Norges miljø- og
biovitenskapelige
universitet

Master's Thesis 2023   60 ECTS
Faculty of Chemistry, Biotechnology and Food Science

Characterization of single cells from breast tumors with mutations implicated in treatment response

Emma Rapp
Master of Science in Biotechnology and Chemistry

Acknowledgements

This thesis is the final project towards my degree in Master of Science in Biotechnology and Molecular biology at the department of Chemistry, Biotechnology and Food science at the Norwegian University of Life Sciences. It was performed in Vessela Kristensen group at Ullevål Oslo University Hospital at the department of Medical Genetics. The work was done under supervision from Grethe Grenaker Alnæs and Vessela Kristensen as external supervisors and Harald Carlsen as internal supervisor from September 2022 to May 2023.

I would firstly like to thank my external supervisors Grethe Grenaker Alnæs and Vessela Kristensen for being good mentors throughout this period, providing valuable advice, feedback, and scientific knowledge throughout the whole period. I want to thank Grethe Grenaker Alnæs for helping me in the lab, and for all the great motivation and advice. Thank you Vessela Kristensen for giving me the opportunity to do my thesis and working in your team. I also want to thank my internal supervisor Harald Carlsen for valuable insight and advice during the writing process of the thesis.

I would like to give a special thanks to all the patients and their contributions to this research by participating in this clinical trial. Thank you, Jürgen Geisler, for pushing forward clinical trials and for your support to the group.

Lastly, I would like to thank my family and peers for all the support and encouragement they have given me throughout my years of study.

Emma Rapp

Emma Rapp

Abstract

Breast cancer has high tumor heterogeneity which can be challenging for individual patients. It is of importance to characterize the mutations in the tumors of breast cancer patients that affect the eradication of the malignant cells because of the treatment. By studying individual cells in the tumor instead of the entire tumor's genetic material, cells of resistance at a single cell level can be identified. It is seen that the mutations which occur simultaneously in each cell, identify clones and the pseudo-times when they arise, and how these clones develop during the treatment.

In the clinical trial NeoLetExe breast cancer patients are treated neoadjuvant with Letrozole and Exemestane sequentially in the following manner: group 1 first receives Letrozole and switches to Exemestane after 2 months, and group 2 first receives Exemestane before switching to Letrozole after 2 months. Biopsies are taken before starting treatment, when changing medication and during surgery. Letrozole and Exemestane are hormone therapy drugs that are used as treatment for breast cancer in postmenopausal patients.

To study individual tumor cells, we employ a unique platform for single-cell analysis. The Mission Bio Tapestri platform is an instrument where single cells are separated and mixed with reagents to perform the analysis in isolated oil droplets, and results in specific parts of the genes being marked at a single cell level. A custom panel of 497 amplicons covering 528 sequence variants was designed in a pilot study. Tumor tissue from three patients at one timepoint (baseline), one patient with three time points (baseline, 2 months, and 4 months) and one patient with two time points (baseline and 4 months) in the NeoLetExe clinical trial was analyzed and shown in this thesis. It was discovered that some genes like ZNF717 were mutated in several or all the samples, and it was observed that clones with certain sequence variants were altered during treatment. Basic software and analysis strategies to analyze the complex data achieved with targeted single-cell sequencing is presented in this thesis, however more efficient analysis strategies should be investigated.

Table of content

Acknowledgements **1**

Abstract **2**

Abbreviations **5**

1. Theory **7**

1.1. The breast – anatomy and cancer incidence *7*

1.1.1 Breast cancer initiation and progression 9

1.2 Classification of breast cancer *11*

1.2.1 Histological classification 12
1.2.2 Grade 13
1.2.3 Stage 13
1.2.4 Biomarkers 14

1.3 Breast cancer subtypes *14*

1.4 Breast cancer heterogeneity *15*

1.5 Treatment *16*

1.5.1 Surgery 17
1.5.2 Radiotherapy 17
1.5.3 Chemotherapy 18
1.5.4 Endocrine therapy 18

1.6 Single cell analyses *19*

2. Aim of the study **22**

3. Materials and method **22**

3.1. Patient information *22*

3.2. Tissue preparation and nuclei extraction *23*

3.2.1. Manual isolation of nuclei from breast tissue 23
3.2.2. Singulator 100 and Percoll clean-up 25

3.3. Quality control *27*

3.4. Mission Bio Tapestri platform *27*

3.4.1. Custom made targeted DNA breast cancer panel 28
3.4.2. Custom single-cell DNA sequencing 29

3.5. Computational analysis *30*

3.6. Own contributions *32*

4. Results **33**

4.1. Nuclei Extraction *33*

4.2. Targeted sequencing *36*

4.2.1. Tapestri Pipeline – sequencing quality data 36
4.2.2. Comparison of variants in bulk-seq and scDNA-seq 37
4.2.3. Interesting variants and clones found by Tapestri targeted single-cell DNA sequencing 42

5. Discussion and future perspectives **46**

5.1. Methodological evaluation *46*

5.2. Variants from WGS and targeted sequencing *48*

5.3. Variants from Tapestri Insights 49

6. Conclusion 52

Sources 53

Appendix 1; Solutions 57

Appendix 2; Supplies 59

Appendix 3; Reagents 61

Appendix 4; Pathology data 63

Appendix 5; NGS and extra library clean up 64

Appendix 6; Example code in Mosaic 68

Appendix 7; Variants used for construction of plots in Mosaic 72

Appendix 8; Variants of interest and %mutated cells 75

Abbreviations

ADH	Atypical ductal hyperplasia
ADO	Allele dropout
AI	Aromatase inhibitor
ALH	Atypical lobular hyperplasia
ALND	Axillary lymph node dissection
BBD	Benign breast disease
BM	Basement membrane
CAN	Copy number aberrations
CNV	Copy number variant
CSC	Cancer stem cell
DCIS	Ductal carcinoma in situ
DNA	Deoxyribonucleic acid
ER	Estrogen receptor
GEP	Gene expression profiling
HCC	Hepatocellular carcinomas
HER2	Human epidermal growth factor 2
HR	Hormone receptor
IDC	Invasive ductal carcinoma
ILC	Invasive lobular carcinoma
LCIS	Lobular carcinoma in situ
LFS	Li-Fraumeni Syndrome
NGS	next generation sequencing
NIR	Nuclei isolation reagent
NSR	Nuclei storage reagent
PCR	Polymerase chain reaction
PI	Propidium Iodide
PR	Progesterone receptor
PRS	Polygenic risk score
RNA	Ribonucleic acid
ROR	Risk of recurrence

scDNA-seq	Single-cell DNA sequencing
scRNA-seq	Single-cell RNA sequencing
SERM	Selective estrogen receptor modulator
SERD	Selective modulators estrogen receptor degrader
SNV	Single nucleotide variant
TDLU	Terminal duct lobular unit
TME	Tumor microenvironment
TNBC	Triple negative breast cancer
VAF	Variant allele frequency
WGS	Whole genome sequencing
WHO	World Health Organization
WT	Wild-type

1. Theory

1.1. The breast – anatomy and cancer incidence

The female breast is a gland, and its function is to produce and secrete milk to nourish offspring. It mainly consists of skin, fat, Cooper ligaments, fibro glandular tissue, lymphatics and neurovascular structures [1] as shown in Figure 1. Mammary glands develop through four stages: embryonic, pubertal, adult, and reproductive. In the embryonic stage the rudimentary ductal tree is formed, which includes secondary branches and a primary duct. During puberty hormones such as estrogen will induce ductal elongation and branching. Each breast has approximately 15-20 lobes, which includes lobules and ducts. Lobules are the functional unit of the breast and is also called terminal duct lobular unit (TDLU). These structures consist of a myoepithelial layer of cells with an inner layer of luminal cells, which can further be differentiated into ductal luminal cells. These cells line the inside of the ducts and secrete milk during lactation [2]. During menopause hormone levels such as estrogen decline, which results in shrinkage of the lobes [3]. Through these phases mammary gland development is regulated by complex network of hormones and local growth factors, such as estrogen and progesterone [4].

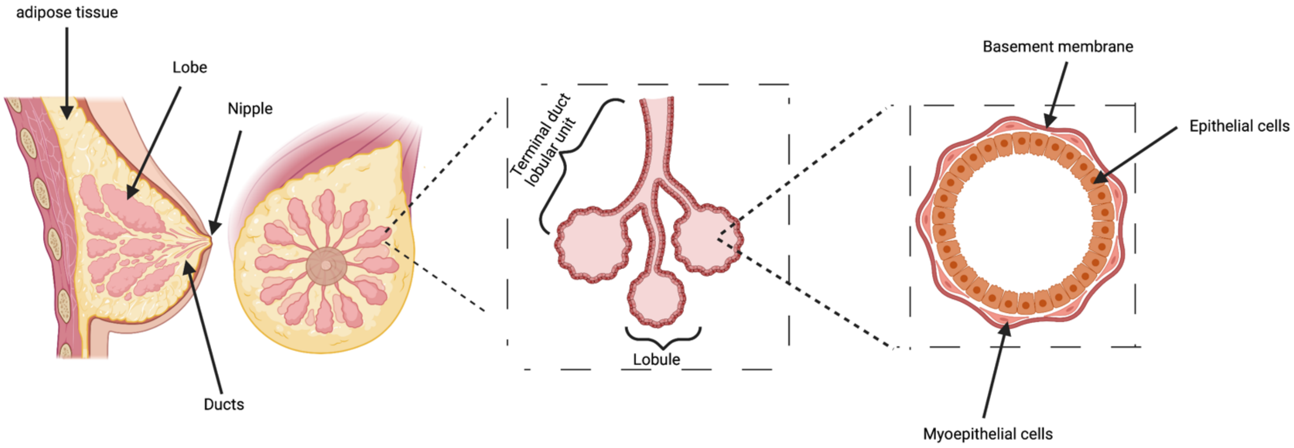


Figure 1. Overview of the anatomy of the female breast, lobes with TDLUs and lobules and the inner layer consisting of the basement membrane, epithelial cells and myoepithelial cells [5].

Most breast cancers originate from the epithelial cells lining the TDLUs, ductal carcinoma and lobular carcinoma [6, 7]. Known risk factors are age, parity alcohol consumption, genetic predisposition, epigenetics, among others (age at menarchy, time of breastfeeding, etc) [8]. Polygenic risk score (PRS) is based on the sum of all risk variants in an individual that can be used to predict breast cancer risks by genetic risk factors. PRS is used in combination with clinical risk factors to provide breast cancer estimates and improved medical management [9]. Risk of recurrence (ROR) score is used to predict the benefits of adjuvant therapy in early-stage breast cancer and the ROR. This score was developed from the PAM50 test and is derived from the expression profile of these 50 genes [10].

Breast cancer accounts for 25% of cancer cases in women worldwide and is the second most fatal type of cancer in women with a mortality rate of 15% of cancer deaths globally in 2018 [11]. In Norway there is a biennial mammography screening program for women in the age 50-69 years where 75% of the invited women attend [12]. In 2021 3991 new incidents were registered in Norway which is a 5.2% increase in incident rates, this is illustrated in Figure 2. This is mainly related to the reduced activity in the screening program in 2020 due to the covid-19 pandemic [13]. Observations show that the mortality rate of breast cancer has decreased since the implementation of the screening program and the five-year relative survival rate was 92.3% in 2021 [14]. The mortality rate is influenced by participation in the screening program, sensitive and specific diagnosis methods, and a wide specter of treatment options [15].

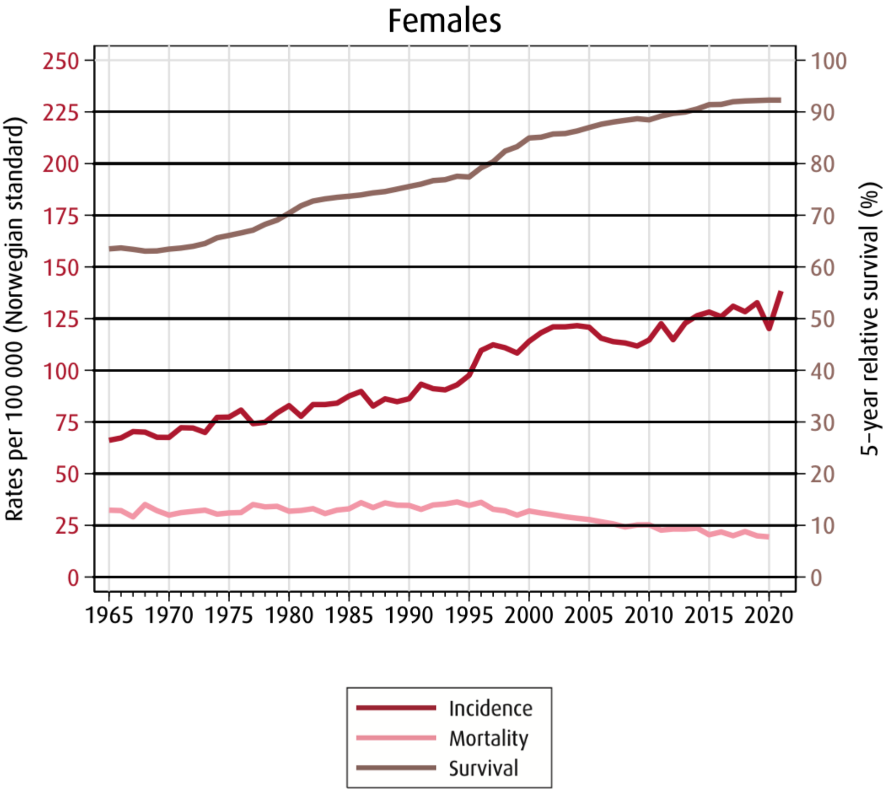


Figure 2. Illustrates incidence and mortality rates and 5-year relative survival proportions in females with breast cancer in Norway [13].

1.1.1 Breast cancer initiation and progression

Initiation of breast cancer is caused by genetic and epigenetic changes in a single cell. The progression of breast cancer involves further genetic changes with clonal expansion and changes in the tumor microenvironment (TME) [16]. Breast cancer progression is considered a developmental process that originates in the TDLUs and continues through stages of increasing proliferation, atypical hyperplasia and carcinoma in situ and ends in invasive breast cancer [17]. The full specter of mechanisms and processes that results in breast cancer is still unknown but there are several hypothesized models, such as the linear evolution model and the diversity evolution model. The linear model suggests that one dominant tumor cell which has growth and survival superiority due to environmental selection gives rise to the cells that comprise the tumor mass. The diversity model suggests that multiple predominant clones can exist within the same tumor from the start [18]. These two models are illustrated in Figure 3. The most common model suggests that invasive breast cancer originates in the TDLU, followed by progression through stages of benign breast disease (BBD). This process upregulates cellular abnormalities and proliferation and can result in atypical hyperplasia [17], manifesting as either atypical ductal hyperplasia (ADH) or atypical lobular hyperplasia (ALH). ADH and ALH is considered a precancerous condition and describes when abnormal cells accumulate in the ducts or lobules in the breast. The condition is known to increase the risk of developing ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) through stages where the epithelium becomes increasingly proliferative [19].

Myoepithelial cells are recognized as a natural tumor suppressor in the breast. They function as gatekeepers of tumor formation in the breast, and they produce the basement membrane (BM) that is a physical barrier around luminal epithelial cells. Studies show that the loss of myoepithelial cells promotes the transition of DCIS to invasive ductal carcinoma (IDC). It has been proposed that a combination of genetic changes in the tumor epithelial cells enables them to invade tissue adjacent to the ducts, and abnormalities in the microenvironment that leads to disruption of the BM and invasion of tissue adjacent to the ducts [17].

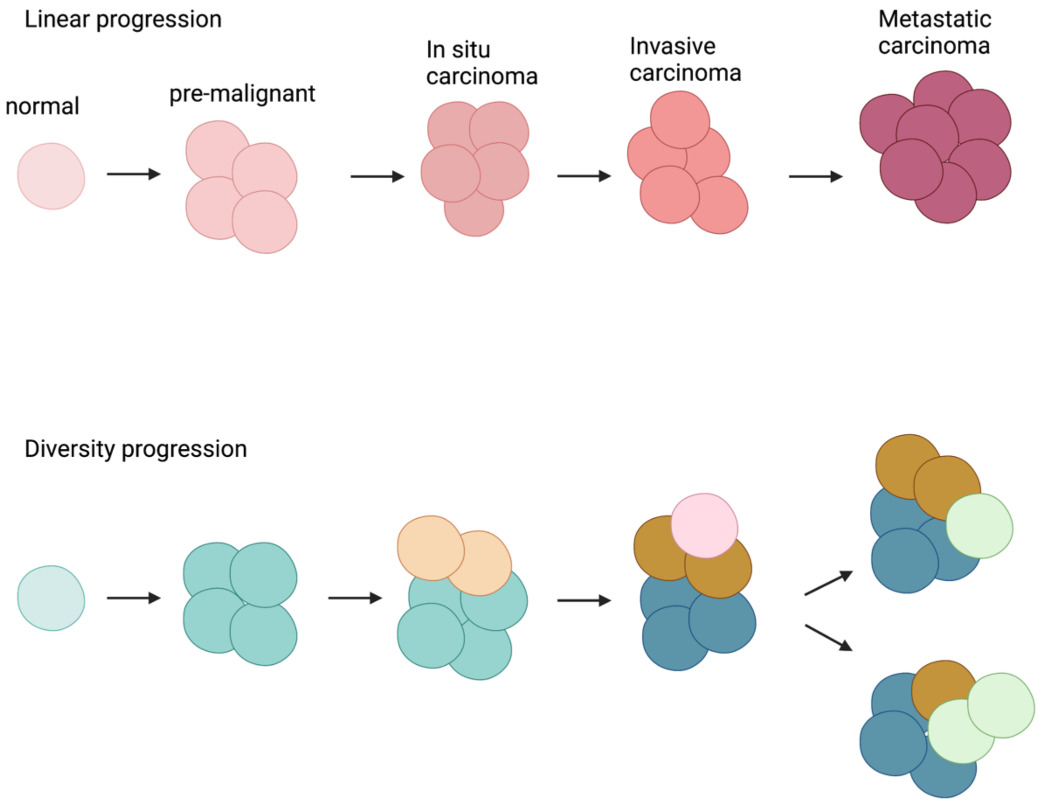


Figure 3. Illustration of the linear and diversity model of breast cancer progression [5] .

1.2 Classification of breast cancer

Breast cancer is a complex disease and show high levels of both intra- and inter-tumoral heterogeneity. To give the optimal diagnosis, treatment options and prognosis, breast cancer is classified based on histomorphology, proteomics, genomic characteristics, and clinical data. The heterogeneity of the subtypes describes different clinical behaviors and biological functions and can occur as differences in biomarker expression, tumor clonal populations and patient specific clinical variables [20]. Breast cancer are classified by histological type, grade, stage (TNM), and biomarkers such as hormone receptor (HR) status. Figure 4 shows how the different classification types affects treatment and prognosis of the patient.

