University of Southern Denmark
Department of Clinical Research

Master's Thesis

Detection of circulating tumor DNA by recalling patient-specific somatic variants in plasma samples using whole-exome sequencing

Detektion af cirkulerende tumor DNA ved at genkalde patientspecifikke somatiske varianter i plasmaprøver ved brug af helexomsekventering

Submitted: 2023-06-30

Characters: 107.321

Degree Programme: Computational Biomedicine, MSc

Author: Jakob Jersild Nielsen
Master of Science Student

Supervisors: Torben A. Kruse
Professor, KI, Clinical Genetics

Mads Thomassen
Professor, KI, Clinical Genetics

Lars Andersen
Post Doc., KI, Clinical Genetics

Kristina M. Koldby
Post Doc., KI, Clinical Genetics

SDU

Acknowledgments

I take this opportunity to express my profound gratitude and deep regard to my supervisor, Torben A. Kruse, and my co-supervisor, Mads Thomassen, for their exceptional guidance, patience, and constant encouragement throughout the duration of this thesis. The completion of this thesis would not have been possible without them.

I would also like to thank Lars Andersen and Kristina M. Koldby for their guidance regarding the bioinformatic aspects of this thesis.

Last but certainly not least, I would like to express my heartfelt gratitude to my family and friends. Their unwavering support, especially in recent years, has been a beacon of strength. The encouragement and love they have given me have fueled my determination and resilience during the most challenging stages of this journey.

Contents

Figures

Tables

Abbreviations

A	Adenine
AS	Alignment scores
AF	Allele frequency
AP	Apurinic/apyrimidinic
AI	Artificial intelligence
BER	Base excision repair
bp	Base pair
BWA	Burrows-Wheeler Aligner
CPU	Central processing unit
cfDNA	Circulating cell-free DNA
CTC	Circulating tumor cell
ctDNA	Circulating tumor DNA
CIGAR	Compact idiosyncratic gapped alignment report
CT	Computed tomography
CNV	Copy number variation
C	Cytosine
DNA	Deoxyribonucleic acid
dNTP	Deoxyribose nucleotide triphosphate
DSBs	Double-stranded breaks
dsDNA	Double-stranded DNA
DP	Dynamic programming
ES	Excited state
FM	Ferragina-Manzini
FPGA	Field programmable gate array
GPU	Graphics processing unit
G	Guanine
HC	High confidence
HR	Homologous recombination
INDEL	Insertion and deletion
IDL	Insertion and deletion loop
LOH	Loss of heterozygosity
ML	Machine learning
MRI	Magnetic resonance imaging
MAPQ	Mapping quality scores
MEM	Maximal exact matches

MMEJ	Microhomology-mediated end joining
MMR	Mismatch repair
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
NMR	Nuclear magnetic resonance
NER	Nucleotide excision repair
PON	Panel of normal
PEP	Polyethylene glycol
PCR	Polymerase chain reaction
ROS	Reactive oxygen species
RNA	Ribonucleic Acid
SAM	S-adenosyl methionine
SGS	Second-generation sequencing
SBS	Sequencing-by-synthesis
SNV	Single nucleotide variant
ssDNA	single-stranded DNA
SPRI	Solid-phase reversible immobilization
SV	Structural variant
SVM	Support vector machine learning
TS	Targeted sequencing
TPU	Tensor processing unit
TGS	Third-generation sequencing
T	Thymine
UV	Ultraviolet
UDI	Unique dual indexes
UMI	Unique molecular identifiers
U	Uracil
VAF	Variant allele frequency
WC	Watson-Crick
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Abstract

Pancreatic cancer is among the deadliest types of cancer. Its high death rate is primarily due to late diagnoses because patients with early-stage disease often are asymptomatic, symptoms generally are non-specificity, and current screening methods are inefficient. In addition, there is a clinical shortcoming of sensitive low-burden disease monitoring approaches, which potentially can be resolved by implementing precision medicine and liquid biopsies. Analysis of ctDNA with next-generation sequencing has been proven effective in cancer prognostics and monitoring. However, its application in cancer diagnostics and screening is currently impeded by technical limitations. This thesis aims to improve upon an existing ctDNA detection tool, MRDetect, by utilizing different somatic nucleotide variant consensus ensembles, leveraging the size difference of ctDNA and circulating cell-free DNA fragments through *in silico* size selection, and integrating different ctDNA detection approaches. It was determined that generating single nucleotide variant consensus ensembles with calls from an increased number of somatic variant callers improved the sensitivity, whereas reducing the level of concordance to 3/7 of the included variant callers further enhanced the sensitivity without an apparent loss of specificity. This amplified the SNV signals, enabling the detection of ctDNA in all four samples with MRDetectSNV. Conversely, the CNV signals were inadequate, with ctDNA detected in only one sample with MRDetectCNV. In silico size selection considerably weakened the SNV signals but enhanced the CNV signals overall. The combination of statistical scores from ctDNA detection approaches was found to improve ctDNA detection. However, for ctDNA detection to achieve clinical relevance, further investigation and development of standardized workflows are necessary.

Resume

Bukspytkirtelkræft er blandt de dødeligste kræftformer. Den høje dødsrate skyldes i stor grad sen diagnose, da patienter i tidligt sygdomsstadie ofte er asymptomatiske, symptomerne generelt er uspecifikke, og de nuværende screeningsmetoder er ineffektive. Derudover er der en klinisk mangel på sensitive metoder til monitorering af lav byrde sygdomme, hvilket potentielt kan løses ved implementering af præcisionsmedicin og flydende biopsier. Analyse af ctDNA med næste generations sekventering har vist sig at være effektiv i kræftprognose og monitorering. Dog er anvendelsen i kræftdiagnostik og screening aktuelt hæmmet af tekniske begrænsninger. Denne afhandling sigter mod at forbedre det eksisterende ctDNA detektionsværktøj, MRDetect, ved at bruge forskellige SNV-konsensus ensembler, udnytte størrelsesforskellen på ctDNA og cirkulerende cellefri DNA-fragmenter gennem in silico størrelsesselektion, og integrere forskellige ctDNA detektionsmetoder. Generering af konsensus ensembler med SNV-kald fra et øget antal somatiske variantkaldere resulterer i forbedret sensitivitet, hvor reduktionen af overensstemmelsesniveauet til 3/7 af de inkluderede variantkaldere forbedrede sensitiviteten yderligere uden tydeligt tab af specificitet. Dette forstærkede SNV-signalerne, hvilket muliggjorde detektion af ctDNA i alle fire prøver med MRDetectSNV. Modsat var CNV signalerne utilstrækkelige, hvilket gjorde at ctDNA kun kunne detekteres i en af prøverne med MRDetectCNV. *In silico* størrelsesselektion svækkede SNV signalerne betydeligt, men forbedrede overordnet set CNV signalerne. Kombinationen af statistiske scoringer fra ctDNA detektionsmetoderne viste sig at forbedre detektionen af ctDNA. Dog er yderligere undersøgelser og standardiseringer nødvendige for, at ctDNA detektion kan opnå klinisk relevans.

1 Aim of study

This thesis aims to evaluate and optimize the sensitivity and specificity of ctDNA detection approaches, focusing on SNV-based and CNV-based methods. This includes an assessment of the impact of using an increased number of variant callers and reduced stringency of consensus ensembles on SNV-based ctDNA detection. Furthermore, the thesis aims to improve the SNV-based and CNV-based ctDNA detection by exploring the effects of in silico DNA fragment length selection. Lastly, the thesis seeks to examine the potential benefits of combining the statistical scores from ctDNA detection approaches.

2 Introduction

2.1 An overview of pancreatic cancer

Cancer is amongst the leading causes of death, estimated to have claimed the lives of 10 million people worldwide in 2020, while approximately 19.3 million new cases occurred (1). In 2019, cancer ranked as the most common cause of premature death in adults below 70 years of age in 57 countries, including North America and most of Europe (1). Pancreatic cancer is considered one of the most lethal cancers and accounts for 1.8% and 4.6% of all cancer incidences and mortalities, respectively (2). Over the past 25 years, pancreatic cancer's incidence, prevalence, and mortality have increased by 55%, 63%, and 53%, respectively (2, 3). A 1.97-fold increase in pancreatic cancer mortality is anticipated by 2060 (2). Pancreatic cancers are classified as localized, locally advanced, and metastatic, with 5-year survival rates of 44%, 15%, and 3%, respectively, while the overall 5-year survival rate is 12% (4). The majority of patients with pancreatic cancer present with locally advanced (30-35%) or metastatic (50%) disease at diagnosis, as patients with early-stage localized pancreatic disease (10-15%) often are asymptomatic (5, 6).

Screening patients for pancreatic cancer can be conducted by detecting serum levels of the sialylated Lewisa blood group antigen CA 19-9 tumor biomarker, which can indicate tumor size and metastatic status (7-9). Notedly, 10% of the population does not express Lewis antigens (10). Suspected pancreatic tumors are commonly evaluated using imaging techniques such as CT and MRI (11-14), where identified discrepancies can be corroborated by immunohistochemistry staining of tissue samples (15, 16), often collected with needle biopsies (16, 17). Treatment of patients with locally advanced or metastatic disease can include chemotherapy and radiochemotherapy to reduce disease progression and increase life expectancy (18, 19). In contrast, treatment of patients with localized pancreatic cancer typically includes surgical resection with neoadjuvant or perioperative therapies (20). Prognostication in cancer conventionally involves using predictive statistical models built on strata (21-23), which are statistical generalizations of cohorts with similar characteristics (24, 25).

Over the past decade, biotechnological advancements have pioneered the innovation of novel analyses of tumor-specific biomarkers, such as CTCs and ctDNA, derived from non-invasive liquid biopsies to improve cancer diagnostics and prognostics (26-30). Presumably, ctDNA shedding results from exhausted phagocytosis of apoptotic and necrotic tumor cells (31-34) but can also occur through

active secretion of extracellular vesicles, such as exosomes and prostasomes, by living tumor cells (34-37). ctDNA typically constitutes between <0.01% and 10% of the total cfDNA while varying between tumor types and clinical stages (38-42). Fragmentation patterns of cfDNA have been observed to be different in cancer patients compared to healthy individuals. Typically, cfDNA fragments from cancer patients are shorter (134-144 bp) than those from healthy individuals (166 bp). Furthermore, cfDNA from cancer patients shows a distinct 10 bp periodicity below 146 bp (43, 44). This reflects nuclease-cleaved nucleosome activity and indicates cfDNA is nonrandomly fragmented (45-47). The smaller size of ctDNA fragments can be leveraged by in vitro and in silico size selection, which may enhance the sensitivity of ctDNA detection. Recent studies have demonstrated 2-fold median ctDNA enrichment by in vitro and in silico size selection of cfDNA fragment sizes ranging between 90-150 bp (48, 49). Furthermore, ctDNA rapidly degrades with a half-life of 30-120 minutes (50-53), which enables real-time tumor burden assessment with greater sensitivity than radiological imaging (54-56).

2.2 Types of DNA lesions and their repair

DNA damage, sometimes called DNA lesions, is broadly categorized as endogenous and exogenous based on their origin. Endogenous DNA damage is naturally occurring in all our cells. It most commonly includes DNA replication errors and DNA engaging in hydrolysis, oxidation, and methylation reactions that damage the DNA bases and sugar backbone (57-59). Exogenous DNA damage results from exposure to physical and chemical agents such as UV light and certain compounds in processed tobacco (60-62). The damages are detected by lesion-specific DNA damage response mechanisms that promote their repair through various pathways (63). Disruptive variants in the genes that encode proteins involved in the DNA damage repair pathways can potentially increase the overall rate of mutagenesis (64-66).

Amongst the most common DNA base lesions are deamination, depurination, depyrimidination, alkylation, and oxidation. Deamination reactions can occur spontaneously, in which A, G, and C are converted to hypoxanthine, xanthine, and U, respectively. Additionally, a particular methylated form of C can be deaminated to T. Depurination and depyrimidination of bases can occur spontaneously or enzymatically, resulting in AP sites (57-59, 67). DNA base methylations can result from spontaneous reactions with SAM or other endogenous alkylating agents (57-59, 68, 69). SAM is usually used as a methyl donor in normal methylation reactions catalyzed by methyltransferases (70). DNA base oxidations can result from reactions with ROS (57-59, 71). The repair of damaged bases typically

occurs through the BER pathway (67, 72-74), while a few base modifications can be reverted through enzyme-mediated direct chemical reversal (75-78). Multiple base damages and bulky lesions, such as intrastrand cross-linked nucleotide dimers, are repaired through the NER pathway (79-81).

In the BER pathway, DNA glycosylases recognize certain modified bases and catalyze their hydrolytic removal (67, 72-74, 82-86), resulting in the creation of AP sites that are removed sequentially by an AP endonuclease and a phosphodiesterase. Spontaneous and chemically induced AP sites are excised similarly (67, 72-74, 86). The pathway is finalized by DNA polymerase gap-filling (67, 72-74, 87) and DNA ligase nick-sealing (67, 72-74, 88, 89). Failure to repair DNA base lesions can resolve in SNVs, as modified bases often engage in mutagenic base pairing during replication. Some base modifications can also result in replication fork stalling, leading to genome instability (67, 90-94).

The NER pathway is initiated upon detection of intrastrand crosslinks and bulky lesions by a multienzyme complex (79, 80, 95-98), followed by the recruitment of two structure-specific endonucleases that nick the strand short distances from the lesion (79, 80, 99, 100). The pathway is finalized by DNA polymerase gap-filling and DNA ligase nick-sealing (79, 80, 101). Persistent bulky lesions can block replication by replication fork stalling (102, 103).

Replication errors are introduced as DNA polymerases are inherently error-prone. The fidelity and processivity of DNA polymerases refer to the ability to replicate a template accurately and continuously, respectively (104, 105). The fidelity of DNA polymerases is essentially governed by free energy changes during the polymerization reaction, discriminating between correct and incorrect nucleotide incorporation by differential free energy barriers (106-108). Yet, nucleotides can be misincorporated consequent to dNTP pool imbalances (109, 110), DNA damage (90, 91), and by adopting wobble and WC-like conformations through processes of tautomerization and ionization (111-119). Mismatched bps in DNA and RNA have been shown to dynamically interchange between wobble configuration and tautomeric and anionic ES with WC-like geometry in a triangular exchange topology (114-116). The tautomeric ES population increases with temperature, whereas the anionic ES population increases with pH (114-116). Anionic bps can additionally exist in an inverted wobble conformation (117-119). Tautomeric and anionic bps are thought to evade fidelity checkpoints because of their WC-like geometry (111-113), unlike wobble bps that are accommodated by conformational changes of the DNA backbone (120-123).

Misaligned strands can arise through slipped-strand, dNTP-stabilized, and misincorporation misalignment. Slipped-strand misalignment typically occurs in genomic regions with direct repeats,

such as microsatellites and simple sequence repeats (124-129). Strand-slippage results from DNA polymerase pausing and disassociation (127, 130), allowing the end of a newly synthesized repetitive strand to separate from its template and rehybridize to other direct repeats, often out-of-frame (127, 129) while forming an extrahelical IDL (128, 129, 131). The process is finalized by DNA polymerase reassociation and replication continuation (127, 129). The dNTP-stabilized and misincorporation misalignment mechanisms can occur in repetitive and non-repetitive DNA sequences (129, 132-135). The dNTP-stabilized misalignment mechanism occurs within the active site of a DNA polymerase, in which a template base is translocated, positioning the next template base within the active site of the polymerase, thereby allowing the supplied dNTP to correctly bind to the next template base (129). This mechanism results in intrahelical bulges where the skipped base is stacked within the DNA double-helix (129, 136). Misincorporation misalignment is instigated by the misincorporation of a nucleotide. If this nucleotide is complementary to the next template base, it may translocate (135). Notedly, this mechanism is induced explicitly by dNTP pool imbalances (109, 135). The frequency of misincorporation misalignment is assumed to be lower than the other two misalignment mechanisms, as it depends on the rate of bp mismatch incorporation and the probability of the mismatched base being complementary to a neighboring template base (129).

The majority of replication errors are typically corrected during replication by exonucleolytic proofreading (137-140), intrinsic to most DNA polymerases, or subsequently by MMR (141, 142). The MMR pathway is initiated upon detection of mismatches and IDLs by lesion-specific heterodimer sensor protein complexes (143-147). Several MMR initiation models have been proposed (148-153). Regardless of the exact initiation mechanism, studies have shown that the sensor protein complex recruits additional proteins, including an enzyme with endonucleolytic activity that makes an incision on the damaged strand and an exonuclease that excises the mismatch from the incision (154-159). The endonucleolytic incision is thought to occur at hemimethylated sites, thereby ensuring excision of the damaged strand, as newly synthesized DNA is unmethylated (159-163). The repair process is finalized by DNA polymerase-catalyzed resynthesis of the resulting ssDNA gap and DNA ligase nick-sealing through the catalyzed formation of phosphodiester bonds between adjacent 5' phosphate and 3' hydroxyl terminal groups (157-159). Uncorrected mismatches and misalignments result in SNVs and INDELs, respectively.

SSBs result from indirect, direct, and topoisomerase-mediated damages (164-167). Indirect damages are induced during the repair of DNA base damages through the BER pathway (167, 168). In

contrast, direct damages result from oxidative damage to the DNA sugar backbone by ROS (57-59, 167). Topoisomerase-mediated damages result from erroneous nicks (166, 167, 169).

Direct and topoisomerase-mediated SSBs are detected by a poly ADP-ribose polymerase that tags the lesions and promotes their repair through a pathway similar to that of BER (166, 170-173), in which damage-specific enzymes process the SSB ends. The ends of direct SSBs are processed by a polynucleotide kinase phosphatase or aprataxin (166, 167, 174, 175), while the ends of topoisomerase-mediated SSBs are processed by a phosphodiesterase (166, 167, 169, 176). The process is finalized by DNA polymerase gap-filling and DNA ligase nick-sealing (166, 167, 177-180). Failure to repair SSBs can result in replication fork stalling (92, 166, 167, 181).

DSBs can result from replication fork stalling and interstrand nucleotide cross-linkage (181). They can be repaired through two major pathways: HR and NHEJ (182-186). The DSB repair pathways are utilized to different extents throughout the cell cycle (185, 186).

HR constitutes a set of subpathways that employ strand invasion to resolve DSBs (187-189). The DSB ends are initially prepared for strand invasion by nucleolytic resection, resulting in long ssDNA overhangs (187, 188, 190). The ssDNA overhangs are subsequently bound by recombinases that mediate strand invasion (187, 188), usually using sister chromatids or homologous chromosomes as templates for DNA synthesis (191, 192). HR can result in LOH through gene conversion and SVs through crossing over (187-189).

In the NHEJ pathway, the DSB ends are bound by a particular protein heterodimer that joins the DSB ends and recruits various end-processing enzymes (193-205). For instance, DNA polymerases and nucleases catalyze the addition and resection of nucleotides, respectively, while DNA ligases seal the remaining nicks (206-208). Both DSB ends are processed independently, where the enzymes can act in any order and multiple times (204-206). Although NHEJ is conservative, it can potentially resolve in very small INDELs and SVs such as translocations (204-206, 209).

The absence of the end-joining protein heterodimer characteristic of NHEJ allows repairing through an alternative pathway of MMEJ (210-213). In this pathway, the DSB ends initially undergo nucleolytic resection, followed by sequential protein-mediated end-bridging and annealing of microhomologies (213-219). Afterward, the remaining overhanging ssDNA strands are removed by an endonuclease (211, 219-221). The process is finalized by DNA polymerase gap-filling and DNA ligase nick-sealing (219, 222-225). MMEJ is considered highly error-prone, potentially resolving in INDELs and SVs such as translocations (225-230).

2.3 Next-generation sequencing

Considerable advancements in high-throughput sequencing technologies with NGS have enabled rapid and cost-effective sequencing of DNA and RNA molecules relative to Sanger sequencing, the first published sequencing technology (231-233). NGS technologies are typically categorized according to read length. Short-read sequencing technologies, such as Illumina and Ion Torrent, are referred to as SGS. In contrast, long-read sequencing technologies, such as Pacific Biosciences and Oxford Nanopore, are referred to as TGS (234, 235). Library construction is a prerequisite to sequencing, where the required library preparation steps depend on the sequencing platform and use case (234, 235). Library preparation for TGS platforms requires minimal steps compared to SGS platforms (234, 235).

For Illumina platforms, the DNA library preparation workflow traditionally comprises DNA fragmentation, end-repair, dA-tailing, adapter ligation, size selection, and PCR amplification (236). The adapters contain indexes unique to the library that enable sample pooling and multiplex sequencing (237). Library preparation for WES and TS includes an additional target enrichment step (236). DNA fragmentation is unnecessary for cfDNA from liquid biopsies, as it sheds from cells as fragments (31).

2.3.1 Illumina library preparation

DNA fragmentation is commonly performed through enzymatic digestion or mechanical shearing. Mechanical DNA fragmentation methods include sonication (238-241), nebulization (242-245), and sink-point shearing (244-250). DNA fragmentation by sonication utilizes acoustic shearing that induces DSBs in high molecular weight DNA through resonance vibrations of ultrasonic sound waves (238-241). Covaris, a popular acoustic shearing device, can produce a variety of fragment sizes ranging from 150 bp to 20 kb through acoustic energy control (251). DNA fragmentation by nebulization and sink-point shearing utilizes hydrodynamic shearing forces to induce DSBs in DNA by repeatedly accelerating DNA through a small cavity (242-250). This acceleration is accomplished with high-pressure air or gas for nebulization (242, 243) and a high-performance liquid chromatography or syringe pump for sink-point shearing (246-250). DNA fragments resulting from nebulization have narrowly distributed sizes ranging from 700 bp to 1330 bp (243), while those resulting from sink-point shearing have wider size distributions ranging from 1 kb to 12 kb (246-250). Despite the general notion that acoustic shearing is random, studies have revealed that the method preferentially

induces DSBs in 5'-CpG-3' dinucleotides (252-255). This sequence specificity of acoustic shearing has been demonstrated to result from sequence-dependent conformational dynamics and is presumably modulated by the intensity of deoxyribose S⇔N interconversion (252-256) but may also pertain to epigenetic mechanisms of d(CpG) methylation (252, 257, 258). Similar sequence specificity has been observed for hydrodynamic shearing, indicating common physicochemical properties associated with the mechanochemical shearing forces (252). Notedly, acoustic shearing has been demonstrated to induce oxidative DNA damage through sonochemical reactions, which can introduce artifactual variants such as C > A/G > T transversions (259-266). Interestingly, these oxidative sonochemical reactions exhibit sequence-dependent preferences (263-266). Additive antioxidants have been proposed to reduce sonochemical oxidative DNA damage (263). Enzymatic DNA fragmentation methods commonly utilize endonucleases (267-270) or transposases (271, 272). Many enzymatic methods utilize DNase I (269), an endonuclease that cleaves phosphodiester bonds through hydrolysis (273, 274). The DNase I activity is strictly dependent on calcium ions (273) and can be activated by divalent metal ions such as magnesium and manganese (273, 274). In the presence of magnesium ions, DNase I cleaves each strand of dsDNA independently; while in the presence of manganese ions, DNase I cleaves both strands of dsDNA at approximately the same site, resulting in both blunt and staggered DSBs (274). While transposases are known to catalyze the genomic transposition of transposon DNA sequences, the activity of a hyperactive Tn5 transposase has been shown to resolve in fragmentation through a cut-and-paste transposition mechanism (271, 272), which, together with numerous other discoveries (275) led in its introduction into library preparation (271, 272). In a process termed tagmentation, the hyperactive Tn5 transposases simultaneously fragments DNA and ligates sequencing adapters (271, 272). Thus, simplifying the traditional library preparation workflow by replacing sequential DNA fragmentation, end-repair, A-tailing and adapter ligation with a single step (271, 272). These enzymatic methods can produce a variety of DNA fragment sizes, dependent on fragmentation conditions such as time and temperature (276-280). A certain degree of sequence specificity has been determined for DNase I (281-283) and Tn5 transposase (284-287). For DNase I, the sequence specificity has been demonstrated to result from sequence-dependent local conformational dynamics pertaining to the flexibility and bendability of DNA (281-283). This was deduced from a positive correlation between the averaged DNase I cleavage intensities and ^{31}P NMR chemical shifts of phosphodiester bonds adjacent 5' to CpA•TpG dinucleotides, known to be highly flexible (283). Despite initial concerns regarding the introduction of sequence bias consequent to the sequence specificity of endonucleases (280), recent

commercial library preparation kits with enzymatic and mechanical DNA fragmentation for Illumina sequencing have proved comparable (288). However, the transposase-based library preparation kit Nextera XT introduced a strong AT bias (288-294), attributable to a symmetric binding motif with a central AT-rich region surrounded by Gs and Cs (289). It was deduced that the Gs and Cs were the main cleavage determinant, as the method exclusively introduces AT bias (289). Furthermore, studies have revealed that enzymatic fragmentation introduces more artifactual variants than mechanical fragmentation (280, 290). Thus, optimization of enzymatic fragmentation is necessary for its viability.

End-repair is a crucial step before adapter ligation in the traditional library preparation workflow, as most DNA fragmentation methods inherently result in a mix of staggered and blunt DSBs (236, 291). The end-repair is performed through enzymatic blunting of the DNA fragment ends with overhanging ssDNA sequences, where a DNA polymerase fills-in recessed 5' ends and exonucleolytically excises 3' overhangs, and a polynucleotide kinase phosphorylates the 5' ends (236, 291-293). Subsequently, the library is primed for adapter ligation through dA-tailing, which involves the enzymatic addition of an A base to the 3' ends of the blunt DNA fragments, commonly with a Klenow Fragment without exonucleolytic activity (236, 291-293). While the A overhang facilitates the adapter ligation by pairing with a complementary T overhang on the adapter, a DNA ligase catalyzes the formation of a phosphodiester bond between the 3' and 5' ends of the DNA fragment and adapter (236, 291, 292). These steps can introduce artifacts such as sequence bias, extension errors, and chimeras. They can generally be affected by factors such as utilized enzymes and incubation conditions, including time and temperature (291, 294). Sequence bias results from the sequence specificity of the involved enzymes, which can lead to preferential end repair and ligation, and thus misrepresentation, of certain sequences (291, 295). Replication errors are artifactual variants that arise from spontaneous mutagenesis, a direct consequence of the fidelity of the enzymes with polymerase activity (296-298). Although the A and T overhangs are supposed to preclude self-ligation and formation of chimeras, hybrids of two or more DNA fragments, incomplete end repair, and DNA ligase fidelity can resolve in chimera formation through, e.g., hybridization of overhangs with high sequence complementarity (294, 299-302).

Size selection is commonly performed with gel electrophoresis (303) or SPRI beads (304-308). The gels used for electrophoresis can be made from agarose or polyacrylamide in addition to a buffer solution (303). Applying an electric current to the gel makes DNA fragments migrate towards the positive charge, where the migration distance depends on the DNA fragment size (303). Smaller DNA fragments migrate the furthest (303). The size selection is performed through DNA ladder-guided exci-

sion of a gel slice (303). As gel electrophoresis tends to be rather time-consuming, carboxylate-coated magnetic SPRI beads, such as AMPure XP, are preferably utilized for size selection, which reversibly binds DNA in the presence of PEG and salt (304-308). Adjusting the PEG and salt concentrations can precipitate variable DNA fragment size ranges (308-310). These SPRI beads are often also used for cleanup throughout the library preparation workflow (307, 308, 311). However, in applications where wide DNA fragment size distributions are undesirable, semi-automated preparative electrophoresis devices such as Labchip XT and Pippin Prep are preferably used (305, 307). Especially Pippin Prep is known to yield very narrowly distributed DNA fragment sizes (305, 307). Notedly, size selection may introduce a coverage bias consequent to the sequence specificity of the DNA fragmentation methods.

Target enrichment for WES is commonly performed using a hybridization capture-based approach, where specially designed probes capture DNA fragments from specific genomic regions. These approaches typically utilize streptavidin-coated magnetic SPRI beads to capture biotinylated probe-target hybrids (304, 311, 312) or microarrays of probes to capture target DNA fragments (312, 313). They can also be used in conjunction (312). These approaches share an initial step of library denaturation to generate ssDNA that enables probe-target hybridization, in addition to finals steps, including washing of the hybrid complexes to remove non-specifically bound DNA and target DNA elution by denaturation, where both denaturation steps usually are conducted by heat (311-313). The introduction of artifacts such as capture bias and off-target enrichment can occur consequent to factors including probe design, hybridization conditions, GC content, and secondary structures (314-340). Probes' length and target sequence complementarity can impact hybridization efficiency and specificity (315-317). Longer probes tend to have high efficiency and low specificity due to more hydrogen bonding opportunities, while shorter probes tend to have high specificity and low efficiency (315-320). In addition, probes with higher target sequence complementarity tend to have increased hybridization efficiency due to increased probe-target hybrid stability (318). In contrast, probes that are complementary to other non-target sequences tend to have reduced hybridization specificity (319, 321, 322). Insufficient target coverage of probes can also reduce efficiency (314). The hybridization conditions pertain to buffer composition, probe concentrations, library amounts, and incubation time and temperature and can also impact both hybridization efficiency and specificity (323-337). For instance, using additive organic solvents can improve the hybridization specificity by altering the melting point of DNA (326). In contrast, higher incubation temperatures generally result in greater specificity than lower incubation temperatures by preventing non-specific hybridization (327-329). Additionally,

sufficient incubation time is necessary for equilibrating, thus maximizing the hybridization efficiency (330, 331). Furthermore, as GC-rich sequences tend to have greater thermal stability than AT-rich sequences due to stronger hydrogen bonding, sufficient temperature and time are necessary to achieve complete denaturation, which otherwise can lead to underrepresentation of GC-rich sequences (332-334). The increased propensity of GC-rich sequences to fold into stable secondary structures can also contribute to their underrepresentation, as secondary structures can prevent proper hybridization (333-337). Conversely, the variable stability of probe-target hybrids, influenced by differing GC content, can lead to the overrepresentation of GC-rich sequences and the underrepresentation of AT-rich sequences (333).

Library enrichment is performed with PCR amplification, comprised of three repeated steps: library denaturation by heat, primer annealing at lower temperatures, and primer extension by a heat-stable DNA polymerase (307, 341). While PCR amplification is often necessary to ensure enough DNA fragments for the sequencing process, it can introduce artifacts such as amplification bias, PCR errors, and chimeras (342-344). Amplification bias pertains to the preferential amplification of certain sequences and is attributable to factors such as GC content and secondary structures (342-346). The increased thermal stability and propensity for forming stable secondary structures of GC-rich sequences can lead to their underrepresentation (342-346). Additives such as betaine and dimethyl sulfoxide can improve the amplification efficiency of GC-rich sequences by reducing the formation of secondary structures (347-353). However, they may consequently reduce the amplification efficiency of AT-rich sequences due to premature termination of the replication process by dissociation of the newly synthesized strand (342-344). PCR errors are replication errors and thus arise consequent to spontaneous mutagenesis and DNA polymerase fidelity (354). Formation of chimeras is thought to occur through template switching, a mechanism similar to strand-slippage, in which the DNA polymerase, upon partial primer extension or premature termination, disassociates from one template strand, reassociates with another template strand that has overlapping sequence complementarity, and continues the extension process (355-357). The frequency of chimera formation pertains to DNA polymerase processivity (344, 358). It depends on factors such as library size, library sequence similarities, library quality, number of amplification cycles, and duration of the extension step (358-360). In addition, DNA damage has been shown to promote template switching (361, 362). A PCR-free library preparation protocol can eliminate these artifacts (363), which may even apply to liquid biopsies, despite low cfDNA yields (363). Alternatively, subjecting the library to fewer PCR cycles and using DNA polymerases with en-

hanced fidelity and processivity can reduce the accumulation of artifacts (343-345, 364). In addition, microdroplet-based PCR enrichment has reportedly achieved coverage uniformity and reproducibility through an intricate workflow (365).

