, and active secretion [7]. Yet, differences in shedding behavior across cancer types as well as across histology within cancer types have been identified in previous publications [10–12]. Particularly in non-small cell lung cancer, work has recently demonstrated how in a subset of lung adenocarcinomas dominated by a low proliferative phenotype, ctDNA shedding was dramatically reduced relative to other histological subtypes{REF new paper}. Besides cancer specific biology, ctDNA release has previously been associated with tumor size, proliferative capacity, rate of cell death, proximity to blood vessels, rate of immune clearance, physiological clearance, and CIN [13,14]. However, the exact contributors on a granular, cancer type- and histology-specific level have not been fully explored.

Technological sensitivity limits pose a challenge to the widespread application of ctDNA in an early detection setting. This raises the question of whether tumors shedding ctDNA in small quantities can be robustly detected.

Here, we investigate the presence of a ctDNA shedding phenotype in a CRC cohort, by analyzing transcriptomic, genomic, and clinical data collected from Stage I-IV CRC patients at Aarhus University Hospital. Additionally, in order to further investigate ctDNA shedding phenotypes across multiple cancer types, we compared transcriptomic profiles of CRC tumors to those of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) tumors from the TRACERx cohort [15], and to the transcriptomic profiles of 6900 tumors representing 24 cancer types available from the Cancer Genome Atlas (TCGA).

RESULTS

Patient and sample characteristics

To investigate the molecular and clinical features associated with ctDNA release, we assembled a cohort of 701 stage I-IV colorectal cancer patients treated locally at Aarhus University Hospital (AUH) with standard-of-care protocols. For these patients, ctDNA measurements were performed prior to treatment (444 ctDNA positive, 257 ctDNA negative). Primary tumor whole exome sequencing (WES) and extended clinical data were available for all patients (Figure 1A, *whole cohort*). Additionally, primary tumor transcriptomic data were available for a subset of patients (n=101, 86 ctDNA positive, 15 ctDNA negative), (Figure 1A, *subcohort*). The ctDNA detection status and clinical characteristics of the subcohort are summarized in Figure 1B, the whole cohort is described in Figure 1C.

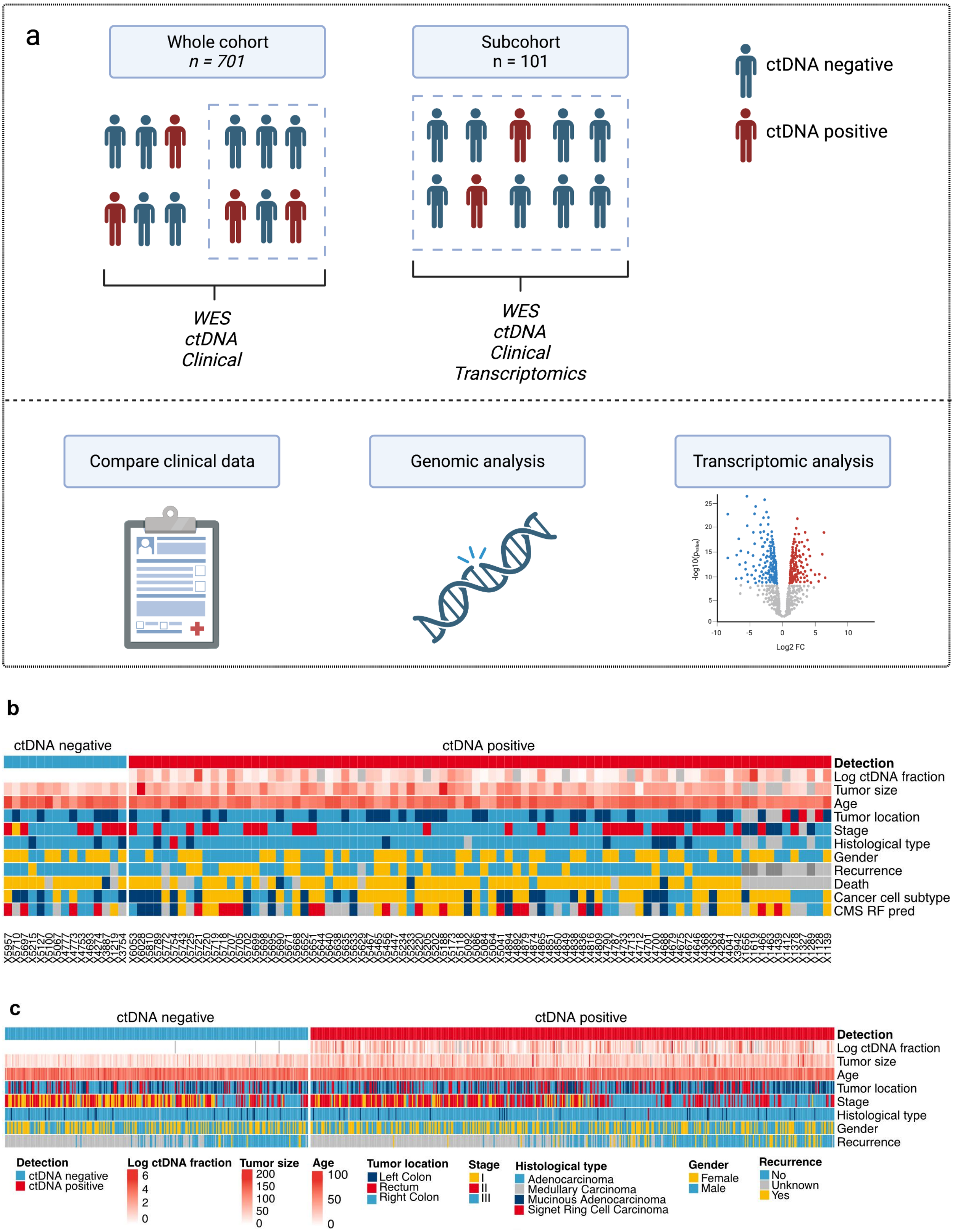


Figure 1: *Overview of patient cohorts. Clinical characteristics and ctDNA status of the subcohort. Figure 1a was created with BioRender.com.*

In order to gain a deeper insight into the contribution of individual clinical characteristics to ctDNA shedding, we performed pairwise statistical analyses comparing ctDNA positive and ctDNA negative patients in the subcohort (Table 1), and repeated the analyses in the whole cohort (Supplementary Table 1). Markedly, we found that the size of the ctDNA positive tumors was significantly larger than that of their ctDNA negative counterparts. The association remained statistically significant within Stage I, Stage II and Stage III patients (Figure 2A-D).

In addition to tumor size, we also observed an association between ctDNA shedding status and molecular subtype, as defined by the consensus molecular subtyping and cancer cell subtyping classifications[6,16]. Specifically, we observed that the Secretory and CMS3 subtypes appear to shed a lower amount of ctDNA compared to other subtypes (Figure 2E-F). Additionally, we noted a significant association between ctDNA detection status and recurrence, showing that ctDNA was detected in a larger fraction of recurrence than non-recurrence patients (Figure 2G). We also observed a significant association between tumor location and ctDNA detection status. However, we found no difference in ctDNA concentration when comparing tumor location in a pairwise manner (Figure 2g, Supplementary Figure 2a). Furthermore, we noted that mismatch repair (MMR) deficient tumors shed a significantly higher amount of ctDNA into the bloodstream compared to their MMR proficient counterparts (Figure 2H).

We found no association between preoperative ctDNA shedding MSI status, histological type, age, gender, and death (Supplementary Figure 2b-f).

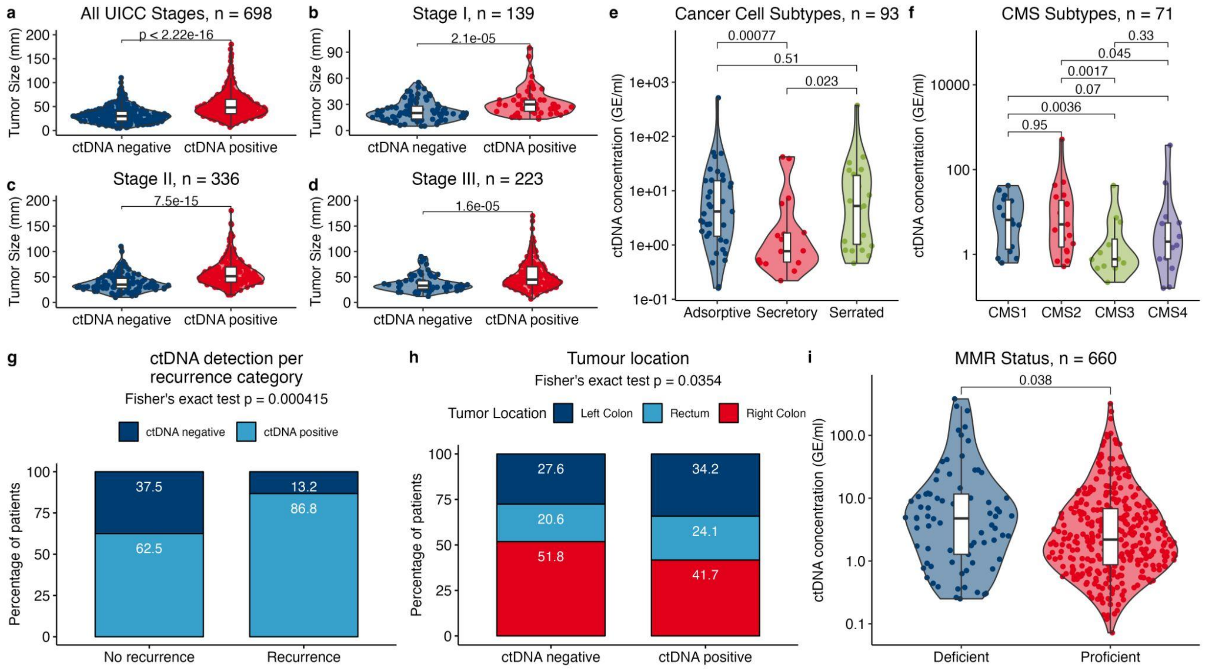
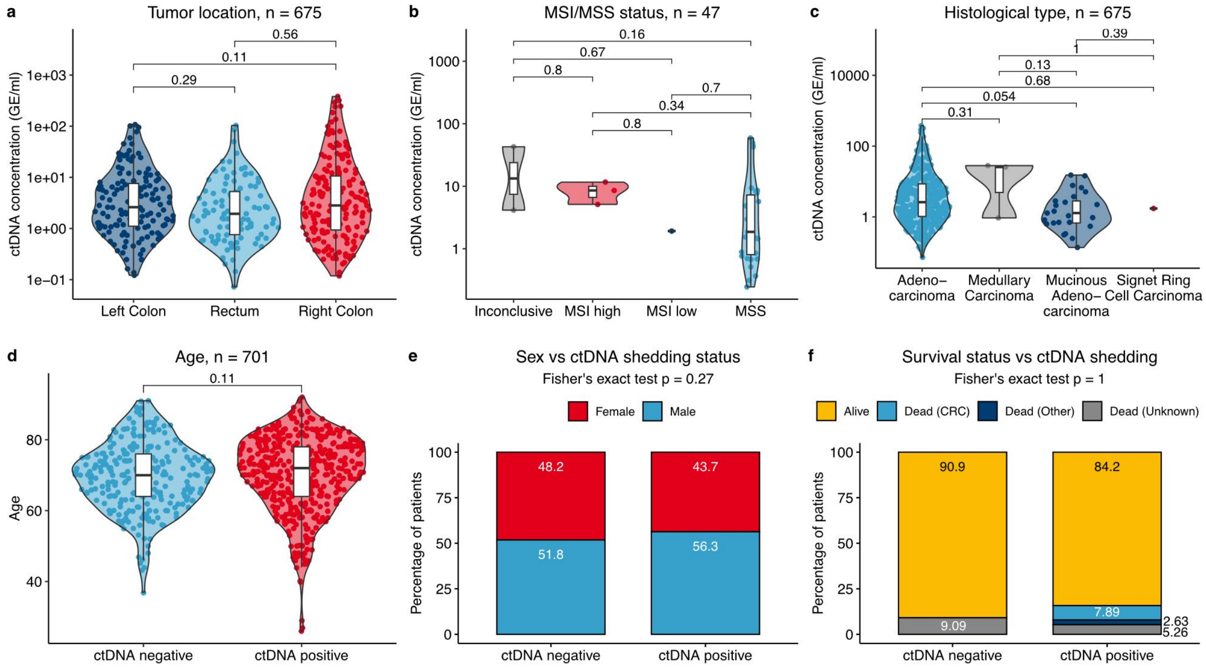


Figure 2: *Comparison of tumor size between ctDNA shedders/non-shedders. Comparison of ctDNA shedding between tumor subtypes,recurrence category, tumor location and MMR status.*



Supplementary Figure 2. *Comparison of tumor location, MSI/MSS status, and histological type against ctDNA concentration. Comparison of ctDNA shedding status against age, sex, and survival status.*

Enhanced proliferative signal can be observed in the ctDNA positive subgroup

Next, we aimed to investigate whether mutations in specific cancer driver genes contribute to ctDNA shedding. Utilizing the WES data collected from the whole cohort, we defined driver mutations by identifying pathogenic mutations in cancer-related genes as described in [17,18].

No cancer gene showed an association between mutational status and ctDNA shedding (Figure 3A, *top 20 most frequently mutated genes shown*). Next we explored if mutations at pathway-level affected ctDNA shedding. Gene mutations were assigned to pathways according to classification previously defined by Sanchez-Vega and colleagues [19]. No significant associations were observed (Figure 3B). Taken together, no evidence was found for a link between mutations in specific genes or pathways and ctDNA shedding in colorectal cancer.

Differential expression analysis of the 86 ctDNA positive and 15 ctDNA negative patients with available transcriptomic profiles did not identify any genes with expression patterns associated with ctDNA status (Figure 3C). However, when we summarized gene expression levels using the Hallmark gene sets from MsigDB to group genes into pathways [20], we observed that a number of proliferative pathways (E2F targets, G2M checkpoint, and MYC targets V1 and V2) were significantly enriched in the ctDNA positive subgroup (Figure 3D). When visualizing the per-patient enrichment scores of the significant pathways, we noted that two distinct shedder subgroups exist in terms of proliferative activation (Figure 3E). We observed that the plasma ctDNA level was significantly higher in high-proliferation relative to low-proliferation ctDNA shedders (Figure 3F). Moreover, high proliferation shedders were significantly larger than non-shedders, but not low proliferation shedders (Figure 3G). Additionally, we noted a significantly different cancer cell and CMS subtype profile among the groups, with the high-proliferation shedders showing an enrichment in Adsorptive and CMS2 subtypes, whereas the low-proliferation shedders, similarly to the nonshedders, displayed the Secretory, CMS3, and CMS4 subtypes with higher frequency (Figure 3H-I).