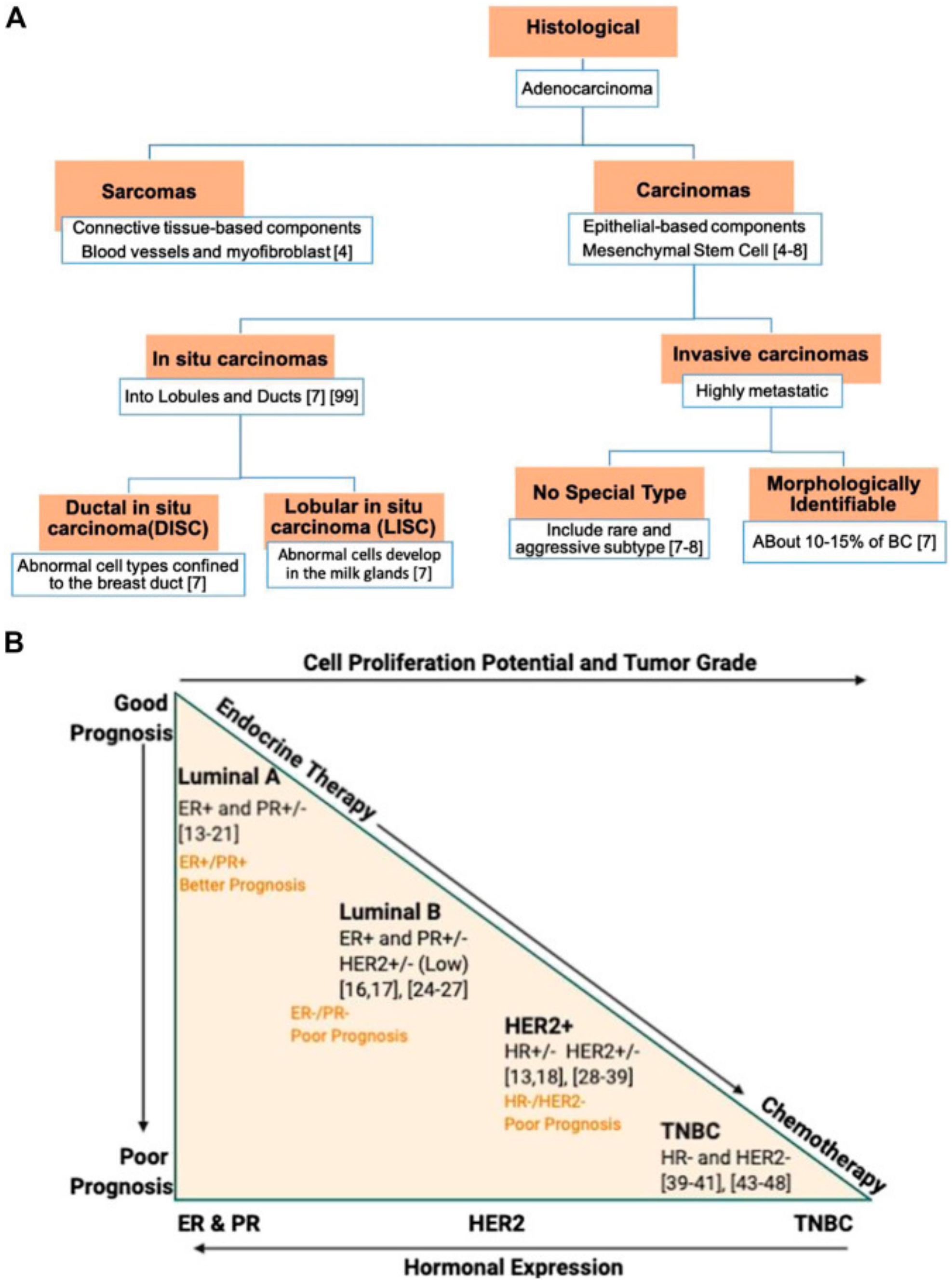


Figure 4. Scheme of breast cancer classification and subtypes: (A) shows the histological stratification. (B) shows the molecular stratification, grading, optimal therapy, and prognosis. Hormone expression shows an inverse proportion to tumor grade and cellular proliferation. Triple negative breast cancer (TNBC) shows no hormonal expression, has higher stage and high cell proliferation and tumor grade, poor prognosis and should be treated with chemotherapy. While Luminal A subtype shows expression of both ER and PR, low cell proliferation and tumor grade, good prognosis and can be treated with endocrine therapy [21] .

1.2.1 Histological classification

Histological classification is essential in breast cancer diagnosis and management and influences treatment decisions and patient outcomes. The histological type refers to the growth pattern of the cancer and the World Health Organization (WHO) defines 21 histological types of breast cancer. WHO divides breast cancer into invasive and non-invasive types. The non-invasive types of breast cancer are DCIS and LCIS. DCIS is characterized by the proliferation of malignant cells within the ducts of the breast with no invasion into the adjacent tissue, while LCIS originates in the lobules of the breast. Invasive breast cancer can be classified into subtypes based on the histological features of the tumor cells. The main subtypes are IDC and invasive lobular carcinoma (ILC). The difference between non-invasive and invasive carcinomas are that non-invasive carcinomas has not invaded surrounding tissue, and invasive carcinomas have spread and invaded adjacent tissue [22].

1.2.2 Grade

IDC are divided into tumor grades that are based on the evaluation of three morphological features mitotic count, tubule or gland formation and nuclear pleomorphism [20]. Histological grading ranges from 1-3 where a higher number indicates faster growing cancers. This provides a simple, cost-effective, and highly accurate method for assessing tumor biological characteristics and patient prognosis. Grade 1 tumors are well-differentiated with high tubule formation and homology to normal breast TDLU, but they have low mitotic count and low degree of nuclear pleomorphism. Grade 2 tumors are moderately differentiated, and grade 3 tumors are poorly differentiated with high degree of cellular pleomorphism and mitosis, and no tubule formation [23] .

1.2.3 Stage

Staging describes how extensive and proliferative the cancer is. The TNM system is the most used cancer staging system, where the T stands for tumor size, N designates if the cancer has spread to adjacent lymph nodes, and the M shows the presence of metastasis. The staging ranges from I-IV where lower number indicates less spread of cancer, and letters where an early letter indicates lower stage. The TNM system uses other key factors to describe the stage of the cancer such as estrogen receptor (ER) status, progesterone receptor (PR) status, human epidermal growth factor 2 (HER2) status and lastly the grade of the cancer [24].

1.2.4 Biomarkers

Biomarkers are measurable molecules with importance in the diagnostic, prognosis, and treatment of the disease. Three biomarkers have shown to be essential in predicting response to specific therapies and in providing prognostic information in breast cancer. These include ER, PR and HER2 [25]. Breast cancers that are HR positive (HR+) can be treated with hormone therapy and is a subtype related to better prognosis. Breast cancers that are hormone HR negative (HR-) do not benefit from hormone therapy as a rule and are often more proliferative and have a poorer prognosis than HR+ breast cancers. Other biomarkers such as the Ki67 gene, which is a cell-proliferation gene has become an important biomarker [26]. There are other less used biomarkers such as the genomic biomarkers of familiar breast cancer, BRCA1 and BRCA2 [27].

1.3 Breast cancer subtypes

Immunohistochemical staining of Ki67, ER and PR status has historically been the main method to classify breast tumors biology. Ki67 expression has been shown to inversely correlate with patient outcome [26]. ER and PR are transcriptional regulators, and their status can predict the patients sensitivity to endocrine therapy. ER and PR positive tumors often have favorable prognosis, while ER negative tumors have a poor prognosis. Growth factor receptors and ligands can also affect cell proliferation and is a key regulator in oncogenesis. One of these is the HER2 gene influencing breast cancer tumor growth. Overexpression of HER2 usually results in aggressive tumors, poor prognosis, and limited response to chemotherapy [20] .

Breast cancer molecular subtypes have been defined based on gene expression profiling. Perou et al. 2000 [28] used DNA microarrays representing 8102 human genes to characterize 65 breast tumor specimens from 42 individuals. They discovered that tumors can be classified into subtypes based on their gene expression profiling (GEP). Four intrinsic molecular subtypes: luminal A, luminal B, *v-erb-b2* (ERBB2)/HER2 gene over-expressing (HER2^{+}), and basal like were proposed by the authors as a classification scheme for breast cancer [29]. These subtypes have shown to have significant differences in incidence, risk factors, prognosis, and treatment. Methylation analyses show that breast cancers can be classified based on DNA methylation status. Tumors can be divided into three clones associated with survival, molecular subtype, ER expression and *TP53* mutation status. The methylation status of 800 cancer related genes were investigated in a study by Rønneberg et al. 2010 [30] and showed that luminal A tumors were evenly distributed between two methylation derived clones that shows significant different prognosis for the individuals [31].

Luminal cancer types comprise about 70% of invasive breast cancer incidences. Luminal A tumors are ER positive and/or PR positive and HER2 negative. This subtype usually has low-grade tumors with good outcome. Luminal A constitutes over 50% of new breast cancer diagnoses and is therefore the most common type of breast cancer [32]. Luminal B tumors are ER+ and/or PR+ and HER2+. This subtype has shown to have lower expression of HRs and higher expression of proliferation markers. Luminal B usually also comprises of higher-grade tumors and worse prognosis than luminal A cancers [33]. HER2-type tumors are ER-, PR- and HER2+. This subtype has high-grade tumors and poor prognosis – a highly proliferative cancer with germline mutations in BRCA1 and BRCA2 in 10-15% of the patients. The last group, basal-like or triple negative breast cancer (TNBC) tumors are ER-, PR- and HER2-. This is the most aggressive type of cancer with high risk of metastasis. TNBC is often a high grade cancer associated with poor prognosis, and 15-20% of incidences are related to germline mutations of BRCA1 or BRCA2 [34].

1.4 Breast cancer heterogeneity

Breast cancer is a disease that displays high degrees of both inter- and intra-tumoral heterogeneity. Intra-tumor heterogeneity indicates that there are several subpopulations of cancer cells, genetically and phenotypically different, that coexist within the same tumor [35]. The main theories that describe the origin and maintenance of tumor heterogeneity are the cancer stem cell (CSC) hypothesis and the clonal evolution/selection model [36]. Both theories consider that cancer tumors originate from single cells that have molecular alterations, developed proliferative potential, and assumes that microenvironment have an impact on the composition of cancer. The CSC hypothesis suggests that only a small fraction of cells can drive the tumor progression and that these cells are naturally therapy resistant. While the clonal evolution model suggests that progression and resistance to therapy follow the Darwinian evolutionary rules. This means that clones that can progress or develop resistance are dependent on mutation rate, population size and proliferation rate [35]. Figure 5 shows how a heterogeneous primary tumor can respond and develop during treatment.

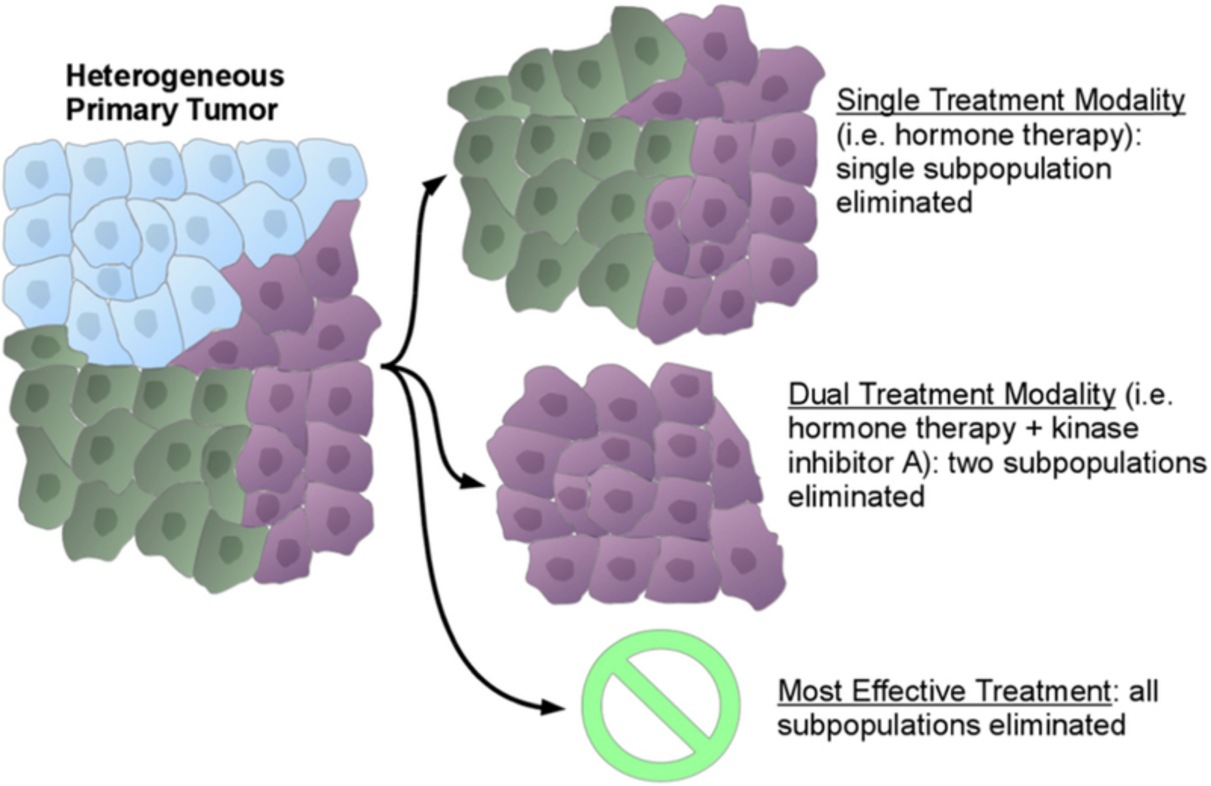

Heterogeneous Primary Tumor
Single Treatment Modality (i.e. hormone therapy): single subpopulation eliminated
Dual Treatment Modality (i.e. hormone therapy + kinase inhibitor A): two subpopulations eliminated
Most Effective Treatment: all subpopulations eliminated

Figure 5. Illustration of intra-tumor heterogeneity in breast cancer, and how subclones may survive during different treatments. In order to eliminate all the cancer cells in heterogenic tumors, one must be able to identify all subclones and tailor therapy to that specific tumor [20] .

1.5 Treatment

Treatment decisions for breast cancer patients is based on pathological subtype, molecular subtype, stage, and histological grade of the tumor. It often involves a combination of chemotherapy, surgery, radiotherapy, and endocrine therapy. However, these therapies will not effectively treat all breast cancer subtypes. Therefore personalized treatment plans and therapies are essential in breast cancer treatment. Neoadjuvant therapy is systemic treatment given before surgery to shrink the cancer.

Adjuvant therapy is administered after surgical intervention or radiation. Adjuvant therapy main aim is to control remaining cancer cells after surgery, reduce recurrence rates and improve long term survival [37]. In breast cancer patients with ER+ and HER2- tumors, one-half of recurrences occur more than 5 years after primary diagnosis [38]. In this chapter therapies most suitable for the patient group studied in this thesis will be described.

1.5.1 Surgery

The goal of breast cancer surgery is to remove as much of the cancer as possible, assess if the cancer has spread to adjacent lymph nodes and to relieve the patients from symptoms of advanced cancer. Mastectomy and lumpectomy are the two main types of breast cancer surgery. Mastectomy is a total excision of the breast, while lumpectomy is a breast-conserving approach where only the primary cancer is removed. Axillary lymph node dissection (ALND) can help determine the cancers potential to spread as metastasis or to the lymph nodes. Knowing this is of great importance for future treatment of the patient [39].

1.5.2 Radiotherapy

Radiotherapy is when high-energy radiance is applied to the whole breast or parts of the breast. Administered neoadjuvantly the goal is often to shrink tumors, make inoperable tumor operable or reduce the need for mastectomy. Patients with breast cancer subtypes that have a high ROR may be treated with radiotherapy adjuvantly. This type of therapy is often used as an attempt to remove any residual cancer cells. Radiotherapy can be given parallel to personalized therapies such as endocrine therapy or antiHER2 therapy [39]. Radiation therapy is used as treatment for all subtypes of breast cancer but is more important when treating TNBC, because of the lack of personalized treatment options.

One of the significant side effects from radiotherapy is cardiotoxicity, which makes it important to minimize radiation exposure to both the heart and lungs [40]. Some techniques that can be used to minimize radiation exposure are breath-holding techniques, optimization of beam angles, partial breast irradiation, intensity-modulated radiotherapy, and usage of multileaf collimator shielding [41].

1.5.3 Chemotherapy

Chemotherapy is given either neoadjuvantly to help shrink the tumor before surgery, or adjuvantly to patients with high ROR or lymph node metastases to remove the remaining cancer cells [37]. Chemotherapy is administration of one or more cytotoxic drugs, such as alkylating agents. These drugs disrupt the cell cycle by binding to the microtubules and disrupts their disassembly resulting in apoptosis. The therapy can also cause breakage of DNA strands and DNA intercalation and disrupt macromolecular biosynthesis.

ER+ breast cancer patients are recommended to receive chemotherapy in combination with endocrine treatment, due to low response when treated with chemotherapy alone. One downside to chemotherapy are the side effects [42]. The common side effects which usually occur 0-6 months into treatment are fatigue, hair loss, cytopenia, neurocognitive dysfunction, muscle-pain, and chemo-induced peripheral neuropathy. After 6 months of treatment chronic or late side effects may occur, and include cardiomyopathy, early menopause, sterility, second cancers and psychosocial impacts [39].

1.5.4 Endocrine therapy

Neoadjuvant endocrine therapy offers a good treatment option to shrink large breast tumors prior to surgery and is the main strategy to treat HR+ invasive breast cancers. Primary endocrine therapy has shown to be as effective as standard neoadjuvant chemotherapy. The aim of endocrine therapy is to stop hormones from fueling the cancer and the treatment targets the ERs directly or the synthesis of estrogen [43]. The three most common types of endocrine therapy are selective estrogen receptor modulators (SERMs), selective modulators estrogen receptor degraders (SERDs), and aromatase inhibitors (AIs).

Aromatase is found in estrogen-sensitive tissues such as the breast and uterus. Expression of aromatase is increased in breast tumors and is related to high estrogen levels. This means that high expression of aromatase will promote ER+ breast cancer

proliferation. AIs inhibits synthesis of estrogen by blocking aromatase enzyme activity. There are two classes of AIs: steroidal AIs and non-steroidal AIs. The preferable aromatase inhibitors for neoadjuvant endocrine therapy in ER+, postmenopausal breast cancer patients are letrozole, anastrozole and exemestane. Letrozole is a non-steroidal AI, while exemestane is a steroidal aromatase-inactivator. Exemestane irreversible binds to aromatase substrate-binding site, while letrozole bind non-covalently to aromatase substrate-binding site and prevents the binding of androgens to the enzyme.
Neoadjuvant endocrine therapy is recognized as one of the best model systems to study treatment response and to study the endocrinology of breast cancer. This treatment is administered orally and only to postmenopausal women [39].

1.6 Single cell analyses

Analysis of the transcriptome has historically been done using bulk RNA-seq, where RNA from all cells in a tissue sample is obtained and an average of the gene expression all the cells is obtained. Single-cell RNA sequencing (scRNA-seq) enables analysis of the transcriptome of a single cell from a tissue sample. This technology can be used to detect subpopulations and their genetic and functional heterogeneity [44].
Standard DNA sequencing, often called "bulk" DNA sequencing homogenizes the DNA content of all the cells in the sample. Genomic signals such as variants, DNA modifications or structural properties of the DNA from one cell or a small clone can easily go undetected using this method [45]. Single-cell DNA sequencing (scDNA-seq) encompasses technologies and approaches that makes it possible to analyze DNA down to each single cell.

scDNA-seq has three core capabilities: fidelity, co-presence, and phenotypic association [45] . Fidelity describes the ability to detect DNA features such as mutations, modifications or other properties that are only present in a small set of cells in the sample. This can also be achieved using bulk sequencing, but with this method mosaic features > 0.5% cannot be detected or be distinguished from sequencing errors [46]. Genetic mosaic features indicates that the sample is composed of more than one genotype as a result of genetic mutations. Mosaicism can be derived into two subgroups: somatic mosaicism and germ-line mosaicism. The key difference between the groups is that germ-line mosaicism is genetically transmissible while somatic mosaicism is not [47]. scDNA-seq is not limited by sequencing error due to it being much lower than the expected signal level of heterozygous DNA. Co-presence describes scDNA-seq capability to ascertain which mosaic DNA features are present in the same cells. This ability is lost in bulk sequencing because the sample is homogenized prior to sequencing. Phenotypic association is the last core capability of scDNA-seq and is the potential to combine it with single-cell phenotyping to identify which cell type and cell state the specific DNA features are present in the sample. These abilities together make it possible to distinguish between tumor and normal cells in tumor samples.

In cancer research this method has many applications such as intra-tumoral heterogeneity, metastasis and invasion, clonal evolution, circulating tumor cell and therapeutic response. As breast cancers are highly heterogenous they contain unique combinations of genetic changes and their intra-tumor heterogeneity can affect how the cancer responds to treatment [45]. scDNA-seq has been used to distinguish subclonal lineage in breast tumors using copy number aberrations (CNAs) [48]. Studies has shown that most tumors contain several important subclonal lineages, and in some cases these subclones has been associated with cancer subtypes. For example, there is a more diverse subclonal environment in ER- breast cancers than in ER+ breast cancers [49].

