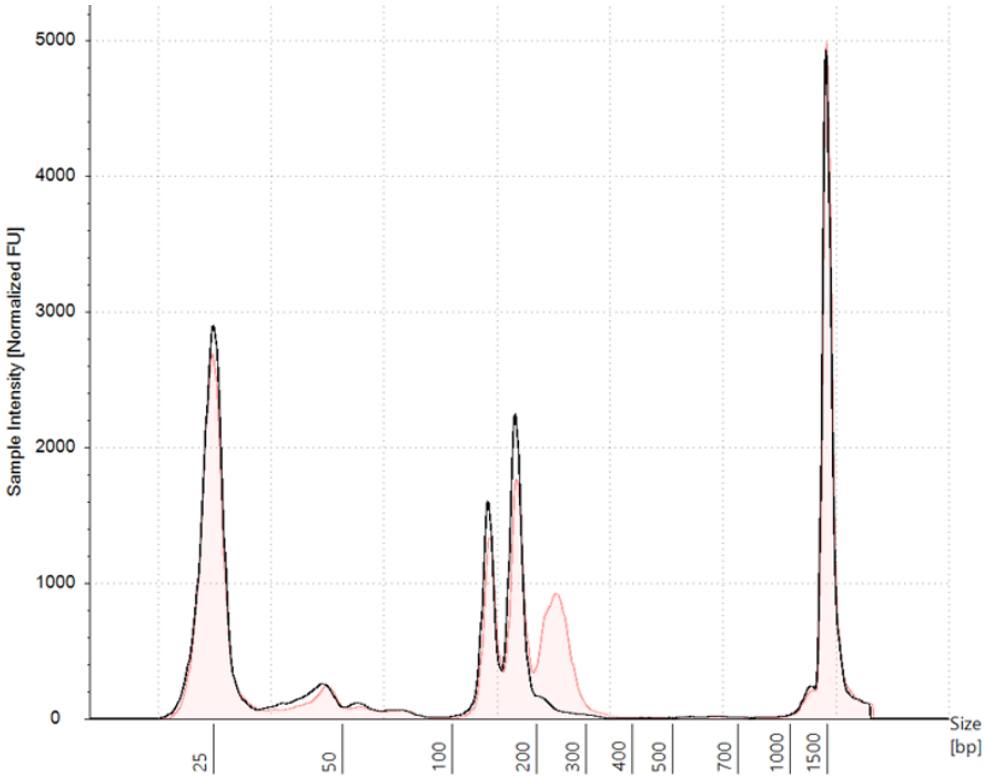


Optimizing a gene panel for enhanced detection of hotspot mutations in endometrial cancer using SiMSen-seq

Master's thesis in Biotechnology

ELMA EKLUND

DEPARTMENT OF LIFE SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Optimizing a gene panel for enhanced detection of hotspot mutations in endometrial cancer using SiMSen-seq

ELMA EKLUND

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life science
Division of Chemical biology
Chalmers University of Technology
Gothenburg, Sweden 2025

Optimizing a gene panel for enhanced detection of hotspot mutations in endometrial cancer using SiMSen-seq
ELMA EKLUND

Supervisors: Anna Linder and Sara Schumacher, Institute of Clinical Sciences, University of Gothenburg

Examiner: Anna Karlsson-Bengtsson, Department of Life Sciences, Chalmers University of Technology

Master's Thesis 2025
Department of Life Sciences
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Telephone +46 31 772 1000

Cover image: Fragment analysis of a test library from the study.

Written in LaTeX
Gothenburg 2025

Optimizing a gene panel for enhanced detection of hotspot mutations in endometrial cancer using SiMSen-seq

ELMA EKLUND
Department of Life Sciences, Chalmers University of Technology

Abstract

Endometrial cancer is the sixth most common cancer among women globally, with over 420,000 cases reported in 2022. Current diagnostic techniques lack sufficient specificity or sensitivity, emphasizing the need for new methods to improve early detection and clinical outcomes. One promising approach is analysis of circulating tumor DNA (ctDNA) in liquid biopsies. This project aimed to optimize a gene panel intended for ctDNA-based mutation detection in endometrial cancer using SiMSen-seq, a next generation sequencing method. The project focused on enhancing the performance and coverage of the panel through complementary assay design and incorporation of additional target sites.

We designed and validated six new assays targeting mutations in five of the most frequently mutated genes in endometrial cancer (*ARID1A*, *CHD4*, *PIK3CA*, *PIK3R1* and *PTEN*) for potential incorporation into the panel. Validation was performed primarily using real-time quantitative PCR and automated electrophoresis, and confirmed compatibility of the new assays with the existing multiplexes, with little observed interference. Based on mutation data from the Catalogue of Somatic Mutations in Cancer, a hypothetical detection coverage was calculated. Estimated, the updated panel would detect an additional 2.6% of mutation-positive endometrial tissue samples, increasing the total coverage to 70.4%. Future efforts should focus on validation using sequencing to confirm individual assay performance and diagnostic utility.

Keywords: gene panel, qPCR, endometrial cancer, multiplexing, liquid biopsy, SiMSen-seq

Acknowledgements

My deepest gratitude to my supervisors Anna Linder and Sara Schumacher who presented me the opportunity to work on this project. I have had immensely fun doing this, and learned more than I could have ever hoped. Thank you for your grand support and inspiring motivation, and for never making me feel like a question is too dumb. And thank you, Isabella, for your moral support and wisdom. I also want to express my gratitude to Karin Sundfeldt for inviting me into the research group, and a big thank you to all the members of the Sundfeldt group, for being so bright, smart and welcoming, and for making meetings interesting and fun. I also want to thank Anna Karlsson-Bengtsson for taking on the role as my examiner at Chalmers. And last, but not least, a one-woman standing ovation to friends and family who made my years at Chalmers to the greatest.

Elma Eklund, Gothenburg, June 2025

List of Acronyms

bp	Base pair
Cq	Quantification cycle
cfDNA	Cell-free DNA
ctDNA	Circulating tumor DNA
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphates
EC	Endometrial cancer
MP	Multiplex
NGS	Next Generation Sequencing
NTC	No template control
PCR	Polymerase Chain Reaction
PTC	Positive template control
qPCR	Quantitative Polymerase Chain Reaction
TVUS	Transvaginal ultrasound
UMI	Unique Molecular Identifier
GENES	
ALK	Anaplastic Lymphoma Kinase
ARID1A	AT-rich interaction domain 1A
CHD4	Chromodomain Helicase DNA-binding Protein 4
PIK3CA	Phosphatidylinositol 3-kinase, catalytic, alpha polypeptide
PIK3R1	Phosphoinositide-3-kinase regulatory subunit 1
PTEN	Phosphatase and tensin homolog

Contents

1
Introduction

Endometrial cancer (EC) is a gynecologic malignancy that arises from the inner lining of the uterus, the endometrium. It is the sixth most common cancer in women, with over 420,000 cases reported globally in 2022 [1], and incidence rising mainly due to an aging population and increasing rates of obesity [2]. Prognosis is highly dependent on the stage at diagnosis. The 5-year survival rate drops drastically from almost 92% at stage I, when the malignancy is confined to the uterine corpus and ovary, to about 15% at stage IV, when the cancer has metastasized [3]. The vast majority of women diagnosed with EC are over the age of 50, with postmenopausal bleeding being the most common symptom, occurring in approximately 90% of cases. As a result, a majority of patients are discovered at an early stage, with about 67% of cases diagnosed when the disease is still localized [4] [2]. In contrast, among women presenting with postmenopausal bleeding the likelihood of an EC diagnosis, regardless of tumor stage, is about 9%, emphasizing the need for diagnostic specificity [4].

Initial diagnosis of suspected EC typically involves transvaginal ultrasound (TVUS), a minimally invasive method that assess the risk of EC by measuring endometrial thickness in women with postmenopausal bleeding [5]. However, TVUS have limited specificity due to endometrial fluctuations in healthy individuals [6]. Because of this, many patients require confirmatory endometrial sampling, which is more invasive and can cause anxiety and physical discomfort [7] [8]. This, in combination with the absence of simple diagnostic methods for targeted screening of EC, calls for new techniques providing sufficient diagnostic accuracy and sensitivity to enable earlier diagnosis and improve treatment outcomes.

One approach that has recently gotten a lot of attention in oncology, is the detection of circulating tumor DNA (ctDNA) as a biomarker in liquid biopsies. This minimally invasive approach holds great potential for early detection [9] and disease monitoring [10] of cancer. Compared to tissue biopsies, liquid biopsies offer a more comprehensive picture of tumor heterogeneity, as tissue samples will only represent a small part of the tumor [11]. Mutations in EC have been successfully detected in ctDNA extracted from fluids from routine Papanicolaou tests [12]. Although it is not yet reliable enough for clinical diagnostics, these finding show the potential of ctDNA-based detection in liquid biopsies for EC detection.

Detecting ctDNA poses a significant challenge due to its low abundance relative to total cell-free DNA (cfDNA). Hence, highly sensitive and specific detection methods are required to distinguish ctDNA from background cfDNA effectively [13]. One method

addressing this issue is simple, multiplexed, PCR-based barcoding of DNA using sequencing (SiMSen-seq). SiMSen-seq is a next generation sequencing (NGS) method capable of detecting genetic variants at frequencies below 0.1% [14], making it highly effective for detection of rare DNA molecules. Thus, this makes SiMSen-seq a suitable method for cancer detection in liquid biopsies, and subsequently an appropriate method for this project.

Using SiMSen-seq, a gene panel targeting hotspot mutations in EC, known as the EC-panel, was previously developed by the Sundfeldt's group [15]. The current panel consists of two multiplexes, or sub-panels, harboring a combined 69 assays targeting sites in 17 commonly mutated genes, such as *TP53*, *ARID1A*, *PTEN*, and *PIK3CA*. Whether the ultimate application of the panel will be as a diagnostic tool for clinical use or primarily for research purposes is yet to be determined. This project focuses on improving the accuracy and performance of the panel through refinement of existing assays and the addition of new target sites to enhance panel coverage, with the goal of refining it for comprehensive EC detection.

1.1 Aim and limitations

This project aimed to optimize a gene panel developed by Sundfeldt's group, with the goal of improving its diagnostic and/or research utility for EC. The panel targets hotspot mutations commonly found in EC and currently comprises 69 single assays with varying levels of performance. The optimization process focused on enhancing panel performance by improving assay sensitivity and specificity through complementary assay design, as well as design and validation of new assays to expand panel coverage by targeting additional mutations of interest.

Due to time constraints, the panel was not tested on clinical samples, only control DNA. This may have led to biased results, as reactions where intact control DNA is used typically performs better. Additionally, sequencing was not included for this same reason, which limited the ability to fully evaluate individual assay and panel performance.

2
Theory

2.1 Circulating tumor DNA in liquid biopsy

A liquid biopsy refers to the analysis of bodily fluids such as urine, blood or saliva, for the detection of biomarkers indicative of disease. A promising biomarker that has recently gotten a lot of attention due to its potential for non-invasive cancer diagnostics, early detection, relapse monitoring, and treatment decisions, is cell-free DNA (cfDNA). cfDNA is fragmented DNA that has been released into the extracellular environment through physiological and pathological mechanisms including apoptosis, necrosis, and active secretion by cells [11]. The majority of cfDNA is the size of about 166 base pairs (bp), which is believed to originate from the mechanism of apoptotic cleavage, where DNA gets degraded by caspases that cleave between nucleosomes [16].

In individuals with cancer, a fraction of the cfDNA originates from tumor cells and is referred to as circulating tumor DNA (ctDNA). Unique cancer-specific mutations can be targeted in the cfDNA and used to track tumor progression or detect remaining cancer after treatment [13]. However, a primary challenge in ctDNA analysis is that ctDNA typically is in much lower abundance relative to cfDNA. The proportion of ctDNA can vary greatly depending on tumor type, disease stage, and also between different bodily fluids, with sightings of less than 0.1% up to almost 90% of total cfDNA in plasma. A person with further progressed cancer often have increased rates compared to healthy individuals, although this is not always the case [17][16]. Therefore, detection of cancer-associated genetic alterations, and the distinction of the ctDNA from the background cfDNA is a challenge which requires highly sensitive and specific detection techniques.

2.2 SiMSen-seq

SiMSen-seq is a next generation sequencing (NGS) method capable of detecting genetic variants at frequencies below 0.1% [14], making it highly effective for detecting ctDNA in liquid biopsies. As opposed to existing barcode methods, SiMSen-seq features a simple and short library construction protocol that allows for low DNA input. The primary source of background in NGS are errors introduced by DNA polymerases during library construction or sequencing. SiMSen-seq addresses this by using Unique Molecular Identifiers (UMIs), barcodes that enables distinction between true genetic variations and polymerase-induced errors (see Figure 2.1). The barcode is protected by a hairpin-loop which remains closed during the initial amplification step, preventing the formation of non-specific products. Subsequently, as the temperature increases, the loop opens, allowing for binding of the

adaptor primers to the barcoded products.

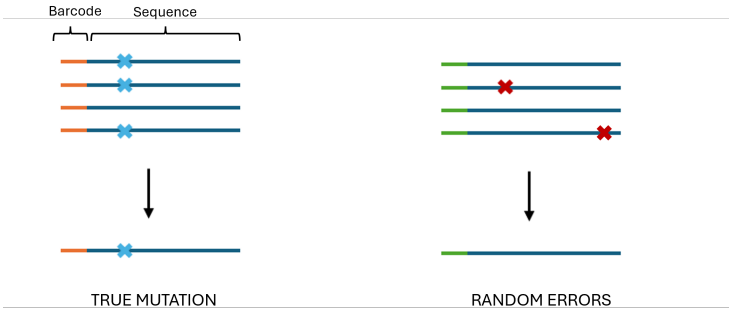


Figure 2.1: A simplification of how the SiMSen-seq method uses barcodes to identify true genetic variations from errors introduced by polymerases during amplification. Genetic alterations are represented by crosses.

The SiMSen-seq workflow includes target assay design and validation followed by addition of barcode sequences, containing the hairpin structure, to the target assay primers. Barcoded assays are then individually validated by making a test library. A first round of barcoding PCR is performed, from which the resulting products are subjected to a second round of adaptor PCR with Illumina sequencing adaptor primers. This is first done with real-time adaptor PCR, and based on results, the last step is performed which is adaptor PCR with high fidelity polymerase. After validation of individual assays, multiplex combinations are tested, followed by library construction and sequencing. The resulting SiMSen-seq data is processed in the UmiErrorCorrect custom software pipeline [18] where the UMI count for each assay can be determined, reflecting their individual performance.