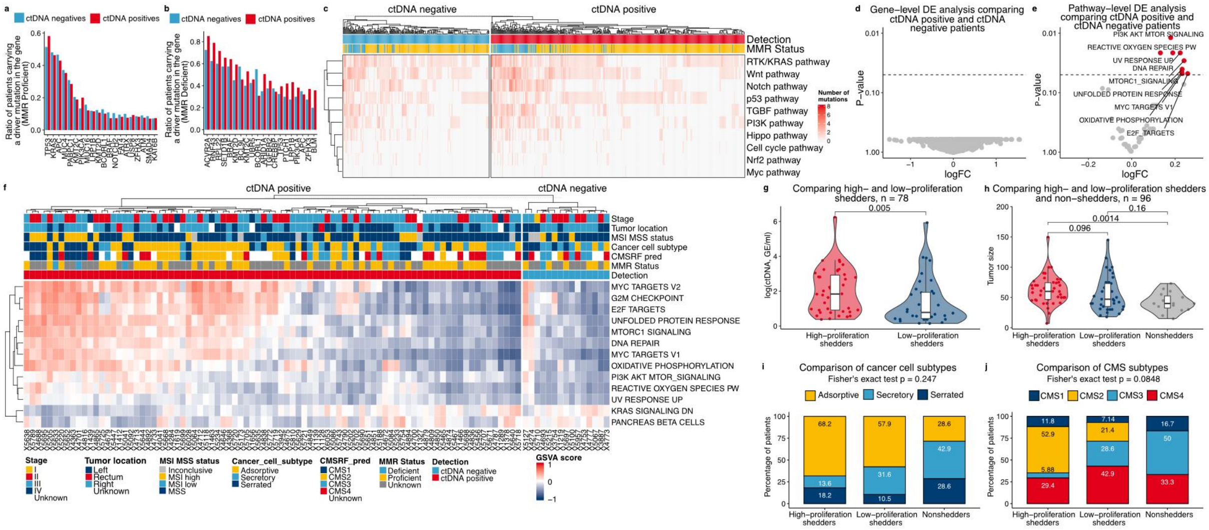


Figure 3: *Mutational and transcriptomic analysis comparing ctDNA shedders with non-shedders.*

Taken together, our findings suggest that the main drivers of ctDNA release in colorectal cancer are tumor size and tumor proliferative capacity. As these findings are similar to previous work in non-small cell lung cancer (NSCLC) (Abbosh et al. 2023), we endeavored to compare colon cancer biology to that of the ctDNA low-shedder and ubiquitous shedder NSCLC histologies.

Proliferation pathway enrichment shows resemblance in CRC and LUSC biology

We compared transcriptomic profiles of 228 colon adenocarcinoma (COAD), 458 LUAD, and 475 LUSC tumors using the TCGA dataset. These were additionally compared to transcriptomic profiles obtained from 339 healthy colon and 313 healthy lung tissues from The Genotype-Tissue Expression (GTEx) project [21,22].

When evaluating proliferation (quantified by the median GSVA enrichment scores of E2F targets, G2M checkpoint, MYC targets V1 and V2 pathways), we found that COAD tumors display a high-proliferative phenotype similar to LUSC tumors (FIgure 4A). Interestingly, we observed that the LUAD tumors display a bimodal distribution of high and low proliferation, supporting the existence of a distinct subset of low-proliferative LUAD tumors (Supplementary figure 2). Furthermore, the high-proliferation LUAD tumors showed proliferation levels comparable to LUSC and COAD, whereas proliferation levels in the low-proliferation subgroup were comparable to healthy colon and lung tissue (Figure 4A). Performing principal component analysis (PCA) based on the proliferation pathways revealed a clustering pattern where COAD and LUSC tumors distinctly separated from healthy colon and lung tissue, whereas LUAD tumors were dispersed between the two clusters (Figure 4B).

Next, we explored whether these patterns were present in our colorectal cancer and previously published NSCLC datasets. This dataset is from [15] and contains 58 LUAD and 31 LUSC patients with NSCLC. In line with the TCGA analysis, we observed that colorectal and LUSC tumors showed high levels of proliferation relative to LUAD tumors (Figure 4C). Furthermore, PCA based on the proliferation pathways aligned with the results obtained using the TCGA dataset, as a similar partition of a CRC-LUSC cluster and a distinct LUAD cluster were found (Figure 4D).

These findings support that COAD and LUSC tumors are ubiquitous ctDNA shedder tumors with a high-proliferative phenotype, whereas among LUAD tumors, there exists a distinct non-shedder, low-proliferation phenotype.

Proliferation associates with ctDNA detection sensitivity across multiple cancer types

These findings led us to explore the association between proliferation and ctDNA release in the context of multiple cancer types. We obtained a ctDNA sensitivity measure from the Circulating Cell-free Genome Atlas (CCGA) study [23] of multi-cancer ctDNA detection and compared it to the proliferation profiles of 8095 tumors of 24 different cancer types from TCGA. The TCGA tumors were grouped to match the cancer types of CCGA according to Supplementary table 2. We observed a significant positive correlation between ctDNA sensitivity and proliferation (Figure 4E, Supplementary Table 3), with Liver/bile-duct and pancreas appearing as strong outliers. Notably, there is a distinct high-proliferation, high-sensitivity group of cancer types in contrast with a low-proliferation, low-sensitivity group. When performing stage stratification, the association loses statistical significance for Stage I and Stage II patients, however, this is likely driven by smaller tumor sizes and the two outlier cancer types noted above (Figure 4F-G). The association remains significant for Stage III patients (Figure 4H).

These results suggest that proliferation may perform as a significant biological contributor to ctDNA shedding on a pan-cancer level.

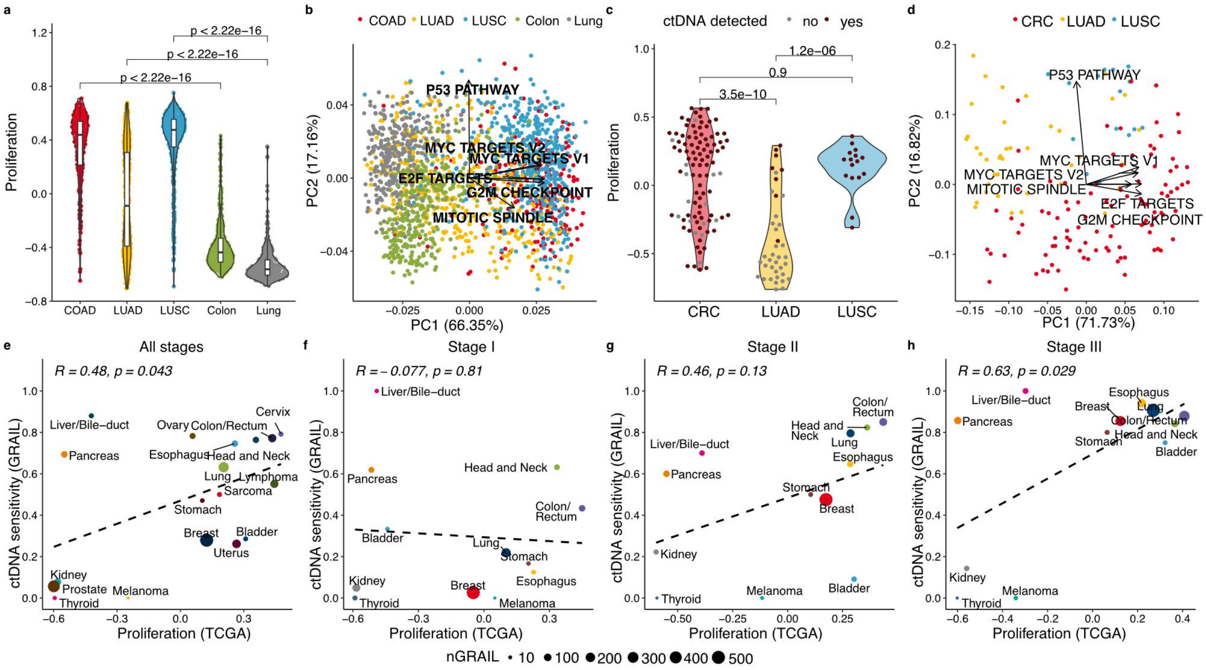
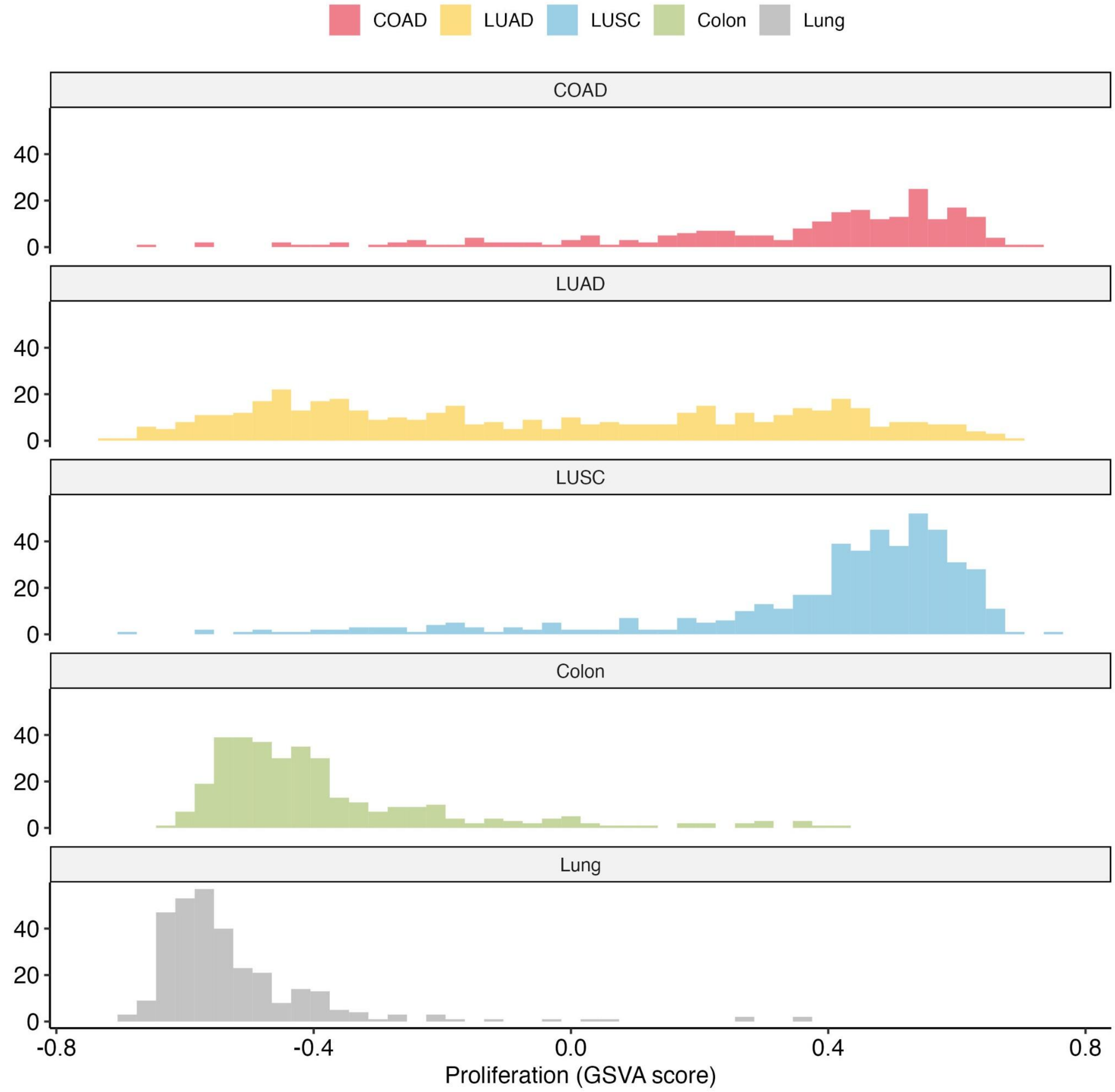


Figure 4: *Comparison proliferation of NSCLC tumors with CRC and healthy tissue. Expanded to pan-cancer level comparing ctDNA sensitivity (CCGA) to proliferation.*



Supplementary Figure 2: *Comparing proliferation between NSCLC, CRC and healthy tissue.*

DISCUSSION

It is generally accepted that ctDNA release originates from cancer cell deaths, and varying levels of ctDNA shedding is commonly attributed to tumor size. While this hypothesis has been confirmed in some cancer types [13,15], other cancer types and subtypes tend to show no ctDNA shedding despite remarkable tumor sizes. Therefore, in this study, we investigated clinical and biological factors contributing to ctDNA shedding in colorectal cancer and further extended our exploration to NSCLC and to the pan-cancer level.

In our locally treated CRC cohort, we observed that ctDNA release is primarily associated with tumor size and might be supported by enrichment in proliferative pathways. We note, however, that tumor's largest diameter, for non-spherical tumors can be an inaccurate proxy of tumor size. Nevertheless, we see a significant association between ctDNA shedding behavior and tumor size as defined in our dataset.

In addition to the association with tumor size and proliferation, we observed that the CMS3 and Secretory subtypes, ones that are associated with less aggressive disease, tend to shed lower amounts of ctDNA. These subtypes are characterized by an enrichment of KRAS mutations and a lower proliferative capacity, suggesting increased metabolic adaptation which might aid prompt clearing of ctDNA from the bloodstream [6 16].

Through our comparative analysis of ctDNA negative and ctDNA positive tumors, we found no other association to clinical factors nor genetic alterations in specific cancer driver genes or pathways. Therefore, based on our analysis, we hypothesize that the main contributor to ctDNA shedding in colorectal cancer is tumor size and proliferative capacity. Furthermore, we postulate that the ctDNA negative tumors in our cohort might release ctDNA in quantities that fall below the limit of detection of the technologies used in our experiments, due to small tumor size and sub-par proliferative capacity that cannot adequately assist the tumor in reaching a sufficiently high shedding rate. Consequently, CRC tumors appear to be ubiquitous shedders, mirroring the findings of Abbosh and colleagues in the context of LUSC tumors.

In addition to analyzing our in-house CRC dataset, we explored a broader view of the proliferative capacity of colon and lung tumors utilizing the TCGA dataset. Our results support previous findings by Abbosh and colleagues [15], indicating that two LUAD subtypes exist, defined by high and low proliferation, respectively. This phenotypic separation is in contrast with the biology of COAD and LUSC, where all tumors appear to be highly proliferative. Comparing proliferation measures between our CRC dataset and the NSCLC dataset from the TRACERx study, we found a similar pattern suggesting that ctDNA shedding is strongly associated with proliferation in LUAD tumors, whereas in LUSC and COAD, all tumors appear highly proliferative and ctDNA shedding is mainly determined by tumor size.

The exploration conducted using the TCGA and GTEx datasets supports the hypothesis of two, high- and low-proliferation subtypes, however, since these datasets do not have any ctDNA information, these analyses are limited to speculative hypotheses about ctDNA shedding. However, the results support that the findings pertaining to the high and low proliferation subtypes in our CRC and NSCLC cohorts are not due to small sample sizes and potential batch effects between the datasets.