While whole genome scDNA-seq has greatly contributed to cancer research already, new methods such as targeted multi-omics scDNA-seq has been developed for deeper analysis on how specific subclonal genotypes associate with treatment response, phenotypes, etc [50] . The Mission Bio Tapestri technology is a microdroplet-based approach for targeted sequencing. It allows for high throughput meaning up to 10 000 cells per sample and has high coverage depth of genomic sites of interest. These capabilities make this a suited method for high resolution studies of important genetic variants within diseases [51]. The Tapestri platform enables analysis of targeted panels for single nucleotide variant (SNV) sequencing, combined SNV and copy number variant (CNV) sequencing, or combined DNA and protein sequencing. Mission Bio provides predesigned panels for several cancers and custom panels are developed using the Tapestri Designer Software. This platform has many applications in analysis of solid tumors, genome editing, biomarker development, and cell and gene therapy. A pipeline for streamlined quality analysis (Tapestri pipeline) and scDNA sequence analysis (Tapestri Insights and Mosaic Jupyter Notebook) is available from Mission Bio.

2. Aim of the study

The main aim of this study is to detect low abundance genetic variations in single cells from breast cancer tumors, and with this information explore the potential mechanisms of adaptation and resistance to the two endocrine treatments letrozole and exemestane. Significant sub-aims are to develop a reliable method to isolate nuclei from fresh frozen tumor samples and implement the Mission Bio Tapestri platform with a scDNA-seq strategy in the R&D section of the Medical Genetics Department thus making it a nationally available analysis platform at the Norwegian Sequencing Center.

3. Materials and method

3.1. Patient information

The NeoLetExe cohort (REK 2015/84) consists of postmenopausal women with locally advanced (T3-T4 and/or N2-3 primary breast cancer) ER+ breast cancer (ER+ in ≥ 10% of cancer cells) neoadjuvantly treated with letrozole and exemestane subsequently.

Patients were randomized to start neoadjuvant therapy with either letrozole or exemestane for 2 months. After 2 months all patients were crossed-over to the alternative therapy for a new 2-month period as shown in Figure 6. Patients received an established dose of 2.5 mg of letrozole daily, and exemestane was given as a 25 mg daily dose. Fresh frozen tumor tissue samples were obtained at three time points: baseline (pre-treatment), after 2 months (treatment cross-over) and after 4 months (at surgery).

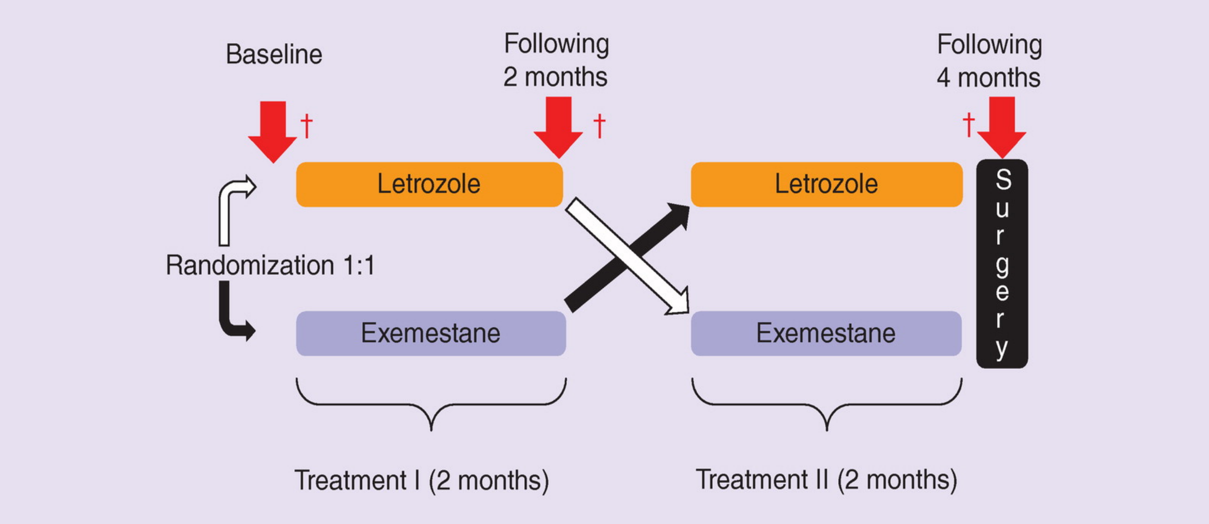


Figure 6. Patients are randomized at the beginning of the trial and is then treated with either letrozole or exemestane for the first treatment period. The patients are then crossed over and is treated with the other AI for the second treatment period, and lastly the patient had surgery to remove the cancer. Biopsies was taken at diagnosis and after 2 and 4 months of treatment.

The samples that are used in this thesis are patient 1, 2 and 3 (baseline), 3 time-points from patient 4 and 2 time-points (baseline and 4 months) from patient 5 (appendix 4). The median age was 81 years (73-84) at diagnosis and all patients had partial response to the treatment. Patient 1 and 3 received letrozole prior to exemestane, patient 2, 4 and 5 received exemestane prior to letrozole. The patients did not have metastases at time of diagnosis and all patients underwent mastectomy and removal of sentinel node.

3.2. Tissue preparation and nuclei extraction

Traditionally, manual methods have been used for extraction of nuclei from fresh frozen tissue. Recent developments have led to automation of the process. Here, two manual techniques, enzymatic dissociation, and mechanical dissociation, as well as the automated Singulator 100 (S2 Genomics) was explored.

3.2.1. Manual isolation of nuclei from breast tissue

3.2.1.1. Enzymatic dissociation

In enzymatic dissociation trypsin, collagenase and dispase are often used to digest tissue and release target cells or nuclei. Which enzymes and concentrations used are depending on tissue type, and when optimized this process can be very efficient especially for more compact and fibrous connective tissue types that may have high quantity of debris [52]. Breast tissue have a high content of adipose cells and the nuclei extraction protocol must be carefully adapted to meet the challenges this creates.

Manual extraction of nuclei from the biopsies was done following the Enzymatic Dissociation Protocol from the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing user guide (Nuclei Extraction From Frozen Tissue For Single-Nuclei DNA Sequencing User Guide). The tissue biopsy (~3x3x3mm) was minced into tiny fragments and then further dissociated in a tissue lysis buffer. The lysis buffer is a mix of trypsin, collagenase and dispase dissociating the tissue and releasing the nuclei from the cells. At the end of the incubation time the enzymatic reaction is stopped by adding a solution containing a trypsin inhibitor and RNase A to help minimize clumping of the nuclei in the suspension. Increased concentrations (2x and 4x) (appendix 1) of enzymes were applied in the lysis mix and extended incubation time (10-30 minutes) for better dissociation. This was followed by several clean-up steps to discard any tissue and cell debris in the sample and included several steps of straining through pre-wetted 50 and 30 µm cell strainers, and centrifugation steps at 500g for 5 minutes at 4 °C. After the last centrifugation the nuclei pellet was resuspended in 50 µL of Mission Bio Cell Buffer. Table 1 shows the nuclei concentration from test tissue and sample tissue using this method.

Table 1. Nuclei concentration in samples from NeoLetExe trial and test tissue, measured on Countess and NucleoCounter® NC-100

Sample	**Measured nuclei concentration (ng/µL)**	**Measuring method**
Test tissue (2x enzyme concentration)	280	NucleoCounter® NC-100
Test tissue (4x enzyme concentration)	360	NucleoCounter® NC-100
Patient-1_baseline	16 100	Countess II FL
Patient-2_baseline	10 800	Countess II FL
Patient-3_baseline	31 900	Countess II FL

3.2.1.2. Mechanical dissociation

Mechanical dissociation is often used for tissues that have strong adhesions that needs to be broken down to extract single cells or nuclei. In mechanical dissociation tissue samples are digested using physical force, such as cutting and crushing. Instruments like pestles or mortars will destroy the tissue into smaller digestible fragments. To maximize the yield using mechanical dissociation the tissue should not be treated too harshly, and over digest the sample so that the nuclei are also digested. Mechanical dissociation is often used in experiments using droplet-based methods [52].

Manual extraction of nuclei from the biopsies was done following the Mechanical Dissociation Protocol from the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing user guide (Nuclei Extraction From Frozen Tissue For Single-Nuclei DNA Sequencing User Guide). Mission Bio has a customer-developed protocol by Martello that uses a commercially available lysis buffer and mechanical dissociation using douncers [52]. The tissue biopsy was minced into smaller fragments, and further dissociated using a douncer. Pestle A was used until all resistance was gone (~30 strokes), this was followed by ~20 strokes with pestle B. The suspension was then incubated on ice for 5 min in chilled Nuclei EZ Lysis buffer to further dissociate all the cells and extract intact nuclei. This was followed by clean-up of cell and tissue debris still present in the sample, which included several steps of straining through pre-wetted 50 and 30 µm cell strainers, and centrifugation steps at 500g for 5 minutes at 4 ^{o}C. After the last centrifugation the nuclei pellet was resuspended in 50 µL of Mission Bio Cell Buffer. A concentration of 410 nuclei/µL was achieved using this method on the test tissue.

3.2.2. Singulator 100 and Percoll clean-up

The Singulator 100 is an automated bench-top system that enables high reproducibility, high yield, and rapid dissociation of solid tissue into single-cell or nuclei suspensions. It can perform cold dissociation, which minimizes expression of stress-related genes in cells and preserves RNA quality in nuclei. With this instrument nuclei can be extracted in 5-6 minutes and single cells can be extracted in 20-60 minutes. S2 Genomics has developed pre-set protocols and pre-formulated reagents for cell and nuclei isolations.

Automated extraction of nuclei was done following the Nuclei Isolation Protocol from Singulator^{TM} 100 Automated Tissue Dissociation System Guide [53]. In this experiment the Preconfigured Standard Nuclei isolation v2 was used. After the instrument was turned on the instrument started cooling down until it reached 4 ^{o}C. The tissue biopsy with mass of <20-100 mg (Table 2) was placed in a pre-chilled cartridge. The cartridge was then loaded into the instrument and the run was started. The tissue is automatically disrupted in S2 Genomics Nuclei Isolation Reagent (NIR), diluted with S2 Genomics Nuclei Storage Reagent (NSR) and strained in a closed system within the cartridge. Incubation of the disrupted tissue is not included in this isolation protocol to avoid over digestion of the nuclei. When the run was completed, the cartridge was removed from the instrument. The sample was then pipetted from the cartridge into a 5 mL LoBind tube, centrifuged at 500g for 5 minutes at 4^{o}C and the supernatant was discarded. Further clean-up was done following the Nuclei Clean-up and Debris Removal Procedure from S2 Genomics [54]. All solutions were kept on ice during the protocol. The nuclei pellet was resuspended in 1 mL of 20% Percoll (appendix 1), and an additional 2 mL of 20% Percoll was added making it a total of 3 mL. The suspension was then centrifuged at 700g for 8 minutes at 4 ^{o}C. The debris cake floating on top in the tube was carefully removed by inserting a serological pipette with two Kim wipes wrapped around it and absorbing around 1-2 mL of the supernatant. The remaining supernatant was removed using a pipette, and the nuclei pellet was resuspended in 50 µL of Mission Bio Cell Buffer. Table 2 shows the mass of tissue biopsies used in the automated extraction method.

Table 2. Amount of tissue from the biopsies that was used in nuclei extraction in Singulator 100.

Sample	Mass of tissue sample (mg)
4_baseline	240
4_2mnd	27
4_4mnd	80
5_baseline	28

5_4mnd	26

3.3. Quality control

After nuclei extraction the concentration of nuclei was measured using the Countess II FL (Countess II FL automated cell counter User Guide) or NucleoCounter® NC-100™ (NC-100). With the Countess II FL systems the isolate is mixed with Trypan blue staining the dead cells (nuclei). The channel inside the NucleoCounter® NC-100™ cassette is covered with immobilized fluorescent dye, Propidium Iodide (PI), binding to DNA in free nuclei. The nuclei suspension should have a concentration of 3000-4000 nuclei/µL for an optimal Mission Bio Tapestri experiment. The quality of the nuclei was inspected in a brightfield microscope and DAPI staining (DAPI solution). Intact and little to no clumping of the nuclei is needed for the downstream procedures.

3.4. Mission Bio Tapestri platform

The Mission Bio Tapestri platform uses microfluid droplet technology to combine cell lysate with barcoding beads attached to gene specific primers to give high-throughput single-cell genomics workflow for targeted DNA sequencing [55]. Figure 7 shows an overview of the library construction. The nuclei are individually partitioned into sub-nanoliter droplets. Barcoding beads and PCR reagents are then added using the Mission Bio Tapestri instrument and DNA cartridge. Cell lysis, protease digestion, cell barcoding and targeted amplification using multiplexed PCR occur within the droplets. The droplets are then disrupted, and the barcoded DNA is extracted for library amplification. Final libraries are purified and sequenced on an Illumina sequencer instrument.

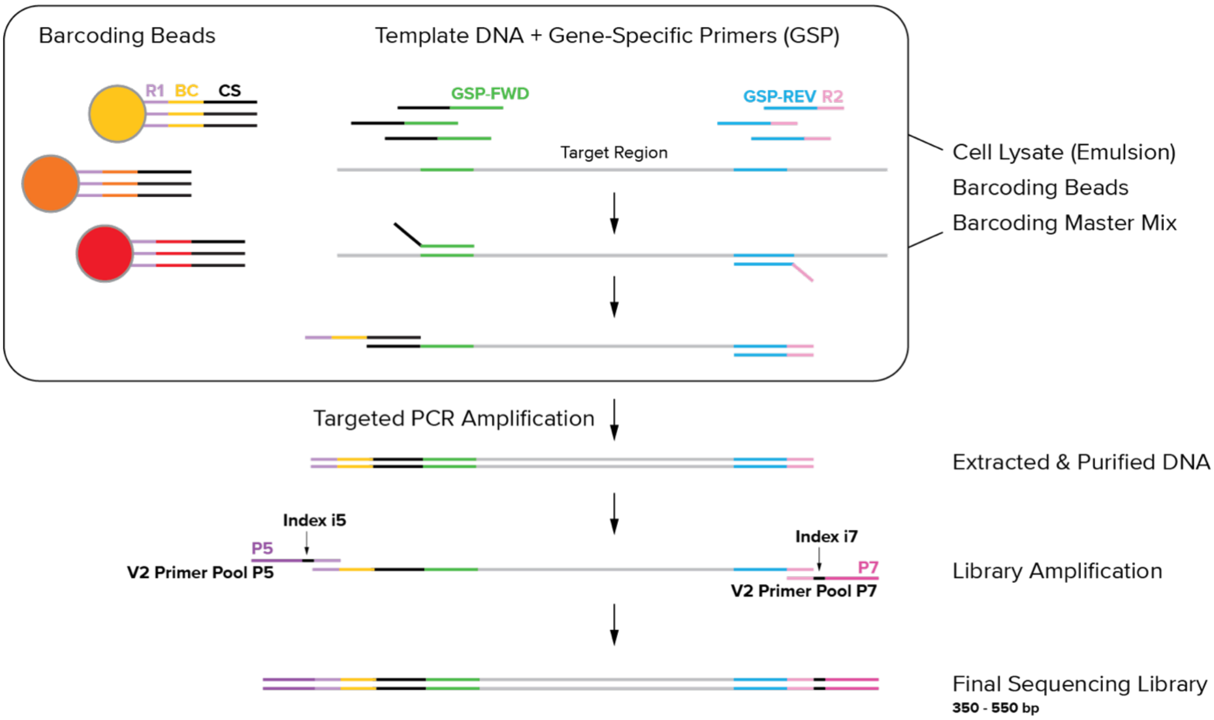


Figure 7. Illustration of library construction in the Mission Bio Tapestri instrument. R1: read 1, BC: barcode, CS: Common sequence GSP-FWD: gene-specific forward primer, GSP-REV: gene-specific reverse primer, P5: P5 Illumina adapter and P7: P7 Illumina adapter [55] .

3.4.1. Custom made targeted DNA breast cancer panel

Mission Bio provides predesigned panels as well as a software to design custom panels (Tapestri Designer). Whole exome sequencing data for the tumors in the NeoLetExe cohort is available and in silico predictions on clone developments through the treatment timepoints are done (unpublished data). Whole genome DNA-seq data was integrated with whole transcriptome RNA-seq data from the tumors to identify cellular characteristics of the subclones and intra-tumor mechanisms of adaptation to letrozole and exemestane. Possible biomarkers to detect treatment response and dissect the heterogeneity of the tumors was proposed. Amplicons covering the mutations found in the bulk sequencing of these tumors were included in the custom targeted DNA sequencing panel, those in COSMIC database prioritized, as well as known cancer mutations, like hotspot ESR1 mutations. In addition, known sequence variants from all chromosomes was hand-picked to enable copy number analysis. The Tapestri Designer presented a panel of 497 amplicons covering 528 known sequence variants (unpublished).

3.4.2. Custom single-cell DNA sequencing

The Genomic protocol in The Tapestri Single-cell DNA Sequencing V2 user guide consists of nine steps (Tapestri Single-Cell DNA Sequencing V2 User Guide). The first step is to prepare cell suspension if blood or cell lines are the material to be analyzed. For fresh frozen solid tissue, a nuclei suspension is obtained as described in chapter 3.2. The second step is to encapsulate the cells or nuclei with the custom-made reverse primer pool and Lysis buffer and create an emulsion in the Tapestri DNA microfluidics cartridge. The optimal cell/nuclei concentration of the sample is 3,000 – 4,000 cell/µL. The third step is lysis and digestion of cells or nuclei, the DNA released, and DNA binding proteins are enzymatically digested on a thermal cycler to make DNA accessible for target amplification. The fourth step is to barcode the cells/nuclei. Here the drops containing encapsulated nuclei lysate are combined with the drops containing the custom-made forward primer pool, barcoding master mix and barcoding beads. The newly constructed drops are distributed into eight PCR collection tubes to create eight cell-barcoding emulsion samples. Step five is UV treatment and targeted PCR amplification. The emulsions are treated with UV light to cleave off barcode-containing forward primers from the barcoding beads before the targeted PCR amplification of the panel amplicons. The sixth step is clean-up of the PCR products. The PCR products are firstly digested using DNA clean up buffer and clean up enzyme. This is followed by Ampure XP library clean-up to remove short fragments like primer dimers from the PCR products. The concentration is measured using a Qubit fluorometer (Qubit), and the quality (fragment size) of the targeted PCR product with an Agilent Tapestation 4200 (D5000 High sensitivity DNA ScreenTape (Agilent Tapestation)). The concentration should be 0.2-4.0 ng/µL and the amplicon fragment size ~370bp. The seventh step is the PCR Targeted Library prep where the P5 and P7 adapter (Illumina) sequences are added to the amplicons for sequencing. The eight step is to quantify and normalize the sequencing library. The concentration of the library should be 0.9-1.3 ng/µL and the fragments size ~450bp. It is also important to take into consideration the percentage of small non-specific fragment ~200bp. The acceptable percentage differ between sequencing instruments and is important to comply to avoid extensive sequencing of small non-specific fragments. Extra Ampure clean-up steps might be necessary. The last and ninth step is 2x 150bp paired-end sequencing of the library which is done using an Illumina sequencing instrument chosen based on the specifications given in the Tapestri protocol like number of pooled samples and number of amplicons in the panel. The sequencing was performed at the Norwegian Sequencing center (OUS-Ullevål) in a ½ SP flowcell (pool of sample 1, 2 and 3) or in a ¼ S4 Novaseq flowcell (pool of sample 4 (3 time-points) and sample 5 (2 time-points). The concentrations and library pool preparation details are shown in appendix 5.

3.5. Computational analysis

Mission Bio have developed an automated pipeline called Tapestri pipeline for analysis of the raw data from the targeted single-cell sequencing. This pipeline aligns the reads and maps them to the human reference genome hg19, deconvolutes cell barcodes and DNA variant calling. The Tapestri pipeline software manages and processes single-cell sequencing data from FASTQ input files. The output from this analysis is a pipeline run report that is used to quality check the sequencing results. The run report displays various metrics from the sequencing run that are used to verify the quality of the sequencing data. These metrics, their description and the desired value of the metrics are shown in Table 3. The output files from the Tapestri pipeline were used in further downstream analysis. Sequencing data from all samples were analyzed using this pipeline.

Table 3. Overview of run report metrics, with description and the desired value of each.

