NTNU
Norwegian University of
Science and Technology

# Focused analysis of genomic alterations in primary breast carcinomas and corresponding metastases

Vilde Naume Solem

Master of Science
Submission date: December 2015
Supervisor: Atle M. Bones, IBI
Co-supervisor: Hege Russnes, Oslo Universitetssykehus

Norwegian University of Science and Technology
Department of Biology

## Acknowledgements

This Master Thesis marks the end of my work as a student at the Department of Biology and at the Teacher Program in Science at the Norwegian University of Technology and Science. The thesis was accomplished at the Department of Cancer Genetics, Institute of Cancer Research at the Norwegian Radium Hospital in Oslo.

I am very grateful to Professor Anne-Lise Børresen-Dale and Gry Geitvik for opening the research facilities to master students and for giving me the opportunity to perform my thesis at the Department of Genetics. My special thanks go to my supervisor Dr. Hege G. Russnes. Thank you for letting me be a part of your research group and for your guidance and encouragement along the way. Your never-ending optimism and your excitement and curiosity for biology have been a true inspiration. I would also like to thank my supervisor at NTNU, Atle Bones, for taking care of the formalities during the thesis. My gratitude goes to engineer Inger Riise Berg for taking the time to teach and help me in the laboratory and for always being available for guidance when problems occurred. Your laughter and high spirits have been highly appreciated in moments of frustrations in the laboratory.

This study had not been possible without the previous efforts of several of the employees at the department. Arne Pladsen, Dr. Lars Ottestad and Turid Gjertsen provided access to the samples and clinical information for all the patients. Thank you for generously sharing all your knowledge with me, and for your tremendous efforts establishing the OsloVal cohort. Inger Riise Bergheim, Puong Vu, Eldri Undlien Due and Veronica Skarpeteig have performed tumor tissue preparations and DNA extractions of the samples. Laxmi Silwal-Pandit and Vegard Nygaard have created scripts necessary for variant calling and annotation and together with Daniel Nebdal guided me through the bioinformatical pipelines. Furthermore, I would like to thank everyone at the Department of Genetics for your support and help and for making me feel welcome at the department.

I also have to express my gratitude to my friends and fellow students in Trondheim (you know who you are) who have contributed to make my time at NTNU so amazing. Thank you Mads Fredrik Heiervang, for being my rock at home and for making me laugh even in

moments when I am grumpy and stressed. Last, but definitely not least, I want to thank my wonderful parents and sister.

Oslo, December, 2015

Vilde Naume Solem

# Abstract

Breast cancer is one of the leading causes of cancer mortality worldwide. Distant metastases are nearly always the direct cause of death, and understanding its relationship to the primary tumor is of great importance. Today, the choice of treatment is, in most cases, based upon histological and molecular analysis of the primary tumor. However, recent studies suggest that the cancer cells may spread to distant organs earlier in tumor progression than previously believed and that genetic alterations can evolve independently in the primary tumor and metastasis. Some studies report of a close resemblance between primary tumors and their corresponding metastasis, while others report of genetic divergence. The conflicting results reflect the need for a deeper understanding of the molecular mechanisms underlying metastatic disease, to be able to improve treatment strategies and clinical outcome.

The objective of this thesis was to investigate whether the somatic mutations found in a lymph node metastasis differ from those in the corresponding primary breast tumor. Targeted sequencing of 20 primary breast tumors and their matched lymph node metastases was performed by the use of Ion Torrent Personal Genome Machine. The Ion AmpliSeq™Cancer Hotspot Panel v2 was utilized, targeting 207 regions in 50 genes found to be frequently mutated in cancer. The genetic variants were compared to public databases such as dbSNP, 1000 genomes and COSMIC. 55% of the tumors were found to harbor at least one somatic mutation (median 0,8, range 0-3). Frequently mutated genes included *TP53* (45%) and *PIK3CA* (25%). The vast majority of the metastases seemed to retain the somatic mutations detected in the primary tumor, but the variant frequencies were slightly different. In three of the patients a *TP53* mutation unique to the primary tumor and/or the lymph node metastasis were revealed. The differences indicate that dissemination may occur at different time points during disease progression, and that analysis of metastatic tumors could provide additional insight affecting treatment decisions. However, as only a small piece of each tumor was analyzed and only selected regions of the cancer genomes were sequenced, no conclusion can be drawn about the time of dissemination or the level of heterogeneity of the tumors. Further analysis of the samples, including copy number analysis and whole genome sequencing, should lead to a broader understanding of breast cancer progression from a local to a metastatic disease.

# Sammendrag

Brystkreft er en av de hyppigste årsakene til dødsfall blant kreftpasienter verden over. Den direkte dødsårsaken er nesten alltid fjernmetastaser, og kunnskap omkring den metastasiske prosessen er derfor svært avgjørende. I dag behandles brystkreft på bakgrunn av histologiske og molekylære analyser av primærtumor. Nyere studier indikerer midlertidig at kreftceller kan spres fra primærtumor mye tidligere i kreftutviklingen enn tidligere antatt, og at genetiske endringer dermed akkumulerer individuelt i metastase og primærtumor. Enkelte studier har demonstrert likheter mellom primærtumor og metastase, mens andre har funnet genetiske forskjeller. De motstridende resultatene reflekterer behovet for en dypere forståelse av de molekylære mekanismene som er involvert i metastatisk utvikling for å kunne rettlede behandlingsvalg bedre.