2.3 Quantitative PCR

The PCR technique enables amplification of nucleic acids using a mixture of DNA template, oligonucleotide primers flanking the template, dNTP, and DNA polymerase. The reaction undergoes thermal cycles of denaturation (separation of double-stranded DNA), annealing (binding of primers to single-stranded DNA), and elongation (DNA synthesis by polymerase), which are repeated until sufficient amplification is achieved [19]. At optimal performance, amplification results in a doubling of DNA per cycle.

Real-time quantitative PCR (qPCR) builds on conventional PCR by incorporating fluorescent probes that bind to the DNA product, enabling real-time monitoring of the amplification by detecting fluorescence emitted from the probe after each cycle [20]. This project used SYBR Green I, which binds to double-stranded DNA [19]

From the qPCR, a fluorescence vs. cycle plot is generated, exemplified in Fig. 2.2, where the quantification cycle (Cq) value is determined at the cycle at which fluorescence surpasses a predefined threshold. Thus, a lower Cq, i.e. a curve appearing earlier in the graph, equals more product, and vice versa. A standard curve is created by measuring a dilution series of DNA and plotting the corresponding Cq values against the logarithm of

the DNA concentrations. The slope of the standard curve can then be used to calculate PCR efficiency using Equation 2.1.

$$Efficiency = (10^{\frac{1}{-slope}} - 1) \cdot 100 \tag{2.1}$$

In addition, when doing qPCR, melting curve analysis is performed. The melting curve represents the denaturation of the double stranded DNA and can be assessed to detect undesired primer-dimers, and other secondary structures, that may have formed during the reaction.

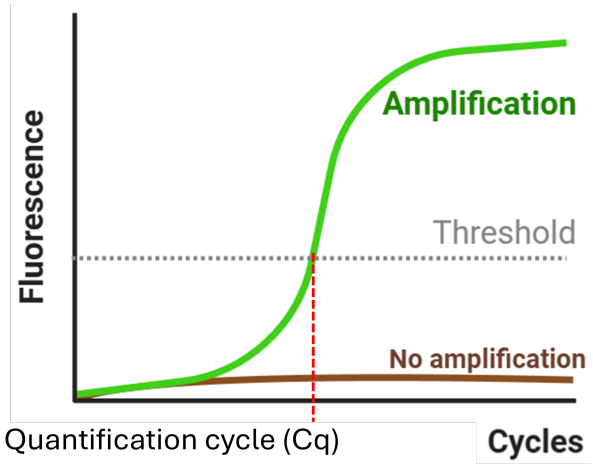
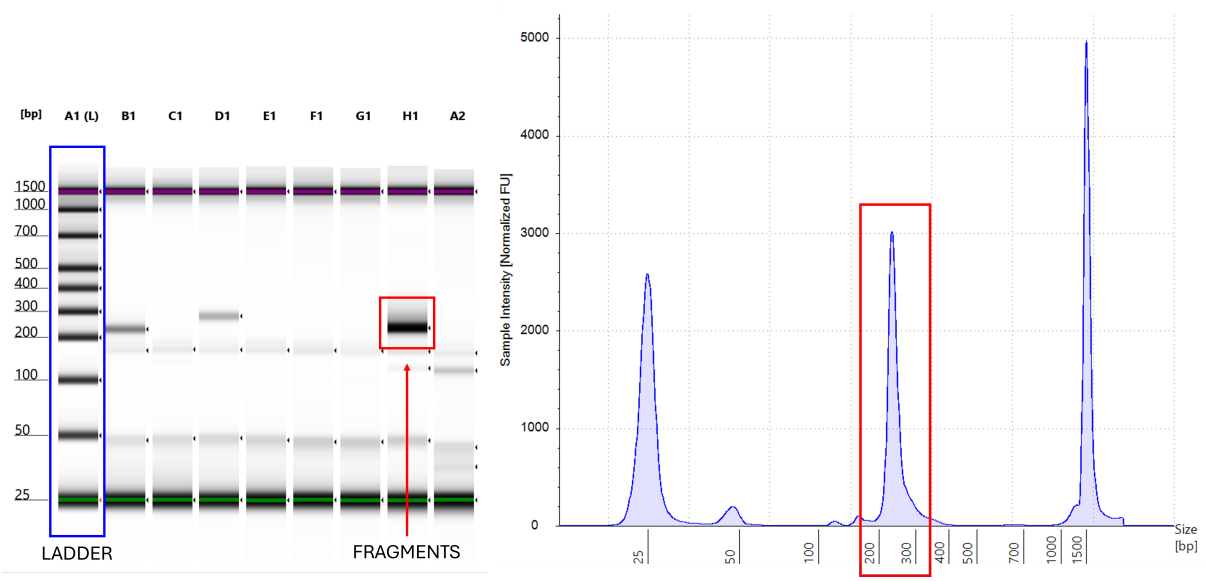


Figure 2.2: Example of an amplification curve from qPCR. The graph represents the florescence units per cycle in a sample with amplification (green curve) contra a sample without amplification (brown curve). The Cq at which the florescence passes the threshold is registered for comparative analysis. Created with BioRender.com

2.4 Fragment analysis

To assess assay performance, evaluate and quantify formation of the correct PCR product and select the appropriate fragments for sequencing, post-PCR samples can be analyzed through fragment analysis. The instrument used for this purpose was the Agilent TapeStation 4200, which performs automated gel electrophoresis [21]. Samples are loaded together with florescent dye into tube strips or a 96-well sample plate, from which the machine transfers them into a disposable "ScreenTape" device containing a gel. Electrodes then apply a current across the gel, generating an electric field in which the negatively charged nucleic acids migrate toward the positive electrode. The migration distance depends on fragment size and shape, with smaller fragments migrating faster, resulting in size-based separation [22]. By then comparing the sample DNA to a reference ladder with known fragment sizes (Figure 2.3a), the size of the DNA fragments including the abundance of each fragment size can be determined. This generates a graph of sample intensity (fluorescent units) to sample size (in bp) of which an example is presented in Figure 2.3b.



(a) (b)

Figure 2.3: (a) Genetic samples migrated through gels depicted by the columns. Horizontal lines of different gray scale are groups of similar sized DNA fragments. Gel A1 contains the reference ladder used to decide the size of the sample fragments. A more pronounced line indicates more fragments. (b) Sample intensity in florescence units, to size in bp, for gel H1. The sample peak marked in red corresponds to the marked fragment-band in a).

The horizontal position of the peak in Figure 2.3b represents the size of the fragments, while the height of the peak corresponds to the abundance of the specific fragment size. The size of the peak, i.e. the integrated peak area, reflects the relative amount of fragment in the sample. For downstream analysis, the desired fragment group can be extracted and purified for later sequencing.

3

Methods

The methodology of this project primarily followed the established SiMSen-seq protocol, integrating the original approach [14] with the updated method [23], including some new custom changes to cycling programs and reagents.

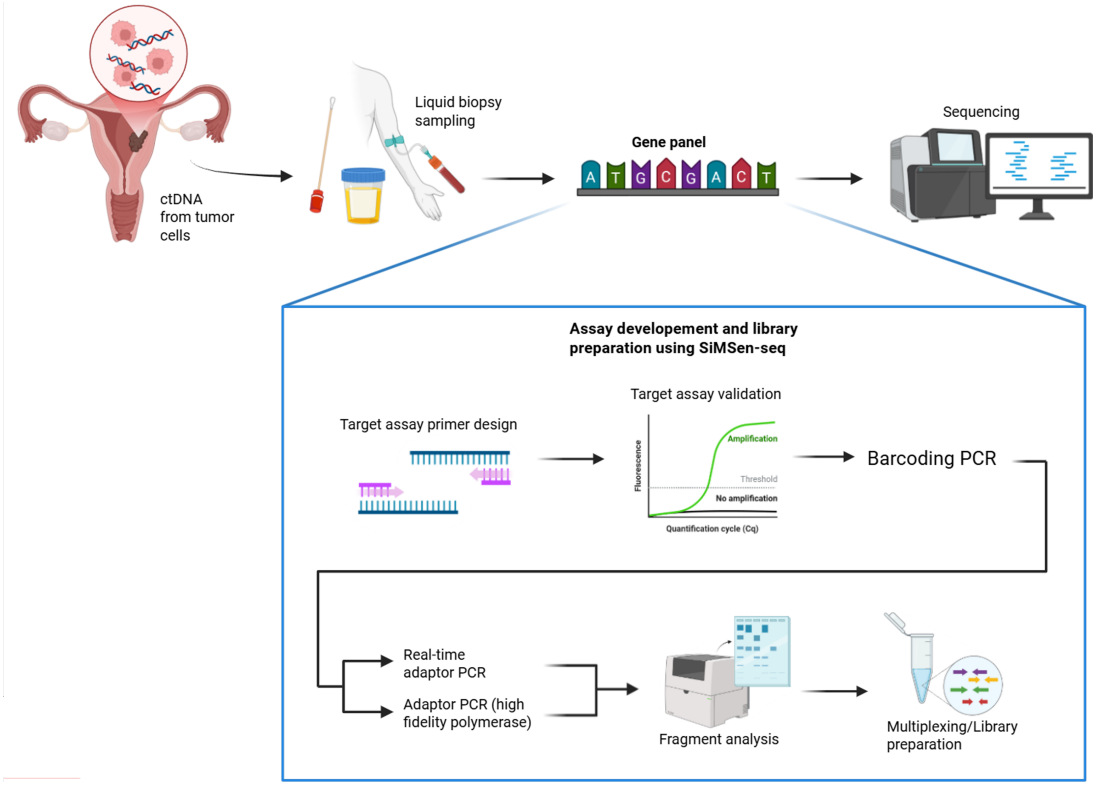


Figure 3.1: Intended workflow for the EC-panel. This project encompass the step of assay development and library preparation using SiMSen-seq. Created with BioRender.com

3.1 Target assay primer design

Frequently mutated sites in EC were previously identified from literature and the Catalogue of Somatic Mutations in Cancer (COSMIC) [24]. Primers flanking these regions of interest was designed using NCBI Primer-BLAST [25], and the initial selection criteria for primers included: amplicon size of 60–80 base pairs, melting temperature (Tm) between 57–63°C, GC content of 50% ± 10% and maximum self and pair complementarity of 3. The designed target assay primers were controlled with the UCSC In-Silico PCR with the

human reference genome hg38 [26]. For all sites of interest, between two and six primer pairs were ordered and tested to increase the chance of finding a suitable assay.

All primers were ordered in tubes of 25 nmole DNA Oligos from (Integrated DNA Technologies, USA). Upon arrival, primers were diluted to 100 μM using TE-buffer (Thermo Fisher Scientific, USA) and stored at +4°C.

3.2 Target assay validation

All reagents used in this method were vortexed and centrifuged prior to use, as well as between dilutions and preparations, unless otherwise specified. The DNA used in this project was obtained from Roche (F. Hoffmann-La Roche AG, Switzerland), supplied pre-solved in aqueous buffer. Upon arrival the concentration of each DNA sample was measured using the Qubit™ 1X dsDNA Broad Range Assay Kit (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. All measurements were performed in triplicate, and the DNA concentration was determined as the average of these values, unless any replicate deviated significantly, in which case it was excluded from the calculation.

Validation of target assays was performed by qPCR, using a five-point DNA dilution series with final concentrations of 0.16, 0.8, 4, 20 and 100 ng/μl. All qPCR reactions were run in triplicate and included no-template controls (NTC) and the reference assay ALK (forward GGGCTTGGGTCGTTGGGCAT and reverse CTCCTTTGCACAGGGGTCTGGG). The qPCR was conducted using the Mic real-time PCR cycler (Bio Molecular Systems, Australia), which can handle up to 48 wells per run. This set-up allowed simultaneous analysis of five concentrations plus NTCs for two target assays, as well as three concentrations (0.8, 4 and 20ng/μl) and NTCs for the reference assay. Each reaction contained the following components (also presented in A.1): 5μl TATAA SYBR GrandMaster Mix (TATAA Biocenter AB, Sweden), 0.4 μl forward/reverse primer mix (10 μM), 2.6 μl nuclease-free water and 2 μl of template DNA at the specific concentration or nuclease-free water for the NTCs, to a total of 10 μl. All reagents were kept on ice during preparation. The qPCR cycling conditions are described in Appendix A.1.2.

A standard curve was created by plotting the Cq values against the logarithm of the DNA series concentrations. The primer performance was then evaluated based on PCR efficiency, which was calculated using the slope of the DNA standard curve according to equation 2.1.

3.3 Barcoded target assay validation

Target assays selected from the initial validation was ordered with the hairpin barcode sequences developed by the Ståhlberg group [23]. The forward barcode primer (GGACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNAAANNANNAAANNNNATGGGAAAGAGTGTCC) containing the barcode of randomized nucleotides, was added to the the 5' end of the forward assay primer, and the reverse

barcode primer (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT) added on the reverse assay primer. All barcoded primers were ordered as 4 nM Ultramer™ DNA Oligos in tube format from (IDT, USA), shipped dry. Upon arrival, primers were stored at +4°C. Prior to use, they were resuspended according to the protocol described in A.2.1, and subsequently stored in -20°C.

3.3.1 Barcoding PCR

Validation of the barcoded primers was performed based on SiMSen-seq library preparation, but with some modifications, such as the Q5 polymerase being exchanged for Platinum SuperFi II Polymerase. Barcoding PCR reactions were assembled in a 96-well microplate. Each reaction consisted of 5X SuperFi II Buffer and 2U/µL Platinum™ SuperFi II DNA Polymerase (Thermo Fisher Scientific, USA), 10mM dNTP mix (Sigma-Aldrich, USA), 5M L-Carnitine (Sigma-Aldrich, USA), primer mix (forward and reverse, 1 µM), nuclease-free water, and genomic DNA. The exact volumes are detailed in Table A.2. All reagents, except the DNA template, were prepared as a mastermix and distributed onto the plate. DNA template or nuclease-free water was then added to each well. All reactions were performed in triplicate, and ALK was included as a reference assay on each plate. Prior to thermal cycling, the micro plate was vortexed and centrifuged. PCR was performed on a T100 Thermal Cycler (Bio-Rad, USA) using the program listed in A.2.2.2. Immediately after completion, the plate was placed on ice and 20 µl of TE buffer was added to each well using a multi-channel pipette to terminate the reaction.