Run report metrics	**Description**	**Desired value**
Cells	Number of cells sequenced	3000-10 000
Panel uniformity	Number of amplicons with mean reads to the amplicon above 0.2x the mean reads per amplicon per cell	>80%
Mean reads/cell/amplicons	Mean reads per cell divided by the number of amplicons in the custom panel	35-150

%DNA read pairs assigned to cells	Percentage of read pairs present in the called cells	>35%
Read quality (QC30)	Percentage of bases with sequencing quality >30%	>80%
% reads mapped to genome	Percentage of read pairs mapped to the genome	>80%
% reads mapped to target	Percentage of read pairs mapped to the insert coordinates of the amplicons in the custom panel	>80%
ADO rate (Allele Drop-Out)	Calculated using germline variants ADO = (Cells with reference calls + Cells with homozygous calls) / Genotyped cells	<15%

Tapestri Insights is a turnkey analysis solution for analysis and visualization tools. This software can be used to review variants and subclones, filter data and construct and export visualizations like UMAP plots, XY plots, violin plots, bar plots and fish plots. The software is equipped with a set of quality filters that filter the single-cell DNA data based on genotype quality, read coverage, mutant (alternate) allele frequency and percentage of mutated cells per variant. This filter removes either cells/entire variants or individual genotypes. Loom files for all samples from the Tapestri pipeline was uploaded and filtered in Tapestri Insights. The software was then used to review variants/subclones for all samples based on set criteria. The variants that were chosen for further downstream analysis in Mosaic were evaluated based on their DANN score, %mutated cells and similarity in variant allele frequency (VAF) scores. DANN is a deep learning approach for annotation of the pathogenicity of genetic variants [56]. The variants of interest should have a DANN score close to 1 which equals high pathogenicity. DANN scores for the variants are obtained from VarSome API. The % mutated cells in the variant should not exceed 90%, because this could be a germline mutation. Insights uses two different methods to calculate the average VAF of each variant, VAF by read count and VAF by cell count. VAF by read count is comparable to conventional bulk sequencing as it considers sequencing reads across all genotyped cells, ignores cell-barcodes and calculates the fraction of mutant sequencing reads. VAF by read count considers alleles across all genotyped cells, ignores cell-barcode identity, and calculates the fraction of mutant alleles. The VAF scores should be comparable and not deviate more than 1.5-fold. If they deviate too much it may indicate copy number alterations or it can be considered an artifact [57].

Mosaic is a python package with a set of tools that can be used to analyze DNA and protein data using data from the Tapestri pipeline. This allows for convenient handling and visualization of single-cell data and exploratory analysis. In this analysis a whitelist of variants handpicked from Tapestri Insights for each sample was used. The variants were clustered by **`Dna.group_by_genotype`** algorithm that clusters cells based on provided variants and returns a data frame of per-clone and per-variant statistics. The algorithm also considers allele dropout out (ADO) to identify false positive clones. These clones are the basis for constructing fishplots and bar plots. An example of code for the mosaic analysis is shown in appendix 6.

3.6. Own contributions

Over the last 9 months I have performed all methods mentioned above from extracting nuclei to the computational analysis, excluding WGS and panel design for the targeted scDNA-seq which was part of a pilot study. I have been included in the optimalization of the nuclei extraction method and have performed the protocol of custom targeted scDNA-seq on the samples. The sequencing was performed by the Norwegian Sequencing Center (OUS-Ullevål).

4. Results

4.1. Nuclei Extraction

Manual extraction is the traditional way to extract nuclei. Both an enzymatic and a mechanical method was carried out on sample tissue to see what would give the highest yield and quality. Nuclei were extracted from the baseline biopsy of patient 1, 2 and 3, three samples (baseline, 2 months and 4 months) from patient 4 and two samples from patient 5 (baseline and 4 months). The nuclei from patients 1, 2 and 3 were extracted using the enzymatic extraction protocol strictly following the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing User guide, while the nuclei from patients 4 and 5 were extracted using the Singulator™ 100 and Percoll clean-up protocol. The concentration of nuclei in samples 1, 2 and 3 were measured with a Countess II FL and samples from patients 4 and 5 were measured using the NucleoCounter NC-100. The concentrations of extracted nuclei are shown in Figure 8.

As described in chapter 3.2.1.1 different concentrations in enzymatic lysis solution was attempted on breast tumor tissue harvested for training. Measured nuclei concentration for the test tissue samples and for patients 1, 2 and 3 are show in Table 1. The concentration of nuclei extracted from the test tissue using 2x enzyme concentration in lysis solution was 280 ng/µL, and for the test tissue sample using 4x enzyme concentration was 360 ng/µL. The concentration of extracted nuclei in the patient samples (patient 1, patient 2 and patient 3) 16 100 ng/µL, 10 800 ng/µL and 31 900 ng/µL respectively.

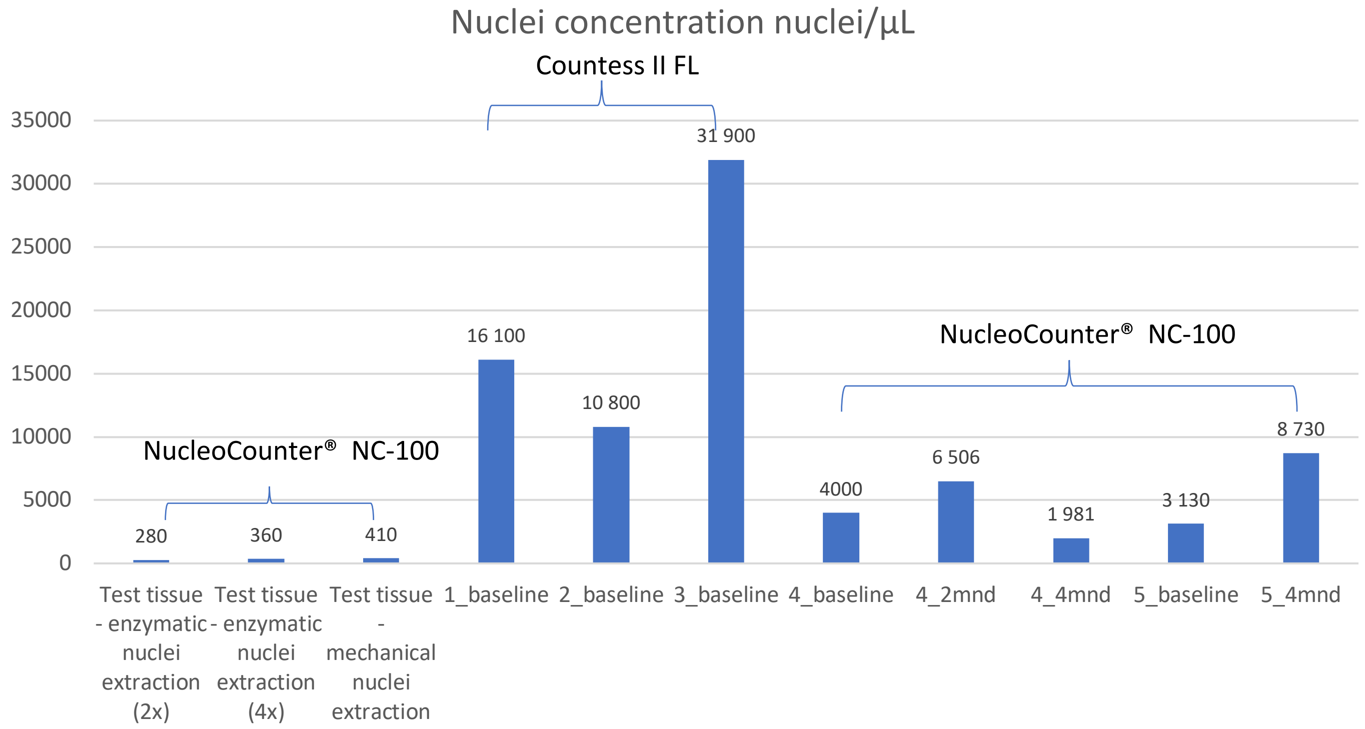


Figure 8. Diagram showing nuclei concentration in test tissue and patient samples using the three different extraction methods and which instrument that was used to measure the nuclei concentration.

The quality of the isolated nuclei was assessed by the shape and intactness of the nuclei membrane, also potential clustering of the nuclei using an inverted microscope (Nikon Eclipse Ts2). DAPI stains the nuclei for localization. Figure 9, figure 10 and figure 11 show both pictures of the nuclei in brightfield and DAPI stained nuclei in the microscope. Figure 9 shows a representative example of nuclei extracted with manual enzymatic dissociation protocol. The isolate does not have visual cell debris or clumps of nuclei; however, the nuclei density is low and unstained droplets in the brightfield picture could be droplets of fat released from adipose cells. Figure 10 show a representative example a tissue with high content of adipocytes, and figure 11 show a representative of a tumor cell dense tissue.

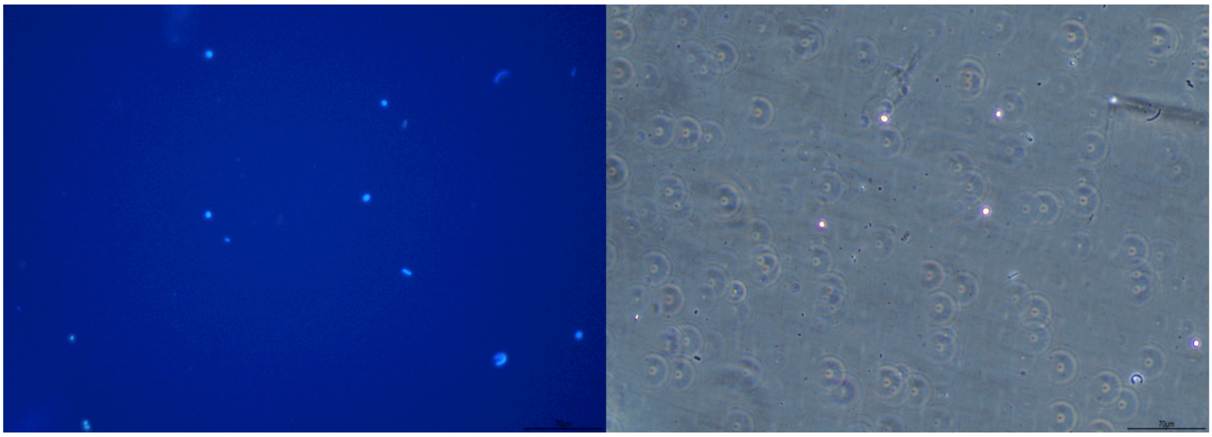


Figure 9. The left side of the figure shows manually extracted nuclei from test tissue (enzymatic) with DAPI staining. The right side of the figure shows the same are but in brightfield. The unstained dots, visual in the brightfield picture could be droplets of fat (20X magnification).

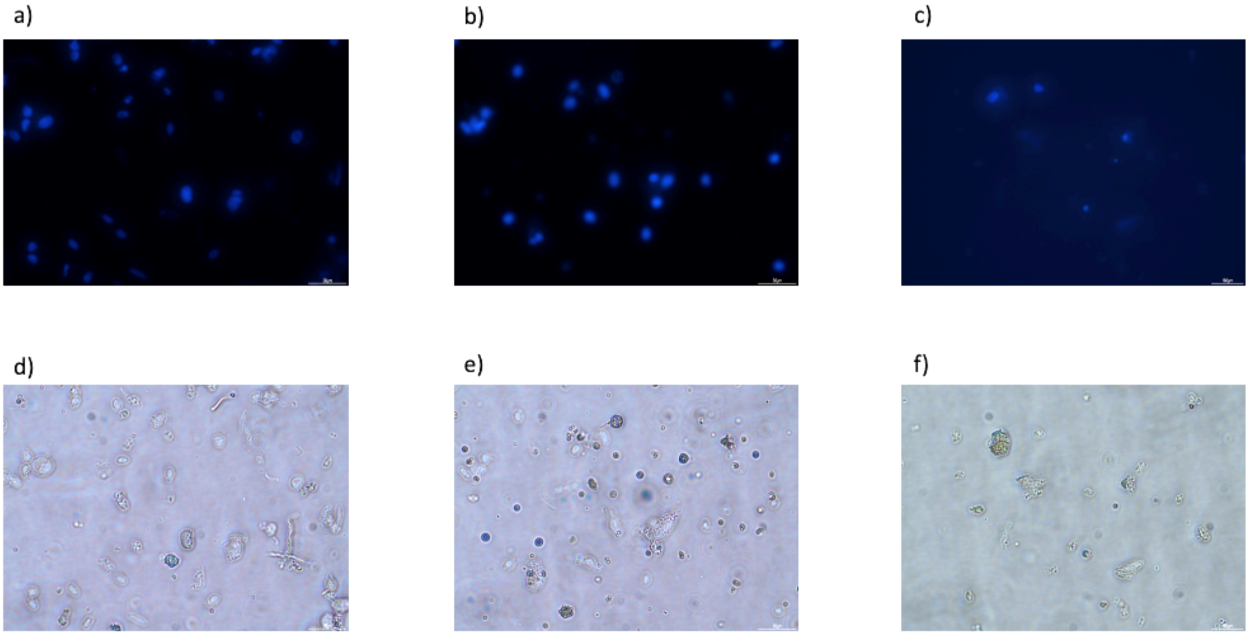


Figure 10. Microscope pictures in brightfield and with DAPI staining of extracted nuclei from patient samples using the Singulator 100™ and Percoll clean-up. a): patient 4 at baseline with DAPI staining, b): patient 4 at 2 months with DAPI staining, c): patient 4 at 4 months with DAPI staining, d): patient 4 at baseline in brightfield, e): patient 4 at 2 months in brightfield and f) patient 4 at 4 months in brightfield.

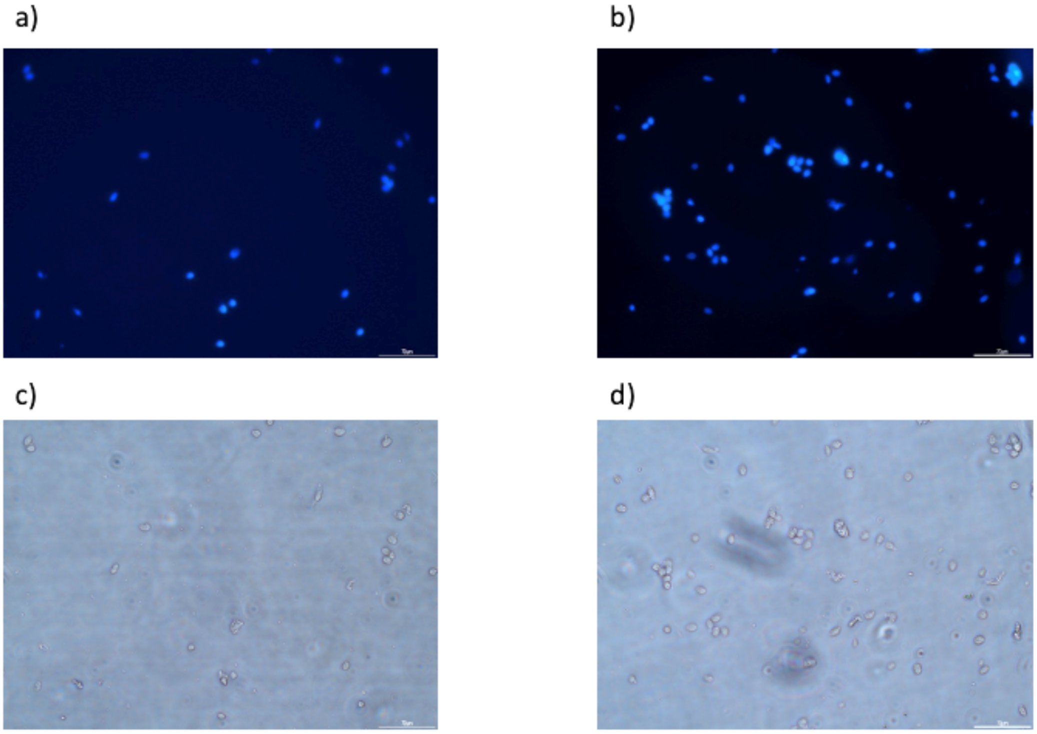


Figure 11. Microscope pictures in brightfield and with DAPI staining of extracted nuclei from patient samples using the Singulator 100™ and Percoll clean-up. a): patient 5 at baseline with DAPI staining, b): patient 5 at 4 months with DAPI staining, c): patient 5 at baseline in brightfield and d) patient 5 at 4 months in brightfield.

4.2. Targeted sequencing

4.2.1. Tapestri Pipeline – sequencing quality data

A FASTQ file with the sequencing data was uploaded to Tapestri pipeline. The reads were aligned and mapped to the human reference genome hg19. Table 4 shows the run report values from the pipeline for samples from patient 1 (one timepoint), patient 2 (one timepoint), patient 3 (one timepoint), patient 4 (three timepoints) and patient 5 (two timepoints). The quality data was compared to the desired values shown in Table 3. All samples had lower number of cells than desired, but the samples from patients 4 and 5 have higher number of cells than the samples from patients 1, 2 and 3. The panel uniformity is low in patients 1, 2 and 3, and over the desired value in patients 4 and 5. The sample from patient 3 was the only sample within the desired value for mean reads/cell/amplicons. Patient 1 and 2 has lower percentage of DNA read pairs assigned to cells than wanted, while the others are above the desired value. All samples are within desired value for read quality (QC30). Patient 2 is the only sample outside of desired value for percentage reads mapped to genome and percentage reads mapped to target. The ADO rate is higher than desired in patients 1, 2, 3 and at one timepoint in patient 4 (4 months).

Table 4. Overview of the run report of the samples in the Tapestri Pipeline with desired values in parentheses.

Sample	**Number of Cells (3 000 – 10 000)**	**Panel uniformity (>80%)**	**Mean reads/cell/amplicons (35-150)**	**% DNA read pairs assigned to cells (>35%)**	**Read quality (Q30) (>80%)**	**% reads mapped to genome (>80%)**	**% reads mapped to target (>80%)**	**ADO rate (>15%)**
patient 1_baseline	462	79.07%	176	32.65%	90.59%	83.50%	81.38%	24.30%
Patient 2_baseline	50	70.02%	352	7.52%	88.22%	72.18%	68.44%	15.20%
Patient 3_baseline	886	74.45%	132	42.29%	91.21%	85.01%	83.11%	15.35%
Patient 4_baseline	2588	86.72%	187	52.77%	89.49%	86.98%	84.20%	13.75%
Patient 4_2mnd	1123	80.89%	509	57.36%	89.38%	88.13%	84.75%	14.45%
Patient 4_4mnd	1919	80.89%	280	49.09%	89.39%	86.35%	82.05%	18.85%
Patient 5_baseline	1792	80.48%	278	54.42%	88.60%	88.28%	84.78%	14.20%
Patient 5_4mnd	2845	80.48%	166	54.27%	88.30%	88.32%	84.87%	14.00%

4.2.2. Comparison of variants in bulk-seq and scDNA-seq

The bulk sequencing and the targeted sequencing was done on tumor tissue from the same patients, but different biopsies. Of the 36 sequence variants constituting the clones found by bulk sequencing of sample 1, 17 clones were successfully analyzed by targeted sequencing, 8 out of 63 clones for sample 2 and 35 out of 142 clones for sample 3. For sample 4 and 5, 15 out of 61 clones and 17 out of 21 clones was found by targeted sequencing as shown in Table 5. Successfully sequenced variants from the targeted sequencing covered 4 out of 5 clones in sample 1, 1 out of 1 clone in sample 2, 7 out of 9 clones in sample 3, 4 out of 4 clones in sample 4 and 4 out of 5 clones in sample 5.

Table 5. number of variants found in the targeted sequencing and the bulk sequencing, and how many variants is found in each clone found in the samples.