Comparing the proliferation levels of the different cancer types included in the TCGA dataset to the ctDNA sensitivity metrics of these cancer types in the CCGA [23] study revealed that the hypothesized association between proliferation and ctDNA shedding potentially expands to multiple cancer types. The correlation, however, is most prominent in Stage III patients, most likely due to the small tumor sizes harbored by Stage I patients and the effect of the two strong outlier cancer types, pancreas and liver/bile-duct. Additionally, the cancer type grouping implemented in the CCGA study might also introduce a bias, as, for instance, LUAD and LUSC, two histologies that are vastly different in terms of biology and ctDNA shedding, are grouped into one lung category in CCGA. Based on our results, these should be separated when analyzing ctDNA shedding and, consequently, we hypothesize that in the absence of this separation, the ctDNA sensitivity estimate of this cancer type might be biased depending on the LUAD/LUSC ratio in CCGA. Analogously, other cancer types with potentially similar patterns would benefit from histology-specific measurements. Therefore, this analysis can be used as an indicator rather than a conclusion that proliferation might be a determining factor between ctDNA shedding and non-shedding phenotypes across cancer types.

In conclusion, our results demonstrate that enhanced proliferative capacity in connection with tumor size may contribute to a higher ctDNA release rate. We anticipate our study to be a starting point for further investigation of ctDNA shedding dynamics in colorectal cancer.

METHODS

Cohort description |

Colorectal Cancer cohort: The CRC patients included in this study is the cohort described in Kabel et al. [14]. Inclusion criteria were UICC stage I-III disease, preoperative ctDNA measurements, and WES of matched primary tumor and buffycoat DNA. For a detailed description of the cohort, sample collection procedure and DNA extraction see Kabel et al. [14].

TRACERx: The NSCLC patients included in this study are from the TRACERx cohort of abbosh et al. [15]. The data was filtered to include only patients with LUAD and LUSC.

The Cancer Genome Atlas (TCGA): Clinical and gene expression data were obtained from TCGA. These were filtered to include primary tumors only and tumors of clinical stage I-III. Clinical stage was grouped as follows; Stage I, Stage IA and Stage IB tumors were grouped into Stage I. Stage II, Stage IIA, Stage IIB and Stage IIC were grouped into Stage II. Stage III, Stage IIIA, Stage IIIB and Stage IIIC were grouped into Stage III. Additionally, cancer types were grouped to match the cancer types of CCGA. Groupings can be found in Supplementary table 3.

The Genotype-Tissue Expression (GTEx): Gene expression data from healthy Colon and healthy Lung tissue were obtained from GTEx [21,22].

The Circulating Cell-free Genome Atlas (CCGA): Sensitivity measures for all included cancer types were obtained from Supplementary table 5 of [23]. Cancer types were filtered to include stage I-III patients only.

RNA sequencing | Total RNA libraries were constructed using KAPA mRNA HyperPreb Kit (Roche) according to the manufacturer's protocol, with either KAPA Dual-Indexed Adapter Kit (Roche) or IDT xGEN dual index UMI adapters (Integrated DNA Technologies (IDT), Inc.). The libraries were sequenced by paired-end sequencing (2x151 base pairs (bp)) on the NovaSeq 6000 platform (Illumina).

In the preprocessing step, adapters were trimmed using cutadapt [24] and subsequently aligned to hg38 with STAR[25]. Arriba [26] and STAR-fusion [27] were used to call gene fusions. The aligned reads were processed with Kallisto [28] and HTseq [29] in order to generate read count and TPM expression values.

Whole Exome Sequencing | DNA libraries were prepared from Peripheral Blood Mononuclear Cells (PBMC), FFPE embedded tissue, or Fresh Frozen (FF) tissue using Twist Library Preparation kit with xGen UDI-UMI adapters (IDT). Prior to preparation, DNA was processed with enzymatic fragmentation for either 10 min (PBMC and FF) or 6 min (FFPE). Libraries were amplified with 7 or 8 cycles of PCR and subsequently captured with the NGS Human Core Exome panel (TWIST Bioscience, ~33 MB). Libraries were sequenced with paired-end sequencing (2x151 bp) on the NovaSeq 6000 platform to a targeted sequencing depth of either 60x (PBMC), 130x (FF tissue) or 150x (FFPE tissue). Reads were demultiplexed using Illumina bcl2fastq to generate fastQ files.
In the preprocessing step, adapters were trimmed using cutadapt [24] (version 3.0) and subsequently aligned to hg38 using BWA-MEM [30] (version 0.7.17). PCR duplicates were removed using Picard [31] (version 2.23.3) MarkDuplicates. Base recalibration and indel realignment was performed using BaseRecalibrator and IndelRealinger from GATK4 [32] (version 4.1.9.0). Somatic SNVs and insertions/deletions were called using GATK4 [32] (version 4.1.9.0) MuTect2 according to GATK best practices. Furthermore, variants were called with Stelka2 [33] (version 2.9.10) and variants discarded by the builtin filters of Mutect2 were retained if they passed the filters of Strelka2.

ctDNA detection |

The ctDNA detection methodology and data included in this paper have been published elsewhere in their own right. Detailed methodology is described in the individual papers. Overall, ctDNA analysis was conducted using either droplet digital PCR (ddPCR), deep targeted sequencing of 12 genes frequently mutated in CRC[14], or Signatera ultradeep multiplex PCR sequencing [34,35]. ddPCR was performed targeting a single patient-specific clonal mutation [14], a single somatic structural variation [8] or three CRC-specific methylation markers (the TriMeth assay)[36]. For detailed description of cfDNA sequencing and bioinformatic preprocessing see Kabel et al[14].

Transcriptomic analysis | Differential gene expression and differential pathway enrichment analyses between ctDNA positive and ctDNA negative patients were conducted by performing limma eBayes. The resulting p-values were adjusted according to the Benjamini-Hochberg method. The genes in the transcriptomic analyses were assigned to pathways with regard to the MSigDB Hallmark gene sets and pathway enrichment was assessed using Gene Set Variation Analysis (GSVA)[20].
GSVA were performed using the gsva function of the GSVA R package [37] min.sz set to 10 and max.sz set to 500. The analysis takes into account 18843 protein coding genes. The list of protein coding genes was retrieved from the HGNC database on April 3rd, 2022.

Genomic analysis | All somatic mutations were annotated to genes using ANNOVAR and the hg38 reference genome. Genes were assigned to pathways using the Sanchez-Vega definition[19]. Comparison of pathway involvement of ctDNA positive and ctDNA negative patients was conducted using Fisher's exact test.

Statistical analysis | All analysis was performed in R version 4.1.2. Visualizations were created using ggplot2 v3.4.2, ComplexHeatmap v2.11.1, and ggAU v1.0.0. Statistical tests were performed using the ggpubr v0.6.0 [38] R package. Differences in clinical and biological associates between ctDNA positive and ctDNA negative patients were tested using Wilcoxon rank sum test. Associations between categorical variables were tested using Fisher's exact test using the stats package (v4.1.2). Correlations between numerical variables were tested using the Pearson correlation coefficient. PCA was performed with the prcomp R function with scale set to TRUE. A P-value below 0.05 was considered significant throughout the analysis.

REFERENCES

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
3. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet* **383**, 1490–1502 (2014).
4. Osterman, E. & Glimelius, B. Recurrence Risk After Up-to-Date Colon Cancer Staging, Surgery, and Pathology: Analysis of the Entire Swedish Population. *Dis. Colon Rectum* **61**, 1016–1025 (2018).
5. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6**, 479–507 (2011).
6. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
7. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).
8. Reinert, T. *et al.* Analysis of circulating tumour DNA to monitor disease

burden following colorectal cancer surgery. *Gut* **65**, 625–634 (2016).

9. Schøler, L. V. *et al.* Clinical Implications of Monitoring Circulating Tumor DNA in Patients with Colorectal Cancer. *Clin. Cancer Res.* **23**, 5437–5445 (2017).

10. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).

11. Ørntoft, M.-B. W. *et al.* Age-stratified reference intervals unlock the clinical potential of circulating cell-free DNA as a biomarker of poor outcome for healthy individuals and patients with colorectal cancer. *Int. J. Cancer* **148**, 1665–1675 (2021).

12. Abbosh, C. *et al.* Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. *Nature* 1–10 (2023).

13. Avanzini, S. *et al.* A mathematical model of ctDNA shedding predicts tumor detection size. *Science Advances* **6**, eabc4308 (2020).

14. Kabel, J. *et al.* Impact of Whole Genome Doubling on Detection of Circulating Tumor DNA in Colorectal Cancer. *Cancers* **15**, (2023).

15. Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).

16. Bramsen, J. B. *et al.* Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. *Cell Rep.* **19**, 1268–1280 (2017).

17. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

18. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).

19. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).

20. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

21. Ward, L. D., Kheradpour, P., Iriarte, B. & Kamvysselis, M. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. (2015).

22. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

23. Klein, E. A. *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**, 1167–1177 (2021).

24. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

25. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

26. Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).

27. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection

from RNA-Seq. *bioRxiv* 120295 (2017) doi:10.1101/120295.

28. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 888 (2016).

29. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945 (2022).

30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* vol. 25 1754–1760 Preprint at https://doi.org/10.1093/bioinformatics/btp324 (2009).

31. Institute, B. Picard tools. (2016).

32. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

33. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).

34. Reinert, T. *et al.* Analysis of Plasma Cell-Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer. *JAMA Oncol* **5**, 1124–1131 (2019).

35. Henriksen, T. V. *et al.* Circulating Tumor DNA in Stage III Colorectal Cancer, beyond Minimal Residual Disease Detection, toward Assessment of Adjuvant Therapy Efficacy and Clinical Behavior of Recurrences. *Clin. Cancer Res.* **28**,

507–517 (2022).

36. Jensen, S. Ø. *et al.* Novel DNA methylation biomarkers show high sensitivity and specificity for blood-based detection of colorectal cancer-a clinical biomarker discovery and validation study. *Clin. Epigenetics* **11**, 158 (2019).

37. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

38. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. Preprint at https://CRAN.R-project.org/package=ggpubr (2022).

Table 1. Subcohort description.

Characteristic	**N**	**ctDNA negative**, N = 15[1]	**ctDNA positive**, N = 86[1]	**p-value**[2]
log_ctDNA	93	0.00 (0.00, 0.00)	1.27 (0.59, 2.67)	<0.001***
Tumor_size	96	40 (32, 52)	55 (40, 74)	0.010*
Tumor_location	96			>0.9
Left Colon		6 (40%)	33 (41%)	
Rectum		0 (0%)	3 (3.7%)	
Right Colon		9 (60%)	45 (56%)	
Stage	100			0.2
I		1 (6.7%)	0 (0%)	
II		7 (47%)	31 (36%)	
III		7 (47%)	49 (58%)	
IV		0 (0%)	5 (5.9%)	
Histological_type	96			0.6
Adenocarcinoma		12 (80%)	71 (88%)	
Mucinous Adenocarcinoma		3 (20%)	8 (9.9%)	
Others		0 (0%)	1 (1.2%)	
Signet Cell Carcinoma		0 (0%)	1 (1.2%)	
age	101	75 (64, 78)	70 (62, 76)	0.3
gender	101			0.11
Female		10 (67%)	38 (44%)	
Male		5 (33%)	48 (56%)	
Recurrence	92			0.14
Irrelevant		0 (0%)	5 (6.4%)	
No		13 (93%)	52 (67%)	

[1] Median (IQR); n (%)

[2] *p<0.05; **p<0.01; ***p<0.001

Characteristic	**N**	**ctDNA negative**, N = 15[1]	**ctDNA positive**, N = 86[1]	**p-value**[2]
Yes		1 (7.1%)	21 (27%)	
Death	87			0.5
Alive		12 (86%)	57 (78%)	
Dead (CRC)		0 (0%)	8 (11%)	
Dead (Other)		1 (7.1%)	3 (4.1%)	
Dead (Unknown)		1 (7.1%)	5 (6.8%)	
Cancer_cell_subtype	101			0.056
Adsorptive		3 (20%)	43 (50%)	
Secretory		7 (47%)	20 (23%)	
Serrated		5 (33%)	23 (27%)	
CMSRF_pred	77			0.12
CMS1		1 (9.1%)	15 (23%)	
CMS2		1 (9.1%)	22 (33%)	
CMS3		5 (45%)	13 (20%)	
CMS4		4 (36%)	16 (24%)	

[1] Median (IQR); n (%)

[2] *p<0.05; **p<0.01; ***p<0.001

Supplementary Table 1. Whole cohort description.

Characteristic	N	**no**, N = 257[1]	**yes**, N = 444[1]	p-value[2]
ctDNA_konc	675	0 (0, 0)	3 (1, 8)	<0.001***
Tumor Size (mm)	698	30 (20, 40)	48 (35, 65)	<0.001***
Tumor Location	701			0.034*
Left Colon		71 (28%)	152 (34%)	
Rectum		53 (21%)	107 (24%)	
Right Colon		133 (52%)	185 (42%)	
UICC Stage	701			<0.001***
1		88 (34%)	53 (12%)	
2		110 (43%)	226 (51%)	
3		59 (23%)	165 (37%)	
Tumor Type	701			0.7
Adenocarcinoma		236 (92%)	414 (93%)	
Medullary Carcinoma		1 (0.4%)	3 (0.7%)	
Mucinous Adenocarcinoma		20 (7.8%)	26 (5.9%)	
Signet Ring Cell Carcinoma		0 (0%)	1 (0.2%)	
Age	701	70 (64, 76)	72 (64, 78)	0.11
Sex	701			0.2
Female		124 (48%)	194 (44%)	
Male		133 (52%)	250 (56%)	
Recurrence	701			0.001**
NA		146 (57%)	225 (51%)	
No		104 (40%)	173 (39%)	
Yes		7 (2.7%)	46 (10%)	
MMR Status	701			0.5

[1] Median (IQR); n (%)

[2] *p<0.05; **p<0.01; ***p<0.001

Characteristic	**N**	**no**, N = 257[1]	**yes**, N = 444[1]	**p-value**[2]
Deficient		40 (16%)	81 (18%)	
NA		7 (2.7%)	8 (1.8%)	
Proficient		210 (82%)	355 (80%)	
log_ctDNA	675	0.00 (0.00, 0.00)	1.27 (0.70, 2.24)	<0.001***
MSI_MSS_status	49			>0.9
Inconclusive		0 (0%)	2 (5.3%)	
MSI_high		1 (9.1%)	4 (11%)	
MSI_low		0 (0%)	1 (2.6%)	
MSS		10 (91%)	31 (82%)	
Death	49			>0.9
Alive		10 (91%)	32 (84%)	
Dead_CRC		0 (0%)	3 (7.9%)	
Dead_other		0 (0%)	1 (2.6%)	
Dead_unknown		1 (9.1%)	2 (5.3%)	

[1] Median (IQR); n (%)

[2] *p<0.05; **p<0.01; ***p<0.001

Part VI

MANUSCRIPT III

Analysis of circulating tumor DNA during checkpoint-inhibition in metastatic melanoma using a tumor-agnostic panel

Analysis of circulating tumor DNA during checkpoint-inhibition in metastatic melanoma using a tumor-agnostic panel

Judit Kisistók[1,2,3], Ditte Sigaard Christensen[1,2,4*], Mads Heilskov Rasmussen[1,2], Lone Duval[5], Ninna Aggerholm-Pedersen[4], Adam Andrzej Luczak[6], Boe Sandahl Sorensen[7], Martin Roelsgaard Jakobsen[8], Trine Heide Oellegaard[2,5$] and Nicolai Juul Birkbak[1,2,3$]*

[1]Department of Molecular Medicine, Aarhus University Hospital, Denmark

[2]Department of Clinical Medicine, Aarhus University, Denmark

[3]Bioinformatics Research Center, Aarhus University, Aarhus, Denmark

[4]Department of Oncology, Aarhus University Hospital, Denmark

[5]Department of Oncology, Goedstrup Hospital, Denmark

[6]Department of Oncology, Aalborg University Hospital, Denmark

[7]Department of Clinical-Biochemistry, Aarhus University Hospital, Denmark

[8]Department of Biomedicine, Aarhus University, Denmark

*These authors contributed equally

$These authors co-supervised

Corresponding authors:

Nicolai Juul Birkbak, nbirkbak@clin.au.dk

Trine Heide Oellegaard, trine.oellegaard@auh.rm.dk

Conflicts of interest and source of funding: none declared.