3.3.2 qPCR

For qPCR, all replicates from the barcoding PCR were used, and each reaction were prepared according to Table A.3. Additionally, two reactions with 2 µl of water, instead of barcoding product, was included as control. qPCR was performed on the Mic Real-Time PCR Cycler (Bio Molecular Systems, Australia) using the thermal profile specified in Appendix A.2.3.2. For each assay, the mean Cq values for the positive template control (PTC) and NTC were calculated. The ΔCq was then determined between the PTC and NTC, and between the PTC and reference assay, ALK.

3.3.3 Adaptor PCR

Adaptor PCR was performed using the first replicate of the PTC and NTC of each assay from the barcoding PCR product. To a 96 well plate, each reaction was prepared by adding Platinum™ SuperFi II PCR Master Mix (2x) (Thermo Fisher Scientific, USA), Illumina Index forward and reverse adaptors (10 µM) (Illumina, USA), nuclease-free water and barcoding PCR product, according to Table A.4. After assembly, the plate was vortexed and centrifuged briefly before cycling on the T100 Thermal Cycler (Bio-Rad, USA) according to the program specified in Appendix A.2.4.2. Following amplification, the plate was immediately put on ice.

3.3.3.1 Fragment analysis

Product from the adaptor PCR were analyzed by automated gel electrophoresis, using the 4200 TapeStation system (Agilent Technologies, USA). The Agilent D1000 ScreenTape Assay was used, and samples were prepared according to the manufacturer's instructions [27]. The expected product size was calculated by adding the amplicon length of an assay plus the length of the forward and reverse barcode primers (70 bp and 34 bp) and forward and reverse Illumina primers (25 bp and 30 bp respectively). The expected product size were then compared to the fragment sizes from the TapeStation to identify the correct target product and confirm correct target amplification. Data analysis was conducted on the associated TapeStation Analysis Software 5.1.

3.4 Multiplexing

Following validation of barcoded target assays, those meeting the approval criteria were incorporated into the full panel for multiplex testing. Since the panel comprises two multiplexes, Unipro UGENE v.44.0 [28] was used to visualize potential primer overlap. Based on this analysis, each assay was assigned to be tested in one of the multiplexes in a way that minimized primer competition.

SiMSen-Seq library construction was used for multiplexing. However, modifications were implemented for panel testing. Previously prepared multiplexes (stored at -80°C) were combined with the newly designed assays to a concentration of 1µl prior to preparation of the barcoded PCR master mix, see Table A.5. The thermal profile used had been previously perfected for the EC panel and is detailed in Appendix A.3.1.2. The remaining steps of the protocol, including adaptor PCR and subsequent analysis, were performed as described in Section 3.3.3, with the exception that the number of cycles for the adaptor PCR was reduced from 30 to 22.

For the potential replacement of one of the assays, the previously designed multiplex was prepared excluding the assay targeted for replacement. The two candidates, the newly designed assay and the original assay, were then individually added to the multiplex and evaluated separately.

3.5 Coverage calculations

For coverage calculations data from COSMIC was used [24]. All data used came from targeted screens in samples originating from endometrial tissue, and was curated for duplicate sample IDs. By comparing genomic coordinates of the target assays to those of the sample mutations provided by COSMIC, the hypothetical coverage of both individual target assays and the whole panel was calculated. The code used for this was scripted in R with the help of ChatGPT [29], and is provided in Appendix C.1.

4
Results

4.1 Target assay primer design

Primers targeting sites of interest in the commonly mutated genes *ARID1A*, *CHD4*, *PIK3CA*, *PIK3R1* and *PTEN* were designed based on COSMIC data. For certain mutation sites, assays had previously been designed and tested without satisfactory results. These sites proved to be particularly challenging to design primers for without exceeding the initial design parameters, which led to some primers being designed with up to five complementary bases, and amplicons that were either shorter or longer than the desired length span (60-80 bp), as reflected in Table B.2.

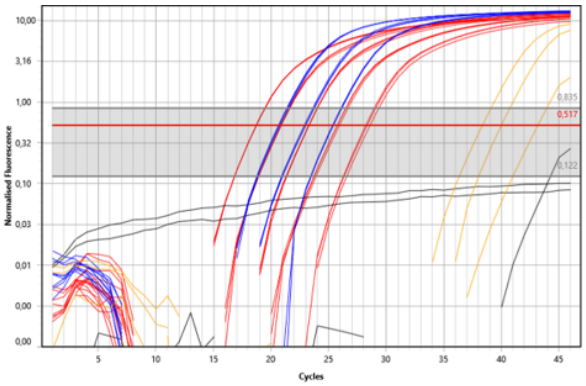
4.2 Target assay validation

Assays were considered acceptable for further validation if their efficiency ranged between 95% and 105%. Efficiencies exceeding this range indicated nonspecific amplification or primer-dimer formation, while values below suggested suboptimal amplification conditions. The efficiencies were compared to the reference assay ALK, with an optimal, previously established, efficiency of 100%. This approach allowed for detection of potential variability between runs. As a result, some primers outside of the preferred range were also selected. In some cases, assays with efficiencies less than 95% were still included due to their relevance in targeting high-priority mutation sites. A summary of all validated target assays, including their efficiency, targeted site of interest and additional characteristics, is provided in Table B.2

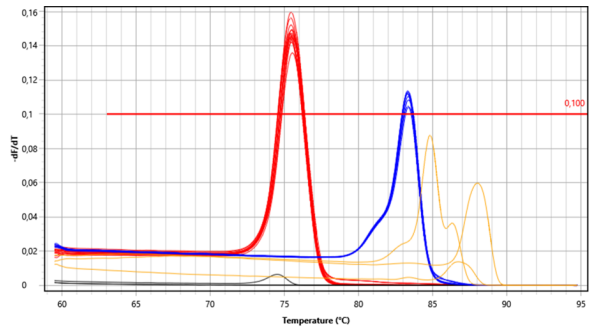
Figure 4.1a presents the cycling graph of an approved assay, PIK3CA_453, which followed the amplification curves of the ALK reference and demonstrated an efficiency of 94.9%. In contrast, the assay PTEN_341 was not approved due to displaying a low efficiency of 87% (B.2) and inadequate performance, as evident by the early deviation from the reference curve indicated in Figure 4.1c.

Additional quality control included assessment of the amplification plots to confirm a state of no or very late non-specific amplification in the NTCs, and melting curve analysis to detect any undesired primer-dimers or off-target products. Any assays exhibiting such signs were excluded, regardless of efficiency values. The melting graph of PIK3CA_453 (Figure 4.1b) showed no evidence of non-specific product amplification in neither assay nor the NTC, further supporting its approval. None of the validated

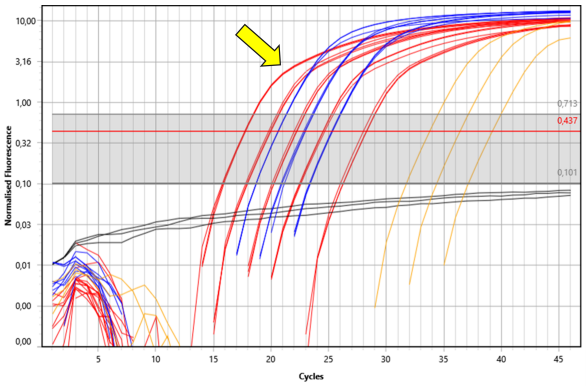
(a) PIK3CA_453



(b) PIK3CA_453



(c) PTEN_341



(d) PTEN_329

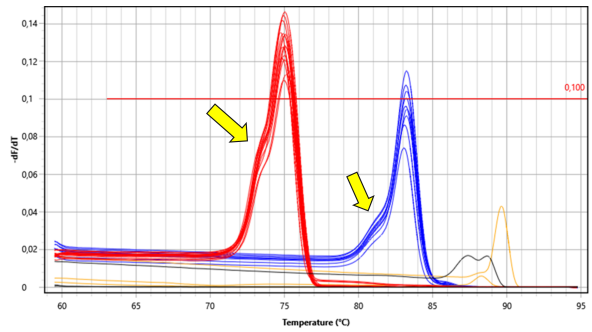


Figure 4.1: qPCR validation of target assays. Graphs depicting target assays (red) with 0.16-100ng/µl DNA, assay NTC (black), reference assay (blue) with 0.8-20ng/µl DNA and reference assay NTC (yellow). (a) Amplification curve of PIK3CA_453. (b) Melting curve of PIK3CA_453. (c) Amplification curve of PTEN_341. Early deviation from control curve indicated by arrow. (d) Melting curve of PTEN_329. Bumps in graph indicated by yellow arrows.

assays showed any additional peaks, though a subtle indication was observed in the melting diagram of PTEN_329, depicted in Figure 4.1d, visible as the slight bump in the assay curve. However, this was too small to be indicative of unwanted product, and something similar can be spotted in the reference assay curve (Figure 4.1d). Instead, the shift could be due to an imbalance in GC-TA content in the primer, causing "chunks" of a TA-richer part to dissociate from the DNA earlier than the rest during the denaturation.

When multiple assays targeting the same site showed acceptable performance, the most suited assay was selected based on efficiency and cycling graph in reference to the ALK. Once all TAs were validated, the selected assays (ARID1A_1989, CHD4_1106, PIK3CA_104_106, PIK3CA_453, PIK3CA_93, PIK3R1_567_572, PTEN_329) were ordered with barcodes for further validation. In addition to these, two additional assays, ARID1A_1989_PP and PIK3CA_345_CancerSEEK [30], which had been previously tested, were included in the final selection.

4.3 Barcoded target assay validation

Barcoded assays were considered qualified if they had a ΔCq where the PTC differed by no more than 1 cycle from the reference assay and if the ΔCq between PTC and NTC was greater than 1 cycle (with NTC coming later than PTC). However, this general guideline was not the only data considered for the inclusion of an assay, and was therefore occasionally overlooked. The ΔCq values for all barcoded target assays are presented in Table B.1.

Further validation was performed using fragment analysis. Based on expected amplicon size, the correct product of the target assays were expected within the 200-300 bp range. Peaks outside this range were considered off-target products. Clear separation between the assay and NTC peak is critical, as this ensures the desired product can be purified prior to sequencing. Integrated peak area, reflecting the amount of product amplified, was also considered carefully as it indicates assay performance. Note that in this step, in contrast to earlier validation, product is observed in all of the NTCs (see Figure 4.2b, 4.2e, 4.2h). This is an expected off-target product resulting from the Illumina primers. It is, however, still important that the NTC appears later in the cycling graph and remains separated in the fragment analysis.

Figure 4.2c shows the fragment analysis diagram for PIK3CA_453, where the product peak at 232 bp exhibited low normalized florescence units, less than a fifth of that observed for the ALK control, indicating weak amplification. The corresponding qPCR cycling plot (Figure 4.2a) showed a ΔCq between the assay PTC and control assay PTC exceeding 1. Based on these results, this assay was not selected for further testing, as was the case for PIK3CA_93. In contrast, PIK3R1_567_572 (Figure 4.2f) showed a product peak of comparable intensity to the control assay, indicating robust amplification. The approval of this assay was further supported by the cycling plot, depicted in Figure 4.2d, which demonstrated a mean Cq less than 1 relative to the control assay.

Some assays, such as PIK3CA_104_106 presented in Figure 4.2i, were accepted for further testing despite a $\Delta Cq > 1$ between the assay PTC and control PTC (Figure 4.2g), granted that their product amplification was sufficient. A strong product amplification does not guarantee optimal performance in the panel, and certain assays targeting key sites were thus prioritized despite unideal results. The fragment analyses for all validated assays are presented in Appendix B.1.

The assays ARIDA_1989 and ARIDA_1989_PP both target the same hotspot site. They performed similarly (see Figure 4.3), and given the importance of the site, both assays were subjected for further analysis in multiplexing.

The final set of target assays qualified for multiplexing included PTEN_329, PIK3CA_104_106, PIK3CA_345_CancerSEEK, PIK3R1_567_572, CHD4_1106 and ARID1A_1989.

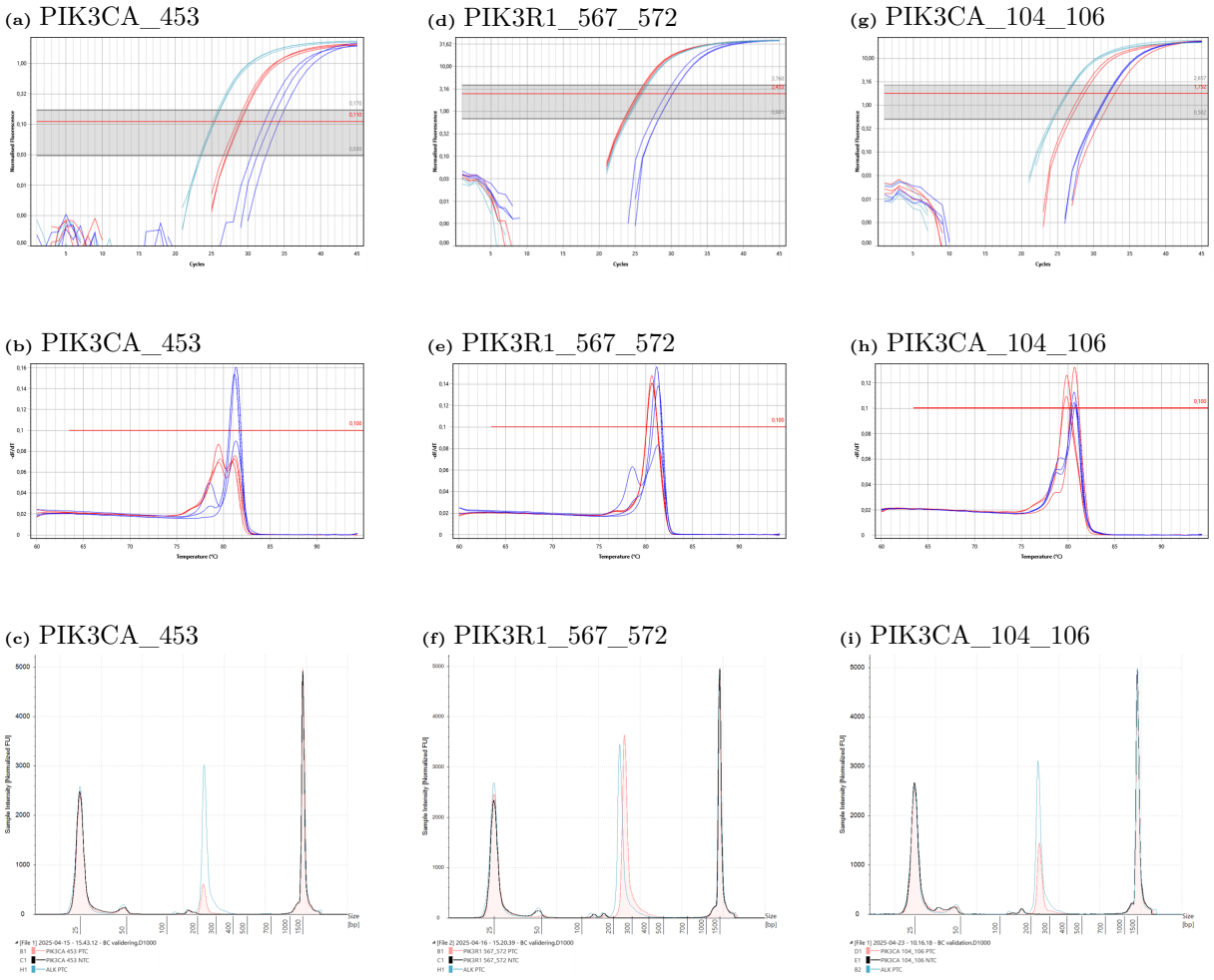


Figure 4.2: Barcoded target assay validation. Cycle and melting curves representing assay PTC (red), NTC (blue) and ALK reference (turquoise). (a) Cycling plot of PIK3CA_453. (b) Melting graph of PIK3CA_453. (c) Fragment analysis of PIK3CA_453. (d) Cycling plot of PIK3R1_567_572. (e) Melting graph of PIK3R1_567_572. (f) Fragment analysis of PIK3R1_567_572. (g) Cycling plot of PIK3CA_104_106. (h) Melting graph of PIK3CA_104_106. (i) Fragment analysis of PIK3CA_104_106.