	Bulk sequencing		**Mission Bio Tapestri targeted single-cell sequencing**
	Clone	**No. of variants**	**No. of bulk variants found**
Patient 1	1	19	10
	2	6	4
	3	2	1
	4	3	0
	5	6	2
Patient 2	1	63	8
Patient 3	1	21	8
	2	57	8
	3	30	9
	4	4	2
	5	9	5
	6	3	0
	7	9	1
	8	3	2
	9	6	0
Patient 4	1	10	4
	2	12	5
	3	17	6
	4	22	9
Patient 5	1	9	7
	2	8	7
	3	1	0
	4	1	1
	5	2	2

For the patients with only one timepoint sequenced the development of the clones during treatment cannot be tracked. Still, several variants from the WGS were successfully sequenced with the Tapestri targeted single-cell method. Variants representing each of the clones from WGS was handpicked, based on DANN score, percentage of mutated cells and similar VAF by cell count/VAF by read count quality to visualize clone formation in sample 1, 2 and 3 (Figure 12) and clone expansion through treatment for sample 4 and 5 (Figure 13 and Figure 14 respectively). The 8 variants chosen for patient 1 gave three main clones in Tapestri Insights (C1, C2 and C3), in addition to small subclones (<1%) and subclones with missing genotype in one or more of the selected variants (missing GT subclones). The missing GT subclones makes out 79% of the clones, the small subclones makes out another 14.94% of the clones, C2 and C3 makes out 1.08% and C1 3.90% of the clones. The 11 chosen variants for patient 2 gave two different clones (C1 and C2), plus small subclones and missing GT subclones. Clones C1 and C2 makes out 2.0% each of the total percentage, while small subclones is 0% and missing GT subclones makes out 96.0%. The 8 variants chosen for patient 3 only gave a wild-type (WT) clone with percentage of mutated allele between 1 and 3, in addition to small subclones and missing GT subclones. The WT clone was 1.58% of the total, small subclones 6.32% and missing GT subclones 92.10%.

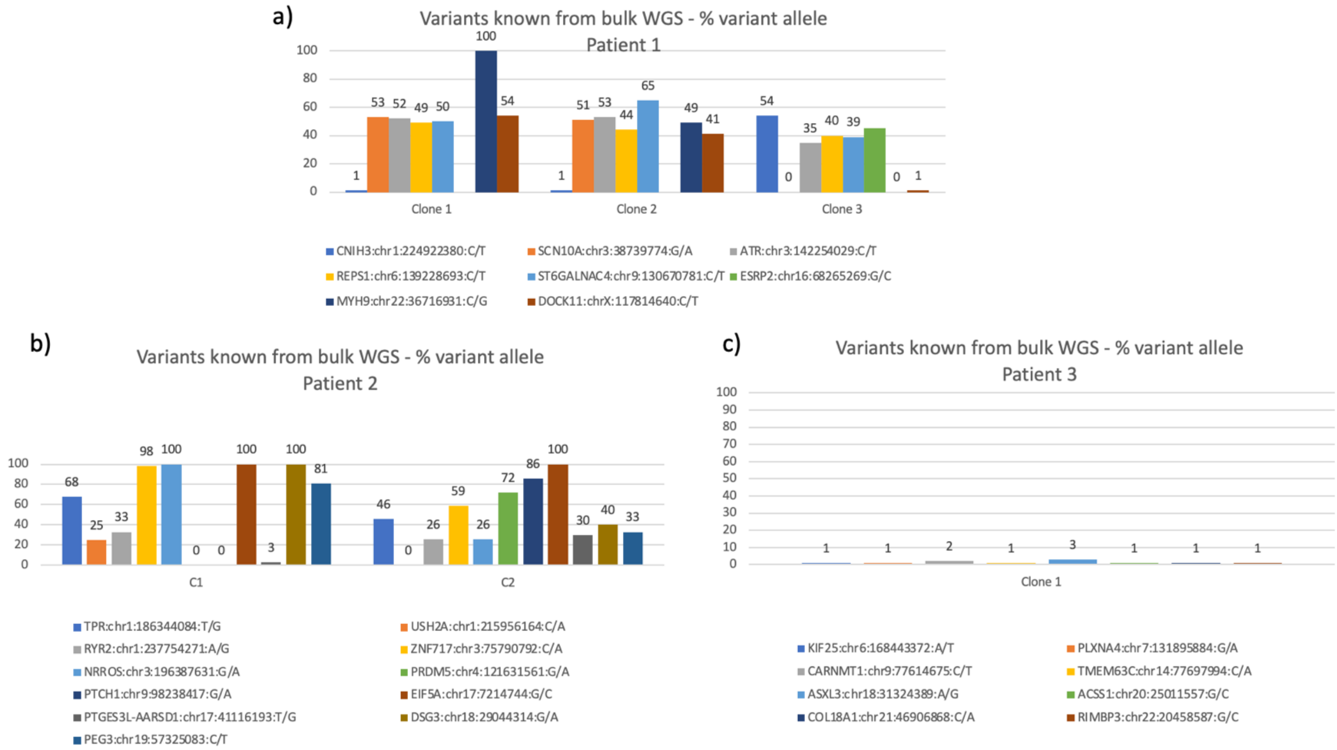


Figure 12. visualization of clone formation of variants known from bulk WGS and the % variant allele per variant chosen in each sample. a) patient 1, b) patient 2 and c) patient 3.

In time course treatment analysis, the potential changes in the composition of the cell clones in the tumor site is observed. Variants from each known clone from the WGS analysis was handpicked, based on DANN score, percentage of mutated cells and similar VAF by cell count/VAF by read count quality in Tapestri Insights, to construct the fishplots and bar plots per timepoint using Mosaic for patients 4 (3 timepoints, Figure 13) and 5 (2 timepoints, Figure 14). The variants picked for sample 4 is ACTR6:chr12:100601481:A/G, NBPF20:chr1:145438939:T/G, chr5:140856734:G/C and chr8:1581155:G/T. The variants picked for sample 5 is CAPN6:chrX:110492213:G/A, HUWE1:chrx:53575011:T/C, SYCP1:chr1:115399240:A/G and UTRN:chr6:144832224:C/T. The bar plots in Figure 13 and Figure 14 show the distribution of the cells in the clones in patients 4 and 5 in the different timepoints the tissue biopsies were obtained. The fishplot from sample 4 show that clones 2, 3 and 4 disappear between timepoint 1 (baseline) and timepoint 2 (2 months) and clone 1, which is the wild-type clone, takes over the entire plot. This is further confirmed in the bar plot where the baseline shows that the cells are distributed over several clones, whereas in the bar plots for 2 months and 4 months it only consists of clone 1 and missing GT subclones. The same trend can be observed in the plots for patient 5, but the clones do not completely disappear by the last timepoint in these samples. All the clones except clone 7 is still present in the sample in the last timepoint (4 months). Data for clones and variants to construct fishplots and bar plots for patients 4 and 5 are shown in appendix 7.

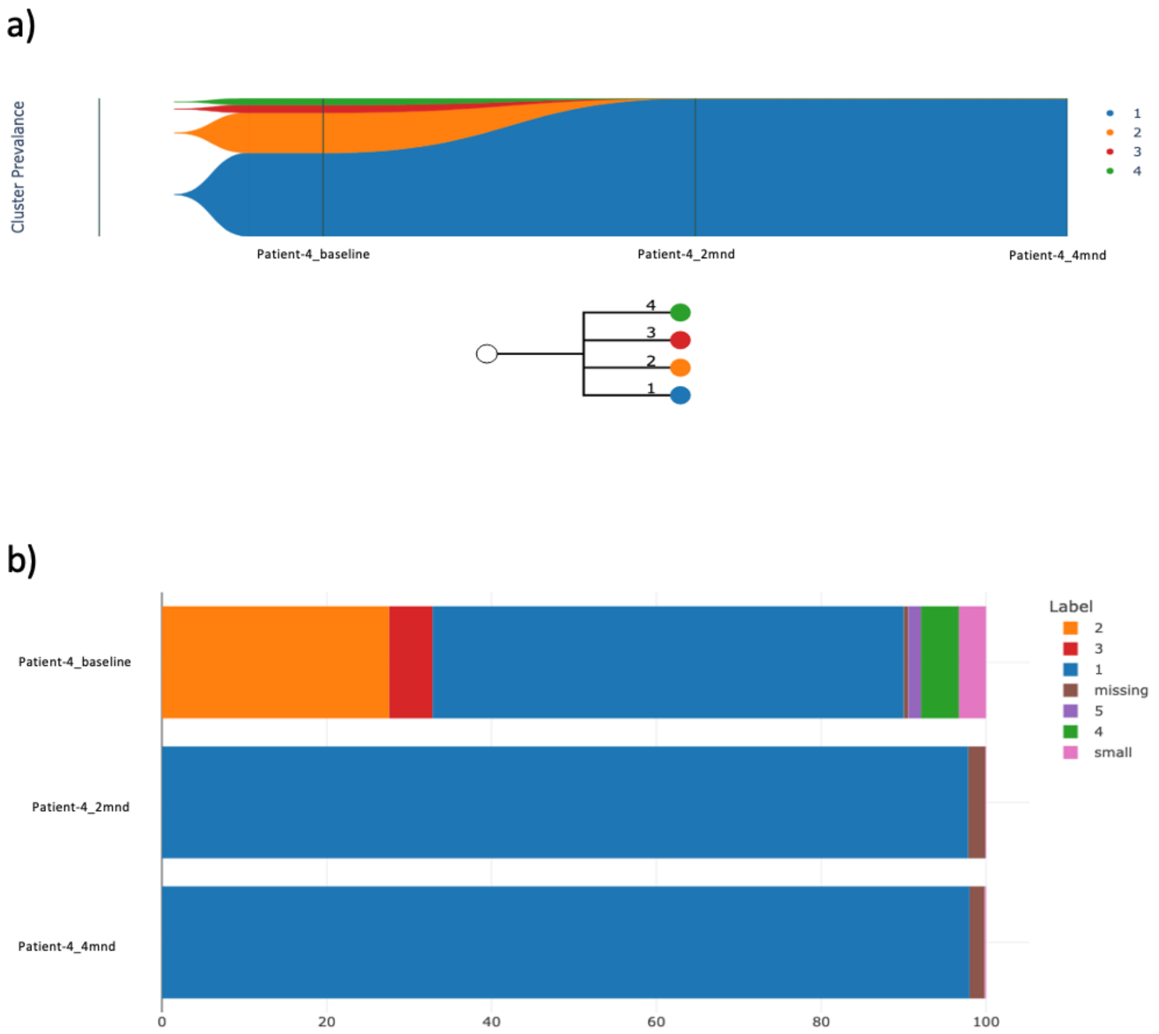


Figure 13. a) fishplot of patient 4 with one variant from each of the different clones from the WGS which was found again in the sample after targeted sequencing. b) bar plots over the distribution of subclones in the samples from patient 4 at baseline (diagnosis), 2 months and 4 months into treatment.

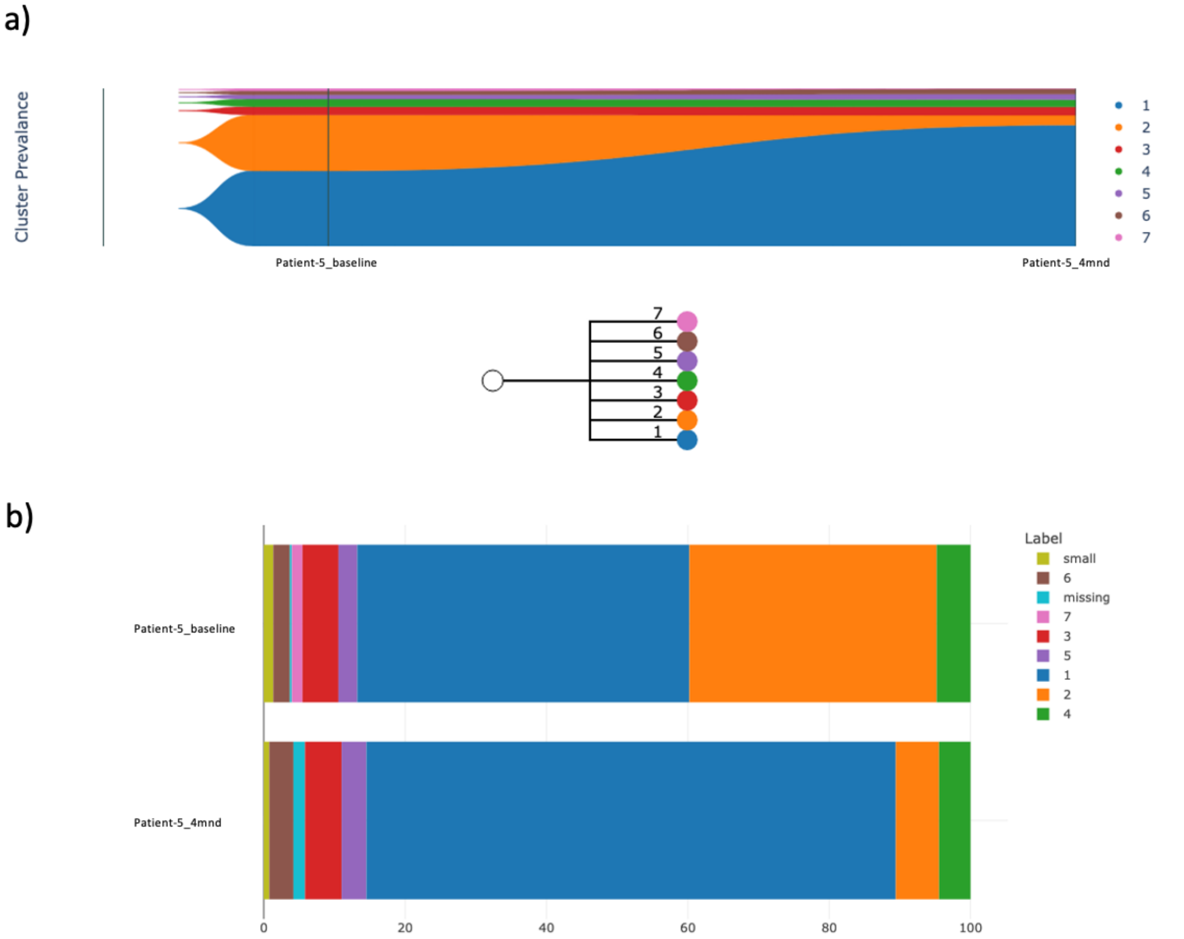


Figure 14. a) fishplot of patient 5 with one variant from each of the different clones from the WGS which was found again in the sample after targeted sequencing. b) bar plots over the distribution of subclones in the samples from patient 5 at baseline (diagnosis) and 4 months into treatment.

4.2.3. Interesting variants and clones found by Tapestri targeted single-cell DNA sequencing

The Tapestri targeted single-cell DNA sequencing data was analyzed solely based on the criteria described in chapter 3.6. Variants with DANN score = 1, % mutated cells > 1 and similar VAF by cell count/VAF by read count from all patients were handpicked for further analysis, some overlap with variants in the WGS analysis is observed.

Figure 15 shows the variants picked for patients 1, 2 and 3 from Tapestri Insights. The 6 variants chosen for patient 1 gave 7 main clones in Tapestri Insights (C1, C2, C3, C4, C5, C6 and C7), in addition to small subclones and missing GT subclones. The missing GT subclones makes out 64.94% of the clones, the small subclones 19.26%, C1 3.46%, C2 and C3 3.25%, C4 1.95%, C5 1.52%, C6 1.30% and C7 1.08%. The 5 chosen variants for patient 2 gave 7 different clones (C1, C2, C3, C4, C5, C6 and C7), plus small subclones and missing GT subclones. All clones except small subclones and missing GT subclones makes out 2.0% each of the total percentage, while small subclones is 0% and missing GT subclones makes out 86.0%. The 4 variants chosen for patient 3 gave 8 clones (C1, C2, C3, C4, C5, C6, C7 and C8), in addition to small subclones and missing GT subclones. The WT clone is defined as C8 and made out 1.02% of the total percentage, C1 7.34%, C2 6.88%, C3 3.16%, C4 and C5 1.47%, C6 and C7 1.24%, the small subclones 6.06% and missing GT subclones 70.09%.

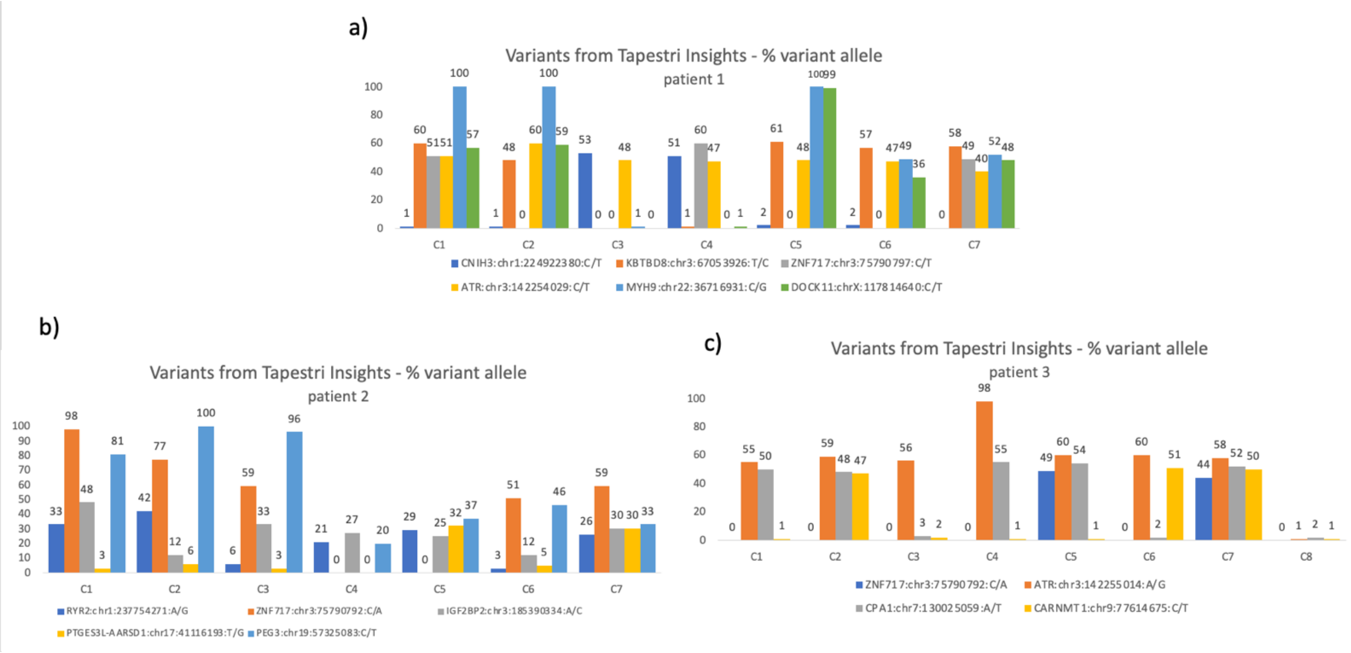


Figure 15. visualization of clone formation of variants handpicked from Tapestri Insights and the % variant allele per variant chosen in each sample. a) patient 1, b) patient 2 and c) patient 3.

In the time course treatment of patient 4 the variants ZNF17:chr13:75790791:G/C, KMT2C:chr7:151921099:C/T and TEX10:chr9:103072628:G/A define 5 different clones and the fishplot and the bar plot show that the presence of these clones does not change extensively during treatment (Figure 16). For patient 5 the variants CALU:chr7:128394413:G/A, CAPN6:chrX:110492213:G/A, NBPF20:chr1:144879090:T/C, UTRN:chr6:14483224:C/T and ZNF717:ch3:75790791:G/C define 8 clones. Figure 17 shows a fish plot and a bar plot that describes the development and distribution of the clones during treatment for patient 5. A cut-off of 8 clones was set for the fish plots to avoid the smallest clones, while the bar plots show the entire clonal profile in the samples found with the chosen variants.