Abstract

Immunotherapy has revolutionized treatment of patients diagnosed with metastatic melanoma, where nearly half of patients receive clinical benefit. However, immunotherapy is also associated with immune-related adverse events, which may be severe and persistent. It is therefore important to identify patients that do not benefit from therapy early. Currently, regularly scheduled CT-scans are used to investigate size changes in target lesions to evaluate progression and therapy response. This study aims to explore if panel-based analysis of circulating tumor DNA (ctDNA) taken at three-week intervals may provide a window into the growing cancer, be used to identify non-responding patients early, and to determine genomic alterations associated with acquired resistance to checkpoint immunotherapy without analysis of tumor tissue biopsies. We designed a gene panel for ctDNA analysis, and sequenced 4-6 serial plasma samples from 24 patients with unresectable stage III or IV melanoma and treated with first-line checkpoint inhibitors enrolled at the Department of Oncology at Aarhus University Hospital, Denmark. TERT was the most mutated gene found in ctDNA and associated with a poor prognosis. We detected more ctDNA in patients with high metastatic load, which indicates that more aggressive tumors release more ctDNA into the bloodstream. While we did not find evidence of specific mutations associated with acquired resistance, we do demonstrate in this limited cohort of 24 patients that untargeted, panel-based ctDNA analysis has potential to be used as a minimally-invasive tool in clinical practice to identify patients where the benefits of immunotherapy outweigh the drawbacks.

The study (#1-10-72-230-19) received ethical approval on 21th of January 2021.

Keywords: *Circulating tumor DNA, melanoma, immunotherapy*

Introduction

Advanced melanoma is an aggressive cancer type with an overall poor survival rate and limited response to traditional cancer therapy regimes such as chemotherapy. Over the past decade, immunotherapy has entered the clinic and has now become a cornerstone of the treatment of melanoma with remarkably improved overall survival rates[1]. Nevertheless, while almost half of the patients will benefit from the treatment and a considerable subset may even become long-term survivors, the remaining patients will experience little to no benefit from the therapy. Considering that immunotherapy is associated with potentially persisting adverse events for the patients, there remains an unmet need for understanding why some patients will benefit from immunotherapy while other patients will not. Currently, this remains unclear. Several studies have explored biomarkers that may predict response to immunotherapy. Potential biomarkers include high PDL1 expression for anti-PD1/PDL1 therapies, and high tumor mutational burden (TMB), both now approved by the United States Food and Drug Administration (FDA) as an indication for using immunotherapy unrelated to diagnosis[2]. Other biomarkers have been reported as associated with immunotherapy response, including clonal TMB[3], an inflammation gene expression signature[4], and a signature based on soluble PD-1[5]. Additionally, somatic mutations of specific genes in the cancer cells have been found to influence the tumor microenvironment and the ability of tumor cells to evade the immune system, and hereby confer immunotherapy resistance[6]. These include inactivating mutations in *PTEN*, the third most frequently mutated gene in melanoma, which has been reported to be associated with resistance to checkpoint inhibition (CPI) in patients suffering from this disease[7].

Circulating tumor DNA (ctDNA) is defined as the fraction of cell-free DNA found in the blood, and is derived from the tumor. In recent years ctDNA has been extensively studied as a non-invasive biomarker in multiple cancer types[8]. Analysis of ctDNA can identify tumor-specific mutations, which reflect the genetic composition of the entire tumor, and ctDNA levels have been shown to correlate with both tumor burden and clinical outcomes during treatment[9]. Targeted deep sequencing of ctDNA at different time points during immunotherapy for genes involved in cancer cells' immune evasion may help resolve clonal diversity and identify the resistant clone. By timing molecular alterations with onset of resistance, it may be possible to decipher one possible contributing factor of resistance to immunotherapy. Several studies have found that patients with detectable ctDNA prior to treatment had worse progression-free survival (PFS) and worse overall survival (OS) than patients with undetectable ctDNA. Moreover, changes in ctDNA levels have been found to correlate with radiologic response, and a decrease in ctDNA level during therapy was shown to be associated with response and longer PFS and OS[10–18]. Taking ctDNA assays into the clinic has also faced challenges due to sensitivity limitations, particularly in patients with metastases at sites protected by the organ blood barrier[19]. Nevertheless, the overall utility of ctDNA as a non-invasive biomarker with insights into tumor biology makes it a valuable tool for real-time monitoring of response during treatment.

In this prospective study we aim to improve the understanding of the tumor biology that drives immunotherapy resistance in metastatic melanoma. We designed a clinical trial, where we hypothesized that patients diagnosed with melanoma and treated with immunotherapy would fall into three groups: (1) those that respond initially and continue to respond (responders), 2. those that fail to ever respond (innate resistance), and 3. those that initially respond, but over time develop resistance (acquired resistance). By using a custom tumor-agnostic ctDNA panel to identify genomic alterations before, during and after immunotherapy, we show how ctDNA levels remain low or undetectable in patients with therapy response, while it is present or increases in patients resistant to therapy. Additionally, by including known melanoma driver genes in the ctDNA panels, we demonstrate how it is possible to acquire novel insights into the biology behind response and resistance to immunotherapy.

MATERIALS AND METHODS

Patients

Clinical characteristics are summarized in Supplementary Table 1 and a full clinical table is found in Supplementary Table 2. Patients with unresectable, previously untreated stage III or IV melanoma who received systemic treatment with immune checkpoint inhibitors were eligible for the study. Key inclusion criterias were absence of uveal melanoma, absence of another primary cancer, and no previous diagnosis with cancer. In total we enrolled 24 patients with metastatic melanoma treated with first-line checkpoint inhibitors at Aarhus University Hospital in 2017. Ten patients received pembrolizumab at a dose 2 mg/kg every 3 weeks and the remaining 14 patients received nivolumab 1 mg/kg plus ipilimumab 3 mg/kg every 3 weeks, followed by maintenance nivolumab 1 mg/kg. Median follow-up time was 794 days, range (63 - 2008).

Disease characteristics and response assessment

Patient demographics and clinicopathologic features included: age, gender, performance status, metastatic sites at baseline, and lactate dehydrogenase (LDH). Elevated LDH level was defined as levels above 205 units/liter (U/L) for patients below the age of 70 and above 255 U/L for patients above the age of 70. Diagnostic tumor biopsies were routinely screened for BRAFV600 status and PD-L1 expression level (</> 1%). Treatment responses were evaluated by Positron emission tomography/computed tomography (PET/CT) scans of the chest, abdomen, and pelvis, and magnetic resonance imaging (MRI) in case of known brain metastases. We defined immunotherapy response groups as either Responders (11/23 patients with long-lasting response to the treatment), Resistance (9/23 patients with no response), and Acquired resistance (4/23 patients with acquired resistance). The definition of acquired resistance encompassed patients that initially showed response on first-line therapy based upon at least one CT evaluation scan, but later progressed or died within 12 months of follow-up. Median time to progression or death for responder patients was 639 days (range 487 - 1362, 5/11 patients progressed or died during follow-up), 121 days for resistant patients (range 30 - 273, 9/9 patients progressed and died during follow-up), and 200 days for acquired resistance patients (range 152 - 334, 4/4 patients progressed or died during follow-up). Median follow-up time was 1893 days for responders (range 778 - 2008), 293 days for resistant patients (range 63 - 1166), and 453 days for acquired resistance patients (range 248 - 809). Additionally, we defined metastatic load groups as high (equal to or more than 3 metastatic sites, 12/24) and low (less than 3 metastatic sites, 12/24). The number of metastatic sites was determined through CT-scans at follow-up times. Survival status was evaluated at the end of the follow-up period (14/24 deceased, 10/24 alive).

Sample collection and preparation

Peripheral blood samples (3 X 10mL Ethylenediamine tetraacetic acid (EDTA) tubes, BDVacutainer, Plymouth, United Kingdom) were obtained at baseline (immediately before treatment initiation) and every 3–4 weeks during treatment for up to one year after treatment initiation. Plasma was isolated from peripheral blood samples within 2–3 hours after blood collection by 1800 X g for 10 min at room temperature (RT). Plasma was cryopreserved at -80ºC.

Circulating-free DNA (cfDNA) extraction

cfDNA was extracted from 4 mL plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The isolated DNA was eluted in a 100 µL elution buffer and stored at -80ºC until analysis.

Circulating tumor DNA analysis and sequencing

A custom gene panel for next-generation sequencing (NGS) for ctDNA analysis was designed using Qiagen's design services covering a total of 40 genes. The 40 genes chosen for the panel are known to be associated with cancer cells susceptibility to immune attack (Supplementary Table 3). The panel covers a total of 150,000 base pairs of genomic content. Sequencing of the ctDNA panel was performed at 15,000x using an Illumina NovaSeq platform. In addition to the whole genes in the panel, the TERT promoter region was sequenced.

ctDNA variant calling and filtering

Variants were called using the Shearwater algorithm from the deepSNV R package [20,21]. Variants were called on a per-sample basis and the set of samples independent of the currently analyzed sample was used as background. Mutant allele frequency (MAF) and p-values per position and alteration were calculated and the most significant alteration as well as the reference allele per position were identified. Only driver alterations that were significant (p <= 0.05) at any time point per patient were retained, the remaining alterations were excluded from further analysis. The retained variants were annotated using Annovar based on the hg38 reference genome. Variants were excluded as potential single nucleotide polymorphisms (SNPs) if their mutant allele frequency exceeded 0.4, if ExAC or gnomAD values exceeded 0.01, or if they were marked as likely SNPs based on high (>0.1) and constant MAF over time. Driver mutations were defined as previously described [22], essentially based on detrimental mutations or frameshifts to known cancer genes as defined by the Catalogue of Somatic Mutations In Cancer (COSMIC)[23] cancer gene census[24]. In particular, the annotation was performed in the following manner:

1. A driver gene list was compiled from genes present in the COSMIC cancer gene census, as well as genes found in large pan-cancer studies.[25]

2. If a gene was listed in the COSMIC cancer gene census as tumor suppressor and 2 out of the 3 computational methods (Sift[26], Polyphen[27], and MutationTaster[28]) identified the variant as stop-gain or predicted deleterious, then it was annotated as a driver mutation
3. If a gene was listed in the COSMIC cancer gene census as an oncogene and an exact match of the variant was found in COSMIC at least 3 times, the variant was again annotated as driver variant.

To further reduce noise from SNPs and low-frequency subclonal mutations, only variants considered likely driver mutations were used in the analysis. A filtered variant table including all driver variants found significant in a given patient at least once is found in Supplementary Table 4.

Sensitivity analysis

We performed in-silico benchmarking by running the Shearwater algorithm on a synthetic dataset. The test data was generated in the following manner:

1. A patient's baseline sample was chosen randomly, and the count matrix was filtered for the BRAF V600E position. From the count matrix containing this position in 93 samples (the random patient's baseline sample and all other samples from the remaining patients), the ranges of counts for A and T nucleotides were obtained for the forward and backward strands.
2. A test matrix was generated where the 93 rows corresponded to samples and the 10 columns corresponded to nucleotides (A, T, G, C, and X corresponding to deletion, forward and backward strands).
3. Nucleotide A (columns 1 and 6) was chosen as reference, and nucleotide T (columns 2 and 7) was chosen as variant.

4. Using the count ranges obtained in Step 1, the counts are increased for the variant nucleotide T in each iteration, starting from 0 to the maximum observed variant count. The A nucleotide counts are set to the maximum observed reference count.
5. Each resulting count matrix is analyzed by Shearwater, yielding a p-value for the variant. The MAFs are calculated by dividing the variant count by the corresponding row sum.

After summarizing over each iteration, we report a median MAF limit of detection of 0.062 (IQR: 0.039 - 0.084, Supplementary Figure 1).

Statistical analysis

Per-patient MAF was calculated by taking the mean MAF of variants annotated as likely driver mutations, per time point. OS was defined as time from treatment initiation to the date of reported death due to any cause. Patients without disease progression or who were still alive at last follow-up were censored at the last follow-up date (15th of August 2021). All analysis was performed in R version 3.6.2, using Tidyverse [29] and ggpubr[30], scales[31], ggrepel[32] for visualizations. For significance testing, Wilcoxon test was used unless otherwise mentioned. P-values less than 0.05 were considered significant. All p-values are two-sided.

Ethics approval and consent to participate

The committees on Biomedical Research Ethics in the Central Region of Denmark approved the study (#1-10-72-230-19). The study was performed in accordance with the Declaration of Helsinki and all patients provided written informed consent.

Data availability statement

The data generated in this study are available within the supplementary data files. Due to privacy laws, access to raw sequencing data is restricted. Raw data can therefore only be made available following approval from the Danish National Committee on Health Research Ethics and the Danish Data Protection Agency. Access requests should be directed to the corresponding author.

RESULTS

Cohort overview

We endeavored to investigate the utility of tracking response to immunotherapy and development of treatment resistance using a custom tumor-agnostic ctDNA panel. For this purpose, we collected plasma samples from a cohort of 24 patients with metastatic melanoma, starting 1st line checkpoint immunotherapy with either pembrolizumab or with a combination of nivolumab and ipilimumab. From all patients, a baseline blood sample was analyzed, followed by additional blood samples drawn during treatment. From these, we purified and analyzed ctDNA using a custom panel and an in-house bioinformatics pipeline (Figure 1). Patients were grouped into Response (lasting response to immunotherapy, 11 patients), Acquired resistance (initial response, then acquired resistance to immunotherapy, 4 patients), and Resistance (no response to immunotherapy, 9 patients).