2.3.2 Illumina sequencing

On Illumina platforms, sequencing is conducted on flow cells in two primary stages: cluster generation and SBS (366-369). In the initial stage of cluster generation, the DNA fragments of the prepared library are immobilized by hybridization of the ligated adapter to a complementary oligonucleotide on the surface of the flow cell. Afterward, clonal clusters are generated through repeated bridge amplification cycles. Cluster generation is finalized by cleavage of the reverse strands (366-369). SBS is conducted through multiple cycles of incorporating proprietary modified dNTPs, which are fluorescently labeled and contain a reversible terminator for DNA polymerase, and fluorescent signal detection by optical imaging. Each cycle is ended by cleavage of the fluorescent dye and the synthesis-blocking terminator from the incorporated dNTP. Once the desired read length is achieved, the index of the forward strand is read by SBS. In paired-end sequencing, the SBS process is repeated for the reverse strand and its index (366-369). Upon sequencing completion, the images gathered for each sequencing cycle are processed to produce fluorescence-intensity measurements, which by default are inferred by the Bustard base calling algorithm (370, 371). However, multiple base calling algorithms have been developed that use different error modeling methodologies and therefore vary in performance (372). Studies have shown that artifacts introduced during library preparation and sequencing result in GC bias and strand bias (288, 291, 295, 343, 373-377). Notedly, artifacts introduced during the sequencing process are platform-specific and can vary dependent on the library preparation workflow. Cluster generation by bridge amplification may introduce biases similar to those from PCR amplification in library preparation, which can affect cluster densities (346, 374, 378, 379). However, differing cluster densities can also result from libraries having too widely ranging DNA fragment sizes (379, 380). Sequencing errors can arise due to the introduction of artifactual variants or artifacts that influence base calling. As previously mentioned, artifactual variants can arise due to DNA damage and DNA polymerase fidelity. Artifacts that may affect base calling include SBS chemistry, signal intensity decay, and color, spatial, and cycle crosstalk (381-391). While utilization of two-channel SBS chemistry in Illumina systems such as MiniSeq, NextSeq, and NovaSeq can generate data faster than four-channel SBS (392, 393), identification of Gs from signal absence can result in overcalling high

confidence Gs due to signal dropout artifacts (381). The fluorescence signal intensity decay can occur consequent to photobleaching (382, 394). However, varying decay rates and inefficient removal of fluorophores can complicate base calling (382, 394). Color crosstalk, or fluorescence bleed-through, refers to the overlaps of fluorophore emission spectra. Spatial crosstalk refers to the color crosstalk of adjacent or mixed clusters (383, 395). This phenomenon has been attributed to a large portion of erroneous base calls and has shown to be cluster-specific and often asymmetric (383). Cycle crosstalk refers to dephasing or asynchronous sequencing (382, 384-387). More specifically, phasing is when part of a cluster falls behind due to the incomplete removal of the terminator. In contrast, pre-phasing is when part of a cluster gets ahead due to the incorporation of multiple nucleotides in one cycle (382, 384-387). Sequence-specific errors have been shown to occur at inverted repeats, homopolymers, and certain motifs consequent to dephasing, DNA polymerase fidelity, and formation of stabile secondary structures (384, 385, 387, 389, 391). In contrast, position-specific errors that increase in frequency towards the end of reads have been attributed accumulation of sequencing artifacts (384, 387-390). Reverse reads have also been shown to have greater error rates than forward reads (389). Furthermore, nucleotide composition fluctuations at the beginning of reads may reflect the sequence specificity of DNA fragmentation (389). Multiplexed sequencing may also introduce an artifact called index crosstalk, or index hopping, where reads are incorrectly assigned to samples during demultiplexing, which can occur consequent to physical index transfer by contaminants such as free adapters and primers, the introduction of artifactual variants in indexes through PCR amplification errors and index misreading by SBS (249-252). However, quality filtering index reads and utilization of adapters with UMIs and UDIs have been shown to eliminate index hopping (396-400) effectively. Additionally, using adapters with UMIs, artifactual variants introduced during PCR amplification and sequencing are readily distinguishable from true variants, reducing the error rate and improving the detection of low-frequency variants in downstream data analysis (396-402). Computational error-correction methods have also proved efficient at eliminating sequencing errors (403).

2.4 Alignment of reads

A core step in NGS data analysis is read alignment. Multiple bioinformatic tools for read alignment with varying implemented mapping algorithms have been published, both open-source and proprietary (404-419). Modern alignment algorithms often utilize index-based approaches to efficiently align reads to a reference genome (420-424). Leveraging data structures such as suffix arrays and

suffix trees to build an index database of a reference genome enables fast and memory-efficient alignment (420-424) with DP (425, 426) or heuristic algorithms (427, 428). The alignment stringency is typically adjustable by various parameters, including penalties and cut-off thresholds (429, 430), which allow control of the trade-off between sensitivity and specificity. AS and MAPQ are calculated for each mapped read by the alignment tool to quantify the sequence similarity to the reference and the probability of misalignment, respectively (404-408, 420-422). Many alignment tools employ clipping, where part of a read is not aligned to the reference genome and is represented in the CIGAR string (420-424).

Two popular aligners, Bowtie2 (407) and BWA (404-406) utilize backward search with the FM-index (431, 432) of Burrows-Wheeler transformed (433) reference genomes for read mapping (404-407). Bowtie2 (407) uses an alignment methodology of seed-and-extend comprised of two steps: a seed-finding step that uses the FM-index to find exact matches to substrings of reads and a seed-extension step with a DP algorithm that can accommodate gaps and mismatches. While the literature does not specify which DP algorithm Bowtie2 has implemented, it does state that the algorithm uses an alignment scoring scheme similar to Needleman-Wunsch and Smith-Waterman (429). BWA consists of three modules with different algorithmic implementations: BWA-backtrack (404) uses the FM-index to find exact matches and matches with a small number of allowed mismatches, BWA-SW (405) uses the FM-index and the Smith-Waterman algorithm for more complex alignments that can include gaps, and BWA-MEM (406) that uses an alignment methodology similar to seed-and-extend comprised of two major steps: identification of MEMs of reads with the FM-index, and gap filling and alignment scoring with the modified Smith-Waterman algorithm.

A recently published aligner, Hierarchical Indexing for Spliced Alignment of Transcripts 2 (HISAT2) (408), has implemented a novel graph-based data structure for the representation of the human genome and a large collection of variants to better encapsulate the genetic diversity of humans and utilize search with the Hierarchical Graph FM-index of the graph-based data structure for read mapping. While the literature mentions that HISAT2 uses DP to align reads, it does not explicitly state its algorithmic implementation.

Benchmarking and comparison studies have shown alignment tools to vary in accuracy and efficiency due to differing algorithmic implementations and trade-off focuses (420-424). Reduced alignment accuracy can generally occur consequent to artifacts introduced during library preparation and sequencing, such as artifactual variants, chimeras, and GC bias, inherent limitations of the se-

quencing platform pertaining to read length, and reference bias (366, 373, 378, 420-422, 434-441). Especially GC bias has been shown to complicate genome assembly and alignment (366, 373, 378, 434, 437). Reads may align equally well to multiple genomic locations due to repetitive sequences in the reference genome, which can complicate the alignment of short reads (420-424). Furthermore, as the current human reference, GRCh38 (442), is a linear composite of haplotypes comprised mostly of sequences from a single individual (443), it does not appropriately represent the genetic diversity of humans. Therefore, alignment accuracies will vary dependent on the resemblance of samples and the reference genome (438-441). To better encapsulate the genetic diversity of humans, studies have suggested constructing a human pangenome reference with a graph-based data structure similar to the one utilized by HISAT2 (443-446). Most recently, a draft human pangenome reference has been published comprising 47 phased diploid assemblies from a cohort of genetically diverse individuals (447, 448). While transitioning to the human pangenome reference can reduce the reference bias (449), optimizing graph-based alignment algorithms' time and storage complexity is vital for improved efficiency (450, 451).

2.5 Somatic variant calling

A common type of genomic analysis is the detection of somatic variants by bioinformatic tools termed somatic variant callers. The somatic variant callers typically detect somatic variants in a paired analysis of aligned NGS reads from a tumor resection or biopsy sample and a matched germline sample, such as a peripheral blood mononuclear cells sample. However, many somatic variant callers support single-sample analysis (452-459). Some tools allow joint variant calling, where evidence is accumulated over multiple samples (454, 459, 460). Somatic variant callers capable of identifying INDELs and SVs often perform local de novo assembly and or realignment of poorly mapped reads, often based on clipping, by internal or third-party algorithms to improve the detection of these types of somatic variants (452-454, 456, 459). Most published somatic variant callers employ heuristic or probabilistic classification algorithms to detect somatic variants. HC calls are often generated through the post-filtration of somatic variants by applying feature and quality metric cutoff thresholds. Some variant callers utilize cross-contamination, strand bias, and sequencing error modeling for post-filtration of somatic variants and exclude germline variants from a PON and common germline variants from a curated database (452-459).

LoFreq (452) employs probabilistic sequencing error modeling as its statistical framework to detect

somatic variants. The algorithm can detect SNVs and INDELs, although, at the time of publication, only the detection of SNVs was implemented. When including the detection of INDELs, performing local realignment with the integrated algorithm is necessary. For the sequencing error modeling, each base at a genomic position is treated as arising from a Bernoulli trial, where reference and variant bases represent the success and failure outcomes, respectively. As each trial is assumed to be independent with an associated sequencing error probability derived from the base quality score, modeling the sequencing errors with a Poisson-binomial distribution allows distinct success probabilities for each variant base. Detected tumor-specific variants, which are significant in the tumor sample but not the germline sample, are further tested for inadequate read coverage in the germline sample using binomial tests with the VAFs from the tumor sample. Significant variants are found not to have arisen consequent to sequencing errors, whereas they are classified as somatic. The somatic variants are false positive filtered by strand bias, in which a two-tailed Fisher's exact test is used to test if the forward and reverse strand counts of the variant-base come from the same distribution as the reference base. Significant variants are removed as they have high strand bias. All the statistical tests are Holm–Bonferroni corrected for multiple testing with a significance level of 0.05 by default.

MuSE (453) can detect SNVs and short INDELs with its probabilistic framework. The workflow identifies candidate variants by applying seven heuristic pre-filters to the tumor and germline samples to exclude false positives likely resulting from sequencing artifacts. Somatic variants are probabilistically identified with the F81 Markov substitution model (461) built on the candidate variants, which describes the evolution from the reference allele to the allelic compositions of the tumor and germline samples, where Bayesian statistics are used to estimate the equilibrium allele frequencies and the evolutionary distance. The somatic variants are filtered with tiered cutoffs computed from a sample-specific error model, where the modeling methodology depends on the template of the data. With WGS data, a two-component Gaussian mixture model is fitted with the estimated equilibrium allele frequencies of the somatic variants in the tumor sample. With WES data, the estimated equilibrium allele frequencies of the somatic variants in the tumor sample are fitted to a beta distribution. The first cutoff tier is determined from the minimal sum of the false positive and false negative probabilities, where each additional tier reduces the cutoff stringency. The sample-specific error model incorporates information on common germline variants from a curated database by increasing the cutoff stringency of somatic variants with genomic positions included in the database.

MuTect2 (454) can detect both SNVs and INDELs, where a Bayesian classifier is used for proba-

bilistic classification of sites by analysis of the read counts of supporting alleles from the tumor and germline samples under two alternate error models: a reference model assuming the site contain no variant and any observed nonreference bases are caused by sequencing errors, and a variant model assuming the site contain a true variant allele in addition to sequencing errors. The models account for sequencing errors by incorporation of the base quality scores. Candidate variants are classified as somatic if the log-likelihood ratio of the models exceeds a predefined threshold. The somatic variants are filtered for false positives according to five heuristic criteria, models of strand and contamination bias, and by excluding germline variants from a generated PON.

SomaticSniper (455) is exclusively designed for the detection of SNVs. Initially, in the workflow, the genotype of each site in the tumor and germline samples is probabilistically determined by modeling each read and its associated mapping quality score using third-party software, MAQ (462). The likelihood of a site not being somatic also called the somatic score, is calculated with Bayesian inference from the genotype probabilities in the tumor and germline samples and case-specific prior probabilities, where sites exceeding a predefined somatic score threshold are classified as somatic. The identified somatic variants are subsequently false positive filtered according to six heuristic criteria.

Strelka2 (456) uses statistical and ML algorithms to detect SNVs, INDELs, and SVs. In the workflow, the reads of the tumor and germline samples are subjected to local de novo assembly and realignment with integrated algorithms. Subsequently, candidate variants are identified through probabilistic genotyping by haplotype modeling with read-associated quality scores to account for sequencing- and alignment-artifacts. The variant occurrence likelihood of the candidate variants is calculated with Bayesian inference from the tumor and germline VAFs and abstract noise terms that represent sequencing, alignment, and assembly errors. The somatic status of the candidate variants is determined from the occurrence likelihood and various associated variant features and quality metrics by a supervised ML classification model, specifically, a random forest model, which is pre-trained on various sequencing conditions. When the ML model is not applied, their somatic status is determined by cutoffs of a set of features.

VarDict (457) can detect SNVs, INDELs, and SVs with its heuristic methodology, in which candidate variants are identified directly by examining the reads in the tumor and germline samples simultaneous with local realignment of the reads by an integrated algorithm. A Fisher's exact test is performed on the VAFs of the candidate variants in the tumor and germline samples to determine if there is a significant difference. For the calculation of the VAFs, only bases with a defined minimum

quality score are regarded. Significant tumor-specific variants are classified as somatic and false positive filtered according to eleven heuristic criteria.

VarScan2 (458) can identify SNVs and INDELs. In the workflow, heuristic genotyping is initially conducted on independently generated pileup files of the tumor and germline samples, where only sites achieving set coverage, quality, and VAF thresholds are genotyped as either heterozygous or homozygous. To determine the variants' somatic status, a one-tailed Fisher's exact test is performed on the read counts of supporting alleles for each site with differing genotypes in the tumor and germline samples, where variants with significant p-values are classified as somatic if genotyped as homozygous in the germline sample. The somatic variants are false positive filtered according to nine heuristic criteria.

Benchmarking and comparison studies have shown somatic variant callers perform inconsistently between datasets and have low concordance due to technical and biological variations and differing algorithmic implementations for somatic variant detection and filtration (463-473). Consequently, it is considered impractical to find a single best-performing somatic variant caller (474). Simplistic consensus approaches can generate ensembles with improved specificity relative to individual variant calls, although sensitivity is sacrificed due to low concordance. The trade-off between sensitivity and specificity can be adjusted based on the level of concordance (473, 475). More complex approaches employ ML models to generate ensembles with superior sensitivity and specificity (474, 476-485). Most of these models are based on supervised ML algorithms, meaning they require training using ground truths commonly obtained from simulated or gold standard datasets (474, 478-485). However, simulations may inadequately encapsulate technical and biological variations, whereas the preferred solution is treating high-confidence calls from gold standard datasets as ground truths, which has proven to yield accurate and robust ML models (485).

In the past few years, bioinformatic tools such as MRDetect (486), INVAR (487), and DREAMS (488) have been developed for the detection of ctDNA in next-generation sequencing data by accumulating signals of patient-specific somatic variants. MRDetect enhances sensitivity and specificity using a pre-trained SVM model to quality filter reads from plasma samples. Contrarily, INVAR models position-specific and sequence-specific errors in patient-specific sequencing data, while DREAMS employs a pre-trained deep neural network error model.

3 Methods and materials

The samples used in this thesis were collected, prepared, and sequenced in a previous project (489) concerning the analysis of ctDNA to identify recurrence in cancer patients.

3.1 Patients and sample collection

Three patients treated for pancreatic cancer at Odense University Hospital, Denmark, were included. Tumor tissue samples were obtained from surgical resections, whereas blood samples were collected before treatment initiation and at the time of recurrence. The study (489), from which the samples originated, was approved by the Reginal Scientific Ethical Committee for Southern Denmark. Additionally, all patients provided written consent.

3.2 Sample preparation and sequencing

Plasma and germline DNA were obtained as described in the thesis (489). Samples were prepared for sequencing using the ThruPLEX Tag-Seq kit (Takara Bio), SeqCap EZ MedExome Target Enrichment Kit (Roche-Nimblegen), and SMARTer Unique Dual Indexes (Takara Bio) according to the manufacturer's instructions. The ThruPLEX Tag-Seq kit contains UMIs that allow for better distinction between artifactual variants and sequence variants. DNA quantities used for the library preparation were as described in the thesis (489). Libraries were sequenced on a NovaSeq 6000 (Illumina) using 2 × 150 bp paired-end reads flow cells (Illumina).

3.3 Alignment, postprocessing, and quality control analysis

The FASTQ files generated were aligned to the GRCh38 reference genome using BWA-MEM2 (v2.2.1) (406). Duplicate alignments were marked by GATK4 (v4.2.0.0) (490) MarkDuplicates. Subsequently, base quality score recalibration was conducted using GATK4 (v4.2.0.0) BaseRecalibrator, and dbSNP (491), 1000 Genomes Phase 1 SNPs (448), and Mills and 1000 Genomes Gold Standard Indels (492) as known sites variation, generating a recalibration table based on multiple covariates. The recalibration table was applied using GATK4 (v4.2.0.0) ApplyBSQR. Afterward, the alignment files were deduplicated according to UMI barcodes using Connor (v0.6.1) (493) and subjected to

another round of base quality score recalibration, yielding the final coordinate sorted alignment files in bam format. From the final alignment files, alignment quality metrics were calculated using GATK4 (v4.2.0.0) BaseRecalibrator, CollectHsMetrics, and CollectInsertSizeMetrics, in addition to Qualimap (v2.2.1) (494). All alignment quality metrics were aggregated using MultiQC (v1.10.1) (495). To simply the subsequent variant calling, bam files of tumor replicates were merged patient-wise using Sambamba (v0.8.0) (496). All plasma samples were filtered according to insert size (90-150 bp) to evaluate the effect in silico DNA fragment length selection has on ctDNA detection in plasma samples when recalling SNVs and CNVs later.

3.4 Single-nucleotide variant calling, postprocessing, and ensemble

For the identification of SNVs, seven variant callers were used with recommended parameters for exome data and in accordance with published documentation and workflows: LoFreq (v2.1.5) (452, 497), MuSE (v1.0.rc) (453, 498), GATK4 (v4.2.0.0) MuTect2 (454, 499), Somatic-Sniper (v1.0.5.0) (455, 500), Strelka (v2.9.10) (456, 501), VarDict-Java (v1.8.2) (457, 502) and VarScan2 (v2.4.4) (458, 503). Calling variants with LoFreq, MuSE, and Strelka was straightforward, as all operations were handled internally. MuTect2 relied on multiple filtrations approaches as specified in the 'somatic short variant discovery' workflow (504), where a PON initially was generated from the germline samples of all three patients and where the raw variants were filtered using a generated cross-sample contamination table and a read orientation model. Somatic-Sniper utilized multiple scripts for variant filtration and various preparation steps, most of which were written in Perl, in addition to samtools (v0.1.16) (505) pileup and bam-readcount (v0.8) (506). VarDict-Java employed an R script for variant filtration and a Perl script for conversion of variants into the standardized vcf format. For variant calling with VarScan2, samtools (v1.9) mpileup was initially used to generate pileups of the matched tumor and germline samples that were used to call variants. The variants called by VarScan2 were filtered for clusters and INDELs. In addition to variant caller specific postprocessing of variants, all triallelic sites were removed using bcftools (v1.12) (505). Two consensus ensembles of SNVs were generated with SomaticSeq (v3.6.2) (479) with minimum mapping quality and minimum base quality set to 10 and 30, respectively. SomaticSeq additionally filtered out common germline variants reported in dbSNP. The first ensemble included calls by LoFreq, MuTect2, and Strelka, where only variants called by all three variant callers were passed. The second, third and fourth ensemble included calls by all seven variant callers where variants called by five, four, and three or more of the variant

callers were passed, respectively. Readcount files were generated from both the ensemble and the variant caller outputs and the matched tumor and germline samples using bam-readcount (v0.8) with minimum mapping quality set to 10 and minimum base quality set to 30. The readcount files were used by vcf-annotation-tools (v3.0.0) (507) vcf-readcount-annotator to annotate the SNVs with corrected AFs.

3.5 Recalling single-nucleotide variants in plasma samples and calculation of Z-scores

Three approaches were used to evaluate the presence of ctDNA through the detection of SNVs in both the filtered and non-filtered plasma samples. Two of the approaches use bam-readcount (v0.8) to generate read count files at the genomic positions of the SNVs from vcf files for each plasma bam file. Two read count files were generated for each plasma bam file; one using default parameters and another with minimum mapping quality set to 10 and minimum base quality set to 30. The read count files were supplied to vcf-annotation-tools (v3.0.0) vcf-readcount-annotator to annotate the SNVs in the vcf files with VAFs from the plasma samples. In the first approach, the mean VAFs were used for statistical evaluation, denoted MeanVAF, while the proportion of VAFs greater than 0 were used for statistical evaluation in the second approach, denoted BinaryVAF. In the third approach, MRDetectSNV (486) was used to recall SNVs in plasma samples, where supplied scripts written in Python2 were utilized for pulling candidate reads from the plasma samples, quality scoring the reads using a pre-trained SVM model, and filtering the reads. In all three approaches, the Z-scores were calculated from the mean and standard deviation of the control samples' SNV signals.

3.6 Copy number analysis and postprocessing

Copy number variations were detected using a VarScan2 (v2.4.4) pipeline. Initially, the data ratio between the merged tumor and matched germline bam files was calculated using samtools (v1.12) flagstat outputs. Additionally, the merged tumor and matched germline bam files were used to generate pileup files using samtools (v1.12) mpileup with minimum alignment quality set to 10, minimum base quality set to 30, and probabilistic realignment disabled. The generated pileup files were processed by VarScan2 (v2.4.4) copynumber with default settings and parameters, except for the calculated data ratio. The outputted copy number segments were filtered for low coverage, excluding segments with a

tumor or normal coverage below 20. Subsequently, the filtered copy number segments were processed by VarScan2 (v2.4.4) copyCaller with default settings and parameters to adjust for GC content and apply amplification and deletion thresholds of log2 ratios of 0.25 and -0.25, respectively. A recenter amount was calculated from the mean log ratio by chromosome and applied in a second round of processing of the filtered copy number segments by copyCaller. Centromere locations were excluded in a filtering process, and circular binary segmentation was performed using the R (v3.6.1) package DNAcopy (v1.60.0) (508) to segment the copy number data into regions of equal copy number and identify genomic regions with abnormal copy number. The segments were further post-processed by recentering the segments according to the median log2 ratio and by adjusting the classification thresholds of the amplifications from 0.25 to 0.3 and deletions from -0.25 to -0.3. The CNVs determined with the VarScan2 were corroborated by CNV profiles of the merged tumor resection samples obtained through an ichor-TITAN workflow (509) that combined the results of ichorCNA (v0.1.0) (510) and TitanCNA (v.1.23.1) (511). The workflow utilized HMMcopy (v1.42.0) (512) for copy number prediction with correction for GC and mappability biases and was executed with default parameters.

3.7 Recalling copy number variations in plasma samples and calculation of Z-scores

A single approach, MRDetectCNA (486), was used for the detection of ctDNA by recalling CNVs in both the filtered and non-filtered plasma samples. Initially, a PON was generated by merging the three germline samples from each patient using Sambamba (v0.8.0). The coverage was determined at each genomic position of interest for the PON and plasma samples using GATK (v3.4.0) DepthOfCoverage with the BadCigar read filter, including reads with deletions. The CNV and neutral intervals were divided into non-overlapping 500 bp bins, and each bin's median coverage was calculated. To allow comparison of samples, the coverage of each bin was normalized by the average coverage of the sample. The average coverage was calculated using samtools (v1.12) depth. Each sample was furthermore subjected to a robust Z-score normalization to account for sample-specific variations in coverage. The CNV signal was calculated as the sum of the differential (plasma - PON) coverage of each bin multiplied by the directionality of the CNV region (+1 for amplifications and -1 for deletions), while the neutral regions serve as background noise. Lastly, Z-scores were calculated for each plasma sample

from the mean and standard deviation of the control samples' CNV signals.

3.8 Combination of statistical scores

The SNV-based ctDNA detection approaches were considered dependent tests, as they utilized similar detection signals for the calculation of statistical scores. An appropriate method for combining their statistical scores would be the harmonic mean p-value (513) that inversely weighs the right-tailed p-values calculated from their Z-scores. The best-performing ensemble was selected for the combination of statistical scores, including both the statistical scores calculated from the signals of the nonfiltered and quality-filtered reads from the MeanVAF and BinaryVAF approaches, in addition to the statistical scores of the MRDetectSNV approach. The statistical tests of the CNV-based ctDNA detection approach, MRDetectCNV, on the total CNV signal, in addition to the separate signals of the amplifications and deletions, were considered dependent tests, whereas the right-tailed p-values calculated from their Z-scores were combined with the harmonic mean p-value method. Both the statistical scores of the raw and post-processed CNV identified by VarScan2 were included. The SNV- and CNV-based ctDNA detection approaches were considered independent tests, as they utilized different signals for the calculation of the statistical scores, and as SNVs and CNVs arise through different mutagenic mechanisms, whereas the product of their probabilities calculated with the harmonic mean p-value method was determined to appropriately represent their joint probability.

4 Results and discussion

4.1 Data quality control

The quality of the sequencing data was examined by gathering general statistics from the final quality control report (Table 1). The GC content for all samples ranged between 50-54% (Table 1). The fraction of the genome with at least 30X and 50X coverage was between 97.4-99.5% and 92.6-99.1%, respectively (Table 1). The median coverage for the germline samples varied from 125X to 134X, while the tumor resection and plasma samples had a median coverage ranging from 151X to 213X and 194X to 376X, respectively (Table 1). All samples, except PC1-18-op-P12 (99.3%), aligned 99.7% or 99.8% of their total sequences (Table 1). Enrichment of the tumor resection and germline samples was observed at 35-38 fold, whereas the plasma samples displayed 44-47 fold enrichment of baited regions (Table 1). The percentage of duplicate reads ranged between 29.4-31.1% for the

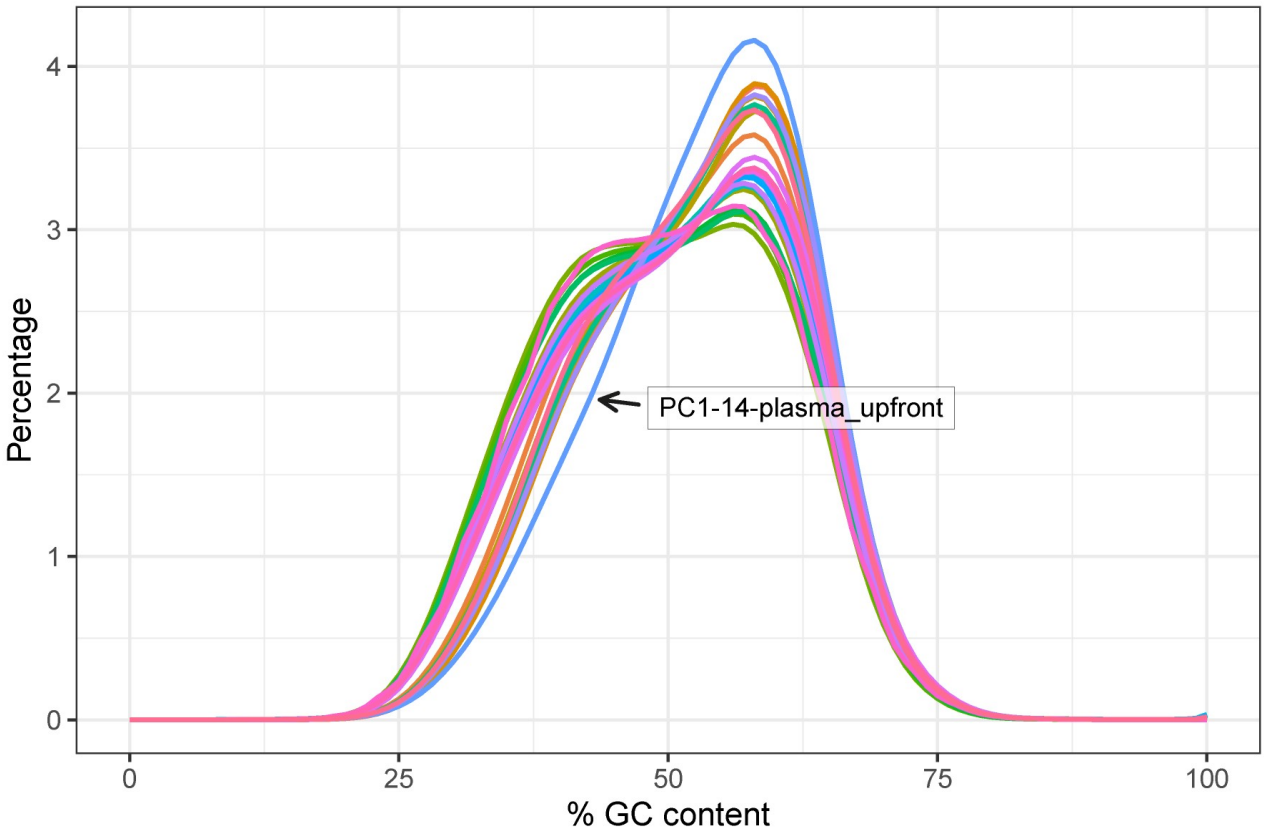

Figure 1: Plot illustrating the average guanine-cytosine content of reads generated with FastQC. Each color represents a different sample. Outliers are pointed out and labeled.

Sample Name	% GC	≥ 30X	≥ 50X	Coverage	% Aligned	Fold Enrichment	% Dups	M Seqs
ECV2-29-plasma_neoadj	53%	99.30%	98.60%	292.0X	99.80%	46	46.70%	175.4
ECV2-31-plasma_neoadj	52%	99.50%	99.10%	376.0X	99.80%	44	51.40%	251.2
ECV2-35-plasma_neoadj	54%	99.10%	98.50%	318.0X	99.80%	46	52.00%	212.9
ECV2-4-plasma_neoadj	53%	99.20%	98.40%	273.0X	99.70%	47	36.70%	137
ECV2-8-plasma_neoadj	53%	99.40%	98.80%	304.0X	99.70%	44	47.80%	197.6
PC1-10-germline	52%	97.40%	92.60%	125.0X	99.70%	37	29.90%	69
PC1-10-op-P1	50%	98.20%	96.30%	168.0X	99.80%	36	37.80%	105.2
PC1-10-op-P2	51%	97.90%	95.40%	151.0X	99.70%	35	31.80%	88.3
PC1-10-op-P3	51%	98.50%	97.20%	199.0X	99.80%	36	35.80%	122.2
PC1-10-op-P4	51%	98.40%	96.40%	162.0X	99.80%	36	33.40%	95.1
PC1-10-plasma_upfront	53%	99.20%	98.60%	294.0X	99.80%	45	56.10%	212.9
PC1-14-germline	52%	97.90%	93.90%	134.0X	99.70%	37	29.40%	74.2
PC1-14-op-P5	52%	98.40%	95.50%	159.0X	99.70%	36	30.30%	92.4
PC1-14-op-P6	52%	98.70%	96.60%	176.0X	99.70%	36	30.10%	101.5
PC1-14-op-P7	52%	98.40%	95.50%	163.0X	99.70%	36	30.30%	94.4
PC1-14-op-P8	52%	98.60%	96.70%	179.0X	99.80%	36	32.60%	106.5
PC1-14-plasma_upfront	54%	98.40%	94.60%	194.0X	99.70%	45	66.30%	199
PC1-14-plasma_recurrence	53%	99.20%	98.00%	205.0X	99.80%	45	62.10%	175.8
PC1-18-germline	52%	97.70%	93.20%	126.0X	99.80%	37	31.10%	70.4
PC1-18-op-P10	53%	98.80%	96.60%	168.0X	99.70%	37	29.50%	96.3
PC1-18-op-P11	53%	99.00%	97.80%	212.0X	99.80%	36	34.40%	135.5
PC1-18-op-P12	52%	99.20%	98.30%	213.0X	99.30%	38	32.20%	169.7
PC1-18-op-P9	53%	98.70%	96.70%	175.0X	99.70%	37	31.10%	106.1
PC1-18-plasma_upfront	53%	99.30%	98.70%	291.0X	99.70%	46	53.50%	199.5

Table 1: General statistics of included samples containing guanine-cytosine content in percentage (% GC), fraction of genome with at least 30X coverage (≥ 30X), fraction of genome with at least 50X coverage (≥ 50X), median coverage (Coverage), percentage mapped reads (% Aligned), fold enrichment (Fold Enrichment), estimated percentage duplicate reads (% Dups), and total sequences in millions (M Seqs).

germline samples, 29.5-37.8% for the tumor resection samples, and 36.7-66.3% for the plasma samples (Table 1). The total sequences in millions were 69-74.2 for the germline samples, 88.3-169.7 for the tumor resection samples, and 137-251.2 for the plasma samples (Table 1).