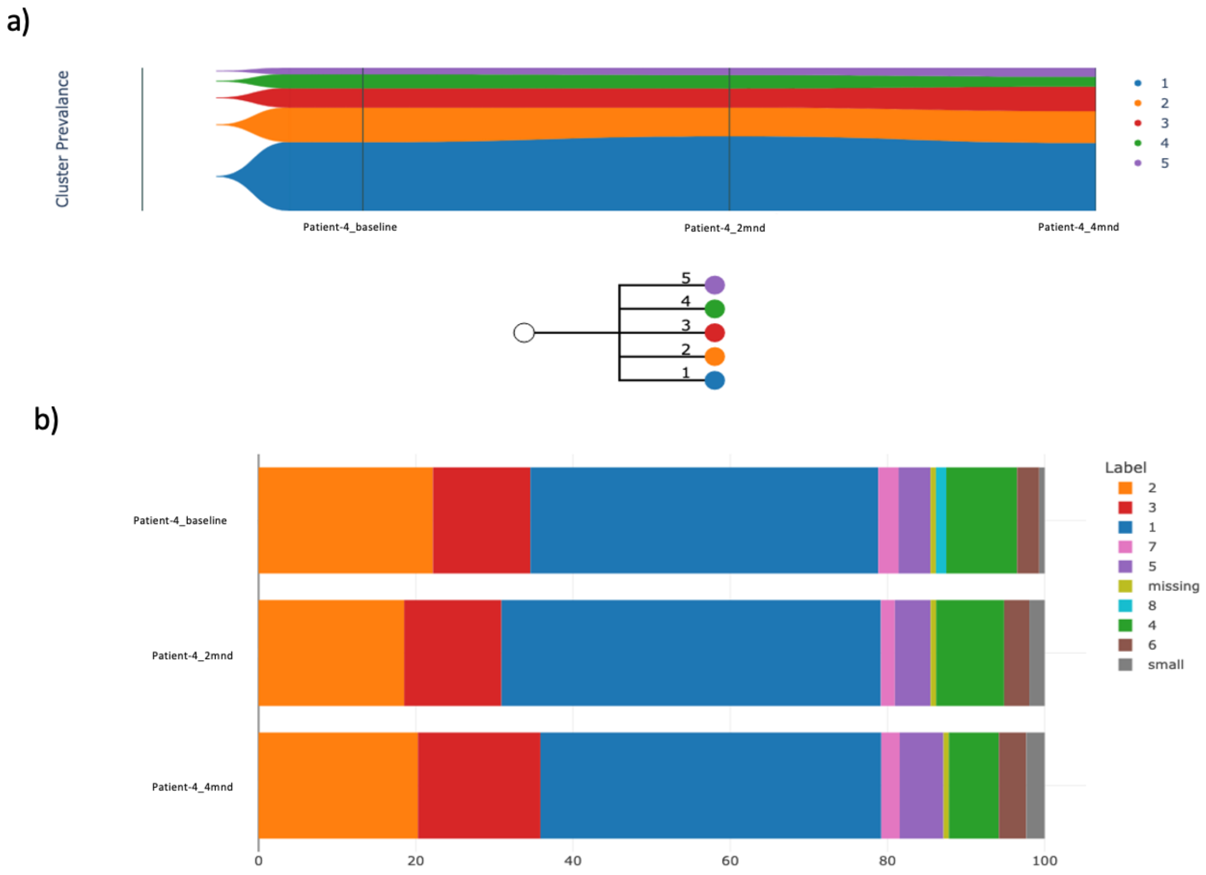


Figure 16. a) fish plot of patient 4 of selected variants from Insights handpicked based on criteria described in chapter 3.6.. b) bar plots over the distribution of subclones in the samples from patient 4 at baseline (diagnosis), 2 moths and 4 months into treatment.

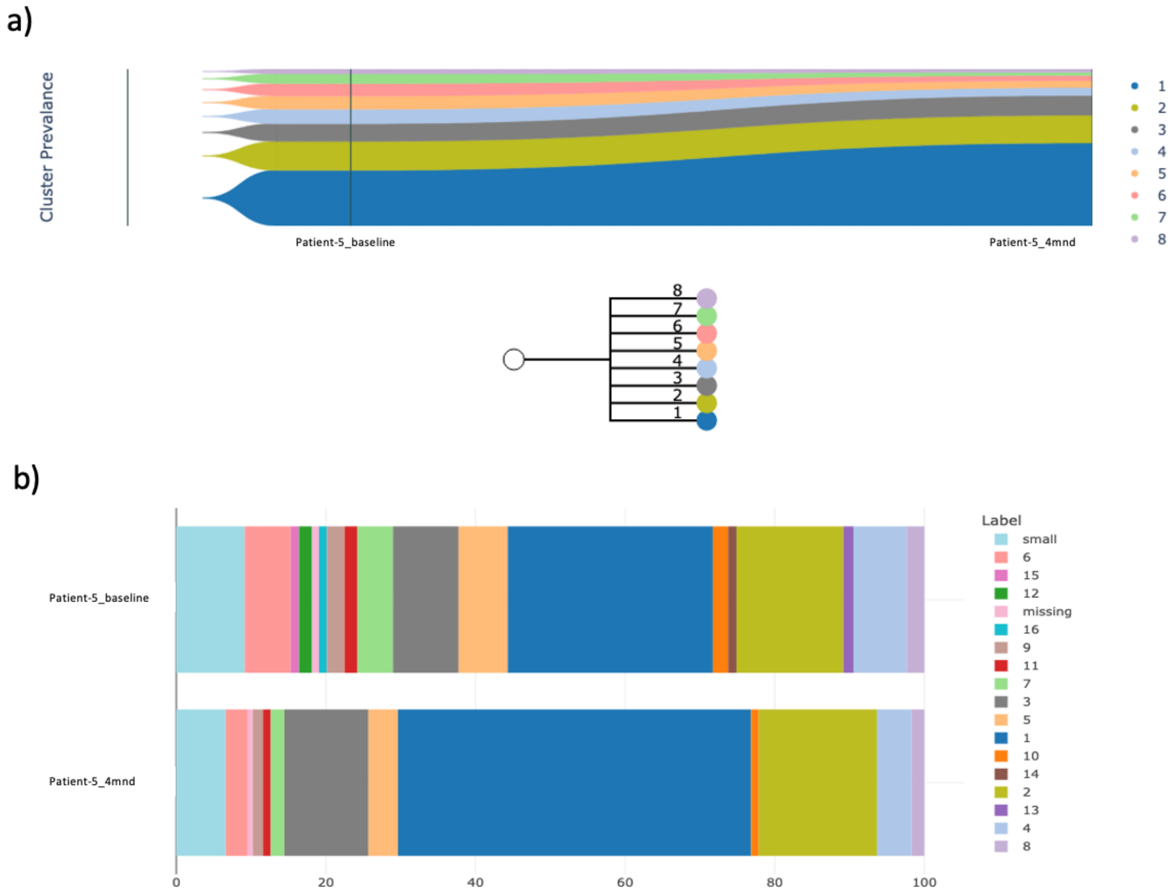


Figure 17. a) fish plot of patient 5 of selected variants from Insights handpicked based on criteria described in chapter 3.6. b) bar plots over the distribution of subclones in the samples from patient 5 at baseline (diagnosis) and 4 months into treatment.

An observation from the targeted single-cell DNA sequencing is the variations in the ZNF717 gene found in all the analyzed patients (appendix 8). Also, variations in the ATR gene, PEG3 gene and PTGES3L-AARSD1 gene are found in the tumor tissue of more than one of the 5 patients. Variants of genes known to commonly be linked to breast cancer such as TP53 and PIK3CA was found as germline mutations in patients 2, 4 and 5.

5. Discussion and future perspectives

5.1. Methodological evaluation

Breast tissue is highly fibrous and usually contains high amounts of adipose cells which challenges the process of fully dissociate the tissue samples into single nuclei without also breaking the nuclei membrane. Optimalization of manual extraction methods (enzymatically and mechanical) was done on training tissue with a higher content of adipose tissue than most tumor biopsies. The quality of the extracted nuclei differed from one tissue sample to another in intactness of the nuclei membrane, adiposeness, and amount of cell debris. These methods did not recover a sufficient number of nuclei for downstream analysis. Adipose cells (20-300 µm in diameter) are usually bigger than epithelial cells (60-100 µm in diameter) so the same size of an adipose tissue piece and an epithelial cell dense tissue piece will have different number of cells. Through the isolation protocol the nuclei from adipose tissue pieces were sticking to the fat droplets and difficult to pellet. Also, the fat may glue debris to the nuclei and because of this the nuclei are not able to migrate through the cell strainers.

Nuclei from three of the tumor biopsies was isolated with the manual enzymatic protocol and the concentration measured by Countess II FL. After sequencing the Tapestri pipeline software was able to distinguish 462 cells from sample 1, 50 cells from sample 2 and 886 cells from sample 3 (Table 4). The Mission Bio Tapestri library prep protocol should give data for 3000-10 000 cells. The low number of cells could be because of a low input of nuclei and in retrospect it was found that the Countess II FL show around a 10X higher nuclei concentration than the NucleoCounter® NC-100™ (data not shown). The automated Singulator™ 100 retrieve a much higher nuclei density in the isolate. Still, the purity varies regarding the tissue quality and remaining cell debris may clog the 40 µm microfluidics channels in the Tapestri cartridge preventing nuclei to be incorporated in the oil droplets. It is not known how fat droplets and other content in the isolate affect the formation of emulsion drops necessary for a successful single-cell analysis. For patient 4 with three time-point samples 2588, 1123 and 1919 cells were successfully analyzed, and for patient 5, with two time-point samples 1792 and 2845 cells was analyzed (Table 4). The number of nuclei loaded in the Tapestri cartridge for these samples were 200 000 for both sample 4 at baseline and at cross-over and approximately 99 000 for the surgery biopsy for patient 4. For patient 5 the number of nuclei loaded was 156 000 for the baseline sample and 200 000 for the surgery biopsy. The low number of nuclei loaded for the surgery biopsy of patient 4 did not affect the number of analyzed cells compared to the other 4 samples where the efficiency of loading nuclei into emulsions could have been affected by fat content or cell debris.

Table 4 in chapter 4.2 shows the run report metrics from all samples analyzed in this thesis. Several of the values in the run report were deviating from the desired values as described in chapter 4.2. The number of cells and mean reads/cell/amplicon are the two metrics where most samples are out of reach of the desired value. Mean reads/cell/amplicon is also called "coverage". The Tapestri pipeline goes over the cells for each amplicon and calculate the reads and average the number. 20% of this number is the cut off value, and amplicons under this value are labeled as "low-performing" amplicons and are left out as cells in the panel uniformity. The barcodes with at least 80% of "good-performing" amplicons are identified as cells in the panel uniformity and run report. This may indicate that the cut off value has become too high in the samples, and several amplicons are not evaluated as "good-performing". Too high value of mean reads/cell/amplicon may also cause false positive amplicons to be included. One way to improve the quality of the sequencing data could be to downsample the FASTQ file and remove reads that does not give useful information. The %DNA read pairs assigned to cells was lower in patients 1 and 2. This can be caused by primer dimers, or that the reads are aligning to barcodes that are not called as cells by the pipeline as expected. Patient 2 is outside of desired values for percentage reads mapped to genome and percentage reads mapped to target. To improve this value one can remove some of the targets, re-sequence the sample or if one has enough tissue do a repeat of the nuclei extraction. The ADO rate was higher in several of the samples. If this is higher than 15% it can give false positive annotations of variants present in the samples.

5.2. Variants from WGS and targeted sequencing

Variants found in the WGS analysis done on biopsies from the same patients were validated in the targeted sequencing. In patient 1 17 out of 36 variants were found, 8 out of 63 in patient 2, 35 out of 142 in patient 3, 15 out of 61 in patient 4 and 17 out of 21 in patient 5. Not all variants were found in the targeted sequencing, this may be caused by the difference in the selection of cells tissue seeing that the WGS were done on different biopsies than the targeted sequencing or that the primers designed by the Tapestri Designer to amplify certain regions did not bind efficiently to the template.

Variants representing each of the clones from WGS was handpicked, based on DANN score, percentage of mutated cells and VAF by cell count/VAF by read count quality. In patient 1 variants from 5 clones was found, but when analyzing one variant from each clone in Tapestri Insights only 3 clones were found using the handpicked variants (figure 12). In patient 2 it was only found variants in one clone in the WGS analysis, and the analysis in Tapestri Insights gave 2 clones with the chosen variants. In patient 3 variants from 9 different clones were found from the WGS analysis, but only one WT clone was found using Tapestri Insights with the chosen variants. The differences in the clones from WGS to the targeted single-cell analysis may be caused because of the low number of detected cells in the samples or by the missing GT subclones in Tapestri Insights. In many of the samples the percentage of missing GT subclones were high. Missing GT subclones are subclones with missing genotypes in one or more of the selected variants. This can indicate that there are several genotypes in the cells that are not included in the custom panel or have missing genotype information. Subclones that contain more than 1% of the total number of cells in the variant are included in clonal formation.

The composition of the cell clones in the tumor site for patients 4 and 5 was analyzed using variants with known clones from the WGS analysis. Figure 14 and figure 15 shows the fishplots and bar plots for patients 4 and 5 respectively. In patient 4 it can be observed that the WT clone, takes over the entire plot in both the fishplot and bar plot. This may indicate that the treatment is effective against the cells harboring the mutations characterizing clone 2, 3 and 4. The same trend can be observed in the plots for patient 5, but the clones does not fully disappear by the last timepoint in these samples. All the clones except clone 7 is still present in the last timepoint (4 months). This can indicate that the treatment is effective to some degree in eradicating the clones but is not able to kill all the cancer cells.

5.3. Variants from Tapestri Insights

In a time course treatment analysis variants from patients 4 and 5 were handpicked from Tapestri Insights based on DANN score, % mutated cells and similar VAF by cell count/VAF by read count. Interesting variants from patients 1, 2 and 3 were also handpicked to analyze further. The variants chosen are described in chapter 4.2.3, and in patients 1, 2 and 3 there were found 7, 7 and 8 clones respectively using these variants. Figure 16 and figure 17 shows fishplot and bar plot constructed in Mosaic using chosen variants for patients 4 and 5 from Tapestri insights. It can be observed that it is minimal change in the clones in both patient 4 and 5. There is a slight increase in clone 1 in patient 5 from the first timepoint (baseline) and the last timepoint (4 months). There was some overlap with variants from the bulk WGS analysis, but most of the ones solely chosen from Insights are not found in the WGS analysis. This may depend on the tissue sample since different biopsies are used for WGS analysis and targeted scDNA-seq. Other variants based on lower DANN score, less pathogenic but variants that would be better suited as markers for the clones or other combinations of variants may have given a better insight in how the tumor reacts to the treatment.

The gene ZNF717 was found to have high DANN score, % mutated genes and similar VAF by cell count/VAF by read count in all samples analyzed. This gene encodes a Kruppel-associated box (KRAB) zinc finger protein, which belongs to a group of transcriptional regulators in mammals. These proteins bind to nucleic acids and are vital in various cellular functions such as cell proliferation, differentiation and apoptosis, and in regulation of viral replication and transcription [58]. ZNF717 has been found interesting in several cancers among these colorectal and liver cancers such as hepatocellular carcinomas (HCC), but not yet proven to be related to breast cancer [59, 60] . Also, variations in the ATR gene, encoding a protein that works as a serine/threonine kinase and DNA damage sensor [61], PEG3, known as a mediator between p53/TP53 and BAX in a neuronal death pathway that is activated by DNA damage [62], and PTGES3L-AARSD1, encoding a protein harboring nucleic acid binding and aminoacyl-tRNA ligase activity [63], are found in the tumor tissue of more than one of the 5 patients [61-63].

ZNF717, ATR, PEG3 and PTGES3L-AARSD1 were mutated in more than one of the samples and may be common denominators for certain breast cancer clones. The variations included in the custom Tapestri panel design is mostly based on findings in the WGS analysis of 24 NeoLetExe study patients. The variants could be sample specific or specific for treatment response, or even specific for further development of the disease. A more extensive analysis of the data from the targeted scDNA-seq of sample 1-5 in addition to an extended cohort will put light on which variants to keep in this panel.

Some other interesting variants found in the samples were germline mutations in genes TP53 and PIK3CA which are known to be linked to breast cancer. TP53 is a tumor suppressor gene and encodes for a cellular tumor antigen protein that acts as a checkpoint control following DNA damage. Germline mutations in the TP53 gene in women have up to 85% chance to develop breast cancer by the age of 60 and can cause a familial cancer predisposition called Li-Fraumeni Syndrome (LFS) [64]. Patients 4 and 5 have intronic TP53 variants that have been reported benign (appendix 8). PIK3CA encodes for proteins that are involved in multiple processes in cells, such as protein synthesis, cell proliferation and survival, glucose homeostasis and DNA repair. Mutations in this gene is highly represented in ER+/HER+ breast cancer [65]. Patient 2 has a missense PIK3CA:p.E545K variant (rs104886003) that is likely pathogenic. This variant is often associated with HER2+ breast cancer, however the pathology report that patient 2 tumor HER2- (appendix 4).

The Tapestri Insights and Mosaic analysis lack the opportunity to analyze the data unsupervised. In Tapestri Insights the number of clones will vary with which and how many variants are analyzed together. With this manual approach the results will be person-dependent and not always correct. There is also a limitation in how many variants (max 10 variants) one can whitelist in Mosaic, making it challenging to decide on which variants to analyze at a time. There are several bioinformatic analysis that can be utilized on the sequencing data from this thesis in the future that would give a better insight in the effect of the treatment on the tumors, and important genes/variants. When analyzing the remaining samples from the NeoLetExe trial, other bioinformatic packages that would give unsupervised analysis of clones will be considered, for instance infSCITE. This software is designed for reconstruction of mutation histories in tumors based on mutation profiles from single-cell sequencing experiments [66].

6. Conclusion

Breast cancer biopsies were successfully sequenced in a targeted scDNA manner with the Mission Bio Tapestri platform; however the concentration and quality of the extracted nuclei are of importance for the number of analyzed cells and other data quality measures. We observed changes from baseline biopsy, through biopsy from cross-over of AI treatment until biopsy at time of surgery for patients 4 and 5 with same variants defining clones in WGS analysis. In both patients the WT clone grew over the timepoints, and the other clones shrunk in size. This indicates that the treatment of the patient is effective for these variants and removes/shrinks the clones with these variants. With variants solely chosen in Tapestri Insights using the guidelines given by this software there was little to no change in the clones. This may suggest that the treatment is not effective towards these variants and does not kill the cancer cells. Some genes were mutated in several or all the samples such as ZNF717, known to have implications in other cancers, for instance gastric cancer.

Targeted scDNA-seq is a promising method in monitoring treatment against specific clones, and to find and monitor clones that are highly pathogenic and with low cell counts. It may be possible to identify clones that are resistant to different types of treatment, and which clones that give a higher probability of recurrence. One of the limitations with this method so far has been the bioinformatics analysis of the sequencing data. A tool like infSCITE for unsupervised analysis of the clones from all filtered variants from the Tapestri pipeline analysis could better resolve the clonal landscape in a biopsy.

Sources

1. Jesinger, R.A., *Breast anatomy for the interventionalist.* Tech Vasc Interv Radiol, 2014. **17**(1): p. 3-9.
2. Javed, A. and A. Lteif, *Development of the human breast.* Semin Plast Surg, 2013. **27**(1): p. 5-12.
3. Walker, R.A. and C.V. Martin, *The aged breast.* J Pathol, 2007. **211**(2): p. 232-40.
4. Chen, W., et al., *Mammary Development and Breast Cancer: a Notch Perspective.* J Mammary Gland Biol Neoplasia, 2021. **26**(3): p. 309-320.
5. BioRender. *BioRender* [cited 2023; Available from: https://www.biorender.com.
6. Wellings, S.R. and H.M. Jensen, *On the origin and progression of ductal carcinoma in the human breast.* J Natl Cancer Inst, 1973. **50**(5): p. 1111-8.
7. Wellings, S.R., H.M. Jensen, and R.G. Marcum, *An atlas of subgross pathology of the human breast with special reference to possible precancerous lesions.* J Natl Cancer Inst, 1975. **55**(2): p. 231-73.
8. Winters, S., et al., *Chapter One - Breast Cancer Epidemiology, Prevention, and Screening*, in *Progress in Molecular Biology and Translational Science*, R. Lakshmanaswamy, Editor. 2017, Academic Press. p. 1-32.
9. Zeinomar, N. and W.K. Chung, *Cases in Precision Medicine: The Role of Polygenic Risk Scores in Breast Cancer Risk Assessment.* Ann Intern Med, 2021. **174**(3): p. 408-412.
10. Lien, T.G., et al., *Sample Preparation Approach Influences PAM50 Risk of Recurrence Score in Early Breast Cancer.* Cancers (Basel), 2021. **13**(23).
11. Houghton, S.C. and S.E. Hankinson, *Cancer Progress and Priorities: Breast Cancer.* Cancer Epidemiol Biomarkers Prev, 2021. **30**(5): p. 822-844.
12. Hofvind, S., Å. Holen, and G. Mangerud, *Mammografiprogrammet – tidligere, i dag og i fremtiden.* Norsk Epidemiologi, 2022. **30**.
13. Norway, C.R.o., *Cancer in Norway 2021 - Cancer incidence, mortality, survival and prevalence in Norway*. 2022: Oslo: Cancer Registry of Norway.
14. Sebuødegård, S., E. Botteri, and S. Hofvind, *Breast Cancer Mortality After Implementation of Organized Population-Based Breast Cancer Screening in Norway.* JNCI: Journal of the National Cancer Institute, 2019. **112**(8): p. 839-846.
15. Nolan, E., G.J. Lindeman, and J.E. Visvader, *Deciphering breast cancer: from biology to the clinic.* Cell, 2023.
16. Polyak, K., *Breast cancer: origins and evolution.* J Clin Invest, 2007. **117**(11): p. 3155-63.
17. Cichon, M.A., et al., *Microenvironmental influences that drive progression from benign breast disease to invasive breast cancer.* J Mammary Gland Biol Neoplasia, 2010. **15**(4): p. 389-97.
18. Polyak, K., *Is breast tumor progression really linear?* Clin Cancer Res, 2008. **14**(2): p. 339-41.
19. Lishman, S.C. and S.R. Lakhani, *Atypical lobular hyperplasia and lobular carcinoma in situ: surgical and molecular pathology.* Histopathology, 1999. **35**(3): p. 195-200.
20. Mueller, C., et al., *Protein biomarkers for subtyping breast cancer and implications for future research.* Expert Rev Proteomics, 2018. **15**(2): p. 131-152.
21. Zubair, M., S. Wang, and N. Ali, *Advanced Approaches to Breast Cancer Classification and Diagnosis.* Front Pharmacol, 2020. **11**: p. 632079.