Figure 1

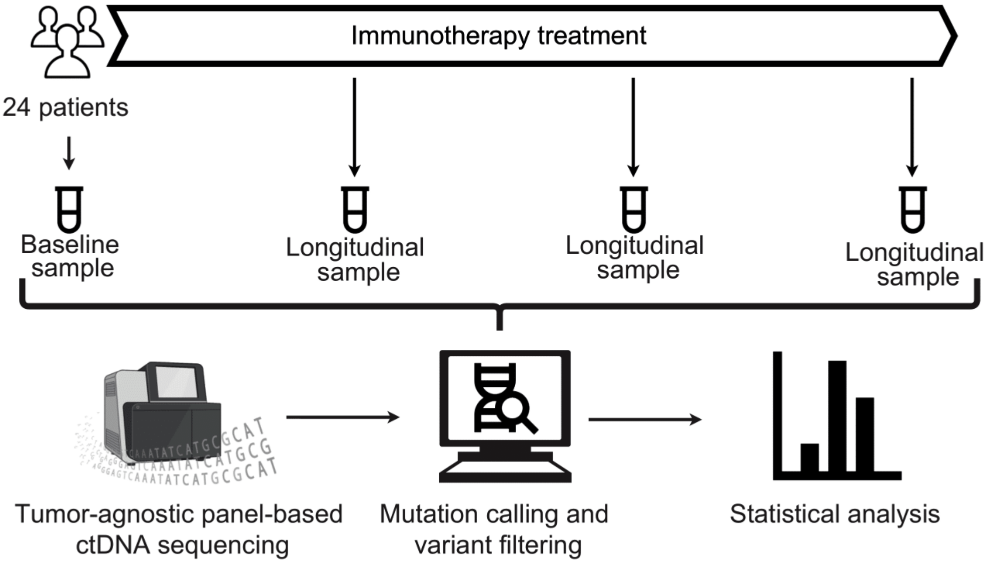


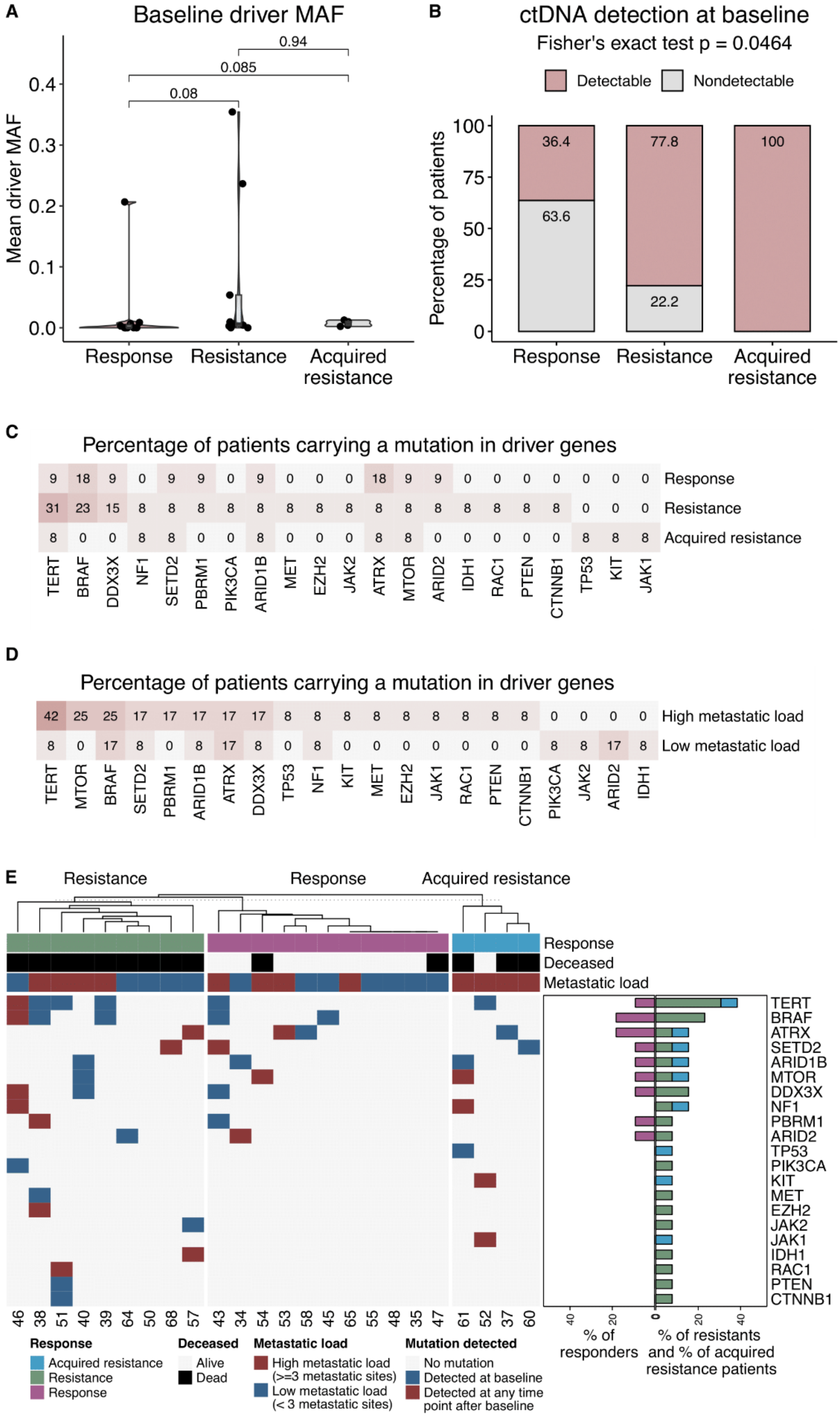
Figure 1. Study overview. 24 patients with metastatic melanoma were enrolled in the study at Aarhus University Hospital in 2017. Prior to receiving systemic treatment with first-line checkpoint inhibitors, a baseline blood sample was taken. Subsequently, 10 patients were selected to receive treatment with Pembrolizumab and the remaining 14 patients received a combination of Nivolumab and Ipilimumab. Over the course of treatment, three additional blood samples were extracted. After cell-free DNA extraction from plasma, circulating tumor DNA was sequenced at 15000x depth using a custom gene panel consisting of 40 genes known to be associated with immunotherapy response. Mutations were called in the resulting ctDNA data by running the Shearwater algorithm on a per-sample basis, using the samples from independent patients as background. Following statistical analyses were performed in R.

ctDNA frequency shows no difference between response groups

Relative to a tumor-informed ctDNA approach, our tumor-agnostic approach had the benefit of not requiring prior tumor DNA sequencing in order to call cancer mutations in plasma. However, this benefit comes with an increased risk of calling germline variants as potential tumor mutations. To minimize this, only variants considered likely cancer driver mutations were included in the downstream analysis (see methods). For all patients, we calculated the mean mutant allele frequency (MAF) for all driver-annotated variants found significant at baseline on a per-patient basis. For 9/24 patients where ctDNA detection was not possible at baseline, we set the mean MAF to 0. In our limited cohort of 24 patients, we observed no differences in baseline ctDNA MAF between patients with a response to treatment versus patients with no response to treatment (Figure 2A); however, when comparing ctDNA detection status at baseline, we observed a significant difference between the three response categories, showing that resistant and acquired resistant patients released ctDNA at baseline with higher likelihood. (p = 0.0464, Figure 2B). Additionally, when performing a survival analysis comparing patients with ctDNA detected at baseline to patients with no ctDNA detected, no significant difference was found, although a trend towards poor outcome for patients with ctDNA can be observed (P = 0.15, Supplementary Figure 2). Considering the limited size of the present cohort, it is likely that in a larger cohort, an association with outcome would be found.

We explored the association between ctDNA MAF and survival. We have found no significant difference in baseline MAF (p = 0.24) or ctDNA detection status at baseline (p = 0.403) when comparing patients who died during the study with those who remained alive (Supplementary Figure 3 A-B). We then investigated whether ctDNA MAF might be associated with metastatic burden. Here, we observed that patients with a high metastatic load harbored significantly higher ctDNA MAF levels at baseline (p = 0.041, Supplementary Figure 4A), supporting an association between cancer burden and ctDNA levels. However, we did not observe a significant difference in the percentage of patients showing ctDNA detection at baseline between high and low metastatic load (p = 0.4, Supplementary Figure 4B).

Figure 2



C Percentage of patients carrying a mutation in driver genes

	TERT	BRAF	DDX3X	NF1	SETD2	PBRM1	PIK3CA	ARID1B	MET	EZH2	JAK2	ATRX	MTOR	ARID2	IDH1	RAC1	PTEN	CTNNB1	TP53	KIT	JAK1
Response	9	18	9	0	9	9	0	9	0	0	0	18	9	9	0	0	0	0	0	0	0
Resistance	31	23	15	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	0	0	0
Acquired resistance	8	0	0	8	8	0	0	8	0	0	0	8	8	0	0	0	0	0	8	8	8

D Percentage of patients carrying a mutation in driver genes

	TERT	MTOR	BRAF	SETD2	PBRM1	ARID1B	ATRX	DDX3X	TP53	NF1	KIT	MET	EZH2	JAK1	RAC1	PTEN	CTNNB1	PIK3CA	JAK2	ARID2	IDH1
High metastatic load	42	25	25	17	17	17	17	17	8	8	8	8	8	8	8	8	8	0	0	0	0
Low metastatic load	8	0	17	8	0	8	17	8	0	8	0	0	0	0	0	0	0	8	8	17	8

Figure 2. Mutation analysis. A) Violinplots showing the mutant allele frequency at baseline. Values shown on the Y axis are calculated by taking the mean MAF over all filtered variants per patient. X axis and color show response category. B) Stacked bar plots showing the percentage of patients per response category where ctDNA was detected at baseline. X axis shows response category, Y axis displays patient percentage, color corresponds to the detection status at baseline. C) Heatmap showing the percentage of patients carrying mutations in panel genes. The percentages are calculated within categories. Color intensity corresponds to the percentage values. D) Heatmap showing the percentage of patients carrying mutations in panel genes, split according to metastatic load categories. Visualization follows panel C). E) Heatmap-barplots showing the driver mutation profile of the cohort. Blue tiles indicate that a mutation was detected at baseline, red tiles indicate mutations that were not found at baseline but were detected at a later time point. Top annotations show patient characteristics: response category (pink: response, green: resistance, blue: acquired resistance), survival status (gray: alive, black: dead) and metastatic load (blue: low, <3 metastatic sites, red: high, >=3 metastatic sites). Barplots on the side show the percentage of patients carrying a driver mutation per response category (pink: response, green: resistance, blue: acquired resistance).

Patients resistant to treatment harbor TERT alterations in ctDNA

Next, we determined the percentage of patients with detectable mutations annotated as drivers in ctDNA in any of the 40 panel genes (methods). Of 40 genes investigated, we found at least one driver mutation in 20 genes in at least one patient. We observed that the telomerase gene *TERT* was the most commonly altered gene found mutated in ctDNA in patients that developed therapy resistance (affecting 5/13 resistant and acquired resistant patients). In comparison, only 1/11 patients with response to therapy harbored a mutation in TERT (Figure 2C and Supplementary Figure 4C). This is consistent with findings from previous studies associating TERT with poor outcome[33], and implicates *TERT* with immunotherapy resistance. This gene was found as most commonly altered in the high metastatic load subgroup as well, compared to their low metastatic load counterparts (Figure 2D and Supplementary Figure 4D). We identified a similar trend when comparing deceased and alive patients, with TERT being the most frequently mutated gene in the deceased subgroup (affecting 4/14 deceased and 2/10 alive patients, Supplementary Figure 3C). BRAF was the second most mutated gene overall, affecting 3/9 resistant patients and 2/11 responders. No significant difference in the frequency of BRAF mutations was observed between responders and resistant patients. When we compared patients with resistance to patients with acquired resistance, we observed no pattern in the overall mutations found in ctDNA arising after initiation of immunotherapy (Figure 2E, red squares indicate mutations only found in later liquid biopsies). Thus in this limited cohort, no mutation to specific genes could be associated with acquired resistance to immunotherapy.

Patients with response to therapy tend to harbor fewer detectable driver mutations in ctDNA

When we compared ctDNA between the patient response groups, ctDNA was detected at baseline in 4/11 (36%) of the responder patients. This compares to 7/9 (78%) and 4/4 (100%) of the resistant patients and patients showing acquired resistance, potentially reflecting a lower initial disease burden among patients responding well to therapy (Figure 3). As expected, all responding patients showed a decline in ctDNA during their treatment, with no patients having detectable ctDNA in their last blood sample. In comparison, ctDNA was found in the last blood sample of 7/9 resistant patients. Patients 50 and 64 were both resistant to therapy, but showed no ctDNA in their last blood sample. No ctDNA was found at any time point for patient 50, which may indicate that the cancer harbored no driver mutations within the panel, making it essentially undetectable by our tumor-agnostic panel. Conversely, a single driver mutation, the chr12:45852805:T alteration affecting ARID2, was found in patient 64 at baseline. This mutation was not found during follow-up, and may represent a minor subclone eliminated by the treatment. Among the patients with acquired resistance, 4/4 showed at least one ctDNA positive blood sample during follow-up. Patient 37 showed detectable ctDNA only at baseline, despite recurrence detected by CT-scan prior to the last blood sample being obtained. This may reflect an overall low disease burden or subclonal selection resulting in outgrowth of a subclone harboring no driver mutations in the ctDNA gene panel. Considering resistant and acquired resistance patients together, we observe that overall 12/13 patients had at least one ctDNA positive sample, indicating that the current tumor-agnostic gene panel, selected to enrich in known cancer genes commonly mutated in melanoma, can detect cancer in 92.3% of patients. In this analysis, patients showing a response to therapy are not included, as we here cannot discern between cancers that are negative due to no driver mutations found within the ctDNA gene panel and cancers that are negative due to therapeutic response to therapy.

Figure 3

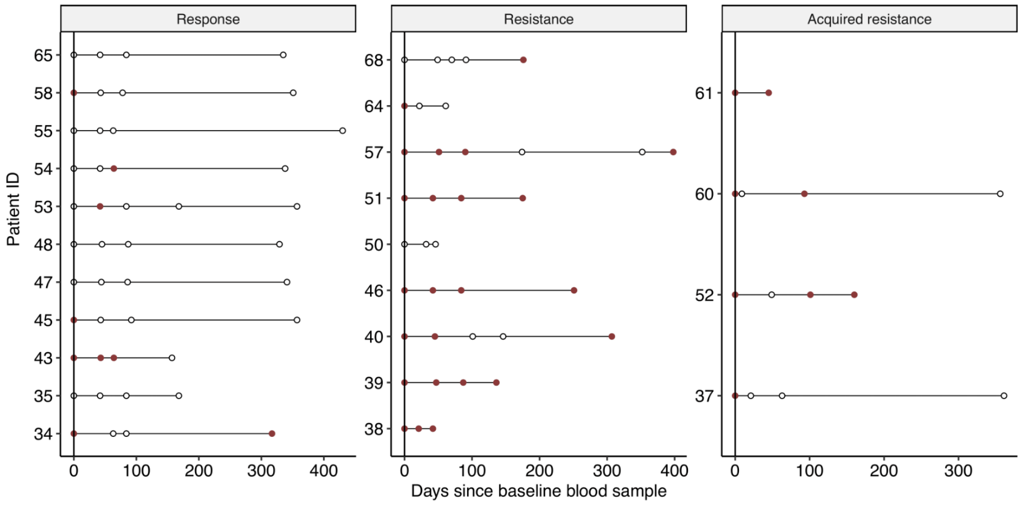


Figure 3. Driver mutation detection per response category. Dotplots showing driver mutation detection status over the course of the study. Y-axis shows patient ID, X-axis shows the days passed since the baseline blood sample was extracted. Dots correspond to blood tests taken, empty dots indicate no driver mutation detected, red dots indicate that driver mutations were detected in the sample.

It has previously been suggested that ctDNA detection likely mostly depends on the total number of cancer cells[34], which is consistent with the finding that patients with a higher metastatic load showed a higher MAF on average (Supplementary Figure 4A). However, in this cohort of limited size, metastatic load showed no association with outcome (Supplementary Figure 5).

Evaluating ctDNA dynamics during therapy

Lastly, we investigated the utility of the ctDNA panel as a biomarker for treatment response in a longitudinal setting. For this purpose, we analyzed the ctDNA MAF in serial samples in all patients, and compared it to treatment response (Figure 4-6). In patients with response to treatment, ctDNA was not detected at any time point in 5/11 patients. In two patients, we detected ctDNA at baseline only. In one patient, patient 43, ctDNA was detected at baseline, and decreased in MAF while on treatment, falling below detection limit as treatment was discontinued (Figure 4).

In patients showing treatment resistance, ctDNA was detected in 7/9 patients at baseline. One additional patient became ctDNA positive as their disease progressed. We observed an increase in ctDNA MAF between baseline and clinical relapse for 4/9 patients (Figure 5).

For patients who acquired resistance to treatment, ctDNA levels were low, but were detected at baseline for all 4 patients (Figure 6). Overall, these observations indicate that in this cohort, ctDNA dynamics alone cannot be used as a reliable biomarker of therapeutic response.

Figure 4

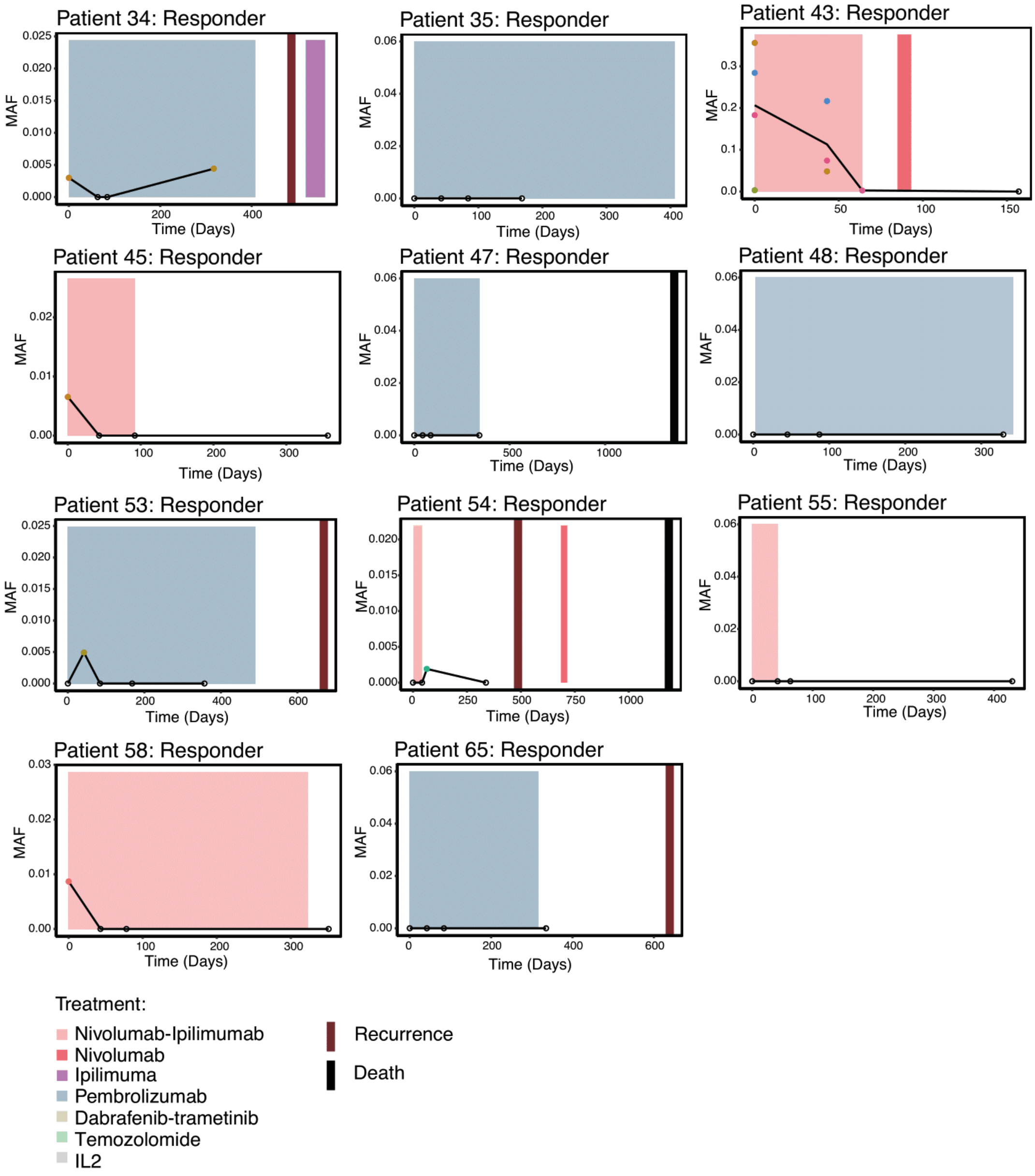


Figure 4. Longitudinal analysis of ctDNA in patients with response. Per-patient plots showing the driver mutations detected over the course of treatment and clinical history. Y axis shows the mutant allele frequency, X axis shows the days since the baseline blood sample. Colored dots indicate detected mutations, empty dots signify that a blood sample was taken at the time point but no mutation was detected. Black line connecting the time points corresponds to the mean MAF over

time. Colored boxes show the type and time frame of treatment. Red vertical lines show the date of clinical progression, black vertical lines show the date of death.

Figure 5

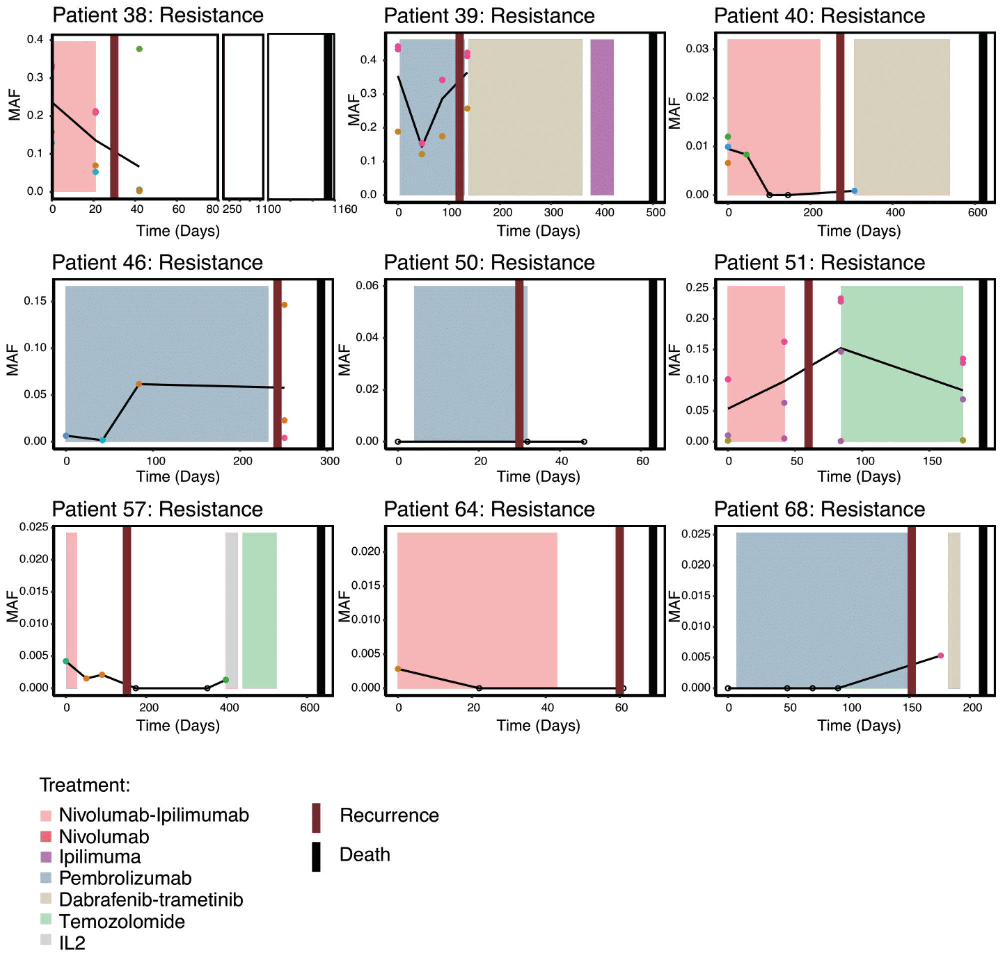


Figure 5. Longitudinal analysis of ctDNA in patients with resistance. Per-patient plots showing the driver mutations detected over the course of treatment and clinical history in patients resistant to treatment. Annotation follows figure 4.

Figure 6

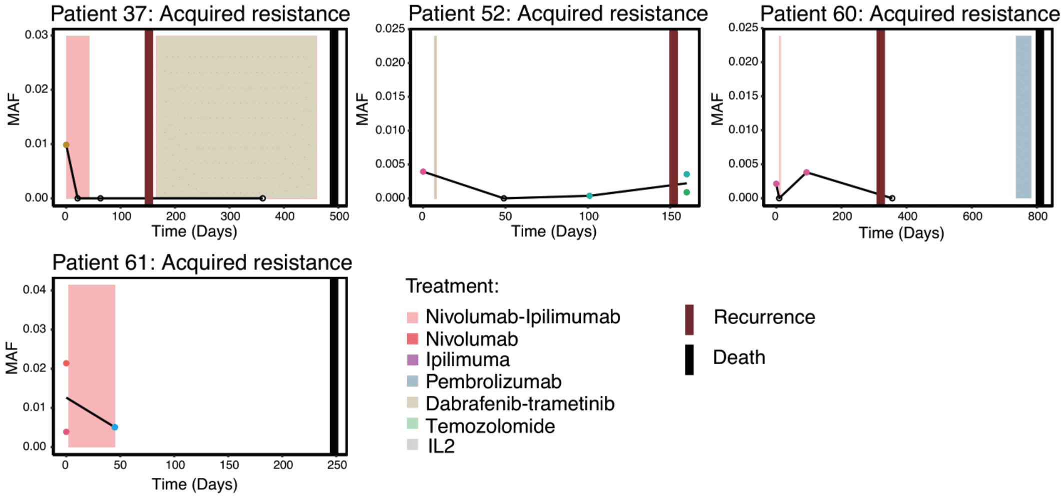


Figure 6. Longitudinal analysis of ctDNA in patients who acquired resistance. Per-patient plots showing the driver mutations detected over the course of treatment and clinical history in patients developing acquired resistance to treatment. Annotation follows figure 4.

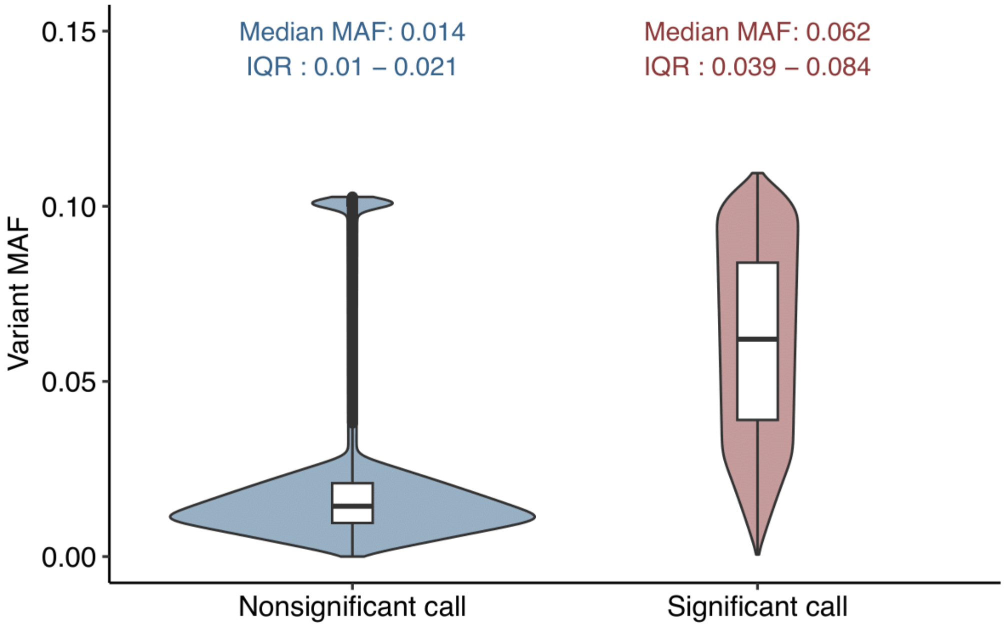
Discussion

In this study we report differences in genomic alterations between patients that have a complete response to immunotherapy compared to patients that have either no response or have developed acquired resistance. By using a unique panel of well-established genes known to be involved with the development of melanoma and checkpoint inhibition response, we have demonstrated how genomic data can be used to analyze and identify certain differences in responses to immunotherapy in patients with metastatic melanoma. Consistent with current literature, we have found that a higher percentage of resistant and acquired resistance patients harbor a mutation in TERT compared to their responder counterparts. While we cannot exclude that TERT mutations may also be found in subclones in tumors not shedding ctDNA, our data indicate that genomic alterations in the TERT gene found in ctDNA can be used as a predictor for poor prognosis and poor response to immunotherapy. Currently, there is a strong focus on investigating ctDNA and exploring and validating the use of ctDNA in clinical practice. One of the strengths of our study is the continuous blood samples obtained during treatment, which has enabled analysis of the dynamics of ctDNA over time. Potentially, continuous blood samples can be used to detect recurrence even before the cancer is detectable on follow-up scans. We observed no differences in overall ctDNA levels between responders, resistant, and acquired resistant patients in this cohort, either at baseline or at any time point during treatment and follow-up. This indicates that ctDNA levels alone may not be sufficient to identify metastatic melanoma patients likely to respond to immunotherapy. However, other studies have demonstrated a correlation between low levels of ctDNA and disease burden, also in metastatic melanoma patients[35]. Thus, our results may indicate a sensitivity issue with tumor-agnostic approaches such as the one applied here. Particularly, the panel gene set was limited to 40 genes, representing a relatively small panel size, which limited sensitivity. Since our study commenced, further experience and technical improvements in ctDNA purification methods have demonstrated improved yields. Particularly, it is today standard to double-spin samples prior to plasma collection as this reduces contaminating nuclear DNA[36]. However, our study was initiated before this was established as a superior methodology, and to ensure uniformity in sample collection, all samples were only subjected to a single round of centrifugation. This may have reduced the total ctDNA yield per sample, and thus negatively affected our ability to detect somatic mutations, particularly in samples with low ctDNA burden. Despite these limitations, we did observe using our tumor agnostic panel that resistant and acquired resistant patients tended to be ctDNA positives at baseline more often than responders..