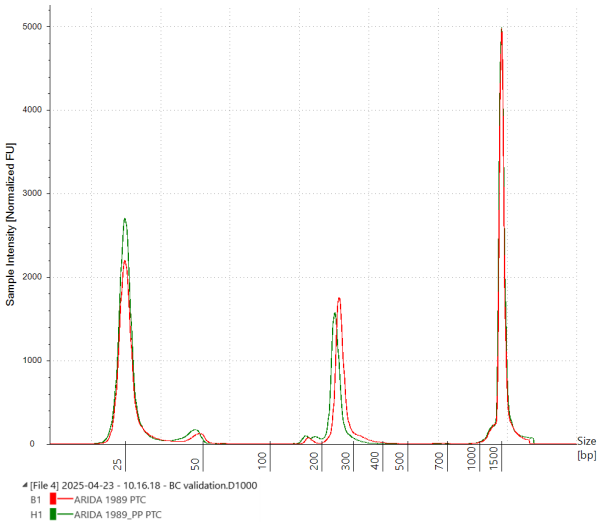


Figure 4.3: Fragment analysis of ARIDA_1989 (red) and ARIDA_1989_PP (green).

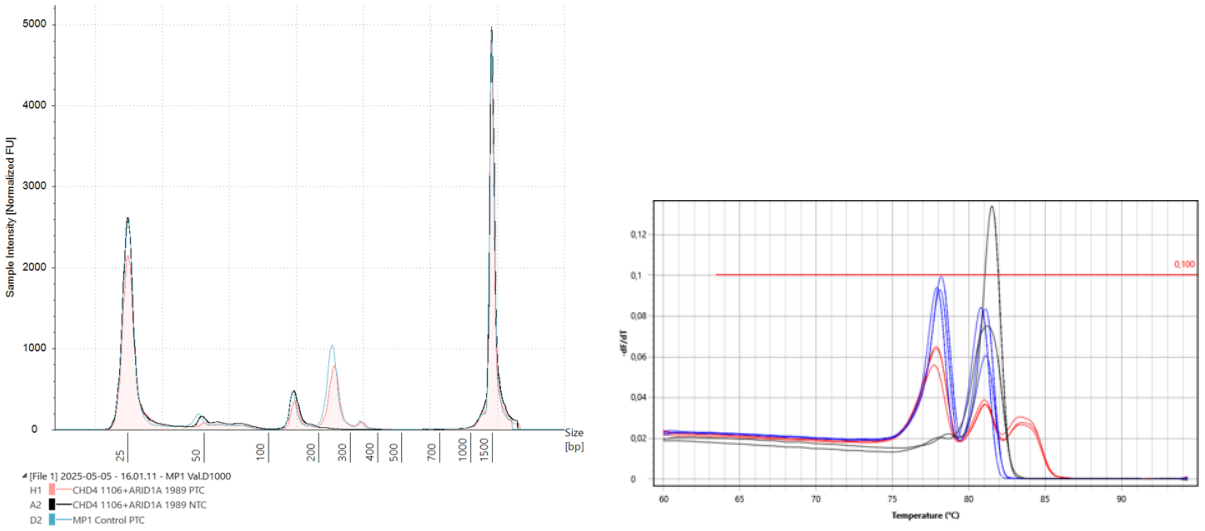
4.4 Multiplexing

Following the individual validation of the barcoded target assays, the assays were validated as a part of the EC-panel by being incorporated into one of its two multiplexes. Specifically, CHD4_1106 and both versions of ARID1A_1989 were tested in Multiplex 1, while PTEN_329, PIK3CA_104_106, PIK3CA_345_CancerSEEK and PIK3R1_567_572 were introduced into Multiplex 2. PIK3R1_567_572 was intended to replace a previous assay, PIK3R1_564, which targeted the same mutation hotspot region but covered fewer mutations overall. For the complete panel, the amplicon sizes ranged from approximately 200 bp to 300 bp, depending on the specific target assay. Which is the span within which product peaks containing the final libraries were expected.

4.4.1 Multiplex 1

The addition of CHD4_1106 and ARID1A_1989 produced a well defined product as illustrated in fig 4.4a. No significant off target-products or interference from the NTC was observed. Although the peak size of the new multiplex peak was slightly lower than that of the control, it indicated a successful library construction and amplification. This was further supported by the corresponding melting curve (Figure 4.4b) which displayed a distinct third peak separate from the NTC.

An overlay of the two ARID1A_1989 variants (Figure 4.5) revealed a slight difference in product peaks, suggesting some variability in performance. However, the ratio between the sub-product peaks and the product peaks appeared similar, possibly reflecting differences in experimental execution rather than inherent assays performance. In conclusion, these results were not definitive.



(a) (b)

Figure 4.4: Validation of CHD4_1106 and ARID1A_1989 in Multiplex 1. (a) Fragment analysis. PTC (pink), NTC (black) and control (blue) which is the Multiplex 1 without the added assays. (b) Melting curve from qPCR. PTC (red), NTC (blue) and dH2O (black).

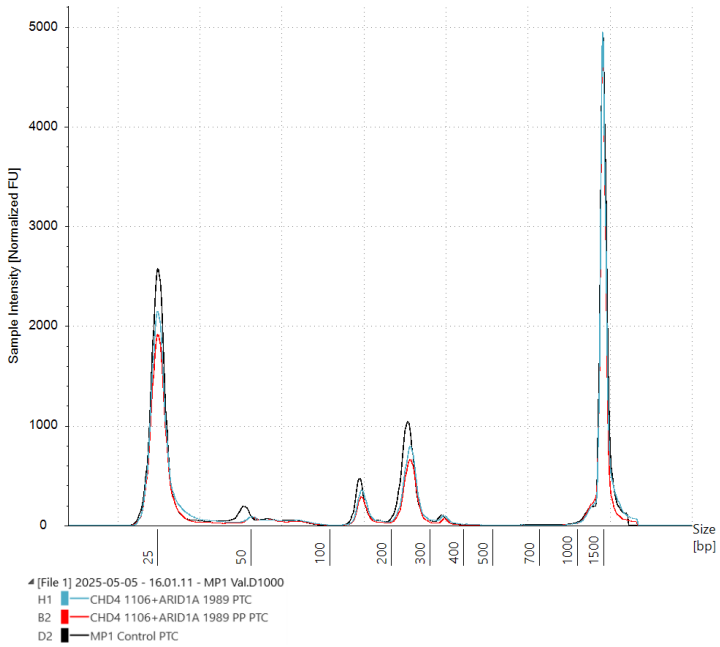


Figure 4.5: Fragment analysis of ARID1A_1989 and ARIDA_1989_PP respectively in Multiplex 1 together with CHD4_1106. Representing ARID1A_1989 (blue) and ARIDA_1989_PP (red). The control (black) represents Multiplex 1 without the additional assays.

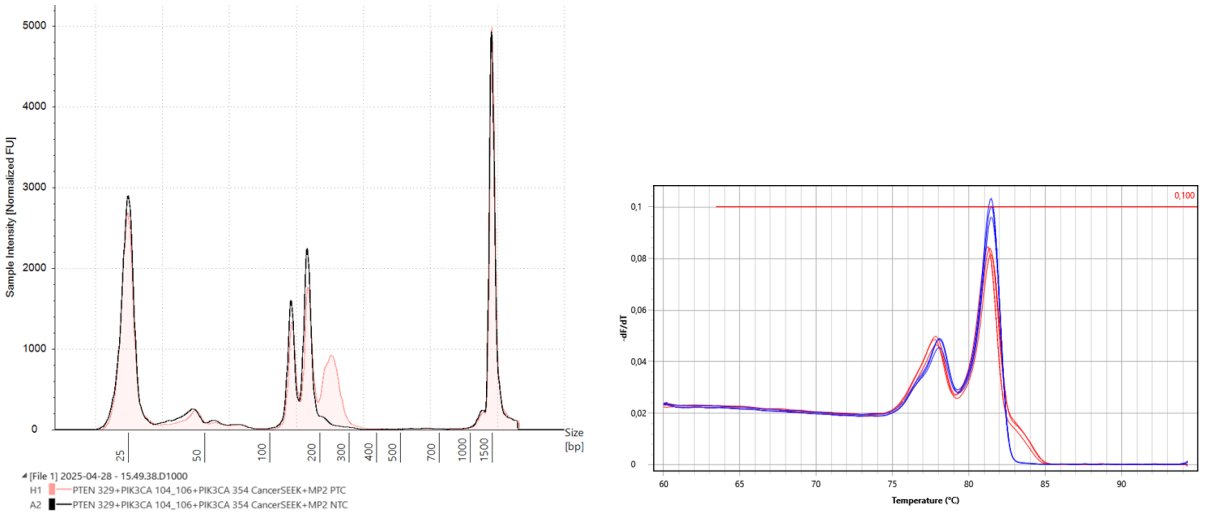


Figure 4.6: Validation of PTEN_329, PIK3CA_104_106 and PIK3CA_345_CancerSeek in Multiplex 2. (a) Fragment analysis. PTC (pink), NTC (black). (b) Melting curve from qPCR illustrating the PTC (red) and NTC (blue).

4.4.2 Multiplex 2

The addition of PTEN_329, PIK3CA_104_106 and PIK3CA_345_CancerSEEK to Multiplex 2 resulted in a clear product peek as illustrated in Figure 4.6a. Minor interference from the NTC was observed, but remained sufficiently distinguished from the target peak, a prerequisite for the subsequent purification of the library. The corresponding melting curve (Fig 4.6b) showed a distinct separation between the PTC and the NTC around 83 Cq, though the peak was relatively small.

For the replacement of PIK3R1_564 with PIK3R1_567_572, both assays were tested in Multiplex 2 aliquots, excluding PIK3R1_564 and including the three newly added assays. Fragment analysis (Figure 4.7), suggested that PIK3R1_567_572 demonstrated a slightly better amplification, or more precisely, interfered less with the multiplex. However, the difference was minor, and sequencing is necessary to validate these results.

4.5 Coverage

Using data from COSMIC, the hypothetical coverage of the EC-panel and individual target assays was calculated. All data presented are based on confirmed mutated samples and were calculated with the ARID1A_1989 assay, not the ARID1A_1989_PP variant. The gene-specific coverages, before and after incorporation of the new assays to the panel, are presented in Figure 4.8. Only coverage for genes targeted by the new assays are shown. The additional coverage gained by the new assays for each gene were as follows: ARID1A 6.3% (to a total of 28.3%), CHD4 11.8% (total 26.2%), PIK3CA 5% (total 72.1%), PIK3R1 13.6% (total 27.6%) and PTEN on 9.6% (total 64.8%). A comprehensive overview of the gene-specific coverage for all final new assays, including ARID1A_1989_PP and PIK3R1_564 is provided in Table B.3.

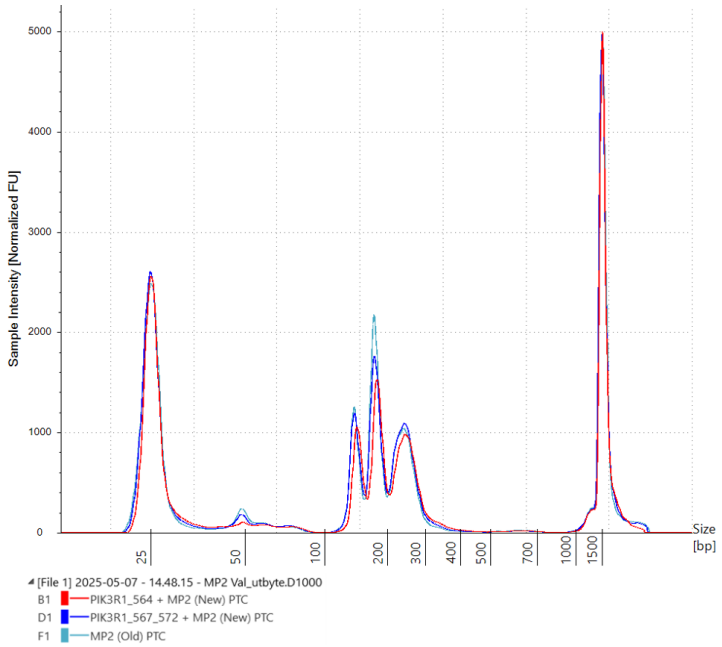


Figure 4.7: Fragment analysis for PIK3R1_564 (red) vs PIK3R1_567_572 (blue), Multiplex 2 control (turquoise).

The total coverage of the EC-panel, across all samples with a detected mutation in any gene screened for (Figure 4.9), would increase by 2.6% with the addition of the new assays, covering a total of 3922 samples. Among these, 94.4% were carcinoma, 2.5% hyperplasia, 1.2% endometriosis and 1.9% were classified as "other" histology types. Within this "other" category, the largest group was adenosarcoma (representing approximately 0.7% of all covered samples), followed by adenomyoma, endometrial polyp, metaplasia, solitary cyst, normal atrophy and benign stromal nodule.

Focusing specifically at samples with confirmed carcinoma (Figure 4.9), which are the primary target of the panel, the additional assays increased coverage by 2.7%, bringing the total panel coverage to 70.8%. In all carcinoma-confirmed samples where a mutation was found in any of the genes targeted by the panel (i.e., the samples potentially detectable by the panel), illustrated by the rightmost pie chart in Figure 4.9, the updated panel hypothetically detected 3.2% more of these samples, rendering a total coverage of 82.9%.

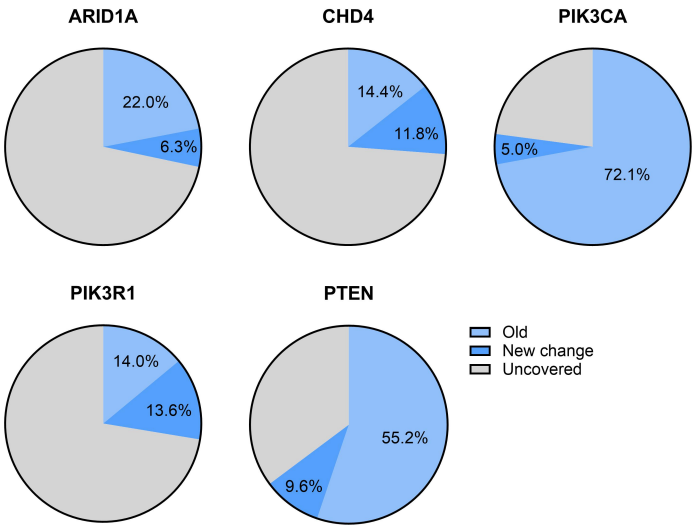


Figure 4.8: Coverage per gene of mutated samples before and after the addition of new target assays to the EC-panel. Coverage presented for the genes *ARID1A*, *CHD4*, *PIK3CA*, *PIK3R1* and *PTEN*.

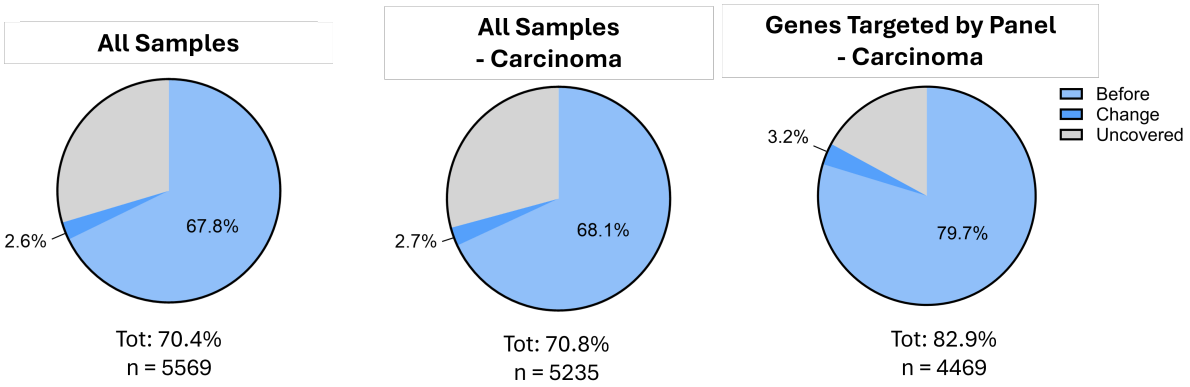


Figure 4.9: From left: coverage over all tested samples of any histology, where a mutation was found in any gene screened for; coverage over all carcinoma-confirmed tested samples where a mutation was found in any gene screened for; coverage over all carcinoma-confirmed samples where a mutation was found in any of the genes targeted by the panel.

5
Discussion

In this project we optimized a gene panel for hotspot mutations in EC by enhancing panel performance and coverage through complementary assay design and incorporation of additional target sites. Six new assays, targeting mutations in five of the most frequently mutated genes in EC (*ARID1A*, *CHD4*, *PIK3CA*, *PIK3R1* and *PTEN*) have been identified for possible inclusion. Five of these, PTEN_329, PIK3CA_104_106, CHD4_1106, ARID1A_1989 and PIK3R1_567_572, are novel designs, while PIK3CA_345_CancerSEEK was adapted from a previous publication [30].

Multiplexing libraries containing new assays exhibited well defined fragment peaks, with acceptable separation from the NTC peaks. This separation was more pronounced in Multiplex 1 compared to Multiplex 2, where a slight overlap was observed. Overall, the new assays demonstrated good compatibility within the multiplexes. However, these results alone are not sufficient to confirm individual assay performance. Sequencing is crucial to evaluate true performance, seeing as the UMI count gives a comprehensive reflection of the assays individual ability to accurately target mutations. Due to time constraints, sequencing was not conducted in this study, and should therefore be a priority for future validation.

The decision regarding which ARID1A_1989 variant to include in the panel remains. The ARID1A 1989 mutation is a frequently observed mutation in EC, and thus an important site to target. As both show similar performance in terms of amplification and coverage (Table B.3), sequencing is necessary to reach a conclusion. For the PIK3R1_567_572 and PIK3R1_564 assays, the choice may be a question of sensitivity versus selectivity. PIK3R1_567_572 offers a markedly higher coverage (20.3%) compared to 5.8% for PIK3R1_564 (Tab B.3). As such, if sequencing were to reveal that PIK3R1_564 performs substantially better, the decision would be whether to prioritize performance or broader mutation detection.

An important limitation of this study was that validation was performed exclusively using test DNA, not patient-derived samples. Synthetic or test DNA typically has lower background, are less fragmented and lacks interfering factors compared to patient DNA. This likely results in more optimistic results of the assay performance. Therefore, validation with patient derived DNA is essential to accurately assess true assay performance before any diagnostic implementation.

The coverage analysis aimed to provide a hypothetical estimate of the EC-panels

diagnostic potential. While it is important to clarify that the presented coverage can not be directly interpreted as detected patients, it gives an adequate overview of the current detection-ability of the panel. The gene-specific coverage analysis demonstrated that the inclusion of the new assays would substantially improve the detection potential of the EC-panel, with an increase of several percentage units overall, and almost a two-fold increase observed for *CHD4* and *PIK3R1*. This highlights the impact and relevance of these targeted regions as important hotspots, and their incorporation into the panel. With the complete panel, including the new assays, the panel would hypothetically detect approximately 70% of mutation-positive samples. Of these, 4.9% corresponds to mutations of benign origin. In a diagnostic setting, this would roughly translate into 192 individuals being incorrectly diagnosed with EC, a false positives rate that must be carefully considered when evaluating the panel's clinical application, as it could cause unnecessary anxiety, medical risks, and costs.

The COSMIC data used for this purpose was corrected for duplicate sample IDs to minimize bias. However, it is important to note that sample IDs correlate to a sample, not to a unique patient. Consequently, multiple sample IDs could originate from the same patient, or possibly the same tumor if multiple samples were collected [24]. Nonetheless, considering the cost of sequencing and the prevalence of research questions focused on inter-patient rather than intra-patient or intra-tumoral variations, the risk of severe overestimation appears low.

Another important consideration is that the COSMIC data set consists of samples from studies performing targeted screenings. Because of this, many mutations may have gone undetected as a result of laying outside the targeted regions of the panels used. In addition, since the EC-panel is largely based on this data, it might be inherently biased to certain groups of mutations. These limitations have a direct implication for clinical use. Because the EC-panel targets a defined set of mutations, it would be inappropriate to solely rely on it to rule out EC in a diagnostic setting. Given the mutational heterogeneity of cancer, alterations could exist in genes or locations not covered by the panel. Therefore, the panel is likely better suited as a confirmatory tool, ideally used in combination with other diagnostic methods.

Moreover, COSMIC data derives from tissue samples, whereas the EC-panel is intended for application in liquid biopsies, which offers both opportunities and challenges. Liquid biopsies can provide a more comprehensive representation of tumor heterogeneity compared to tissue biopsies, which increases the chance of detecting cancer-specific mutations. However, it presents more challenges in terms of fragmented DNA and rare DNA molecules, the latter specifically in early-stage disease. The EC-panel addresses these challenges through a short amplicon design optimized for highly fragmented DNA. In contrast, tissue biopsy panels often rely on longer amplicons, which provide higher specificity, but do not work on shorter DNA fragments such as ctDNA. Short amplicon assays, on the other hand, can still effectivley target longer DNA fragments, making the EC-panel suitable not only for liquid biopsy, but also for tissue samples, especially those with fragmented DNA, such as formalin-fixed paraffin-embedded specimens.

Nevertheless, the small fraction of ctDNA relative to cfDNA remains a major obstacle. Although high ctDNA-levels have been observed in some cases, levels are generally very low, especially in early-stage disease, necessitating highly sensitive methods and robust performing assays. Furthermore, ctDNA abundance can vary between bodily fluids, and thus the type of liquid biopsy source may affect detection rates. For instance, vaginal swabs have shown promise for EC detection, but other sources, such as urine or blood, might provide more comprehensive data. Notably, analysis of plasma-derived cfDNA has shown to be affected by age-related clonal hematopoiesis [31], mutations which could confound the results. This would be of extra relevance for the EC-panel, considering the main demographic for EC is older ages. Thus, correcting against the buffy coat would be essential to minimize false positives.

The clinical utility of the EC-panel is yet to be established. The current false-positive rate, even if reduced, would likely translate into a large quantity of patients if applied at the population level, limiting its utility as a general screening tool. Moreover, since the majority of EC cases are already found at an early stage, the health benefits of population-wide screening would likely not justify the costs. At this stage, the panel shows potential as a diagnostic tool for targeted screening of high-risk patients, monitoring and treatment decisions using liquid biopsies, or as an additional confirmatory tool in tissue biopsy diagnostics, ultimately enabling earlier diagnosis and improved treatment outcomes.

6
Conclusion

An EC gene panel was optimized by incorporating five additional target assays, as well as replacing one existing assay, in five of the most mutated genes in EC: *ARID1A*, *CHD4*, *PIK3CA*, *PIK3R1* and *PTEN*. The addition of the new assays is intended to expand the coverage of the panel, thereby increasing its sensitivity and enabling the detection of more patients. Inital validation of the assays produced promising results, with no significant interference observed when introduced into the existing multiplexes. However, compatibility alone does not equal a strong individual assay performance. As such, further validation through sequencing is needed to establish whether the new target assays sufficiently target mutations and enhance panel coverage, and are to be included in the future panel.

In conclusion, this study provides optimistic evidence that the updated EC panel offers substantial potential for research and clinical applications in EC diagnostics. These findings support its continued development and refinement to contribute to earlier detection and improved treatment outcomes for patients.

References

[1] F. Bray, M. Laversanne, H. Sung, *et al.*, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[2] B. Gu, X. Shang, M. Yan, *et al.*, "Variations in incidence and mortality rates of endometrial cancer at the global, regional, and national levels, 1990–2019," *Gynecologic oncology*, vol. 161, no. 2, pp. 573–580, 2021.

[3] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023, ISSN: 0007-9235. DOI: 10.3322/caac.21763.

[4] M. A. Clarke, B. J. Long, A. D. M. Morillo, M. Arbyn, J. N. Bakkum-Gamez, and N. Wentzensen, "Association of endometrial cancer risk with postmenopausal bleeding in women: A systematic review and meta-analysis," *JAMA internal medicine*, vol. 178, no. 9, pp. 1210–1222, 2018.

[5] B. Long, M. A. Clarke, A. D. M. Morillo, N. Wentzensen, and J. N. Bakkum-Gamez, "Ultrasound detection of endometrial cancer in women with postmenopausal bleeding: Systematic review and meta-analysis," *Gynecologic oncology*, vol. 157, no. 3, pp. 624–633, 2020.

[6] I. Jacobs, A. Gentry-Maharaj, M. Burnell, *et al.*, "Sensitivity of transvaginal ultrasound screening for endometrial cancer in postmenopausal women: A case-control study within the ukctocs cohort," *The lancet oncology*, vol. 12, no. 1, pp. 38–48, 2011.

[7] N. Colombo, C. Creutzberg, F. Amant, *et al.*, "Esmo-esgo-estro consensus conference on endometrial cancer: Diagnosis, treatment and follow-up," *International Journal of Gynecological Cancer*, vol. 26, no. 1, pp. 2–30, 2016.

[8] K. Charoenkwan and C. Nantasupha, "Methods of pain control during endometrial biopsy: A systematic review and meta-analysis of randomized controlled trials," *Journal of Obstetrics and Gynaecology Research*, vol. 46, no. 1, pp. 9–30, 2020.

[9] W. Feng, N. Jia, H. Jiao, *et al.*, "Circulating tumor dna as a prognostic marker in high-risk endometrial cancer," *Journal of translational medicine*, vol. 19, pp. 1–11, 2021.

[10] C. Casas-Arozamena, A. Vilar, J. Cueva, *et al.*, "Role of cfdna and ctdna to improve the risk stratification and the disease follow-up in patients with endometrial cancer: Towards the clinical application," *Journal of Experimental & Clinical Cancer Research*, vol. 43, no. 1, p. 264, 2024.

[11] Z. Qin, V. A. Ljubimov, C. Zhou, Y. Tong, and J. Liang, "Cell-free circulating tumor dna in cancer," *Chinese journal of cancer*, vol. 35, pp. 1–9, 2016.

[12] Y. Wang, L. Li, C. Douville, *et al.*, "Evaluation of liquid from the papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers," *Science translational medicine*, vol. 10, no. 433, eaap8793, 2018.

[13] L. A. Diaz Jr and A. Bardelli, "Liquid biopsies: Genotyping circulating tumor dna," *Journal of clinical oncology*, vol. 32, no. 6, pp. 579–586, 2014.

[14] A. Ståhlberg, P. M. Krzyzanowski, M. Egyud, S. Filges, L. Stein, and T. E. Godfrey, "Simple multiplexed pcr-based barcoding of dna for ultrasensitive mutation detection by next-generation sequencing," *Nature protocols*, vol. 12, no. 4, pp. 664–682, 2017.

[15] S. Schumacher, *Construction of endometrial carcinoma-associated dna libraries for application on liquid biopsies using simsen-seq*, Medicinaregatan, 2021.

[16] A. Kustanovich, R. Schwartz, T. Peretz, and A. Grinshpun, "Life and death of circulating cell-free dna," *Cancer Biology & Therapy*, vol. 20, no. 8, pp. 1057–1067, 2019, ISSN: 1538-4047. DOI: 10.1080/15384047.2019.1598759. [Online]. Available: https://dx.doi.org/10.1080/15384047.2019.1598759.

[17] R. B. Corcoran and B. A. Chabner, "Application of cell-free dna analysis to cancer treatment," *New England Journal of Medicine*, vol. 379, no. 18, pp. 1754–1765, 2018.

[18] T. Österlund, S. Filges, G. Johansson, and A. Ståhlberg, "Umierrorcorrect and umianalyzer: Software for consensus read generation, error correction, and visualization using unique molecular identifiers," *Clinical chemistry*, vol. 68, no. 11, pp. 1425–1435, 2022.

[19] M. Kubista, J. M. Andrade, M. Bengtsson, *et al.*, "The real-time polymerase chain reaction," *Molecular aspects of medicine*, vol. 27, no. 2-3, pp. 95–125, 2006.

[20] J. S. Dymond, "Explanatory chapter: Quantitative pcr," in *Methods in enzymology*, vol. 529, Elsevier, 2013, pp. 279–289.

[21] A. Technologies, *Agilent 4200 tapestation system manual*, 10/2022, SD-UF0000087 Rev. E, Waldbronn, Germany: Agilent Technologies, Inc., 2022. [Online]. Available: https://www.agilent.com.

[22] R. Westermeier, *Electrophoresis in practice: a guide to methods and applications of DNA and protein separations*. John Wiley & Sons, 2016.

[23] P. Micallef, M. L. Santamaría, M. Escobar, *et al.*, "Digital sequencing is improved by using structured unique molecular identifiers," *Genome Biology*, vol. 26, no. 1, p. 37, 2025.

[24] J. G. Tate, S. Bamford, H. C. Jubb, *et al.*, "Cosmic: The catalogue of somatic mutations in cancer," *Nucleic acids research*, vol. 47, no. D1, pp. D941–D947, 2018.

[25] N. C. for Biotechnology Information (NCBI), *Primer-blast: A tool for finding specific primers*, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/tools/primer-blast/.

[26] G. Perez, G. P. Barber, A. Benet-Pages, *et al.*, "The ucsc genome browser database: 2025 update," *Nucleic Acids Research*, vol. 53, no. D1, pp. D1243–D1249, 2025.

[27] Agilent Technologies, *D1000 screentape assay for tapestation systems quick guide*, 01/2021, For Research Use Only. Not for use in Diagnostic Procedures., Agilent Technologies, Germany, 2021. [Online]. Available: https://www.agilent.com/genomics/tapestation.

[28] K. Okonechnikov, O. Golosova, M. Fursov, and U. Team, "Unipro ugene: A unified bioinformatics toolkit," *Bioinformatics*, vol. 28, no. 8, pp. 1166–1167, 2012.

[29] OpenAI, *Chatgpt: Language model (may 2025 version)*, 2025. [Online]. Available: https://chat.openai.com.

[30] J. D. Cohen, L. Li, Y. Wang, *et al.*, "Detection and localization of surgically resectable cancers with a multi-analyte blood test," *Science*, vol. 359, no. 6378, pp. 926–930, 2018.

[31] D. Magee, V. Domenyuk, J. Abraham, *et al.*, "Characterization of plasma cell-free dna variants as of tumor-or clonal hematopoiesis-origin in 16,812 advanced cancer patients," *Clinical Cancer Research*, 2025.

A

Appendix - Protocols

This Appendix presents protocols of reagents and cycling conditions not portrayed in the method.

A.1 Target assay primer validation

A.1.1 Reagents

Table A.1: Regents and volumes per reaction used for qPCR in validation of target assays.

Reagent	**[µl]**
SYBR Green	5
Fwd/Rev Primer Mix (10uM)	0.4
Nuclease-free water	2.6
Genomic DNA* or H2O	2
Total	**10**

∗Different concentrations based on dilution series. In this project 0.08, 0.4, 2, 10 and 50 ng/μl was used, resulting in the final concentrations of 0.16, 0.8, 4, 20 and 100 ng/μl.

A.1.2 Cycling conditions

1. **Initial Hold:** 98°C for 3 s
2. **Cycling (45 cycles):**
   (a) 98°C for 10 s
   (b) 60°C for 30 s
   (c) 72°C for 20 s
3. **Melting Curve:** 59.5°C to 95°C at 0.1°C/s

A.2 Barcoded target assay validation

A.2.1 Resuspension of Oligos (Barcoded primers)

1. Spin down dried DNA oligos in the tube with benchtop centrifuge for 10 sec.
2. Add TE buffer to each oligo tube, to reach the concentration of 100 µM, according to the oligo amount in each tube.
3. Vortex the tube for 10 sec, and spin the tube for 10 sec.
4. Incubate the tube in the heating block in 55°C for 3 min.
5. Vortex the tube for 10 sec, and spin the tube for 10 sec.
6. Store the 100 µM oligo solution in -20°C freezer.

A.2.2 Barcode PCR

A.2.2.1 Reagents

Table A.2: Volume per reagent for one reaction of Barcode PCR. The final concentration of genomic DNA was (20 ng/ul).

Reagent	**Volume 1x (µl)**
SuperFi Buffer (5x)	2
10mM dNTP mix	0,2
Nuclease-free water	4,375
L-Carnatine (5M)	1
Primer mix fw/rw (1uM)*	0,4
Platinum SuperFi Pol (2U/ul)	0,025
Genomic DNA sample (10 ng/ul)	2
Total volume	**10**

∗Forward and reverse primers pre-diluted to a concentration of 1µl

A.2.2.2 Cycling conditions

1. **Initial Hold:** 98°C for 30 s
2. **Cycling (3 cycles):**
   (a) 98°C for 10 s
   (b) 60°C for 3 min
   (c) 72°C for 30 s
3. **Final Hold:** 4°C ∞

A.2.3 Barcode qPCR

A.2.3.1 Reagents

Table A.3: Regents and volumes per reaction used for qPCR in validation of barcoded target assays.

Reagent	Volume x1 (µl)
SYBG mix (2x)	5
Illumina index FW 10µM	0,4
Illumina index RW 10µM	0,4
Nuclease-free water	2,2
Product from Barcoding	2
Total volume	**10**

A.2.3.2 Cycling conditions

1. **Initial Hold:** 98°C for 3:00 min
2. **Cycling (45 cycles):**
   (a) 98°C for 10 s
   (b) 80°C for 1 s
   (c) 72°C for 30 s
   (d) 76°C for 30 s
3. **Melting Curve:** 60°C to 95°C at 0.2°C/s

A.2.4 Adaptor PCR

A.2.4.1 Reagents

Table A.4: Volume per reagent for one reaction of Adaptor PCR.

Reagent	Volume [µL]
SuperFi MM (2x)	20
Illumina index RW 10uM	1,6
Illumina index FW 10uM	1,6
Nuclease-free water	6,8
Product from Barcode PCR	10
Total volume	**40**

A.2.4.2 Cycling conditions

1. **Initial Hold:** 98°C for 3 min
2. **Cycling (30 cycles):**
   (a) 98°C for 10 s
   (b) 80°C for 1 s*
   (c) 72°C for 30 s*
   (d) 76°C for 30 s*
3. **Final Hold:** 4°C ∞

*With ramping at 0.2 °C/s

A.3 Multiplexing

A.3.1 Barcode PCR

A.3.1.1 Reagents

Table A.5: Volume per reagent for one reaction of Barcode PCR when creating a library/multiplexing. The final concentration of genomic DNA was (20 ng/ul).

Reagent	**Volume 1x (µl)**
SuperFi Buffer (5x)	2
10mM dNTP mix	0,2
Nuclease-free water	4,375
L-Carnatine (5M)	1
Primer(s) fw/rw + MP Mix (1uM)*	0,4
Platinum SuperFi Pol (2U/ul)	0,025
Genomic DNA sample (10 ng/ul)	2
Total volume	**10**

∗Forward and reverse primers pre-diluted together with multiplex aliquotes to a concentration of 1µl for all components

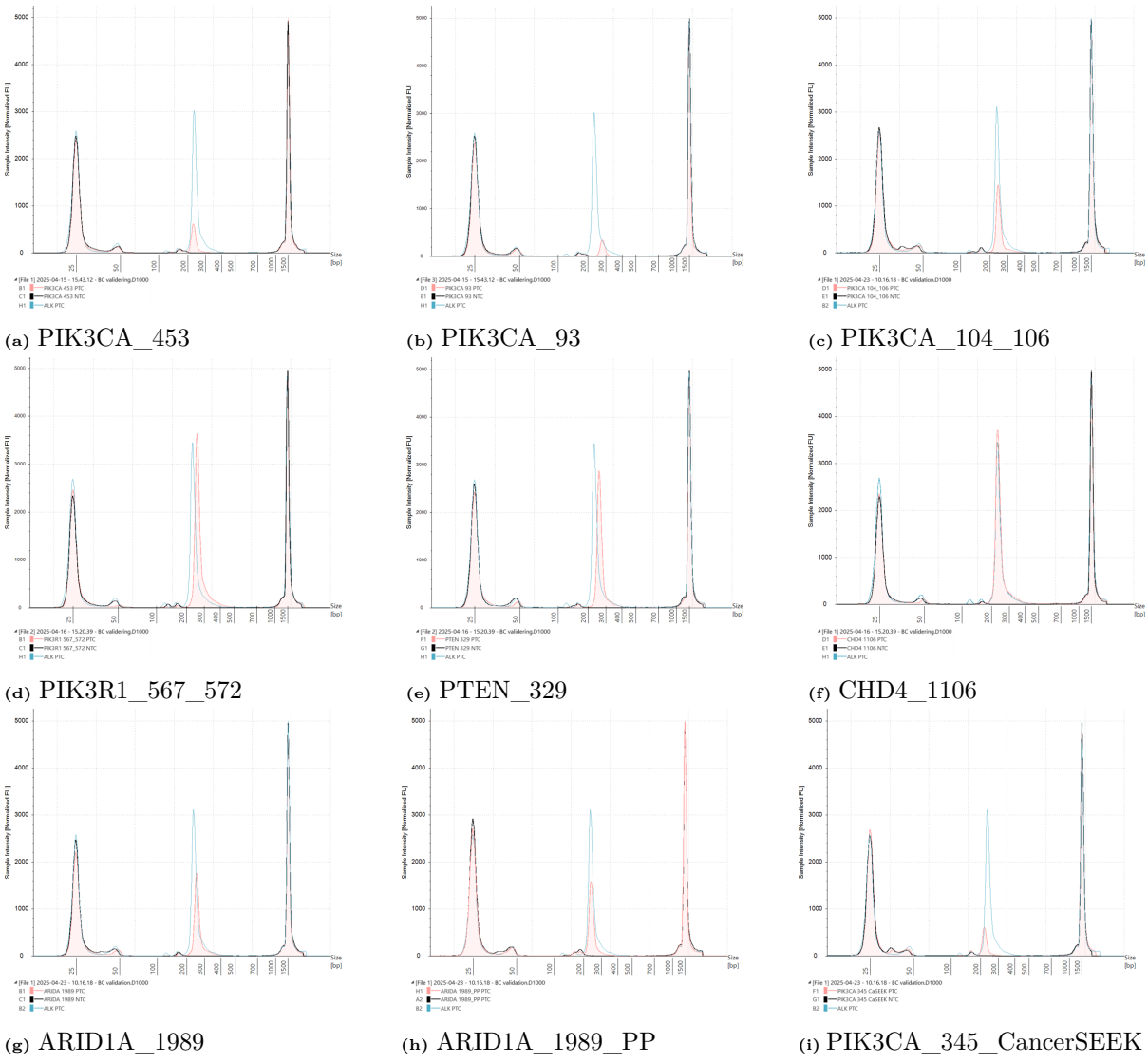
A.3.1.2 Cycling conditions

1. **Initial Hold:** 98°C for 30 s
2. **Cycling (5 cycles):**
   (a) 98°C for 10 s
   (b) 62°C for 6 min
   (c) 72°C for 30 s
3. **Final Hold:** 4°C ∞

B

Appendix - Data

B.1 Barcoded assay validation



(a) PIK3CA_453 (b) PIK3CA_93 (c) PIK3CA_104_106

(d) PIK3R1_567_572 (e) PTEN_329 (f) CHD4_1106

(g) ARID1A_1989 (h) ARID1A_1989_PP (i) PIK3CA_345_CancerSEEK

Figure B.1: Fragment analysis of PIK3CA_453, PIK3CA_93, PIK3CA_104_106, PIK3R1_567_572, PTEN_329, CHD4_1106, ARID1A_1989, ARID1A_1989_PP and PIK3CA_345_CancerSEEK. Graphs illustrating assay PTC (pink), NTC (black) and ALK reference (blue).

Table B.1: SiMSen-Seq data for barcoded target assay validation, calculated from qPCR. All mean values represents the mean Cq value of the triplicates for said thing.

Assay	PTC mean	NTC mean	ΔCq (NTC-PTC)	Validation >|1|	Control PTC mean	ΔCq (PTC-Control)	Validation <|1|
ARID1A_1989	26,72	33,49	6,77	Yes	26,25	0,48	Yes
PIK3CA_104_106	28,40	32,00	3,59	Yes	26,25	2,16	No
PIK3CA_345_CancerSEEK	29,67	34,69	5,02	Yes	26,25	3,42	No
ARID1A_1989_PP	27,82	29,84	2,02	Yes	26,25	1,57	No
PIK3R1_567_572	25,39	30,11	4,71	Yes	25,79	-0,40	Yes
CHD4_1104_1106	26,28	34,77	8,49	Yes	25,79	0,49	Yes
PTEN_329	26,80	33,04	6,24	Yes	25,79	1,01	No
PIK3CA_453	28,79	33,21	4,42	Yes	25,29	3,51	No
PIK3CA_93	29,66	33,97	4,31	Yes	25,29	4,37	No

B.2 Efficiency and Coverage

Table B.2: Overview of all validated target assays. Showing information on the most occuring mutation of the targeted site of interest, calculated efficiency, as well as the ALK reference assay efficency for intra-batch comparison. Assays targeting genes in *ARID1A*, *CHD4*, *PIK3CA*, *PIK3R1* and *PTEN*. Assays submitted for barcode validation marked in bold.

Gene	Site	Chromosomal Position	AA mutation	CDS mutation	Mutation type	Forward primer (5'->3')	Reverse primer (5'->3')	Amplicon size	Amplicon	Assay Efficiency 0.16-100 ng/ul DNA	ALK Efficiency 4-20 ng/ul DNA
ARID1A	ARID1A 1989	1:26779863..26779863	p.R1989*	c.5965C>T	Substitution - Nonsense	**GCGTCTGTGTGTCCAATACC**	**GATGAGCAGCAGCCCTGG**	**89**	**chr1:26779840+26779928**	**91.20**	**94.67**
						GCGTCTGTGTGTCCAATACC	TGAGCAGCAGCCCTGG	87	chr1:26779840+26779926	90.51	94.67
	ARID1A 543	1:26731403..26731403	p.Q543*	c.1627C>T	Substitution - Nonsense	CGGCTCCATACCCCTCCC	GCCTGTGGCTGTGAGTAGG	76	chr1:26731393+26731468	86.44	93.95
						GGCTCCATACCCCTCCCA	GCCTGTGGCTGTGAGTAGG	75	chr1:26731394+26731468	87.08	93.95
						CGGCTCCATACCCCTCC	GCCTGTGGCTGTGAGTAG	76	chr1:26731393+26731468	86.06	93.16
						CATCAGCAGTCCCCGGC	GGCTGGCTCTGGGGGT	68	chr1:26731380+26731447	95.91	97.44
	ARIDA 2158	1:26780370..26780370	p.R2158*	c.6472C>T	Substitution - Nonsense	TATGGTGCGCTTCCTCAGTG	CAGTACCACAGCCATCTCCC	61	chr1:26780348+26780408	102.57	99.63
						GGTGCGCTTCCTCAGTG	CAGCAGTACCACAGCCATCT	61	chr1:26780351+26780411	103.28	124.73
						ATGGTGCGCTTCCTCAGT	CAGTACCACAGCCATCTCC	60	chr1:26780349+26780408	100.62	124.73
						TGGTGCGCTTCCTCAGTG	CACAGCCATCTCCCGGC	53	chr1:26780350+26780402	88.39	90.21
						GGTGCGCTTCCTCAGTG	CAGTACCACAGCCATCTCCC	58	chr1:26780351+26780408	86.16	90.21
						GCCGCCTGGAGAAGTTGT	CAGCAGTACCACAGCCATCT	89	chr1:26780323+26780411	82.45	93.16
CHD4	CHD4 1106	12:6591493..6591493	p.R1105W	c.3313C>T	Substitution - Missense	TAAACAACTCCTTCTCTCTCACC	TCGATGGTGGAATCACTGGG	76	chr12:6591444+6591519	87.09	113.00
					Substitution - Missense	AAACAACTCCTTCTCTCTCACCA	GTGGAATCACTGGGAACATGC	69	chr12:6591445+6591513	86.55	113.00
						AAACAACTCCTTCTCTCTCACC	**CGATGGTGGAATCACTGGGA**	**74**	**chr12:6591445+6591518**	**98.97**	**107.93**
PIK3CA	PIK3CA 93	3:179199102..179199102	p.R93W	c.277C>T	Substitution - Missense	**ACTCAAGAAGCAGAAAGGGAAGA**	**CACGGTTGCCTACTGGTTCA**	**106**	**chr3:179199045+179199150**	**91.37**	**94.39**
					Substitution - Missense	TCAAGAAGCAGAAAGGGAAGAA	ACGGTTGCCTACTGGTTCA	103	chr3:179199047+179199149	94.42	94.39
	PIK3CA 104_106	3:179199136..179199136	p.P104L	c.311C>T	Substitution - Missense	CCTTCGGCTTTTTCAACCCT	TTGAGGATCTTTTCTTCACGGT	69	chr3:179199097+179199165	97.57	92.96
		3:179199142..179199142	p.G106V	c.317G>T	Substitution - Missense	**GACTTTGTGACCTTCGGCTT**	**TCGATTGAGGATCTTTTCTTCACG**	**83**	**chr3:179199087+179199169**	**90.43**	**92.96**
						TGTGACCTTCGGCTTTTTCAAC	CGATTGAGGATCTTTTCTTCACGG	77	chr3:179199093+179199169	96.66	99.47
						TGTGACCTTCGGCTTTTTCAAC	TTGAGGATCTTTTCTTCACGGT	74	chr3:179199093+179199166	95.31	99.47
	PIK3CA 345	3:179203765..179203765	p.N345K	c.1035T>A	Substitution - Missense	TCTTTGTGCAACCTACGTGA	AGCATCAGCATTTGACTTTACCT	67	chr3:179203744+179203810	94.48	96.07
					Substitution - Missense	TTCTTTGTGCAACCTACGTGA	AGCATCAGCATTTGACTTTACC	68	chr3:179203743+179203810	94.07	88.93
						ATTCTTTGTGCAACCTACGTGA	GCATCAGCATTTGACTTTACCTT	68	chr3:179203742+179203809	90.06	88.93
	PIK3CA 453	3:179210291..179210291	p.E453K	c.1357G>A	Substitution - Missense	TCTGGAAAAATGGCTTTGAATCTT	ACACCAATAGGGTTCAGCAA	71	chr3:179210246+179210316	97.95	96.83
						ATCTGGAAAAATGGCTTTGAATCT	**ACACCAATAGGGTTCAGCAAA**	**72**	**chr3:179210245+179210316**	**94.91**	**99.00**
						GGAAAAATGGCTTTGAATCTTTGGC	TTGATCCAGTAACACCAATAGGG	79	chr3:179210249+179210327	95.94	104.48
						GAAAAATGGCTTTGAATCTTTGGC	TGATCCAGTAACACCAATAGGG	77	chr3:179210250+179210326	94.04	99.00
						TGGAAAAATGGCTTTGAATCTTTG	ACACCAATAGGGTTCAGCAAAT	69	chr3:179210248+179210316	86.99	96.07
PIK3R1	PIK3R1 379	5:68293310..68293310	p.G376R	c.1126G>A	Substitution - Missense	TGTTTTTACAGGAAAGGGGGA	TGGGTCAGAGAAGCCATATTTC	78	chr5:68293292+68293369	117.73	84.39
						GTTTTTACAGGAAAGGGGGAAAT	TAATGGGTCAGAGAAGCCATATT	80	chr5:68293293+68293372	118.96	94.20
						TGTTTTTACAGGAAAGGGGGAA	AATGGGTCAGAGAAGCCATATT	80	chr5:68293292+68293371	147.30	95.43
						TTGTTTTTACAGGAAAGGGGGA	ATGGGTCAGAGAAGCCATATTT	80	chr5:68293291+68293370	116.68	97.12
	PIK3R1 567_572	5:68295278..68295278	p.K567E	c.1699A>G	Substitution - Missense	**GACTTGAAGAAGCAGGCAGC**	**TGGTCTCTCGTCTTTCTCAGC**	**95**	**chr5:68295221+68295315**	**102.95**	**107.93**
						AGACTTGAAGAAGCAGGCAGC	TTGGTCTCTCGTCTTTCTCAGC	97	chr5:68295220+68295316	104.88	96.32
PTEN	PTEN 329	10:87961079..87961082	p.N329Kfs*14	c.987_990del	Deletion - Frameshift	AGCGTGCAGATAATGACAAGGA	TGGAGAAAAGTATCGGTTGGCT	98	chr10:87961012+87961109	101.68	96.32
						GCGTGCAGATAATGACAAGGA	TGGAGAAAAGTATCGGTTGGC	97	chr10:87961013+87961109	99.83	103.47
						AGCGTGCAGATAATGACAAGG	GGAGAAAAGTATCGGTTGGCT	97	chr10:87961012+87961108	98.33	103.47
						AGCGTGCAGATAATGACAAGGA	**GAAAAGTATCGGTTGGCTTTGT**	**94**	**chr10:87961012+87961105**	**97.80**	**99.62**
	PTEN 341	10:87961113..87961113	p.F341V	c.1021T>G	Substitution - Missense	AAAGACAAAGCCAACCGATACT	ACAAGTCAACAACCCCCACAA	79	chr10:87961080+87961158	105.01	121.65
						GACAAAGCAAATAAAGACAAAGCC	CAACAACCCCCACAAAATGT	85	chr10:87961068+87961152	101.63	121.65
						ACAAAGCAAATAAAGACAAAGCCA	GTCAACAACCCCCACAAAATGT	86	chr10:87961069+87961154	85.30	92.63
						ACAAAGCAAATAAAGACAAAGCCA	CAACAACCCCCACAAAATGT	84	chr10:87961069+87961152	84.43	92.63
						TAAAGACAAAGCCAACCGATACT	CAAGTCAACAACCCCCACAA	79	chr10:87961079+87961157	87.53	98.12
						AAAGACAAAGCCAACCGATACTT	AGTCAACAACCCCCACAAAATG	76	chr10:87961080+87961155	87.50	98.12

Table B.3: Coverage per gene for the individual target assays tested in multiplexe: ARID1A_1989, ARID1A_1989_PP, CHD4_1106, PIK3CA_104_106, PIK3CA_345_CancerSEEK, PIK3R1_567_572, PIK3R1_564 and PTEN_329. Calculations based on data from COSMIC.

Gene	Assay	Target region size (bp)	Covered samples	Mutated samples (per gene)	Tested samples (per gene)	Coverage % (Mutated samples)	Coverage % (Tested samples)
ARID1A	ARID1A_1989	50	48	601	1411	8	3.4
	ARID1A_1989_PP	34	47	601	1411	7.8	3.3
CHD4	CHD4_1106	32	19	160	423	11.9	4.5
PIK3CA	PIK3CA_104_106	38	36	1452	4633	2.5	0.8
	PIK3CA_345_CancerSEEK	15	59	1452	4633	4.1	1.3
PIK3R1	PIK3R1_567_572	53	70	344	2546	20.3	2.7
	PIK3R1_564	20	70	344	2546	5.8	0.8
PTEN	PTEN_329	49	215	1771	3999	12.1	5.4

C

Appendix - Source Code

C.1 Code for Coverage calculations

Code used for Coverage calculations. Written in programming language R.

```
shell("cls")       # Clears console (Windows only, If Mac: shell("clear")
    )
rm(list = ls())    # Clears environment

setwd("C:/Users/Benny/OneDrive - Chalmers/ExjobbVT25/Assay Coverage")

library(dplyr)           # Data manipulation
library(tidyr)           # Column splitting / reshaping
library(readxl)          # Read Excel files
library(GenomicRanges)
library(openxlsx)        # Create and style Excel workbooks
library(stringr)         # String operations
library(ggplot2)         # Plotting

# 1. LOAD SAMPLES TESTED PER GENE
samples_tested_df <- read_excel("Samples_tested_per_gene.xlsx")

# Check that required columns are present
if (!all(c("Gene samples", "Samples tested") %in% colnames(samples_
    tested_df))) {
  stop(" 'Samples_tested_per_gene.xlsx' must include columns: 'Gene
    samples' and 'Samples tested'")
}

# 2. LOAD TARGET ASSAY COORDINATES
TAs_coordinates_all <- read_excel("TargetAssays.xlsx", sheet = 1) %>%
  mutate(
    Target.region.Start = as.numeric(Target.region.Start),
    Target.region.End   = as.numeric(Target.region.End),
    Target.region.Start = replace_na(Target.region.Start, 0),
    Target.region.End   = replace_na(Target.region.End,   0)
  )

# Validate required columns
if (!all(c("Gene", "Target.region.Start", "Target.region.End") %in%
    colnames(TAs_coordinates_all))) {
  stop(" 'TestCoverage_TACoords.xlsx' must include columns: 'Gene', '
    Target.region.Start', 'Target.region.End'")
}

```

```
# 3. LOAD COSMIC DATA
cosmic_data_raw <- read.csv("COSMIC_ALL_SAMPLES_Mutated.csv", header = TRUE)

# Check required columns
required_cols <- c("Gene.Name", "Sample.ID", "AA.Mutation", "Genomic.Co.ordinates")
if (!all(required_cols %in% colnames(cosmic_data_raw))) {
  stop(" COSMIC file must contain columns: Gene.Name, Sample.ID, AA.Mutation, Genomic.Co.ordinates")
}

# Filter COSMIC data to only include genes found in TA coordinates
valid_genes <- unique(str_extract(TAs_coordinates_all$Gene, "^[^_]+"))
cosmic_data <- cosmic_data_raw %>% filter(Gene.Name %in% valid_genes)

# Split COSMIC data into list of dataframes by gene name
cosmic_data_by_gene <- split(cosmic_data, cosmic_data$Gene.Name)


# 4. SET UP OUTPUT WORKBOOK
wb <- createWorkbook()
addWorksheet(wb, "Assay_Level_Coverage")
addWorksheet(wb, "Gene_Level_Coverage")

# Style for bold headers
headerStyle <- createStyle(textDecoration = "bold")

# Containers for results = memory storage
all_gene_coverage <- list()
all_gene_assay_coverage <- tibble()   #Tibble = more advanced data frame. Cleaner, more consistent way of handling data.
all_mut_samples <- c()
all_covered_samples <- c()


# 5. LOOP OVER EACH GENE
for (gene_name in names(cosmic_data_by_gene)) {

  # Extract mutation data for the current gene
  mut_data <- cosmic_data_by_gene[[gene_name]]

  # Retrieve the number of samples tested for this gene from another dataframe
  samples_tested <- samples_tested_df %>%
    filter(`Gene samples` == gene_name) %>%
    pull(`Samples tested`)

  # If no sample data found, skip this gene
  if (length(samples_tested) == 0) {
    message(" No samples tested found for gene: ", gene_name, " Skipping.")
    next
  }

  # Clean and transform mutation data:
```

```
86  # - Remove prefix from Genomic.Co.ordinates
87  # - Split into start and end positions
88  # - Convert to numeric and replace NA with 0
89  # - Extract amino acid mutation position using regex
90  # - Create a column for overlapping target assays (TAs)
91  mut_data <- mut_data %>%
92    mutate(Genomic.Co.ordinates = sub(".*:", "", Genomic.Co.ordinates))
     %>%
93    separate(Genomic.Co.ordinates, into = c("Start", "End"), sep = "\\..
    ") %>%
94    mutate(
95      Start = as.numeric(Start),
96      End = as.numeric(End),
97      Start = replace_na(Start, 0),
98      End = replace_na(End, 0),
99      AA_mutation_pos = as.numeric(str_extract(AA.Mutation, "(?<=p\\.\\D
    {1})(\\d+)", group = 1)),
100     overlapping_TA = NA_character_
101   )
102
103 # Store all sample IDs with mutations
104 all_mut_samples <- c(all_mut_samples, mut_data$Sample.ID)
105
106 # Filter and initialize target assay coordinates related to this gene
107 TAs_coordinates <- TAs_coordinates_all %>%
108   filter(str_detect(Gene, paste0("^", gene_name, "_"))) %>%
109   mutate(mutation_count = 0)
110
111 # Skip if no assay coordinates exist for this gene
112 if (nrow(TAs_coordinates) == 0) {
113   message(" No TA coordinates found for gene: ", gene_name, " Skipping
    .")
114   next
115 }
116
117 # Check for genomic overlaps: for each mutation, find overlapping TA
     regions
118 mut_data <- mut_data %>%
119   rowwise() %>%
120   mutate(
121     overlapping_TA = {
122       overlaps <- TAs_coordinates %>%
123         filter(Target.region.Start <= End, Target.region.End >= Start)
    %>%
124         pull(Gene)
125       if (length(overlaps) > 0) paste(unique(overlaps), collapse = ";
    ") else NA_character_
126     }
127   ) %>%
128   ungroup()
129
130 # Special case: if Start/End are missing (0), use AA mutation position
    to find overlap
131 for (i in seq_len(nrow(mut_data))) {
132   row <- mut_data[i, ]
133   if (row$Start == 0 && row$End == 0 && !is.na(row$AA_mutation_pos)) {
```

```
      ov <- c()
      for (j in seq_len(nrow(TAs_coordinates))) {
        ta <- TAs_coordinates[j, ]
        ta_muts <- filter(mut_data,
                          Start >= ta$Target.region.Start,
                          End <= ta$Target.region.End)
        if (nrow(ta_muts) > 0) {
          min_aa <- min(ta_muts$AA_mutation_pos, na.rm = TRUE)
          max_aa <- max(ta_muts$AA_mutation_pos, na.rm = TRUE)
          if (!is.na(min_aa) &&
              row$AA_mutation_pos >= min_aa &&
              row$AA_mutation_pos <= max_aa) {
            ov <- c(ov, ta$Gene)
          }
        }
      }
      if (length(ov) > 0) {
        mut_data$overlapping_TA[i] <- paste(unique(ov), collapse = "; ")
      }
    }
  }

  # Get all sample IDs with at least one overlapping TA
  covered_samples_gene <- mut_data %>%
    filter(!is.na(overlapping_TA)) %>%
    pull(Sample.ID)

  # Update the global covered sample list
  all_covered_samples <- c(all_covered_samples, covered_samples_gene)

  # Increment mutation counts in TAs based on overlapping mutations
  for (i in seq_len(nrow(mut_data))) {
    s <- mut_data$overlapping_TA[i]
    if (!is.na(s)) {
      genes <- str_split(s, ";\\s*")[[1]]
      for (g in genes) {
        idx <- which(TAs_coordinates$Gene == g)
        if (length(idx)) {
          TAs_coordinates$mutation_count[idx] <-
            TAs_coordinates$mutation_count[idx] + 1
        }
      }
    }
  }

  # OUTPUT: write mutation data per gene to Excel sheet
  mut_data_out <- select(mut_data, -AA_mutation_pos)  # Drop helper
   column
  sht1 <- paste0(gene_name, "_Mutations")
  addWorksheet(wb, sht1)
  writeData(wb, sht1, mut_data_out)

  # Apply conditional formatting to overlapping_TA column:
  # Orange = overlapping assays, Green = one assay
  col_TA <- which(names(mut_data_out) == "overlapping_TA")
  let <- int2col(col_TA)
```

```
189   conditionalFormatting(wb, sht1,
190                         cols = col_TA,
191                         rows = 2:(nrow(mut_data_out)+1),
192                         rule = paste0('LEN(',let,'2)-LEN(SUBSTITUTE(',
      let,'2,";",""))>=1'),
193                         style = createStyle(bgFill="orange"))
194   conditionalFormatting(wb, sht1,
195                         cols = col_TA,
196                         rows = 2:(nrow(mut_data_out)+1),
197                         rule = paste0('AND(',let,'2<>"",LEN(',let,'2)-
      LEN(SUBSTITUTE(',let,'2,";",""))=0)'),
198                         style = createStyle(bgFill="lightgreen"))
199
200   # Summary stats for this gene
201   mutated_samples <- n_distinct(mut_data$Sample.ID)
202   covered_samples <- n_distinct(covered_samples_gene)
203
204   # Calculate coverage per assay (TA) for this gene
205   coverage_per_assay <- TAs_coordinates %>%
206     transmute(
207       Assay = Gene,
208       Gene = gene_name,
209       Target.region.Start,
210       Target.region.End,
211       Covered.samples = mutation_count,
212       Mutated.samples = mutated_samples,
213       Coverage_mutated_samples = round(Covered.samples / Mutated.samples
      * 100, 1),
214       Samples_tested = samples_tested,
215       Coverage_tested_samples = round(Covered.samples / Samples_tested *
      100, 1)
216     )
217
218   # Append to overall TA coverage data
219   all_gene_assay_coverage <- bind_rows(all_gene_assay_coverage, coverage
      _per_assay)
220
221   # Store overall coverage summary for the gene
222   coverage_per_gene <- tibble(
223     Gene = gene_name,
224     Covered_samples_gene_tot = covered_samples,
225     Mutated_samples = mutated_samples,
226     Coverage_mutated_samples_tot = round(covered_samples / mutated_
      samples * 100, 1),
227     Samples_tested = samples_tested,
228     Coverage_tested_samples_tot = round(covered_samples / samples_tested
      * 100, 1)
229   )
230   all_gene_coverage[[gene_name]] <- coverage_per_gene
231 }
232
233
234 # 6. FINAL SUMMARY SHEETS FOR COVERAGE
235
236 # Add assay coverage sheet
237 writeData(wb, "Assay_Level_Coverage", all_gene_assay_coverage)
```

```
238
239 # Alternating background color by gene
240 gene_column <- all_gene_assay_coverage$Gene
241 row_colors <- ifelse(as.numeric(as.factor(gene_column)) %% 2 == 0, "#
    FFFFFF", "lightgray")
242 for (i in seq_along(row_colors)) {
243   conditionalFormatting(wb, "Assay_Level_Coverage",
244                         cols = 1:ncol(all_gene_assay_coverage),
245                         rows = i + 1,
246                         rule = "TRUE",
247                         style = createStyle(bgFill = row_colors[i]))
248 }
249
250 # Add gene-level summary sheet
251 coverage_per_gene_all <- bind_rows(all_gene_coverage)
252
253 total_samples_tested <- samples_tested_df %>%
254   filter(`Gene samples` == "TOTAL_ENDOMETRIUM_ALL_TESTED") %>%
255   pull(`Samples tested`)
256 unique_total_samples_mutated <- unique(cosmic_data_raw$Sample.ID)
257 total_samples_mutated <- length(unique_total_samples_mutated)
258
259 unique_mut_samples     <- unique(all_mut_samples)
260 unique_covered_samples <- unique(all_covered_samples)
261 total_mutated_samples  <- length(unique_mut_samples)
262 total_covered_samples  <- length(unique_covered_samples)
263
264 total_coverage_panel_genes <- tibble(
265   Gene = "TOTAL over genes targeted by panel",
266   Covered_samples_gene_tot = total_covered_samples,
267   Mutated_samples = total_mutated_samples,
268   Coverage_mutated_samples_tot = round(total_covered_samples / total_
    mutated_samples * 100, 1),
269 )
270
271 total_coverage_all <- tibble(
272   Gene = "TOTAL over all genes",
273   Covered_samples_gene_tot = total_covered_samples,
274   Mutated_samples = total_samples_mutated,
275   Coverage_mutated_samples_tot = round(total_covered_samples / total_
    samples_mutated * 100, 1),
276   Samples_tested = total_samples_tested,
277   Coverage_tested_samples_tot = round(total_covered_samples / total_
    samples_tested * 100, 1)
278 )
279
280 coverage_per_gene_all <- bind_rows(coverage_per_gene_all, total_coverage
    _panel_genes)
281 coverage_per_gene_all <- bind_rows(coverage_per_gene_all, total_coverage
    _all)
282
283 writeData(wb, "Gene_Level_Coverage", coverage_per_gene_all)
284
285 # Gray background for TOTAL row
286 total_row <- nrow(coverage_per_gene_all) + 1
287 gray_style <- createStyle(bgFill = "#D9D9D9", textDecoration = "bold")
```

```
conditionalFormatting(wb, "Gene_Level_Coverage",
                      cols = 1:ncol(coverage_per_gene_all),
                      rows = (total_row-1):total_row,
                      rule = "TRUE",
                      style = gray_style)


# 7. HISTOLOGY SUMMARY FOR COVERED SAMPLES

# Check if Histology column exists
if (!"Histology" %in% colnames(cosmic_data_raw)) {
  warning("No 'Histology' column found in COSMIC data. Skipping histology summary.")
} else {

  # Create total sample counts per Histology across all samples (not just covered)
  Total_Histology_Counts <- cosmic_data_raw %>%
    distinct(Sample.ID, Histology) %>%
    group_by(Histology) %>%
    summarise(Total_Sample_Count = n(), .groups = "drop")

  # Filter to covered samples only and count per Histology
  Histology_Summary <- cosmic_data_raw %>%
    filter(Sample.ID %in% all_covered_samples) %>%
    group_by(Histology) %>%
    summarise(Covered_Sample_Count = n_distinct(Sample.ID), .groups = "drop") %>%
    arrange(desc(Covered_Sample_Count))

  # Join with total counts
  Histology_Summary <- Histology_Summary %>%
    left_join(Total_Histology_Counts, by = "Histology")

  # Calculate percentage of covered samples
  total_covered_histology <- sum(Histology_Summary$Covered_Sample_Count)
  Histology_Summary <- Histology_Summary %>%
    mutate(Percentage.of.covered.samples = round(Covered_Sample_Count / total_covered_histology * 100, 1))

  # Add to workbook
  addWorksheet(wb, "Histology_Summary")
  writeData(wb, "Histology_Summary", Histology_Summary)

  # Formatting
  addStyle(wb, "Histology_Summary", style = headerStyle, rows = 1, cols = 1:4, gridExpand = TRUE)
  freezePane(wb, "Histology_Summary", firstActiveRow = 2)
  setColWidths(wb, "Histology_Summary", cols = 1:4, widths = "auto")

  # Breakdown of "Other" histology by subtype

  if ("Histology.Subtype.1" %in% colnames(cosmic_data_raw)) {

    # Total counts for subtypes across all samples (not just covered), where Histology == "Other"
```

```
338     Total_Other_Subtypes <- cosmic_data_raw %>%
339       filter(Histology == "Other") %>%
340       distinct(Sample.ID, Histology.Subtype.1) %>%
341       group_by(Histology.Subtype.1) %>%
342       summarise(Total_Sample_Count = n(), .groups = "drop")
343 
344     # Covered sample counts for subtypes
345     other_subtypes <- cosmic_data_raw %>%
346       filter(Sample.ID %in% all_covered_samples, Histology == "Other") %>%
347       group_by(Histology.Subtype.1) %>%
348       summarise(Other_Sample_Count = n_distinct(Sample.ID), .groups = "drop") %>%
349       arrange(desc(Other_Sample_Count))
350 
351     # Join with total counts
352     other_subtypes <- other_subtypes %>%
353       left_join(Total_Other_Subtypes, by = "Histology.Subtype.1") %>%
354       mutate(Percentage.of.covered.samples = round(Other_Sample_Count / total_covered_histology * 100, 1))
355 
356     # Row to insert subtitle
357     start_row <- nrow(Histology_Summary) + 4
358     writeData(wb, "Histology_Summary", "Breakdown of 'Other' histology by subtype:",
359               startRow = start_row - 1, colNames = FALSE)
360 
361     # Write the subtype breakdown table
362     writeData(wb, "Histology_Summary", other_subtypes,
363               startRow = start_row, colNames = TRUE)
364 
365     # Style for header
366     addStyle(wb, "Histology_Summary", style = headerStyle,
367              rows = start_row, cols = 1:ncol(other_subtypes), gridExpand = TRUE)
368 
369     setColWidths(wb, "Histology_Summary", cols = 1:ncol(other_subtypes), widths = "auto")
370 
371   } else {
372     message(" Column 'Histology.Subtype.1' not found: skipping subtype breakdown.")
373   }
374 }
375 
376 
377 # 8. ORDER SHEETS AND SAVE
378 
379 # Apply formatting to all sheets
380 for (sheetName in names(wb)) {
381   sheetData <- tryCatch(
382     openxlsx::readWorkbook(wb, sheet = sheetName),
383     error = function(e) NULL
384   )
385 
386   if (!is.null(sheetData)) {
```

```
387     n_cols <- ncol(sheetData)
388 
389     # Bold header row
390     addStyle(wb, sheet = sheetName, style = headerStyle, rows = 1, cols
        = 1:n_cols, gridExpand = TRUE)
391 
392     # Freeze first row
393     freezePane(wb, sheet = sheetName, firstActiveRow = 2)
394 
395   }
396 }
397 
398 # Auto-adjust column widths
399 setColWidths(wb, "Assay_Level_Coverage", cols = 1:n_cols, widths = "auto
        ")
400 setColWidths(wb, "Gene_Level_Coverage", cols = 1:n_cols, widths = "auto"
        )
401 
402 wb$sheetOrder <- append(wb$sheetOrder[wb$sheetOrder != which(names(wb)
        == "Histology_Summary")],
403                         which(names(wb) == "Histology_Summary"),
404                         after = 2)
405 
406 outname <- paste0("MultiGene_TA_Coverage_", Sys.Date(), ".xlsx")
407 saveWorkbook(wb, outname, overwrite = TRUE)
408 cat(" Written: ", outname, "\n")
```

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

CHALMERS