22. Cserni, G., *Histological type and typing of breast carcinomas and the WHO classification changes over time.* Pathologica, 2020. **112**(1): p. 25-41.
23. Rakha, E.A., et al., *Breast cancer prognostic classification in the molecular era: the role of histological grade.* Breast Cancer Res, 2010. **12**(4): p. 207.
24. Hortobagyi, G.N., S.B. Edge, and A. Giuliano, *New and Important Changes in the TNM Staging System for Breast Cancer.* Am Soc Clin Oncol Educ Book, 2018. **38**: p. 457-467.
25. Gamble, P., et al., *Determining breast cancer biomarker status and associated morphological features using deep learning.* Communications Medicine, 2021. **1**(1): p. 14.
26. Zhang, A., et al., *The Role of Ki67 in Evaluating Neoadjuvant Endocrine Therapy of Hormone Receptor-Positive Breast Cancer.* Front Endocrinol (Lausanne), 2021. **12**: p. 687244.
27. Walsh, M.F., et al., *Genomic Biomarkers for Breast Cancer Risk.* Adv Exp Med Biol, 2016. **882**: p. 1-32.
28. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
29. Zhang, X., *Molecular Classification of Breast Cancer: Relevance and Challenges.* Archives of Pathology & Laboratory Medicine, 2022. **147**(1): p. 46-51.
30. Rønneberg, J.A., et al., *Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer.* Mol Oncol, 2011. **5**(1): p. 61-76.
31. Fleischer, T., et al., *DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival.* Oncotarget, 2016. **8**(1).
32. Gao, J.J. and S.M. Swain, *Luminal A Breast Cancer and Molecular Assays: A Review.* Oncologist, 2018. **23**(5): p. 556-565.
33. Ades, F., et al., *Luminal B breast cancer: molecular characterization, clinical management, and future perspectives.* J Clin Oncol, 2014. **32**(25): p. 2794-803.
34. Schnitt, S.J., *Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy.* Mod Pathol, 2010. **23 Suppl 2**: p. S60-4.
35. Martelotto, L.G., et al., *Breast cancer intra-tumor heterogeneity.* Breast Cancer Res, 2014. **16**(3): p. 210.
36. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-13.
37. Fisusi, F.A. and E.O. Akala, *Drug Combinations in Breast Cancer Therapy.* Pharm Nanotechnol, 2019. **7**(1): p. 3-23.
38. Dowling, R.J.O., et al., *Toronto Workshop on Late Recurrence in Estrogen Receptor-Positive Breast Cancer: Part 1: Late Recurrence: Current Understanding, Clinical Considerations.* JNCI Cancer Spectr, 2019. **3**(4): p. pkz050.
39. Burguin, A., C. Diorio, and F. Durocher, *Breast Cancer Treatments: Updates and New Challenges.* J Pers Med, 2021. **11**(8).
40. Cheng, Y.J., et al., *Long-Term Cardiovascular Risk After Radiotherapy in Women With Breast Cancer.* J Am Heart Assoc, 2017. **6**(5).
41. Taylor, C.W. and A.M. Kirby, *Cardiac Side-effects From Breast Cancer Radiotherapy.* Clin Oncol (R Coll Radiol), 2015. **27**(11): p. 621-9.

42. Peto, R., et al., *Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials.* Lancet, 2012. **379**(9814): p. 432-44.
43. Howlader, N., et al., *US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status.* J Natl Cancer Inst, 2014. **106**(5).
44. Xu, K., et al., *Integrative analyses of scRNA-seq and scATAC-seq reveal CXCL14 as a key regulator of lymph node metastasis in breast cancer.* Human Molecular Genetics, 2021. **30**(5): p. 370-380.
45. Evrony, G.D., A.G. Hinch, and C. Luo, *Applications of Single-Cell DNA Sequencing.* Annu Rev Genomics Hum Genet, 2021. **22**: p. 171-197.
46. Sloan, D.B., et al., *Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods.* Trends Biotechnol, 2018. **36**(7): p. 729-740.
47. Foulkes, W.D. and F.X. Real, *Many mosaic mutations.* Curr Oncol, 2013. **20**(2): p. 85-7.
48. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
49. Baslan, T., et al., *Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing.* Elife, 2020. **9**.
50. Ruff, D.W., et al., *High-Throughput Multimodal Single-Cell Targeted DNA and Surface Protein Analysis Using the Mission Bio Tapestri Platform.* Methods Mol Biol, 2022. **2386**: p. 171-188.
51. Zhang, H., et al., *Application of high-throughput single-nucleus DNA sequencing in pancreatic cancer.* Nature Communications, 2023. **14**(1): p. 749.
52. MissionBio. *Nuclei Extraction From Frozen Tissue For Single-Nuclei DNA Sequencing (User Guide)*. 2021; Available from: https://support.missionbio.com/hc/article_attachments/4421562098967/Protocol_Nuclei_Extraction_from_Frozen_Tissue_User_Guide_RevE_.pdf.
53. Genomics, S., *Singulator™ 100 System: Automated Tissue Dissociation System Guide*. 2022: S2 Genomics.
54. Genomics, S., *Cell Clean-up and Debris Removal Procedure*
55. MissionBio. *Tapestri Single-Cell DNA Sequencing V2 (user guide)*. 2020 [cited 2022 26/08/22]; Available from: https://support.missionbio.com/hc/article_attachments/360063349594/Tapestri_Single-Cell_DNA_Sequencing_V2_User_Guide_PN_3354H.pdf.
56. Quang, D., Y. Chen, and X. Xie, *DANN: a deep learning approach for annotating the pathogenicity of genetic variants.* Bioinformatics, 2015. **31**(5): p. 761-3.
57. MissionBio. *Difference in «VAF by Read Count», «VAF by cell count»?* 2022; Available from: https://support.missionbio.com/hc/en-us/articles/360042326414-Advanced-filtering.
58. NIH. *ZNF717 - zinc finger protein 717 (human)*. 2023 29/05/23 [cited 2023 09/05/23]; Available from: https://www.ncbi.nlm.nih.gov/gene/100131827.
59. Duan, M., et al., *Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma revealed by single-cell genome sequencing.* Cell Res, 2018. **28**(3): p. 359-373.
60. Liang, Y., et al., *Discovery of Aberrant Alteration of Genome in Colorectal Cancer by Exome Sequencing.* The American Journal of the Medical Sciences, 2019. **358**(5): p. 340-349.

61. Chevarin, M., et al., *The "extreme phenotype approach" applied to male breast cancer allows the identification of rare variants of ATR as potential breast cancer susceptibility alleles.* Oncotarget, 2023. **14**: p. 111-125.
62. Otsuka, S., et al., *Aberrant promoter methylation and expression of the imprinted PEG3 gene in glioma.* Proc Jpn Acad Ser B Phys Biol Sci, 2009. **85**(4): p. 157-65.
63. NIH. *PTGES3L-AARSD1(human)* 2023 29/05/23 [cited 2023 09/05/23]; Available from: https://www.ncbi.nlm.nih.gov/gene/100885850.
64. Schon, K. and M. Tischkowitz, *Clinical implications of germline mutations in breast cancer: TP53.* Breast Cancer Res Treat, 2018. **167**(2): p. 417-423.
65. Fusco, N., et al., *PIK3CA Mutations as a Molecular Target for Hormone Receptor-Positive, HER2-Negative Metastatic Breast Cancer.* Frontiers in Oncology, 2021. **11**.
66. Jahn, K., J. Kuipers, and N. Beerenwinkel, *Tree inference for single-cell data.* Genome Biol, 2016. **17**: p. 86.

Appendix 1; Solutions

Tissue Lysis Solution (2x concentration):

Table 6. 2x concentration tissue lysis solution used in Enzymatic Dissociation Protocol from the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing user guide.

Reagent	**Stock Concentration**	**Final Concentration**	**Volume for 2 mL**
Trypsin-EDTA (0.25%) phenol red	2.5 mg/mL (0.25%)	0.006 mg/mL (0.006%)	48 µL
Collagenase	8 mg/mL	0.2 mg/mL	50 µL
Dispase II	100 mg/mL	0.2 mg/mL	4 µL
Spermine Solution (pH 7.6)			1.898 mL

Tissue Lysis Solution (4x concentration):

Table 7. 4x concentration tissue lysis solution used in Enzymatic Dissociation Protocol from the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing user guide.

Reagent	**Stock Concentration**	**Final Concentration**	**Volume for 2 mL**
Trypsin-EDTA (0.25%) phenol red	2.5 mg/mL (0.25%)	0.12 mg/mL (0.012%)	96 µL
Collagenase	8 mg/mL	0.4 mg/mL	100 µL
Dispase II	100 mg/mL	0.4 mg/mL	8 µL
Spermine Solution (pH 7.6)			1.796 mL

S.I.P Percoll solution:

Table 8. S.I.P Percoll solution used in the Nuclei Clean-up and Debris Removal Procedure from S2 Genomics.

Reagent	Volume
10x PBS (-Ca/Mg)	500 µL
Percoll	4.5 mL

20% Percoll solution:

Table 9. 20% Percoll solution solution used in the Nuclei Clean-up and Debris Removal Procedure from S2 Genomics.

Reagent	Volume
S.I.P Percoll	2 mL
NSR	8 mL

Appendix 2; Supplies

Enzymatic and mechanical nuclei extraction:

Table 10. Supplies used in the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing user guide.

Equipment	**Supplier (Part number)**
pH meter	VWR (662-1861)
Refrigerated Centrifuge	Beckman Coulter (B08895)
Tube Vortexer	Thermo Fisher Scientific (88880017TS)
Fluorescent microscope EVOS FL	Life Technologies
5 mL DNA LoBind Eppendorf tubes	Eppendorf (30108310)
50 mL tubes	VWR (89004-364)
1.5 mL DNA LoBind Eppendorf tubes	Eppendorf (022431021)
15 mL tubes	Thermo Fisher Scientific (339651)
50 µm Celltrics ® cell strainer	Sysmex (04-004-2327)
30 µm Celltrics ® cell strainer	Sysmex (04-004-2316)
Sterile Petri Dishes	VWR (664160)
Dounce Tissue Grinder Set	Sigma (D8938-1SET)
NucleoCounter® NC-100™	ChemoMetec
SPRUCE: Sterile Disposable Scalpel With Carbon Steel Blade	VWR (BRAUBA815SU_P)

Singulator™ 100 and Nuclei Clean-up and Debris Removal Procedure:

Table 11. Supplies used when extracting nuclei with Singulator™ 100 and the Nuclei Clean-up and Debris Removal Procedure from S2 Genomics.

Equipment	**Supplier (Part number)**
Refrigerated Centrifuge	Beckman Coulter (B08895)
NucleoCounter® NC-100™	ChemoMetec
Singulator™ 100	S2 Genomics (100-067-764)
Nuclei Dissociation Chamber	S2 Genomics (100-060-817)
5 mL DNA LoBind Eppendorf tubes	Eppendorf (30108310)
15 mL tubes	Thermo Fisher Scientific (339651)
1.5 mL DNA LoBind Eppendorf tubes	Eppendorf (022431021)

Tapestri Single-cell DNA sequencing:

Table 12. Supplies used in the targeted Single-cell DNA sequencing with The Tapestri Single-cell DNA Sequencing V2 user guide.

Equipment	**Supplier (Part number)**
Mission Bio Tapestri Instrument	Mission Bio (191335)
Tapestri Single-Cell DNA Cartridge Kit	Mission Bio (PN 046459)
0.2 mL PCR Axygen MAXYmum Recovery PCR Tubes	Axygen (PCR-02-L-C)
Axygen Gel Tips	Axygen (TGL200RD57R)
TipOne RPT ultra low retention filter tip	USA Scientific (1180-8810)
Nuclease free Microcentrifuge Tubes, 1.5 mL	Eppendorf (0030108035)
0.2 mL PCR Tubes	USA Scientific (1402-8120)
Nucleocounter® NC-100™	ChemoMetec
4200 Tapestation system	Agilent (G2991BA)
Qubit Fluorometer	Qubit: Thermo Fisher (Q33216)
Qubit Assay tubes	Thermo Fisher (Q32856)
Pipettes, 1 µL – 1000 µL	Mettler-Toledo
Centrifuge	Eppendorf
Vortex mixer	Thermo Fisher Scientific (88880017TS)
Thermal cycler	Thermo Fisher Scientific
0.2 mL 8-strip PCR tube Magnetic Separation Stand	Seqmatic (TM-700)
6-Tube Magnetic Separation Rack	New England Biolabs (S1506S)
½ SP flowcell	Illumina
¼ S4 Novaseq flowcell	Illumina

Appendix 3; Reagents

Enzymatic and mechanical nuclei extraction:

Table 13. Reagents used in the nuclei extractions protocols from the Nuclei Extraction from Frozen Tissue For Single-Nuclei DNA Sequencing user guide.

Component name	**Supplier (Part number)**
DAPI solution 1 mg/mL	Thermo Fisher Scientific (62248)
Trypsin inhibitor from chicken egg white, Type II-O	Sigma (T9253)
Sodium citrate tribasic dehydrate	Sigma (C8532)
Spermine Tetrahydrochloride	Sigma (S1141)
Tris (Hydroxymethyl) aminomethane	Sigma (252859)
IGEPAL CA-630	SIGMA (I8896)
Trypsin-EDTA (0.25%), phenol red	Thermo Fisher Scientific (25200072)
Collagenase	Worthington (CLS-7 LS005332)
Dispase II	Gibco (17105-041)
Ribonuclease A from bovine pancreas, type I-A	Sigma (R4875-100mg)
1x DPBS, no calcium, no magnesium	Thermo Fisher Scientific (14190-136)
Dry ice in pellets	N/A
HCL, Molecular Biology grade	Sigma (H1758)
UltraPure™ BSA (50 mg/mL)	Thermo Fisher Scientific, AM2618
Nuclei EZ lysis buffer	MilliporeSigma / Sigma Aldrich (N3408)

Singulator™ 100 and Percoll clean-up:

Table 14. Reagents used when extracting nuclei using the Singulator™ 100 and Percoll clean-up..

Component name	**Supplier (Part number)**
Percoll	Sigma-Aldrich (P1644-25ML)
NSR – Nuclei Storage Reagent	S2 Genomics
NIR- Nuclei Isolation Reagent	S2 Genomics

10x PBS (-Ca/Mg)	Thermo Fisher Scientific (70011044)
DAPI solution 1 mg/mL	Thermo Fisher Scientific (62248)
BSA	Thermo Fisher Scientific (AM2618)

Targeted Single-Cell DNA sequencing:

Table 15. Reagents used when following the The Tapestri Single-cell DNA Sequencing V2 user guide.

Component name	**Supplier (Part Number)**
AMPure XP Reagent	Beckman Coulter (A63880)
DPBS w/o Ca^{2+}/Mg^{2+} (1X)	Gibco (14190-144)
Qubit® dsDNA HS Assay Kit	Qubit® (Q32851)
Ethanol, Molecular Biology Grade	Sigma (E7023)
Agilent High Sensitivity D5000 ScreenTape Agilent High Sensitivity D5000 Reagents	Matriks AS (5067-5592) Agilent Technologies (5067-5593)
Sequencing Reagent Kit (NovaSeq 6000)	Illumina
Tapestri Single-Cell DNA Core Ambient Kit v2	Mission Bio (MB51-0007)
Tapestri Single-Cell DNA Core – 20 Kit	Mission Bio (MB51-0010)
Tapestri Single-Cell DNA Bead Kit	Mission Bio (MB51-0009)
Tapestri Single-Cell DNA Custom Kit	Mission Bio (PN 145936)

Appendix 4; Pathology data

Table 16. Pathology data for patients 1-5. NST is invasive breast cancer and stands for no special type, and SNEC stands for small cell neuroendocrine carcinoma. EXE and LET under AI-sequence stands for exemestane and letrozole respectively, while MAST under surgery stands for mastectomy.

Patient	Age a.d.	BC type	grade	cT	Size (mm)	ER	PGR	HER-2	AI-sequence	M at diag.	Surgery	Size (mm) at surgury	Positive nodes/nodes checked	Response
1	81	NST	II	cT4	42	100 %	<10%	Negative	EXE-LET	M0	MAST+ axilla	30	5/9	PR
2	73	NST	II	cT4	40	100 %	90 %	Negative	LET-EXE	M0	MAST+ sentinel node	15	1/2	PR
3	78	NST	I	cT4	45	>50%	neg.	Negative	EXE-LET	M0	MAST+ sentinel node	25	0/1	PR
4	84	NST	II	cT4	50	>50%	>10%	Negative	LET-EXE	M0	MAST+ sentinel node	11	1/9	PR
5	82	SNEC with mucinous parts	II	cT4	65	90 %	90 %	Negative	LET-EXE	M0	MAST+ sentinel node	25	0/2	PR

Appendix 5; NGS and extra library clean up

Table 17 shows an overview of the library pools, information about the nuclei suspensions, sequencing, and targeted PCR. The library pool clean-up procedures are explained in this appendix, and Figure 18 shows a good example of wanted electropherogram after the library clean-up.

Table 17. Overview over the NGS set up for patients 1-3 (pool 1), and patients 4 and 5 (pool 2). The index number refers to index primer which are listed in The Tapestri Single-cell DNA Sequencing V2 user guide.