In our work, we found significant differences in ctDNA MAF when we compared patients with high and low metastatic load. This indicates that, in line with current literature, patients harboring higher metastatic load will shed more ctDNA into circulation due to a higher cancer cell burden[34]. Additionally, we observed an enhanced signal of TERT mutations in the high load group, which is consistent with already published work[33] associating TERT with poor prognosis. Potentially, TERT can act as a biomarker for identifying patients likely resistant to immunotherapy, however, this needs to be further validated in a larger cohort.

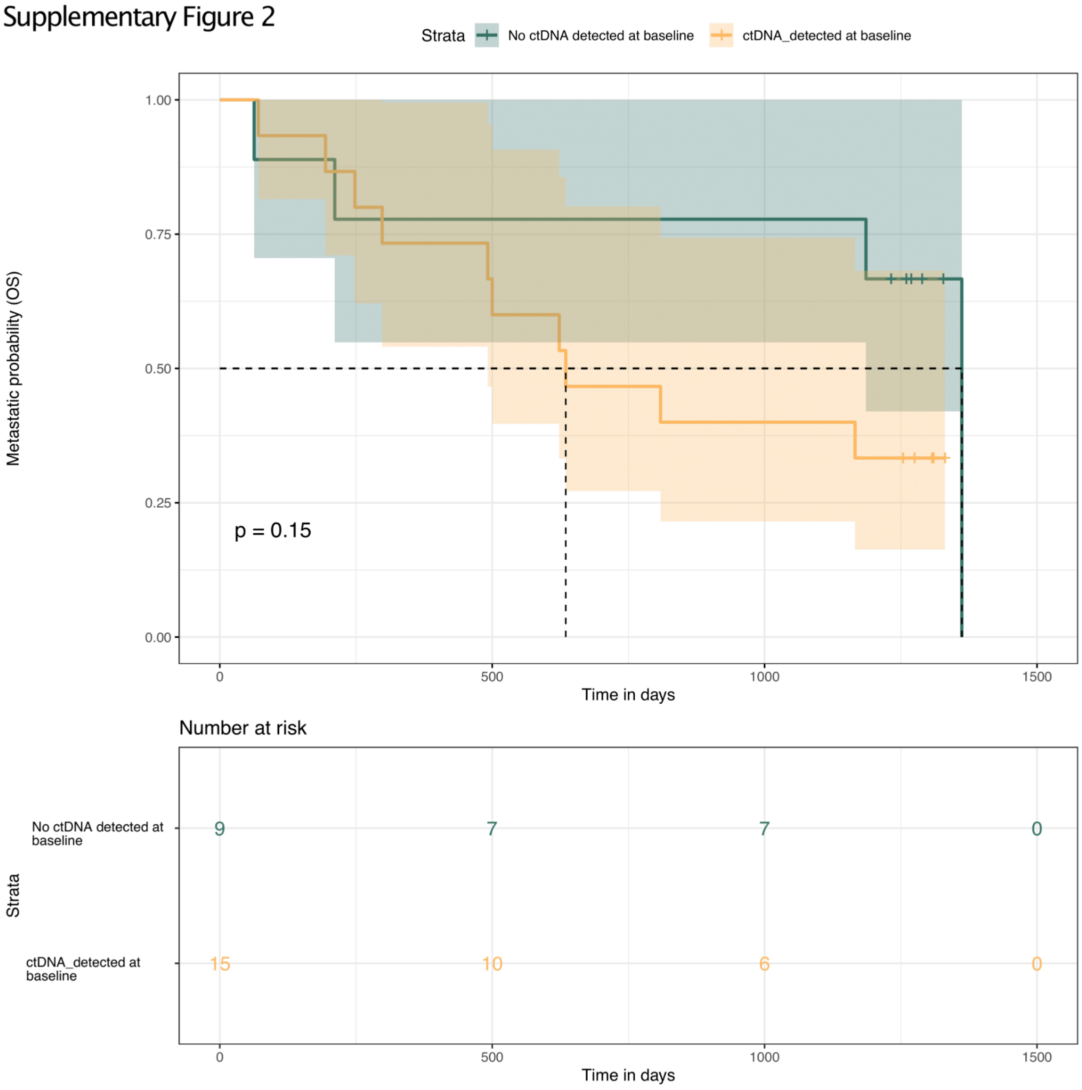
A major limitation to our study is the small cohort size as well as a lack of tumor biopsies or germline control samples which is a challenge for ctDNA mutation calling and makes it difficult to evaluate the performance of our variant filtering and noise reduction. While we use independent samples as background for variant calling and known SNP databases to filter out normal alterations, we expect that some melanoma-specific variants remain undetected or are excluded during filtering. Nevertheless, after meticulous analysis of the data, we here show how a tumor agnostic panel ctDNA can be used to inform about tumor biology and cancer progression, and we believe our study may serve as a starting point for deeper investigations into the utility of ctDNA in metastatic melanoma and the biology of treatment response.

Supplementary Figure 1



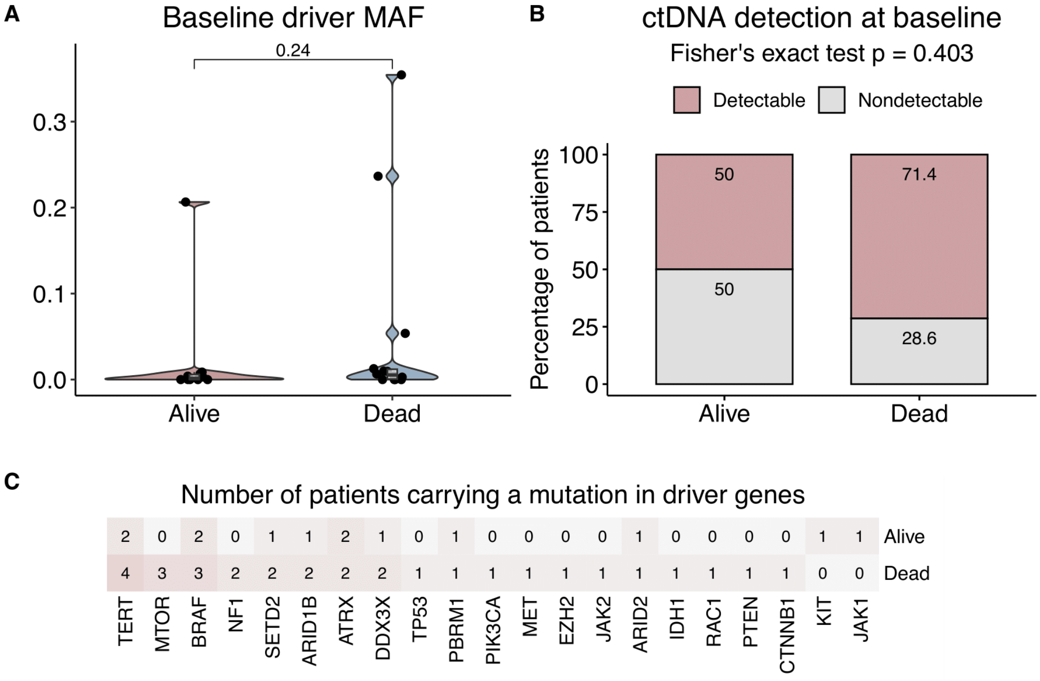
Supplementary Figure 1. Violin-boxplot showing the results of the in silico sensitivity benchmarking results of our in-house bioinformatics pipeline.

Supplementary Figure 2



Supplementary Figure 2. Kaplan-Meier curve showing overall survival comparing patients with ctDNA detected at baseline to patients without detected ctDNA at baseline.

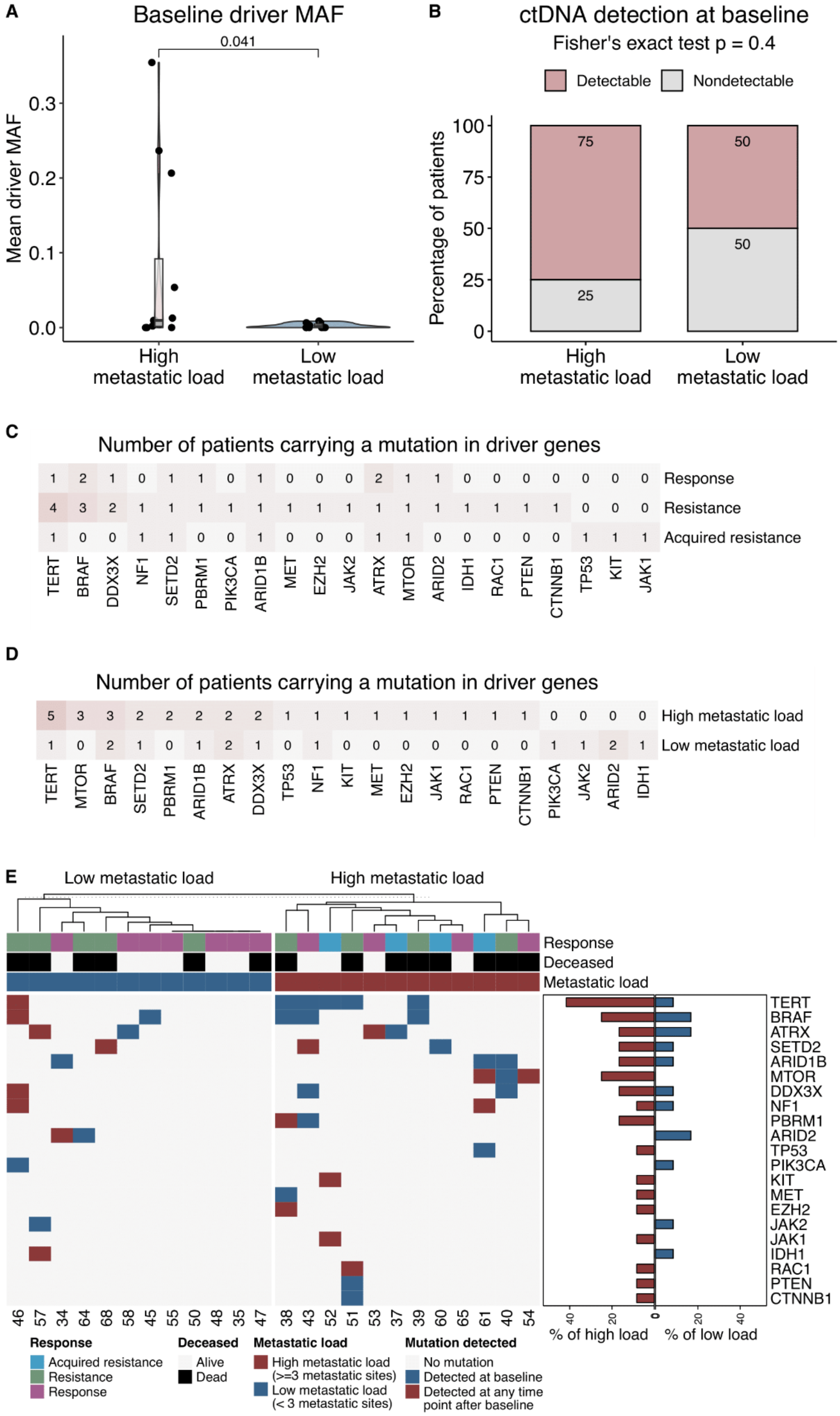
Supplementary Figure 3



	TERT	MTOR	BRAF	NF1	SETD2	ARID1B	ATRX	DDX3X	TP53	PBRM1	PIK3CA	MET	EZH2	JAK2	ARID2	IDH1	RAC1	PTEN	CTNNB1	KIT	JAK1
Alive	2	0	2	0	1	1	2	1	0	1	0	0	0	0	1	0	0	0	0	1	1
Dead	4	3	3	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	0	0

Supplementary Figure 3. Mutation analysis comparing patients who deceased over the course of treatment with patients who were alive at the last follow-up. A) Violinplots showing the mutant allele frequency at baseline. Values shown on the Y axis are calculated by taking the mean MAF over all filtered variants per patient. X axis and color show survival status. B) Stacked bar plots showing the percentage of patients per survival category where ctDNA was detected at baseline. X axis shows survival status, Y axis displays patient percentage, color corresponds to the detection status at baseline. C) Heatmap showing the number of patients carrying mutations in panel genes. Color intensity corresponds to the values.

Supplementary Figure 4



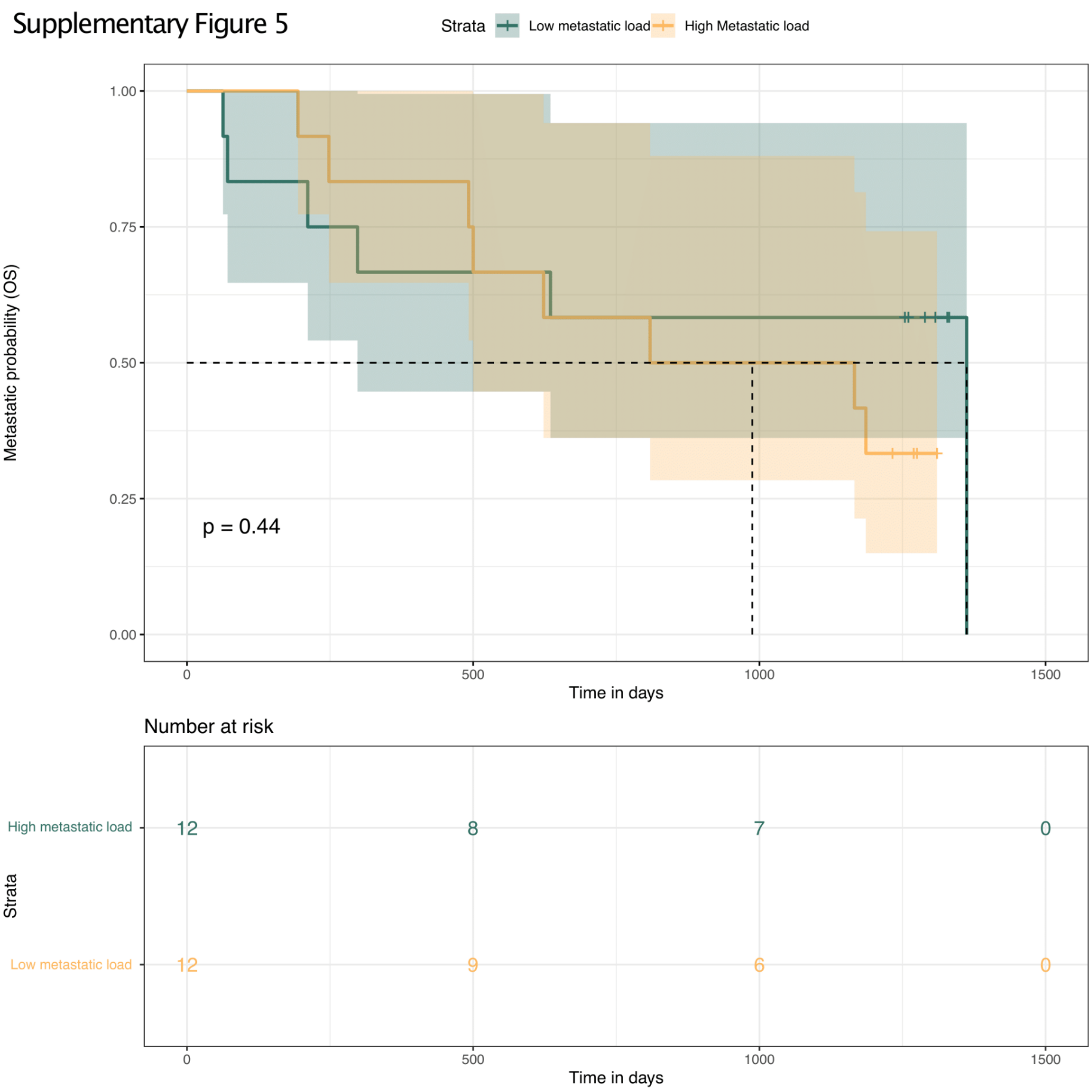
Number of patients carrying a mutation in driver genes

	TERT	BRAF	DDX3X	NF1	SETD2	PBRM1	PIK3CA	ARID1B	MET	EZH2	JAK2	ATRX	MTOR	ARID2	IDH1	RAC1	PTEN	CTNNB1	TP53	KIT	JAK1
Response	1	2	1	0	1	1	0	1	0	0	0	2	1	1	0	0	0	0	0	0	0
Resistance	4	3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
Acquired resistance	1	0	0	1	1	0	0	1	0	0	0	1	1	0	0	0	0	0	1	1	1

Number of patients carrying a mutation in driver genes

	TERT	MTOR	BRAF	SETD2	PBRM1	ARID1B	ATRX	DDX3X	TP53	NF1	KIT	MET	EZH2	JAK1	RAC1	PTEN	CTNNB1	PIK3CA	JAK2	ARID2	IDH1
High metastatic load	5	3	3	2	2	2	2	2	1	1	1	1	1	1	1	1	1	0	0	0	0
Low metastatic load	1	0	2	1	0	1	2	1	0	1	0	0	0	0	0	0	0	1	1	2	1

Supplementary Figure 4. Mutation analysis. A) Violinplots showing the mutant allele frequency at baseline. Values shown on the Y axis are calculated by taking the mean MAF over all filtered variants per patient. X axis and color show metastatic load category. B) Stacked bar plots showing the percentage of patients per metastatic load category where ctDNA was detected at baseline. X axis shows metastatic load category, Y axis displays patient percentage, color corresponds to the detection status at baseline. C) Heatmap showing the number of patients carrying mutations in panel genes, split according to response categories. Color intensity corresponds to the values. D) Heatmap showing the number of patients carrying mutations in panel genes, split according to metastatic load categories. Visualization follows panel C). E) Heatmap-barplots showing the driver mutation profile of the cohort. Blue tiles indicate that a mutation was detected at baseline, red tiles indicate mutations that were not found at baseline but were detected at a later time point. Top annotations show patient characteristics: response category (pink: response, green: resistance, blue: acquired resistance), survival status (gray: alive, black: dead), and metastatic load (blue: low, <3 metastatic sites, red: high, >=3 metastatic sites). Barplots on the side show the percentage of patients carrying a driver mutation per metastatic load category (red: high, blue: low).



Supplementary Figure 5. Kaplan-Meier curve showing overall survival comparing patients with high metastatic load to patients with low metastatic load.

References

1. Weiss, S. A., Wolchok, J. D. & Sznol, M. Immunotherapy of Melanoma: Facts and Hopes. *Clin. Cancer Res.* **25**, 5191–5201 (2019).

2. Wang, Y. *et al.* FDA-Approved and Emerging Next Generation Predictive

Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Front. Oncol.* **11**, 683419 (2021).

3. Litchfield, K. *et al.* Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614.e14 (2021).
4. Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, (2018).
5. Pedersen, J. G. *et al.* Increased Soluble PD-1 Predicts Response to Nivolumab plus Ipilimumab in Melanoma. *Cancers* **14**, (2022).
6. Keenan, T. E., Burke, K. P. & Van Allen, E. M. Genomic correlates of response to immune checkpoint blockade. *Nat. Med.* **25**, 389–402 (2019).
7. Peng, W. *et al.* Loss of PTEN Promotes Resistance to T Cell-Mediated Immunotherapy. *Cancer Discov.* **6**, 202–216 (2016).
8. Keller, L., Belloum, Y., Wikman, H. & Pantel, K. Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond. *Br. J. Cancer* **124**, 345–358 (2021).
9. Abbosh, C. *et al.* Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. *Nature* (2023) doi:10.1038/s41586-023-05776-4.
10. Lipson, E. J. *et al.* Circulating tumor DNA analysis as a real-time method for monitoring tumor burden in melanoma patients undergoing

treatment with immune checkpoint blockade. *J Immunother Cancer* **2**, 42 (2014).

11. Lee, D. J. & Faries, M. B. *Practical Manual for Dermatologic and Surgical Melanoma Management*. (Springer Nature, 2020).

12. Gray, E. S. *et al.* Circulating tumor DNA to monitor treatment response and detect acquired resistance in patients with metastatic melanoma. *Oncotarget* **6**, 42008–42018 (2015).

13. Cabel, L. *et al.* Circulating tumor DNA changes for early monitoring of anti-PD1 immunotherapy: a proof-of-concept study. *Ann. Oncol.* **28**, 1996–2001 (2017).

14. Christensen, E. *et al.* Early Detection of Metastatic Relapse and Monitoring of Therapeutic Efficacy by Ultra-Deep Sequencing of Plasma Cell-Free DNA in Patients With Urothelial Bladder Carcinoma. *Journal of Clinical Oncology* vol. 37 1547–1557 Preprint at https://doi.org/10.1200/jco.18.02052 (2019).

15. Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).

16. Pedersen, J. G. *et al.* Inflammatory Cytokines and ctDNA Are Biomarkers for Progression in Advanced-Stage Melanoma Patients Receiving Checkpoint Inhibitors. *Cancers* **12**, (2020).

17. Reinert, T. *et al.* Analysis of Plasma Cell-Free DNA by Ultradeep

Sequencing in Patients With Stages I to III Colorectal Cancer. *JAMA Oncol* **5**, 1124–1131 (2019).

18. Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* **26**, 1114–1124 (2020).
19. Khaddour, K. *et al.* Case report: Real-world experience using a personalized cancer-specific circulating tumor DNA assay in different metastatic melanoma scenarios. *Front. Oncol.* **12**, (2022).
20. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
21. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
22. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
23. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
24. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
25. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes

across 21 tumour types. *Nature* **505**, 495–501 (2014).

26. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
27. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
28. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
29. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
30. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. Preprint at https://rpkgs.datanovia.com/ggpubr/ (2020).
31. Hadley, W. & Seidel, D. scales: Scale functions for visualization. Preprint at https://CRAN.R-project.org/package=scales (2019).
32. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. Preprint at https://CRAN.R-project.org/package=ggrepel (2021).
33. Gandini, S. *et al.* TERT promoter mutations and melanoma survival: A comprehensive literature review and meta-analysis. *Crit. Rev. Oncol. Hematol.* **160**, 103288 (2021).
34. Avanzini, S. *et al.* A mathematical model of ctDNA shedding predicts

tumor detection size. *Sci Adv* **6**, (2020).

35. McEvoy, A. C. *et al.* Correlation between circulating tumour DNA and metabolic tumour burden in metastatic melanoma patients. *BMC Cancer* **18**, 726 (2018).

36. Trigg, R. M., Martinson, L. J., Parpart-Li, S. & Shaw, J. A. Factors that influence quality and yield of circulating-free DNA: A systematic review of the methodology literature. *Heliyon* **4**, e00699 (2018).

Part VII

DECLARATIONS

AARHUS
UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Judit Kisistók

This declaration concerns the following article/manuscript:

Title:	Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA
Authors:	Christopher Abbosh, Alexander M. Frankell, Thomas Harrison, Judit Kisistok, Aaron Garnett, Laura Johnson, Selvaraju Veeriah, Mike Moreau, Adrian Chesh, Tafadzwa L. Chaunzwa, Jakob Weiss, Morgan R. Schroeder, Sophia Ward, Kristiana Grigoriadis, Aamir Shahpurwalla, Kevin Litchfield, Clare Puttick, Dhruva Biswas, Takahiro Karasaki, James R. M. Black, Carlos Martínez-Ruiz, Maise Al Bakir, Oriol Pich, Thomas B. K. Watkins, Emilia L. Lim, Ariana Huebner, David A. Moore, Nadia Godin-Heymann, Anne L'Hernault, Hannah Bye, Aaron Odell, Paula Roberts, Fabio Gomes, Akshay J. Patel, Elizabeth Manzano, Crispin T. Hiley, Nicolas Carey, Joan Riley, Daniel E. Cook, Darren Hodgson, Daniel Stetson, J. Carl Barrett, Roderik M. Kortlever, Gerard I. Evan, Allan Hackshaw, Robert D. Daber, Jacqui A. Shaw, Hugo J. W. L. Aerts, Abel Licon, Josh Stahl, Mariam Jamal-Hanjani, TRACERx Consortium, Nicolai J. Birkbak, Nicholas McGranahan, Charles Swanton

The article/manuscript is: Published ☒ Accepted □ Submitted □ In preperation □

If published, state full reference: Abbosh, C., Frankell, A.M., Harrison, T. et al. Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. Nature 616, 553–562 (2023). https://doi.org/10.1038/s41586-023-05776-4

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes □ If yes, give details:

Your contribution

Please rate (A-F) your contribution to the elements of this article/manuscript, **and** elaborate on your rating in the free text section below.

A. Has essentially done all the work (>90%)
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution (<10%)
F. N/A

Category of contribution	Extent (A-F)
The conception or design of the work:	F
Free text description of PhD student's contribution ***(mandatory)*** The study was conceived and already in progress prior to me joining the research efforts.	
The acquisition, analysis, or interpretation of data:	C

Free text description of PhD student's contribution ***(mandatory)*** I built a data pipeline for analyzing a variety of data types for the section "Biology of ctDNA detection". Additionally, I was responsible for creating visualizations for this section, in particular, for Figure 2.	
Drafting the manuscript:	D
Free text description of PhD student's contribution ***(mandatory)*** I assisted in drafting and editing the section pertaining to my work, including figure legends and methods.	
Submission process including revisions:	D
Free text description of PhD student's contribution ***(mandatory)*** I assisted in addressing questions pertaining to my work during the revision process, including making changes in the analyses and editing figures for publication. Additionally, I helped review sections of the paper prior to resubmission.	

Signatures of first- and last author, and main supervisor

Date	Name	Signature
10/5-23	Charles Swanton	
10/5-23	Christopher Abbosh	
10/5-23	Nicolai Juul Birkbak	Nicolai Birkbak

Date: 10/5-23

Judit Kisistók

Signature of the PhD student

AARHUS UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Judit Kisistók

This declaration concerns the following article/manuscript:

Title:	Exploring the biology of ctDNA release in colon cancer
Authors:	Judit Kisistók, Laura Andersen, Tenna Vesterman Henriksen, Jesper Bertram Bramsen, Thomas Reinert, Trine Block Mattesen, Nicolai Juul Birkbak, Claus Lindbjerg Andersen

The article/manuscript is: Published □ Accepted □ Submitted □ In preperation ☒

If published, state full reference:

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes □ If yes, give details:

Your contribution

Please rate (A-F) your contribution to the elements of this article/manuscript, **and** elaborate on your rating in the free text section below.

A. Has essentially done all the work (>90%)
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution (<10%)
F. N/A

Category of contribution	Extent (A-F)
The conception or design of the work:	D
*Free text description of PhD student's contribution **(mandatory)*** I contributed to defining and study concept and the analysis methodology.	
The acquisition, analysis, or interpretation of data:	B
*Free text description of PhD student's contribution **(mandatory)*** I built a data pipeline for analyzing and visualizing transcriptomic, genomic, and clinical data.	
Drafting the manuscript:	B
*Free text description of PhD student's contribution **(mandatory)*** I assisted in writing the Introduction, Results, Discussion and Methods sections.	
Submission process including revisions:	F

AARHUS
UNIVERSITY

Free text description of PhD student's contribution ***(mandatory)***
We have not submitted the manuscript yet.

Signatures of first- and last author, and main supervisor

Date	Name	Signature
24-05-2023	Claus Lindbjerg Andersen	
23-05-2023	Nicolai Juul Birkbak	
25-05-2023	Judit Kisistok	

Date: 25-05-2023

Signature of the PhD student

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Judit Kisistók

This declaration concerns the following article/manuscript:

Title:	Analysis of circulating tumor DNA during checkpoint-inhibition in metastatic melanoma using a tumor-agnostic panel
Authors:	Judit Kisistók, Ditte Sigaard Christensen, Mads Heilskov Rasmussen, Lone Duval, Ninna Aggerholm-Pedersen, Adam Andrzej Luczak, Boe Sandahl Sorensen, Martin Roelsgaard Jakobsen, Trine Heide Oellegaard, Nicolai Juul Birkbak

The article/manuscript is: Published □ Accepted ☒ Submitted □ In preperation □

If published, state full reference:

If accepted or submitted, state journal: Melanoma Research

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No □ Yes ☒ If yes, give details: It has been included in the PhD thesis of Ditte Sigaard Christensen.

Your contribution

Please rate (A-F) your contribution to the elements of this article/manuscript, **and** elaborate on your rating in the free text section below.

A. Has essentially done all the work (>90%)
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution (<10%)
F. N/A

Category of contribution	Extent (A-F)
The conception or design of the work:	D
Free text description of PhD student's contribution ***(mandatory)*** I contributed to defining the analysis pipeline.	
The acquisition, analysis, or interpretation of data:	B
Free text description of PhD student's contribution ***(mandatory)*** I built a variant calling pipeline for processing the mutation data and a data pipeline for analysing and visualizing the genomic and clinical data.	
Drafting the manuscript:	C
Free text description of PhD student's contribution ***(mandatory)*** I assisted in writing the Results sections.	
Submission process including revisions:	C

Free text description of PhD student's contribution ***(mandatory)***
I helped drafting the rebuttal letter and carried out supplementary analyses requested by the reviewer.

Signatures of first- and last author, and main supervisor

Date	Name	Signature
24/5-2023	Nicolai Juul Birkbak	Nicolai Birkbak
25/5-2023	Judit Kisistok	

Date: 25/5-2023

Signature of the PhD student