To evaluate whether the GC content of the samples was at an acceptable level, a bed file containing genomic regions covered by the target enrichment kit covered was used to approximate the expected GC content from the reference genome. The reference genome's GC content was approximated to 51%. All the samples' GC contents were considered within an acceptable range from the reference (Table 1), taking into account technical and biological variations. However, the upfront PC1-14 plasma sample was identified as a potential outlier upon examination of the samples per sequence GC content (Figure 1).

The fraction of the genome with at least 30X and 50X coverage, and median coverage, was deemed acceptable as the sample coverage distributions formed Poisson-like distributions (Figure 2). However, the upfront PC1-14 plasma sample was again identified as an outlier (Figure 2).

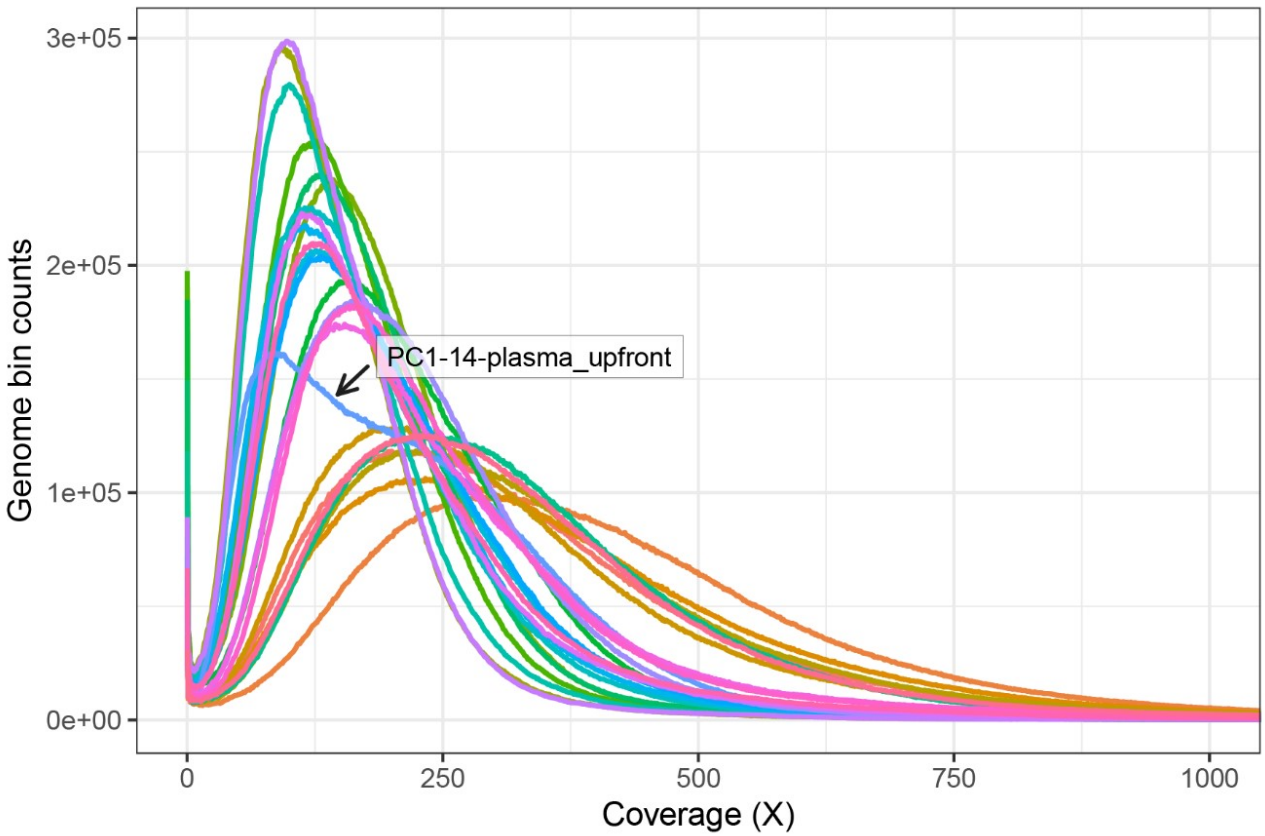

Figure 2: Plot illustrating sample coverage distributions generated with Qualimap BamQC. Each color represents a different sample. Outliers are pointed out and labeled.

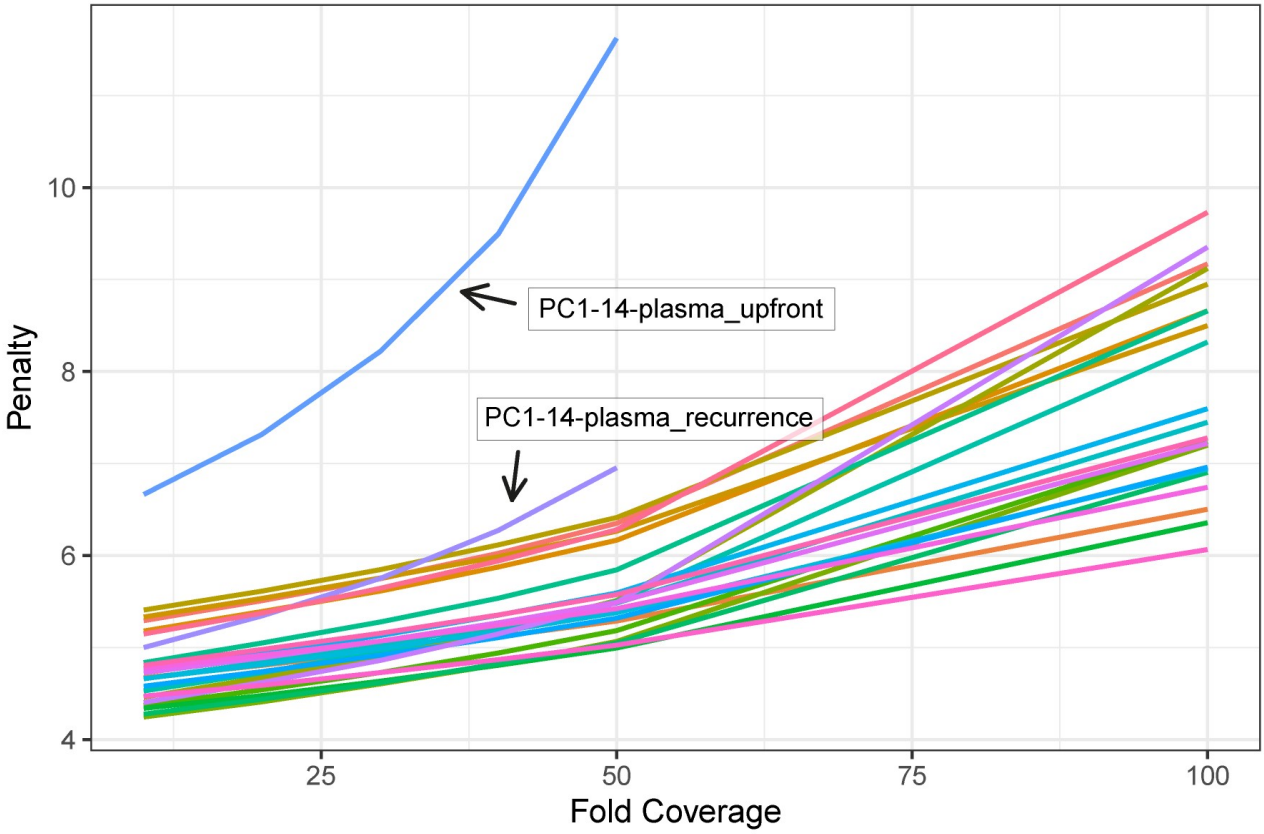

Figure 3: Plot illustrating the "hybrid selection penalty" incurred to get 80% of target bases to a given coverage. Data generated with Picard. Each color represents a different sample. Outliers are pointed out and labeled.

Examination of the hybrid selection penalty generated by Picard also marked the PC1-14 plasma recurrence sample as an outlier, as both the PC1-14 plasma upfront and recurrence samples only reached 50-fold coverage (Figure 3). However, the PC1-14 plasma upfront sample showed higher penalties at the given fold coverages than the PC1-14 plasma recurrence (Figure 3).

Upon analyzing the percentage of mapped reads for each sample, successful alignment was confirmed. However, it should be noted that a previous study found a positive relationship between the percentage of mapped reads and the false alignment rate (406). Misalignments could become of concern when detecting somatic variants in a paired analysis of the tumor resections and germline samples and when recalling somatic variants in plasma samples. The false alignment rate could be reduced by increasing the alignment stringency, which would reduce the percentage of mapped reads. The fold enrichment, or the fold by which the baited regions were amplified above the genomic background, was greater for the plasma samples than the tumor resections and blood samples. This

could be explained by the insert size distribution of the samples, where the plasma samples have smaller insert size distribution centers than the tumor resections and blood samples (Figure 4). Larger fragments could potentially lead to greater genomic background, as larger parts of the fragments might be mapped outside of the baited regions. Upon further examination of the insert size distributions, both the PC1-14 plasma upfront and recurrence samples displayed fewer reads around the distribution center than the other plasma samples (Figure 4), which could indicate the cfDNA was contaminated with genomic DNA or that a greater proportion of the cfDNA derived from necrotic tumor cells or polynucleosomes (31), thus they were marked as outliers. Genomic DNA contamination might negatively impact downstream analyses but might be rectifiable by *in silico* size selection of the plasma samples' reads. DNA obtained from tumor resections and blood samples were subjected to ultrasonic fragmentation, whereas the plasma cfDNA was shed fragmented as part of the cellular degradation, accounting for the clear differences in insert size distributions (Figure 4).

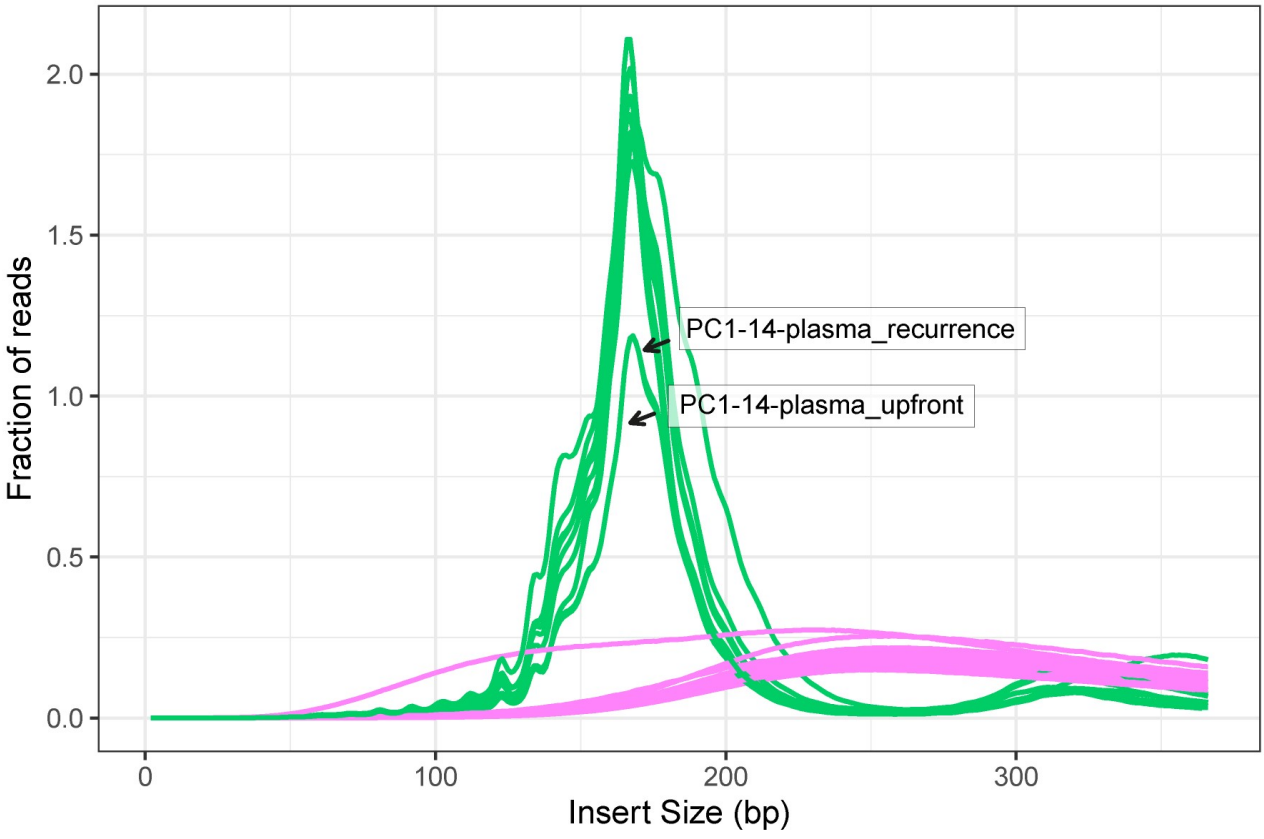

Figure 4: Plot illustrating insert size distributions. Data generated with QualiMap. Plasma samples are colored green, whereas tumor resections and blood samples are colored magenta. Outliers are pointed out and labeled.

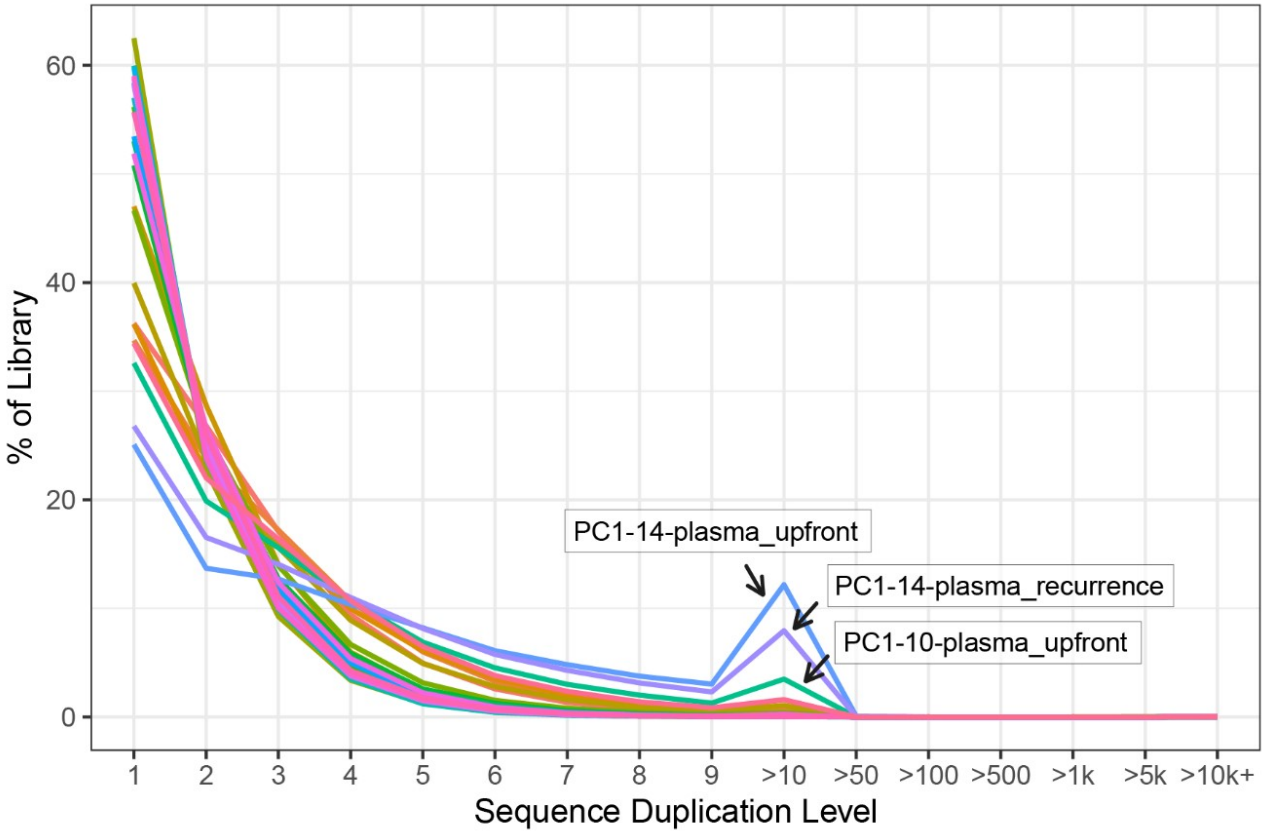

Figure 5: Plot illustrating the relative level of duplication found for every sequence. Data generated with FastQC. Each color represents a different sample. Outliers are pointed out and labeled.

The number of duplicate reads in plasma samples was higher than in tumor resections and blood samples, potentially indicating low input DNA amounts (Figure 5). As the plasma samples probably had a low input DNA amount, more PCR rounds were required to reach the desired coverage. Three samples were flagged as potential outliers for sequence duplication: the PC1-14 plasma upfront and recurrence samples in addition to the PC1-14 plasma upfront sample (Figure 5). The PC1-14 plasma upfront and recurrence samples had high sequence duplication levels for a larger proportion of the library than the rest of the samples (Figure 5). The PC1-10 plasma upfront sample was flagged due to its high sequence duplication level at >10 (Figure 5). This suggested low cfDNA input amounts.

The mean base quality score across each position in the reads (Figure 6A) and the distribution of the reads' mean sequence base quality score (Figure 6B) were used to examine the samples for sequencing artifacts. All samples exhibited great base quality metrics, eliminating the need for further investigation. However, the PC1-18-op-P12 tumor resection sample was marked as an outlier from its deviating quality scores. The batch effect was disregarded, as the samples were analyzed together.

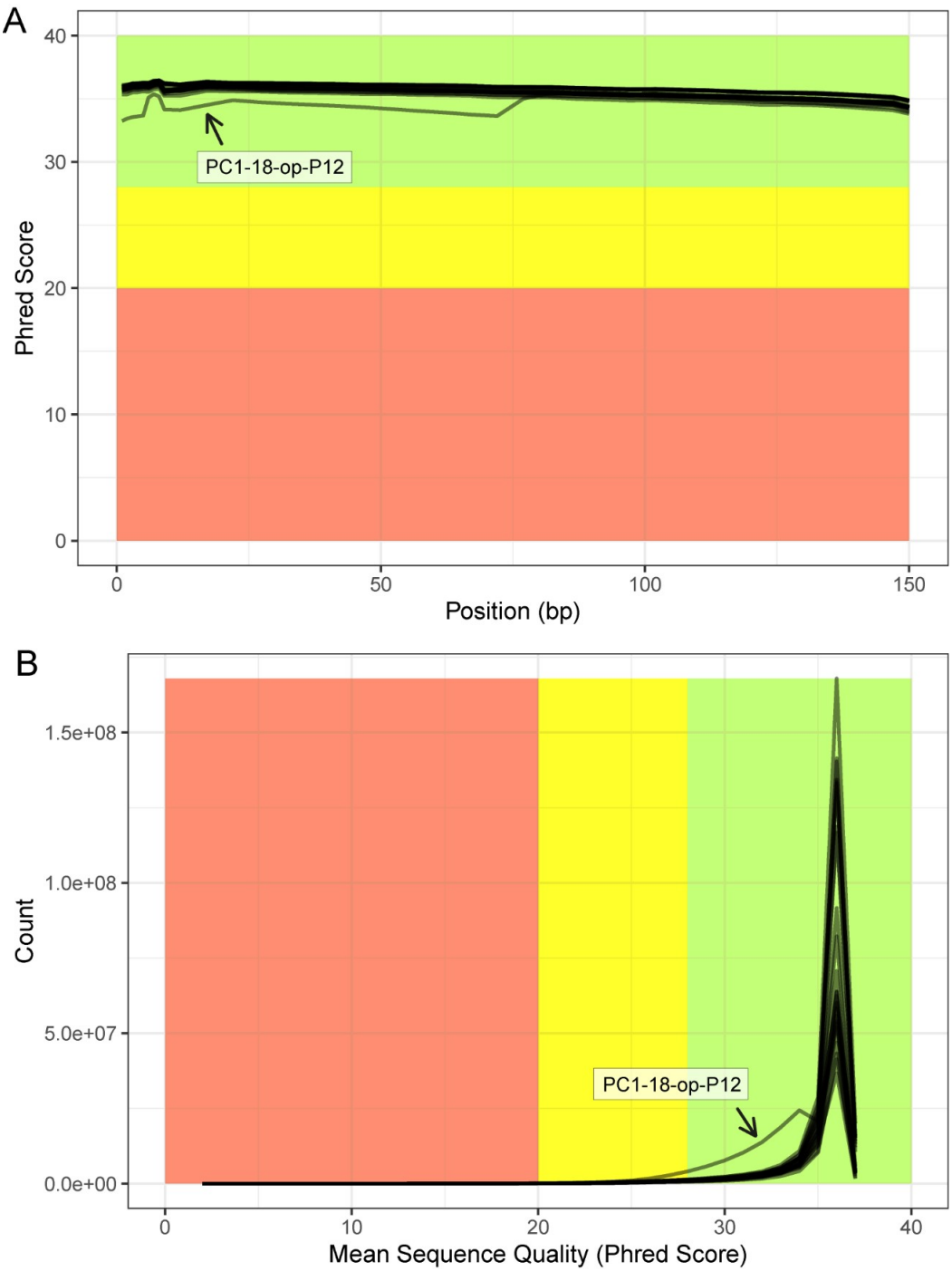

Figure 6: Plots illustrating sequence quality. Data generated with FastQC. All samples are colored black. Phred Scores greater than 28 are considered good. Outliers are pointed out and labeled. **(A)** The mean quality value across each base position in the read. **(B)** The number of reads with average quality scores.

Sample Name	Median VAF (% TC)	ichorCNA-TitanCNA (% TC)
PC1-10-op	63.2	54.13
PC1-14-op	58.5	49.35
PC1-18-op	65.4	54.10

Table 2: Approximated tumor contents of merged tumor resection samples determined from the median VAF of the ensembles and through an ichorCNA-TitanCNA pipeline.

Two approaches were used to approximate the tumor content in the surgical resections: the median VAF of the SNVs identified by the ensembles and an ichorCNA-TitanCNA pipeline (Table 2). The tumor content was estimated to be 63.2%, 58.5%, and 65.4% from the median VAF of the SNVs identified by the ensembles for PC1-10, PC1-14, and PC1-18, respectively (Table 2). The ichorCNA-TitanCNA pipeline estimated the tumor content to be 54.13%, 49.35%, and 54.10% for PC1-10, PC1-14, and PC1-18, respectively (Table 2). The tumor content was lower for all the merged tumor resection samples when approximated by ichorCNA-TitanCNA, relative to when determined from the median VAF of the SNVs identified by the ensembles. This discrepancy could be due to copy number variants, where genomic regions might deviate from 2N. The ichorCNA-TitanCNA pipeline considered ploidy when approximating tumor content. Thus, its tumor content estimations might be closer to the true tumor contents than those determined from the median of VAF of the SNVs identified by the ensembles.

4.2 Evaluation of variant callers and ensembles

The performance of the seven variant callers was initially evaluated by examining the total and unique number of identified SNVs, in addition to the number of identified SNVs with a VAF greater than zero in the germline samples (Table 3). Most of the included variant callers could call both SNVs and INDELs; however, as INDELs were not recalled in the used ctDNA detection approaches, only the identified SNVs were examined. As shown in Table 3, MuSE, Strelka, and VarDict consistently identified the most SNVs, although the majority of the SNVs were not identified by any of the other variant callers. The SNVs identified by MuTect2 and VarScan2 had similar, relatively low percentages of unique SNVs across all patients, despite VarScan2 identifying three times more SNVs than MuTect2 (Table 3). LoFreq identified few SNVs but had the lowest percentage of unique SNVs of all the variant callers in all patients (Table 3). SomaticSniper identified the fewest SNVs of all the variant callers in all three patients while having among the greatest percentage of unique SNVs (Table 3). Additionally,

	PC1-10					PC1-14					PC1-18				
	Total	Unique		VAF_G > 0		Total	Unique		VAF_G > 0		Total	Unique		VAF_G > 0	
	(#)	(#)	(%)	(#)	(%)	(#)	(#)	(%)	(#)	(%)	(#)	(#)	(%)	(#)	(%)
LoFreq	120	6	5	7	5.8	80	3	3.8	2	2.5	55	9	16.4	2	3.6
MuSE	803	345	43	21	2.6	624	314	50.3	15	2.4	553	298	53.9	19	3.4
MuTect2	174	31	17.8	29	16.7	112	31	27.7	20	17.9	98	32	32.7	24	24.5
SomaticSniper	32	25	78.1	23	71.9	18	14	77.8	14	77.8	18	14	77.8	8	44.4
Strelka	1345	858	63.8	173	12.9	1306	992	76	234	17.9	1368	1131	82.7	329	24
VarDict	955	772	80.8	336	35.2	771	641	83.1	256	33.2	899	821	91.3	323	35.9
VarScan2	530	110	20.8	36	6.8	326	96	29.4	34	10.4	292	71	24.3	21	7.2

Table 3: Total and unique number of identified SNVs, in addition to the number of identified SNVs with a VAF greater than zero in the matched germline sample, for each variant caller and patient.

LoFreq and MuSE had the lowest proportion of SNVs with a VAF greater than zero in the germline samples, followed by VarScan2. Strelka and MuTect2 had a similar modest proportion of SNVs with a VAF greater than zero in the germline samples across all patients, even though Strelka identified up to approximately 13 times more SNVs than MuTect2 (Table 3). Approximately a third of the SNVs identified by VarDict had a VAF greater than zero in the germline samples in the three patients (Table 3). SomaticSniper had the greatest proportion of SNVs with a VAF greater than zero in the germline samples in all patients (Table 3).

Ideally, when evaluating the performance of variant callers, the identified SNVs should be compared to a list of true positives. However, as true positives are virtually impossible to distinguish from false positives in real data definitively, it was assumed that the false positive rate would increase with the proportion of unique SNVs and SNVs with a VAF greater than zero in the germline samples. From these assumptions regarding the false positive rate, LoFreq, MuSE, MuTect2, and VarScan2 were evaluated to have the highest specificities, followed by Strelka and VarDict. Additionally, SomaticSniper was evaluated to have the lowest specificity and could be disregarded in future studies, especially considering its poor software optimization and, thus, long processing time. It should be noted that SomaticSniper was one of the first variant callers that was developed (455), which could explain the poorer performance. Furthermore, the differing total and unique number of SNVs identified by each variant caller were explained by varying statistical and classification algorithms employed for variant detection, in addition to differences in the features and quality metrics upon which the variants were filtered.

For each variant caller, the VAFs of their identified SNVs were recalculated from read counts with fixed quality metric thresholds in the merged tumor resection and matched germline samples and plotted in histograms for PC1-10 (Figure 7) to support the initial evaluation of the variant callers. LoFreq, MuTect2, and VarScan2 were mostly sensitive to common variants, as they predominantly identified SNVs distributed around a VAF of 0.3 in the tumor sample (Figure 7). The SNVs unique to LoFreq and MuTect2 primarily had VAFs of approximately 0.05 in the tumor sample, while the SNVs unique to VarScan2 were distributed in three clusters around VAFs of 0.05, 0.2, and 0.5 in the tumor sample (Figure 7). The SNVs identified by LoFreq and MuTect2 with a VAF greater than zero in the germline sample were sporadically distributed in the tumor sample, although for MuTect2, some were clustered around a VAF of 0.05 in the tumor sample (Figure 7). For VarScan2, the SNVs with a VAF greater than zero in the germline sample were clustered around a VAF of 0.25 in the tumor sample

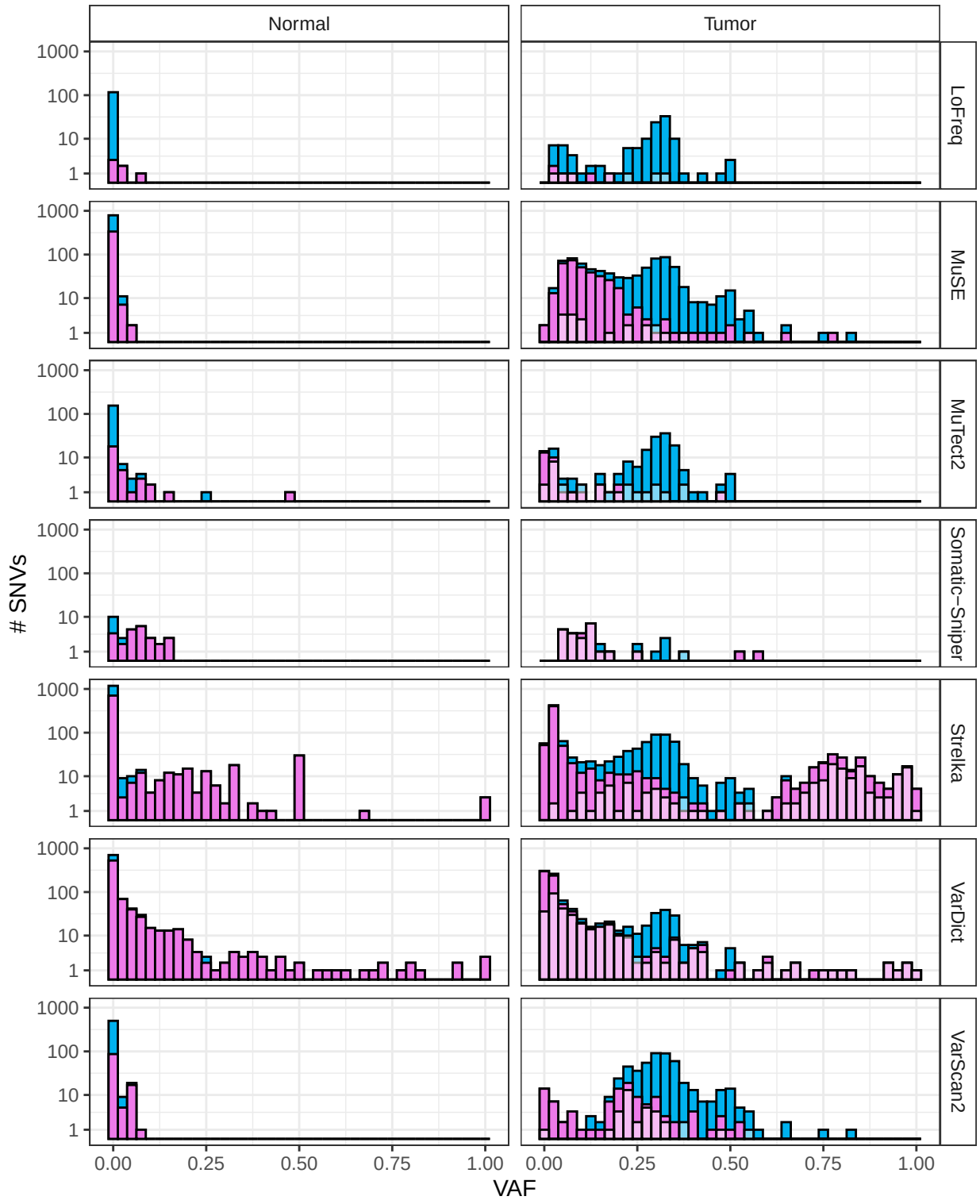

Figure 7: Histograms of the VAF of the SNVs for each variant caller in the merged tumor resection sample, tumor, and the matched germline sample, normal, for the patient, PC1-10. SNVs identified by the variant caller are color blue, whereas SNVs unique to the variant caller is colored magenta. SNVs with a VAF greater than zero in the matched germline sample are plotted on top with a transparent white color, and thus appear in a lighter magenta and blue color. The VAFs are calculated from reads with a minimum mapping quality of 10 and a minimum base quality of 30. The y-axis is pseudo $\log_{10}$ scaled to properly illustrate the SNV counts of each bin.

(Figure 7). The SNVs identified by LoFreq all had VAFs less than 0.05 in the germline sample (Figure 7). The SNVs identified by MuTect2 and VarScan2 mostly had VAFs less than 0.1 in the germline sample, while few of those identified by MuTect2 had greater VAFs (Figure 7). MuSE, SomaticSniper, Strelka2, and VarDictJava displayed greater sensitivity towards low-frequency variants than LoFreq, MuTect2, and VarScan2, as many, if not most of their identified SNVs were distributed around a VAF of 0.05-0.1 in the tumor sample, while they additionally identified common variants distributed around VAF centers of 0.3 and 0.5 in the tumor sample (Figure 7). Strelka additionally identified many high-frequency SNVs clustered around a VAF of 0.8 in the tumor sample (Figure 7). The majority of the SNVs unique to MuSE, SomaticSniper, Strelka2, and VarDictJava were part of their low-frequency clusters (Figure 7). Additionally, all the high-frequency SNVs identified by Strelka2 were unique (Figure 7). The SNVs identified by MuSE with a VAF greater than zero in the germline sample were clustered around VAFs of 0.1 and 0.3 in the tumor sample (Figure 7). At the same time, for SomaticSniper, they were primarily distributed around a VAF of 0.1 in the tumor sample (Figure 7). In the germline sample, the SNVs identified by MuSE mostly had VAFs less than 0.05, while those identified by SomaticSniper had VAFs up to 0.2 (Figure 7). Strelka2 and VarDictJava showed wide VAF distributions in the tumor sample of their SNVs with a VAF greater than zero in the germline sample (Figure 7). However, the SNVs predominantly had VAFs ranging from 0.6 to 1.0 for Strelka2 and from 0.0 to 0.25 for VarDictJava (Figure 7). In the germline sample, the VAF distributions of the SNVs identified by Strelka2 and VarDictJava were equally wide, although most of their SNVs had VAFs less than 0.25 (Figure 7). Similar VAF distributions were observed for PC1-14 (Figure A.1) and PC-18 (Figure A.2).

While the impact of the variant callers' use of different statistical and classification algorithms for the detection of somatic variants was observable already from the initial evaluation, it became even more evident from their distinct VAF distribution patterns. These patterns were used for further evaluation, in conjunction with the initial assumptions regarding the false positive rate, as they reflected the sensitivity and specificity of each variant caller. The initial assumptions were thought to remain applicable, as the VAF distributions of the unique SNVs and the SNVs with a VAF greater than zero in the germline samples commonly overlapped for each variant caller. Thus, LoFreq, MuSE, MuTect2, and VarScan2 were evaluated to have the highest specificities, whereas MuSE, Strelka2, VarDictJava, and VarScan2 were evaluated to have the highest sensitivities. SomaticSniper was determined to have both the lowest sensitivity and specificity, where its SNVs had suspiciously high VAFs in the germline

sample relative to the tumor sample. Strelka2 and VarDictJava were deduced to have amongst the lowest specificities, only superseded by SomaticSniper, as they had such wide VAF distributions in the germline samples. It should be noted that false positive variant calls generally could arise from a combination of technical and biological factors, such as sequencing, alignment and bioinformatic artifacts, cross-contamination, normal cell contamination, and tumor heterogeneity. However, many sequencing artifacts should have been rectified during deduplication with UMIs. LoFreq, Strelka2, and VarDict attempted to correct alignment artifacts by performing local de novo assembly and or realignment of suspected misaligned reads. This could partially explain the wide VAF distributions of SNVs identified by Strelka2 and VarDictJava in the germline samples, as the VAFs were recalculated from aligned reads that had not been subjected to the variant callers local de novo assembly and realignment algorithms. While this also was the case for the aligned reads of the plasma samples, the recalculated VAFs were thought to better reflect recalling somatic variants in plasma samples. From the analysis of sequencing depth distributions of the SNVs identified by each variant caller, it was evident that sequencing and alignment artifacts contributed to false positive variant calls in LoFreq, MuSE, MuTect2, SomaticSniper, and VarDictJava, as their identified SNVs with a VAF greater than zero in the germline samples tended to have high coverage in both samples. However, the SNVs identified by Strelka2 and VarScan2 with a VAF greater than zero in the germline samples had low coverage in the germline samples, whereas germline variants could have been misclassified as somatic. Each variant caller could be optimized by increasing the stringency of their heuristic filtration criteria or through further post-processing of their identified variants by applying additional universal features and quality metric filters. However, this was not explored, as it was deemed beyond the scope of this thesis. Instead, the applied approach for false positive filtration was consensus ensembles.

Similar to the variant callers, the consensus ensembles were initially assessed by examining the total and unique number of identified SNVs and the number of SNVs with a VAF greater than zero in the germline samples (Table 4). As shown in Table 4, all three SomaticSeq ensembles contained more SNVs then the MRDetect ensemble in all patients. Additionally, increased stringency of the SomaticSeq ensembles reduced the total number of identified SNVs, while it increased the number of SNVs with a VAF greater than zero in the germline samples (Table 4). Interestingly, the total number of identified SNVs and the number of SNVs with a VAF greater than zero in the germline samples for each of the ensembles differed greatly between patients (Table 4).

Considering the probability of ctDNA detection in plasma samples has been shown to increase

	PC1-10					PC1-14					PC1-18				
	Total	Unique		VAF_G > 0		Total	Unique		VAF_G > 0		Total	Unique		VAF_G > 0	
	(#)	(#)	(%)	(#)	(%)	(#)	(#)	(%)	(#)	(%)	(#)	(#)	(%)	(#)	(%)
MRDetect	93	0	0	3	3.2	54	0	0	1	1.9	42	0	0	1	2.4
SomaticSeq5	125	0	0	3	2.4	58	0	0	1	1.7	53	0	0	1	1.9
SomaticSeq4	142	0	0	8	5.6	71	0	0	1	1.4	55	0	0	1	1.8
SomaticSeq3	152	10	6.6	9	5.9	82	11	13.4	2	2.4	57	2	3.5	1	1.8

Table 4: Total and unique number of identified SNVs, in addition to the number of identified SNVs with a VAF greater than zero in the matched germline sample, for each ensemble and patient.

with the number of SNVs in statistical models (486), and with the assumptions regarding the false positive rate from the initial evaluation of the variant callers, the SomaticSeq5 ensemble would be an ideal candidate to proceed with for recalling SNVs in plasma samples. Notedly, false positives could bias recalling SNVs in plasma samples, as the inclusion of germline variants could greatly increase the detection signal in matched plasma samples. Sequencing and alignment artifacts could additionally bias the detection signals. The exclusion of SNVs with a VAF greater than zero in the germline samples could increase the specificity of the ensembles; however, for the sake of simplicity, postprocessing of SNVs in the ensembles was not conducted.

To support the deductions of the initial evaluation of the ensembles, the SNV AFs in the merged tumor resection and matched germline samples were plotted in histograms for each of the ensembles for the patient PC1-10 (Figure 8). The SNVs identified by each of the ensembles for the patient PC1-10 were primarily distributed around a VAF of 0.32 in the merged tumor resection sample, while few SNVs were observed with a VAF of approximately 0.5 in the merged tumor resection sample (Figure 8). The SNVs that were identified by the SomaticSeq5 ensemble but not by the MRDetect ensemble were mainly distributed around a VAF of 0.32 in the merged tumor resection sample, while few had a VAF of approximately 0.5 in the merged tumor resection sample (Figure 8). The SNVs that were identified by the SomaticSeq4 ensemble but not by the SomaticSeq5 ensemble mostly had a VAF ranging from 0.2 to 0.4 in the merged tumor resection sample, while some were distributed around a VAF of 0.05 in the tumor resection sample (Figure 8). It should be noted that the SNVs that had a VAF greater than zero in the germline sample had a VAF ranging from 0.2 to 0.4 in the merged tumor resection sample and were distributed around a VAF of approximately 0.02 in the germline sample (Figure 8). The SNVs that were identified by the SomaticSeq3 ensemble, but not the SomaticSeq4 ensemble, all had a VAF less than 0.2 in the merged tumor resection sample, while the majority were distributed around a VAF of 0.03 in the merged tumor resection sample (Figure 8). Among these SNVs, one had a VAF greater than zero in the germline sample. Its VAF was approximately 0.05 in the germline sample and 0.2 in the merged tumor resection sample (Figure 8). Similar VAF distributions were observed for PC1-14 (Figure A.3) and PC-18 (Figure A.4).

The SNVs identified by the MRDetect and SomaticSeq5 ensembles for the patient PC1-10 were considered reliable and thus most likely true positives, primarily evaluated from the VAF of the SNVs in the germline sample. It should be noted that the SomaticSeq5 ensemble seemingly had greater sensitivity than the MRDetect ensemble without any evident loss of specificity. The SNVs identified by

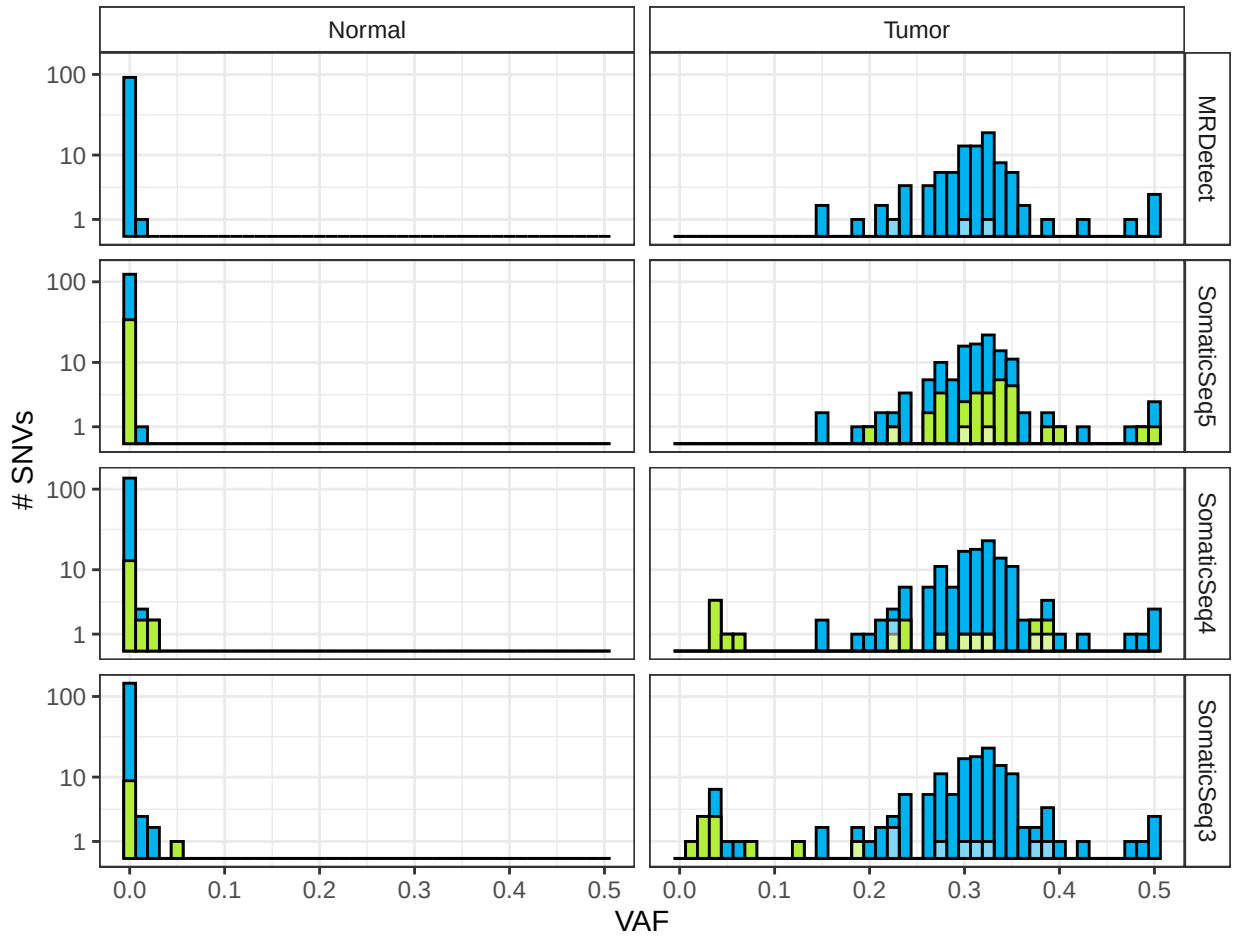

Figure 8: Histograms of the VAF of the SNVs for each ensemble in the merged tumor resection sample, tumor, and the matched germline sample, normal, for the patient, PC1-10. SNVs identified by the ensembles are colored blue. SNVs not identified by the above ensemble are colored green. SNVs with a VAF greater than zero in the matched germline sample are plotted on top with a transparent white color, and thus appear in a lighter blue and green color. The VAFs are calculated from reads with a minimum mapping quality of 10 and a minimum base quality of 30. The y-axis is pseudo $\log_{10}$ scaled to properly illustrate the SNV counts of each bin.

the SomaticSeq4 and SomaticSeq3 ensembles were largely considered reliable, although the increase in SNVs with a VAF greater than zero in the germline sample was concerning, especially as some of their VAFs in the germline sample were relatively high. Examination of allele counts revealed that the SNVs with a VAF greater than zero in the germline sample typically had a single forward or reverse read with the non-reference allele (data not included) that could have originated from misalignments or cross-contamination, whereas they most likely also could be considered true positives. The reduced stringency of the SomaticSeq4 and SomaticSeq3 ensembles, therefore, seemingly further increased the sensitivity without apparently reduced specificity, whereas the SomaticSeq3 ensemble was evaluated

to be the best-performing ensemble. Similar deductions were made from observations of the VAF distributions of identified SNVs for each ensemble in the merged tumor resection sample and matched germline sample of the patients PC1-14 and PC1-18. A commonality of the ensembles for the three patients was that the SomaticSeq5 ensembles identified more ancient somatic variants than the MRDetect ensembles, while the reduced stringency of the SomaticSeq4 and SomaticSeq3 ensembles relative to the SomaticSeq5 ensembles progressively identified more subclonal somatic variants.

4.3 Detection of circulating tumor DNA by recalling single-nucleotide variants in plasma samples

For evaluation of the MeanVAF, BinaryVAF, and MRDetectSNV approaches used for recalling SNVs in plasma samples, the Z-scores calculated from their respective detection signals were plotted for each detection approach and ensemble (Figure 9). The SomaticSeq3 and SomaticSeq5 ensembles were both included to investigate the potential effect of reduced consensus stringency. The MRDetect ensemble was additionally included and compared to the SomaticSeq3 and SomaticSeq5 ensembles to assess the potential effect of utilizing ensembles generated from an increased number of variant callers. The potential effect of read quality filtering was investigated by performing the 'by hand' analyses on nonfiltered and quality-filtered reads. To investigate the potential effect of in silico DNA fragment length selection, the three approaches were performed on both the nonfiltered and insert-size filtered plasma samples. Using the MeanVAF approach with the MRDetect ensemble, it was possible to detect ctDNA in two nonfiltered plasma samples without read-quality filters (Figure 9). This was increased to three when applying read-quality filters (Figure 9). Amongst the insert-size filtered plasma samples, only one was detected for ctDNA without read-quality filters, while two were detected with read-quality filters (Figure 9). With the SomaticSeq5 ensemble, ctDNA was detected using the MeanVAF approach in two nonfiltered plasma samples, both with and without read-quality filters (Figure 9). One insert-size filtered plasma sample was detected for ctDNA with and without read-quality filters (Figure 9). Two nonfiltered plasma samples were detected for ctDNA both with and without read quality filters, using the MeanVAF approach with the SomaticSeq3 ensemble (Figure 9). A single insert size filtered plasma sample was detected for ctDNA without read quality filters, while none was detectable when applying read quality filters (Figure 9). With the BinaryVAF approach, ctDNA was detected with the MRDetect ensemble in three nonfiltered plasma samples with and without

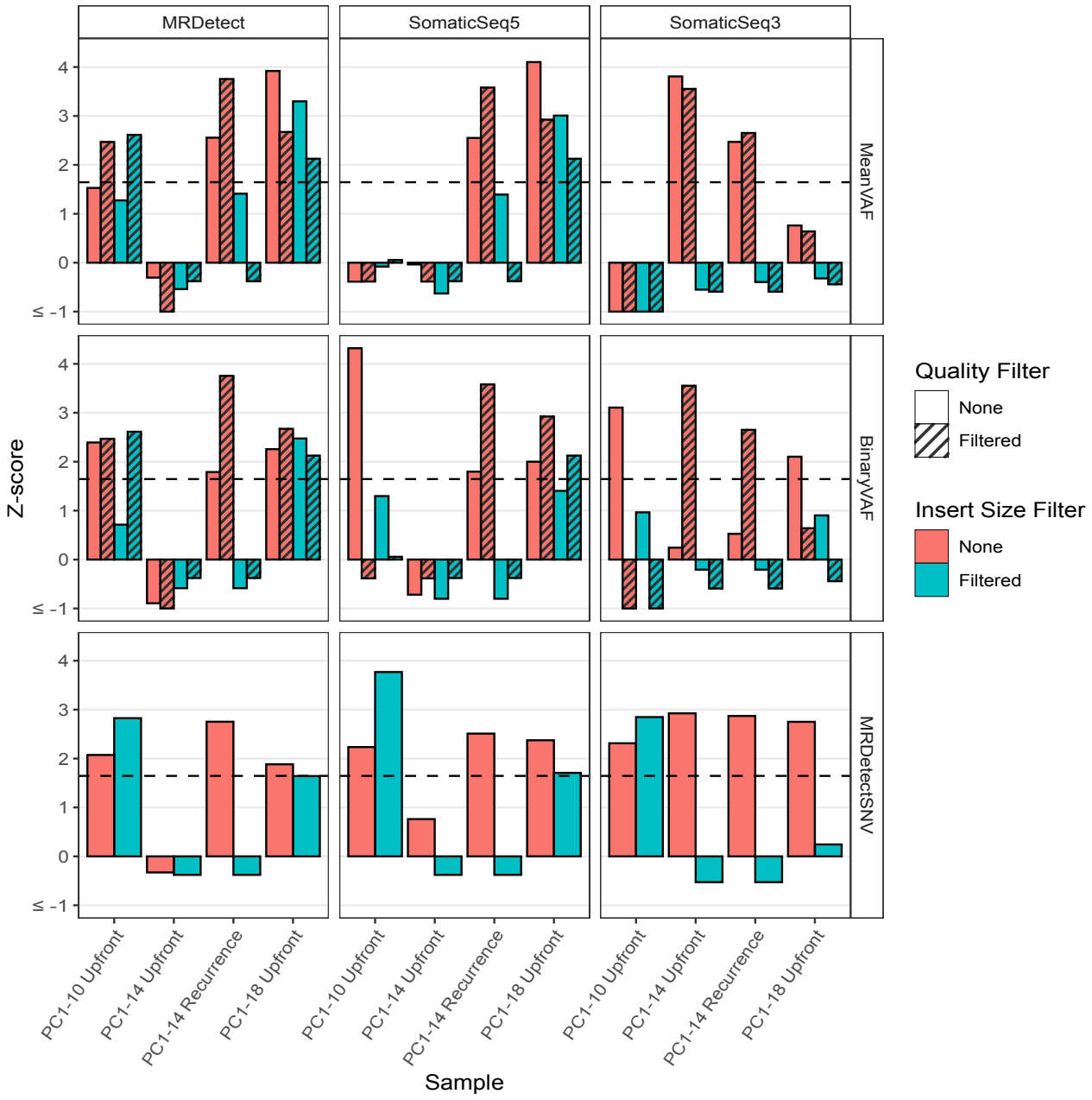

Figure 9: Comparison of ctDNA detection approaches (vertical) with different ensembles (horizontal) illustrated as bar plots of sample Z-Scores. The Z-Scores are calculated from the mean and standard deviation of 6-7 control samples. Z-Scores calculated from sample reads that have not been quality filtered are illustrated as hollow bars, while those calculated from sample reads that have been filtered according to base quality (Q > 30) and mapping quality (MAPQ > 10) are illustrated as striped bars. Z-Scores calculated from samples that have not been filtered according to insert size are colored red, while those that are calculated from samples that have been filtered according to insert size (90-150 bp) are colored green. A line is drawn horizontally on all plots to indicate the significance level (α=0.05; right-tailed).

read quality filters, while one and two insert size filtered plasma samples were detected for ctDNA without and with read quality filters, respectively (Figure 9). Using the BinaryVAF approach with the SomaticSeq5 ensemble, ctDNA was detected in three nonfiltered plasma samples without read quality filters, whereas two were detected for ctDNA with read quality filters (Figure 9). With the SomaticSeq3 ensemble, the BinaryVAF approach could detect ctDNA in two nonfiltered plasma samples both with and without read-quality filters (Figure 9). One insert-size filtered plasma sample was detected for ctDNA without read-quality filters, while none was detected with read-quality filters (Figure 9). When using the MRDetectSNV approach with the MRDetect ensemble, three nonfiltered plasma samples were detected for, whereas one insert-size filtered plasma sample could be detected for ctDNA (Figure 9). Using the MRDetectSNV approach with the SomaticSeq5 ensemble, three nonfiltered plasma samples were detected for ctDNA, while amongst the insert-size filtered plasma samples, two could be detected for ctDNA (Figure 9). With the SomaticSeq3 ensemble, the MRDetectSNV approach could detect ctDNA in all four nonfiltered samples, while only one insert-size filtered plasma sample was detected for ctDNA (Figure 9).

As evident from the Z-score bar plots of the MeanVAF and BinaryVAF approaches of ctDNA detection in plasma samples, filtering reads according to base quality and mapping quality did overall improve the sample Z-scores, especially when used with the MRDetect ensembles. While the read quality filtering should correct for sequencing and alignment errors, the quality metric thresholds used for filtering the reads might be insufficient, as evident from the few sample Z-scores that were greatly reduced by the read quality filtering. Profiling and tweaking the read quality filtering metrics could further improve ctDNA detection through these approaches. When comparing the Z-scores of the insert size filtered plasma samples to the Z-scores of the nonfiltered plasma samples, it was indisputable that the insert size selection overall resulted in greatly reduced performance of all approaches, as nearly all sample Z-scores were reduced for all detection approaches with all ensembles, with the exception the PC1-10 upfront plasma sample of all three ensembles with the MRDetectSNV approach. Although the selected insert size interval obviously negatively influenced the ctDNA detection approaches, profiling the insert sizes of reads with somatic variants might provide greater insight into whether in silico DNA fragment length, selection could improve upon SNV-based ctDNA detection approaches. From examinations of the results of the three SNV-based ctDNA detection approaches, it was evident that the MRDetectSNV approach had the best and most consistent performance across the three ensembles, with the best-performing ensemble being SomaticSeq3 that was able to detect ctDNA in

all four samples. The SVM model utilized by the MRDetectSNV approach for read quality filtering, in addition to the provided blacklist of somatic variants, was thought to be the cause of the better performance and consistency. Using plasma samples from healthy individuals instead of cross-patient plasma samples could improve the statistical scores, as common driver variants were hypothesized to potentially result in greater detection signals in the cross-patient plasma samples. It should be noted that the MeanVAF approach would be error-prone to false positive SNVs, whereas it should ideally be used with a stringent ensemble method. This error proneness was partially corrected for in the BinaryVAF approach, as each identified SNV contributed equally to the signal. Considering the MRDetectSNV approach used the site detection rate as a signal, it was thought that it also would be error-prone to false positive SNVs.

4.4 Evaluation of copy number variants

To assess the CNVs identified by the VarScan2 pipeline, the log2 ratio of copy numbers for each CNV segment in the autosomes was plotted for each of the three patients (Figure 10). For the patient PC1-10, the most prominent segments of amplifications were identified on chromosomes 1, 3, 8, 14, 15, 16, and 22, while the most prominent segments of deletions were identified on chromosomes 10, 14, 17, and 18 (Figure 10A). For the patient PC1-14, the most prominent segments of amplifications were identified on chromosomes 3, 9, 10, and 18, while the segments of deletions were identified on chromosomes 1, 3, 4, 5, 9, 17, 18, 19, and 22 (Figure 10B). For the patient PC1-18, the only large segment of amplifications was identified on chromosome 19, while the most prominent segments of deletions were identified on chromosomes 6, 8, 10, 11, 12, 19, and 21 (Figure 10C). Additionally, three neutral CNV segments identified on chromosomes 5, 8, and 12 were observed to have greater log2 ratios than otherwise observed for neutral segments (Figure 10C). Additional minor segments of amplifications and deletions were identified throughout the autosomes of the three patients (Figure 10). The log2 changes were observed to be slightly skewed for all patients, evident from the log2 change in allele proportions of the neutral segments (Figure 10).

The most prominent CNV segments identified by VarScan2 were corroborated by findings from the ichor-TITAN CNA pipeline regarding the total and allelic copy numbers in the merged tumor resection samples of the patients PC1-10 (Figure A.5), PC1-14 (Figure A.6), and PC1-18 (Figure A.7), whereas they were considered true positives. The additional minor segments of amplifications and deletions throughout the autosomes of the patients were thought to be artifacts of low regional coverage, as

all genomic positions with a coverage of at least 20 in both the merged tumor resection and matched germline samples were included in the VarScan2 CNA. These minor segments were thus considered false positives and could be algorithmically rectified by segment collapsing involving the calculation of weighted averages for improved specificity. From visual examinations, multiple neutral segments were evaluated to be misclassified as amplifications in the patients PC1-10 and PC1-14 and deletions in the patient PC1-18. The three neutral segments in the patient PC1-18 with greater log2 ratios than otherwise observed for neutral segments were evaluated to be misidentified and should have been classified as amplifications. This was corroborated by findings from the ichorCNA-TitanCNA pipeline regarding the total and allelic copy numbers in the merged tumor resection sample of patient PC1-18. The misidentification of segments was most certainly the consequence of the generally observed skew of the log2 ratios. From a technical perspective, the bias of the skewed log2 ratios could have been caused by low regional coverage, which would be evident from non-systematic variations in segment

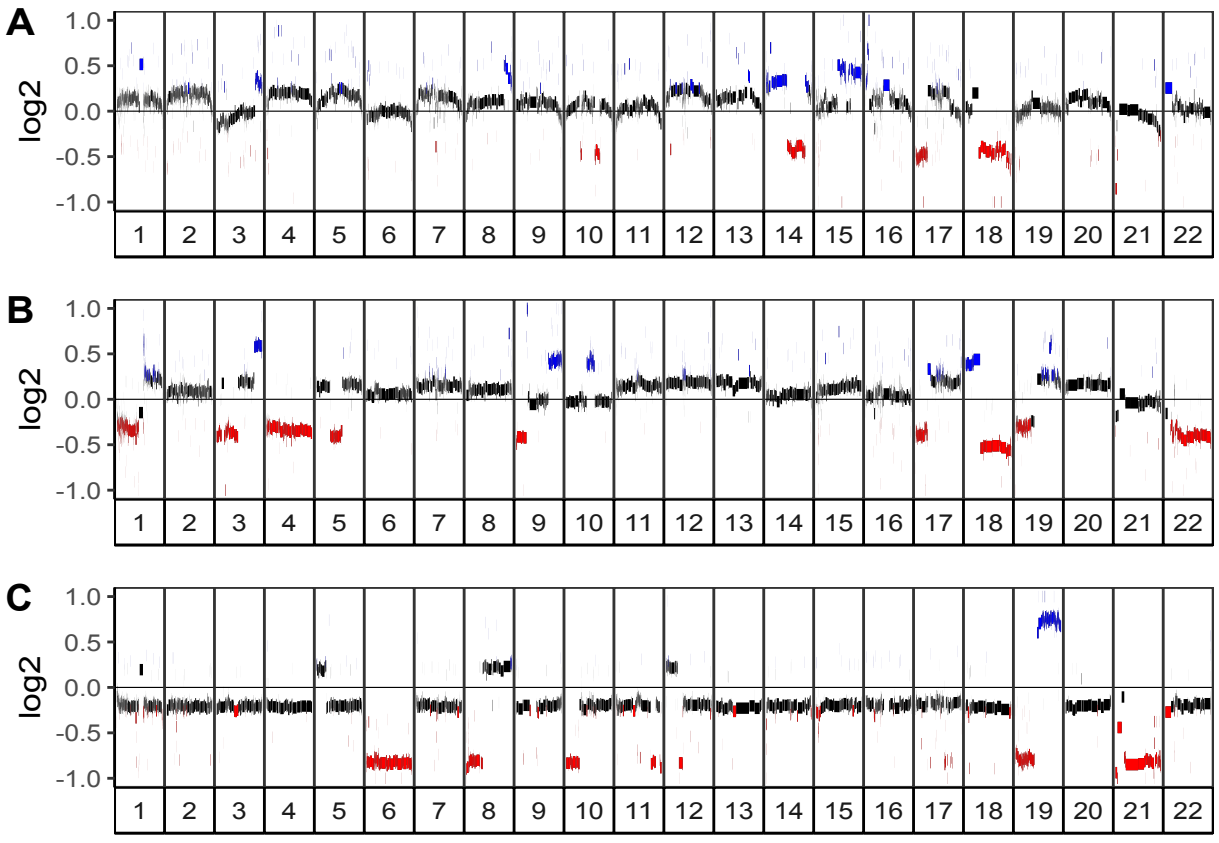

Figure 10: Plots illustrating the log2 change in copy numbers for each autosome calculated through the VarScan2CNA pipeline for the patients PC1-10 (A), PC1-14 (B), and PC1-18 (C). Neutral segments are colored black, while amplifications (log2 > 0.25) and deletions (log2 < -0.25) are colored blue and red, respectively.

log2 ratios. The bias of the skewed log2 ratios could also be caused by the by-chromosome mean log2 ratio centering approach employed by the VarScan2 pipeline, which especially would be evident from the majority of the log2 ratios of neutral segments deviating from a log2 of zero. Arguably, centering according to the median log2 ratio would have been a more robust methodology that could have increased both the sensitivity and specificity. The segment classification could additionally be optimized through adjustments of the log2 ratio thresholds. From a biological standpoint, the bias of the skewed log2 ratios could have been caused by differing clonal and subclonal ploidy, which especially would be evident from clusters of segments with systematically differing log2 ratios. After recentering the segments according to the median log2 ratio and altering the classification thresholds from 0.25 to 0.3 for amplifications and from -0.25 to -0.3 for deletions, the majority of the suspected false positives biased by the skewed log2 ratios were rectified (Figure A.8). Additionally, further investigation of tumor heterogeneity and more detailed CNV profiling was deemed beyond the scope of this thesis, as the approach used for recalling CNVs in plasma samples, MRDetectCNV, only required CNV segments to be classified as amplifications, deletions, and neutral.

4.5 Detection of circulating tumor DNA by recalling copy number variants in plasma samples

For assessment of the MRDetectCNV ctDNA detection approach, the Z-scores calculated from the total CNV signals, in addition to the separate signals of the amplifications and deletions, were plotted for each of the patients' plasma samples (Figure 11). To investigate the potential effect of segment misclassification, the analysis was performed for both the raw and post-processed CNVs identified by VarScan2. Additionally, the potential effect of in silico DNA fragment size selection was investigated by assessment of the Z-scores calculated from the signals of the nonfiltered and insert size filtered plasma samples. With the raw CNVs identified by VarScan2, none of the nonfiltered plasma samples were detected for ctDNA when using the total CNV signal and the signal of the deletions, while one nonfiltered plasma sample was detected for ctDNA when using the signal of the amplifications (Figure 11). Additionally, with the raw CNVs identified by VarScan2, one insert-size filtered plasma sample was detected for ctDNA using the total signal of the CNVs, while none insert size filtered plasma samples were detected for ctDNA when using the signals of the deletions and amplifications separately (Figure 11). With the post-processed CNVs identified by VarScan2, ctDNA was not detected in any of

the nonfiltered plasma samples using the total CNV signal and the signal of the deletions, while one nonfiltered plasma sample was detected for ctDNA when using the signal of the amplifications (Figure 11). None of the insert-size filtered plasma samples were detected for ctDNA with the post-processed CNVs identified by VarScan2 when using either of the signals (Figure 11).

The further postprocessing of the CNVs identified by VarScan2 was determined to overall improve the Z-scores of the plasma samples, whereas misclassification of segments was deduced to generally negatively impact CNV-based ctDNA detection. However, it should be noted that further post-processing greatly reduced the Z-score of the PC1-14 upfront insert size filtered plasma sample calculated from the total CNV signal. Reanalyzing the samples with different classification thresholds could further elucidate the importance of high sensitivity versus high specificity. The insert size filtered plasma samples were evaluated to overall improve the Z-scores of the plasma samples, especially

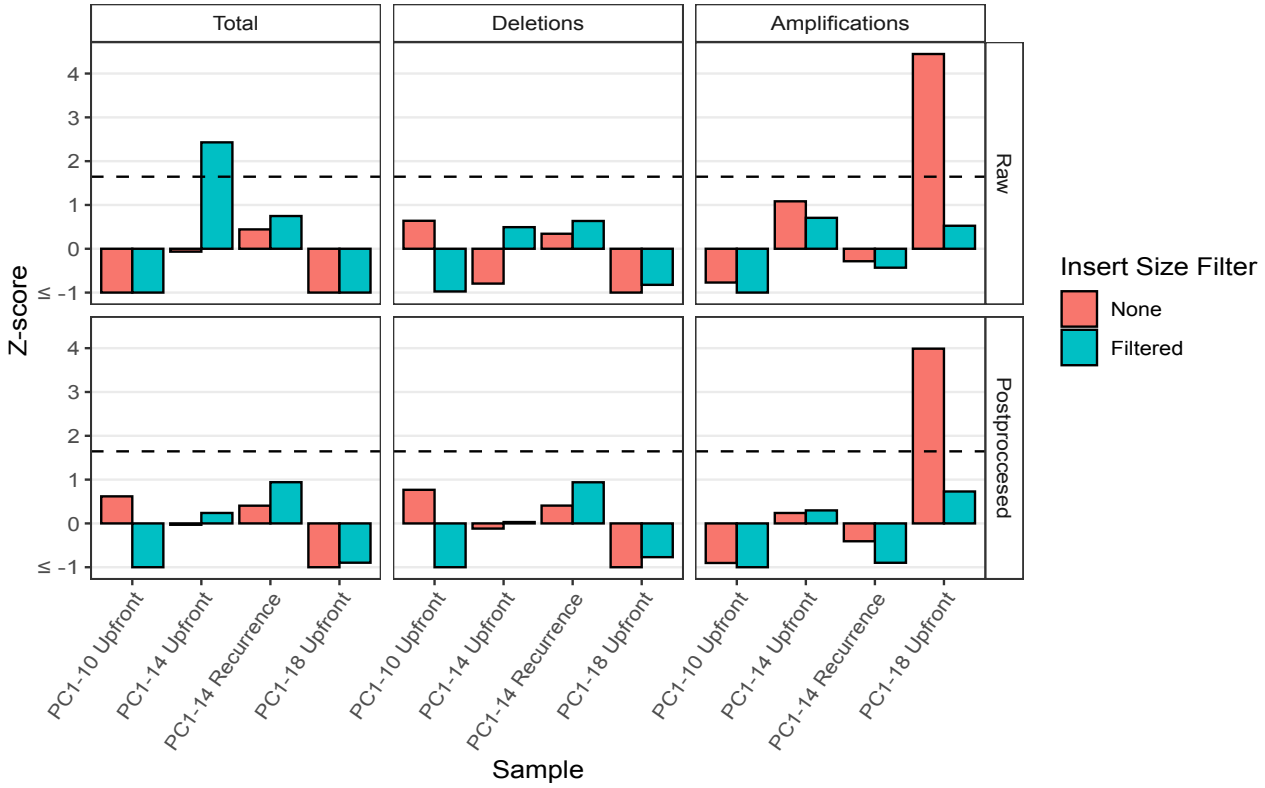

Figure 11: Comparison of Z-scores calculated from total signal of the deletions and amplifications, in addition to the separate signals of the deletions and amplifications. All signals are calculated with MRDetectCNA from the raw and postprocessed copy number variants determined by VarScan2. Z-scores calculated from the signal of reads from the plasma samples filtered according to insert size (90-150 bp) and the non-filtered plasma samples are colored green and red, respectively. A line is drawn horizontally on all plots to indicate the significance level (α=0.05; right-tailed).

deduced from the increased Z-scores of the PC1-14 plasma samples, whereas in silico DNA fragment length, selection could be a viable option in CNV-based ctDNA detection approaches. Notedly, the insert size filtration was observed to greatly reduce the Z-score calculated from the signal of the amplifications in the PC1-18 upfront plasma sample. Profiling the insert sizes of the genomic regions of the CNVs could elucidate a more optimal insert size filtration interval. Using plasma samples from healthy individuals as controls instead of cross-patient plasma samples could additionally improve the statistical scores by lowering the population means and standard deviation from which the Z-scores were calculated, as the CNVs in the cross-patient plasma samples were hypothesized to have distorted the signals. Using plasma samples from healthy individuals as controls could elucidate an optimal significance level threshold for acceptance of the null hypothesis in the statistical tests.

4.6 Combination of statistical scores

The joint probabilities of the SNV- and CNV-based ctDNA detection approaches were converted into Z-scores and plotted for each of the patients' samples (Figure 12). With the combined statistical scores of the SNV- and CNV-based ctDNA detection approaches, all four nonfiltered plasma samples

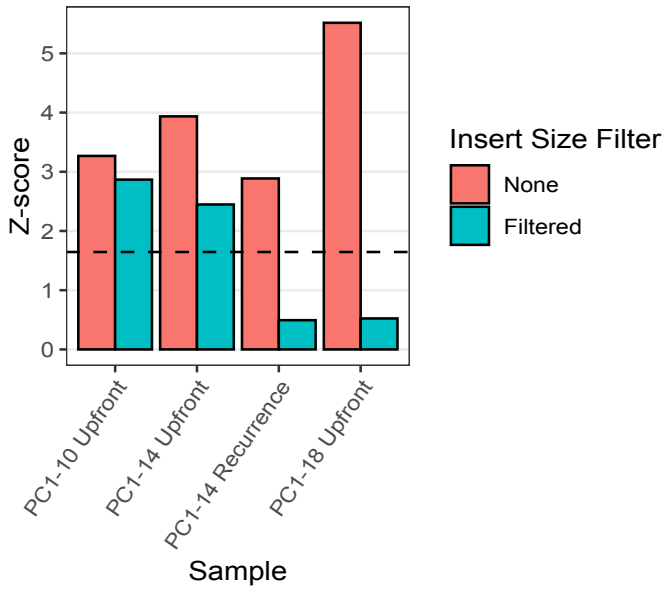

Figure 12: Graphical illustration of Z-scores calculated from the joint probabilities of the SNV- and CNV-based ctDNA detection approaches. Z-Scores calculated from samples that have not been filtered according to insert size are colored red, while those that are calculated from samples that have been filtered according to insert size (90-150 bp) are colored green. A line is drawn horizontally on all plots to indicate the significance level (α=0.05; right-tailed).

were detected for ctDNA. In contrast, two insert-size filtered plasma samples were detected for ctDNA (Figure 12).

As previously mentioned, although the insert size filtration caused a reduction in the statistical scores of the plasma samples, in silico DNA fragment length selection was evaluated to have shown promise. Further investigation would, however, be required. Upon selecting the methodologies for combining the statistical scores of the ctDNA detection approaches, the dependencies of the statistical tests were well considered to calculate an appropriate representation of the joint probability of ctDNA detection. The methodology of combining statistical scores would allow the integration of additional ctDNA detection approaches. Utilization of the joint probability of ctDNA detection could become of clinal relevance in prognostics by improved detection of minimal residual disease for predicting recurrence, in addition to evaluation of treatment resistance.

5 Conclusion

It was evident from the comparisons of the ensembles used for the SNV-based ctDNA detection approaches that the inclusion of SNVs identified by an increased number of variant callers and reduced stringency of the consensus ensembles increased the sensitivity without apparent loss of specificity. The optimal stringency was determined to be 3/7 of the included variant callers. Increased sensitivity of ensembles was evaluated to improve SNV-based ctDNA detection, deduced from the number of plasma samples detected for ctDNA. MRDetectSNV was determined to perform best among the included SNV-based ctDNA detection approaches, as it consistently detected ctDNA in most plasma samples with each included ensemble.

Examination of the CNV profiles of VarScan2 revealed inefficient segment recentering, which was attributed to the workflow's mean log2 ratio centering methodology. The further post-processing by segment recentering with the median log2 ratio and adjustments of the log2 ratio CNV classification thresholds was determined to have increased the sensitivity and specificity of the analyses. From the comparison of the statistical score of the MRDetectCNV ctDNA detection approach with the raw and post-processed CNVs of VarScan2, the misclassification of segments was deduced to impact CNV-based ctDNA detection negatively. However, general poor performance was observed for both the raw and post-processed CNVs.

The in silico DNA fragment length selection method by insert size filtration yielded mixed results, as it reduced to statistical scores of the plasma samples with the SNV-based ctDNA detection approaches but increased the statistical scores of the plasma samples with the CNV-based ctDNA detection approach. Further investigation would be required to determine the effect of in silico DNA fragment length selection on ctDNA detection approaches.

Combining the statistical scores from the ctDNA detection approaches was determined to be an appropriate methodology for representing the joint probability of ctDNA detection in plasma samples that could become of clinical relevance regarding improved recurrence prediction by detection of minimal residual disease in addition to evaluation of treatment resistance.

6 Further perspectives

While WES in this thesis has been demonstrated to be efficient at the mutational characterization of tumors with various variant callers, broadening of sequencing with WGS would allow the identification of larger quantities of somatic variants. The increased breadth of sequencing could also be beneficial for detecting ctDNA by improved use of the limited input material. It should be noted that utilizing WGS instead of WES as a sequencing approach would reduce the depth of coverage, which could reduce the specificity of the variant callers. The potential reduction in the specificity of the variant callers could be overcome by generating consensus ensembles with increased stringency, consequently resulting in reduced sensitivity. The ensemble generation could be further optimized by the inclusion of more variant callers. Additionally, considering erroneous mapping of reads are a major concern when calling variants, using multiple aligners in conjunction with multiple variant callers could further increase the sensitivity (475). With the advancements in the application of supervised machine learning for somatic mutation classification, consensus ensembles could be considered outmoded, as machine learning algorithms are capable of increasing the specificity with minimal loss of sensitivity by filtering variants upon numerous features and quality metrics. Retaining high sensitivity and specificity of the identified somatic variants is especially import for the SNV-based ctDNA detection approaches, as recalling larger quantities of somatic variants in plasma samples has been shown to increase the probability of ctDNA detection in simulations (486), while false positives can result in erroneous detection signals. The utilization of supervised machine learning for the filtration of false positives would require knowing the true positives, whereas such models inevitably must be trained on either simulated or thoroughly examined real data. Considering training models on simulated data could introduce biases, real data would be preferred. Training on high-confidence calls from gold-standard datasets has proven capable of building accurate and robust machine-learning models (485, 514). However, training on high-confidence calls generated by a consensus approach might be more useful to properly encapsulate sample-specific technical and biological variations.

The CNV analysis could be further extended by including additional variant callers and comparing their CNV profiles. An algorithmic combination of the CNVs identified by different variant callers could potentially increase sensitivity and specificity, thus improving upon the CNV-based ctDNA detection approach. Another option for optimization of the CNV-based ctDNA detection approach would be combing the statistical scores calculated from the detection signals of the different variant

callers' CNV profiles with a methodology such as the harmonic mean p-value. As the field of liquid biopsies expands, new approaches for detection of ctDNA in plasma samples are certain to be developed. They could be used in addition to those used in this thesis for improved detection of ctDNA. In contrast, dependencies of the statistical tests should be carefully considered to appropriately represent their joint probability. As briefly mentioned in the results section, the detection of ctDNA could become of clinical relevance in prognostics by predicting recurrence and evaluating treatment response. The development of standardized bioinformatics procedures for ctDNA detection in plasma samples would, however, be imminent for consistency and reproducibility and should be based on the results of a greater case-control study.

The mutational characterization of the tumors could be extended by the inclusion of variant callers for the detection of INDELs and SVs. It should be noted that the detection of SVs on short-read NGS data, as used in this thesis, is inherently limited by the length of the generated reads, whereas incorrect and failed mapping of reads could be of concern (515). Using long-read sequencing technologies would allow accurate detection of complex SVs by improved sequence assembly (515). Although identifying INDELs and SVs would not benefit the currently available ctDNA detection approaches, the extended mutational characterization could be clinically relevant in personalized cancer treatments by targeted therapy against driver variants. A more comprehensive mutational characterization of tumors could additionally aid the identification of novel driver variants and genes by somatic variant recurrence (516, 517) and pathogenicity prediction (518-521). As recently demonstrated (522), identifying novel driver variants could lead to developing novel targeted therapies enabled by the revolutionary application of AI for in silico drug discovery and development. The applied AI algorithm, AlphaFold, can predict protein structures from amino acid sequences with unprecedented accuracy (523-525), from which ligand-protein interactions of generated chemical compounds can be simulated with various computational methodologies (526).

Considering the availability of four distinct tumor resection samples for each patient included in this thesis, it would be possible to investigate the subclonal heterogeneity of the tumors by comparison of the mutational landscapes of the distinct tumor resection samples. Although it is unlikely to become clinically relevant, it could further our understanding of tumor heterogeneity and evolution. However, using single-cell DNA sequencing would arguably be more efficient for profiling tumor heterogeneity.

While various analyses of somatic variants in tumor and plasma samples could become of clinical relevance, including only well-optimized algorithms for such analyses in a clinical setting would benefit

the patient outcome by reducing the total data processing time. This could especially be a key factor for personalized cancer treatments and the evaluation of treatment resistance. Most bioinformatic tools available nowadays have been software optimized through various optimization techniques, focusing on algorithm runtime in addition to memory and storage usage to varying extents, where their algorithmic efficiencies are classified according to the Big-O notation. Additionally, most available bioinformatic tools employ multithreaded computational parallelization for accelerated runtime, in which algorithms simultaneously execute operations on multiple CPU threads. Many bioinformatic tools could, however, be further optimized by employing heterogenous hardware acceleration in conjunction with computational parallelization, where all available processing units are leveraged for maximized performance. Such processing units could be GPUs, TPUs, and FPGAs, in addition to CPUs. However, it should be noted that GPUs, TPUs, and FPGAs can only handle basic mathematical operations, whereas CPUs are excellent at handling complex operations. Utilization of NVIDIA Clara Parabricks, a GPU-accelerated computational genomics application framework, has recently been demonstrated to perform a WGS analysis for a 30X human genome on an 8×A100 GPU server over 80 times faster than CPU-based workflows on the same server (527). The total power consumptions required for the WGS analyses calculated from the rated maximum power consumption of the CPUs (528) and GPUs (529) in the server suggest that heterogenous hardware acceleration could reduce the power consumption 10-fold relative to CPU-based workflows. Thus optimization of bioinformatic tools with heterogenous hardware acceleration would be both time- and cost-effective, although it should be noted that server-grade GPUs are rather expensive.

Bibliography

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians.n/a(n/a).

2. Lippi G, Mattiuzzi C. The global burden of pancreatic cancer. Arch Med Sci. 2020;16(4):820-4.

3. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392(10159):1789-858.

4. Survival Rates for Pancreatic Cancer American Cancer Society [Available from: https://www.cancer.org/cancer/pancreatic-cancer/detection-diagnosis-staging/survival-rates.html.

5. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA: A Cancer Journal for Clinicians. 2021;71(1):7-33.

6. Kelsen DP, Portenoy R, Thaler H, Tao Y, Brennan M. Pain as a predictor of outcome in patients with operable pancreatic carcinoma. Surgery. 1997;122(1):53-9.

7. Poruk KE, Gay DZ, Brown K, Mulvihill JD, Boucher KM, Scaife CL, et al. The clinical utility of CA 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. Curr Mol Med. 2013;13(3):340-51.

8. Maithel SK, Maloney S, Winston C, Gönen M, D'Angelica MI, Dematteo RP, et al. Preoperative CA 19-9 and the yield of staging laparoscopy in patients with radiographically resectable pancreatic adenocarcinoma. Ann Surg Oncol. 2008;15(12):3512-20.

9. Karachristos A, Scarmeas N, Hoffman JP. CA 19-9 Levels Predict Results of Staging Laparoscopy in Pancreatic Cancer. Journal of Gastrointestinal Surgery. 2005;9(9):1286-92.

10. Scarà S, Bottoni P, Scatena R. CA 19-9: Biochemical and Clinical Aspects. Adv Exp Med Biol. 2015;867:247-60.

11. Chu LC, Goggins MG, Fishman EK. Diagnosis and Detection of Pancreatic Cancer. The Cancer Journal. 2017;23(6).

12. Phoa SSKS, Reeders JWAJ, Rauws EAJ, de Wit L, Gouma DJ, Laméris JS. Spiral computed tomography for preoperative staging of potentially resectable carcinoma of the pancreatic head. British Journal of Surgery. 1999;86(6):789-94.

13. Phoa SS, Reeders JW, Rauws EA, De Wit L, Gouma DJ, Laméris JS. Spiral computed tomography for preoperative staging of potentially resectable carcinoma of the pancreatic head. Br J Surg. 1999;86(6):789-94.

14. Jin Z, Li X, Cai L. Assessing the resectability of pancreatic ductal adenocarcinoma: comparision of dual-phase helical CT arterial portography with conventional angiography. Chin Med Sci J. 2001;16(1):40-5.

15. de Bono JS, Harris JR, Burm SM, Vanderstichele A, Houtkamp MA, Aarass S, et al. Systematic study of tissue factor expression in solid tumors. Cancer Rep (Hoboken). 2023;6(2):e1699.

16. Farrell JJ, Wong JL, Ken B, Baker M, Maney T. 215 The Routine Clinical Yield of Molecular Analysis for Precision Medicine in Pancreatic Cancer Using Pancreatic Biopsy Fine Needle Aspiration (FNA) Material. Room for Improvement? Gastrointestinal Endoscopy. 2016;83(5, Supplement):AB131-AB2.

17. Yoshizawa N, Yamada R, Sakuno T, Inoue H, Miura H, Takeuchi T, et al. Comparison of endoscopic ultrasound-guided fine-needle aspiration and biopsy with 22-gauge and 25-gauge needles for the "precision medicine" of pancreatic cancer: A retrospective study. Medicine (Baltimore). 2018;97(24):e11096.

18. Luo G, Guo M, Liu Z, Xiao Z, Jin K, Long J, et al. Blood Neutrophil–Lymphocyte Ratio Predicts Survival in Patients with Advanced Pancreatic Cancer Treated with Chemotherapy. Annals of Surgical Oncology. 2015;22(2):670-6.

19. Liu J, Yu L, Ding W. Efficacy and safety of Kanglaite injection combined with radiochemotherapy in the treatment of advanced pancreatic cancer: A PRISMA-compliant meta-analysis. Medicine (Baltimore). 2019;98(32):e16656.

20. Seufferlein T, Ettrich TJ. Treatment of pancreatic cancer-neoadjuvant treatment in resectable pancreatic cancer (PDAC). Transl Gastroenterol Hepatol. 2019;4:21.

21. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457-81.

22. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187-220.

23. Peto R, Peto J. Asymptotically Efficient Rank Invariant Test Procedures. Journal of the Royal Statistical Society Series A (General). 1972;135(2):185-207.

24. Halabi S, Owzar K. The importance of identifying and validating prognostic factors in oncology. Semin Oncol. 2010;37(2):e9-18.

25. Molinaro AM, Wrensch MR, Jenkins RB, Eckel-Passow JE. Statistical considerations on prognostic models for glioma. Neuro Oncol. 2016;18(5):609-23.

26. Nikanjam M, Kato S, Kurzrock R. Liquid biopsy: current technology and clinical applications. Journal of Hematology & Oncology. 2022;15(1):131.

27. Chang C-M, Lin K-C, Hsiao N-E, Hong W-A, Lin C-Y, Liu T-C, et al. Clinical application of liquid biopsy in cancer patients. BMC Cancer. 2022;22(1):413.

28. Xin L, Yue Y, Zihan R, Youbin C, Tianyu L, Rui W. Clinical application of liquid biopsy based on circulating tumor DNA in non-small cell lung cancer. Front Physiol. 2023;14:1200124.

29. Saha S, Araf Y, Promon SK. Circulating tumor DNA in cancer diagnosis, monitoring, and prognosis. Journal of the Egyptian National Cancer Institute. 2022;34(1):8.

30. Stadler JC, Belloum Y, Deitert B, Sementsov M, Heidrich I, Gebhardt C, et al. Current and Future Clinical Applications of ctDNA in Immuno-Oncology. Cancer Res. 2022;82(3):349-58.

31. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch R-D, et al. DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells1. Cancer Research. 2001;61(4):1659-65.

32. Zent CS, Elliott MR. Maxed out macs: physiologic cell clearance as a function of macrophage phagocytic capacity. Febs j. 2017;284(7):1021-39.

33. Pinney JJ, Rivera-Escalera F, Chu CC, Whitehead HE, VanDerMeid KR, Nelson AM, et al. Macrophage hypophagia as a mechanism of innate immune exhaustion in mAb-induced cell clearance. Blood. 2020;136(18):2065-79.

34. Stroun M, Lyautey J, Lederrey C, Olson-Sand A, Anker P. About the possible origin and mechanism of circulating DNA: Apoptosis and active DNA release. Clinica Chimica Acta. 2001;313(1):139-42.

35. Minciacchi VR, Zijlstra A, Rubin MA, Di Vizio D. Extracellular vesicles for liquid biopsy in prostate cancer: where are we and where are we headed? Prostate Cancer Prostatic Dis. 2017;20(3):251-8.

36. Thakur BK, Zhang H, Becker A, Matei I, Huang Y, Costa-Silva B, et al. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. Cell Research. 2014;24(6):766-9.

37. Kahlert C, Melo SA, Protopopov A, Tang J, Seth S, Koch M, et al. Identification of Double-stranded Genomic DNA Spanning All Chromosomes with Mutated KRAS and p53 DNA in the Serum Exosomes of Patients with Pancreatic Cancer*. Journal of Biological Chemistry. 2014;289(7):3869-75.

38. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6(224):224ra24.

39. Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci Transl Med. 2016;8(346):346ra92.

40. Chaudhuri AA, Chabon JJ, Lovejoy AF, Newman AM, Stehr H, Azad TD, et al. Early Detection of Molecular Residual Disease in Localized Lung Cancer by Circulating Tumor DNA Profiling. Cancer Discov. 2017;7(12):1394-403.

41. Schwarzenbach H, Stoehlmacher J, Pantel K, Goekkurt E. Detection and Monitoring of Cell-Free DNA in Blood of Patients with Colorectal Cancer. Annals of the New York Academy of Sciences. 2008;1137(1):190-6.

42. Lanman RB, Mortimer SA, Zill OA, Sebisanovic D, Lopez R, Blau S, et al. Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. PLoS One. 2015;10(10):e0140712.

43. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci U S A. 2015;112(11):E1317-25.

44. Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, et al. Fragment Length of Circulating Tumor DNA. PLoS Genet. 2016;12(7):e1006162.

45. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci U S A. 2018;115(46):E10925-e33.

46. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164(1-2):57-68.

47. Prat A, Brasó-Maristany F, Martínez-Sáez O, Sanfeliu E, Xia Y, Bellet M, et al. Circulating tumor DNA reveals complex biological features with clinical relevance in metastatic breast cancer. Nature Communications. 2023;14(1):1157.

48. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10(466).

49. Liu Y, Liu Y, Wang Y, Li L, Yao W, Song Y, et al. Increased detection of circulating tumor DNA by short fragment enrichment. Transl Lung Cancer Res. 2021;10(3):1501-11.

50. Beiter T, Fragasso A, Hudemann J, Niess AM, Simon P. Short-term treadmill running as a model for studying cell-free DNA kinetics in vivo. Clin Chem. 2011;57(4):633-6.

51. Lo YM, Zhang J, Leung TN, Lau TK, Chang AM, Hjelm NM. Rapid clearance of fetal DNA from maternal plasma. Am J Hum Genet. 1999;64(1):218-24.

52. Lau TW, Leung TN, Chan LY, Lau TK, Chan KC, Tam WH, et al. Fetal DNA clearance from maternal plasma is impaired in preeclampsia. Clin Chem. 2002;48(12):2141-6.

53. Yu SC, Lee SW, Jiang P, Leung TY, Chan KC, Chiu RW, et al. High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing. Clin Chem. 2013;59(8):1228-37.

54. Coakley M, Garcia-Murillas I, Turner NC. Molecular Residual Disease and Adjuvant Trial Design in Solid Tumors. Clin Cancer Res. 2019;25(20):6026-34.

55. Øgaard N, Reinert T, Henriksen TV, Frydendahl A, Aagaard E, Ørntoft MW, et al. Tumour-agnostic circulating tumour DNA analysis for improved recurrence surveillance after resection of colorectal liver metastases: A prospective cohort study. Eur J Cancer. 2022;163:163-76.

56. Henriksen TV, Tarazona N, Frydendahl A, Reinert T, Gimeno-Valiente F, Carbonell-Asins JA, et al. Circulating Tumor DNA in Stage III Colorectal Cancer, beyond Minimal Residual Disease Detection, toward Assessment of Adjuvant Therapy Efficacy and Clinical Behavior of Recurrences. Clin Cancer Res. 2022;28(3):507-17.

57. Dizdaroglu M, Jaruga P, Birincioglu M, Rodriguez H. Free radical-induced damage to DNA: mechanisms and measurement1, 2 1This article is part of a series of reviews on "Oxidative DNA Damage and Repair." The full list of papers may be found on the homepage of the journal. 2Guest Editor: Miral Dizdaroglu. Free Radical Biology and Medicine. 2002;32(11):1102-15.

58. Gates KS. An overview of chemical processes that damage cellular DNA: spontaneous hydrolysis, alkylation, and reactions with radicals. Chem Res Toxicol. 2009;22(11):1747-60.

59. Dizdaroglu M, Jaruga P. Mechanisms of free radical-induced damage to DNA. Free Radical Research. 2012;46(4):382-419.

60. Irigaray P, Belpomme D. Basic properties and molecular mechanisms of exogenous chemical carcinogens. Carcinogenesis. 2010;31(2):135-48.

61. Ikehata H, Ono T. The Mechanisms of UV Mutagenesis. Journal of Radiation Research. 2011;52(2):115-25.

62. Mavragani IV, Nikitaki Z, Kalospyros SA, Georgakilas AG. Ionizing Radiation and Complex DNA Damage: From Prediction to Detection Challenges and Biological Significance. Cancers (Basel). 2019;11(11).

63. Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environ Mol Mutagen. 2017;58(5):235-63.

64. Bouwman P, Jonkers J. The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. Nat Rev Cancer. 2012;12(9):587-98.

65. Ghosal G, Chen J. DNA damage tolerance: a double-edged sword guarding the genome. Transl Cancer Res. 2013;2(3):107-29.

66. Wolters S, Schumacher B. Genome maintenance and transcription integrity in aging and disease. Front Genet. 2013;4:19.

67. Krokan HE, Bjørås M. Base excision repair. Cold Spring Harb Perspect Biol. 2013;5(4):a012583.

68. Barrows LR, Magee PN. Nonenzymatic methylation of DNA by S-adenosylmethionine in vitro. Carcinogenesis. 1982;3(3):349-51.

69. Näslund M, Segerbäck D, Kolman A. S-Adenosylmethionine, an endogenous alkylating agent. Mutation Research Letters. 1983;119(3):229-32.

70. Struck AW, Thompson ML, Wong LS, Micklefield J. S-adenosyl-methionine-dependent methyltransferases: highly versatile enzymes in biocatalysis, biosynthesis and other biotechnological applications. Chembiochem. 2012;13(18):2642-55.

71. Henle ES, Linn S. Formation, Prevention, and Repair of DNA Damage by Iron/Hydrogen Peroxide*. Journal of Biological Chemistry. 1997;272(31):19095-8.

72. Lindahl T. New class of enzymes acting on damaged DNA. Nature. 1976;259(5538):64-6.

73. Nicholl ID, Nealon K, Kenny MK. Reconstitution of Human Base Excision Repair with Purified Proteins. Biochemistry. 1997;36(24):7557-66.

74. Parikh SS, Mol CD, Tainer JA. Base excision repair enzyme family portrait: integrating the structure and chemistry of an entire DNA repair pathway. Structure. 1997;5(12):1543-50.

75. Yi C, He C. DNA repair by reversal of DNA damage. Cold Spring Harb Perspect Biol. 2013;5(1):a012575.

76. Müller M, Carell T. Structural biology of DNA photolyases and cryptochromes. Curr Opin Struct Biol. 2009;19(3):277-85.

77. Brettel K, Byrdin M. Reaction mechanisms of DNA photolyase. Current Opinion in Structural Biology. 2010;20(6):693-701.

78. Drabløs F, Feyzi E, Aas PA, Vaagbø CB, Kavli B, Bratlie MS, et al. Alkylation damage in DNA and RNA—repair mechanisms and medical significance. DNA Repair. 2004;3(11):1389-407.

79. Schärer OD. Nucleotide excision repair in eukaryotes. Cold Spring Harb Perspect Biol. 2013;5(10):a012609.

80. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. Understanding nucleotide excision repair and its roles in cancer and ageing. Nature Reviews Molecular Cell Biology. 2014;15(7):465-81.

81. Tubbs JL, Latypov V, Kanugula S, Butt A, Melikishvili M, Kraehenbuehl R, et al. Flipping of alkylated DNA damage bridges base and nucleotide excision repair. Nature. 2009;459(7248):808-13.

82. Otterlei M, Warbrick E, Nagelhus TA, Haug T, Slupphaug G, Akbari M, et al. Post-replicative base excision repair in replication foci. Embo j. 1999;18(13):3834-44.

83. Haushalter KA, Todd Stukenberg P, Kirschner MW, Verdine GL. Identification of a new uracil-DNA glycosylase family by expression cloning using synthetic inhibitors. Current Biology. 1999;9(4):174-85.

84. Neddermann P, Jiricny J. Efficient Removal of Uracil from G.U Mispairs by the Mismatch-Specific Thymine DNA Glycosylase from HeLa Cells. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(5):1642-6.

85. Van Der Kemp PA, Thomas D, Barbey R, xe, gine, De Oliveira R, et al. Cloning and Expression in Escherichia coli of the OGG1 Gene of Saccharomyces cerevisiae, Which Codes for a DNA Glycosylase that Excises 7,8-Dihydro-8-Oxoguanine and 2,6-Diamino-4-Hydroxy-5-N-Methylformamidopyrimidine. Proceedings of the National Academy of Sciences of the United States of America. 1996;93(11):5197-202.

86. Kubota Y, Nash RA, Klungland A, Schär P, Barnes D, Lindahl T. Reconstitution of DNA base excision-repair with purified human proteins: interaction between DNA polymerase beta and the XRCC1 protein. The EMBO journal. 1996;15(23):6662-70.

87. Sobol RW, Horton JK, Kühn R, Gu H, Singhal RK, Prasad R, et al. Requirement of mammalian DNA polymerase- in base-excision repair. Nature. 1996;379(6561):183-6.

88. Wei Y-F, Robins P, Carter K, Caldecott K, Pappin DJC, Yu G-L, et al. Molecular Cloning and Expression of Human cDNAs Encoding a Novel DNA Ligase IV and DNA Ligase III, an Enzyme Active in DNA Repair and Recombination. Molecular and Cellular Biology. 1995;15(6):3206-16.

89. Nash RA, Caldecott KW, Barnes DE, Lindahl T. XRCC1 Protein Interacts with One of Two Distinct Forms of DNA Ligase III. Biochemistry. 1997;36(17):5207-11.

90. Shrivastav N, Li D, Essigmann JM. Chemical biology of mutagenesis and DNA repair: cellular responses to DNA alkylation. Carcinogenesis. 2010;31(1):59-70.

91. Fu D, Calvo JA, Samson LD. Balancing repair and tolerance of DNA damage caused by alkylating agents. Nat Rev Cancer. 2012;12(2):104-20.

92. Zhou W, Doetsch PW. Effects of abasic sites and DNA single-strand breaks on prokaryotic RNA polymerases. Proc Natl Acad Sci U S A. 1993;90(14):6601-5.

93. Thompson PS, Cortez D. New insights into abasic site repair and tolerance. DNA Repair (Amst). 2020;90:102866.

94. Laverty DJ, Averill AM, Doublié S, Greenberg MM. The A-Rule and Deletion Formation During Abasic and Oxidized Abasic Site Bypass by DNA Polymerase . ACS Chem Biol. 2017;12(6):1584-92.

95. Masutani C, Sugasawa K, Yanagisawa J, Sonoyama T, Ui M, Enomoto T, et al. Purification and cloning of a nucleotide excision repair complex involving the xeroderma pigmentosum group C protein and a human homologue of yeast RAD23. Embo j. 1994;13(8):1831-43.

96. Nishi R, Okuda Y, Watanabe E, Mori T, Iwai S, Masutani C, et al. Centrin 2 stimulates nucleotide excision repair by interacting with xeroderma pigmentosum group C protein. Mol Cell Biol. 2005;25(13):5664-74.

97. Riedl T, Hanaoka F, Egly JM. The comings and goings of nucleotide excision repair factors on damaged DNA. Embo j. 2003;22(19):5293-303.

98. Fu I, Mu H, Geacintov NE, Broyde S. Mechanism of lesion verification by the human XPD helicase in nucleotide excision repair. Nucleic Acids Research. 2022;50(12):6837-53.

99. Staresincic L, Fagbemi AF, Enzlin JH, Gourdin AM, Wijgers N, Dunand-Sauthier I, et al. Coordination of dual incision and repair synthesis in human nucleotide excision repair. Embo j. 2009;28(8):1111-20.

100. Fagbemi AF, Orelli B, Schärer OD. Regulation of endonuclease activity in human nucleotide excision repair. DNA Repair (Amst). 2011;10(7):722-9.

101. Ogi T, Limsirichaikul S, Overmeer RM, Volker M, Takenaka K, Cloney R, et al. Three DNA Polymerases, Recruited by Different Mechanisms, Carry Out NER Repair Synthesis in Human Cells. Molecular Cell. 2010;37(5):714-27.

102. Minca EC, Kowalski D. Replication fork stalling by bulky DNA damage: localization at active origins and checkpoint modulation. Nucleic Acids Res. 2011;39(7):2610-23.

103. Iyer DR, Rhind N. Replication fork slowing and stalling are distinct, checkpoint-independent consequences of replicating damaged DNA. PLOS Genetics. 2017;13(8):e1006958.

104. Kunkel TA. DNA Replication Fidelity*. Journal of Biological Chemistry. 2004;279(17):16895-8.

105. Reha-Krantz LJ, Woodgate S, Goodman MF. Engineering processive DNA polymerases with maximum benefit at minimum cost. Frontiers in Microbiology. 2014;5.

106. Oertell K, Harcourt EM, Mohsen MG, Petruska J, Kool ET, Goodman MF. Kinetic selection vs. free energy of DNA base pairing in control of polymerase fidelity. Proceedings of the National Academy of Sciences. 2016;113(16):E2277-E85.

107. Kuznetsova AA, Fedorova OS, Kuznetsov NA. Structural and Molecular Kinetic Features of Activities of DNA Polymerases. Int J Mol Sci. 2022;23(12).

108. Joyce CM, Benkovic SJ. DNA Polymerase Fidelity: Kinetics, Structure, and Checkpoints. Biochemistry. 2004;43(45):14317-24.

109. Kumar D, Abdulovic AL, Viberg J, Nilsson AK, Kunkel TA, Chabes A. Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. Nucleic Acids Res. 2011;39(4):1360-71.

110. de Paz AM, Cybulski TR, Marblestone AH, Zamft BM, Church GM, Boyden ES, et al. High-resolution mapping of DNA polymerase fidelity using nucleotide imbalances and next-generation sequencing. Nucleic Acids Research. 2018;46(13):e78-e.

111. Wang W, Hellinga HW, Beese LS. Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. Proceedings of the National Academy of Sciences - PNAS. 2011;108(43):17644-8.

112. Bebenek K, Pedersen LC, Kunkel TA. Replication infidelity via a mismatch with Watson-Crick geometry. Proc Natl Acad Sci U S A. 2011;108(5):1862-7.

113. Koag MC, Nam K, Lee S. The spontaneous replication error and the mismatch discrimination mechanisms of human DNA polymerase . Nucleic Acids Res. 2014;42(17):11233-45.

114. Kimsey IJ, Petzold K, Sathyamoorthy B, Stein ZW, Al-Hashimi HM. Visualizing transient Watson–Crick-like mispairs in DNA and RNA duplexes. Nature. 2015;519(7543):315-20.

115. Kimsey IJ, Szymanski ES, Zahurancik WJ, Shakya A, Xue Y, Chu C-C, et al. Dynamic basis for dG·dT misincorporation via tautomerization and ionization. Nature. 2018;554(7691):195-201,A-L.

116. Szymanski ES, Kimsey IJ, Al-Hashimi HM. Direct NMR Evidence that Transient Tautomeric and Anionic States in dG·dT Form Watson-Crick-like Base Pairs. J Am Chem Soc. 2017;139(12):4326-9.

117. Rangadurai A, Szymanski ES, Kimsey I, Shi H, Al-Hashimi HM. Probing conformational transitions towards mutagenic Watson-Crick-like G·T mismatches using off-resonance sugar carbon R(1) relaxation dispersion. J Biomol NMR. 2020;74(8-9):457-71.

118. Johnson SJ, Beese LS. Structures of mismatch replication errors observed in a DNA polymerase. Cell. 2004;116(6):803-16.

119. Xia S, Konigsberg WH. Mispairs with Watson-Crick base-pair geometry observed in ternary complexes of an RB69 DNA polymerase variant. Protein Sci. 2014;23(4):508-13.

120. Allawi HT, SantaLucia J, Jr. NMR solution structure of a DNA dodecamer containing single G.T mismatches. Nucleic Acids Res. 1998;26(21):4925-34.

121. Patel DJ, Kozlowski SA, Marky LA, Rice JA, Broka C, Dallas J, et al. Structure, dynamics, and energetics of deoxyguanosine . thymidine wobble base pair formation in the self-complementary d(CGTGAATTCGCG) duplex in solution. Biochemistry. 1982;21(3):437-44.

122. Hare D, Shapiro L, Patel DJ. Wobble dG X dT pairing in right-handed DNA: solution conformation of the d(C-G-T-G-A-A-T-T-C-G-C-G) duplex deduced from distance geometry analysis of nuclear Overhauser effect spectra. Biochemistry. 1986;25(23):7445-56.

123. Hunter WN, Brown T, Kneale G, Anand NN, Rabinovich D, Kennard O. The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. J Biol Chem. 1987;262(21):9962-70.

124. Tautz D, Schlötterer. Simple sequences. Curr Opin Genet Dev. 1994;4(6):832-7.

125. Wells RD. Molecular basis of genetic instability of triplet repeats. J Biol Chem. 1996;271(6):2875-8.

126. Sinden RR, Potaman VN, Oussatcheva EA, Pearson CE, Lyubchenko YL, Shlyakhtenko LS. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. J Biosci. 2002;27(1 Suppl 1):53-65.

127. Viguera E, Canceill D, Ehrlich SD. Replication slippage involves DNA polymerase pausing and dissociation. Embo j. 2001;20(10):2587-95.

128. Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 1987;4(3):203-21.

129. Baptiste BA, Jacob KD, Eckert KA. Genetic evidence that both dNTP-stabilized and strand slippage mechanisms may dictate DNA polymerase errors within mononucleotide microsatellites. DNA Repair (Amst). 2015;29:91-100.

130. Murat P, Guilbaud G, Sale JE. DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. Genome Biology. 2020;21(1):209.

131. Joshua-Tor L, Frolow F, Appella E, Hope H, Rabinovich D, Sussman JL. Three-dimensional structures of bulge-containing DNA fragments. J Mol Biol. 1992;225(2):397-431.

132. Efrati E, Tocco G, Eritja R, Wilson SH, Goodman MF. Abasic translesion synthesis by DNA polymerase beta violates the "A-rule". Novel types of nucleotide incorporation by human DNA polymerase beta at an abasic lesion in different sequence contexts. J Biol Chem. 1997;272(4):2559-69.

133. Kunkel TA. The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. Journal of Biological Chemistry. 1985;260(9):5787-96.

134. Tippin B, Kobayashi S, Bertram JG, Goodman MF. To slip or skip, visualizing frameshift mutation dynamics for error-prone DNA polymerases. J Biol Chem. 2004;279(44):45360-8.

135. Bebenek K, Kunkel TA. Frameshift errors initiated by nucleotide misincorporation. Proc Natl Acad Sci U S A. 1990;87(13):4946-50.

136. Ling H, Boudsocq F, Woodgate R, Yang W. Crystal Structure of a Y-Family DNA Polymerase in Action: A Mechanism for Error-Prone and Lesion-Bypass Replication. Cell. 2001;107(1):91-102.

137. Vanderstraeten S, Van den Brûle S, Hu J, Foury F. The Role of 3-5 Exonucleolytic Proofreading and Mismatch Repair in Yeast Mitochondrial DNA Error Avoidance. The Journal of biological chemistry. 1998;273(37):23690-7.

138. Longley MJ, Nguyen D, Kunkel TA, Copeland WC. The Fidelity of Human DNA Polymerase with and without Exonucleolytic Proofreading and the p55 Accessory Subunit. The Journal of biological chemistry. 2001;276(42):38555-62.

139. Kroutil LC, Register K, Bebenek K, Kunkel TA. Exonucleolytic Proofreading during Replication of Repetitive DNA. Biochemistry. 1996;35(3):1046-53.

140. Roberts JD, Thomas DC, Kunkel TA. Exonucleolytic proofreading of leading and lagging strand DNA replication errors. Proc Natl Acad Sci U S A. 1991;88(8):3465-9.

141. Modrich P. Mechanisms in E. coli and Human Mismatch Repair (Nobel Lecture). Angewandte Chemie International Edition. 2016;55(30):8490-501.

142. Sherrer SM, Penland E, Modrich P. The mutagen and carcinogen cadmium is a high-affinity inhibitor of the zinc-dependent MutL endonuclease. Proc Natl Acad Sci U S A. 2018;115(28):7314-9.

143. Su SS, Modrich P. Escherichia coli mutS-encoded protein binds to mismatched DNA base pairs. Proc Natl Acad Sci U S A. 1986;83(14):5057-61.

144. Grilley M, Welsh KM, Su SS, Modrich P. Isolation and characterization of the Escherichia coli mutL gene product. J Biol Chem. 1989;264(2):1000-4.

145. Harfe BD, Jinks-Robertson S. DNA mismatch repair and genetic instability. Annu Rev Genet. 2000;34:359-99.

146. Schofield MJ, Hsieh P. DNA mismatch repair: molecular mechanisms and biological function. Annu Rev Microbiol. 2003;57:579-608.

147. Li G-M. DNA mismatch repair and cancer. FBL. 2003;8(4):997-1017.

148. Allen DJ, Makhov A, Grilley M, Taylor J, Thresher R, Modrich P, et al. MutS mediates heteroduplex loop formation by a translocation mechanism. The EMBO Journal. 1997;16(14):4467-76.

149. Gradia S, Acharya S, Fishel R. The Human Mismatch Recognition Complex hMSH2-hMSH6 Functions as a Novel Molecular Switch. Cell. 1997;91(7):995-1005.

150. Guarné A, Ramon-Maiques S, Wolff EM, Ghirlando R, Hu X, Miller JH, et al. Structure of the MutL C-terminal domain: a model of intact MutL and its roles in mismatch repair. The EMBO Journal. 2004;23(21):4134-45.

151. Junop MS, Obmolova G, Rausch K, Hsieh P, Yang W. Composite Active Site of an ABC ATPase: MutS Uses ATP to Verify Mismatch Recognition and Authorize DNA Repair. Molecular Cell. 2001;7(1):1-12.

152. Yang W, Junop MS, Ban C, Obmolova G, Hsieh P. DNA mismatch repair: from structure to mechanism. Cold Spring Harb Symp Quant Biol. 2000;65:225-32.

153. Hombauer H, Campbell Christopher S, Smith Catherine E, Desai A, Kolodner Richard D. Visualization of Eukaryotic DNA Mismatch Repair Reveals Distinct Recognition and Repair Intermediates. Cell. 2011;147(5):1040-53.

154. Qiu R, Sakato M, Sacho EJ, Wilkins H, Zhang X, Modrich P, et al. MutL traps MutS at a DNA mismatch. Proc Natl Acad Sci U S A. 2015;112(35):10914-9.

155. Kadyrov FA, Dzantiev L, Constantin N, Modrich P. Endonucleolytic function of MutLalpha in human mismatch repair. Cell. 2006;126(2):297-308.

156. Ortega J, Lee GS, Gu L, Yang W, Li G-M. Mispair-bound human MutS–MutL complex triggers DNA incisions and activates mismatch repair. Cell Research. 2021;31(5):542-53.

157. Yuan F, Gu L, Guo S, Wang C, Li GM. Evidence for involvement of HMGB1 protein in human DNA mismatch repair. J Biol Chem. 2004;279(20):20935-40.

158. Guo S, Zhang Y, Yuan F, Gao Y, Gu L, Wong I, et al. Regulation of Replication Protein A Functions in DNA Mismatch Repair by Phosphorylation*. Journal of Biological Chemistry. 2006;281(31):21607-16.

159. Zhang Y, Yuan F, Presnell SR, Tian K, Gao Y, Tomkinson AE, et al. Reconstitution of 5-Directed Human Mismatch Repair in a Purified System. Cell. 2005;122(5):693-705.

160. Mardenborough YSN, Nitsenko K, Laffeber C, Duboc C, Sahin E, Quessada-Vial A, et al. The unstructured linker arms of MutL enable GATC site incision beyond roadblocks during initiation of DNA mismatch repair. Nucleic Acids Res. 2019;47(22):11667-80.

161. Putnam CD. Evolution of the methyl directed mismatch repair system in Escherichia coli. DNA Repair. 2016;38:32-41.

162. Lee JY, Chang J, Joseph N, Ghirlando R, Rao DN, Yang W. MutH Complexed with Hemi- and Unmethylated DNAs: Coupling Base Recognition and DNA Cleavage. Molecular Cell. 2005;20(1):155-66.

163. Liu J, Hanne J, Britton BM, Bennett J, Kim D, Lee J-B, et al. Cascading MutS and MutL sliding clamps control DNA diffusion to activate mismatch repair. Nature. 2016;539(7630):583-7.

164. Hegde ML, Hazra TK, Mitra S. Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. Cell Res. 2008;18(1):27-47.

165. Wang JC. Cellular roles of DNA topoisomerases: a molecular perspective. Nat Rev Mol Cell Biol. 2002;3(6):430-40.

166. Caldecott KW. XRCC1 and DNA strand break repair. DNA Repair. 2003;2(9):955-69.

167. Andres SN, Schellenberg MJ, Wallace BD, Tumbale P, Williams RS. Recognition and repair of chemically heterogeneous structures at DNA ends. Environ Mol Mutagen. 2015;56(1):1-21.

168. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993;362(6422):709-15.

169. Plo I, Liao Z-Y, Barceló JM, Kohlhagen G, Caldecott KW, Weinfeld M, et al. Association of XRCC1 and tyrosyl DNA phosphodiesterase (Tdp1) for the repair of topoisomerase I-mediated DNA lesions. DNA Repair. 2003;2(10):1087-100.

170. D'Amours D, Desnoyers S, D'Silva I, Poirier GG. Poly(ADP-ribosyl)ation reactions in the regulation of nuclear functions. Biochem J. 1999;342 (Pt 2)(Pt 2):249-68.

171. Davidovic L, Vodenicharov M, Affar EB, Poirier GG. Importance of poly(ADP-ribose) glycohydrolase in the control of poly(ADP-ribose) metabolism. Exp Cell Res. 2001;268(1):7-13.

172. Lin Y, Bai L, Cupello S, Hossain MA, Deem B, McLeod M, et al. APE2 promotes DNA damage response pathway from a single-strand break. Nucleic Acids Res. 2018;46(5):2479-94.

173. Abbotts R, Wilson DM, 3rd. Coordination of DNA single strand break repair. Free Radic Biol Med. 2017;107:228-44.

174. Harris JL, Jakob B, Taucher-Scholz G, Dianov GL, Becherel OJ, Lavin MF. Aprataxin, poly-ADP ribose polymerase 1 (PARP-1) and apurinic endonuclease 1 (APE1) function together to protect the genome against oxidative damage. Hum Mol Genet. 2009;18(21):4102-17.

175. Tumbale P, Williams JS, Schellenberg MJ, Kunkel TA, Williams RS. Aprataxin resolves adenylated RNA-DNA junctions to maintain genome integrity. Nature. 2014;506(7486):111-5.

176. El-Khamisy SF, Saifi GM, Weinfeld M, Johansson F, Helleday T, Lupski JR, et al. Defective DNA single-strand break repair in spinocerebellar ataxia with axonal neuropathy-1. Nature. 2005;434(7029):108-13.

177. Nocentini S. Comet Assay Analysis of Repair of DNA Strand Breaks in Normal and Deficient Human Cells Exposed to Radiations and Chemicals. Evidence for a Repair Pathway Specificity of DNA Ligation. Radiation Research. 1995;144(2):170-80.

178. Cotner-Gohara E, Kim IK, Hammel M, Tainer JA, Tomkinson AE, Ellenberger T. Human DNA ligase III recognizes DNA ends by dynamic switching between two DNA-bound states. Biochemistry. 2010;49(29):6165-76.

179. Lan L, Nakajima S, Oohata Y, Takao M, Okano S, Masutani M, et al. In situ analysis of repair processes for oxidative DNA damage in mammalian cells. Proc Natl Acad Sci U S A. 2004;101(38):13738-43.

180. Mortusewicz O, Rothbauer U, Cardoso MC, Leonhardt H. Differential recruitment of DNA Ligase I and III to DNA repair sites. Nucleic Acids Res. 2006;34(12):3523-32.

181. Andrei K. Single-Strand Interruptions in Replicating Chromosomes Cause Double-Strand Breaks. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(15):8241-6.

182. Pfeiffer P, Goedecke W, Obe G. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. Mutagenesis. 2000;15(4):289-302.

183. Lieber MR. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem. 2010;79:181-211.

184. Sonoda E, Hochegger H, Saberi A, Taniguchi Y, Takeda S. Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. DNA Repair. 2006;5(9):1021-9.

185. Her J, Bunting SF. How cells ensure correct repair of DNA double-strand breaks. Journal of Biological Chemistry. 2018;293(27):10502-11.

186. Rothkamm K, Krüger I, Thompson LH, Löbrich M. Pathways of DNA double-strand break repair during the mammalian cell cycle. Mol Cell Biol. 2003;23(16):5706-15.

187. Li X, Heyer WD. Homologous recombination in DNA repair and DNA damage tolerance. Cell Res. 2008;18(1):99-113.

188. Elbakry A, Löbrich M. Homologous Recombination Subpathways: A Tangle to Resolve. Front Genet. 2021;12:723847.

189. Guirouilh-Barbat J, Lambert S, Bertrand P, Lopez BS. Is homologous recombination really an error-free process? Front Genet. 2014;5:175.

190. Fiorentini P, Huang KN, Tishkoff DX, Kolodner RD, Symington LS. Exonuclease I of Saccharomyces cerevisiae functions in mitotic recombination in vivo and in vitro. Mol Cell Biol. 1997;17(5):2764-73.

191. Kadyk LC, Hartwell LH. Sister chromatids are preferred over homologs as substrates for recombinational repair in Saccharomyces cerevisiae. Genetics. 1992;132(2):387-402.

192. Soutoglou E, Dorn JF, Sengupta K, Jasin M, Nussenzweig A, Ried T, et al. Positional stability of single double-strand breaks in mammalian cells. Nat Cell Biol. 2007;9(6):675-82.

193. Doherty AJ, Jackson SP. DNA repair: How Ku makes ends meet. Current Biology. 2001;11(22):R920-R4.

194. Zhang Z, Zhu L, Lin D, Chen F, Chen DJ, Chen Y. The three-dimensional structure of the C-terminal DNA-binding domain of human Ku70. J Biol Chem. 2001;276(41):38231-6.

195. Fell VL, Schild-Poulter C. Ku regulates signaling to DNA damage response pathways through the Ku70 von Willebrand A domain. Mol Cell Biol. 2012;32(1):76-87.

196. Uematsu N, Weterings E, Yano K, Morotomi-Yano K, Jakob B, Taucher-Scholz G, et al. Autophosphorylation of DNA-PKCS regulates its dynamics at DNA double-strand breaks. J Cell Biol. 2007;177(2):219-29.

197. Yano K, Morotomi-Yano K, Wang SY, Uematsu N, Lee KJ, Asaithamby A, et al. Ku recruits XLF to DNA double-strand breaks. EMBO Rep. 2008;9(1):91-6.

198. Grundy GJ, Rulten SL, Zeng Z, Arribas-Bosacoma R, Iles N, Manley K, et al. APLF promotes the assembly and activity of non-homologous end joining protein complexes. Embo j. 2013;32(1):112-25.

199. Kanno S, Kuzuoka H, Sasao S, Hong Z, Lan L, Nakajima S, et al. A novel human AP endonuclease with conserved zinc-finger-like motifs involved in DNA strand break responses. Embo j. 2007;26(8):2094-103.

200. Macrae CJ, McCulloch RD, Ylanko J, Durocher D, Koch CA. APLF (C2orf13) facilitates nonhomologous end-joining and undergoes ATM-dependent hyperphosphorylation following ionizing radiation. DNA Repair (Amst). 2008;7(2):292-302.

201. Mari PO, Florea BI, Persengiev SP, Verkaik NS, Brüggenwirth HT, Modesti M, et al. Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. Proc Natl Acad Sci U S A. 2006;103(49):18597-602.

202. Costantini S, Woodbine L, Andreoli L, Jeggo PA, Vindigni A. Interaction of the Ku heterodimer with the DNA ligase IV/Xrcc4 complex and its regulation by DNA-PK. DNA Repair (Amst). 2007;6(6):712-22.

203. Nick McElhinny SA, Snowden CM, McCarville J, Ramsden DA. Ku recruits the XRCC4-ligase IV complex to DNA ends. Mol Cell Biol. 2000;20(9):2996-3003.

204. Weterings E, Chen DJ. The endless tale of non-homologous end-joining. Cell Research. 2008;18(1):114-24.

205. Davis AJ, Chen DJ. DNA double strand break repair via non-homologous end-joining. Transl Cancer Res. 2013;2(3):130-43.

206. Stinson BM, Moreno AT, Walter JC, Loparo JJ. A Mechanism to Minimize Errors during Non-homologous End Joining. Molecular Cell. 2020;77(5):1080-91.e8.

207. Ma Y, Pannicke U, Schwarz K, Lieber MR. Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. Cell. 2002;108(6):781-94.

208. Mahajan KN, Nick McElhinny SA, Mitchell BS, Ramsden DA. Association of DNA polymerase mu (pol mu) with Ku and ligase IV: role for pol mu in end-joining double-strand break repair. Mol Cell Biol. 2002;22(14):5194-202.

209. Richardson C, Jasin M. Frequent chromosomal translocations induced by DNA double-strand breaks. Nature. 2000;405(6787):697-700.

210. Boulton SJ, Jackson SP. Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. The EMBO Journal. 1996;15(18):5093-103.

211. Ma J-L, Kim EM, Haber JE, Lee SE. Yeast Mre11 and Rad1 Proteins Define a Ku-Independent Mechanism To Repair Double-Strand Breaks Lacking Overlapping End Sequences. Molecular and Cellular Biology. 2003;23(23):8820-8.

212. Chiruvella KK, Liang Z, Wilson TE. Repair of double-strand breaks by end joining. Cold Spring Harb Perspect Biol. 2013;5(5):a012757.

213. Wang H, Xu X. Microhomology-mediated end joining: new players join the team. Cell & Bioscience. 2017;7(1):6.

214. Sfeir A, Symington LS. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends in Biochemical Sciences. 2015;40(11):701-14.

215. McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends in Genetics. 2008;24(11):529-38.

216. Pannunzio NR, Li S, Watanabe G, Lieber MR. Non-homologous end joining often uses microhomology: Implications for alternative end joining. DNA Repair. 2014;17:74-80.

217. Sinha S, Villarreal D, Shim EY, Lee SE. Risky business: Microhomology-mediated end joining. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2016;788:17-24.

218. Truong LN, Li Y, Shi LZ, Hwang PY, He J, Wang H, et al. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. Proc Natl Acad Sci U S A. 2013;110(19):7720-5.

219. Seol JH, Shim EY, Lee SE. Microhomology-mediated end joining: Good, bad and ugly. Mutat Res. 2018;809:81-7.

220. Decottignies A. Microhomology-mediated end joining in fission yeast is repressed by pku70 and relies on genes involved in homologous recombination. Genetics. 2007;176(3):1403-15.

221. Fishman-Lobell J, Haber JE. Removal of nonhomologous DNA ends in double-strand break recombination: the role of the yeast ultraviolet repair gene RAD1. Science. 1992;258(5081):480-4.

222. Liang L, Deng L, Nguyen SC, Zhao X, Maulion CD, Shao C, et al. Human DNA ligases I and III, but not ligase IV, are required for microhomology-mediated end joining of DNA double-strand breaks. Nucleic Acids Research. 2008;36(10):3297-310.

223. Kent T, Chandramouly G, McDevitt SM, Ozdemir AY, Pomerantz RT. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase . Nat Struct Mol Biol. 2015;22(3):230-7.

224. Hogg M, Sauer-Eriksson AE, Johansson E. Promiscuous DNA synthesis by human DNA polymerase . Nucleic Acids Res. 2012;40(6):2611-22.

225. Villarreal DD, Lee K, Deem A, Shim EY, Malkova A, Lee SE. Microhomology directs diverse DNA break repair pathways and chromosomal translocations. PLoS Genet. 2012;8(11):e1003026.

226. Simsek D, Jasin M. Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4-ligase IV during chromosomal translocation formation. Nat Struct Mol Biol. 2010;17(4):410-6.

227. Simsek D, Brunet E, Wong SY, Katyal S, Gao Y, McKinnon PJ, et al. DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation. PLoS Genet. 2011;7(6):e1002080.

228. Zhang Y, Jasin M. An essential role for CtIP in chromosomal translocation formation through an alternative end-joining pathway. Nat Struct Mol Biol. 2011;18(1):80-4.

229. Meaburn KJ, Misteli T, Soutoglou E. Spatial genome organization in the formation of chromosomal translocations. Semin Cancer Biol. 2007;17(1):80-90.

230. Kuhfittig-Kulle S, Feldmann E, Odersky A, Kuliczkowska A, Goedecke W, Eggert A, et al. The mutagenic potential of non-homologous end joining in the absence of the NHEJ core factors Ku70/80, DNA-PKcs and XRCC4-LigIV. Mutagenesis. 2007;22(3):217-33.

231. Wetterstrand KA. The Cost of Sequencing a Human Genome: National Human Genome Research Institute; 2021 [updated 01-11. Available from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

232. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977;74(12):5463-7.

233. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. 1975;94(3):441-8.

234. Pervez M, Ul Hasnain M, Abbas S, Moustafa M, Aslam N, Shah S. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. BioMed Research International. 2022;2022:1-12.

235. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Human Immunology. 2021;82(11):801-11.

236. TruSeq DNA Sample Preparation Guide: Illumina; [Protocol]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqdna/TruSeq_DNA_SamplePrep_Guide_15026486_C.pdf.

237. Sample Multiplexing Overview: Illumina; [Available from: https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing.html.

238. Tseng Q, Lomonosov AM, Furlong EEM, Merten CA. Fragmentation of DNA in a submicroliter microfluidic sonication device. Lab on a Chip. 2012;12(22):4677-82.

239. Deininger PL. Random subcloning of sonicated DNA: Application to shotgun DNA sequence analysis. Analytical Biochemistry. 1983;129(1):216-23.

240. Kasoji SK, Pattenden SG, Malc EP, Jayakody CN, Tsuruta JK, Mieczkowski PA, et al. Cavitation Enhancing Nanodroplets Mediate Efficient DNA Fragmentation in a Bench Top Ultrasonic Water Bath. PLOS ONE. 2015;10(7):e0133014.

241. Sambrook J, Russell DW. Fragmentation of DNA by Sonication. Cold Spring Harbor Protocols. 2006;2006(4):pdb.prot4538.

242. Lentz YK, Worden LR, Anchordoquy TJ, Lengsfeld CS. Effect of jet nebulization on DNA: identifying the dominant degradation mechanism and mitigation methods. Journal of Aerosol Science. 2005;36(8):973-90.

243. Sambrook J, Russell DW. Fragmentation of DNA by Nebulization. Cold Spring Harbor Protocols. 2006;2006(4):pdb.prot4539.

244. Yew FFH, Davidson N. Breakage by hydrodynamic shear of the bonds between cohered ends of -DNA molecules. Biopolymers. 1968;6(5):659-79.

245. Shui L, Sparreboom W, Spang P, Roeser T, Nieto B, Guasch F, et al. High yield DNA fragmentation using cyclical hydrodynamic shearing. RSC Advances. 2013;3(32):13115-8.

246. Thorstenson YR, Hunicke-Smith SP, Oefner PJ, Davis RW. An automated hydrodynamic process for controlled, unbiased DNA shearing. Genome research. 1998;8(8):848-55.

247. Oefner PJ, Hunicke-Smith SP, Chiang L, Dietrich F, Mulligan J, Davis RW. Efficient Random Subcloning of DNA Sheared in a Recirculating Point-Sink Flow System. Nucleic Acids Research. 1996;24(20):3879-86.

248. Nesterova IV, Hupert ML, Witek MA, Soper SA. Hydrodynamic shearing of DNA in a polymeric microfluidic device. Lab on a Chip. 2012;12(6):1044-7.

249. Joneja A, Huang X. A device for automated hydrodynamic shearing of genomic DNA. BioTechniques. 2009;46(7):553-6.

250. Shui L, Bomer JG, Jin M, Carlen ET, van den Berg A. Microfluidic DNA fragmentation for on-chip genomic analysis. Nanotechnology. 2011;22(49):494013.

251. DNA/RNA Shearing for NGS Covaris [Available from: https://www.covaris.com/dna-rna-shearing-for-ngs/.

252. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep. 2014;4:4532.

253. Grokhovsky SL, Il'icheva IA, Nechipurenko DY, Golovkin MV, Panchenko LA, Polozov RV, et al. Sequence-specific ultrasonic cleavage of DNA. Biophys J. 2011;100(1):117-25.

254. Grokhovsky SL. Specificity of DNA cleavage by ultrasound. Molecular Biology. 2006;40(2):276-83.

255. Nechipurenko YD, Golovkin MV, Nechipurenko DY, Il'icheva IA, Panchenko LA, Polozov RV, et al. Characteristics of ultrasonic cleavage of DNA. Journal of Structural Chemistry. 2009;50(5):1007-13.

256. Il'icheva IA, Nechipurenko DY, Grokhovsky SL. Ultrasonic cleavage of nicked DNA. J Biomol Struct Dyn. 2009;27(3):391-8.

257. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci U S A. 2013;110(16):6376-81.

258. Doerfler W, Böhm P. DNA Methylation: Basic Mechanisms: Springer Science & Business Media; 2006.

259. McKee JR, Christman CL, O'Brien WD, Jr., Wang SY. Effects of ultrasound on nucleic acid bases. Biochemistry. 1977;16(21):4651-4.

260. Chandran J, Aravind UK, Aravindakumar CT. Sonochemical transformation of thymidine: A mass spectrometric study. Ultrasonics Sonochemistry. 2015;27:178-86.

261. McAuley-Hecht KE, Leonard GA, Gibson NJ, Thomson JB, Watson WP, Hunter WN, et al. Crystal structure of a DNA duplex containing 8-hydroxydeoxyguanine-adenine base pairs. Biochemistry. 1994;33(34):10266-70.

262. Beard WA, Batra VK, Wilson SH. DNA polymerase structure-based insight on the mutagenic properties of 8-oxoguanine. Mutat Res. 2010;703(1):18-23.

263. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, 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 Res. 2013;41(6):e67.

264. Saito I, Nakamura T, Nakatani K, Yoshioka Y, Yamaguchi K, Sugiyama H. Mapping of the Hot Spots for DNA Damage by One-Electron Oxidation: Efficacy of GG Doublets and GGG Triplets as a Trap in Long-Range Hole Migration. Journal of the American Chemical Society. 1998;120(48):12686-7.

265. Margolin Y, Cloutier JF, Shafirovich V, Geacintov NE, Dedon PC. Paradoxical hotspots for guanine oxidation by a chemical mediator of inflammation. Nat Chem Biol. 2006;2(7):365-6.

266. Margolin Y, Shafirovich V, Geacintov NE, DeMott MS, Dedon PC. DNA sequence context as a determinant of the quantity and chemistry of guanine oxidation produced by hydroxyl radicals and one-electron oxidants. J Biol Chem. 2008;283(51):35569-78.

267. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Research. 1981;9(13):3015-27.

268. Hoheisel JD, Nizetic D, Lehrach H. Control of partial digestion combining the enzymes dam methylase and MboI. Nucleic Acids Research. 1989;17(23):9571-82.

269. Ignatov KB, Blagodatskikh KA, Shcherbo DS, Kramarova TV, Monakhova YA, Kramarov VM. Fragmentation Through Polymerization (FTP): A new method to fragment DNA for next-generation sequencing. PLoS One. 2019;14(4):e0210374.

270. Wong K-K, Markillie LM, Saffer JD. A novel method for producing partial restriction digestion of DNA fragments by PCR with 5-methyl-CTP. Nucleic Acids Research. 1997;25(20):4169-71.

271. Syed F, Grunenwald H, Caruccio N. Next-generation sequencing library preparation: simultaneous fragmentation and tagging using in vitro transposition. Nature Methods. 2009;6(11):i-ii.

272. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 2010;11(12):R119.

273. Pan CQ, Lazarus RA. Ca2+-dependent activity of human DNase I and its hyperactive variants. Protein Sci. 1999;8(9):1780-8.

274. Wiame I, Remy S, Swennen R, Sági L. Irreversible Heat Inactivation of DNase I without RNA Degradation. BioTechniques. 2000;29(2):252-6.

275. Reznikoff WS. Transposon Tn5. Annu Rev Genet. 2008;42:269-86.

276. Hostetter G, Kim SY, Savage S, Gooden GC, Barrett M, Zhang J, et al. Random DNA fragmentation allows detection of single-copy, single-exon alterations of copy number by oligonucleotide array CGH in clinical FFPE samples. Nucleic Acids Research. 2010;38(2):e9-e.

277. KAPA HyperPlus Kits: Roche; [Product Information]. Available from: https://sequencing.roche.com/global/en/products/group/kapa-hyperplus-kits.html.

278. sparQ DNA Frag & Library Prep Kit: Quantabio; [Protocol]. Available from: https://www.quantabio.com/media/contenttype/IFU-122.1_REV_04_95194_sparQ_DNA_Frag_Library_Prep_Kit_1119.pdf.

279. Nextera DNA Library Prep Reference Guide: illumina; [Protocol]. Available from: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nexteradna/nextera-dna-library-prep-reference-guide-15027987-01.pdf.

280. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One. 2011;6(11):e28240.

281. Brukner I, Jurukovski V, Savic A. Sequence-dependent structural variations of DNA revealed by DNase I. Nucleic Acids Res. 1990;18(4):891-4.

282. Brukner I, Sánchez R, Suck D, Pongor S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. Embo j. 1995;14(8):1812-8.

283. Heddi B, Abi-Ghanem J, Lavigne M, Hartmann B. Sequence-Dependent DNA Flexibility Mediates DNase I Cleavage. Journal of Molecular Biology. 2010;395(1):123-33.

284. Sasakawa C, Carle GF, Berg DE. Sequences Essential for Transposition at the Termini of IS50. Proceedings of the National Academy of Sciences of the United States of America. 1983;80(23):7293-7.

285. Johnson RC, Reznikoff WS. DNA sequences at the ends of transposon Tn5 required for transposition. Nature. 1983;304(5923):280-2.

286. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. Tn5 / IS50 Target Recognition. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(18):10716-21.

287. Shevchenko Y, Bouffard GG, Butterfield YSN, Blakesley RW, Hartley JL, Young AC, et al. Systematic sequencing of cDNA clones using the transposon Tn5. Nucleic Acids Research. 2002;30(11):2469-77.

288. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Research. 2019;26(5):391-8.

289. Segerman B, Ástvaldsson Á, Mustafa L, Skarin J, Skarin H. The efficiency of Nextera XT tagmentation depends on G and C bases in the binding motif leading to uneven coverage in bacterial species with low and neutral GC-content. Front Microbiol. 2022;13:944770.

290. Tanaka N, Takahara A, Hagio T, Nishiko R, Kanayama J, Gotoh O, et al. Sequencing artifacts derived from a library preparation method using enzymatic fragmentation. PLoS One. 2020;15(1):e0227427.

291. Zhang A, Li S, Apone L, Sun X, Chen L, Ettwiller LM, et al. Solid-phase enzyme catalysis of DNA end repair and 3' A-tailing reduces GC-bias in next-generation sequencing of human genomic DNA. Sci Rep. 2018;8(1):15887.

292. Ausebel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, et al. Enzymatic Manipulation of DNA and RNA. Current Protocols in Molecular Biology: John Wiley & Sons, Inc.; 2003.

293. Klenow H, Overgaard-Hansen K. Proteolytic cleavage of DNA polymerase from Escherichia Coli B into an exonuclease unit and a polymerase unit. FEBS Lett. 1970;6(1):25-7.

294. Lodes M. Chimera-Free Library Prep for NGS Platforms. Genetic Engineering & Biotechnology News. 2012;32(2):20-1.

295. Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, Prado JL, et al. Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. PLOS ONE. 2013;8(10):e78575.

296. Keohavong P, Thilly WG. Fidelity of DNA polymerases in DNA amplification. Proc Natl Acad Sci U S A. 1989;86(23):9253-7.

297. Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. PCR Methods Appl. 1991;1(1):17-24.

298. Bertram JG, Oertell K, Petruska J, Goodman MF. DNA polymerase fidelity: comparing direct competition of right and wrong dNTP substrates with steady state and pre-steady state kinetics. Biochemistry. 2010;49(1):20-8.

299. Luo J, Bergstrom DE, Barany F. Improving the fidelity of Thermus thermophilus DNA ligase. Nucleic Acids Res. 1996;24(15):3071-8.

300. Nilsson SV, Magnusson G. Sealing of gaps in duplex DNA by T4 DNA ligase. Nucleic Acids Res. 1982;10(5):1425-37.

301. Goffin C, Bailly V, Verly WG. Nicks 3' or 5' to AP sites or to mispaired bases, and one-nucleotide gaps can be sealed by T4 DNA ligase. Nucleic Acids Res. 1987;15(21):8755-71.

302. Wu DY, Wallace RB. Specificity of the nick-closing activity of bacteriophage T4 DNA ligase. Gene. 1989;76(2):245-54.

303. Barril P, Nates S, Reddy PR, Raju N, Yilmaz M, Ozic C, et al. Gel Electrophoresis. Rijeka: IntechOpen; 2012.

304. DeAngelis MM, Wang DG, Hawkins TL. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 1995;23(22):4742-3.

305. Quail MA, Gu Y, Swerdlow H, Mayho M. Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. Electrophoresis. 2012;33(23):3521-8.

306. Lennon NJ, Lintner RE, Anderson S, Alvarez P, Barry A, Brockman W, et al. A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biology. 2010;11(2):R15.

307. Quail MA, Swerdlow H, Turner DJ. Improved protocols for the illumina genome analyzer sequencing system. Curr Protoc Hum Genet. 2009;Chapter 18:Unit 18.2.

308. Liu D, Li Q, Luo J, Huang Q, Zhang Y. An SPRI beads-based DNA purification strategy for flexibility and cost-effectiveness. BMC Genomics. 2023;24(1):125.

309. Borgström E, Lundin S, Lundeberg J. Large scale library generation for high throughput sequencing. PLoS One. 2011;6(4):e19119.

310. Lundin S, Stranneheim H, Pettersson E, Klevebring D, Lundeberg J. Increased throughput by parallelization of library preparation for massive sequencing. PLoS One. 2010;5(4):e10029.

311. Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biology. 2011;12(1):R1.

312. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27(2):182-9.

313. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nature Genetics. 2007;39(12):1522-7.

314. Wang VG, Kim H, Chuang JH. Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts. PLoS One. 2018;13(10):e0204912.

315. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol. 2001;19(4):342-7.

316. Relógio A, Schwager C, Richter A, Ansorge W, Valcárcel J. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Res. 2002;30(11):e51.

317. Kolpashchikov DM. A binary DNA probe for highly specific nucleic Acid recognition. J Am Chem Soc. 2006;128(32):10625-8.

318. Tijssen P. Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization with Nucleic Acid Probes. Theory and Nucleic Acid Preparation: Elsevier; 1993.

319. Chou CC, Chen CH, Lee TT, Peck K. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Res. 2004;32(12):e99.

320. Liu WT, Guo H, Wu JH. Effects of target length on the hybridization efficiency and specificity of rRNA-based oligonucleotide microarrays. Appl Environ Microbiol. 2007;73(1):73-82.

321. Evertsz EM, Au-Young J, Ruvolo MV, Lim AC, Reynolds MA. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. Biotechniques. 2001;31(5):1182, 4, 6 passim.

322. Xu W, Bak S, Decker A, Paquette SM, Feyereisen R, Galbraith DW. Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of Arabidopsis thaliana. Gene. 2001;272(1-2):61-74.

323. Wang H, Li J, Liu H, Liu Q, Mei Q, Wang Y, et al. Label-free hybridization detection of a single nucleotide mismatch by immobilization of molecular beacons on an agarose film. Nucleic Acids Res. 2002;30(12):e61.

324. Yao G, Tan W. Molecular-beacon-based array for sensitive DNA analysis. Anal Biochem. 2004;331(2):216-23.

325. Wang X, Yun W, Dong P, Zhou J, He P, Fang Y. A controllable solid-state Ru(bpy)(3)(2+) electrochemiluminescence film based on conformation change of ferrocene-labeled DNA molecular beacon. Langmuir. 2008;24(5):2200-5.

326. Dave N, Liu J. Fast molecular beacon hybridization in organic solvents with improved target specificity. J Phys Chem B. 2010;114(47):15694-9.

327. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239(4839):487-91.

328. DeLong EF, Wickham GS, Pace NR. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. Science. 1989;243(4896):1360-3.

329. Amann RI, Krumholz L, Stahl DA. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. J Bacteriol. 1990;172(2):762-70.

330. Peterson AW, Heaton RJ, Georgiadis RM. The effect of surface probe density on DNA hybridization. Nucleic Acids Res. 2001;29(24):5163-8.

331. Sartor M, Schwanekamp J, Halbleib D, Mohamed I, Karyala S, Medvedovic M, et al. Microarray results improve significantly as hybridization approaches equilibrium. BioTechniques. 2004;36(5):790-6.

332. Valko L, Gersi P. An approximate approach to DNA denaturation. Gen Physiol Biophys. 1995;14(6):491-502.

333. Gharaibeh RZ, Fodor AA, Gibas CJ. Software note: using probe secondary structure information to enhance Affymetrix GeneChip background estimates. Comput Biol Chem. 2007;31(2):92-8.

334. Chauhan K, Singh AR, Kumar S, Granek R. Can one detect intermediate denaturation states of DNA sequences by following the equilibrium open–close dynamic fluctuations of a single base pair? The Journal of Chemical Physics. 2022;156(16).

335. Cederquist KB, Stoermer Golightly R, Keating CD. Molecular beacon-metal nanowire interface: effect of probe sequence and surface coverage on sensor performance. Langmuir. 2008;24(16):9162-71.

336. Lima WF, Monia BP, Ecker DJ, Freier SM. Implication of RNA structure on antisense oligonucleotide hybridization kinetics. Biochemistry. 1992;31(48):12055-61.

337. Gao Y, Wolf LK, Georgiadis RM. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. Nucleic Acids Res. 2006;34(11):3370-7.

338. Archer MJ, Long N, Lin B. Effect of probe characteristics on the subtractive hybridization efficiency of human genomic DNA. BMC Research Notes. 2010;3(1):109.

339. Chen X, Liu N, Liu L, Chen W, Chen N, Lin M, et al. Thermodynamics and kinetics guided probe design for uniformly sensitive and specific DNA hybridization without optimization. Nature Communications. 2019;10(1):4675.

340. Guo Y, Long J, He J, Li C-I, Cai Q, Shu X-O, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012;13(1):194.

341. Lorenz TC. Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. J Vis Exp. 2012(63):e3998.

342. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 2015;43(21):e143.

343. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12(2):R18.

344. Nagai S, Sildever S, Nishi N, Tazawa S, Basti L, Kobayashi T, et al. Comparing PCR-generated artifacts of different polymerases for improved accuracy of DNA metabarcoding. Metabarcoding and Metagenomics. 2022;6:27-39.

345. Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, et al. Minimizing polymerase biases in metabarcoding. Molecular Ecology Resources. 2018;18(5):927-39.

346. Stein A, Takasuka TE, Collings CK. Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? Nucleic Acids Res. 2010;38(3):709-19.

347. Choi JS, Kim JS, Joe CO, Kim S, Ha KS, Park YM. Improved cycle sequencing of GC-rich DNA template. Exp Mol Med. 1999;31(1):20-4.

348. Henke W, Herdel K, Jung K, Schnorr D, Loening SA. Betaine Improves the PCR Amplification of GC-Rich DNA Sequences. Nucleic Acids Research. 1997;25(19):3957-8.

349. Chakrabarti R, Schutt CE. The enhancement of PCR amplification by low molecular-weight sulfones. Gene. 2001;274(1):293-8.

350. Yi SUN, Hegamyer G, Colburn NH. PCR-direct sequencing of a GC-rich region by inclusion of 10 % DMSO : application to mouse c-jun. BioTechniques. 1993;15(3):372-4.

351. Weissensteiner T, Lanchbury JS. Strategy for Controlling Preferential Amplification and Avoiding False Negatives in PCR Typing. BioTechniques. 1996;21(6):1102-8.

352. Turner SL, Jenkins FJ. Use of deoxyinosine in PCR to improve amplification of GC-rich DNA. Biotechniques. 1995;19(1):48-52.

353. Pratyush DD, Tiwari S, Kumar A, Singh SK. A new approach to touch down method using betaine as co-solvent for increased specificity and intensity of GC rich gene amplification. Gene. 2012;497(2):269-72.

354. McInerney P, Adams P, Hadi MZ. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. Mol Biol Int. 2014;2014:287430.

355. Klug J, Wolf M, Beato M. Creating chimeric molecules by PCR directed homologous DNA recombination. Nucleic Acids Res. 1991;19(10):2793.

356. Bradley RD, Hillis DM. Recombinant DNA sequences generated by PCR amplification. Molecular Biology and Evolution. 1997;14(5):592-3.

357. Odelberg SJ, Weiss RB, Hata A, White R. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res. 1995;23(11):2049-57.

358. Lahr DJ, Katz LA. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques. 2009;47(4):857-66.

359. Wang GCY, Wang Y. The frequency of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from different bacterial species. Microbiology (Reading). 1996;142 (Pt 5):1107-14.

360. Liu J, Song H, Liu D, Zuo T, Lu F, Zhuang H, et al. Extensive recombination due to heteroduplexes generates large amounts of artificial gene fragments during PCR. PLoS One. 2014;9(9):e106658.

361. Pääbo S, Irwin DM, Wilson AC. DNA damage promotes jumping between templates during enzymatic amplification. J Biol Chem. 1990;265(8):4718-21.

362. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. Nucleic Acids Research. 2019;47(2):e12-e.

363. Beagan JJ, Drees EEE, Stathi P, Eijk PP, Meulenbroeks L, Kessler F, et al. PCR-Free Shallow Whole Genome Sequencing for Chromosomal Copy Number Detection from Plasma of Cancer Patients Is an Efficient Alternative to the Conventional PCR-Based Approach. J Mol Diagn. 2021;23(11):1553-63.

364. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, et al. Optimal enzymes for amplifying sequencing libraries. Nature Methods. 2012;9(1):10-1.

365. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nature Biotechnology. 2009;27(11):1025-31.

366. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

367. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci U S A. 2006;103(52):19635-40.

368. Illumina Sequencing Technology: Illumina; 2010 [Available from: https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.

369. An Introduction to Next-Generation Sequencing Technology: Illumina; 2017 [Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

370. Calling Sequencing SNPs: Illumina; 2010 [Available from: https://www.illumina.com/Documents/products/technotes/technote_snp_caller_sequencing.pdf.

371. Ye C, Hsiao C, Corrada Bravo H. BlindCall: ultra-fast base-calling of high-throughput sequencing data by blind deconvolution. Bioinformatics. 2014;30(9):1214-9.

372. Cacho A, Smirnova E, Huzurbazar S, Cui X. A Comparison of Base-calling Algorithms for Illumina Sequencing Technology. Briefings in Bioinformatics. 2016;17(5):786-95.

373. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature Methods. 2009;6(4):291-5.

374. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72.

375. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol. 2015;76(2-3):166-75.

376. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, et al. Comparison of Sample Preparation Methods Used for the Next-Generation Sequencing of Mycobacterium tuberculosis. PLoS One. 2016;11(2):e0148676.

377. Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics. 2012;13(1):666.

378. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research. 2008;36(16):e105.

379. How to achieve more consistent cluster density on Illumina sequencing platforms: Illumina; 2023 [updated 19-05. Available from: https://knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000001481.

380. Nextera Library Validation and Cluster Density Optimization: Illumina; 2014 [Available from: https://www.illumina.com/documents/products/technotes/technote_nextera_library_validation.pdf.

381. Andrews S. Illumina 2 colour chemistry can overcall high confidence G bases. QC Fail. 2016.

382. Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. 2009;10(8):R83.

383. Wang B, Wan L, Wang A, Li LM. An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters. Scientific Reports. 2017;7(1):41348.

384. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinform. 2021;3(1):lqab019.

385. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39(13):e90.

386. Using a PhiX Control for HiSeq® Sequencing Runs: Illumina; 2013 [Available from: https://www.illumina.com/content/dam/illumina-support/documents/products/technotes/technote_phixcontrolv3.pdf.

387. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports. 2018;8(1):10950.

388. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research. 2015;43(6):e37-e.

389. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17(1):125.

390. Victoria Wang X, Blades N, Ding J, Sultana R, Parmigiani G. Estimation of sequencing error rates in short reads. BMC Bioinformatics. 2012;13(1):185.

391. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics. 2011;12:451.

392. Illumina CMOS Chip and One-Channel SBS Chemistry Illumina2018 [Available from: https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf.

393. Faster sequencing and data processing: Illumina; [Available from: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html.

394. Widengren J, Rigler R. Mechanisms of photobleaching investigated by fluorescence correlation spectroscopy. Bioimaging. 1996;4(3):149-57.

395. Yu Y, Chen Q, Wen L, Hu X, Zhang H-F. Spatial optical crosstalk in CMOS image sensors integrated with plasmonic color filters. Opt Express. 2015;23(17):21994-2003.

396. Frank MS, Fuß J, Steiert TA, Streleckiene G, Gehl J, Forster M. Quantifying sequencing error and effective sequencing depth of liquid biopsy NGS with UMI error correction. BioTechniques. 2021;70(4):226-32.

397. Streubel A, Stenzinger A, Stephan-Falkenau S, Kollmeier J, Misch D, Blum TG, et al. Comparison of different semi-automated cfDNA extraction methods in combination with UMI-based targeted sequencing. Oncotarget. 2019;10(55):5690-702.

398. Yukai H, Nan H, Rui L, Bing W, Xiaohong D, Chunyan Y, et al. Ultra-deep sequencing with unique molecular identifier(UMI) for detection of ctDNA by fragment profiling using machine learning. Journal of Clinical Oncology. 2022;40($16_{s}uppl$) : $e15508 - e$.

399. Filges S, Yamada E, Ståhlberg A, Godfrey TE. Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes. Sci Rep. 2019;9(1):3503.

400. Sequencing Accuracy with Unique Molecular Identifiers: Illumina; [Available from: https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/unique-molecular-identifiers.html.

401. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 2016;17(7):239.

402. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017;27(3):491-9.

403. Mitchell K, Brito JJ, Mandric I, Wu Q, Knyazev S, Chang S, et al. Benchmarking of computational error-correction methods for next-generation sequencing data. Genome Biology. 2020;21(1):71.

404. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

405. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.

406. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Ithaca: Cornell University Library, arXiv.org; 2013.

407. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357-9.

408. Sirén J, Välimäki N, Mäkinen V. Indexing graphs for path queries with applications in genome research. IEEE/ACM Trans Comput Biol Bioinformatics. 2014;11(2):375–88.

409. Bushnell B. BBMap short read aligner, and other bioinformatic tools. SourceForge2022 [updated 10-06. Available from: https://sourceforge.net/projects/bbmap/.

410. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nature Methods. 2012;9(12):1185-8.

411. Grzegorz MB, Jean T-M, Danielle T-M, Ben B, Thomas LM. Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. bioRxiv. 2018:390013.

412. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.

413. Powerful tool designed for mapping of short reads onto a reference genome from Illumina, Ion Torrent, and 454 NGS platforms.: NovoCraft; [Available from: https://www.novocraft.com/products/novoalign/.

414. Siragusa E, Weese D, Reinert K. Fast and accurate read mapping with approximate seeds and multiple backtracking. Nucleic Acids Research. 2013;41(7):e78-e.

415. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics. 2011;27(18):2518-28.

416. Otto C, Stadler PF, Hoffmann S. Lacking alignments? The next-generation sequencing mapper segemehl revisited. Bioinformatics. 2014;30(13):1837-43.

417. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

418. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research. 2019;47(8):e47-e.

419. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 2013;14(4):R36.

420. Musich R, Cadle-Davidson L, Osier MV. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Frontiers in Plant Science. 2021;12.

421. Donato L, Scimone C, Rinaldi C, D'Angelo R, Sidoti A. New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies. Neural Comput Appl. 2021;33(22):15669-92.

422. Kumar S, Agarwal S, Ranvijay. Fast and memory efficient approach for mapping NGS reads to a reference genome. Journal of Bioinformatics and Computational Biology. 2019;17(02):1950008.

423. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV. Benchmarking short sequence mapping tools. BMC Bioinformatics. 2013;14(1):184.

424. Kim J, Ji M, Yi G. A Review on Sequence Alignment Algorithms for Short Reads Based on Next-Generation Sequencing. IEEE Access. 2020;8:189811-22.

425. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195-7.

426. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443-53.

427. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

428. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

429. Langmead B, Kim D, Charles R, Chen N-C, Wilks C, Antonescu V. Bowtie 2 Fast and sensitive read alignment [Manual]. Available from: https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml.

430. Li H. Manual Reference Pages - bwa 2013 [updated 08-03. Manual]. Available from: https://bio-bwa.sourceforge.net/bwa.shtml.

431. Ferragina P, Manzini G. Indexing compressed text. Journal of the ACM. 2005;52(4):552-81.

432. Ferragina P, Manzini G, editors. Opportunistic data structures with applications. Proceedings 41st Annual Symposium on Foundations of Computer Science; 2000 12-14 Nov. 2000.

433. Burrows M. A block-sorting lossless data compression algorithm. SRC Research Report, 124. 1994.

434. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience. 2020;9(2):giaa008.

435. Edgar R, Haas B, Clemente J, Quince C, Knight R. UCHIIME improves sensitivity and speed of chimera detection. Bioinformatics (Oxford, England). 2011;27:2194-200.

436. Pervez MT, Babar ME, Nadeem A, Aslam M, Awan AR, Aslam N, et al. Evaluating the Accuracy and Efficiency of Multiple Sequence Alignment Methods. Evolutionary Bioinformatics. 2014;10:EBO.S19199.

437. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One. 2013;8(4):e62856.

438. Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol. 2020;21(1):250.

439. Chen NC, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. Genome Biol. 2021;22(1):8.

440. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS genetics. 2019;15(7):e1008302-e.

441. Oliva A, Tobler R, Cooper A, Llamas B, Souilmi Y. Systematic benchmark of ancient DNA read mapping. Briefings in bioinformatics. 2021;22(5).

442. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27(5):849-64.

443. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604(7906):437-46.

444. Nguyen N, Hickey G, Zerbino DR, Raney B, Earl D, Armstrong J, et al. Building a Pan-Genome Reference for a Population. Journal of Computational Biology. 2015;22(5):387-401.

445. Adam MN, Erik G, Benedict P. A Graph Extension of the Positional Burrows-Wheeler Transform and its Applications. bioRxiv. 2016:051409.

446. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, et al. Building the sequence map of the human pan-genome. Nature Biotechnology. 2010;28(1):57-63.

447. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312-24.

448. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

449. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. 2018. p. 875-9.

450. Jain C, Zhang H, Gao Y, Aluru S. On the Complexity of Sequence-to-Graph Alignment. Journal of Computational Biology. 2020;27(4):640-54.

451. Feng Z, Luo Q. Accelerating Sequence-to-Graph Alignment on Heterogeneous Processors. Proceedings of the 50th International Conference on Parallel Processing; Lemont, IL, USA: Association for Computing Machinery; 2021. p. Article 26.

452. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189-201.

453. Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biology. 2016;17(1):178.

454. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013;31(3):213-9.

455. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28(3):311-7.

456. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nature Methods. 2018;15(8):591-4.

457. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Research. 2016;44(11):e108-e.

458. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76.

459. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018;16:15-24.

460. Mutect2 Broad Institute: GATK Team; 2023 [updated 20-03. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/13832710384155-Mutect2.

461. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981;17(6):368-76.

462. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851-8.

463. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244.

464. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep. 2017;7:43169.

465. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 2013;5(10):91.

466. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Medicine. 2013;5(3):28.

467. Roberts ND, Kortschak RD, Parker WT, Schreiber AW, Branford S, Scott HS, et al. A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics. 2013;29(18):2223-30.

468. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, et al. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J Mol Diagn. 2014;16(1):75-88.

469. Cai L, Yuan W, Zhang Z, He L, Chou KC. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep. 2016;6:36540.

470. Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLoS One. 2016;11(3):e0151664.

471. Wang Q, Kotoula V, Hsu P-C, Papadopoulou K, Ho JWK, Fountzilas G, et al. Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data. BMC Medical Genomics. 2019;12(9):181.

472. Vijayan V, Yiu SM, Zhang L. Improving somatic variant identification through integration of genome and exome data. BMC Genomics. 2017;18(Suppl 7):748.

473. Goode DL, Hunter SM, Doyle MA, Ma T, Rowley SM, Choong D, et al. A simple consensus approach improves somatic mutation prediction accuracy. Genome Med. 2013;5(9):90.

474. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med Genomics. 2019;12(1):63.

475. Callari M, Sammut S-J, De Mattos-Arruda L, Bruna A, Rueda OM, Chin S-F, et al. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. Genome Medicine. 2017;9(1):35.

476. Kim SY, Jacob L, Speed TP. Combining calls from multiple somatic mutation-callers. BMC Bioinformatics. 2014;15:154.

477. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics. 2014;15:104.

478. Huang W, Guo YA, Muthukumar K, Baruah P, Chang MM, Jacobsen Skanderup A. SMuRF: portable and accurate ensemble prediction of somatic mutations. Bioinformatics. 2019;35(17):3157-9.

479. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biology. 2015;16(1):197.

480. Wang M, Luo W, Jones K, Bian X, Williams R, Higson H, et al. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci Rep. 2020;10(1):12898.

481. Ainscough BJ, Barnell EK, Ronning P, Campbell KM, Wagner AH, Fehniger TA, et al. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Nature Genetics. 2018;50(12):1735-43.

482. Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, et al. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. Bioinformatics. 2012;28(2):167-75.

483. Spinella J-F, Mehanna P, Vidal R, Saillour V, Cassart P, Richer C, et al. SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genomics. 2016;17(1):912.

484. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional neural networks for accurate somatic mutation detection. Nature Communications. 2019;10(1):1041.

485. Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nature Biotechnology. 2021;39(9):1141-50.

486. Zviran A, Schulman RC, Shah M, Hill STK, Deochand S, Khamnei CC, et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. Nat Med. 2020;26(7):1114-24.

487. Wan JCM, Heider K, Gale D, Murphy S, Fisher E, Mouliere F, et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. Science Translational Medicine. 2020;12(548):eaaz8084.

488. Christensen MH, Drue SO, Rasmussen MH, Frydendahl A, Lyskjær I, Demuth C, et al. DREAMS: deep read-level error model for sequencing data applied to low-frequency variant calling and circulating tumor DNA detection. Genome Biol. 2023;24(1):99.

489. Koldby KM. Circulating tumor DNA and structural variation detection in patients with gastroesophageal cancer: University of Southern Denmark; 2020.

490. Van G, O'Connor BD. Genomics in the cloud : using Docker, GATK, and WDL in Terra: O'reilly Media; 2020.

491. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 2000;28(1):352-5.

492. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16(9):1182-90.

493. Gates C. Connor [Available from: https://github.com/umich-brcf-bioinf/Connor.

494. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.

495. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.

496. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032-4.

497. Wilm A. LoFreq Fast and sensitive variant calling from next-gen sequencing data [Documentation]. Available from: https://csb5.github.io/lofreq/commands/.

498. MuSE [Documentation]. Available from: https://github.com/wwylab/MuSE.

499. Mutect2 [Documentation]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360056969692-Mutect2.

500. Larson DE, Abbott TE, Harris CC. SomaticSniper User Manual 2011 [updated 26-10. Documentation]. Available from: https://gmt.genome.wustl.edu/packages/somatic-sniper/documentation.html.

501. Saunders C, Bekritsky M. Strelka User Guide 2018 [Documentation]. Available from: https://github.com/Illumina/strelka/blob/v2.9.x/docs/userGuide/README.md.

502. VarDictJava [Documentation]. Available from: https://github.com/AstraZeneca-NGS/VarDictJava.

503. VarScan [Documentation]. Available from: https://varscan.sourceforge.net/somatic-calling.html.

504. Caetano-Anolles D. Somatic short variant discovery (SNVs + Indels) 2023 [updated 20-03. Workflow]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels.

505. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008.

506. Khanna A, Larson DE, Srivatsan SN, Mosior M, Abbott TE, Kiwala S, et al. Bam-readcount – rapid generation of basepair-resolution sequence metrics. ArXiv. 2021.

507. Kiwala S, Miller C. VCF Annotation Tools (VAtools) [Available from: https://vatools.readthedocs.io/en/latest/.

508. Seshan VE, Olshen A. DNAcopy: DNA copy number data analysis [Available from: https://www.bioconductor.org/packages/release/bioc/html/DNAcopy.html.

509. Ha G. Snakemake workflow for TITAN [Workflow]. Available from: https://github.com/gavinha/TitanCNA/tree/master/scripts/snakemake.

510. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications. 2017;8(1):1324.

511. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014;24(11):1881-93.

512. Lai D, Ha G. HMMcopy: A package for bias-free copy number estimation and robust CNA detection in tumour samples from WGS HTS data 2023 [Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/HMMcopy/inst/doc/HMMcopy.pdf.

513. Wilson DJ. The harmonic mean p-value for combining dependent tests. Proc Natl Acad Sci U S A. 2019;116(4):1195-200.

514. Sahraeian SME, Fang LT, Karagiannis K, Moos M, Smith S, Santana-Quintero L, et al. Achieving robust somatic mutation detection with deep learning models derived from reference data sets of a cancer sample. Genome Biology. 2022;23(1):12.

515. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biology. 2019;20(1):246.

516. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455(7216):1069-75.

517. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008;321(5897):1801-6.

518. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69(16):6660-7.

519. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013;29(5):647-8.

520. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS One. 2013;8(10):e77945.

521. Han Y, Yang J, Qian X, Cheng WC, Liu SH, Hua X, et al. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic Acids Res. 2019;47(8):e45.

522. Ren F, Ding X, Zheng M, Korzinkin M, Cai X, Zhu W, et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. Chemical Science. 2023;14(6):1443-52.

523. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

524. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590-6.

525. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Applying and improving AlphaFold at CASP14. Proteins: Structure, Function, and Bioinformatics. 2021;89(12):1711-21.

526. Brogi S, Ramalho TC, Kuca K, Medina-Franco JL, Valko M. Editorial: In silico Methods for Drug Design and Discovery. Frontiers in Chemistry. 2020;8.

527. Braunstein V, Burnett G. GPU-Accelerated Tools Added to NVIDIA Clara Parabricks v36 for Cancer and Germline Analyses [Internet]. NVIDIA Developer Technical Blog 2021. [cited 2023]. Available from: https://developer.nvidia.com/blog/gpu-accelerated-tools-added-to-nvidia-clara-parabricks-v3-6-for-cancer-and-germline-analyses/.

528. Kennedy K. AMD EPYC 7742 Benchmarks and Review Simply Peerless 2019 [updated 09-12. Available from: https://www.servethehome.com/amd-epyc-7742-benchmarks-and-review-simply-peerless/.

529. NVIDIA A100 PCIe 80 GB [technical specifications]. Available from: https://www.techpowerup.com/gpu-specs/a100-pcie-80-gb.c3821.

Appendix

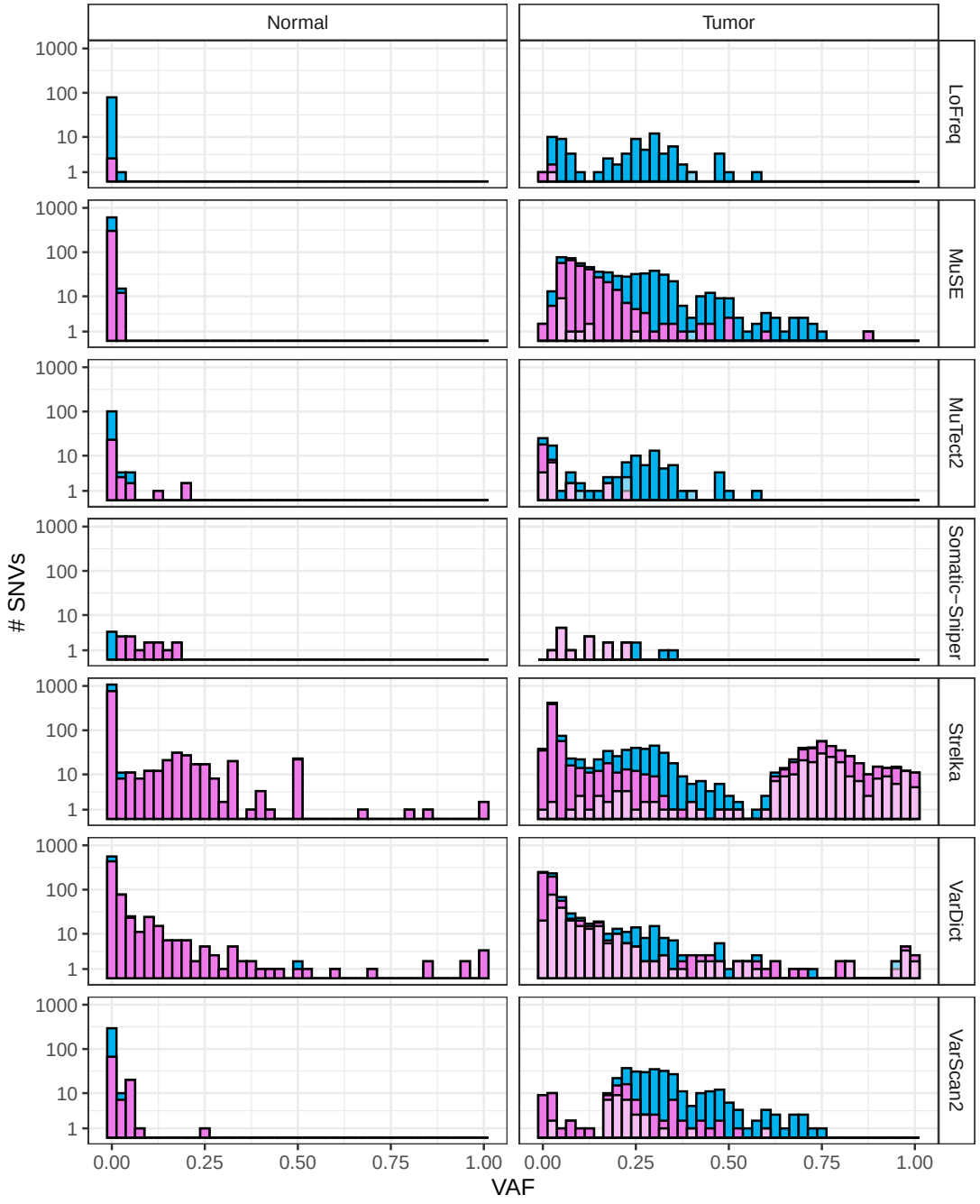

Figure A.1: Histograms of the VAF of the SNVs for each variant caller in the merged tumor resection sample, tumor, and the matched germline sample, normal, for the patient, PC1-14. SNVs identified by the variant caller are color blue, whereas SNVs unique to the variant caller is colored magenta. SNVs with a VAF greater than zero in the matched germline sample are plotted on top with a transparent white color, and thus appear in a lighter magenta and blue color. The VAFs are calculated from reads with a minimum mapping quality of 10 and a minimum base quality of 30. The y-axis is pseudo $\log_{10}$ scaled to properly illustrate the SNV counts of each bin.

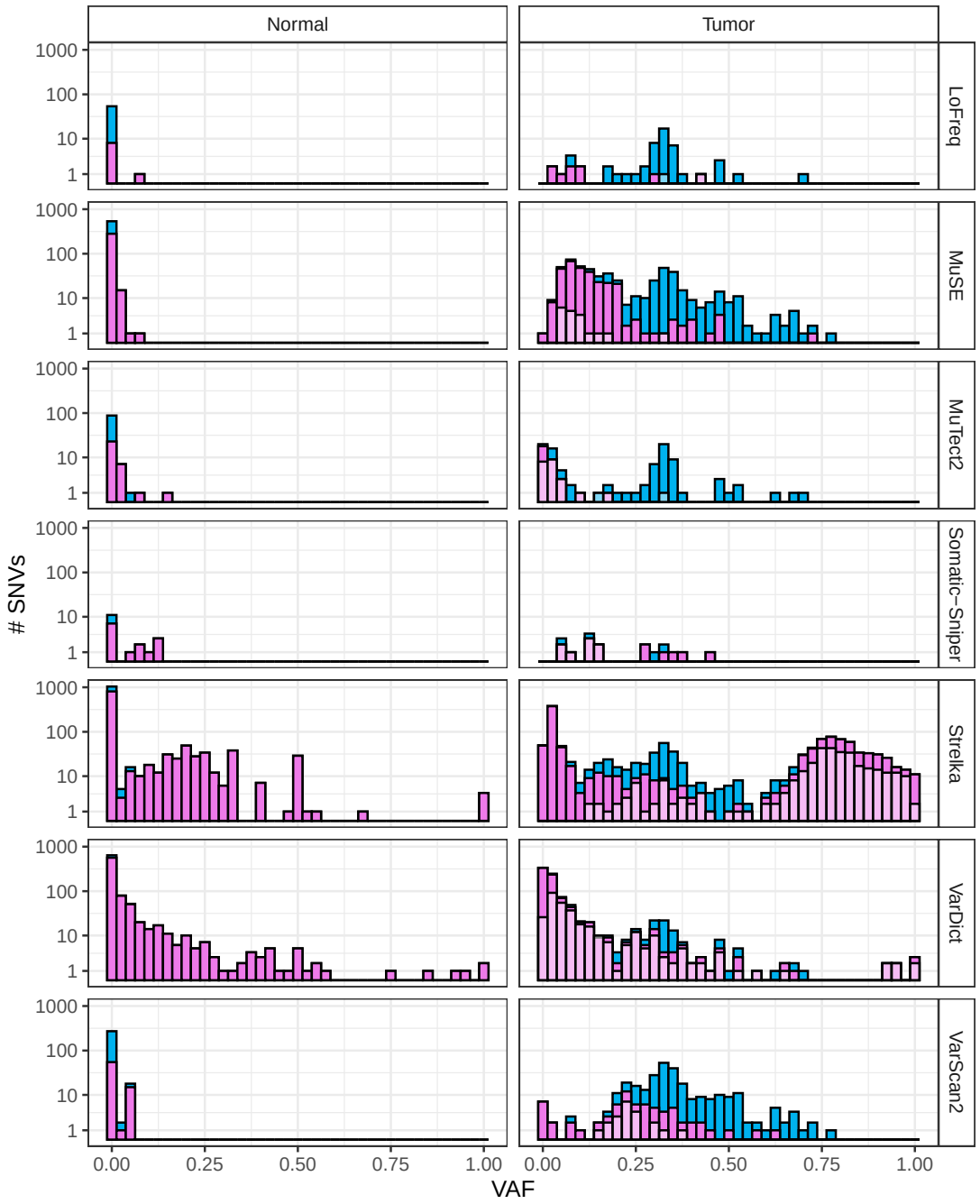

Figure A.2: Histograms of the VAF of the SNVs for each variant caller in the merged tumor resection sample, tumor, and the matched germline sample, normal, for the patient, PC1-18. SNVs identified by the variant caller are color blue, whereas SNVs unique to the variant caller is colored magenta. SNVs with a VAF greater than zero in the matched germline sample are plotted on top with a transparent white color, and thus appear in a lighter magenta and blue color. The VAFs are calculated from reads with a minimum mapping quality of 10 and a minimum base quality of 30. The y-axis is pseudo $\log_{10}$ scaled to properly illustrate the SNV counts of each bin.

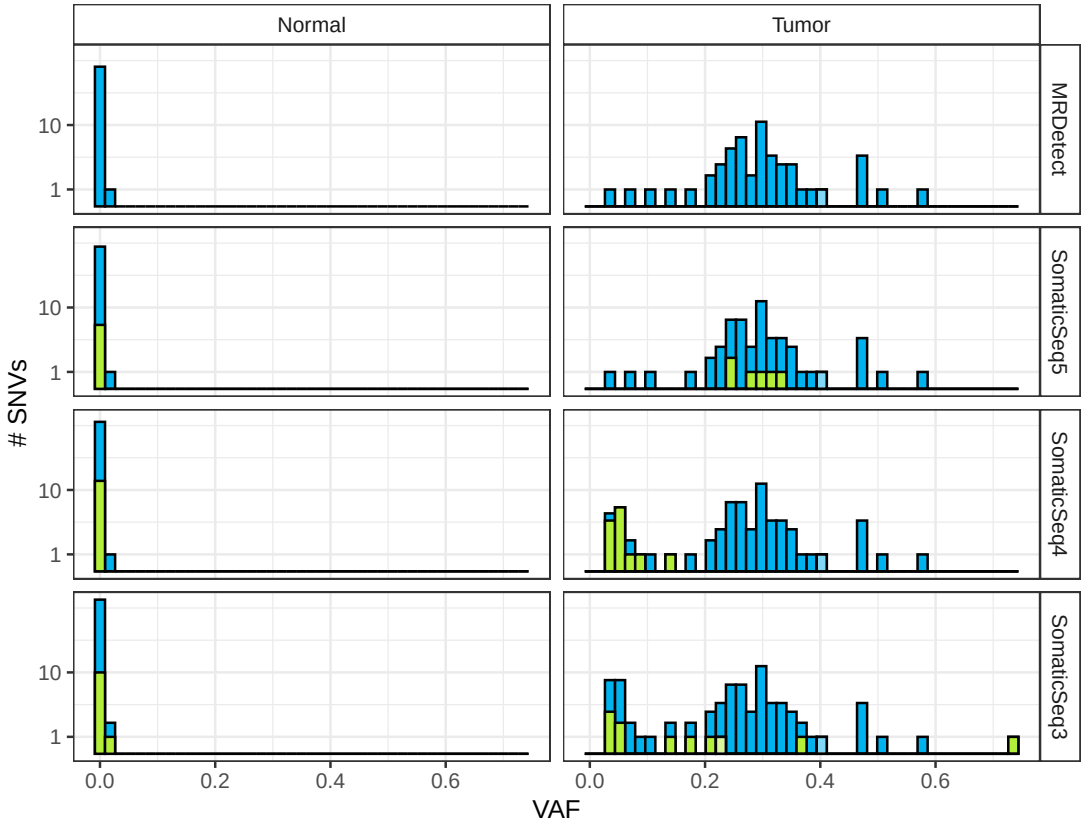

Figure A.3: Histograms of the VAF of the SNVs for each ensemble in the merged tumor resection sample, tumor, and the matched germline sample, normal, for the patient, PC1-14. SNVs identified by the ensembles are colored blue. SNVs not identified by the above ensemble are colored green. SNVs with a VAF greater than zero in the matched germline sample are plotted on top with a transparent white color, and thus appear in a lighter blue and green color. The VAFs are calculated from reads with a minimum mapping quality of 10 and a minimum base quality of 30. The y-axis is pseudo $\log_{10}$ scaled to properly illustrate the SNV counts of each bin.

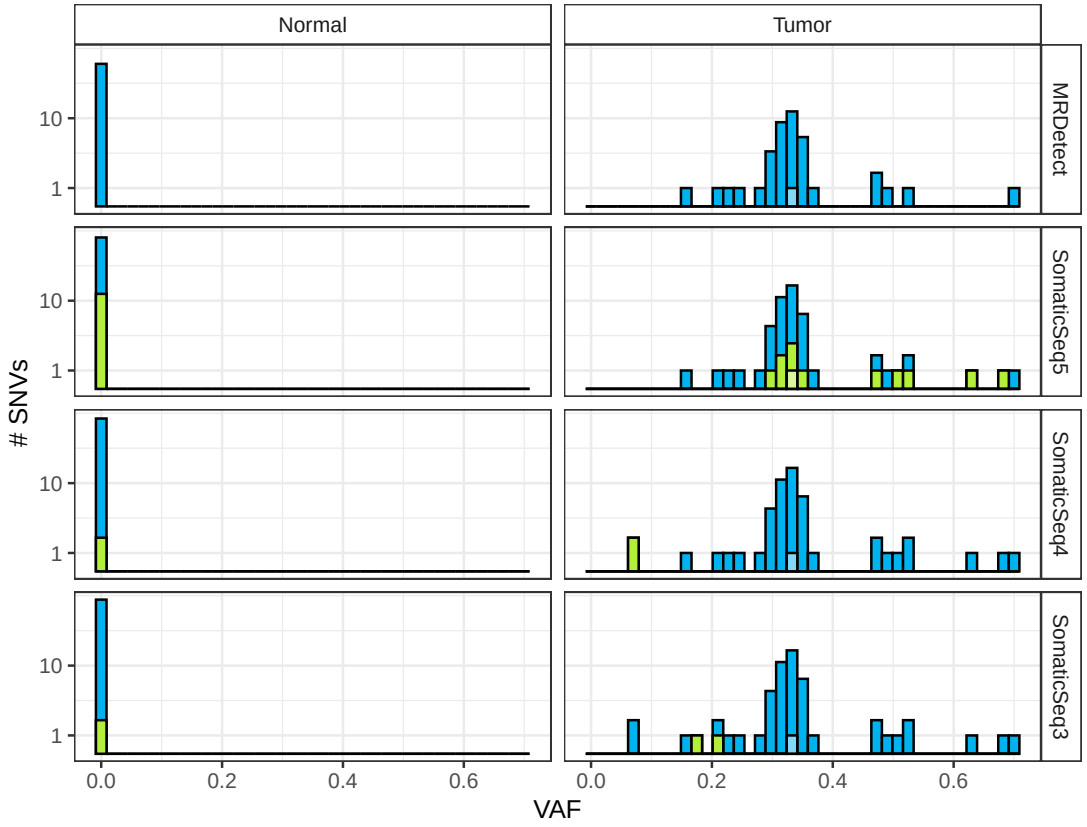

Figure A.4: Histograms of the VAF of the SNVs for each ensemble in the merged tumor resection sample, tumor, and the matched germline sample, normal, for the patient, PC1-18. SNVs identified by the ensembles are colored blue. SNVs not identified by the above ensemble are colored green. SNVs with a VAF greater than zero in the matched germline sample are plotted on top with a transparent white color, and thus appear in a lighter blue and green color. The VAFs are calculated from reads with a minimum mapping quality of 10 and a minimum base quality of 30. The y-axis is pseudo $\log_{10}$ scaled to properly illustrate the SNV counts of each bin.

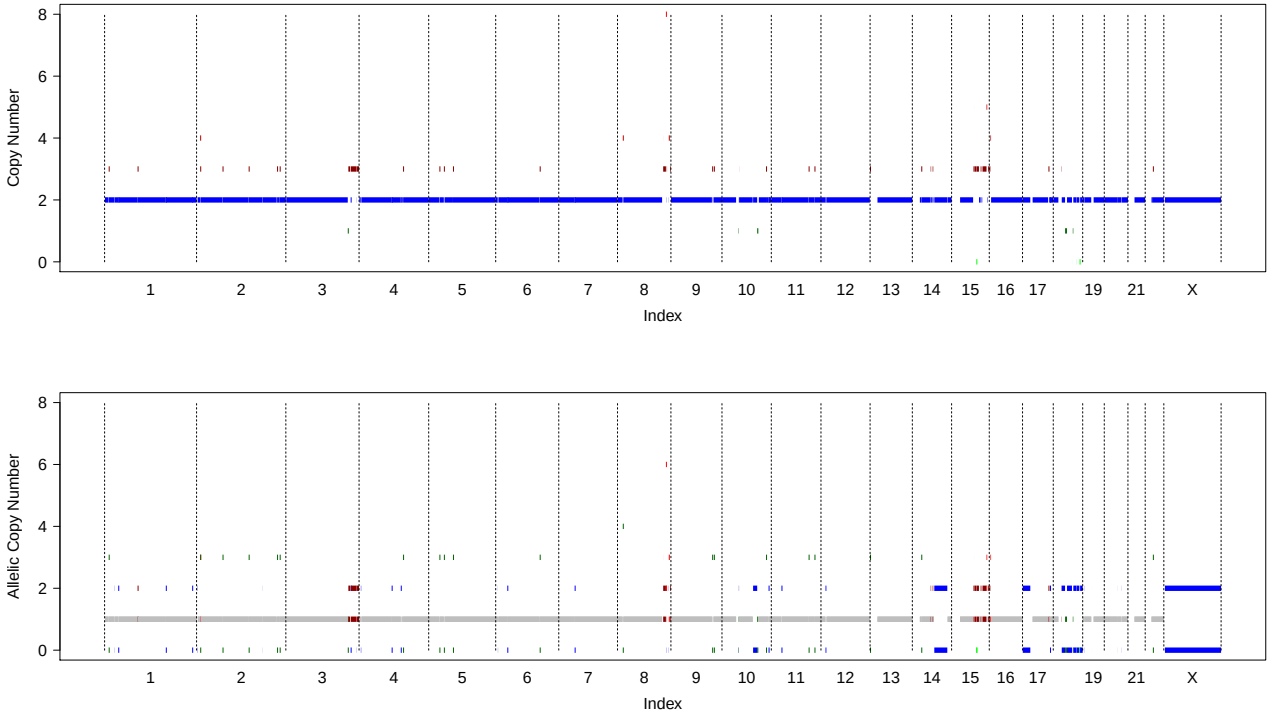

Figure A.5: Plots illustrating the copy numbers (top) and allelic copy numbers (bottom) for the chromosomes of the patient PC1-10, calculated with a ichorCNA-TitanCNA workflow. LOH, copy neutral LOH, and amplifications are colored green, blue, and red, respectively.

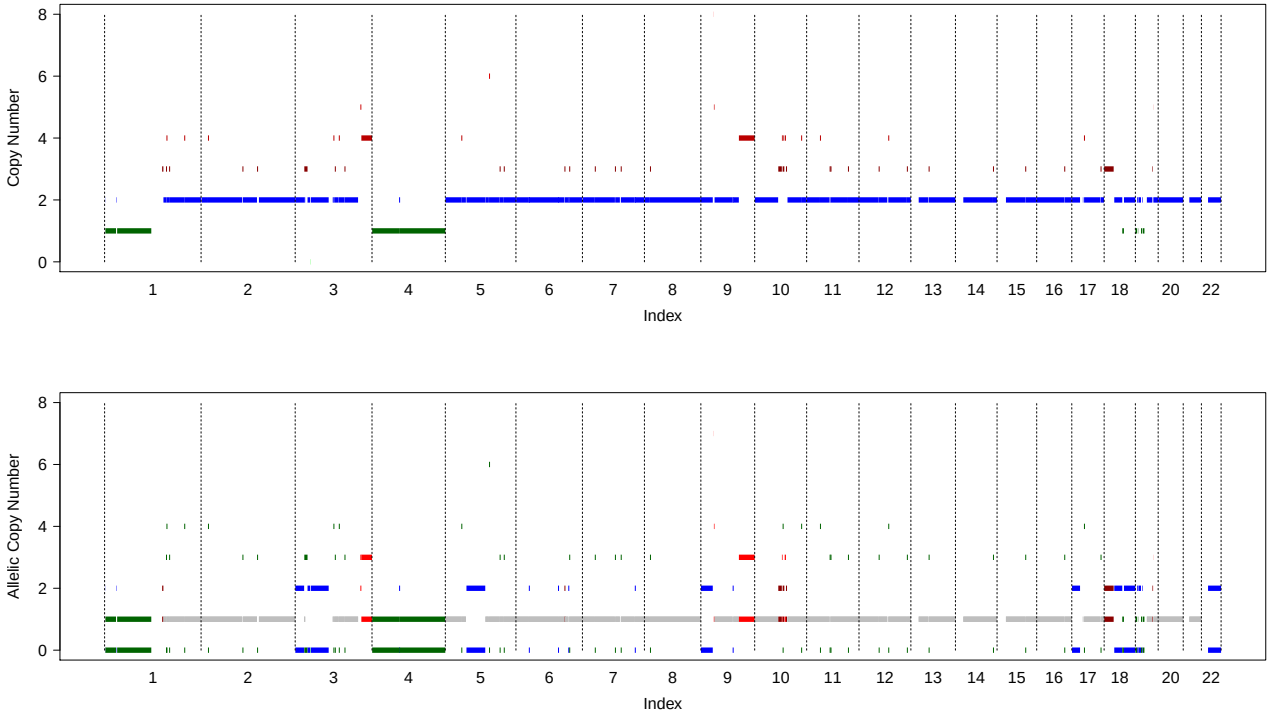

Figure A.6: Plots illustrating the copy numbers (top) and allelic copy numbers (bottom) for the chromosomes of the patient PC1-14, calculated with a ichorCNA-TitanCNA workflow. LOH, copy neutral LOH, and amplifications are colored green, blue, and red, respectively.

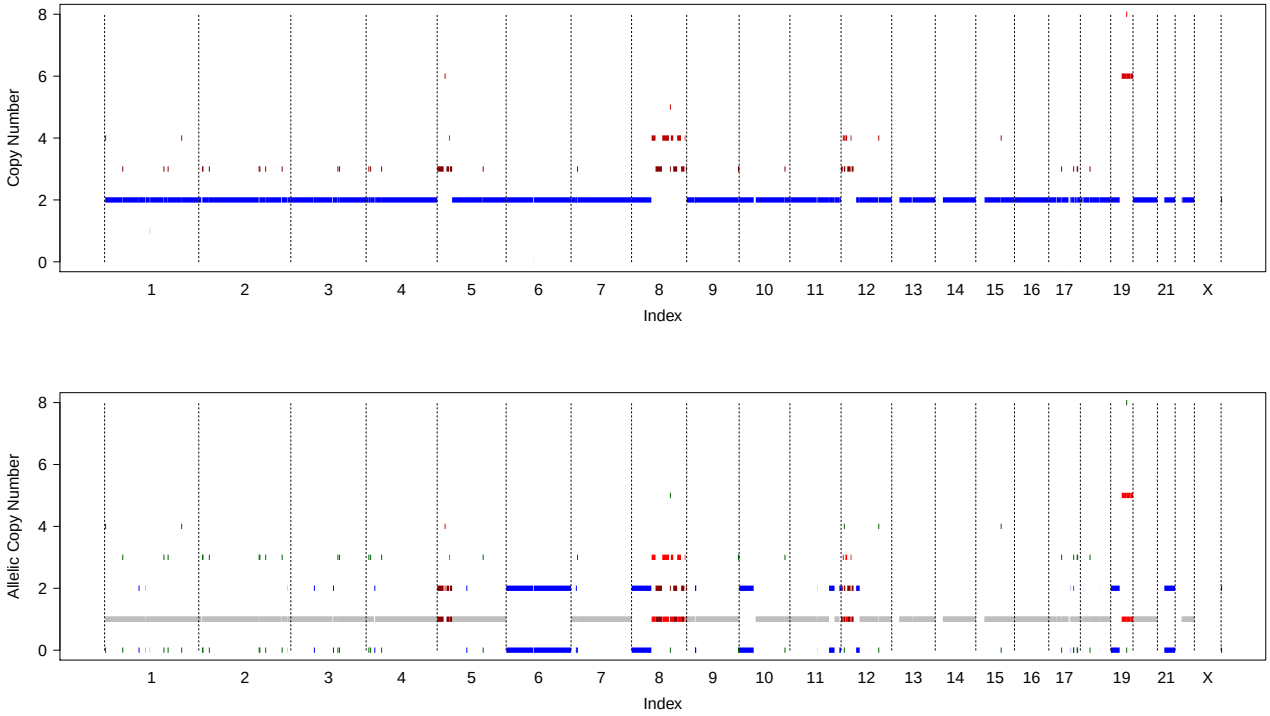

Figure A.7: Plots illustrating the copy numbers (top) and allelic copy numbers (bottom) for the chromosomes of the patient PC1-18, calculated with a ichorCNA-TitanCNA workflow. LOH, copy neutral LOH, and amplifications are colored green, blue, and red, respectively.

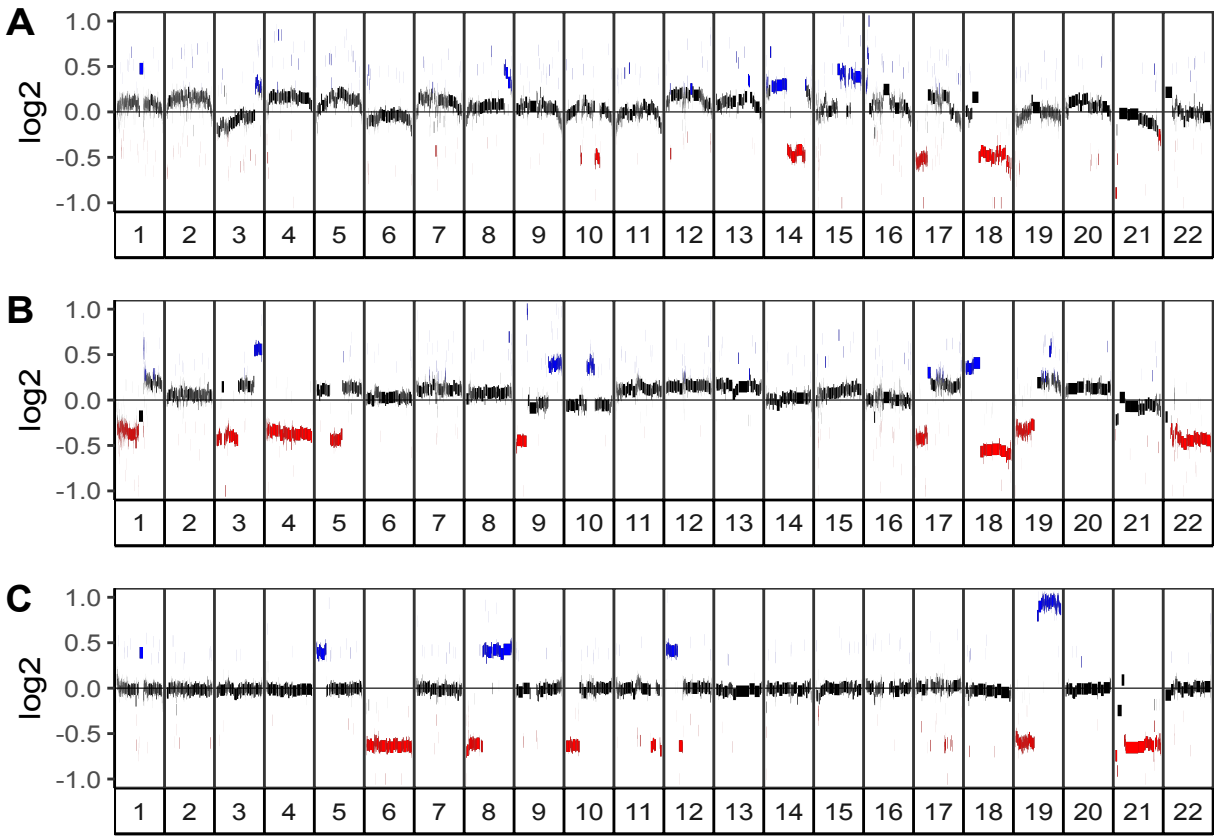

Figure A.8: Plots illustrating the log2 change in copy numbers for each autosome calculated through the VarScan2CNA pipeline, recentered according to the median log2 ratio, for the patients PC1-10 (A), PC1-14 (B) and PC1-18 (C). Neutral segments are colored black, while amplifications (log2 > 0.3) and deletions (log2 < -0.3) are colored blue and red, respectively.