	Sample	Tissue comments	Nuclei isolation method	Tissue weight (mg)	Nuclei measure method	Nuclei concentration, nuclei/µl	Volume nuclei suspension	Add cell buffer (µl) to	Targeted PCR Qubit (ng/µl)	Index no.	Sequencing information	No. of cycles	TapeStation conc. (ng/µl)	Primer dimer peak, nmol/l	Amplicon peak, nmol/l	TapeStation quality	Library Qubit (ng/µl)
Pool 1*	1_baseline		Manual-enzymatic		Countess	16 100	12.4	37.6	0.406	1	½ SP flowcell (XP workflow)	10	1.42	0.606	4.1	12.9	1.2
	2_baseline		Manual-enzymatic		Countess	10 800	18.5	31.5	0.222	3	½ SP flowcell (XP workflow)	12	1.43	0.854	4.06	17.4	1.33
	3_baseline		Manual-enzymatic		Countess	31 900	6.3	43.7	0.384	2	½ SP flowcell (XP workflow)	10	1.53	0.265	4.63	5.4	1.44
Pool 2**	4_baseline	Very, very fatty	Singulator 100	30+70+140	NucleoCounter	4 000	All used	0	0.654	5	¼ S4 Novaseq flowcell	12	20.1	1.4	18.7	7.0	14.8
	4_2months	Dispersed coloration in tissue. Precipitate in nuclei isolate.	Singulator 100	27	NucleoCounter	6 506	All used	0	0.536	7	¼ S4 Novaseq flowcell	12	13	0.472	11.7	3.9	9.5
	4_4months	Dispersed coloration in tissue. Precipitate in nuclei isolate.	Singulator 100	80	NucleoCounter	1 981	All used	0	0.872	4	¼ S4 Novaseq flowcell	12	15.9	0.838	13.4	5.9	11.3
	5_baseline	Good	Singulator 100	28	NucleoCounter	3 130	All used	0	0.782	6	¼ S4 Novaseq flowcell	10	17.2	1.58	53.7	2.9	13
	5_4months	Good	Singulator 100	26	NucleoCounter	8 730	25	25	0.846	8	¼ S4 Novaseq flowcell	10	19.5	1.92	66.9	2.8	15.2

Library pool preparations Pool 1:

All three samples underwent, in separate tubes, an extra cleanup due to high concentrations of short fragments, following this protocol:

1. Add 16 µL of nuclease-free water to 9 µL sample for a total volume of 25 µL.
2. Mix and quick-spin to collect the contents.
3. Add 18 µL (0.72x) of Ampure XP reagent, at room temperature and well-mixed, to the above sample.
4. Vortex for 5 seconds and quick-spin to collect the contents.
5. Incubate the tube at room temperature for 5 minutes.
6. Place on the magnet and wait 5 minutes for the beads to separate from the solution.
7. Without removing the tube from the magnet, remove the clear liquid from the tube and discard.
8. Add 100 µL of the freshly prepared 80 % ethanol, wait 30 seconds, and remove 100 µL of ethanol without disturbing the Ampure beads.
9. Repeat Step 8 once for a total of two wash cycles.
10. Remove all residual ethanol from the tube. Take the tube off the magnet and do a quick spin. Place the tube back on the magnet with the caps open and remove any residual ethanol.
11. Dry the Ampure bead pellets in the tubes on the magnet by incubating at room temperature for 2 – 5 minutes. *Avoid overdrying the beads.*
12. Remove the tube from the magnet. Add 9 µL of nuclease-free water into the tube. Vortex and quick-spin to collect the contents.
13. Incubate the tubes at room temperature for 2 minutes.
14. Place the tube onto the magnet and wait for at least 2 minutes or until the solutions are clear.
15. Transfer 8 µL of purified PCR product from the tube to a new 0.2 mL PCR.

The purified libraries were measured for concentration and quality using TapeStation 4200 and 5nM of each was combined prior to sequencing in a ½ SP flowcell with a XP workflow (The NovaSeq™ Xp workflow provides flexibility and control without sacrificing data quality or yield (illumina.com)) for low concentration samples at the Norwegian Sequencing center.

Pool 2

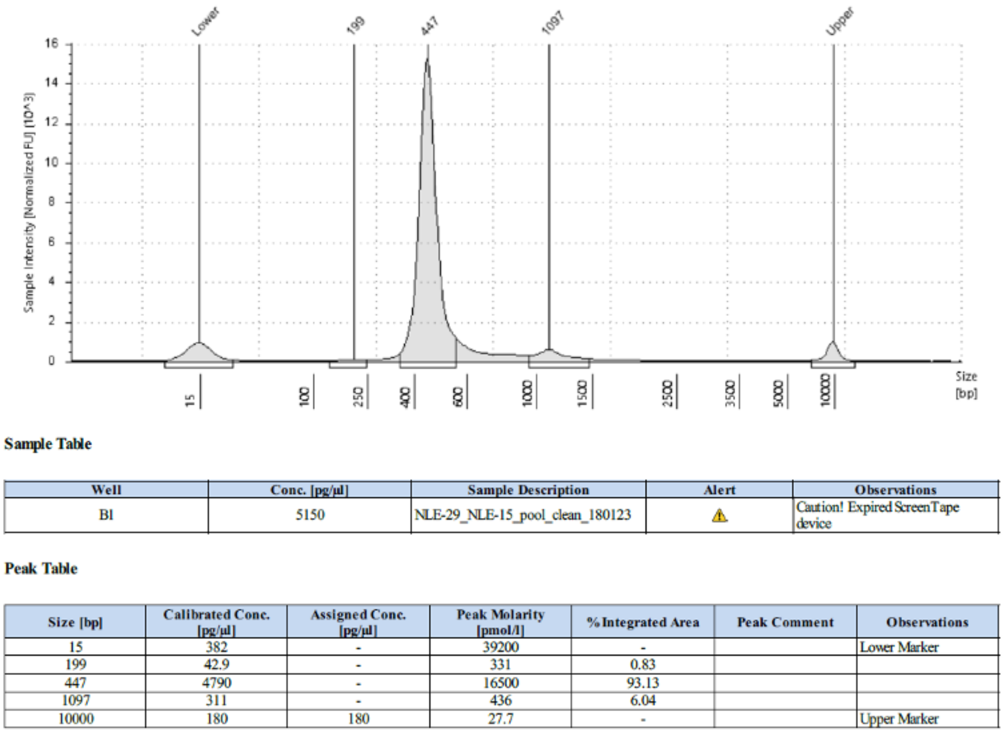
The tree timepoints for patient 4 and two timepoints for patient 5 were pooled prior to an extra cleanup step

Library	µl
4_baseline	4.46
4_2months	7
4_4months	6.09
5_baseline	5.25
5_4months	4.22
	27.02

Cleanup protocol:

1. Add 48µl nuclease-free water to 27.02µl pool (total volume 75µl)
2. Mix and quick-spin to collect the contents.
3. Add 54 µL (0.72x) of Ampure XP reagent, at room temperature and well-mixed, to the above sample.
4. Vortex for 5 seconds and quick-spin to collect the contents.
5. Incubate the tube at room temperature for 5 minutes.
6. Place on the magnet and wait 5 minutes for the beads to separate from the solution.
7. Without removing the tube from the magnet, remove the clear liquid from the tube and discard.
8. Add 200 µL of the freshly prepared 80 % ethanol, wait 30 seconds, and remove 100 µL of ethanol without disturbing the Ampure beads.
9. Repeat Step 8 once for a total of two wash cycles.
10. Remove all residual ethanol from the tube. Take the tube off the magnet and do a quick spin. Place the tube back on the magnet with the caps open and remove any residual ethanol.
11. Dry the Ampure bead pellets in the tubes on the magnet by incubating at room temperature for 2 – 5 minutes. *Avoid overdrying the beads.*
12. Remove the tube from the magnet. Add 27 µL of nuclease-free water into the tube. Vortex and quick-spin to collect the contents.
13. Incubate the tubes at room temperature for 2 minutes.
14. Place the tube onto the magnet and wait for at least 2 minutes or until the solutions are clear.
15. Transfer 25 µL of purified PCR product from the tube to a new 0.2 mL PCR and use the last 2µl for TapeStation.

The purified library pool quality measure detected 1% short fragments.



Sample Table

Well	Conc. [pg/µl]	Sample Description	Alert	Observations
B1	5150	NLE-29_NLE-15_pool_clean_180123	⚠	Caution! Expired ScreenTape device

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
15	382	-	39200	-		Lower Marker
199	42.9	-	331	0.83		
447	4790	-	16500	93.13		
1097	311	-	436	6.04		
10000	180	180	27.7	-		Upper Marker

Figure 18. Electropherogram from Tapestation analysis of patient 4 after library pool clean-up. The fragment at 447 bp is the desired peak, and the peak at approximately 190 bp indicates primer dimer.

Appendix 6; Example code in Mosaic

Example code for identification of clones based on a whitelist of variants and construction of fishplot and bar plot:

Multisample analysis

Load data

```
In [1]:
```

```
# Import mosaic libraries
import missionbio.mosaic as ms

# Import these to display entire dataframes
from IPython.display import display, HTML

# Import graph_objects from the plotly package to display figures when saving the notebook as an HTML
# Import numpy for statistics
import plotly as px
import plotly.graph_objects as go
import numpy as np

# Import additional packages for specific visuals
import missionbio.mosaic.utils as mutils
import matplotlib.pyplot as plt

# Import the colors
from missionbio.mosaic.constants import COLORS
import seaborn as sns

import plotly.offline as pyo
pyo.init_notebook_mode()
```

```
In [2]:
```

```
# Check version; this notebook is designed for Mosaic 2.2 or higher
print(ms.__version__)
```

```
2.2
```

```
In [3]:
```

```
h5path = r”/Users/patrikw/Multisample_analysis/merged_NLE-15.dna.h5”

# Select respectively Bulk or Tapestri
# From Bulk white list
group = ms.load(h5path, raw=False, apply_filter=False, whitelist=[‘chr12:100601481:A/G’,’chr1:145438
939:T/G’,’chr5:140856734:G/C’,’chr8:1581155:G/T’])

# From Tapestri white list (Grethe)
# group = ms.load(h5path, raw=False, apply_filter=False, whitelist=[‘chr3:75790791:G/C’,’chr7:151921099:C
/T’,’chr9:103072628:G/A’])

print(group)
# Print the list of samples in the group object
[s.name for s in group.samples]
```

```
Loading, /Users/patrikw/Multisample_analysis/merged_NLE-15.dna.h5
Loaded in 4.2s.
Group of 3 samples
```

subsetting data for variants of interest

```
In [4]:
```

```
def filt(sample):
   filt_vars = sample.dna.filter_variants()
   return filt_vars

# dna_vars = group.apply(filt)
dna_vars = group.apply(filt)



# Check the number of filtered variants. When using the default filters, the number of
# variants is likely smaller compared to the originally loaded variants due to the more
# stringent filtering criteria (e.g., vaf_ref=5, vaf_hom=95, vaf_het=35).

For I in range(len(group.samples)):
   sample = group.samples[i]
   print(sample.name)
   print(“Number of variants:”, len(dna_vars[i]))
   print(dna_vars[i], “\n”)
```

```
In [5]:
```

```
# Select respectively Bulk or Tapestri
# From Bulk
final_vars = [‘chr12:100601481:A/G’,’chr1:145438939:T/G’,’chr5:140856734:G/C’,’chr8:1581155:G/T’]

# # From Tapestri
# final_vars = [‘chr3:75790791:G/C’,’chr7:151921099:C/T’,’chr9:103072628:G/A’]
```

In [6]:

```
len(final_vars)
```

```
4
```

In [7]:

```
for sample in group:
  print(sample.dna.shape)
```

In [8]:

```
for sample in group:
  print(set(final_vars).issubset(set(sample.dna.ids())))
```

In [9]:

```
for sample in group:
  sample.dna = sample.dna[sample.dna.barcodes(), final_vars]
```

Annotation addition

In [10]:

```
for sample in group:
  annotation = sample.dna.get_annotations()



  for col, content in annotation.items():
    sample.dna.add_col_attr(col, content.values)
```

In [11]:

```
ann = annotation.sort_values(by=[“DANN”, “Coding impact”], ascending=False)

display(HTML(ann.to_html()))
```

In [12]:

```
# Add annotation to the id names
for sample in group:
  sample.dna.set_ids_from_cols([“Gene”, “id”])
  # Another xample:
  # sample.dna.set_ids_from_cols([“Gene”, “CHROM”, “POS”, “REF”, “ALT”])
```

```
    # Annotations are now added to the variants
    print(sample.name, “\n”, sample.dna.ids(), “\n”)
```

clustering

In [13]:

```
variants_of_interest = sample.dna.ids()
def cluster(sample):
    clone_table = sample.dna.group_by_genotype(variants_of_interest) #group_missing=True, min_clone_size =1, layer=”NGT_FILTERED”, show_plot=True
    return clone_table

tables = group.apply(cluster)

[display(HTML(t.to_html())) for t in tables]
```

Fishplot and barplot

In [16]:

```
group.fishplot(
    “dna”,
    sample_order=[“NLE-15_baseline”,”NLE-15_2mnd”,’NLE-15_4mnd’],
    # labels=[“0-1-1”, “1-1-2”],
    # labels=[“1”, “2”,”3”,”4”,”5”],
    labels=[“1”, “2”,”3”,”4”],
    parents=[None,None,None,None]
    )
```

In [17]:

```
# Draw a barplot for the dna labels

group.barplot(
    “dna”,
    sample_order=[“NLE-15_baseline”,”NLE-15_2mnd”,’NLE-15_4mnd’],
    # labels=[“1”,”2”,”3”,”4”,”5”,”6”,”small”,”missing”],
    # labels=[“None”,”2”,”3”,”4”,”5”,”6”,”small”,”missing”],
    percentage=True)
```

Appendix 7; Variants used for construction of plots in Mosaic

Patient 4:

Table 18. Variants for patient 4 chosen for further analysis known clones from the WGS.

Variant known from bulk	**WT**	**C2**	**C3**	**Small Subclones (12)**	**Missing GT Subclones (45)**
ACTR6:chr12:100601481:A/G	WT (1%)	WT (1%)	Het (39%)	-	Missing in 4.65% of clones
NBPF20:chr1:145438939:T/G	WT (0%)	Het (50%)	Het (49%)	-	Missing in 4.67% of clones
chr5:140856734:G/C	WT (0%)	WT (2%)	Het (44%)	-	Missing in 5.42% of clones
chr8:1581155:G/T	WT (0%)	WT (2%)	Het (50%)	-	Missing in 36.07% of clones
Total	2107 (37.42%)	857 (15.22%)	59 (1.05%)	92 (1.63%)	2515 (44.67%)
Patient-4_baseline.cells	482 (18.62%)	857 (33.11%)	59 (2.28%)	89 (3.44%)	1101 (42.54%)
Patient-4_2mnd.cells	522 (46.48%)	0 (0.00%)	0 (0.00%)	1 (0.09%)	600 (53.43%)
Patient-4_4mnd.cells	1103 (57.48%)	0 (0.00%)	0 (0.00%)	2 (0.10%)	814 (42.42%)

Table 19. Variants from patient 4 chosen solely from Tapestri Insights based on DANN score, %mutated cells and VAF-scores.

Variant from Tapestri selection	**C1**	**C2**	**C3**	**C4**	**WT**	**Small Subclones (8)**	**Missing GT Subclones (27)**
ZNF717:chr3:75790791:G/C	WT (0%)	Het (39%)	WT (0%)	WT (0%)	WT (0%)	-	Missing in 15.22% of clones
KMT2C:chr7:151921099:C/T	Het (42%)	Het (45%)	Het (41%)	WT (4%)	WT (3%)	-	Missing in 24.44% of clones
TEX10:chr9:103072628:G/A	Het (30%)	Het (31%)	WT (4%)	Het (29%)	WT (4%)	-	Missing in 49.52% of clones
Total	936 (16.63%)	292 (5.19%)	246 (4.37%)	146 (2.59%)	86 (1.53%)	105 (1.87%)	3819 (67.83%)
Patient-4_baseline.cells	420 (16.23%)	154 (5.95%)	112 (4.33%)	73 (2.82%)	49 (1.89%)	46 (1.78%)	1734 (67.00%)
Patient-4_2mnd.cells	236 (21.02%)	62 (5.52%)	23 (2.05%)	35 (3.12%)	10 (0.89%)	20 (1.78%)	737 (65.63%)
Patient-4_4mnd.cells	280 (14.59%)	76 (3.96%)	111 (5.78%)	38 (1.98%)	27 (1.41%)	39 (2.03%)	1348 (70.24%)

Patient 5:

Table 20. Variants for patient 5 chosen for further analysis known clones from the WGS.

Variant known from bulk	C1	WT	C3	C4	C5	C6	C7	Small Subclones (20)	Missing GT Subclones (56)
CAPN6:chrX:110492213:G/A	Het (52%)	WT (1%)	Het (51%)	WT (2%)	Het (50%)	Het (57%)	Hom (98%)	-	Missing in 13.89% of clones
HUWE1:chrX:53575011:T/C	WT (0%)	WT (1%)	WT (1%)	WT (0%)	WT (0%)	WT (0%)	WT (0%)	-	Missing in 1.85% of clones
SYCP1:chr1:115399240:A/G	Het (53%)	WT (1%)	Het (51%)	Het (52%)	WT (3%)	Hom (98%)	Het (58%)	-	Missing in 11.93% of clones
UTRN:chr6:144832224:C/T	WT (0%)	WT (0%)	Het (48%)	WT (0%)	WT (0%)	WT (0%)	WT (0%)	-	Missing in 3.47% of clones
Total	1740 (37.52%)	910 (19.62%)	248 (5.35%)	116 (2.50%)	104 (2.24%)	79 (1.70%)	69 (1.49%)	111 (2.39%)	1260 (27.17%)
Patient-5_4mnd.cells	1618 (56.87%)	128 (4.50%)	1 (0.04%)	95 (3.34%)	88 (3.09%)	58 (2.04%)	62 (2.18%)	10 (0.35%)	785 (27.59%)
Patient-5_baseline.cells	122 (6.81%)	782 (43.64%)	247 (13.78%)	21 (1.17%)	16 (0.89%)	21 (1.17%)	7 (0.39%)	101 (5.64%)	475 (26.51%)

Table 21. Variants from patient 5 chosen solely from Tapestri Insights based on DANN score, %mutated cells and VAF-scores.

Variant from Tapestri selection	C1	C2	C3	C4	WT	C6	C7	C8	Small Subclones (39)	Missing GT Subclones (192)
CALU:chr7:128394413:G/A	Het (51%)	WT (0%)	Het (50%)	Het (51%)	WT (0%)	Het (48%)	WT (0%)	Het (49%)	-	Missing in 10.39% of clones
CAPN6:chrX:110492213:G/A	Het (53%)	WT (1%)	Het (51%)	Het (51%)	WT (1%)	Het (53%)	WT (1%)	WT (3%)	-	Missing in 13.89% of clones
NBPF20:chr1:144879090:T/C	Het (32%)	Het (37%)	WT (4%)	Het (31%)	WT (4%)	Het (30%)	Het (34%)	Het (32%)	-	Missing in 41.13% of clones
UTRN:chr6:144832224:C/T	WT (0%)	WT (0%)	WT (0%)	WT (0%)	WT (0%)	Het (49%)	WT (0%)	WT (0%)	-	Missing in 3.47% of clones
ZNF717:chr3:75790791:G/C	WT (0%)	WT (0%)	WT (0%)	Het (37%)	WT (0%)	WT (0%)	Het (38%)	WT (0%)	-	Missing in 13.69% of clones
Total	632 (13.63%)	343 (7.40%)	160 (3.45%)	117 (2.52%)	95 (2.05%)	84 (1.81%)	61 (1.32%)	47 (1.01%)	269 (5.80%)	2829 (61.01%)

Patient-5_4mnd.cells	587 (20.63%)	46 (1.62%)	142 (4.99%)	112 (3.94%)	17 (0.60%)	0 (0.00%)	6 (0.21%)	40 (1.41%)	135 (4.75%)	1760 (61.86%)
Patient-5_baseline.cells	45 (2.51%)	297 (16.57%)	18 (1.00%)	5 (0.28%)	78 (4.35%)	84 (4.69%)	55 (3.07%)	7 (0.39%)	134 (7.48%)	1069 (59.65%)

Appendix 8; Variants of interest and %mutated cells

Table 22. %mutated cells of variants of interest in patients 1-5.

	Gene Variants													
Sample	**ATR:chr3:142254029:C/T**	**ATR:chr3:142255014:A/G**	**ZNF717:chr3:75790791:G/C**	**ZNF717:chr3:75790792:C/A**	**ZNF717:chr3:75790797:C/T**	**ZNF717:chr3:75790811:G/T**	**ZNF717:chr3:75790837:C/T**	**PEG3:chr19:57325083:C/T**	**PTGES3L:chr17:41116193:T/G**	**PIK3CA:chr3:178936091:G/A**	**TP53:chr17:7577407:A/C**	**TP53:chr17:7577427:G/A**	**TP53:chr17:7578115:T/C**	**TP53chr17:7576841:A/G**
Patient 1	86%	-	19%	18%	39%	39%	11%	77%	9%	96%	-	-	-	-
Patient 2	-	-	77%	77%	-	-	-	88%	21%	-	-	-	-	-
Patient 3	87%	14%	13%	20%	20%	-	-	-	-	-	-	-	-	-
Patient 4	-	-	21%	20%	30%	-	-	-	-	-	-	-	-	92%
Patient 5	-	-	15%	14%	31%	31%	11%	-	-	-	97%	97%	100%	-

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway