Next generation molecular diagnostics using ultrasensitive sequencing

Stefan Filges

Department of Laboratory Medicine
Institute of Biomedicine
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg, 2022

Cover illustration: Liquid biopsy
By Stefan Filges. Created with BioRender.com.

Next generation molecular diagnostics using ultrasensitive sequencing

stefan.filges@gu.se

ISBN 978-91-8009-658-4 (PRINT)
ISBN 978-91-8009-659-1 (PDF)

Printed in Borås, Sweden 2022
Printed by Stema Specialtryck AB, Borås

"Today is only one day in all the days that will ever be. But what will happen in all the other days that ever come can depend on what you do today."

—Ernest Hemingway

Abstract

Massively parallel sequencing enables the exploration of the genetic heterogeneity within microbial, viral and tumour cell populations. Detecting circulating tumor DNA in blood and other body fluids has the potential to revolutionize molecular diagnostics. However, these liquid biopsies typically contain only minute amounts of highly degraded DNA and standard sequencing approaches lack the resolution to detect rare genetic variants. The overall goal of this thesis was to develop an ultrasensitive sequencing approach with single molecule resolution that requires only minimal amounts of material. To this end, we developed the simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing protocol (SiMSen-Seq). SiMSen-Seq achieves ultrasensitive detection of nucleotide variants by attaching unique molecular identifiers to target DNA molecules using PCR primers. SiMSen-Seq is enabled by highly optimized reaction conditions and the use of a stem-loop structure that prevents the UMI from forming non-specific PCR products. We showed that ultrasensitive variant detection is attained mainly by using UMI, while gains in sensitivity from using high-fidelity polymerases were minor. We also demonstrated that oligonucleotide quality is essential in numerous molecular applications, including SiMSen-Seq. Next generation diagnostics tools also demand optimized preanalytical conditions to achieve the necessary variant detection sensitivity, while remaining fast, simple, and cost efficient. Therefore, we established a workflow for cell-free DNA analysis and developed quantitative PCR-based quality controls to evaluate each experimental step. We also developed a bioinformatics pipeline for processing any type of targeted sequencing data containing unique molecular identifiers, including barcode clustering, error correction, variant calling, and visualization. Next, we used SiMSen-Seq in applications requiring ultrasensitive mutant detection. We first employed SiMSen-Seq to experimentally confirm that UV light rapidly induces highly recurrent mutations within a specific promotor motif. These mutations remained sub-clonal even after weeks of cell culture, arguing against a tumour-driving role. Our results highlight the importance of sequence context for the interpretation of somatic variants in cancer. We also showed that ctDNA can be used as a clinical biomarker for tumour burden and to monitor treatment efficacy in uveal melanoma. Patients with high ctDNA levels had worse overall survival, demonstrating the clinical utility of circulating tumour-DNA-based liquid biopsy analysis. In conclusion, we showed that SiMSen-Seq is a simple, flexible, low-DNA input protocol that enables rare variant detection to address a multitude of clinical and basic research questions.

Keywords: Liquid biopsy, cell-free DNA, circulating tumour DNA, molecular diagnostics, next-generation sequencing, unique molecular identifiers, melanoma

Sammanfattning på Svenska

Utvecklingen av nya DNA-sekvenseringsmetoder, kallat ”massivt parallell sekvensering”, gör det möjligt att utforska den genetiska heterogeniteten i mikrobiella, virala och tumörcellspopulationer. Detta har potential att revolutionera den molekylära diagnostiken genom att bland annat möjliggöra snabb och minimalt invasiv detektion av cirkulerande tumör-DNA (ctDNA) från blodprover och andra kroppsvätskor. Sådana vätskebiopsier innehåller dock vanligtvis endast små mängder mycket nedbrutet DNA och traditionella sekvenseringsmetoder saknar upplösning för att upptäcka sällsynta genetiska subpopulationer. Det övergripande syftet med denna avhandling var att utveckla och tillämpa SiMSen-Seq, en ultrakänslig sekvenseringsteknik som använder sig av en teknik innefattande unika molekylära identifierare (UMI) för att avlägsna fel som uppstår under den experimentella processen och därmed kan särskilja genetiska varianter i patientens DNA från bakgrunden. Vi har också utvecklat bioinformatiska analysverktyg för bearbetning, tolkning och visualisering av UMI-baserade sekvenseringsdata. Nästa generations diagnostikverktyg kräver mycket optimering av arbetsflöden för att uppnå den nödvändiga känsligheten för detektion av genetiska varianter och samtidigt vara snabba, enkla och kostnadseffektiva. Vi visade att hög polymerasfidelitet minskade SiMSen-Seq-felen, men att felkor-rigering med hjälp av UMI är betydligt högre än de förbättringar som uppnås genom att använda polymeraser med högre fidelitet. Utöver detta, visade vi även på relevansen av att ta hänsyn till ytterligare polymerasegenskaper utöver fidelitet vid utveckling av ultrakänsliga sekvenseringsanalyser. Vi visade också på många faktorer som påverkar kvaliteten på oligonukleotider, en essentiell byggsten i den experimentella processen, vilket är viktigt i många molekylära tillämpningar, inklusive klinisk diagnostik. Vi har vidare upprättat ett arbetsflöde för att analysera DNA som cirkulerar fritt i blodet och utvecklat kvalitetskontroller för att utvärdera varje experimentellt steg. Vi använde sedan SiMSen-Seq för att experimentellt bekräfta att UV-ljus inducerar många återkommande mutationer inom ett specifikt DNA-motiv, vilket belyser vikten av sekvenskontext för tolkningen av genetiska varianter i cancer. Vi ger också bevis för att ctDNA kan användas som en klinisk biomarkör för tumörbörda och behandlingseffektivitet vid uvealt melanom. Patienter med höga ctDNA-nivåer uppvisade sämre total överlevnad och ctDNA vid baslinjen var lägre för patienter som svarade på behandling an de som inte gjorde det, vilket visar på den kliniska nyttan av ctDNA-baserad vätskebiopsianalys.

List of papers

This thesis is based on the following eight studies, referred to in the text by their Roman numerals. Papers I – V relate to the development of ultrasensitive sequencing, while papers VI – VIII revolve around its applications in melanoma diagnostics.

A. Developing ultrasensitive sequencing for next-generation diagnostics

I. A Ståhlberg, PM Krzyzanowski, M Egyud, S Filges, L Stein, TE Godfrey. *Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing.* Nature protocols (2017)

II. S Filges, E Yamada, A Ståhlberg, TE Godfrey. *Impact of polymerase fidelity on background error rates in next-generation sequencing with unique molecular identifiers/barcodes.* Scientific reports (2019)

III. S Filges, P Mouhanna, A Ståhlberg. *Digital quantification of oligonucleotide synthesis errors.* Clinical Chemistry (2021)

IV. T Österlund, S Filges, G Johansson, A Ståhlberg, *UMIErrorCorrect and UMIVisualizer: Software for Consensus Read Generation, Error Correction and Visualization using Unique Molecular Identifiers* (Manuscript)

V. G Johansson, D Andersson, S Filges, J Li, A Muth, T Godfrey, A Ståhlberg. *Considerations and quality controls when analyzing cell-free tumor DNA.* Biomolecular Detection and Quantification (2019)

B. Molecular diagnostics in melanoma

VI. N Fredriksson, K Elliott, S Filges, J Van den Eynden, A Ståhlberg, E Larsson. *Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature.* PLoS genetics (2017)

VII. K Elliott, M Boström, S Filges, M Lindberg, J Van den Eynden, A Ståhlberg, A Clausen, E Larsson. *Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers.* PLoS genetics (2018)

VIII. L Ny, H Jespersen, V Sah, J Karlsson, S Alsén, S Filges, C All-Eriksson, B Andersson, Roger Olofsson Bagge, A Carneiro, H Helgadottir, M Levin, I Ljuslinder, U Stierner, A Ståhlberg, G Ullenhag, Lisa Nilsson, J Nilsson. *Preclinical and phase 2 assessment of combined histone deacetylase and PD-1 inhibition in metastatic uveal melanoma.* Nature Communications (2021)

Other relevant papers *not* included in the thesis:

I. S Bjursten, C Vannas, S Filges, F Puls, A Pandita, H Fagman, A Ståhlberg, Max Levin. *Response to BRAF/MEK Inhibition in A598-T599insV BRAF Mutated Melanoma.* Case reports in oncology (2019)

II. C Vannas, S Bjursten, S Filges, H Fagman, A Ståhlberg, M Levin. *Dynamic ctDNA evaluation of a patient with BRAF V600E metastatic melanoma demonstrates the utility of ctDNA for disease monitoring and tumor clonality analysis.* Acta Oncologica (2020)

III. M Egyud, P Sridhar, A Devaiah, E Yamada, S Saunders, A Ståhlberg, S Filges, P Krzyzanowski, I Kalatskaya, W Jiao, L Stein, S Jalisi, T Godfrey. *Plasma circulating tumor DNA as a potential tool for disease monitoring in head and neck cancer.* Head & Neck (2018)

IV. M Egyud, M Tejani, A Pennathur, J Luketich, P Sridhar, E Yamada, A Ståhlberg, S Filges, P Krzyzanowski, J Jackson, I Kalatskaya, W Jiao, G Nielsen, Z Zhou, V Litle, L Stein, T Godfrey. *Detection of Circulating Tumor DNA in Plasma: A Potential Biomarker for Esophageal Adenocarcinoma.* The Annals of Thoracic Surgery (2019)

V. H He, E Stein, Y Konigshofer, T Forbes, F Tomson, R Garlick, E Yamada, T Godfrey, T Abe, K Tamura, M Borges, M Goggins, S Elmore, M Gulley, J Larson, L Ringel, B Haynes, C Karlovich, M Williams, A Garnett, A Ståhlberg, S Filges, L Sorbara, M Young, S Srivastava, K Cole. *Multilaboratory Assessment of a New Reference Material for Quality Assurance of Cell-Free Tumor DNA Measurements.* The Journal of Molecular Diagnostics (2019)

Contents

1

Introduction

The human genome was sequenced for the first time in 2001, after a decade long effort that cost 2.7 billion dollars[1,2]. Today, a genome can be sequenced in a few days for less than 1000$. Due to this reduction in cost, the number of individual people with at least partially sequenced genomes has grown from 2 in 2007 to over 30 million in 2021[3]. The actual information content of a haploid human genome requires only 760 megabytes of storage. However, to ensure even coverage and high confidence, each base must be sequenced dozens of times, increasing the required data volume to >100 gigabytes per genome. By some estimates genomic data will be generated at a rate of several exabytes per year in 2025[4], where an exabyte corresponds to one billion gigabytes.

This astronomical amount of data can only be produced due to the advances in genome research that were made in the last century. Although deoxyribonucleic acid (DNA) was first discovered in 1869[5], it took until 1944 to understand that DNA, not proteins, contain the hereditary information[6]. Every subsequent decade saw further milestones in our understanding of the genome. The DNA double helix structure, and thus a mechanism for DNA replication, was unravelled in 1953[7]. The genetic code that enables the translation of information encoded in DNA into proteins was deciphered by 1966[8]. The 1960s also saw the first sequencing of nucleic acids, beginning with a 77 nucleotides long yeast tRNA in 1965[9]. The same method was used to sequence the first whole genome in 1976 - that of bacteriophage MS2, containing only 3569 nucleotides[10].

The next technological breakthrough came in 1977, when Frederick Sanger developed the eponymous "Sanger sequencing" technique. Sanger sequencing uses fluorescently labelled nucleotides that prevent binding of further nucleotides to the DNA template. The DNA is amplified using both labelled and unlabelled nucleotides yielding fragments of various size each ending with a labelled nucleotide. These fragments can then be separated and visualized using gel-electrophoresis[11]. An automated approach for Sanger sequencing was first introduced by Applied Biosystems in 1985, allowing sequencing with much higher throughput and making it the first generation of sequencing to become widely used[12]. As the Sanger technique is limited to templates of less than 1000 bp, shotgun sequencing approaches were developed that combine many short fragments *in silico*[13,14]. Shotgun

sequencing was also used in the assembly of the first human genome during the Human Genome Project (1990-2001)[1,2].

1.1 Development of high throughput sequencing

As became obvious from the time and effort it took to sequence just a single human genome, further advances in genome biology required new technologies with even greater throughput. Pyrosequencing became the first in a wave of second generation sequencing approaches to be commercialized by 454 Life Sciences in 2005 and was later acquired by Roche[15]. Pyrosequencing uses fragmented genomic DNA to which adapter sequences are ligated. Then each fragment is bound to a 28 µm diameter bead under reaction conditions that facilitate the binding of only a single template molecule per bead. DNA on each bead is then amplified using emulsion PCR, before the bead containing the enriched template is embedded in a picolitre-sized well of a micro titre plate. This allows the parallel sequencing of hundreds of thousands of 110 bp long stretches of DNA, so called reads, on a single plate. At each sequencing cycle a different nucleotide species is added to the plate. Every time a nucleotide is incorporated into a DNA fragment a free pyrophosphate molecule is released which triggers light emission from luciferase in that fragment's well.

A similar technology was developed by Solexa in 2006 and acquired by Illumina in 2007. The two main differences to Pyrosequencing are that template enrichment is performed on a solid phase flow cell using bridge amplification and all four nucleotides can be added at once in each cycle instead of being added iteratively[16] (Figure 1.1). As in Pyrosequencing, DNA templates need to be fragmented and contain adapters for hybridization onto the flow cell surface, which can be incorporated by either ligation or PCR. Each template is then enriched in an isothermal bridge amplification resulting in a cluster of roughly 1000 copies[17]. Fluorescently labelled nucleotides containing a reversible terminator molecule similar to Sanger sequencing are then added in each cycle. The terminator molecule prevents binding of more than one nucleotide per fragment at a time. At the end of each cycle the terminator is cleaved off, resulting in a distinct fluorescence signal for each type of nucleotide, while simultaneously providing space for binding of another nucleotide in the following cycle (Figure 1.1). Initially, only short reads of 35 bp could be analysed using the Illumina/Solexa platform, although more than 30 million reads could be generated in parallel, compared to 200,000 for Pyrosequencing[18].

The third major technology to reach the market was SOLiD in 2009, which uses sequencing by ligation instead of sequencing by synthesis as 454 or Illumina. The original SOLiD approach could sequence 100 million 35 bp long reads with roughly 10 times lower error rates compared to Illumina, albeit with lower specificity and a bias towards under representing AT sequences[16].

Although its read length was eventually increased to 1000 bp, Pyrosequencing was discontinued by Roche in 2016, after Illumina also increased its maximum read length to several hundred bp[16]. However, IonTorrent, which was introduced in 2010 based on 454 technology is still widely used and now owned by ThermoFisher. IonTorrent uses a non-optical detection method with the same workflow as Pyrosequencing but detects H^+ ions released after nucleotide insertion instead of light emission[19].

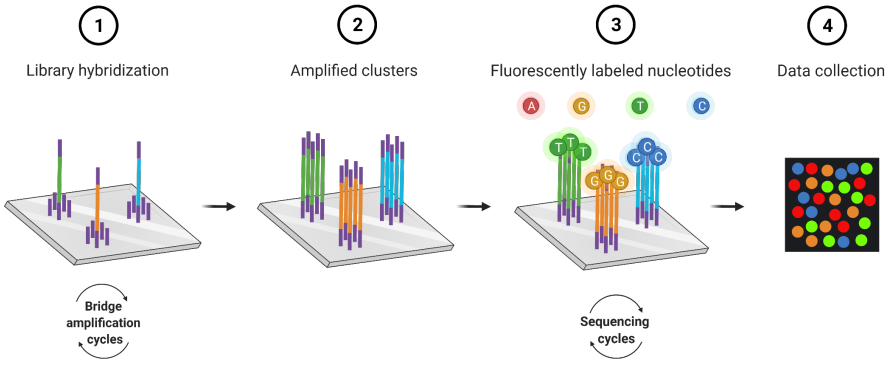

Figure 1.1: Illumina sequencing by synthesis. Created with BioRender.com.

As of 2022, Illumina has become the clear market leader for short read, high throughput sequencers, due to its mature technology and wide array of instruments[16,18]. Illumina platforms have the highest throughput and longest reads of currently available second-generation sequencing systems. Illumina's latest NovaSeq instrument can generate up to 20 billion paired end 250 bp reads, resulting in the lowest cost per base of any sequencing technology. Despite the predominance of the Illumina platform, it comes with several disadvantages[18]. The random clustering on the flow cell requires tightly controlled loading concentrations with overloading resulting in merged clusters, reducing quality and yield. Additionally, samples with low sequence complexity like amplicons or 16s metagenomic libraries require the addition of a reference material in large quantities, reducing the available number of reads per sample. Illumina platforms perform relatively well for homopolymer regions introducing few insertions and deletions (InDels), although they have a bias towards under representing both AT- and GC-rich regions and producing substitution errors[20–22].

The SOLiD 5500xl platform has the second highest throughput after Illumina with 1.8 billion reads and the lowest error rate, despite an AT bias. However, SOLiD also has the shortest reads of all technologies with a maximum of 75 bp and a much smaller range of available library preparation kits and instruments[18]. The most recent IonTorrent platform, Ion S5 550, has comparatively low throughput with 130 million, single-end 200 bp reads but has the fastest run times, making it particularly suitable for targeted and clinical point-of-care sequencing[23]. IonTorrent also struggles with sequencing homopolymer regions and has as bias towards producing InDels[16].

The most recent development in sequencing technology was the introduction of "third generation" long read sequencers by Pacific Biosciences and Oxford Nanopore in 2011 and 2014, respectively[16]. Initially, their relatively high cost per base, high error rate and relatively low throughput limited the range of applications where long read sequencing offered an advantage over state-of-the-art second generation technologies[16]. However, long read technologies proved ideal for genome assembly, sequencing long stretches of homopolymers and full-length RNA. Furthermore, the Oxford Nanopore MinION device is only 2 x 10 cm in size and can interface with a personal computer using USB, making

it by far the most portable sequencing technology[16]. This has already been used for tracking viral disease outbreaks in real-time at the point-of-care, especially in hard-to-reach regions[24]. While long-read sequencing is still mostly relegated to niche applications, the field is evolving rapidly and Oxford Nanopore's new PromethION has throughput comparable to large Illumina sequencers[25]. Similarly, data quality is also improving with Pacific Biosciences reaching >99% base accuracy using circular consensus sequencing[16,26]. Both technologies can also use native DNA, avoiding biases introduced by amplification[25]. Further reduction in cost and improved accuracy may lead to a wider adoption of long read sequencers in the near future[25].

1.2 Ultrasensitive sequencing

Genetic variations are a central feature of biology, driving evolution by generating diversity and contributing to the development of many diseases[27–29]. The availability of affordable high throughput sequencing (HTS) has enabled the exploration of the genetic heterogeneity within microbial[30,31], viral[32,33] and tumour cell[34–36] populations. This also revealed the existence of rare subpopulations of cells underlying anti-cancer[37] and anti-microbial drug resistance[38] , adaptive immunity[39,40], and genetic mosaicism[41–44], explaining the phenotypic variability of many diseases.

In principle, any size of genetic subpopulation can be detected using deep sequencing, where each region of the genome is analysed ten thousand times or more. However, background error rates limit the detection of rare subpopulations. Traditional Sanger sequencing has high fidelity and low cost but low throughput and a sensitivity of only 15% variant allele frequency (VAF) and is therefore not suitable for detecting rare mutations[45]. HTS has potentially genome-scale throughput and is much more sensitive than Sanger sequencing. Yet, even with deep coverage, HTS has a limit of detection of 1% VAF mainly due to polymerase errors incurred during library preparation and sequencing itself[45–47]. Other potential sources of error are DNA damage from heat[48], ultrasonic shearing[49] or formalin fixation[50–52].

Robust calling of variants below 1% allele frequency remains challenging[47], despite the availability of computational filters that remove low-quality sequences[53], handle sequence misalignment[54], identify noise patterns from oxidative damage or DNA polymerases[55], and the development of statistical background error models[56,57]. Nevertheless, deep sequencing has been used in a variety of fields, where a limit of detection well below 0.1% would be desirable, including metagenomics[58,59], forensics[60], paleo-genomics[61,62], prenatal screening[63–65], immunology[66], epigenetics[67–69], and the detection of circulating tumour DNA (ctDNA)[70–73].

1.2.1 Error correction using molecular barcodes

To address the challenges of ultra-rare variant allele detection, a plethora of methods have been developed, each with their own strengths and weaknesses[74,75]. PCR-based detection methods have very high sensitivity, especially digital PCR[76]. However, PCR assays have limited capacity for multiplexing and low throughput. Sequencing-based approaches use

over-sampling to analyse the same original molecule multiple times and subsequently construct an error-free consensus sequence. Examples of this are SafeSeqS[70,71], SiMSen-Seq[72,73], CAPP-Seq[77,78] and DuplexSeq[46,79], which utilize molecular barcoding to remove erroneous sequence variants.

Barcodes are also referred to as unique molecular identifiers (UMI). They are attached to each original sample molecule at the initial stages of analysis, making it possible to trace all amplified molecules back to their origin[46,79]. Barcodes can be attached to target DNA using PCR[70,72] or ligation[46] (exogenous) or can be inferred from the template itself[80,81] (endogenous). If a wildtype template molecule is amplified accurately, all products containing the same barcode will appear free of mutations, while a PCR error will lead to an apparent mixture of mutation-free and mutated daughter molecules (Figure 1.2). All amplified molecules containing the same barcode are called a barcode family. This makes it possible to differentiate a true mutation, which must be present in practically all molecules of barcode family, from an artefact that is present only in a smaller subset (Figure 1.2). This approach reduces error by a factor of up to 1000 times, enabling single-molecule analysis and limits of detection well below 0.1% variant allele frequency[70,71,73,79–81].

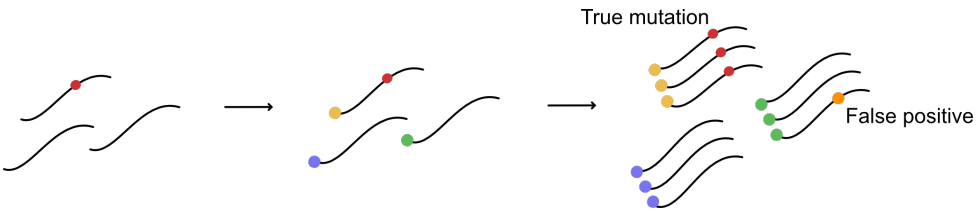

Figure 1.2: Molecular barcoding. Molecules in a sample (left) are barcoded by attaching UMI (middle) and subsequently amplified (right). Daughter molecules are grouped by UMI into barcode families. True mutations must be present in all molecules of a given family, while errors only exist in a subfraction.

Various UMI designs have been reported[72,82], although the optimal length and structure of the UMI depends on the specific application. Generally, the number of available UMI should greatly exceed the number of template molecules. Typical samples will include thousands of template molecules, but a twelve-nucleotide long, fully randomized UMI yields 16.7 million possible sequence combinations, making it highly unlikely that two different template molecules receive the same barcode[83,84].

UMI are commonly attached to target molecules using either ligation or PCR. Ligation-based barcoding approaches are often more complex than PCR-based strategies. They are more time consuming, have lower efficiency, require target enrichment and involve more cleaning steps that lead to loss of material[46,77]. On the other hand, ligation approaches can easily cover broad ranges of DNA that would necessitate overlapping PCR assays, which may introduce unspecific products[85]. In scenarios where template DNA is highly fragmented, PCR assays need to be very short or they might fail to amplify a large portion of the available templates[86], whereas intact DNA will require fragmentation prior to ligation which may itself introduce biases[49]. Although a working PCR protocol may be easier and quicker to perform than a ligation-based protocol, as it allows barcoding and target enrichment in a single step, multiplexed PCR panels may require substantial development and optimization beforehand[87,88]. Thus, UMI ligation followed by hybridization capture

may be preferable over PCR-based target enrichment for larger panels.

1.2.2 Variant calling using UMI data

HTS can generate millions of reads per sample and requires specialized bioinformatic tools for processing and analysis. Standardized analysis workflows exist for variant calling in exome and whole genome sequencing data. These can remove PCR duplicates, handle alignment errors and sometimes integrate results from multiple variant calling algorithms to improve fidelity of mutation calls[89–91]. However, they do not take UMI information into account and therefore remain fundamentally limited by the errors incurred during library preparation and sequencing. Several UMI-based variant calling algorithms for targeted amplicon sequencing have been published, including DeepSNVMiner[92], MAGERI[93], smCounter2[94], UMI-VarCal[95], and debarcer[72]. Other methods exist for capture-based approaches, such as iDES[78], or for long read sequencing with UMI[96].

DeepSNVMiner and debarcer are conceptually similar, relying only on heuristic thresholds to determine variant alleles. DeepSNVMiner has a default minimum barcode family size of 5, where mutations are called if they are present in at least 40% of reads in a barcode family. Debarcer allows variable consensus depth cut-offs and requires 100% of reads to have the same variant for barcode families under 20 reads to be called as a true variant. Neither algorithm uses a statistical model for background error rates, nor can they provide an estimate for the confidence in the called variants.

MAGERI and smCounter2 estimate background error rates using a beta-binomial distribution[97]. In this model, the number of sequencing errors x is assumed to follow a binomial distribution $x \sim Bin(d, p)$, which itself depends on the sequencing depth d and error rate p, which is modelled as a random variable derived from a Beta distribution $p \sim Beta(\alpha, \beta)$. However, MAGERI assumes a universal beta distribution at all sites, whereas smCounter2 uses site-specific estimates for each type of base substitution. MAGERI is a self-contained pipeline that does not require external tools but is significantly slower and requires dedicated computing resources[95]. smCounter2 was developed specifically for the QIAseq targeted sequencing protocol and is not available as an open-source software anymore.

UMI-VarCal estimates background noise from base quality scores. Variants are then called using a Poisson test. UMI-VarCal does not take FASTQ files as inputs and thus depends on the user to perform sequence alignment separately. UMI-VarCal does not require any external data set to model background error rates and uses sequencing quality scores instead. However, most sequencing errors should be corrected by UMI clustering alone, whereas errors introduced during the first cycle of barcoding or chemical modifications introduced during library preparation can have high base quality scores and therefore might be better estimated using a statistical distribution[93].

1.3 Liquid biopsy

Many tumours are now characterized on a molecular level by sequencing DNA obtained from a tissue biopsy which is also used for histological assessment of the tumour[98–100].

Tissue biopsies are invasive procedures, that only reflect a single time-point in the tumour development, may not give a full picture of tumour heterogeneity[101] and have been implicated in promoting tumour spread[102,103]. Evolution can significantly alter the genetic landscape of a tumour over time[104,105] , promoting therapy resistance[37,106,107] and genetic divergence of primary tumours and metastases[108–111].

Liquid biopsies are the sampling and analysis of body fluids and offer a compelling alternative or addition to standard tissue biopsies[71,74,112–114]. Body liquids such as blood[71,115–117], urine[118–120], or sputum[120] can be obtained cost-effectively and using non- or minimally invasive procedures, which allows frequent longitudinal sampling[115,121–123]. Additionally, liquid biomarkers, whether they be DNA[70–72], proteins[71,124,125], RNAs[126–132], metabolites[133] or circulating tumour cells (CTCs)[134–136] are likely to give a more comprehensive picture of the disease than a single tumour biopsy[137,138] (Figure 1.3).

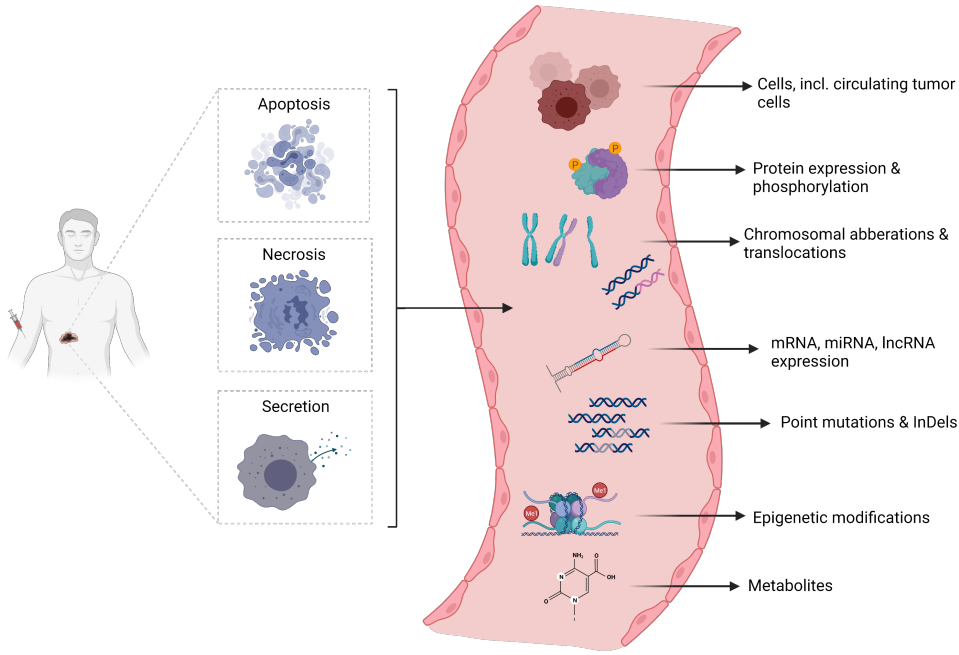

Figure 1.3: Analytes available from liquid biopsies. Adapted from Wan *et al.*[74] and created with BioRender.com.

1.3.1 Cell-free DNA

The presence of cell-free DNA (cfDNA) in the blood was first described in 1948[117]. However, its potential as a clinical biomarker was only established over 40 years later when circulating tumour DNA was discovered in the circulation of cancer patients[139]. Since then, the analysis of cfDNA has matured and is now used clinically for non-invasive pre-natal testing[140,141] and *EGFR* mutation testing in non-small cell lung cancer[142]. Beyond these already established techniques, growing evidence suggests the wide-spread applicability of cfDNA based biomarkers for the detection and minimally invasive monitoring of solid

tumours, patient stratification, discovery of therapy-resistant clones and comprehensive profiling of spatio-temporal genomic heterogeneity of metastatic disease[71,143]. Besides oncology, cfDNA is investigated in fields as varied as diabetes[144,145], transplant rejection[146] and sepsis[147].

Origin of cfDNA

Current evidence suggests that cell-free DNA is released by both neoplastic and non-neoplastic cells into the surrounding body liquids through processes such us apoptosis, necrosis, or active cellular secretion[148] (Figure 2). Typically cell-free DNA is highly fragmented with an average size of 167 bp or multiples thereof, corresponding to the size(s) of individual nucleosomes[149,150]. This is consistent with caspase-dependent cleavage of DNA and therefore most cfDNA likely originates from apoptotic cells[151–153]. Longer DNA fragments (>1000 bp) are thought to be derived from necrotic cells[151] or to have been incorporated in exosomes[154,155]. Notably, physiological factors such as pregnancy[156] or exercise[157] and disorders such as cancer[158], physical trauma[159], inflammation[160], obesity[161], or cardiovascular disease[162] may impact the total amount of cfDNA.

Clearance of cfDNA

CfDNA is rapidly cleared from the blood stream and has an estimated mean half-life between 15 minutes and 2.5 hours[156,163]. A number of cfDNA clearance systems have been described, although the exact mechanism remains unclear[164,165]. Several organs are responsible for clearing cfDNA, primarily the liver, spleen and kidneys[165]. Animal experiments indicate that nucleosomes are rapidly trapped in the liver and that the kidney plays only a smaller role in the clearance of apoptotic cfDNA[166]. This is further supported by data showing that patients with chronic renal failure do not show increased levels of cfDNA[167]. Yet, tissue distribution of DNase I activity in mice is similar between liver and kidney and nearly twice as large in the spleen[168]. The same study showed that urine contains the highest DNAse I activity, nearly 20 times more than liver or kidney, likely explaining the high level of degradation of cfDNA from urine compared to plasma. In the blood cfDNA is degraded predominantly by DNAse I, factor VII-activating protease and factor H, although their overall effect on cfDNA degradation are likely minor[165,169].

Function of cfDNA

Several studies also suggest a functional role of cfDNA as an immune system modulator[170]. Extracellular mitochondrial DNA activates leucocytes and stimulates the release of pro-inflammatory cytokines. These cytokines can then lead to changes in the tumour microenviroment through induction of tumour promoting macrophages[165]. Cell line studies have suggested that cfDNA can also perform regulatory roles through horizontal gene transfer, although effects are highly cell-type dependent and usually transient[165,171,172]. Besides apopotosis and necrosis neutrophil extracellular trap release (NETosis) is another source of cfDNA. NETs were shown to promote the development of the metastatic niche by protecting tumour cells and stimulating cell proliferation[165,173].

1.3.2 Circulating tumour DNA

Circulating tumour DNA (ctDNA) represents the fraction of total cfDNA originating from neoplastic cells, which can be differentiated from normal tissue cfDNA through genetic or epigenetic alterations that have occurred in the tumour[143,158]. Levels of total cfDNA in presumably healthy individuals or patients with minimal disease burden are highly variable, but ordinarily around 2 - 10 ng cfDNA per mL of plasma[174]. Increased levels of total cfDNA are sometimes observed in cancer patients with severe metastatic disease[151,158]. The tissue origin of cfDNA in healthy individuals is heavily skewed towards various blood cell types and endothelial cells with less than 6% of cfDNA originating in the other major organs[175].

For most clinically interesting applications the tumour burden will be low and only a small fraction of the total cfDNA will come from the tumour (∼0.1% or less). At a ctDNA fraction of 0.1% this translates to approximately 0.54 - 2.7 tumour-derived molecules per mL of plasma, assuming roughly 270 haploid genomes per ng of genomic DNA[176,177]. Depending on the location of the tumour(s), different body liquids may be more appropriate sources of ctDNA, such as cerebrospinal fluid for cancers of the central nervous system[178,179], urine for prostate[119] and bladder cancer[118], Papanicolaou test smears for ovarian cancer[180], sputum for lung cancer[120] and saliva for head-and-neck cancers[181,182].

Due to the rarity of ctDNA in a typical sample even a method with single molecule resolution sampling error can lead to false negative result. Since the amount of sample is usually the limiting factor, this implies that several strategies could improve the sensitivity of ctDNA detection: Development of methods with single molecule resolution to capture all available molecules, utilization of multiple independent targets and improvement of sample recovery with optimized pre-analytical workflows.

Detection of ctDNA

Methods that have been used extensively for the detection of ctDNA include qPCR, droplet digital PCR (ddPCR), and high-throughput sequencing with and without UMI error correction. Each of these have different cost, sensitivity, flexibility, and scalability. PCR-based ctDNA detection approaches are cheap, relatively easy to perform and have among the highest sensitivities of any method, especially ddPCR and BEAMing[76,183–186]. However, they are limited to analysing a small number of individual loci and are therefore preferably used in scenarios where mutations are already known, such as detecting highly recurrent hotspot mutations[74]. Conversely, exome sequencing has genome scale, but is limited to mutations >1% allele frequency[187]. Whole genome sequencing has the same limitations for the detection of small somatic variants as exome sequencing, and it would be prohibitively expensive to perform genome-wide deep sequencing, although large genomic rearrangements can be detected with high sensitivity using PARE[188–190].

Therefore, targeted-sequencing approaches have been developed to combine the throughput and flexibility of sequencing with relatively moderate cost. Since they do not require knowledge about existing mutations, they can be used to search for *de novo* resistance mutations and monitor clonal evolution[74]. Conventional deep sequencing panels without UMI error correction such as Ampliseq have limited sensitivity[191,192], but UMI-based approaches

such as SafeSeqS, SiMSen-Seq or CAPP-Seq have very high sensitivities, comparable to ddPCR[70–73,77]. Their flexibility enables the design of patient-specific panels that target multiple loci at the same time, which may further increase their sensitivity[74]. The main disadvantages with targeted sequencing over PCR-based approaches are longer turnaround time, expensive equipment and the need of bioinformatics expertise[74,193].

Another challenge is biological background noise that cannot be removed using UMI or other digital approaches. Clonal haematopoiesis of indeterminate potential (CHIP) leads to the age-related, non-malignant expansion of blood cells caused by common oncogenic mutations, which can be detected by ultrasensitive methods[44,194,195]. Genes associated with CHIP include *TP53*, *DNMT3A*, *ASXL1*, *TET2*, *JAK2*, *SF3B1*, *CBL*, *GNAS*, and *IDH*[194–196]. Sequencing matched blood cell DNA from the same patient may partly alleviate this issue by identifying these background mutations.

1.3.3 Liquid biopsy using ultrasensitive sequencing

Most liquid biopsy studies, including the work presented in this thesis, are based on blood samples. Typically, whole blood is collected in specialized collection tubes and plasma is isolated by centrifugation from which cell-free DNA can then be extracted. Plasma is preferable over serum as the latter may contain contaminating DNA from damaged blood cells[197]. Numerous preanalytical factors influence the yield, integrity, and purity of cfDNA, including blood collection tubes, storage temperature and extraction method[75,198,199]. Blood collected in ordinary EDTA tubes must be processed within hours, plasma should be stored at -80°C and repeated freeze/thaw cycles should be avoided[198]. Extended storage may also reduce cfDNA yield[200]. Specialized blood collection tubes have been developed to preserve cfDNA in cases where sample logistics does not allow immediate plasma isolation, although they may negatively affect yield and the ability to perform other analyses on the same sample[201–204]. A plethora of cfDNA extraction methods have been used in the literature and most either use magnetic beads or silica membranes to isolate cfDNA[205–207]. Novel liquid phase extraction protocols may be able to improve cfDNA extraction yield by 60% over established solid-phase methods[208]. However, the lack of standardized protocols and operating procedures makes it difficult to compare results from different studies[205,209,210]. After cfDNA isolation, any of the methods discussed above may be used for ctDNA detection.

The most relevant use cases for liquid biopsies in oncology are (i) screening or early detection (ii) molecular profiling and prognostication and (iii) treatment monitoring for detection of recurrence and identification of therapy resistance mutations (Figure 1.4). Screening may be the most difficult of the three, as early-stage tumours will be small and overall ctDNA levels extremely low. Therefore, it may not be possible to achieve the sensitivity necessary for cancer screening using small hotspot mutations panels. Commercial applications for cancer screening are still in development and one of the most prominent is GRAIL's Galleri platform, which uses differential methylation profiles in ctDNA instead of mutations. Analysing thousands of differentially methylated sites instead of a small number of mutations might significantly improve sensitivity[211].

The majority of clinical studies involving ctDNA analysis have been retrospective.

They have shown for many different cancer types that ctDNA levels are correlated with tumour burden and that residual ctDNA immediately after treatment is a sign of worse survival outcomes[74,193]. Recurrence and therapy-resistant clones have also been detected using ctDNA, often weeks or months before recurrence was confirmed on imaging[212,213]. All of this suggests that ctDNA can be used to evaluate clinically relevant parameters. However, the paucity of prospective clinical data means that the actual patient benefit of ctDNA-guided clinical decision making remains uncertain[214].

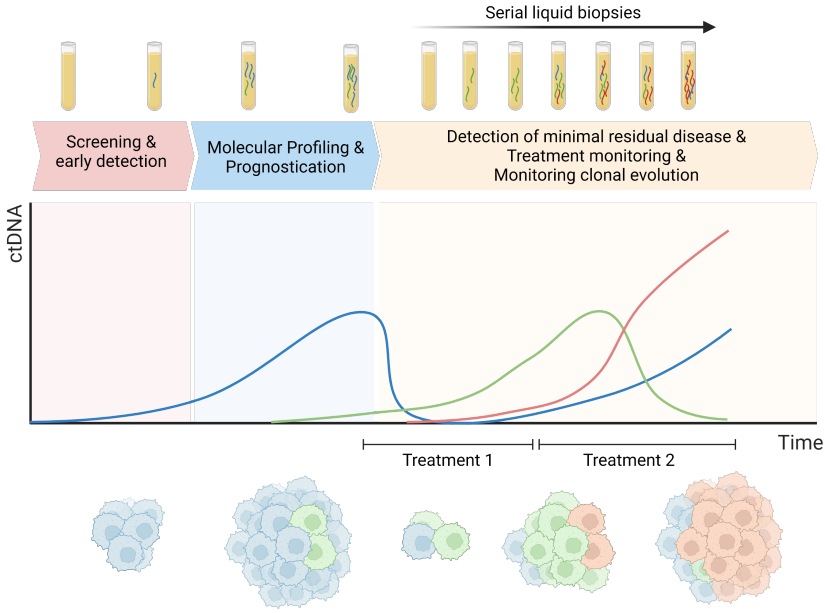

Figure 1.4: Clinical liquid biopsy applications. Adapted from Wan *et al.*[74] and created with BioRender.com.

Currently there only a few commercially available ctDNA platforms have gained FDA-approval. FoundationOne Liquid CDx[215] and Guardant360 CDx[216,217] use targeted HTS panels for mutation profiling across multiple genes, whereas Cobas *EGFR* Mutation Test v2[218,219] and therascreen *PIK3CA* RGQ PCR kit[220,221] utilize qPCR to detect a small selection of clinically relevant mutations in *EGFR* or *PIK3CA*, respectively. The Cobas *EGFR* test has been used as a companion diagnostic in 1st and 2nd line EGFR tyrosine kinase inhibitor therapy for the detection of treatment resistance mutation such as *EGFR* T790M[218]. The therascreen *PIK3CA* test is used instead to identify *PIK3CA* positive breast cancer patients who will respond to alpeslib[221].

1.4 Melanoma

Cancers arise from the transformation of a normal cell into a tumour cell through the accumulation of genetic and epigenetic alterations over time[104,105]. All tumours share

hallmark properties such as replicative immortality, avoidance of immune destruction, genomic instability, and invasiveness[222]. Tumour entities are genetically heterogeneous both between and within different cancer types and a single tumour may consist of many genetically distinct subclones[36,108,223–225]. This heterogeneity means that, given enough time, almost inevitably individual clones will arise that are resistant to even the best therapy[226,227]. Novel therapies, such as BRAF inhibitors and immunotherapy initially show impressive clinical results for many patients. Yet, in most cases these results are short-lived and patients eventually suffer from tumour recurrence, therapy resistance and fatal metastatic spread of the disease[226,228,229].

Melanomas are cancers of the melanocytes which occur primarily in the skin[230] and eye, called uveal melanoma[231]. Skin melanomas are characterized by a strong UV-radiation induced mutation signature and an overall high tumour mutational burden. Despite this large number of mutations per tumour there is only a moderate number of highly recurrent driver mutations[232,233]. Over 80% of skin melanomas are driven by mutually exclusive mutations in *BRAF*, *NRAS*, *KIT* or *NF1*[233]. Other recurrent mutations occur in the tumour suppressors *TP53*, *PTEN* and the oncogenes *RAC1*, *PIK3CA*, *HRAS* and *WT1*[232]. In contrast, uveal melanomas typically lack a distinct UV-signature and are instead driven by mutually exclusive hotspot mutations in *GNA11*, *GNAQ*, *CYSLTR1* and *PLCB4*, all of which occur rarely in skin melanoma[231]. The tumour suppressor genes *BAP1* and *SF3B1* are also frequently mutated in uveal melanoma[231].

1.4.1 UV-induced mutations

Absorption of ultraviolet (UV) photons by DNA causes the formation of dimers between adjacent pyrimidine bases (C or T), primarily by inducing cyclobutane pyrimidine dimers (CPD) and pyrimidine 6-4 pyrimidone photoproducts (6-4PP), although CPDs are far more common. Pyrimidine dimers distort the DNA double helix and may result in stalling of the replication fork and fatal double-strand breaks[234]. To prevent this, they can be repaired with nucleotide excision repair (NER), where the dimer is excised from the affected strand and subsequently repaired using information from the opposite strand[235]. However, when NER is impaired, or DNA replication is attempted before NER was active, specialized DNA polymerases may be recruited to a stalled replication fork and enable replication by trans-lesion synthesis (TLS). At this stage, mutations are incurred at CPD sites by two different mechanisms: Some TLS polymerases may insert an adenine on the opposite strand of the lesion, which is then matched with a thymine. Alternatively, spontaneous deamination of cytosine to uracil (the RNA equivalent of thymine) causes the DNA polymerase to correctly pair an adenine with the uracil, resulting in a C→T transition after replication (Figure 1.5)[234].

1.4.2 Molecular diagnostics in melanoma

Histopathological analysis of the primary tumour biopsy remains the gold standard for diagnosis and staging of melanoma. However, tissue pathology requires many years of training and carefully defined diagnostic criteria. Nevertheless, many histological assessments have low concordance between different pathologists, especially for separating

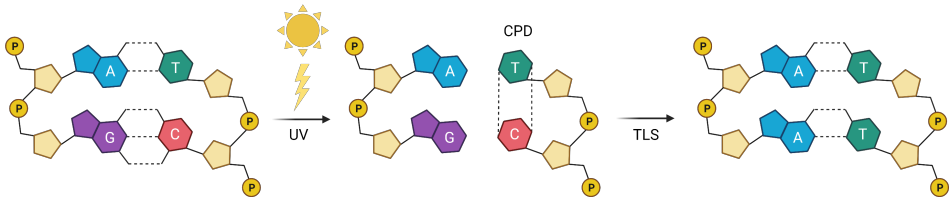

Figure 1.5: Mutations induced by UV radiation. UV radiation causes the formation of CPDs and subsequent TLS over CPDs leads to C→T mutations. Created with BioRender.com.

early stage tumours[236]. Molecular markers are therefore increasingly used to supplement histopathological assessment for diagnosis, by delineating tumour subtypes and guiding clinical decision making[237]. Uveal melanoma is often asymptomatic and typically diagnosed by imaging alone, although tissue biopsies may be used for prognostication and mutational profiling[238,239]. Hence, there is a great need for improved diagnosis and minimally invasive biomarkers for prognostication and treatment monitoring.

Serum protein levels of lactate dehydrogenase (LDH) and S100B are also correlated with clinical outcomes[240] and are widely used for prognostication[241]. Although LDH has relatively low specificity, while S100B is considered to be more specific and reliable, both are easily measured biomarkers and predictive of poor response in late-stage melanomas[242]. However, while S100B is a good marker for skin melanoma it has no predictive value in uveal melanoma[243].

Mutational status can be a potent predicator of treatment response. For instance, *NRAS* positive tumours will be unresponsive to BRAF inhibitors[244], since *BRAF* and *NRAS* mutations are mutually exclusive. Since melanomas have highly recurrent mutations it is a suitable tumour type for ctDNA detection and numerous studies were able to detect ctDNA in melanoma patients. Some have used qPCR[245], ddPCR[246] or BEAMing[247] to detect BRAF/NRAS mutations in melanoma patients and had generally high specificity, although the sensitivity was limited and ctDNA was not detectable even in some patients with advanced disease[245,248,249]. This highlights the difficulty of detecting mutations from limited sample material when only a single mutation is analysed. Although ctDNA levels correlate with tumour burden, the fact that even patients with confirmed advanced disease can be ctDNA negative suggests that the dynamics of ctDNA release and clearance are complex[115,158,196].

Other studies have used CAPP-Seq combined with ddPCR[250] or targeted SiMSen-Seq panels[115,251,252]. ctDNA has also been used to monitor therapy response, where both pre- and post-treatment presence of ctDNA has been associated with worse survival outcome[249,253]. In some cases, ctDNA was detectable several months prior to radiological evidence of progression[254] and thus may be useful clinical tool to identify melanoma patients with high risk of recurrence. Liquid biopsies may also be used in cases where primary tumour material is not available[250].

Immunotherapy has led to drastically improved outcomes in a subset of melanoma patients, although mechanisms and biomarkers of resistance to immunotherapy remain unclear[255]. Therefore, identifying patients that will benefit from immunotherapy or detecting markers of resistance could greatly improve the pool of patients receiving

immunotherapy[256]. Immunotherapy can also cause pseudo-progression where responding tumours actually appear to be increasing in size on imaging and ctDNA might offer a more robust measure of tumour burden in these cases[257]. One study showed that patients with no detectable ctDNA after eight weeks of immunotherapy had significantly improved survival[246]. Persistent levels of ctDNA during immunotherapy can be predictive of survival outcomes[258] and differences in ctDNA clearance after two cycles of therapy can distinguish treatment responders and non-responders[259]. Similar results were reported by Ashida *et al.* who also noted that ctDNA was a superior predictor of treatment response over LDH[260]. Similarly, several studies showed that ctDNA was highly correlated with serum S100 levels and an equivalent or superior marker of tumour burden for skin melanoma[261,262].

2

Aims

The overall goal of this thesis was to develop and implement ultrasensitive sequencing technologies in clinical and research applications that require the ability to detect single nucleotide changes in individual molecules. The specific aims for each study are:

Paper I: To develop a simple, flexible, and generic ultrasensitive sequencing protocol for targeted sequencing.

Paper II: To define the properties of different DNA polymerases in ultrasensitive sequencing.

Paper III: To characterize oligonucleotide synthesis errors and determine the role of oligonucleotide synthesis strategies, purity grades, batch, and sequence context for the use of oligonucleotides in high-performance applications, such as ultrasensitive sequencing.

Paper IV: To develop a bioinformatical pipeline and tools for processing, analysis, and visualization of barcoded sequencing data.

Paper V: To develop a workflow for analysing liquid biopsies using ultrasensitive sequencing, including the use of quality controls.

Paper VI: To characterize the role of sequence context-dependent hotspot promoter mutations from UV exposure.

Paper VII: To understand the role of DNA repair for the emergence of hotspot promoter mutations in melanoma.

Paper VIII: To develop and apply a tailor-made ultrasensitive sequencing approach to quantify ctDNA levels in uveal melanoma patients enrolled in a phase II clinical trial investigating the efficacy of combined epigenetic and immunotherapy.

3

Results and Discussion

Ultrasensitive sequencing technologies need to fulfil three main criteria to become an integral part of next generation diagnostic tools. Firstly, they need to achieve single-molecule resolution by removing or suppressing polymerase errors introduced during library preparation and sequencing. Secondly, they need to be flexible, scalable and work with limited and challenging sample types to meet clinical requirements. Lastly, the protocol needs to be simple, fast, and cost-efficient to reach widespread use. In paper I, we described SiMSen-Seq, a method for multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by sequencing. Then, we studied technical aspects of ultrasensitive sequencing focusing on DNA polymerases (paper II) and synthetic oligonucleotides (paper III). In paper IV, we developed a generic platform for processing and analysing ultrasensitive sequencing data. Ultimately, the successful real-world application depends on preanalytical factors, which are discussed in paper V. Together these studies provide a framework for developing molecular diagnostics applications using ultrasensitive sequencing from sample collection to data analysis. In papers VI - VIII we used SiMSen-Seq to address questions in melanoma research that required the use of ultrasensitive techniques.

3.1 SiMSen-Seq

Error correction with unique molecular identifiers (UMI) enables ultrasensitive sequencing. UMI approaches have been described previously with methods using either ligation or PCR to incorporate UMI into sequencing libraries. We chose a PCR-based barcoding strategy for SiMSen-Seq, as PCR requires less DNA input than ligation, is simpler as it does not require target capture and subsequent purification steps and can be customized relatively easily through multiplexing. SiMSen-Seq was inspired by the development of the SafeSeqS protocol, which also employs barcoding using UMI embedded in PCR primers[70]. However, the original version of the SafeSeqS protocol used only single amplicons, required multiple rounds of PCR, primer digestion and clean-up, as well as polyacrylamide gel and

bead purification steps. At the time of SiMSen-Seq's development all published approaches for ultrasensitive UMI sequencing used long, multistep workflows that were challenging and costly to implement with increased risk of losing rare variant alleles during multiple purification steps, ligation, or target capture. SiMSen-Seq improves upon earlier barcoding protocols by introducing two key innovations:

A main issue with PCR-based barcoding is the excessive formation of unspecific products that outcompete specific products due to the many complementary sequences generated by random UMI. If several assays are multiplexed together this gets further exacerbated, making such reactions infeasible. In contrast to earlier barcoding approaches, SiMSen-Seq primers contain a temperature-dependent stem-loop that closes during PCR annealing temperatures to protect the UMI from primer-to-primer binding, thereby reducing non-specific product formation drastically (Figure 3.1).

Secondly, we made use of optimized PCR conditions developed for single-cell PCR[263] by combining low primer concentrations with extended annealing times, reduced enzyme concentration, PCR additives and extended PCR extension time to improve specificity and amplification uniformity for multiplexed assays.

3.1.1 SiMSen-Seq library construction

SiMSen-Seq libraries are constructed in a two-step workflow where UMI are incorporated into template molecules during an initial barcoding PCR, the product of which is then amplified in a second PCR with Illumina adapter primers. The final PCR products are subsequently purified, inspected using capillary gel-electrophoresis and sequenced (Figure 3.1).

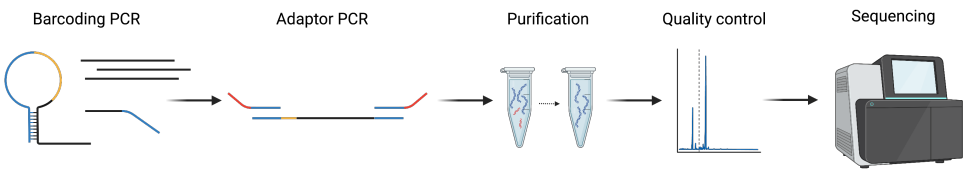

Figure 3.1: SiMSen-Seq protocol. Barcoding PCR primers are protected by a temperature-dependent stem-loop, enabling PCR using random UMI. Products from the barcoding reaction are used as templates in a second PCR to incorporate Illumina adapter sequences. Final libraries are purified using magnetic beads and sequenced after passing quality control. Created with BioRender.com based on paper I[73].

The barcoding PCR is performed for three cycles, resulting in six unique barcodes per original template molecule (For details see: Supplementary Figure 5 in paper I). The primer concentration is only 40 nM, approximately 10 times less than in ordinary PCR protocols, requiring an extended annealing time of 6 minutes for efficient product formation, compared to 20 – 30 seconds in standard PCR. The hairpin loop is closed at the annealing temperature of 62 °C preventing further formation of unspecific products. We successfully used various high fidelity DNA polymerases for the barcoding step, including Accuprime and Phusion in paper I and Platinum in subsequent studies[115,264], although the optimal choice of enzyme may require considering other factors besides fidelity, as

discussed in paper II. The quality of the primers is critical for SiMSen-Seq performance, especially that of the forward primer during the barcoding PCR, due to its length and random components. Although we recommended the use of oligonucleotide purification in the original protocol[73], this may not always be enough as we have shown in paper III[265].

After the completion of barcoding the reaction is kept at higher temperatures and Tris-EDTA buffer and a heat sensitive protease is added to completely halt the reaction before cooling to prevent spurious amplification during the cooling process. One third of the products from this reaction, *i.e.*, two barcodes per original molecule, are transferred to a second PCR reaction and amplified using P5 and P7 adapter primers containing an index sequence for sample deconvolution after sequencing. The adapter PCR is performed at higher temperatures with Q5 polymerase, so that the stem-loop remains open and does not interfere with amplification. Libraries are then purified using magnetic beads and analysed using capillary gel-electrophoresis to ensure purity and correct sizing. Purified libraries can be sequenced on any Illumina instrument.

3.1.2 Ultrasensitive mutation detection

Raw sequencing data is processed bioinformatically by aligning reads to the reference genome, grouping reads with the same UMI into barcode families and forming an error-corrected consensus sequence. We originally used the software debarcer, developed together with the first version of the experimental protocol[72,73] although we subsequently developed a refined version called UMIErrorCorrect, which is described in paper IV. Error correction through consensus read formation is very efficient at eliminating sequencing errors already at a minimum consensus family size of 3 (consensus 3). That is consensus reads are formed from all barcode families where the number of reads in the family is greater or equal to the minimum family size. Increasing the minimum family size will further improve error correction but comes at the cost of losing molecules that have not been amplified sufficiently. Typically, minimum family sizes between 3 and 10 will result in good error correction while maintaining adequate coverage. Even though most errors can be corrected in such manner, errors that were introduced during the first cycle of barcoding cannot be differentiated from true mutations and some residual background noise remains (Figure 3.2).

Generally, errors are reduced about 7 times on average, but for some positions errors are reduced over 1000 times. However, the degree of potential error correction depends on sequencing depth, consensus threshold and overall levels of background noise[72,73]. At very low allele frequencies, when only few mutant reads will be present, the allele frequency of consensus data may be slightly higher for a few positions than in raw reads due to sampling error, as can be seen in Figure 3.2. However, at these low allele frequencies such small differences are not practically meaningful.

Furthermore, error rates are not constant across the genome but can also depend on nucleotide context as will be discussed in papers VI and VII. Several methods for modelling background errors using statistics are discussed in paper IV and although all of them have advantages and disadvantages, UMIErrorCorrect enabled us to detect rare variant alleles in uveal melanoma patients (Figure 3.3; details in paper VIII). As can be seen in Figure

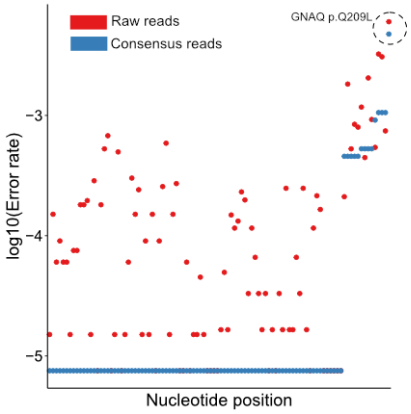

Figure 3.2: SiMSen-Seq error correction. Error rates for raw and consensus reads for all nucleotides in amplicons generated from a uveal melanoma patient. Each dot denotes the error for a position in one of the amplicons. Positions with no error were set to half of the lowest detected variant allele frequency. The patient specific *GNAQ* mutation is highlighted.

3.3, many positions contain background noise. However, here this corresponds to only 1 - 3 reads, while the expected variant was called based on 9 out of 3800 analysed molecules (0.24%). Thus, SiMSen-Seq provides a simple and flexible protocol for ultrasensitive sequencing applications in next-generation diagnostics.

Future development of SiMSen-Seq may include alternative UMI structures and dual barcoding,*i.e.* attaching barcodes on both strands. Improving the UMI setup may reduce failure rates of barcoding primers due to synthesis errors and limit the formation of unspecific products. Dual indexing can be used to increase multiplexing capacity and paired-end sequencing could further increase error correction by leveraging the fact that both reads in a mate-pair are derived from the same molecule. *In silico* modelling of primer interactions could be used to optimize the assay design workflow and accelerate the development of larger panels.

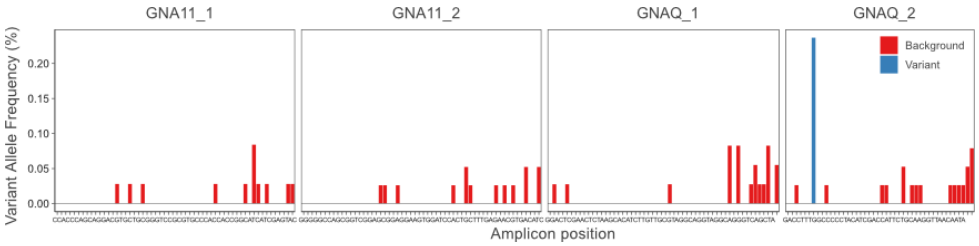

Figure 3.3: SiMSen-Seq variant detection. Data from four amplicons used to sequence cell-free DNA of a uveal melanoma patient. With data from paper VIII.

3.2 A bioinformatic analysis pipeline for barcoded sequencing data

There is a lack of open-source UMI-variant callers that are easy to use, do not require dedicated computing resources and can handle multiple barcoding strategies and enrichment chemistries, while also providing a statistical background error model appropriate for both PCR and ligation-based barcoding approaches. Therefore, in paper IV, we developed UMIErrorCorrect, a generic pipeline for analysing amplicon sequencing data with UMI and UMIAnalyzer, an R-package for data analysis and visualization.

3.2.1 UMIErrorCorrect

UMIErrorCorrect is a pipeline for UMI clustering and variant calling using raw FASTQ or pre-processed alignment files as inputs. An overview of the pipeline is shown in Figure 3.4. It can be run as a single command *run_umierrorcorrect.py*, or alternatively each step may be performed separately. The pipeline supports both single and paired-end data as well as single and dual barcoding strategies. In the first step raw sequencing data in FASTQ format is pre-processed by extracting the UMI and spacer sequences before aligning reads to any indexed reference genome using bwa-mem[54].

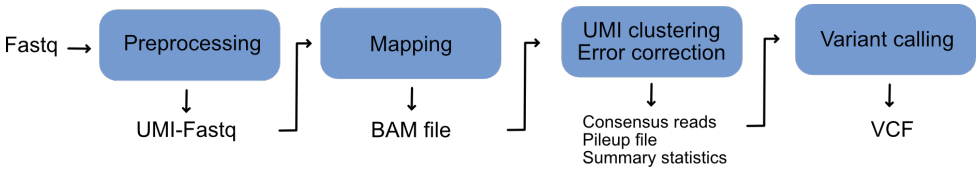

Figure 3.4: UMIErrorCorrect pipeline. UMIErrorCorrect performs reads processing and alignment, followed by UMI clustering and consensus read generation. Output files can be further analysed using the R-package UMIAnalyzer or Shiny app UMIVisualizer.

In the next step UMI sequences are clustered into barcode families, the number of which will also be an estimate of the number of original molecules. However, errors in the UMI can cause barcode families to split or merge and need to be accounted for when clustering UMIs. Different strategies for modelling UMI errors have been proposed and each may introduce a bias that leads to over- or underestimation of the true number of UMI, as described in UMI-tools[266]. UMIErrorCorrect removes UMI errors using the 'adjacency' network method from UMI-tools. First, it builds a network of all UMI that are within an edit distance of one and then selects the most populous node (the largest barcode family) as a lead node in the network. All adjacent nodes are grouped together with that lead node and removed from the network. If all nodes in the network are accounted for at this point, the sequence of the lead node is considered the true UMI sequence, and all other nodes are merged with it (Figure 3.5). However, this will not account for all nodes in a complex network where two different barcodes may collide, and each be within an edit distance of one of a third barcode but not of each other, e.g. AAAT ↔ ACAT ↔ ACGT. In such a case the two largest nodes are set as lead nodes and all their immediate neighbours, excluding other lead nodes, are grouped together. This is repeated until all

nodes in the network are accounted for. The edit distance threshold can be set to 0 to disallow any errors in the UMI or set to a number >1 to allow more sequencing errors in the UMI.

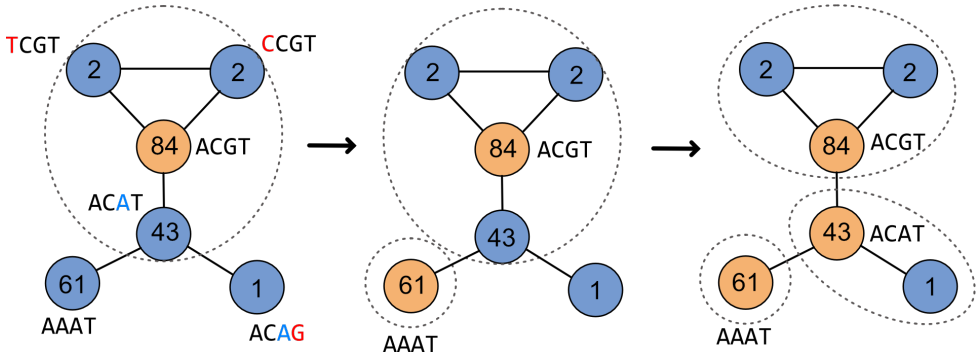

Figure 3.5: UMI clustering for barcodes with an edit distance of one. Lead nodes are highlighted in orange and numbers indicate the size of each barcode family. Assuming that all distinct UMI are derived from different molecules would result in six barcode families, a likely overestimate. Conversely, clustering all UMI that are within an edit distance of one together (left network) would result in only a single family based on the most populous node (ACGT). Adjacency clustering finds the most likely subgroups within this network based on node size and yields a more representative estimate of three barcode families (right network). Adapted from UMI-tools[266].

Then a consensus sequence is created for every position in each barcode family. For insertions and deletions at least 60% of reads need to have the insertion or deletion. For mismatches the base quality score is used to calculate the allele with the highest probability under the prior assumption that each base is equally likely. Consensus reads are saved as a binary alignment map (BAM) file and a pileup is performed, summarizing the allele counts for each base at every position at consensus cut-off of 0, 1, 2, 3, 4, 5, 7, 10, 20 and 30, where consensus 0 refers to all mapped reads and consensus 1 refers to all collapsed UMI sequences including singletons.

3.2.2 Variant calling using a statistical background error model

Then UMIErrorCorrect uses a variant calling algorithm based on a site-specific beta-binomial background error model like smCounter2, using the variant allele counts for each position. We estimated the hyper parameters of the Beta distribution α and β by fitting a typical, publicly available, SiMSen-Seq data set that was generated using 80 ng genomic DNA, excluding all known variants. For each position with k variant reads we then calculate the fraction P of j beta-binomial simulations in which the background error x is greater than the observation k at the sequencing depth n and error rate p_i using

$$P = P_{Betabin}(x \geq k|n, \alpha, \beta) = j^{-1} \cdot \sum_{i=1}^{j} P_{Bin}(x \geq k|n, p_i),$$

where each p_i is drawn from a Beta distribution with hyper parameters α and β. That is, we estimate the probability of having $\geq k$ false positive reads under the null hypothesis of the position being wildtype. This P-value is then converted to a Q-score using

$Q = -10 \cdot log_{10}(P)$, with a default cut-off of Q20, which corresponds to a false-positive rate of 0.01. Thus, we first estimate the background error rate from an external data set (Figure 3.6) and then use the resulting Beta distribution to generate background errors for each position. We then calculate the probability of the observed error rate arising by chance. If this is false positive probability is lower than a predefined threshold, a true somatic mutation is called for that position (Figure 3.6).

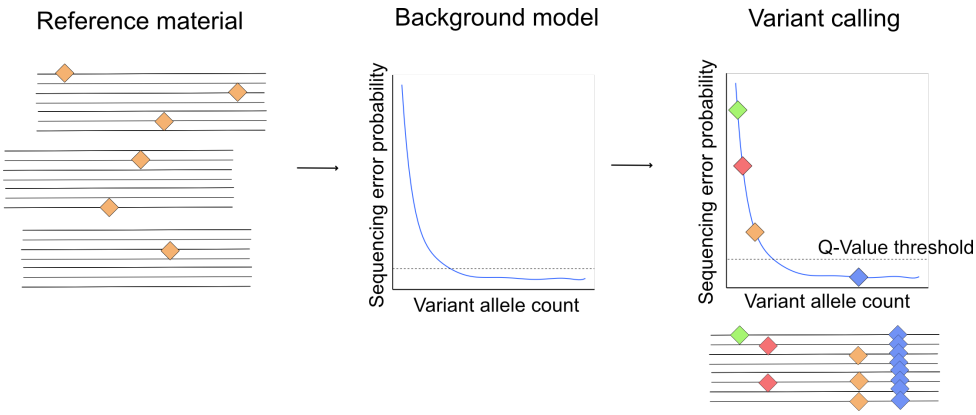

Figure 3.6: UMIErrorCorrect variant caller. A variant calling algorithm using beta binomial background model with significant variants marked in blue.

Unlike MAGERI and smCounter2, UMIErrorCorrect does not try to model each type of substitution error individually but uses the maximum non-reference count instead. Indeed, despite using 300 ng of reference material, it was not possible for smCounter2 to model most types of transitions and any transversion individually[94], but instead conservatively assumed that all mutations follow the most common error distributions, G → A for transitions and C → T for transversions. The type of nucleotide changes observed also depends on enrichment chemistry and choice of DNA polymerase as discussed in paper II. CAPP-Seq, which uses hybridization capture, exhibits predominantly A → C and G → T errors, which are less common in PCR-based studies[78,94]. Thus, modelling of individual substitution types is difficult for rare types of substitutions and prone to biases from experimental design. However, UMIErrorCorrect's model may also overestimate the error for rare types of nucleotide changes, since the major non reference allele error is likely biased towards the most common types of substitutions. Therefore, it remains challenging to build a universal background error model and it may be better to generate a new set of parameters for each type of experiment to account for differences in enrichment method, DNA polymerases and mean coverage.

To validate UMIErrorCorrect and UMIAnalyzer we used a 5-plex SiMSen-Seq panel to analyse Seracare cell-free DNA reference material at four different variant allele frequencies (VAF): 1%, 0.25%, 0.125% and 0% (wildtype). We could detect all known mutations and their expected VAF and observed a mean background noise of 0.013% across all nucleotides, excluding known variants. The highest individual nucleotide error was 0.244%. Compared to the background noise from the non-variant positions all variants at 1% VAF and 0.25% VAF samples were significant. For the 0.125% samples the PIK3CA_b and the

TP53_a assays were significant, but the KIT and TP53_b assays were non-significantly above background (p-values 0.106 and 0.061). For consensus 3 reads UMIErrorCorrect's beta-binomial variant caller had sensitivities of 93.3%, 100% and 100% to detect 0.125% VAF, 0.25% VAF and 1% VAF respectively at Q $\geq$ 10, with > 98% specificity. At Q $\geq$ 15 the sensitivity for 0.125% VAF decreased to 46.7%, while the sensitivity at 0.25% and 1% VAF remained at 100% with a specificity of 100% for all VAF, comparing favorably with smCounter2's 92% sensitivity for SNVs at 0.5% VAF[94] and UMI-VarCal which did not detect any mutations below 0.3% VAF[95].

Using a consensus threshold for detecting known mutations UMIErrorCorrect, just like DeepSNVMiner and debarcer, detects known variant alleles at $\leq$ 0.1% VAF. However, heuristic approaches were noted to have a much higher false positive rate than other variant callers[94]. Thus, ideally, background error rates should be considered by putting variant calling into a statistical framework, even though this is likely to provide a more conservative estimate of true mutations. For clinical applications this might also be more appropriate to minimize the chance of affecting the treatment of patients due to a positive variant call. Currently, UMIErrorCorrect allows the user to choose between using heuristic cut-off or a beta binomial variant caller. The latter currently uses a simple estimation of background error rates using major alternate allele frequencies and might be expanded to nucleotide-change-specific error rates using enough reference material to build a universal error model. However, for many applications it might be more sensible to create an error model optimized for each experimental setup, including sample type, enrichment chemistry and average UMI depth. This could also be used to account for run-to-run variation and data generated using different sequencing platforms.

Lastly, we used UMIErrorCorrect to analyse publicly available data generated using the Roche Avenio, QIAseq and Archer panels, demonstrating the utility of UMIErrorCorrect as a generic variant calling platform for targeted sequencing data using UMI. UMIErrorCorrect and UMIAnalyzer provide the toolbox to process, analyze, and visualize practically any type of UMI setup, requiring only raw FASTQ files as inputs.

3.3 Impact of polymerase fidelity on background error rates in massively parallel sequencing

Even though high-fidelity DNA polymerases are widely used for the construction of ultrasensitive sequencing libraries, their impact on various steps of library preparation have hitherto remained unexplored. In paper II we evaluated the impact of different DNA polymerases on sequencing data generated with SiMSen-Seq. Our results provide insights into the evaluation of DNA polymerases in ultrasensitive sequencing beyond the consideration of fidelity alone.

We compared DNA polymerases with reported fidelities between 1x to >100x relative to *Taq* polymerase in the SiMSen-Seq barcoding PCR and found that polymerase fidelity had no significant impact on raw read error rates. However, UMI corrected consensus sequences showed the expected decline in errors with increased polymerase fidelity, although the improvement in error correction did not scale linearly with estimated fidelity, possibly due

short length of the sequences analyzed. Since raw errors were unaffected by fidelity in the barcoding step, this implies that most errors are incurred later in the library construction or during sequencing, which can be corrected using UMI. We found that error rates at the same position were correlated between enzymes, suggesting that sequence context influences polymerase error. GC-rich regions are known to be more difficult to amplify, although none of the target sequences in this study had a particularly high GC content.

We additionally identified technical artefacts in the BRAF amplicon, which had been reported previously in other published data sets. Since these errors were not corrected, this indicates that they could have been caused by physical base modifications or real biological variation.

We also compared the impact of standard *Taq* polymerase versus a high-fidelity polymerase in the adapter PCR. The choice of polymerase had no significant influence on raw errors (before UMI error correction), implying that most errors are incurred later, *i.e.* during sequencing. However, we found that for this comparison consensus errors were seemingly smaller for the lower fidelity *Taq* polymerase compared to the high-fidelity enzyme. Ideally, consensus reads should only contain errors introduced during the first cycle of barcoding, since errors that occur later on, including the adapter PCR, are corrected by using UMI information. Thus, the fidelity of the adaptor PCR should not actually affect the consensus error, while the first cycle barcoding PCR error (*i.e.* the consensus error) should be similar for both enzymes, since the same polymerase was used for barcoding in both cases. And indeed, we found that this conundrum was caused by amplification bias of the regular *Taq* polymerase, resulting in 50% fewer barcode families compared to the high-fidelity enzyme and therefore artificially reduced error rates due to rare allele variants being under-represented in the final library.

Our data showed that increased fidelity during barcoding does reduce the amount of background noise, although it may be less than expected for small amplicon panels or if barcoding was performed for more than 3 cycles. We further showed that most errors are incurred during sequencing and that these can be corrected using UMI-based consensus formation. Given the differences in cost between high and low fidelity polymerases as well as other polymerase properties such as multiplexing ability, PCR efficiency, amplification bias, sensitivity to inhibitors and ability to amplify GC-rich areas, we suggest that fidelity alone should not necessarily be the main criterion for choosing an enzyme in ultrasensitive sequencing applications.

Although most errors that are incurred during sequencing can be corrected using UMI, the error rate of the sequencing technique and average UMI family size directly impact the amount of background noise. This means that higher error rates during sequencing, due to methodology or run-to-run variation, will increase the amount of small barcode families where all reads contain the same error and will therefore result in a false positive. Increasing the minimum family size required for greater error correction (consensus depth) can reduce the background noise but comes at the cost of reducing the number of molecules analysed. This trade-off may occur especially in cases where family sizes are kept small on purpose to reduce sequencing cost.

3.4 Digital quantification of oligonucleotide synthesis errors

Oligonucleotides are a critical and increasingly sophisticated component of many molecular techniques. Despite their widespread use, errors incurred during oligonucleotide synthesis and their impact on experimental outcomes have not been studied in depth. In paper III we used SiMSen-Seq to characterize and compare the error profiles in synthetic oligonucleotides across multiple manufacturers, templates, batches, and purification methods. We found that all types of oligonucleotides contained errors, with deletion errors being the most common, and some samples containing over 12% truncated molecules. Although we expected to see deletion errors, as they are known artefacts of chemical oligonucleotide synthesis, we also found substitution errors in all types of oligonucleotides significantly above the levels found in biologically produced DNA.

We further showed that for one sequence configuration errors increased in the direction of synthesis (Figure 3.7), but we did not observe this directional bias for another template where the sequence was inverted. This implies that sequence context influences the error profiles of chemical oligonucleotide synthesis, perhaps because of secondary structure formation of the nascent oligonucleotide chain. We also found that some positions have aberrantly high error rates specific to individual batches, which suggests that failures during individual synthesis cycles can drastically increase error rates at specific positions. We also observed that overall error rates across amplicons differ between batches and that these batch-to-batch differences were in some cases greater than the effect of purification.

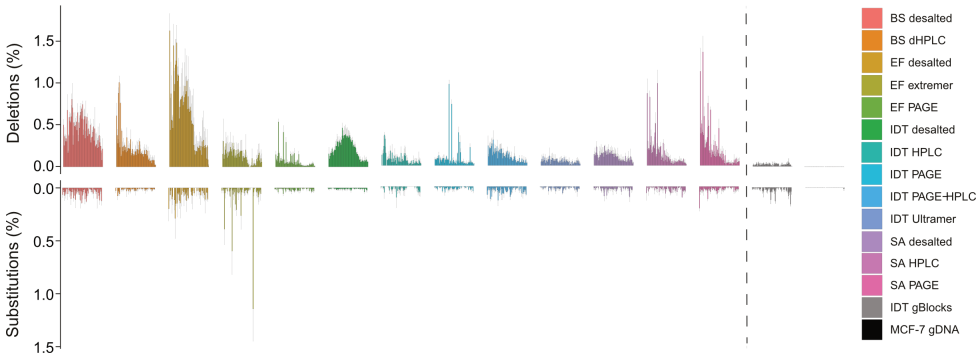

Figure 3.7: Errors in synthetic oligonucleotides. Errors are shown for each combination of oligonucleotide manufacturer and purity grade. Genomic DNA was obtained from the breast cancer cell line MCF-7. BS: BioSearch, EF: Eurofins, IDT: Integrated DNA Technologies, SA: Sigma-Aldrich. Adapted from paper III[265].

3.4.1 Oligonucleotide quality is critical for molecular techniques

These findings have potentially far-reaching implications for all techniques that use oligonucleotides in ultrasensitive applications. While errors levels are low enough not to impact the performance of standard PCR applications, they can severely impact more sensitive techniques such as SiMSen-Seq. We showed that the quality of oligonucleotides used

during the barcoding PCR affects the quality of SiMSen-Seq library preparation, possibly due to errors interfering with the stem-loop structure.

Oligonucleotide errors, especially when they occur at very high frequencies, may also produce off-target products, reduce on-target binding and lead to amplicon switching in multiplex PCR (Figure 3.8). The latter may be particularly severe in case where degenerate primers are used, for example in metagenomic studies[267]. Oligonucleotides have also been used for cloning applications and oligonucleotides specifically developed for these, such as Ultramers and gBlocks from IDT, had indeed among the lowest error rates of all studied oligonucleotides. However, all of them still carried some errors and gBlocks had in fact higher levels of substitution errors than some "regular" oligonucleotides. Even at low frequencies these might negatively impact the result of cloning studies by artificially introducing unwanted genetic variants, especially if rare subpopulations are studied.

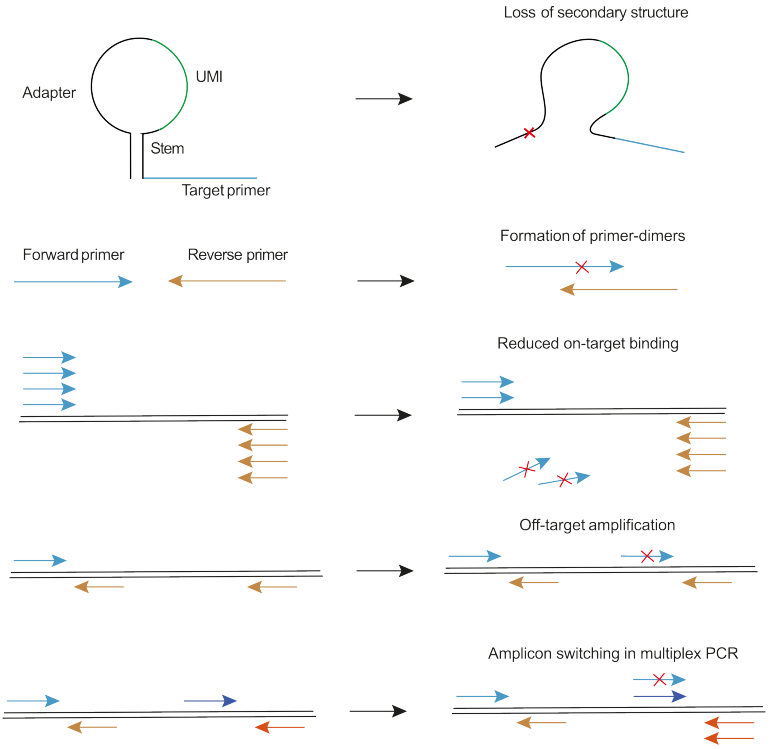

Figure 3.8: Consequences of oligonucleotide synthesis errors. Oligonucleotide errors are marked with red crosses. Adapted from paper III[265].

3.5 Preanalytical considerations in liquid biopsy analysis

Analysing cfDNA is challenging due to the limited amount of cfDNA in a typical sample, highly fragmented source material and the low fraction of ctDNA present in early stage

or minimal residual disease. Ultrasensitive sequencing techniques, such as SiMSen-Seq, have been developed to allow rare variant allele detection, but they remain limited by the small amounts of sample available. Therefore, optimized sample handling and extraction workflows are required to maximize ctDNA yield and increase analytical sensitivity, regardless of the method used. In paper V we developed a framework for consideration of preanalytical factors in liquid biopsy studies to improve their sensitivity and robustness.

For few molecules, the probability P of observing k mutant molecules given an average concentration of λ mutant molecules per reaction can be modelled using the Poisson distribution[75]

$$P(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Thus, even if a sample contains (on average) a single molecule ($\lambda = 1$), there is a 37% chance that no molecule will be detected ($k = 0$). Increasing the sample volume by 3 times reduces the chance of a false-negative result to 5%. Alternatively, analysing multiple independent targets has the same effect as increasing the amount of sample. For example, the ability to detect at least one ctDNA molecule in a sample where at least one ctDNA molecule is present on average increases from 63% for a single assay to 95% for three and to 99.3% for five independent assays.

Although total cfDNA levels may be increased for patients with advanced disease or due to other factors, for many applications such as screening, relapse detection or minimal residual disease detection the amount of cfDNA will be comparable to that of healthy people. Reported levels for cfDNA in putatively healthy people vary widely with a median of 1640 copies[268] per mL plasma (or 5.43 ng/mL) and a range of 100 – 4000 copies/mL. Therefore, the median amount of cfDNA that can be obtained from a typical blood draw of 5 mL plasma will be 8200 haploid genome equivalents. That means that the median, theoretical limit of detection for an individual mutant molecule will be 0.012% variant allele frequency for a single assay. In practice, losses due to sampling, DNA extraction, sample concentration, DNA fragmentation (for PCR-based assays) or ligation efficiency will reduce the number of available molecules. We showed that only 49% of cfDNA was amplifiable by first quantifying the amount of cfDNA fluorometrically and then measuring the actually amplified material with qPCR. Thus, qPCR-based cfDNA quantification more accurately reflects the available material for analysis.

Furthermore, we showed experimentally that amplicon length correlates with yield for randomly fragmented DNA but less so for cfDNA. Since cfDNA fragmentation is not the same for all regions in the genome due to nucleosome positioning[269], it may be optimal to analyse less fragmented regions where possible. However, since most oncogenes are actively expressed, and therefore lie in more fragmented open chromatin, this may not always be feasible. DNA fragmentation also differs between analytes. For example, urinary cfDNA is more fragmented than cfDNA obtained from blood samples and is highly dependent on the time the sample was obtained[270].

Improper storage or sample handling may result in the release of contaminating cellular DNA into the plasma, potentially confounding later analysis. We developed a qPCR-based detection assay to measure the ratio of short and long DNA fragments, where the presence

of long DNA fragments indicates cellular DNA contamination.

Another difficulty is the presence of inhibitors. Many clinical samples contain potent PCR inhibitors such as haemoglobin in haemolysed blood samples[271] or melanin in melanoma tissue[272]. Because samples often need to be concentrated prior to analysis, this may not only lead to losses due to technical inefficiencies but simultaneously increase the concentration of inhibitors. We showed that is possible to quantify PCR inhibition using a simple qPCR assay that can also be used to quantify the amount of available cfDNA in the same reaction. PCR additives may also be used to facilitate library preparation and counteract sample inhibition[272]. Other factors that influence cfDNA detection are the interplay of blood collection tubes and cfDNA extraction method and we showed in paper V how changing extraction method reduced sample inhibition.

3.5.1 Challenges in liquid biopsy analysis

Any ultrasensitive molecular diagnostic platform needs to address numerous preanalytical and analytical challenges and requires workflows that are carefully optimized at each step to achieve the necessary sensitivity and specificity (Figure 3.9). Optimal sample collection and handling will depend on the availability of infrastructure for processing and biobanking, as EDTA tubes are preferable but may not be feasible because samples cannot be processed quickly enough. CfDNA extraction methods need to be evaluated based on the yield, presence of inhibitors and ability to scale and automate. Analytical techniques need to be sensitive enough, but simultaneously require robust, easy, fast, and cheap protocols to be considered for routine clinical applications. Especially sequencing-based protocols require optimized bioinformatics tools so that non-specialists can quickly analyse and interpret data for clinical decision making.

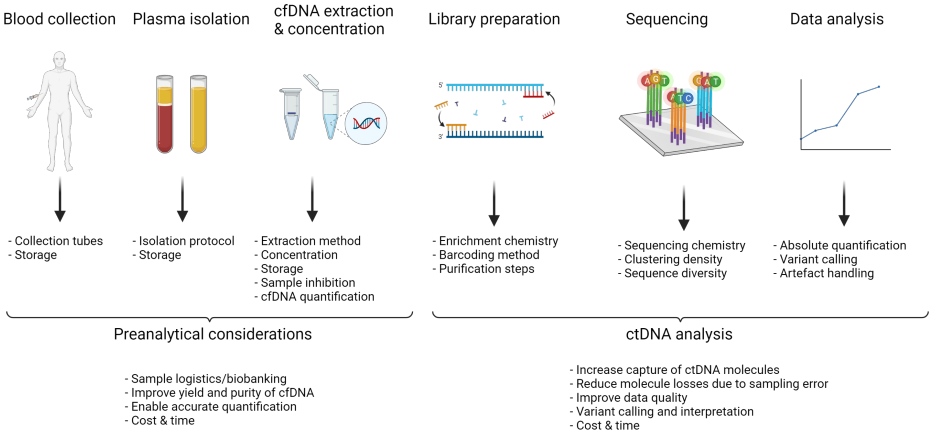

Figure 3.9: Liquid biopsy workflow. At each step several factors affect the yield and quality of ctDNA analysis and should be considered by users and developers. Preanalytical challenges revolve around increasing yield and purity of cfDNA, analytical challenges focus on increasing assay sensitivity and specificity. Created with BioRender.com.

3.6 Recurrent promoter mutations – a novel melanoma biomarker

Identifying disease-specific biomarkers, especially when they are functional and clinically relevant, is necessary for results from ultrasensitive molecular tools to have real-world impact. High-throughput sequencing techniques have allowed the generation of large amounts of genomic data, enabling the identification of recurrent driver mutations in many cancer types, although these can also be accompanied by non-functional passenger mutations. Functional somatic regulatory mutations have been described in the *TERT* promoter and whole genome sequencing studies of melanomas have identified recurrent promoter mutations in other genes, although their impact on gene function remained unclear. In paper VI we showed that frequently recurring promoter mutations in melanoma occur due to sequence context-dependent sensitivity to UV radiation without positive selection. We then tried to elucidate the origin of these mutations in paper VII and showed that they are caused by elevated pyrimidine dimer formation.

3.6.1 Sequence context-dependent mutation rates

In paper VI we mapped recurrent promoter mutations in publicly available whole genome sequencing data from 38 melanomas. We found that 27/32 (84%) recur-rent promoter mutations, that were present in at least 4/38 tumours, occurred within the context of the ETS transcription factor binding motif TTCCG[273]. These mutations were predominantly C→T or CC→TT transitions, which are compatible with UV-damage-induced cylcobutane pyrimidine dimer (CPD) formation. The most frequently recurring mutation was observed in 11/38 (29%) tumours in the transcription start site of the gene *RPL13A* and we also identified the previously described *TERT* promoter mutations. We then confirmed these mutations in an independent melanoma data set[274] as well as data from squamous cell carcinomas and sun-exposed skin[42,275]. Strikingly, we could not detect these mutations, except the *TERT* promoter mutations, in 13 non-UV-exposed cancer types. Furthermore, RNA-seq data showed that these recurrent promoter mutations did not alter the expression of nearby genes, arguing against a functional role for these mutations, although this analysis may be limited by the small cohort analysed.

We therefore proposed a model where the observed hotspot mutations are caused by elevated mutation rates due to sequence-context induced sensitivity to UV radiation. To confirm our model experimentally, we exposed melanoma cells and keratinocytes to daily UV treatment for 5 or 10 weeks and then used SiMSen-Seq to study promoter mutations in *RPL13A* and *DPH3*. We observed elevated levels of the expected mutations, consistent with UV exposure, specifically at the TTCCG motif. We did not observe them in the control cells and at much lower levels at other potentially vulnerable cytosines within the studies amplicons. Despite the daily exposure to UV light for several weeks, the mutations remained sub-clonal, strongly arguing against positive selection (Figure 3.10). For the *DPH3* hotspot region we could also detect unusual C→A and C→G substitutions, which had also been observed in the tumor data.

Thus, we demonstrated that recurrent promoter mutations in melanoma are not caused

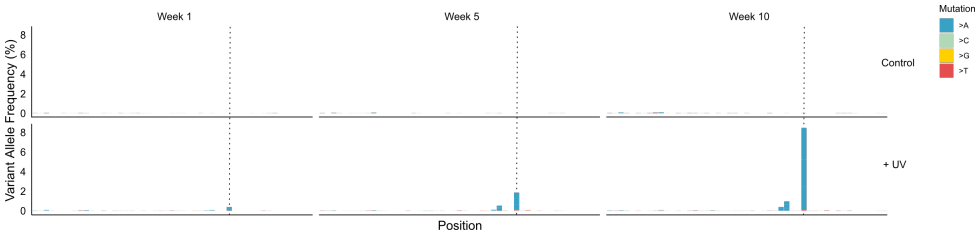

Figure 3.10: Mutations depend on sequence context. UV-induced hotspot mutations in the RPL13 promoter after 1, 5 and 10 weeks of UV exposure. Chromosome positions are shown on the x-axis and each bar represents a genomic position. The ETS-binding motif is highlighted with a dashed line. Only weeks 5 and 10 were shown in paper VI.

by positive selection but are the result of a sequence motif conferring in-creased sensitivity to UV radiation. These results have important implications for the interpretation of somatic mutations in cancer and highlight the role of sequence patterns and genomic context for the emergence and clinical relevance of somatic mutations in regulatory regions.

3.6.2 Mechanisms underlying non-coding mutations

The mechanism of these elevated mutation rates remained unclear, although locally impaired nucleotide excision repair (NER) caused by binding of ETS transcription factors had been proposed as a possible explanation[276]. However, our genomic data of squamous cell carcinomas from paper VI, which lack global NER, contradicts this hypothesis. In paper VII we aimed to elucidate this phenomenon that potentially explains a substantial proportion of non-coding recurrent variants in human malignancies.

We used a cohort of 221 publicly available melanoma genomes and again found that highly recurrent promoter mutations were primarily associated within the context of the ETS-binding motif TTCCG. As in paper VI we again observed a strong correlation of recurrent hotspot mutations with overall mutational burden, consistent with our model of passenger mutations. The fact that these mutations predominantly occurred near highly expressed genes indicates that ETS-binding, rather than properties of the sequence itself, are responsible for these mutations.

We again used the *RPL13A* hotspot region as model to study the role of NER for the emergence of promoter mutations experimentally. *RPL13A* was the most frequent non-coding mutation and almost as common at the BRAF V600E mutation. We therefore exposed cell lines with both intact NER and homozygous mutations in DNA repair components required for global NER, transcription coupled NER (TC-NER), and lesion verification in both global and TC-NER. Even small doses of UV caused extreme cell mortality and forced us to use a single two second dose of UV followed by three weeks of recovery. We subsequently performed ultrasensitive sequencing and despite the minuscule dose we again detected elevated mutation rates at the expected hotspot sites in cells with intact NER and all mutant cell lines (Figure 2 in paper VII). However, the overall mutation burden was $<0.5\%$ in all samples and close to the background noise in some samples.

These data, combined with our previous results from tumours lacking NER, imply that global NER is not responsible for these elevated mutation rates. Furthermore, TC-NER is

not active at transcription start sites where these mutations predominantly occur and can therefore also be ruled out as a mechanism for hotspot promoter mutations. Consequently, we investigated whether a purely biochemical mechanism based on elevated pyrimidine dimer formation could better explain their origin and generated a genome-wide map of CPDs in human melanoma cells immediately after UV exposure and before DNA repair processes is active. We observed a peak of CPD formation at the expected hotspots that was absent in naked DNA without bound proteins and non-UV-exposed control DNA. These mutations occurred predominantly in two cytosines directly next to the core ETS-binding motif CCTTCCG, where they do not interfere with DNA binding. Therefore, when the site is occupied by a transcription factor, CPDs are formed between the two pyrimidines inducing C→T mutations upon UV exposure, especially in the second base. Although observed mutations in the center of the motif should interfere with transcription factor binding it is known that many ETS factors are known oncogenes and ETS-binding might be enabled by *TERT* promoter mutations[277]. Another study by Mao *et al.*[278] also mapped CPD formation in ETS-binding sties which are fully congruent with our data and additionally suggested a mechanism for CPD formation in the ETS-DNA complex using structural data.

We thus demonstrated that highly recurrent mutations at ETS-binding sites arise by elevated pyrimidine dimer formation and are likely non-functional passenger mutations. This is supported by their correlation with overall mutational burden, little association with known cancer-associated genes and experimental evidence that suggests lack of positive selection. Despite the apparent lack of functional significance of these hotspot promoter mutations, their high recurrence between tumours, the large number of ETS sites in the genome and specificity to UV exposure, may still make them an interesting biomarker in melanoma and other cancers with a strong UV signature. Although they can occur in non-cancer tissue they may nonetheless be used as a complement to monitor known oncogenic mutations, such as BRAF V600E during the treatment of melanoma to improve the sensitivity of ctDNA detection by increasing the number of targets. More interestingly, they may be used to estimate the levels of UV damage in otherwise healthy people for risk-stratification and targeted skin cancer screening programs.

3.7 Next generation diagnostics in the clinical management of melanoma

Metastatic uveal melanoma is a cancer with high mortality and currently lacks established and effective treatments. Therefore, we conducted the multicenter, phase II clinical trial (PEMDAC) to test the effectiveness of combined HDAC inhibition with entinostat in combination with immune-checkpoint inhibition with pembrolizumab. In paper VIII we showed that a small subset of patients can profit from this new treatment regimen and that responders were predominantly *BAP1* wildtype patients. We also retrospectively evaluated ctDNA as biomarker for treatment monitoring and found that high levels of ctDNA are significantly associated with worse survival outcomes.

3.7.1 PEMDAC trial - rational & outcome

PD-1 inhibition has greatly improved the outcome of patients with metastatic cutaneous melanoma, although their efficacy in uveal melanoma has been disappointing[279,280]. Early phase trials have shown that entinostat and pembrolizumab may be effective in treating PD1-inhibition resistant cutaneous melanoma or lung cancer[281,282]. This provided the rational to evaluate the combined HDAC and PD-1 inhibition in the clinic to leverage the positive immune-stimulatory effects of both drugs in uveal melanoma patients.

Twenty-nine metastatic uveal melanoma patients were enrolled in the trial with a median age of 70 (range, 34 - 83). We observed a partial response in four patients, resulting in an overall response rate of 14%. This compares favourably with reports of anti-PD-1 monotherapy, whose reported response rate was <5%[279,280]. Eight patients (28%) showed clinical benefit and the median overall survival of 13.4 months was longer than the historical average[283], although this needs to be interpreted cautiously due to the relatively small patient cohort. One-year progression-free survival was 17% with a median of 2.1 months.

3.7.2 ctDNA as a biomarker for uveal melanoma

We developed a uveal melanoma specific hotspot mutation panel for SiMSen-Seq covering six hotspot driver mutations in *GNAQ*, *GNA11*, *CYSLTR1* and *PLCB4*. Practically all uveal melanomas are mutated in one of these hotspots[231], which guarantees that we will have at least one potentially existing mutation in each patient. However, since they are mutually exclusive[231], this fundamentally limits the sensitivity of our analysis as discussed above and in paper V. *BAP1* is a tumour suppressor and has no clear mutation hotspots, therefore it would be necessary to cover the whole gene or develop patient-specific assays for *BAP1* to be a meaningful addition to the panel. We instead designed six additional assays covering frequently recurrent mutations in *SF3B1*, increasing the size of the panel to a 12-plex, keeping the panel size and sequencing cost limited, while simultaneously increasing the chance of multiple mutations per patient.

Survival analysis

We obtained longitudinal plasma samples from before commencement of treatment and at each round of therapy for ctDNA analysis with SiMSen-Seq. In line with previous studies, we found that low baseline ctDNA relative to the median was predictive of overall survival (Figure 3.11a), but not progression-free survival, probably because of the short median progression-free survival in our cohort. Both high ctDNA and increased LDH are good predictors of overall survival (Figure 3.11).

Additionally, one can use the Cox proportional hazards model to evaluate differences in survival. The Cox model uses the hazard function $h(t)$ to describe the risk of death at time t,

$$h(t) = h_0(t) \cdot exp(\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_i \cdot x_i),$$

where the hazard depends on i variables $x_1, x_2, ..., x_i$ and their exponentiated coefficients $exp(\beta_i)$ are called hazard ratios. The term h_0 describes the baseline hazard, *i.e.* the hazard

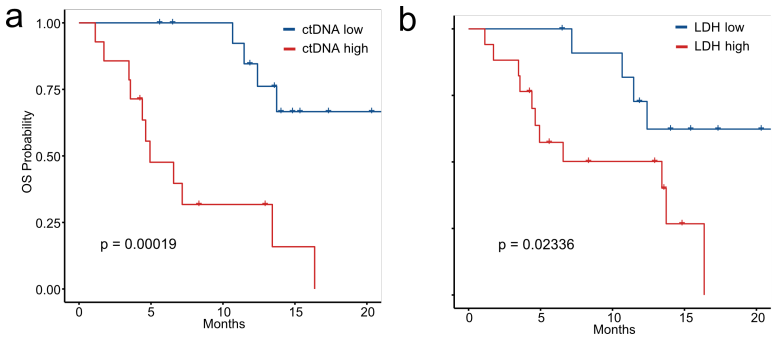

Figure 3.11: Biomarkers for uveal melanoma survival. Kaplan-Meier analysis of **(a)** ctDNA and **(b)** LDH at baseline. ctDNA is classified relative to the median and LDH relative to the upper level of normal. The statistical significance of differences between survival curves was assessed using log-rank tests without adjusting for multiple testing. OS: Overall survival. Adapted from paper VIII[115].

if all $x_i = 0$[284]. ctDNA seems to be slightly superior in both univariate and a multiple Cox regression model, which uses both LDH and ctDNA as covariates (Table 3.1). However, the loss of significance for LDH in the multiple regression model needs to be interpreted carefully due to the correlation between LDH and ctDNA (Pearson's $r = 0.75$, $p < 0.001$). In both models high baseline ctDNA conferred a more than 7 times increase in the risk of death compared to below median levels of ctDNA. This data is in agreement with previous results[260], suggesting that ctDNA detection is superior to LDH as a biomarker of survival in metastatic uveal melanoma. However, the small number of mutations per patient limit sensitivity of ctDNA and including other biomarkers such as the eight protein expression signature used in CancerSEEK[71] may improve detection sensitivity.

Table 3.1: Cox regression models. Results of univariate and multiple Cox regression for the PEMDAC data. β: Cox coefficient; HR: hazard ratio; CI: confidence interval.

	Univariate regression				Multiple regression			
Variable	β	**HR**	**95% CI**	**P-Value**	β	**HR**	**95% CI**	**P-Value**
ctDNA	1.98	7.23	(2.22 - 23.59)	0.001	2.03	7.62	(1.48 - 39.129)	0.015
LDH	1.23	3.57	(1.11 - 11.49)	0.033	-0.07	0.93	(0.19 - 4.61)	0.927

cfDNA and ctDNA between response groups

Total cell-free DNA levels at baseline were nominally lower in responders compared to patients with progressive or stable disease (Figure 3.12b). Baseline mutant reads per mL plasma were nominally, but not significantly, lower between partial responders and progressive disease (Figure 3.12c). However, the difference between median mutant reads per mL plasma during the treatment, excluding baseline data, was significant between responders and progressive disease (Figure 3.12d). Due to the small number of responders, especially the baseline comparison may be susceptible to outliers.

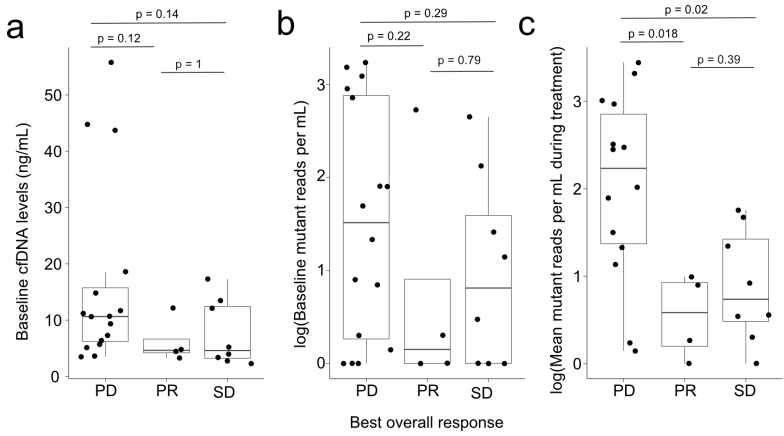

Figure 3.12: ctDNA and survival in uveal melanoma. (a) Cell-free DNA levels[a], **(b)** Baseline mutant reads per mL plasma **(c)** Mean mutant reads per mL plasma. Differences between response groups were assessed using Wilcoxon rank-sum tests and considered significant at $P < 0.05$. For baseline levels each dot corresponds to a single sample, for mutant reads during treatment each dot represents the mean for all available samples from each patient. PD: Progressive disease; PR: Partial response; SD: Stable disease. Adapted from paper VIII.[115]

Interestingly, even though the longitudinal patterns of ctDNA broadly correspond to treatment response, many patients had a complex pattern of ctDNA 'spikes' at some points (Figure 3.13). For instance, one patient with stable disease (3-010) and one responder (1-004) were ctDNA negative at most time points but had substantial levels of detectable ctDNA in one or two samples, respectively. We speculate that these patterns may be caused by proliferative waves within the tumour. Other studies have shown that immunotherapy sometimes induces a delayed response even after patients have already progressed[285]. Since cfDNA is cleared from the body within a day, changes in tumour activity can quickly lead to an increase or decrease of ctDNA. Tumour cells may enter a state of dormancy as a result of changes in the tumour microenviroment that create growth-antagonistic affects, such as hypoxia or immune activation[286]. The complex interplay between tumour environment, immune activity and drug treatment will affect cfDNA shedding and thus ctDNA detection sensitivity.

We found that all patients with progression of disease had detectable ctDNA at some time point and 75% (12/16) were ctDNA positive at baseline. Of the four patients with PD who were baseline negative, three had drastically increased levels of ctDNA during therapy (Figure 3.13). The remaining patient (3-012) had very low levels of ctDNA at one time point during treatment but no detectable ctDNA at any other time point. However, despite an early progression this patient had a longer than median overall survival of 17.3 months.

We also extracted DNA from pre-treatment FFPE tissue and identified the canonical *GNAQ* and *GNA11* driver mutations as well as mutations in the tumour suppressor *BAP1*. We could also identify mutations in the ctDNA of five patients where exome sequencing

[a]Erratum: In paper VIII Supplementary Figure 1d the cell-free DNA levels shown here in sub-figure 3.12a were inadvertently used instead of the mutant counts per mL plasma shown in sub-figure 3.12b. This did not affect the results, main text or conclusions drawn in the paper.

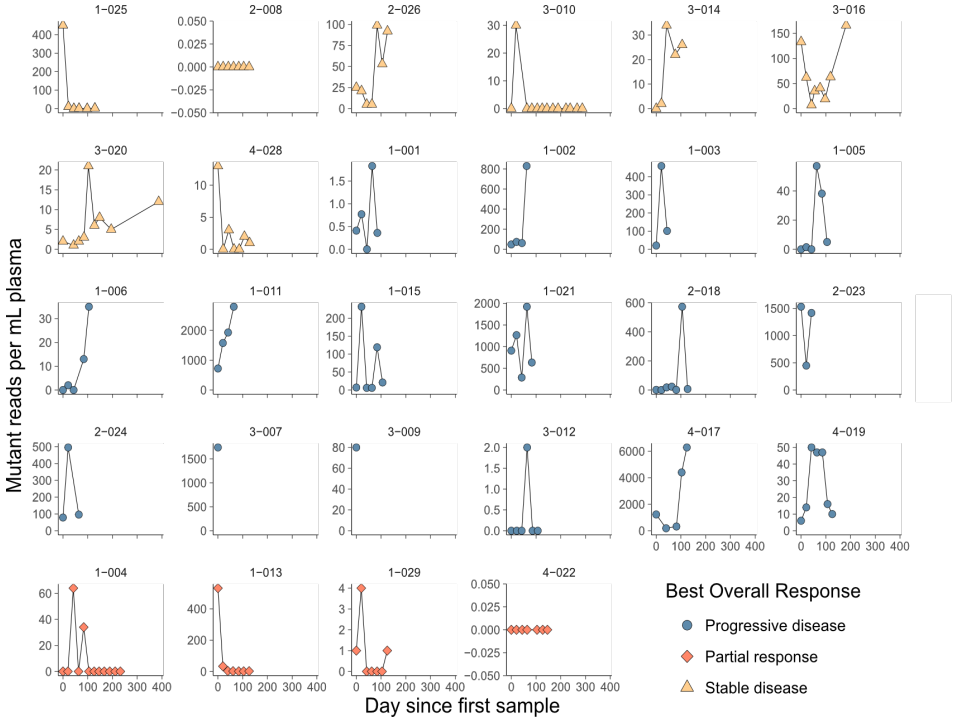

Figure 3.13: Longitudinal ctDNA measurements. Each data point represents the number of mutant reads from a single assay covering any of the mutually exclusive hotspots in *GNAQ* and *GNA11*. Adapted from paper VIII[115].

failed, highlighting the added value of ctDNA analysis. Three of the four responders had *BAP1* wildtype tumours and patients with wildtype *BAP1* had nominally longer survival and lower tumour burden based on ctDNA and LDH. Future studies may want to combine epigenetic and immunotherapy in a prospective randomized trial compared with immunotherapy alone. These should also investigate the role of *BAP1* mutational status and evaluate ctDNA for risk stratification.

4

Conclusions

Ultrasensitive sequencing is required to understand rare molecular events in numerous areas of research and diagnostics, including oncology, immunology, virology, and ageing research. We developed SiMSen-Seq, a method for ultrasensitive DNA sequencing with unique molecular identifiers. In papers I – V we established a generic ultrasensitive sequencing platform that can be applied to various applications, including liquid biopsies. We then used SiMSen-Seq in papers VI – VIII to reveal the sequence-dependent increase of promoter mutations after UV exposure and demonstrated the clinical utility of ctDNA as a biomarker for monitoring uveal melanoma treatment efficacy. Our specific conclusions are:

Paper I: We developed an ultrasensitive DNA sequencing protocol for simple, multiplexed, PCR-based barcoding of DNA for sensitive mutation detection using sequencing called SiMSen-Seq. It enables the detection of rare variants below 0.1% allele frequency and can be customized for larger multiplexed panels using specific design and optimization guidelines.

Paper II: We showed that high polymerase fidelity reduced error in barcoded sequencing, but that error correction using barcodes is orders of magnitude higher than the improvements gained by using higher fidelity polymerases. Therefore, polymerases with lower fidelity may provide adequate error correction while providing other beneficial characteristics such as higher yield, resistance to PCR inhibitors and improved PCR efficiency, depending on the application.

Paper III: We showed that choice of manufacturer, synthesis strategy, purity, batch, and sequence context significantly affect oligonucleotide qualities. We found that batch effects often outweigh increased oligonucleotide purity and could be a major contributor to poor oligonucleotide performance. We further showed that high-performance oligonucleotides are essential in numerous molecular applications, including clinical diagnostics.

Paper IV: We developed UMIErrorCorrect, a generic bioinformatics pipeline for the

analysis of barcoded sequencing data and a comprehensive toolset called UMIVisualizer, for the analysis, interpretation and visualization of data processed with UMIErrorCorrect.

Paper V: We established a workflow for cell-free DNA analysis and developed qPCR-based quality controls to evaluate each experimental step.

Paper VI: We showed bioinformatically that frequently recurring promoter mutations in melanoma occur almost exclusively at cytosines flanked by a distinct sequence signature and used SiMSen-Seq to experimentally confirm that UV light induces these recurrent mutations, highlighting the importance of sequence context for the interpretation of somatic variants in cancer.

Paper VII: We found that ETS-binding sites exhibit mutation hotspots with in-creased cyclobutane pyrimidine dimer formation rapidly after UV exposure and thus before DNA repair, contributing to recurrent mutations in melanoma. We also show experimentally, using SiMSen-Seq, that impaired nucleotide-excision repair is not the underlying mechanism for elevated mutation rates at these positions. Due to their high specificity for UV damage these mutations may eventually be used as biomarker for melanoma screening.

Paper VIII: We showed that combined epigenetic and immunotherapy can cause tumour regression in a small subset of patients diagnosed with metastatic uveal melanoma and that ctDNA can be used as a clinical biomarker for tumour burden and treatment efficacy in melanoma. Patients with high ctDNA levels showed worse overall survival and ctDNA at baseline was lower for responders than non-responders, demonstrating the clinical utility of ctDNA-based liquid biopsy analysis.

5

Future prospects

Simsen-seq performed favourably in an inter-laboratory comparison of ultrasensitive methods[287] and has been used for mutation detection in liquid biopsies from oesophageal cancer[288], skin[251,252] and uveal melanoma[115], colorectal cancer[289], and head and neck cancer[290]. However, the field of molecular diagnostics is rapidly evolving, and new technologies face many challenges to meet the requirements for real-world adaptation. More sensitive and abundant markers than mutations are needed for cancer screening and technologies need to become simpler and more cost-efficient. Adoption of liquid biopsies in the routine care will also depend on the outcome of dedicated, prospective clinical trials that provide evidence for increased clinical benefit through the use of liquid biopsies.

5.1 Emerging biomarkers and diagnostics

While the modern concept of liquid biopsies began with the study of circulating tumour cells and has since expanded to include ctDNA, miRNA and proteins, many of these biomarkers individually still have limited sensitivity that prevent their clinical use. Novel diagnostics methods like CancerSEEK therefore combine multiple biomarkers such as ctDNA and protein detection and use a machine-learning algorithm to integrate the data into a cancer detection score[71]. While the sensitivity for non-metastatic cancers overall was still relatively low (70%), CancerSEEK had >99% specificity[71], could predict tumour location and could double the number of breast cancer patients identified through screening in breast cancer[291]. Thus, multi-analyte or multi-omics approaches can potentially harness the strengths of different analytes resulting in a much more powerful integrated analysis. A drawback is the increased experimental and analytical complexity of multi-omics approaches, especially in a screening setting where potentially millions of tests need to be performed at low cost to have substantial public health benefits.

Epigenetic modifications are pervasive across the genome and there are hundreds to thousands of differentially methylated sites between tumour and normal tissues. Therefore, many recent studies have investigated the detection of methylated ctDNA, which has

potentially exquisite sensitivity due to the large number of available sites. This may be particularly important for cancer screening, where ctDNA will be low and since methylation pattern are tissue-specific this may allow tumour localization as well as early detection[175,292]171,284. Differentially methylated sites may also be useful complement to mutation detection during therapy, especially for tumour forms with low tumour mutational burden and tumour entities with few recurrent mutations. A recent study found that combined genome-wide detection of DNA fragmentation patterns, which are associated with nucleosome positioning[269,293], and ddPCR for the detection of a tumour-specific fusion gene was superior to either method alone[294] This is also supported by companies, such as GRAIL, moving towards epigenetic markers for their cancer screening products. However, individual methylated sites may carry little biological information, whereas many mutations have meaningful clinical impact, such as making the tumour sensitive or resistant to certain therapeutic interventions. Furthermore, at very low disease burden and early-stage disease it may be difficult to distinguish the tumour pattern, which may mostly resemble the tissue-of-origin, from other sources of increased methylated cell-free DNA[295].

Ultrasensitive sequencing can also be used in immunology applications, where instead of rare mutations, the focus in on quantifying rare populations of immune receptors[66]. A recent study combined immune receptor profiling, RNA sequencing and targeted ctDNA analysis in a multi-omics approach to provide mechanistic insights into rare lung cancer variants[296]. Thus, the future will likely be a multi-biomarker paradigm involving different combinations of markers to achieve the necessary sensitivity, specificity, and clinical utility for each specific application.

5.2 Clinical adaptation of liquid biopsies

Before a liquid biopsy assay can be used in the clinic it needs to meet three criteria: analytical validation, clinical validation, and clinical utility[114,297]. There is now an ample amount of retrospective data demonstrating both the analytical and clinical validity of liquid biopsy assays, showing the correlation of the ctDNA and survival outcomes and that ctDNA is detectable at very low levels long before recurrence is seen on imaging. However, there is limited data on whether liquid biopsy guided clinical decision making provides clinical utility, *i.e.* that it can actually improve patient outcomes[214,298]. This requires large prospective clinical trials with different arms comparing ctDNA guided interventions with conventional treatment protocols. Among others, such trials are currently under way in colon[299] and breast cancer[300], evaluating the benefit of additional adjuvant therapy in ctDNA positive patients. Clinical trials are expensive and lengthy endeavours and for 25 trials in colon cancer that are evaluating the impact of liquid biopsies, which started between 2007-2020, results are not expected until the mid to late 2020s for most studies[299]. Even after demonstrating clinical utility, assays need to obtain regulatory approval, achieve a favorable cost-benefit ratio and require investment in laboratory and human resources, including training of hospital staff in performing, analysing and interpreting data[114]. Thus, clinical adaptation will remain slow, lag technical innovation, and will likely first occur in a small set of specific clinical scenarios before wider adoption is possible.

Acknowledgements

All the work herein could not have been possible without the contributions from a myriad of people to whom I want to express my great appreciation and thanks:

First and foremost, to my main supervisor, Anders Ståhlberg, for taking me into your group and giving me the chance to work on such exciting projects. Your curiosity, positive energy and passion for research are truly inspiring! My co-supervisors, Erik Larsson Lekholm and Göran Landberg, for collaborations and providing valuable insights into other fields of science.

All the members, past and present, of the Ståhlberg lab. Daniel and Gustav for all your help, fun and extended discussions both on and off topic. Emma for introducing me to the lab when I started and always being extremely helpful with anything that needed doing. Soheila for always being sarcastic. Helena, Lisa, Maria, Manuel, Parmida, Pia, Peter, Tobias, and all former members for creating such a fantastic work environment and making the lab a great place to be, especially in the moments when the work is least fun.

Colleagues at Cancer Center and elsewhere for fika and lunch talks, meetings and collaborations, especially from the Åman, Landberg and Dalin groups. Anna and Karoline, for sharing the PhD journey, friendship and patiently listing to frequent rants. Agnieszka, Dorota and Jana for your friendship and humour. Kerryn from Erik Larsson's group for the great collaboration, fun talks and proof-reading my thesis. Also, Ágota, Christoffer, Elena, Emma P, Emil, Elin, Kristell, Malin, Mandy, Maryam, Mohamed, Pernilla, Sara, Simona, Tamara, and all the others at SCCR who made this a great time.

Lars Ny and Jonas Nilsson for your collaboration on the PEMDAC paper and the patients who donated their samples for the benefit of others. Tony Godfrey for laying the ground work to this thesis together with Anders at your lab in Boston.

The 'Lundberg crew' for all the good times during our Master's and beyond. All my friends be they here in Gothenburg, back in Germany or wherever you are, for your friendship, support and most importantly fun, whenever it was needed.

Meinen Eltern Brigitte und Wolfgang. Danke, dass ihr mir immer die Freiheit gegeben habt Dinge auf meine Art zu machen.

I would also like to thank the following, whose generous funding support made this research possible: The Assar Gabrielsson Research Foundation, Johan Jansson Foundation for Cancer Research, Wilhelm and Martina Lundgren Foundation, the Lions Cancer Research Fund of Western Sweden and the Adlerbertska foundation, the Region Västra Götaland, Swedish Cancer Society, Swedish Research Council, Swedish Childhood Cancer Foundation, Knut and Alice Wallenberg Foundation, the Swedish state under the agreement between the Swedish government and the county councils, the ALF-agreement, and Sweden's Innovation Agency.

Glossary

ASXL1 Putative Polycomb group protein ASXL1.
BAM Binary alignment map.
BAP1 BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase).
BEAMing Beads, emulsion, amplification, magnetics.
BRAF v-Raf murine sarcoma viral oncogene homolog B.
CAPP-Seq CAncer Personalized Profiling by deep Sequencing.
CBL Casitas B-lineage Lymphoma.
cfDNA Cell-free DNA.
CHIP Clonal hematopoiesis of indeterminate potential.
CPD Cyclobutane pyrimidine dimer.
ctDNA Circulating tumor DNA.
CYSLTR1 Cysteinyl leukotriene receptor 1.
ddPCR Droplet digital polymerase chain reaction.
DNA Deoxyribonucleic acid.
DNase I Deoxyribonuclease I.
DNMT3A DNA (cytosine-5)-methyltransferase 3A.
EDTA Ethylenediaminetetraacetic acid.
EGFR Epidermal growth factor receptor.
FDA United States Food and Drug Administration.
GNA11 Guanine Nucleotide-binding protein subunit Alpha-11.
GNAQ Guanine Nucleotide-binding protein G(q) subunit Alpha.
GNAS Guanine Nucleotide binding protein, Alpha Stimulating activity polypeptide.
HDAC Histone deacytylase.
HRAS GTPase HRas.
RAC1 Ras-related C3 botulinum toxin substrate 1.
HTS High-throughput sequencing.
IDH Isocitrate dehydrogenase.
JAK2 Janus kinase 2.
KIT Proto-oncogene c-KIT.
LDH Lactate dehydrogenase.
MAGERI Molecular tAgged GEnome Re-sequencing pIpeline.
NER Nucleotide excision repair.
NETosis Neutrophil extracellular trap release.
NF1 Neurofibromin 1.
NRAS Neuroblastoma RAS viral oncogene homolog.
PARE Personalized analysis of rearranged ends.
PCR Polymerase chain reaction.
PIK3CA Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha.
PLCB4 1-Phosphatidylinositol-4,5-bisphosphate phosphodiesterase beta-4.
PTEN Phosphatase and tensin homolog.

qPCR Quantitative polymerase chain reaction.
SafeSeqS Safe sequencing system.
SF3B1 Splicing factor 3B subunit 1.
SiMSen-Seq Simple Multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation Sequencing.
TC-NER Transcription-coupled nucleotide excision repair.
TET2 Tet methylcytosine dioxygenase 2.
TLS Trans-lesion synthesis.
TP53 Tumor protein p53.
UMI Unique molecular identifier.
UV Ultraviolet.
VAF Variant allele frequency.
WT1 Wilms' tumor protein.

Bibliography

1. Craig Venter, J. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome: International Human Genome Sequencing Consortium. *Nature* **412,** 565–566 (2001).
3. Jones, K. M. *et al.* Complicated legacies: The human genome at 20. *Science (New York, N.Y.)* **371,** 564–569 (2021).
4. Stephens, Z. D. *et al.* Big data: Astronomical or genomical? *PLoS Biology* **13,** 1–11 (2015).
5. Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278,** 274–288 (2005).
6. Avery, O. Y., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acids fraction isolated from Pneumococcus type III. *Journal of Experimental Medicine* **79,** 137–158 (1944).
7. Watson, J. & Crick, F. A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953).
8. Nirenberg, M. *et al.* The RNA code and protein synthesis. *Cold Spring Harbor symposia on quantitative biology* **31,** 11–24 (1966).
9. Holley, R. *et al.* Structure of a Ribonucleic Acid. *Science* **147** (1965).
10. Fiers, W *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature* **260,** 500–507 (1976).
11. Sanger, F, Nicklen, S & Coulson, A. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74,** 5463–5467 (1977).
12. Martin, W. J. *et al.* Automation of dna sequencing: A system to perform the sanger dideoxysequencing reactions. *Bio/Technology* **3,** 911–915 (1985).
13. Anderson *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 1–18 (1981).
14. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* **6,** 2601–2610 (1979).
15. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
16. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17,** 333–351 (2016).

17. Buermans, H. P. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease* **1842,** 1932–1941 (2014).

18. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* **30,** 418–426 (2014).

19. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).

20. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10** (2009).

21. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36** (2008).

22. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39** (2011).

23. Malapelle, U. *et al.* Ion Torrent next-generation sequencing for routine identification of clinically relevant mutations in colorectal cancer patients. *Journal of Clinical Pathology* **68,** 64–68 (2015).

24. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).

25. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* **21,** 597–614 (2020).

26. Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* **38,** e159–e159 (2010).

27. Kennedy, S. R., Loeb, L. A. & Herr, A. J. Somatic mutations in aging, cancer and neurodegeneration. *Mechanisms of Ageing and Development* **133,** 118–126 (2012).

28. Luria, S; Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **1943,** 491–511 (1943).

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

30. Hyman, R. W. *et al.* The dynamics of the vaginal microbiome during infertility therapy with in vitro fertilization-embryo transfer. *Journal of Assisted Reproduction and Genetics* **29,** 105–115 (2012).

31. Milani, C. *et al.* Gut microbiota composition and Clostridium difficile infection in hospitalized elderly individuals: A metagenomic study. *Scientific Reports* **6,** 1–12 (2016).

32. Nasu, A. *et al.* Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. *PLoS ONE* **6,** e24907 (2011).

33. Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research* **21,** 1616–1625 (2011).

34. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 13081–13086 (2008).

35. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

36. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–726 (2013).

37. Schmitt, M. W., Loeb, L. A. & Salk, J. J. The influence of subclonal resistance mutations on targeted cancer therapy. *Nature Reviews Clinical Oncology* **13,** 335–347 (2016).

38. Fisher, R. *et al.* Deep Sequencing Reveals Minor Protease Resistance Mutations in Patients Failing a Protease Inhibitor Regimen. *Journal of Virology* **86,** 6231–6237 (2012).

39. Boyd, S. D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science Translational Medicine* **1,** 1–16 (2010).

40. Howie, B. *et al.* High-throughput pairing of T cell receptor α and β sequences. *Science Translational Medicine* **7,** 1–12 (2015).

41. Vijg, J. Somatic mutations, genome mosaicism, cancer and aging. *Current Opinion in Genetics and Development* **26,** 141–149 (2014).

42. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. en. *Science* **349,** 1483–1489 (2015).

43. Fernández, L. C., Torres, M. & Real, F. X. Somatic mosaicism: on the road to cancer. *Nature Reviews Cancer* **16,** 43–55 (2015).

44. Acuna-Hidalgo, R. *et al.* Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *American Journal of Human Genetics* **101,** 50–64 (2017).

45. Rohlin, A. *et al.* Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques. *Human Mutation* **30,** 1012–1020 (2009).

46. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 14508–13 (2012).

47. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl* **1** (2014).

48. Chen, G., Mosier, S., Gocke, C. D., Lin, M. T. & Eshleman, J. R. Cytosine Deamination Is a Major Cause of Baseline Noise in Next-Generation Sequencing. *Molecular Diagnosis and Therapy* **18,** 587–593 (2014).

49. Chen; L., Liu; P., Jr.; T. C. E. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **361,** 752–756 (2018).

50. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research* **41,** 1–12 (2013).

51. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clinical Chemistry* **61,** 64–71 (2015).

52. Arbeithuber, B., Makova, K. D. & Tiemann-Boege, I. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA research : an international journal for rapid publication of reports on genes and genomes* **23,** 547–559 (2016).

53. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38,** 1767–1771 (2009).

54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754–1760 (2009).

55. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biology* **20,** 1–15 (2019).

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

57. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31,** 213–219 (2013).

58. MacKelprang, R. *et al.* Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480,** 368–371 (2011).

59. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nature Reviews Genetics* **20,** 341–355 (2019).

60. De la Puente, M. *et al.* Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. *Forensic Science International: Genetics* **45,** 102213 (2020).

61. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9,** 1056–1082 (2014).

62. Gansauge, M. T., Aximu-Petri, A., Nagel, S. & Meyer, M. Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. *Nature Protocols* **15,** 2279–2300 (2020).

63. Bianchi, D. W. Circulating fetal DNA: its origin and diagnostic potential-a review. eng. *Placenta* **25 Suppl A,** S93–s101 (2004).

64. Lo, Y. M. & Chiu, R. W. Prenatal diagnosis: progress through plasma nucleic acids. *Nat Rev Genet* **8,** 71–77 (2007).

65. De Franco, E. *et al.* Analysis of cell-free fetal DNA for non-invasive prenatal diagnosis in a family with neonatal diabetes. *Diabetic Medicine* (2016).

66. Johansson, G. *et al.* Ultrasensitive DNA Immune Repertoire Sequencing Using Unique Molecular Identifiers. *Clinical chemistry* **66,** 1228–1237 (2020).

67. Pisanic 2nd, T. R. *et al.* DREAMing: a simple and ultrasensitive method for assessing intratumor epigenetic heterogeneity directly from liquid biopsies. eng. *Nucleic Acids Res* **43,** e154 (2015).

68. Chen, C. *et al.* Ultrasensitive DNA hypermethylation detection using plasma for early detection of NSCLC: A study in Chinese patients with very small nodules. *Clinical Epigenetics* **12,** 1–11 (2020).

69. Liang, N. *et al.* Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nature Biomedical Engineering* **5,** 586–599 (2021).

70. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 9530–9535 (2011).

71. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359,** 926–930 (2018).

72. Ståhlberg, A. *et al.* Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Research,* gkw224 (2016).

73. Ståhlberg, A. *et al.* Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. *Nature Protocols* **12,** 664–682 (2017).

74. Wan, J. C. *et al. Liquid biopsies come of age: Towards implementation of circulating tumour DNA* 2017.

75. Johansson, G *et al.* Considerations and quality controls when analyzing cell-free tumor DNA. *Biomol Detect Quantif* **17,** 100078 (2019).

76. Vogelstein, B & Kinzler, K. W. Digital PCR. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 9236–41 (1999).

77. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature Medicine* **20,** 548–554 (2014).

78. Newman, A. M. *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nature biotechnology* **34,** 547–55 (2016).

79. Salk, J. J., Schmitt, M. W. & Loeb, L. A. *Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations* 2018.

80. Li, C *et al.* INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* **5,** 34 (2016).

81. Lou, D. I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences* **110,** 19872–19877 (2013).

82. Bystrykh, L. V. & Belderbos, M. E. Clonal analysis of cells with cellular barcoding: When numbers and sizes matter. *Methods in Molecular Biology* **1516,** 57–89 (2016).

83. Schmitt, M. W., Fox, E. J. & Salk, J. J. Risks of double-counting in deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 2014 (2014).

84. Best, K., Oakes, T., Heather, J. M., Shawe-Taylor, J. & Chain, B. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. en. *Scientific reports* **5,** 14629 (2015).

85. Schenk, D., Song, G., Ke, Y. & Wang, Z. Amplification of overlapping DNA amplicons in a single-tube multiplex PCR for targeted next-generation sequencing of BRCA1 and BRCA2. *PLoS ONE* **12,** 1–16 (2017).

86. Lanman, R. B. *et al.* Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. *PLoS ONE* **10** (ed Hoheisel, J. D.) e0140712 (2015).

87. Kechin, A. *et al.* NGS-PrimerPlex: High-throughput primer design for multiplex polymerase chain reactions. *PLoS Computational Biology* **16,** 1–12 (2020).

88. Wingo, T. S., Kotlar, A. & Cutler, D. J. MPD: Multiplex primer design for next-generation targeted sequencing. *BMC Bioinformatics* **18,** 1–5 (2017).

89. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 254–260 (2009).

90. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43,** 11.10.1–11.10.33 (2013).

91. Nystedt, B. *et al.* Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* **9,** 1–20 (2020).

92. Andrews, T. D., Jeelall, Y., Talaulikar, D., Goodnow, C. C. & Field, M. A. DeepSNVMiner: A sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. *PeerJ* **2016,** 1–13 (2016).

93. Shugay, M. *et al.* MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS Computational Biology* **13,** 13–17 (2017).

94. Xu, C. *et al.* Smcounter2: An accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers. *Bioinformatics* **35,** 1299–1309 (2019).

95. Sater, V. *et al.* UMI-VarCal: A new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. *Bioinformatics* **36,** 2718–2724 (2020).

96. Karst, S. M. *et al.* using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods* **18** (2021).

97. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research* **41,** e89 (2013).

98. Conway, J. R., Warner, J. L., Rubinstein, W. S. & Miller, R. S. Next-Generation Sequencing and the Clinical Oncology Workflow: Data Challenges, Proposed Solutions, and a Call to Action. *JCO Precision Oncology,* 1–10 (2019).

99. Donoghue, M. T. A., Schram, A. M., Hyman, D. M. & Taylor, B. S. Discovery through clinical sequencing in oncology. *Nature Cancer* **1** (2020).

100. Malone, E. R., Oliva, M., Sabatini, P. J., Stockley, T. L. & Siu, L. L. *Molecular profiling for precision cancer therapies* 2020.

101. Bozbiyik, O. *et al.* Reliability of fine needle aspiration biopsy in large thyroid nodules. *Turkish Journal of Surgery* **33,** 10–13 (2017).

102. Boyum, J. H. *et al.* Incidence and Risk Factors for Adverse Events Related to Image-Guided Liver Biopsy. *Mayo Clinic Proceedings* **91,** 329–335 (2016).

103. Shyamala, K, Girish, H. & Murgod, S. Risk of tumor cell seeding through biopsy and aspiration cytology. *Journal of International Society of Preventive and Community Dentistry* **4,** 5 (2014).

104. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61,** 759–767 (1990).

105. Tomasetti, Cristian; Li, Lui; Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355,** 1330–1334 (2017).

106. Engelman, J. A. & Settleman, J. Acquired resistance to tyrosine kinase inhibitors during cancer therapy. *Current Opinion in Genetics and Development* **18,** 73–79 (2008).

107. Katayama, R. *et al.* Mechanisms of acquired crizotinib resistance in ALK-rearranged lung Cancers. *Science translational medicine* **4,** 120ra17 (2012).

108. Kuukasjärvi, T. *et al.* Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Research* **57,** 1597–1604 (1997).

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

110. Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nature Medicine* **22,** 298–305 (2016).

111. Sprouffske, K. *et al.* Genetic heterogeneity and clonal evolution during metastasis in breast cancer patient-derived tumor xenograft models. *Computational and Structural Biotechnology Journal* **18,** 323–331 (2020).

112. Rolfo, C. & Russo, A. Liquid biopsy for early stage lung cancer moves ever closer. *Nature Reviews Clinical Oncology* **17,** 523–524 (2020).

113. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring cancer-genetics in the blood. *Nature reviews. Clinical oncology* **10,** 472–84 (2013).

114. Ignatiadis, M & Dawson, S. J. Circulating tumor cells and circulating tumor DNA for precision medicine: dream or reality? eng. *Ann Oncol* **25,** 2304–2313 (2014).

115. Ny, L. *et al.* The PEMDAC phase 2 study of pembrolizumab and entinostat in patients with metastatic uveal melanoma. *Nature Communications,* 1–10 (2021).

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

117. Mandel, P & Metais, P. Les acides nucléiques du plasma sanguin chez l'homme. *C R Seances Soc Biol Fil* **142,** 241–243 (1948).

118. Ferro, M. *et al.* Liquid biopsy biomarkers in urine: A route towards molecular diagnosis and personalized medicine of bladder cancer. *Journal of Personalized Medicine* **11,** 1–17 (2021).

119. Oshi, M. *et al.* Urine as a source of liquid biopsy for cancer. *Cancers* **13,** 1–15 (2021).

120. Wu, Z. *et al.* Differences in the genomic profiles of cell-free DNA between plasma, sputum, urine, and tumor tissue in advanced NSCLC. *Cancer Medicine* **8,** 910–919 (2019).

121. Keup, C. *et al.* Longitudinal multi-parametric liquid biopsy approach identifies unique features of circulating tumor cell, extracellular vesicle, and cell-free DNA characterization for disease monitoring in metastatic breast cancer patients. *Cells* **10,** 1–22 (2021).

122. Ococks, E *et al.* Longitudinal tracking of 97 esophageal adenocarcinomas using liquid biopsy sampling. *Annals of Oncology* **32,** 522–532 (2021).

123. Khan, K. H. *et al.* Longitudinal Liquid Biopsy and Mathematical Modeling of Clonal Evolution Forecast Time to Treatment Failure in the PROSPECT-C Phase II Colorectal Cancer Clinical Trial. *CANCER DISCOVERY* (2018).

124. Petrera, A. *et al.* Multiplatform Approach for Plasma Proteomics: Complementarity of Olink Proximity Extension Assay Technology to Mass Spectrometry-Based Protein Profiling. *Journal of Proteome Research* **20,** 751–762 (2021).

125. Zhou, J. *et al.* High-throughput single-EV liquid biopsy: Rapid, simultaneous, and multiplexed detection of nucleic acids, proteins, and their combinations. *Science Advances* **6,** 1–14 (2020).

126. Chang, J. W. C. *et al.* Transcriptomic analysis in liquid biopsy identifies circulating PCTaire-1 mRNA as a biomarker in NSCLC. *Cancer Genomics and Proteomics* **17,** 91–100 (2020).

127. Drula, R., Ott, L. F., Berindan-Neagoe, I., Pantel, K. & Calin, G. A. *Micrornas from liquid biopsy derived extracellular vesicles: Recent advances in detection and characterization methods* 2020.

128. Galardi, A. *et al.* Exosomal MiRNAs in pediatric cancers. *International Journal of Molecular Sciences* **20** (2019).

129. Sciandra, M. *et al.* Circulating miR34a levels as a potential biomarker in the follow-up of Ewing sarcoma. *Journal of Cell Communication and Signaling* **14,** 335–347 (2020).

130. Piao, X. M., Cha, E. J., Yun, S. J. & Kim, W. J. *Role of exosomal miRNA in bladder cancer: A promising liquid biopsy biomarker* 2021.

131. Baassiri, A. *et al.* Exosomal non coding RNA in LIQUID biopsies as a promising biomarker for colorectal cancer. *International Journal of Molecular Sciences* **21** (2020).

132. Tellez-Gabriel, M. & Heymann, D. Exosomal lncRNAs: The newest promising liquid biopsy. *Cancer Drug Resistance* **2,** 1002–1017 (2019).

133. Lee, B. *et al.* Integrated RNA and metabolite profiling of urine liquid biopsies for prostate cancer biomarker discovery. *Scientific Reports* **10,** 1–17 (2020).

134. Alix-Panabières, C. & Pantel, K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer discovery* **6,** 479–91 (2016).

135. Heidrich, I., Abdalla, T. S., Reeh, M. & Pantel, K. *Clinical applications of circulating tumor cells and circulating tumor DNA as a liquid biopsy marker in colorectal cancer* 2021.

136. Bu, Jiyoon; Hee Lee, Tae; Poellmann, Michael J.; Rawding, Piper A.; Jeong, Woo-Jin; Hong, Rachel S.; Hyun, Sung Hee; Eun, Hyuk Soo; Hong, S. Tri-modal liquid biopsy : Combinational analysis of circulating tumor cells , exosomes , and cell-free DNA using machine learning algorithm. *Clinical and translational medicine,* 1–6 (2021).

137. Huang, H. M. & Li, H. X. *Tumor heterogeneity and the potential role of liquid biopsy in bladder cancer* 2021.

138. Parikh, A. R. *et al.* Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nature Medicine* **25,** 1415–1421 (2019).

139. Stroun, M *et al.* Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46,** 318–322 (1989).

140. Kaimal, A. J., Norton, M. E. & Genetic, C. A. Clinical Management Guidelines for Obstetrician – Gynecologists Screening for Fetal Chromosomal. *ACOG Practice Bulletin* **136,** 48–69 (2020).

141. Kostenko, E. *et al.* Clinical and Economic Impact of Adopting Noninvasive Prenatal Testing as a Primary Screening Method for Fetal Aneuploidies in the General Pregnancy Population. *Fetal Diagnosis and Therapy* **45,** 413–423 (2019).

142. Kwapisz, D. The first liquid biopsy test approved . Is it a new era of mutation testing for non-small cell lung cancer ? *Annals of translational medicine* **5,** 1–7 (2017).

143. Diaz, L. A. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **32,** 579–86 (2014).

144. Njeim, R. & Azar, W. S. NETosis contributes to the pathogenesis of diabetes and its complications. *Journal of molecular endocrinology* (2018).

145. Zhang, K., Lin, G., Han, Y., Xie, J. & Li, J. *Circulating unmethylated insulin DNA as a potential non-invasive biomarker of beta cell death in type 1 Diabetes: a review and future prospect* 2017.

146. De Vlaminck, I. *et al.* Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Science translational medicine* **6,** 241ra77 (2014).

147. Ullrich, E. *et al.* Evaluation of host-based molecular markers for the early detection of human sepsis. *Journal of Biotechnology* **310,** 80–88 (2020).

148. Grabuschnig, S. *et al. Putative origins of cell-free DNA in humans: A review of active and passive nucleic acid release mechanisms* 2020.

149. Thierry, A. R., El Messaoudi, S., Gahan, P. B., Anker, P. & Stroun, M. Origins, structures, and functions of circulating DNA in oncology. *Cancer and Metastasis Reviews* **35,** 347–376 (2016).

150. Giacona, M. B. *et al.* Cell-free DNA in human blood plasma: Length measurements in patients with pancreatic cancer and healthy controls. *Pancreas* **17,** 89–97 (1998).

151. Jahr, S *et al.* DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer research* **61,** 1659–65 (2001).

152. Nagata, S, Nagase, H, Kawane, K, Mukae, N & Fukuyama, H. *Degradation of chromosomal DNA during apoptosis* 2003.

153. Stroun, M, Lyautey, J, Lederrey, C, Mulcahy, H. E. & Anker, P. Alu repeat sequences are present in increased proportions compared to a unique gene in plasma/serum DNA: evidence for a preferential release from viable cells? eng. *Ann N Y Acad Sci* **945,** 258–264 (2001).

154. Kahlert, C. *et al.* Identification of Double- stranded Genomic DNA Spanning All Chromosomes with Mutated KRAS and p53 DNA in the Serum Exosomes. *Journal of Biological Chemistry* **289,** 3869–3875 (2014).

155. Williams Caitlin; Rodriguez-Barrueco, ruth; Silva, Jose M; Zhang, Weija; Hearn, Stephen; Elemento, Olivier; Paknejad, Navid;Manova-Todorova, Katia;Welte, Karl; Bromberg, Jaqueline; Peinado, Hector; Lyden, D. Double-stranded DNA in exosomes : a novel biomarker in cancer detection. *Cell research* **24,** 766–769 (2014).

156. Lo, Y. M. *et al.* Rapid clearance of fetal DNA from maternal plasma. *American journal of human genetics* **64,** 218–24 (1999).

157. Hummel, E. M. *et al.* Cell-free DNA release under psychosocial and physical stress conditions. *Translational Psychiatry* **8** (2018).

158. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science translational medicine* **6,** 224ra24 (2014).

159. Rodrigues Filho, E. M., Ikuta, N, Simon, D & Regner, A. P. Prognostic value of circulating DNA levels in critically ill and trauma patients. eng por. *Rev Bras Ter Intensiva* **26,** 305–312 (2014).

160. Zwirner, K. *et al.* Circulating cell-free DNA : A potential biomarker to differentiate inflammation and infection during radiochemotherapy. *Radiotherapy and Oncology* **129,** 575–581 (2018).

161. Haghiac, M *et al.* Increased death of adipose cells, a path to release cell-free DNA into systemic circulation of obese women. eng. *Obesity (Silver Spring)* **20,** 2213–2219 (2012).

162. Polina, I. A., Ilatovskaya, D. V. & Deleon-pennell, K. Y. Cell free DNA as a diagnostic and prognostic marker for cardiovascular diseases. *Clinica Chimica Acta* **503,** 145–150 (2020).

163. Stewart, C. M. & Tsui, D. W. Y. Circulating cell-free DNA for non-invasive cancer management. *Cancer Genetics* **228,** 169–179 (2018).

164. Celec, P., Vlková, B., Lauková, L., Bábíčková, J. & Boor, P. Cell-free DNA: the role in pathophysiology and as a biomarker in kidney diseases. *Expert Reviews in Molecular Medicine* **20,** 1–14 (2018).

165. Kustanovich, A., Schwartz, R., Peretz, T. & Grinshpun, A. Life and death of circulating cell-free DNA. *Cancer Biology and Therapy* **20,** 1057–1067 (2019).

166. Gauthier, V. J., Tyler, L. N. & Mannik, M. Blood clearance kinetics and liver uptake of mononucleosomes in mice. *Journal of immunology (Baltimore, Md. : 1950)* **156,** 1151–6 (1996).

167. Korabecna, M. *et al.* Cell-Free Plasma DNA during Peritoneal Dialysis and Hemodialysis and in Patients with Chronic Kidney Disease. *Annals of the New York Academy of Sciences* **1137,** 296–301 (2008).

168. Koizumi, T. *Tissue Distribution of Deoxyribonuclease I (DNase I) Activity Level in Mice and its Sexual Dimorphism* 1995.

169. Stephan, F. *et al.* Cooperation of factor vii-activating protease and serum dnase i in the release of nucleosomes from necrotic cells. *Arthritis and Rheumatology* **66,** 686–693 (2014).

170. Korabecna, M. *et al.* Cell-free DNA in plasma as an essential immune system regulator. *Scientific Reports* **10,** 1–10 (2020).

171. Trejo-Becerril, C. *et al.* Cancer Progression Mediated by Horizontal Gene Transfer in an In Vivo Model. *PLoS ONE* **7,** 1–12 (2012).

172. Lee, T. H. *et al.* Barriers to horizontal cell transformation by extracellular vesicles containing oncogenic H-ras. *Oncotarget* **7,** 51991–52002 (2016).

173. Albrengues, J. *et al.* Neutrophil extracellular traps produced during inflammation awaken dormant cancer cells in mice. eng. *Science (New York, N.Y.)* **361** (2018).

174. Alborelli, I. *et al.* Cell-free DNA analysis in healthy individuals by next-generation sequencing: a proof of concept and technical validation study. *Cell Death and Disease* **10** (2019).

175. Moss, J *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* **9,** 5068 (2018).

176. Gregory, T. R. A bird's-eye view of the C-value enigma: Genome size, cell size, and metabolic rate in the class aves. *Evolution* **56,** 121–130 (2002).

177. Tiersch, T. R., Chandler, R. W., Wachtel, S. S. & Elias, S. Reference standards for flow cytometry and application in comparative studies of nuclear DNA content. *Cytometry* **10,** 706–710 (1989).

178. Mattox, A. K., Yan, H. & Bettegowda, C. The potential of cerebrospinal fluid-based liquid biopsy approaches in CNS tumors. *Neuro-Oncology* **21,** 1509–1518 (2019).

179. Escudero, L., Martínez-Ricarte, F. & Seoane, J. Ctdna-based liquid biopsy of cerebrospinal fluid in brain cancer. *Cancers* **13,** 1–15 (2021).

180. Kinde, I. *et al.* Evaluation of DNA from the papanicolaou test to detect ovarian and endometrial cancers. *Science Translational Medicine* **5** (2013).

181. Aro, K., Wei, F., Wong, D. T. & Tu, M. Saliva liquid biopsy for point-of-care applications. *Frontiers in Public Health* **5** (2017).

182. Cheng, J., Nonaka, T., Ye, Q., Wei, F. & Wong, D. T. W. in *Salivary Bioscience: Foundations of Interdisciplinary Saliva Research and Applications* (eds Granger, D. A. & Taylor, M. K.) 157–175 (Springer International Publishing, Cham, 2020).

183. Thress, K. S. *et al.* EGFR mutation detection in ctDNA from NSCLC patient plasma: A cross-platform comparison of leading technologies to support the clinical development of AZD9291. *Lung Cancer* **90,** 509–515 (2015).

184. Li, M., Diehl, F., Dressman, D., Vogelstein, B. & Kinzler, K. W. BEAMing up for detection and quantification of rare sequence variants. *Nature Methods* **3,** 95–97 (2006).

185. Ntzifa, A., Kroupis, C., Haliassos, A. & Lianidou, E. A pilot plasma-ctDNA ring trial for the Cobas® EGFR Mutation Test in clinical diagnostic laboratories. *Clinical Chemistry and Laboratory Medicine* **57,** E97–E101 (2019).

186. Kim, Y., Shin, S. & Lee, K. A. A Comparative Study for Detection of EGFR Mutations in Plasma Cell-Free DNA in Korean Clinical Diagnostic Laboratories. *BioMed Research International* **2018** (2018).

187. Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497,** 108–112 (2013).

188. Leary, R. J. *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. *Science translational medicine* **2,** 20ra14 (2010).

189. Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Science translational medicine* **4,** 162ra154 (2012).

190. Heitzer, E. *et al.* Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Medicine* **5,** 1–16 (2013).

191. Frenel, J. S. *et al.* Serial next-generation sequencing of circulating cell-free DNA evaluating tumor clone response to molecularly targeted drug administration. *Clinical Cancer Research* **21,** 4586–4596 (2015).

192. Lebofsky, R. *et al.* Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. *Molecular oncology* **9,** 783–90 (2015).

193. Siravegna, G., Marsoni, S., Siena, S. & Bardelli, A. *Integrating liquid biopsies into the management of cancer* 2017.

194. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine* **371,** 2488–2498 (2014).

195. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine* **371,** 2477–2487 (2014).

196. Hrebien, S *et al.* Early ctDNA dynamics as a surrogate for progression-free survival in advanced breast cancer in the BEECH trial, 1–8 (2019).

197. Pittella-silva, F., Chin, M., Chan, T., Nagayama, S. & Miyauchi, E. Plasma or Serum : Which Is Preferable for Mutation Detection in Liquid Biopsy ? *Clin Chem* **957** (2020).

198. El Messaoudi, S., Rolet, F., Mouliere, F. & Thierry, A. R. Circulating cell free DNA: Preanalytical considerations. *Clinica Chimica Acta* **424,** 222–230 (2013).

199. Bronkhorst, A. J., Aucamp, J. & Pretorius, P. J. Cell-free DNA: Preanalytical variables. *Clinica Chimica Acta* **450,** 243–253 (2015).

200. Sozzi, G. *et al.* Effects of Prolonged Storage of Whole Plasma or Isolated Plasma DNA on the Results of Circulating DNA Quantification Assays. *Journal of the National Cancer Institute* **97,** 24–26 (2005).

201. Diaz, I. M., Nocon, A., Mehnert, D. H., Fredebohm, J. & Diehl, F. Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing. *PLoS ONE,* 1–18 (2016).

202. Kang, Q. *et al.* Comparative analysis of circulating tumor DNA stability In K3EDTA, Streck, and CellSave blood collection tubes. *Clinical Biochemistry* **49,** 1354–1360 (2016).

203. Schmidt, B. *et al.* isolation and characterization of cell-free plasma DNA from tumor patients. *Clinica Chimica Acta* **469,** 94–98 (2017).

204. Gahlawat, A. W. *et al.* Evaluation of storage tubes for combined analysis of circulating nucleic acids in liquid biopsies. *International Journal of Molecular Sciences* **20** (2019).

205. Devonshire, A. S. *et al.* Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification. *Analytical and bioanalytical chemistry* **406,** 6499–512 (2014).

206. Streubel, A. *et al.* Comparison of different semi-automated cfDNA extraction methods in combination with UMI-based targeted sequencing. **10,** 5690–5702 (2019).

207. Andersson, D., Kristiansson, H., Kubista, M. & Ståhlberg, A. *Ultrasensitive circulating tumor DNA analysis enables precision medicine: experimental workflow considerations* 2021.

208. Janku, F. *et al.* A novel method for liquid-phase extraction of cell-free DNA for detection of circulating tumor DNA. *Scientific Reports* **11,** 1–9 (2021).

209. Connors, D. *et al.* Critical Reviews in Oncology / Hematology International liquid biopsy standardization alliance white paper. *Critical Reviews in Oncology / Hematology* **156,** 103112 (2020).

210. Geeurickx, E. & Hendrix, A. Targets , pitfalls and reference materials for liquid biopsy tests in cancer diagnostics. *Molecular Aspects of Medicine* **72,** 100828 (2020).

211. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology* **31** (2020).

212. Christensen, E. *et al.* Monitoring Treatment Response and Metastatic Relapse in Advanced Bladder Cancer by Liquid Biopsy Analysis Associate Editor : *European Urology* **3,** 541–542 (2017).

213. Tie, J. *et al.* Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Science Translational Medicine* **8,** 346ra92 (2016).

214. Merker, J. D. *et al.* Circulating Tumor DNA Analysis in Patients With Cancer: American Society of Clinical Oncology and College of American Pathologists Joint Review. *J Clin Oncol* **36,** 1631–1641 (2018).

215. Holmes, F. A., Levin, M. K., Cao, Y., Balasubramanian, S. & Ross, J. S. Comutation of PIK3CA and TP53 in Residual Disease After Preoperative Anti-HER2 Therapy in ERBB2 (HER2) -Ampli fi ed Early Breast Cancer abstract. *JCO Precision Oncology* **2,** 1–26 (2019).

216. Hong, D. S. *et al.* KRASG12C Inhibition with Sotorasib in Advanced Solid Tumors. *The New England journal of medicine,* 1207–1217 (2020).

217. Aggarwal, C. *et al.* Clinical Implications of Plasma-Based Genotyping with the Delivery of Personalized Therapy in Metastatic Non-Small Cell Lung Cancer. *JAMA Oncology* **5,** 173–180 (2019).

218. Heeke, S. *et al.* Critical Assessment in Routine Clinical Practice of Liquid Biopsy for EGFR Status Testing in Non e Small-Cell Lung Cancer : A Single-. *Clinical Lung Cancer* (2019).

219. Keppens, C. *et al.* Detection of EGFR Variants in Plasma A Multilaboratory Comparison of a Real-Time PCR EGFR Mutation Test in Europe. *The Journal of Molecular Diagnostics* **20,** 483–494 (2018).

220. Narayan, P. *et al.* Approval Summary : Alpelisib Plus Fulvestrant for Patients with HR-positive , HER2-negative , PIK3CA-mutated , Advanced or Metastatic Breast Cancer. *Clinical Cancer Research,* 1842–1850 (2021).

221. Martínez-sáez, O. *et al.* Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Research* **9,** 1–9 (2020).

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

223. Vogelstein, B *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).

224. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501,** 328–337 (2013).

225. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501,** 355–364 (2013).

226. Grimm, J. *et al.* BRAF inhibition causes resilience of melanoma cell lines by inducing the secretion of FGF1. *Oncogenesis* **7** (2018).

227. Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486,** 532–6 (2012).

228. Villanueva, J., Vultur, A. & Herlyn, M. *Resistance to BRAF inhibitors: Unraveling mechanisms and future treatment options* 2011.

229. Gide, T. N., Wilmott, J. S., Scolyer, R. A. & Long, G. V. Primary and Acquired Resistance to Immune Checkpoint Inhibitors in Metastatic Melanoma. *Clin Cancer Res* **24** (2018).

230. Schadendorf, D. *et al.* Melanoma. *Nature Reviews Disease Primers* **1,** 1–20 (2015).

231. Jager, M. J. *et al.* Uveal melanoma. *Nature Reviews Disease Primers* **6,** 18–20 (2020).

232. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150,** 251–263 (2012).

233. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545,** 175–180 (2017).

234. Ikehata, H. & Ono, T. The mechanisms of UV mutagenesis. *Journal of Radiation Research* **52,** 115–125 (2011).

235. Vaughn, C. M. & Sancar, A. in *DNA Damage{,} DNA Repair and Disease: Volume 2* 1–23 (The Royal Society of Chemistry, 2021).

236. Deacon, D. C., Smith, E. A. & Judson-Torres, R. L. Molecular Biomarkers for Melanoma Screening, Diagnosis and Prognosis: Current State and Future Prospects. *Frontiers in Medicine* **8** (2021).

237. Yang, K., Oak, A. S., Slominski, R. M., Brożyna, A. A. & Slominski, A. T. Current molecular markers of melanoma and treatment targets. *International Journal of Molecular Sciences* **21** (2020).

238. Bustamante, P. *et al.* Circulating tumor DNA tracking through driver mutations as a liquid biopsy-based biomarker for uveal melanoma. *Journal of Experimental and Clinical Cancer Research* **40,** 1–16 (2021).

239. Jin, E. & Burnier, J. V. Liquid Biopsy in Uveal Melanoma: Are We There Yet? *Ocular Oncology and Pathology* **7,** 1–16 (2021).

240. Deckers, E. A. *et al.* The association between active tumor volume, total lesion glycolysis and levels of S-100B and LDH in stage IV melanoma patients. *European Journal of Surgical Oncology* **46,** 2147–2153 (2020).

241. Gogas, H. *et al.* Biomarkers in melanoma. *Annals of Oncology* **20,** 8–13 (2009).

242. Wagner, N. B., Forschner, A., Leiter, U., Garbe, C. & Eigentler, T. K. S100B and LDH as early prognostic markers for response and overall survival in melanoma patients treated with anti-PD-1 or combined anti-PD-1 plus anti-CTLA-4 antibodies. *British Journal of Cancer* **119,** 339–346 (2018).

243. Rodríguez, M. F. *et al.* Blood biomarkers of Uveal Melanoma: Current perspectives. *Clinical Ophthalmology* **14,** 157–169 (2020).

244. Randic, T., Kozar, I., Margue, C., Utikal, J. & Kreis, S. NRAS mutant melanoma : Towards better therapies. *Cancer Treatment Reviews* **99,** 102238 (2021).

245. Schreuer, M. *et al.* Quantitative assessment of BRAF V600 mutant circulating cell-free tumor DNA as a tool for therapeutic monitoring in metastatic melanoma patients treated with BRAF/MEK inhibitors. *Journal of translational medicine* **14,** 95 (2016).

246. Cabel, L *et al.* Circulating tumor DNA changes for early monitoring of anti-PD1 immunotherapy : a proof-of-concept study Original article. *Changes in serum IL-8 levels reflect and predict response to anti-PD-1 treatment in melanoma and NSCLC* **28,** 1996–2001 (2017).

247. Rowe, S. P. *et al.* From validity to clinical utility : the influence of circulating tumor DNA on melanoma patient management in a real-world setting. *Molecular Oncology* **12,** 1661–1672 (2018).

248. Bidard, F. C. *et al.* Detection rate and prognostic value of circulating tumor cells and circulating tumor DNA in metastatic uveal melanoma. *International Journal of Cancer* **134,** 1207–1213 (2014).

249. Herbreteau, G. *et al.* Circulating tumour DNA is an independent prognostic biomarker for survival in metastatic BRAF or NRAS-mutated melanoma patients. *Cancers* **12,** 1–13 (2020).

250. Kaneko, A. *et al.* Liquid biopsy-based analysis by ddPCR and CAPP-Seq in melanoma patients. *Journal of Dermatological Science* **102,** 158–166 (2021).

251. Bjursten, S. *et al.* Response to BRAF/MEK Inhibition in A598-T599insV BRAF Mutated Melanoma. *Case Reports in Oncology* **12,** 872–879 (2019).

252. Vannas, C *et al.* Dynamic ctDNA evaluation of a patient with BRAFV600E metastatic melanoma demonstrates the utility of ctDNA for disease monitoring and tumor clonality analysis. *Acta Oncologica* **59,** 1388–1392 (2020).

253. Lee, J. H. *et al.* Pre-operative ctDNA predicts survival in high-risk stage III cutaneous melanoma patients. *Annals of Oncology,* 815–822 (2019).

254. 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).

255. Tie, J. Tailoring immunotherapy with liquid biopsy. *Nature Cancer* **1,** 857–859 (2020).

256. Fattore, L. *et al.* The Promise of Liquid Biopsy to Predict Response to Immunotherapy in Metastatic Melanoma. *Frontiers in Oncology* **11,** 1–12 (2021).

257. Seremet, T. *et al.* Illustrative cases for monitoring by quantitative analysis of BRAF/NRAS ctDNA mutations in liquid biopsies of metastatic melanoma patients who gained clinical benefits from anti-PD1 antibody therapy. *Melanoma Research* **28,** 65–70 (2018).

258. Lee, J. H. *et al.* Association Between Circulating Tumor DNA and Pseudoprogression in Patients With Metastatic Melanoma Treated With Anti-Programmed Cell Death 1 Antibodies. *JAMA Oncol* **4,** 717–721 (2018).

259. Bratman, S. V. *et al.* Personalized circulating tumor DNA analysis as a predictive biomarker in solid tumor patients treated with pembrolizumab. *Nature Cancer* **1,** 873–881 (2020).

260. Ashida, Atsuko; Sakaizawa, Kaori; Uhara, Hisashi; Okuyama, R. Circulating Tumour DNA for Monitoring Treatment Response to Anti-PD-1 Immunotherpay in Melanoma Patients. *Acta Dermato-Venerologica,* 1212–1218 (2017).

261. Knuever, J. *et al.* The use of circulating cell-free tumor DNA in routine diagnostics of metastatic melanoma patients. *Scientific Reports,* 1–8 (2020).

262. Forschner, A. *et al.* Circulating tumor DNA correlates with outcome in metastatic melanoma treated by braf and mek inhibitors – results of a prospective biomarker study. *OncoTargets and Therapy* **13,** 5017–5032 (2020).

263. Andersson, D. *et al.* Properties of targeted preamplification in DNA and cDNA quantification. *Expert Review of Molecular Diagnostics* (2015).

264. Filges, S., Yamada, E., Ståhlberg, A. & Godfrey, T. E. Impact of Polymerase Fidelity on Background Error Rates in Next- Generation Sequencing with Unique Molecular Identifiers / Barcodes. *Scientific reports* **9,** 1–7 (2019).

265. Filges, S., Mouhanna, P. & Ståhlberg, A. Digital Quantification of Chemical Oligonucleotide Synthesis Errors. *Clin Chem* **11,** 1–11 (2021).

266. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27,** 491–499 (2017).

267. Hugerth, L. W. *et al.* DegePrime , a Program for Degenerate Primer Design for Broad-Taxonomic-Range PCR in Microbial Ecology Studies. *Applied and Environmental Microbiology* (2014).

268. Meddeb, R., Al, Z., Dache, A. & Thezenas, S. Quantifying circulating cell-free DNA in humans, 1–16 (2019).

269. Mouliere, F. *et al.* Fragmentation patterns and personalized sequencing of cell-free DNA in urine and plasma of glioma patients. *EMBO Molecular Medicine* **13,** 1–14 (2021).

270. Zhang, J. *et al.* Presence of Donor- and Recipient-derived DNA in Cell-free Urine Samples of Renal Transplantation Recipients : Urinary DNA Chimerism. **1746,** 1741–1746 (1999).

271. Sidstedt, M *et al.* Inhibition mechanisms of hemoglobin, immunoglobulin G, and whole blood in digital and real-time PCR. *Anal Bioanal Chem* **410,** 2569–2583 (2018).

272. Giambernardi, T. A., Rodeck, U. & Klebe, R. J. Bovine serum albumin reverses inhibition of RT-PCR by melanin. *BioTechniques* **25,** 564–566 (1998).

273. Wei, G. H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO Journal* **29,** 2147–2160 (2010).

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

275. Zheng, C. L. *et al.* Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9,** 1228–1234 (2014).

276. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532,** 264–267 (2016).

277. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science* **339,** 957–959 (2013).

278. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nature Communications* **9** (2018).

279. Algazi, A. P. *et al.* Clinical outcomes in metastatic uveal melanoma treated with PD-1 and PD-L1 antibodies. *Cancer* **122,** 3344–3353 (2016).

280. Mignard, C. *et al.* Efficacy of Immunotherapy in Patients with Metastatic Mucosal or Uveal Melanoma. *Journal of Oncology* **2018** (2018).

281. Agarwala, S. S. *et al.* Efficacy and safety of entinostat (ENT) and pembrolizumab (PEMBRO) in patients with melanoma progressing on or after a PD-1/L1 blocking antibody. *Journal of Clinical Oncology* **36,** 9530 (2018).

282. Hellmann, M *et al.* OA05.01 Efficacy/Safety of Entinostat (ENT) and Pembrolizumab (PEMBRO) in NSCLC Patients Previously Treated with Anti-PD-(L)1 Therapy. *Journal of Thoracic Oncology* **13,** S330 (2018).

283. Khoja, L. *et al.* Meta-analysis in metastatic uveal melanoma to determine progression free and overall survival benchmarks: An international rare cancers initiative (IRCI) ocular melanoma study. *Annals of Oncology* **30,** 1370–1380 (2019).

284. Bradburn, M. J., Clark, T. G., Love, S. B. & Altman, D. G. Survival Analysis Part II: Multivariate data analysis- An introduction to concepts and methods. *British Journal of Cancer* **89,** 431–436 (2003).

285. Lim, A. M. *et al.* Delayed Response After Confirmed Progression (DR) and Other Unique Immunotherapy-Related Treatment Concepts in Cutaneous Squamous Cell Carcinoma. *Frontiers in Oncology* **11** (2021).

286. Butturini, E., de Prati, A. C., Boriero, D. & Mariotto, S. Tumor dormancy and interplay with hypoxic tumor microenvironment. *International Journal of Molecular Sciences* **20** (2019).

287. He, H. J. *et al.* Multilaboratory Assessment of a New Reference Material for Quality Assurance of Cell-Free Tumor DNA Measurements. *J Mol Diagn* (2019).

288. Egyud, M. *et al.* Detection of Circulating Tumor DNA in Plasma: A Potential Biomarker for Esophageal Adenocarcinoma. *Annals of Thoracic Surgery* **108,** 343–349 (2019).

289. Moser, T. *et al.* On-treatment measurements of circulating tumor DNA during FOLFOX therapy in patients with colorectal cancer. *npj Precision Oncology* **4** (2020).

290. Egyud, M. *et al.* Plasma circulating tumor DNA as a potential tool for disease monitoring in head and neck cancer. *Head and Neck* **41,** 1351–1358 (2019).

291. Lennon, A. M. *et al.* Feasibility of blood testing combined with PET-CT to screen for cancer and guide intervention. *Science* **369** (2020).

292. Caggiano, C. *et al.* Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nature Communications* **12** (2021).

293. Cristiano, S *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570,** 385–389 (2019).

294. Peneder, P. *et al.* Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nature Communications* **12,** 1–16 (2021).

295. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173,** 291–304.e6 (2018).

296. Chen, K. *et al.* Multiomics Analysis Reveals Distinct Immunogenomic Features of Lung Cancer with Ground-Glass Opacity. *American Journal of Respiratory and Critical Care Medicine,* 1–55 (2021).

297. Heitzer, E., Haque, I. S., Roberts, C. E. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics* **20,** 71–88 (2019).

298. Cree, I. A. *et al.* The evidence base for circulating tumour DNA blood-based biomarkers for the early detection of cancer: A systematic mapping review. *BMC Cancer* **17,** 1–17 (2017).

299. Parisi, A. *et al.* What is known about theragnostic strategies in colorectal cancer. *Biomedicines* **9,** 1–29 (2021).

300. Liu, Q. *A Prospective, Phase II Trial Using ctDNA to Initiate Post-operation Boost Therapy After NAC in TNBC* 2020.