Målet med denne studien var å undersøke om de genetiske endringene som finnes i en primærtumor og metastase er like eller ulike. Målrettet sekvensering av DNA fra 20 primærtumorer og tilhørende lymfeknute-metastaser ble gjennomført ved hjelp av Ion Torrent Personal Genome Machine. Ion AmpliSeq™Cancer Hotspot Panel v2 ble benyttet, og 207 sekvenser fra til sammen 50 gener som ofte er mutert ved kreft, ble amplifisert og sekvensert. De genetiske variantene ble karakterisert ved å sammenlikne dem med data fra offentlige databaser slik som dbSNP, 1000 genomes og COSMIC. En eller flere somatiske mutasjoner ble detektert i 55% av svulstene (gjennomsnitt 0,8, variasjonsbredde 0-3). Genene med høyest antall mutasjoner var *TP53* (45%) og *PIK3CA* (25%). De fleste mutasjonene som ble detektert i primærtumor ble også funnet i den korresponderende metastasen, men frekvensen av varianten var i de fleste tilfeller noe ulik. For tre av pasientene ble det funnet mutasjoner som kun var tilstede i primærtumor eller metastase, alle i genet *TP53*. Resultatene indikerer at spredning kan skje ved ulike tidspunkt, og at molekylære undersøkelser av metastaser kan gi informasjon som ikke er tilgjengelig ved analyse av primærtumor alene. Ettersom bare en liten bit av svulstene ble undersøkt og bare en svært begrenset del av genomet ble sekvensert, kan det ikke trekkes noen konklusjon om når metastasen skilte lag fra primærtumor, og heller ikke omkring nivået av tumorheterogeneitet. Videre analyser av prøvene, inkludert kopitalls analyser og helgenom-sekvensering, vil gi en bredere forståelse av svulstenes genomiske endringer og om heterogenitet og evolusjonsprosesser.

## Table of Contents

## Abbreviations

| | |
|---|---|
| AI | Allelic imbalance |
| AKT | V-akt Murine Thymoma Viral Oncogene Homolog 1 |
| BAF | B Allele Frequency |
| BAM | Binary Alignment Map |
| *BRCA* | Breast Cancer, Early Onset Gene |
| *CDH1* | Cadherin 1 |
| CHP2 | Cancer Hotspot Panel version 2 |
| CIN | Chromosomal instability |
| CNA | Copy Number Alteration |
| COSMIC | Catalogue of Somatic Mutations in Cancer |
| CTC | Circulating Tumor Cells |
| dbSNP | Single Nucleotide Polymorphism Database |
| DCIS | Ductal Carcinoma *in situ* |
| ddNTP | Dideoxynucleotide Triphosphates |
| dNTP | Deoxynucleotide Triphosphate |
| DTC | Disseminated Tumor Cells |
| EGFR | Epidermal Growth Factor Receptor |
| EMT | Epithelial-Mesenchymal Transition |
| ER | Estrogen Receptor |
| FRET | Fluorescence Resonance Energy Transfer |
| GNAS | GNAS Complex Locus |
| GWAS | Genome-wide Association Studies |
| HE | Hematoxylin and Eosin |
| HER2 | Human Epidermal Growth factor Receptor 2 |
| HR | Homologous Recombination |
| IARC | International Agency for Research on Cancer |
| ICGC | International Genome Consortium |
| IDC | Invasive Ductal Carcinoma |
| IGV | Integrative Genomics Viewer |
| IHC | Immunohistochemistry |
| ILC | Invasive Lobular Carcinoma |

| | |
|---|---|
| ISP | Ion Sphere Particle |
| Ki-67 | Marker of Proliferation Ki-67 |
| LCIS | Lobular Carcinoma *in situ* |
| Log R | Log Ratio |
| LNM | Lymph Node Metastasis |
| LOH | Loss of Heterozygosity |
| NBCG | Norsk Bryst Cancer Gruppe |
| NGS | Next Generation Sequencing |
| PAM50 | Prediction Analysis of Microarray |
| PCF | Piecewise Constant Fitting |
| PCR | Polymerase Chain Reaction |
| PGM | Personal Genome Machine |
| PgR | Progesterone Receptor |
| *PIK3CA* | Phosphatidylinositol-4,5-Bisphosphate 3-kinase Catalytic Subunit Alpha |
| $PIP_3$ | Phosphatidylinositol-3,4,5-Triphosphate |
| PMP | Paramagnetic Particle |
| *PTEN* | Phosphate and Tensin Homolog |
| RTK | Receptor Tyrosine Kinase |
| SNP | Single Nucleotide Polymorphism |
| *STK11* | Serine/Threonine Kinase 11 |
| TCGA | The Cancer Genome Atlas |
| TDLU | Terminal Duct Lobular Unit |
| TNBC | Triple Negative Breast Cancers |
| TNM | Tumor Node Metastasis |
| *TP53* | Tumor Protein p53 |
| TVC | Torrent Variant Caller |
| VCF | Variant Call Format |

# 1 Introduction

## 1.1 Principles of cancer

Breaking the most fundamental rules of cell behavior by which multicellular organisms normally operate, cancer cells may be defined by two key properties: they do not respond to the signals that normally control cell growth and death, and they invade areas originally reserved for other cells. Cancer is a genetic disease, and the transition from a normal cell to a malignant cell is driven by the accumulation of changes in the cell's DNA during subsequent cell divisions. Both genetic and epigenetic alterations are thought to contribute to the phenotype of cancer, including point mutations, copy number changes, rearrangements and changes in DNA modifications such as methylations. Most of the changes are somatic, as they are not present in the normal cells of a patient, but inherited mutations in genes that control genome integrity are also of importance[1,2].

The process of life in a multicellular organism is dependent on a complex interplay between cells. Under normal conditions, the processes of genomic replication, cell growth and division are under strict control. The production and release of growth promoting and growth inhibiting signals are carefully regulated, ensuring a homeostasis of cell number. DNA repair systems work to maintain the genomic integrity of the cells, and when errors are detected, cell cycle checkpoints are activated and induce cell cycle arrest. If the damage is not to repair, the cells will be eliminated by programmed cell death[2,3].

Cancer cells acquire the ability to sustain chronic proliferation in a number of alternative ways, often involving an overproduction of growth factors that bind to cell surface receptors or by constitutive activation of different signaling pathways involved in cell proliferation. The tumor cells may also circumvent the negative control of growth inhibitory signals, by inactivating some of the essential tumor suppressor genes that function to control and limit cell growth and proliferation. The p53 protein, encoded by the gene *TP53,* is an example of a critical gatekeeper of cell cycle progression. Mutations in the gene are frequently linked to the cancer cells ability to limit or resist cell death by apoptosis[2,3].

There are restrictions for how many times a normal cell can divide. The telomers protecting the ends of chromosomes shorten progressively with each DNA replication, eventually shortened to the extent that critical genetic information will be lost if further division take place. Replicative senescence is then triggered, and the cell will stop dividing. The majority of cancer cells express the enzyme producing telomeric repeats, Telomerase, and are thus able to maintain the telomeric length and acquire replicative immortality[2,3].

However, it is not the growth of the local tumor that is the leading cause of death of cancer, it is the distant metastases. To be able to spread to other organs, the cancer cells have to invade through the basement membrane as well as being able to survive in different microenvironments. An epithelial-mesenchymal transition (EMT) seems to be an important driver of the ability of transformed epithelial cells to invade neighboring tissue. EMT is an embryonic development program exploited by cancer cells, that causes loss of intercellular adhesion and epithelial polarization as well as gain of the migratory properties of mesenchymal cells[3,4].

The establishment of metastases in secondary organs is a complex and probably a slow operation. The metastatic process is traditionally depicted as a multistep process, beginning with local invasion from the primary site to the surrounding tissue. The cancer cells may then enter the circulation by invading the blood or lymphatic vessels through the endothelial lining (intravasation), and are conveyed to new environments by the circulatory system[2,5]. These circulating tumor cells (CTCs) may eventually escape from the lumina of the vessels (extravasation) and settle in secondary organs such as lymph nodes, liver, bone and lungs, as disseminated metastatic tumor cells (DTCs)[5,6]. If the cancer cells manage to survive and continue to proliferate at the secondary site, the cells form micro-metastases and possibly macroscopic tumors (figure 1).

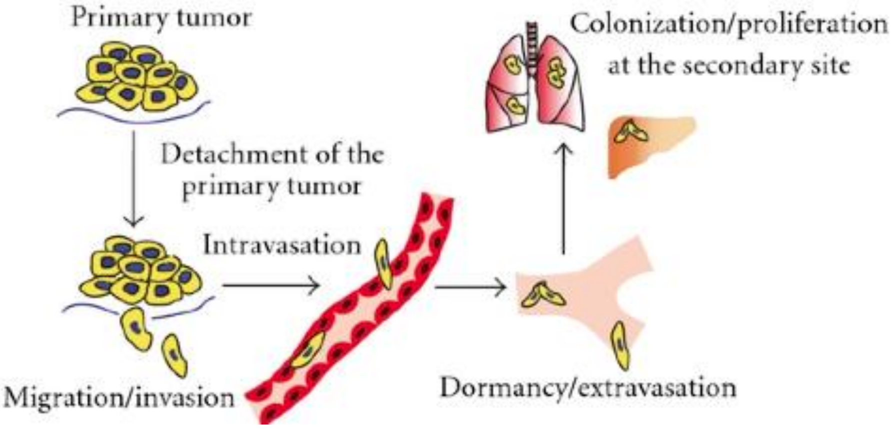

Primary tumor
Detachment of the primary tumor
Intravasation
Migration/invasion
Colonization/proliferation at the secondary site
Dormancy/extravasation

**Figure 1:** Illustration of the metastatic process: local invasion at the primary site, followed by dissemination and intravasation, survival in the circulation, extravasation and finally colonization at a secondary site[2].

Hanahan and Weinberg proposed in 2000 six hallmarks of cancer that together constitute the main biological capabilities acquired during the development of tumors[3]. In 2011 the hallmarks were revised, and the list of essential alterations for malignant growth were expanded to include 10 hallmark features of cancer: self-sufficiency in growth signaling, insensitivity to growth suppressors, tissue invasion and metastasis, induction of angiogenesis, genome instability, limitless replicative potential, ability to evade apoptosis, avoiding immune destruction, recruitment of tumor-promoting inflammatory cells that facilitate tumor progression, and reprogramming of energy metabolism to better fuel cell proliferation (Figure 2). Cancer is an incredibly diverse and complex disease, and not all cancer cells will hold the same set of hallmarks. The order and the means by which the hallmarks are acquired will vary among types and subtypes of cancer[3].

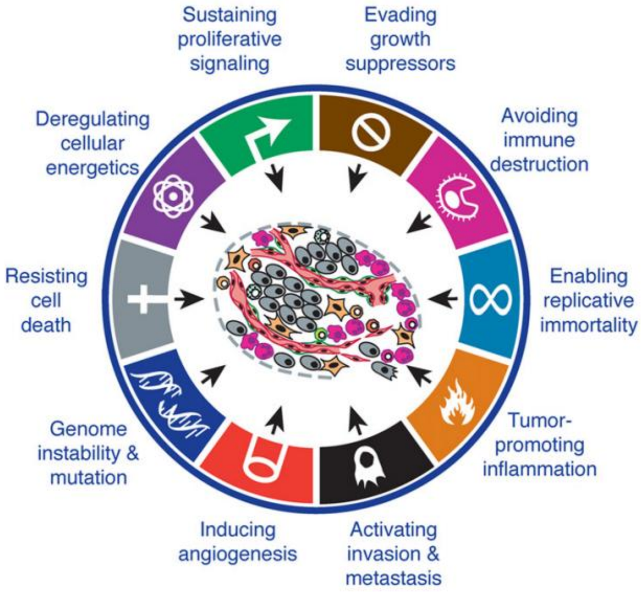

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

**Figure 2**: The 10 hallmarks of cancer as proposed by Hanahan and Weinberg. Modified from[3].

## 1.2 Breast cancer

### 1.2.1 Incidence

Breast cancer is the second most common cancer in the world, and the most frequent malignancy in females worldwide, with an estimated 1,67 million new cases diagnosed in 2012 (25 % of all cancers) and 522 000 deaths[7]. In 2013, 3220 new incidences were reported in Norway, comprising more than 20% of all female cancer incidences, and the disease caused the death of 630 women[8]. There has been a decrease in the number of breast cancer related deaths the last decades, for which early diagnosis and improved treatments are likely to have played a role. Still, based on normal life expectancy for Norwegian woman, 1 in 12 will develop breast cancer during their life time[9]. Figure 3 displays trends in incidence, mortality and survival of breast cancers in Norway.

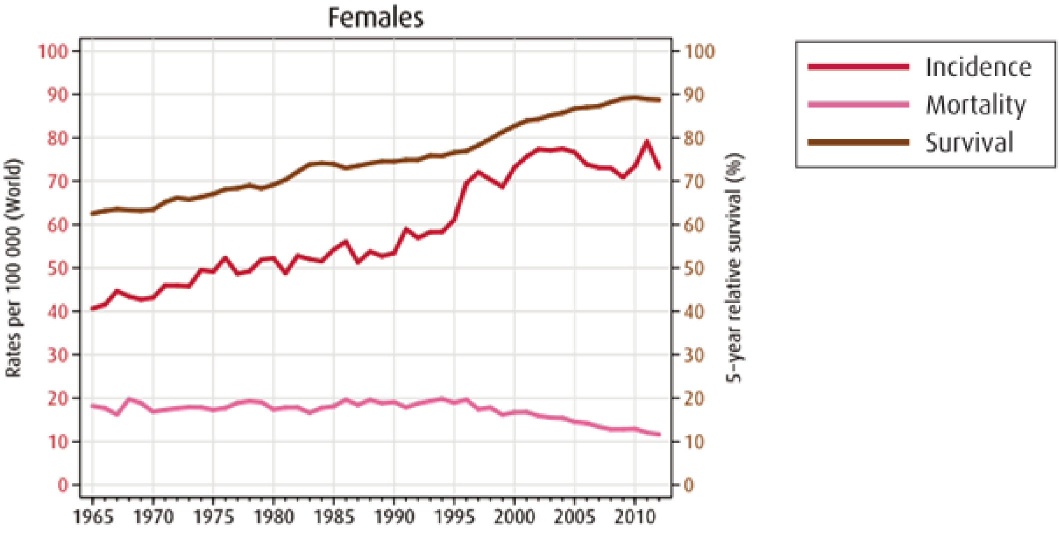


**Figure 3:** Trends in incidence, mortality and survival of breast cancer in Norway in 1965-2012[9].

### 1.2.2 Risk factors

Both genetic and environmental factors are determinants of breast cancer risk. A number of breast cancer associated genes have been identified to predispose for increased risk, but the majority of the variants are relatively common and display a low penetrance[10]. Most of the genetic variations to be found in a population are caused by single nucleotide polymorphisms (SNPs). These are single nucleotide variants that have a minor allele frequency of more than 1% in the general population. Because mutations with an unfavorable effect on an organism's fitness will be eliminated by natural selection in evolution, only neutral (or almost neutral) mutations will accumulate in the genome[11]. A particular SNP do not cause a disorder, but genome-wide association studies (GWAS) have identified several loci that may predispose for breast cancer, at most associated with a modest increase in risk[12].

Rare germ line mutations in certain tumor suppressor genes may cause a high risk of breast cancer, and mutations in six genes have been identified to have a high-penetrance; *BRCA1*, *BRCA2*, *TP53*, *PTEN*, *STK11* and *CDH1* (figure 4)[10,13]. Cells that lack the *BRCA1/2* genes are unable to sense DNA damage properly and to repair DNA damage by homologous recombination[14]. A mutation in *PTEN* will disrupt normal rates of mitosis, and leads to what is known as the Cowden syndrome[15]. *TP53* encodes one of the central proteins regulating cell cycle progression, and both germline and somatic mutations in the gene predispose to a wide specter of cancers[16]. However, all of the high-penetrance mutations are rare in the population, and familial breast cancer constitutes only 5-10% of total breast cancer. The majority of all breast cancers are sporadic and result from the acquisition of numerous somatic mutations[10,17].

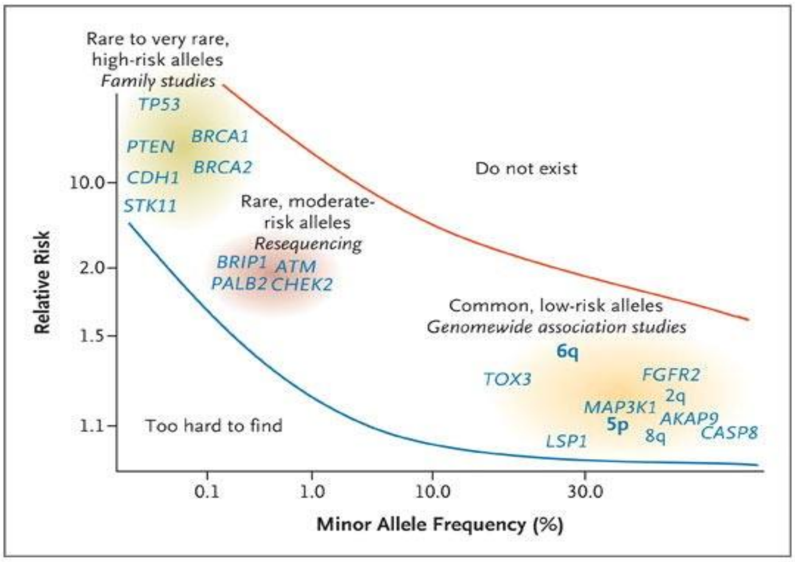


**Figure 4:** Breast cancer susceptibility loci and genes. High risk genes are highlighted in green/yellow, moderate-penetrance genes in red and low-risk genes in orange. No genes have been identified above the red line or below the blue line[15].

Non-genetic factors are also involved in the causation of breast cancer, including menstrual and reproductive history, alcohol intake, body mass index and physical activity[10]. The risk of developing breast cancer is greatly affected by hormonal influences. Menarche at an early age and a late menopause are associated with increased risk, as they both affect the total estrogen and progesterone exposure time[18,19]. Estrogen, mediated through the estrogen receptor (ER), is a steroid hormone that is an important stimulator of cell proliferation and regulator of cell differentiation in breast epithelium[14,20]. Pregnancy does also affect the risk of developing breast cancer, giving a short term increased risk, followed by a long term decrease in breast cancer risk. However, the risk of breast cancer is dependent on the age at pregnancy, as getting pregnant after the age of 35 increases the risk permanently[19,21]. The risk factors are however differently associated with specific breast cancer subtypes (see section 1.2.4.3)[22].

### 1.2.3 Anatomy of the Breast

The female breast includes 6-10 interlacing duct systems, surrounded by a dense stroma mixed with fat tissue (figure 5A). Each duct system, originating in the nipple, branches into gradually smaller ducts, which end in a terminal duct lobular unit (TDLU). The TDLU is a grape-like cluster of acini, which constitute a milk producing lobule.

The ductal system is lined by a double cell layer, an inner (luminal) epithelium and outer (basal) myoepithelium (figure 5B). The former having absorptive or secretory functions and the latter having contractile like properties that assist in milk ejection. The basal cells also produce and maintain the basal membrane[23].

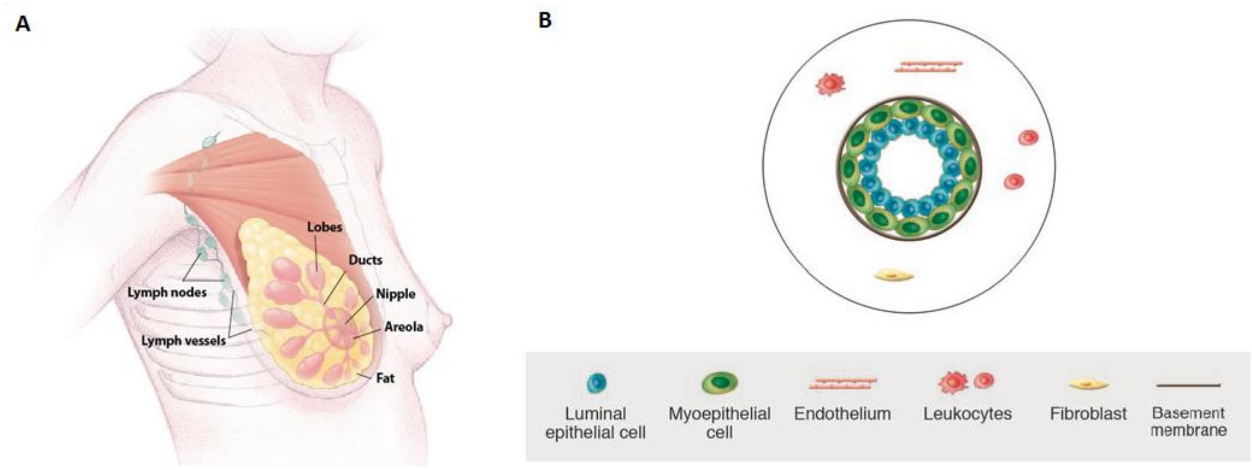


**Figure 5: A:** Anatomy of the breast[24]. **B:** Outline of the two layer duct epithelium, with luminal epithelium, the basal layer of myoepithelial cells and the surrounding basal lamina. Modified from[25].

The development of the human breast is a progressive process, starting during embryonic development. The major portion of growth occur during puberty, but the development and differentiation are not completed before the end of the first full term pregnancy and lactation period. Hence, the breast is the only human organ that is not fully developed after puberty[19,21].

### 1.2.4 Breast cancer classification

Breast cancer is a highly heterogeneous disease, and there are differences in phenotype, molecular alterations, and clinical features. Classification of the disease into subgroups is thus important to allow individualization of treatment and to predict the clinical outcome of a patient[26].

#### 1.2.4.1 Histopathological classification – type, grade and stage

Breast carcinoma arises from the epithelium of the mammary gland. The carcinomas are divided into preinvasive carcinoma (ductal carcinoma *in situ*, DCIS, or lobular carcinoma *in situ*, LCIS), or invasive carcinoma. The difference is whether the neoplastic cells have invaded through the outer myoepithelial cells and into the surrounding stroma, or not (figure 6)[23]. All together there are more than 22 histological subtypes of invasive breast carcinoma, referring to different cellular phenotypes and architectural growth patterns of the tumor, but invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) are the most frequent types[26]. Malignant transformations of stromal-, vascular- or fat cells in the breast (breast sarcomas) are extremely rare[27].

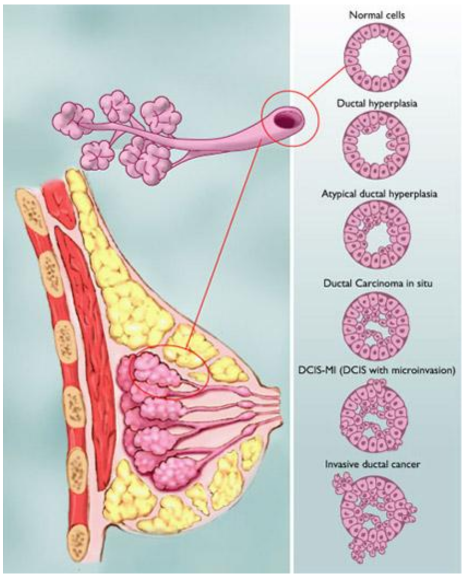


**Figure 6**: To the left: A cross section of the human breast. To the right: The development from normal cells to invasive ductal carcinoma. A cross section of a normal duct is depicted on the top, with a monolayer of epithelial cells, surrounded by the basement membrane at the external side and the lumen on the internal side. An excessive growth of normal-looking cells (ductal hyperplasia) or abnormal-looking cells (atyplical ductal hyperplasia) are benign proliferations that may be precursors of breast cancer. A malignant transformation of the epithelial cells lead to local growth of cancer cells in the lumen (ductal carcinoma in situ) and eventually an invasion into the normal surrounding breast tissue (invasive ductal carcinoma)[28].

Histological grading is a prognostic classification system applied on invasive carcinoma that takes into account three different histological features of the tumor:

- The proportion of cells with tubule formation
- The degree of nuclear polymorphism (an evaluation of size and shape of the nucleus in the tumor cells)
- The mitotic rate (how many cells with visible mitotic figures are present).

Each feature is given a score from 1 to 3, and the scores are then combined to give a grade of 1 (total score 3 to 5), 2 (total score 6 or 7) or 3 (total score of 8 or 9). Grade 3 tumors are the least differentiated, with the most aggressive phenotype and worst prognosis[26].

In the clinical setting, the breast carcinomas are also staged based on the size of the primary tumor as well as on the extent the cancer has spread in the body. Three parameters together constitute a TNM-status: tumor size (T), presence and extent of cancer cells in lymph nodes (N) and presence of distant metastatic sites (M)[26].

#### 1.2.4.2 Molecular markers

The molecular markers used in clinical practice world-wide are estrogen receptor (ER), progesterone receptor (PgR) and human epidermal growth factor receptor 2 (HER2). ER, PgR and HER2 expression status are usually determined by the use of immunohistochemistry (IHC), and are utilized to categorize breast cancer into different therapeutic groups: the hormone receptor positive group, the HER2 amplified group and the triple negative breast cancers (TNBCs)[26,29].

ER and PgR are ligand-activated transcription factors that stimulate the growth breast epithelium when bound by estrogen and progesterone respectively[26,30]. In normal breast, ER expression is restricted to a small subset of luminal cells[5]. However, ER expression is found to be elevated in a large proportion of cancerous cells, and about 80% of all breast carcinomas are dependent on estrogen and a functional estrogen receptor for growth. These tumors are thus abbreviated ER-positive breast carcinomas. Approximately 40% of these ER-positive tumors are PgR-negative[26].

Growth factor receptors play an essential role in both proliferation and cell survival. In breast cancer biology, the HER2 have been studied the most, which is a member of the HER family of transmembrane receptor tyrosine kinases (RTK). About 15% of invasive breast cancers have HER2 gene amplification and/or protein over expression which is associated with accelerated cell growth and proliferation[14,29]. Approximately 10-15% of breast cancers are negative for all of these three receptors, so-called triple-negative breast cancers (TNBC)[26].

#### 1.2.4.3 Molecular classification

The rapid development of high throughput technology, first by genomic and expression microarray technology and later by next generation sequencing, has made it possible to study molecular alterations in cancer cells in much more detail, and an improved taxonomy of cancer have been proposed. By analyzing the expression patterns of around 550 genes that displayed the greatest variation between different patients and the least variation between samples from the same patients, Perou et al. classified invasive breast carcinoma into different biological subgroups, often referred to as the intrinsic subtypes of breast cancer. Five subtypes were identified: luminal A, luminal B, HER2-enriched, basal-like and normal-like[31]. The classification has revealed differences in incidence, survival and response to treatment[31–33].

The luminal A group is the most common subtype, representing 50-60% of all breast cancers. It is characterized by strong hormone receptor positivity, low proliferation rates, negative HER2 and a low histological grade. These tumors are associated with a lower relapse rate and an improved prognosis than the other intrinsic subtypes. The luminal B tumors display a more aggressive phenotype, being of a higher histological grade, having hormone receptor positivity of a varying degree and a high proliferation rate[29,34].

Relative to luminal A, patients with HER2-enriched or Basal-like subtypes display a poor outcome, and they are both highly proliferative. HER2- enriched tumors have HER2 protein overexpression and/or HER2 gene amplification. The basal-like tumors are often triple-negative, and have a high frequency of *TP53* mutations. The normal-like tumors account for only 5%-10% of breast carcinomas. These tumors are frequently classified as triple negative, but they are otherwise poorly characterized. There are doubts about their existence as a true breast cancer subtype[34].

A gene expression assay named PAM50 (Prediction Analysis of Microarray), has further been developed and validated for robust classification of intrinsic subtypes[35]. In total, 50 genes were included and a standardized method for classification was developed. The PAM50 assay also generates a risk of relapse score that predict a patient's probability of disease recurrence. The development of PAM50 for use in predictive analysis is a significant contribution to prognostic and predictive analysis[34,35].

### 1.2.5 Treatment

In Norway, guidelines for diagnosis, treatment and follow up of breast cancer patients are given by NBCG (Norsk Bryst Cancer Gruppe), and are included in the guidelines provided by the Norwegian directorate of Health guidelines (www.nbcg.no). The primary treatment of breast cancer is surgery, aiming to gain local control of the disease. This may be either mastectomy (complete removal of the breast) or lumpectomy (breast conserving surgery) combined with radiation therapy[36].

In addition to surgery and radiation therapy, systemic treatment may be given either adjuvant (after surgery) to minimize the risk of recurrence or neoadjuvant (before surgery) to shrink large tumors to a size possible to operate. TNM-status, tumor stage, hormone receptor status, HER2 status, Ki67 expression (to estimate level of proliferation) and menopausal status are all decisive for choice of treatment[36,37].

Several different types of systemic treatment can be offered breast cancer patients in accordance with the guidelines. Different classes of chemotherapeutics are available, all causing cell death by apoptosis by interfering with processes involved in cell division[38]. Targeted therapy binds and inhibits a specific molecular target, crucial for maintaining a proliferative pathway in the cancer cell. In the case of ER-positive cancers, the patients are offered hormone therapy in the form of tamoxifen or aromatase-inhibitors. Tamoxifen is an estrogen antagonist that inhibits ER activation. Aromatase inhibitors inhibit the enzyme responsible for estrogen synthesis in post-menopausal women[39]. Another targeted therapy is the monoclonal antibody trastuzumab (Herceptin) given to patients with HER2-positive tumors. Trastuzumab is an antibody that blocks the dimerization of HER2 receptors, inhibits their kinase activity and reduces cell proliferation. Other agents against HER2 are also available[14,37].

### 1.2.6 Genetic alterations in breast cancer

#### 1.2.6.1 Somatic point mutations

There are several types of somatic mutations in cancer, including substitutions of one base by another and insertions or deletions of small segments of DNA[1]. Alterations of nucleotides within the coding region of a gene may lead to an amino acid substitution (missense mutation), a premature stop codon and a truncated gene product (nonsense mutation), a change within the splice-recognition site (splice-site mutation) or no change at all (silent mutation). Insertions or deletions of a small number of bases (indels) may result in a frameshift mutation and an abnormal protein product[40].

The somatic mutations in the genome of cancer cells may be classified into “driver” mutations, which confer growth advantage and are important for the process of carcinogenesis, and “passenger” mutations with no clear functional consequence. It is likely that most cancers carry more than one driver mutation, and that the number differs between cancer types[1]. Two major groups of driver genes are frequently altered in human tumors. Dominant cancer genes, known as proto-oncogenes, encode proteins that normally enhance cell division or inhibit cell death. These genes require only one of the two parental alleles to be mutated, and the resulting protein will usually be constantly activated. Recessive cancer genes, known as tumor suppressor genes, encode proteins that normally limit cell division or promote cell death. These genes need mutations or inactivation of both parental alleles, and may result in an elimination or inactivation of the protein[41].

Somatic mutations may be caused by different mechanisms, such as defective DNA repair, enzymatic modifications of DNA, inaccurate DNA replication or the exposure to mutagens of both internal and external origin. Different mutational mechanisms have been found to generate different combinations of mutation types, which can be detected as a mutational “signature”[42]. Genome-wide profiling of somatic mutations in breast cancer have demonstrated a great variation in numbers and types of mutations between tumors, indicating that the mutational processes that generate these genomic landscapes may vary[43,44]. In most cancers, the mutations are caused by more than mutational mechanisms.

Several genes are identified as recurrently mutated genes in breast cancer, where *TP53* and *PIK3CA* are the two most frequently mutated genes. The majority of *TP53* mutations lead to a single amino acid change in the central region of the p53 protein, generating variants that to various extents have lost their tumor-suppressive functions[45]. *PIK3CA* is a proto-oncogene encoding the catalytic subunit (p110α) of the PI3K enzyme. When activated by receptor tyrosine kinases (RTKs), PI3K catalyzes the phosphorylation of inositol lipids to phosphatidylinositol-3,4,5-triphosphate ($PIP_3$) in the cell membrane. $PIP_3$ is an important lipid second messenger activating AKT and interfere with other pathways resulting in inhibition of apoptosis and promotion of cell growth, cell motility and proliferation. Cancer-associated *PIK3CA* gene mutations result in production of an altered p110α subunit that thus allows increased PI3K signaling and abnormal proliferation of cells[46]. Many genes display subtype-specific patterns of mutation, as they are more diverse in luminal A and luminal B tumors, than within basal-like and HER2-enriched subtypes (figure 7)[43].

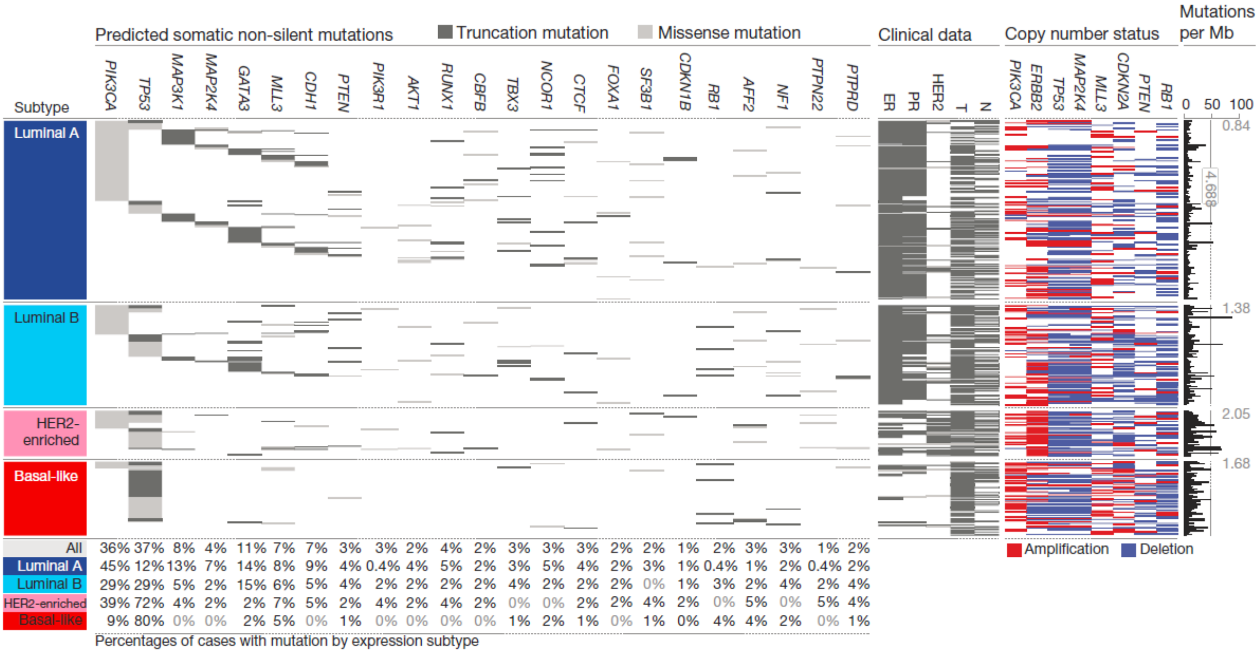


**Figure 7:** Recurrently mutated genes in luminal A, luminal B, HER2-enriched and basal-like breast cancers and correlation with genomic and clinical features. The panel on the left shows patterns of non-silent somatic mutations and frequencies of the mutated genes in the different subtypes. The middle panel shows clinical features; receptor status (ER, PgR and HER2), tumor size (T) and node status (N). The dark grey color indicates positive or T2-4, white indicates negative or T1, and light grey indicate that the information was not available. The right panel shows genes with frequent copy number amplifications (red) or deletions (blue). The diagram on the far right shows the rate of non-silent mutations per tumor sample (mutations per megabase) as well as the average mutations rate for each subtype[43].

### 1.2.6.2 Ploidy and copy number alterations

Ploidy is a measure of the total genomic DNA content in a cell. A normal human somatic cell is diploid, and contains two sets of all 23 chromosomes. When cancer cells are rapidly dividing, mistakes in the distribution of chromosomes may occur due to several defects such as spindle attachment defects, multipolar spindles, defects in chromosome cohesion or impairment of the mitotic checkpoint response. Abnormal chromosome content – also known as aneuploidy – is thus a common feature of cancer cells. Many cells display chromosomal instability (CIN), as they frequently lose and gain whole chromosomes or parts of chromosomes, during divisions[47].

Copy number alterations (CNAs) occur frequently in breast cancer and define important genetic events driving tumorgenesis. CNAs are changes in the structure of the chromosomes, including amplifications, duplications, deletions, inversions and translocations. The copy number of a DNA sequence may increase from the two copies normally present in a diploid cell, up to several hundred copies. Copy number reduction may on the other hand lead to the loss of a DNA sequence from the cancer genome[1]. Frequently observed CNAs in breast cancer include gain of chromosomal regions on 1q, 8q, 17q and 20q, and loss of regions on 1p, 8p, 13q, 16q and 17p. Known oncogenes and tumor suppressor genes such as HER2, EGFR, BRCA1/2 and TP53 reside in these regions[48,49].

The mechanisms that contribute to the different types of CNAs are only partly uncovered, and double strand breaks of the DNA and erroneous replications seem to be important factors. The strategies used to solve the errors are important for which type CNA that take place, and three general mechanisms have been proposed that can explain the majority of the structural changes: homologous recombination (HR), non-replicative non-homologous recombination and replications based repair mechanisms[50].

## 1.3 Tumor heterogeneity

All cancers are believed to derive from a single cell that starts to behave abnormally, and increased rates of cell division can explain the acquisition of somatic mutations in the genome. In the classic view of cancer development, some of these genetic aberrations will give certain cells a selective advantage to undergo clonal expansion, and the fittest clone will eventually outgrow the other cells and come to dominate the cellular composition[2,51]. However, studies using high-resolution sequencing technology have demonstrated that there is extensive variation not only between tumors (intertumor heterogeneity) but also within tumors (intratumor heterogeneity). Yates et al. studied the spatial distribution of subclones in breast carcinomas by sequencing 8 needle biopsy samples from distinct quadrants in each of 12 primary tumors. In 10 out of 12 carcinomas at least one mutation was found to be present in only a small section of the tumor[52]. By performing single cell sequencing Navin and colleagues demonstrated that some breast carcinomas are composed of multiple genetically divergent subclones[53]. All together, separate regions of the same tumor may contain subclones with different somatic mutations, gene expression signatures, DNA ploidy and copy number changes[52,54,55].

Different models explaining intratumor heterogeneity are proposed (figure 8). The clonal evolution model, presented by Nowell in 1976, suggests that tumors evolve from one (monoclonal) or several (polyclonal) subpopulations. In this model all clones have the potential to proliferate and are subjects for natural selection. In the cancer stem cell model, there is a hierarchical organization within the tumor. Only a small fraction of the cells initiate tumor progression, and may give rise to different subpopulations within the tumor. Tumors showing a high degree of intratumor heterogeneity are thought to follow the mutator phenotype model. This model suggests that tumors evolve by a gradual and random accumulation of mutations as the tumor grows, leading to a high degree of intratumor variation, rather than clonal subpopulations[56].

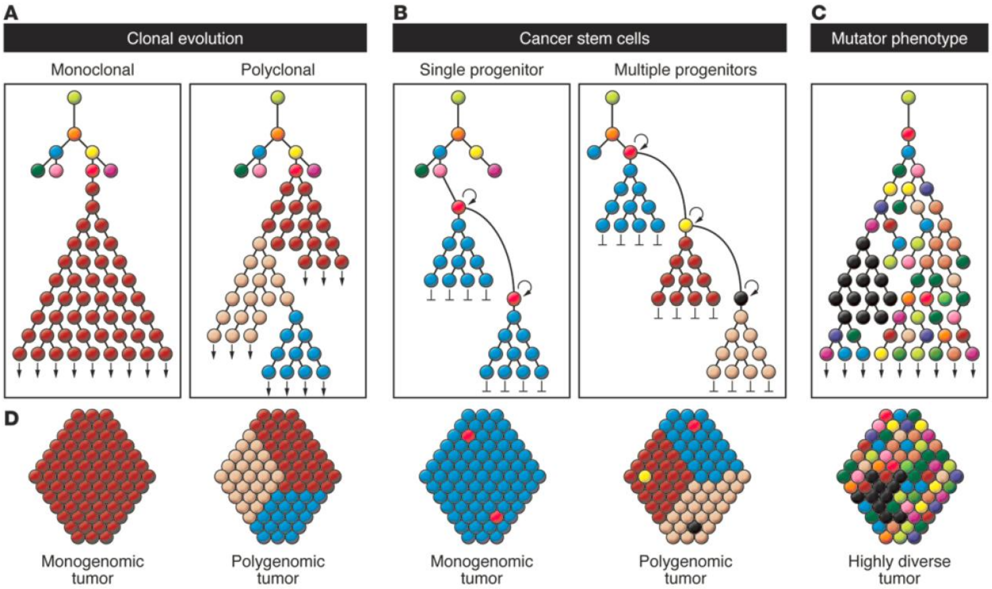


**Figure 8:** Three hypothetical models for tumor evolution explaining intratumor heterogeneity: The clonal evolution model (**A**), the cancer stem cell model (**B**) and the mutator phenotype model (**C**). Different subpopulations of tumor cells (**D**) will result from the distinct models[56].

The concept of intratumor heterogeneity has important implications for both diagnosis and disease management. As genetic alterations may be found in only a fraction of the tumor cells, the genetic information extracted from single tumor-biopsy samples may underestimate the genetic diversity of tumors as a whole, and the reservoir of different cells from which resistant tumor cells can be selected are probably very large.

## 1.4 Tumor progression and metastatic disease

Metastatic disease is the cause of most deaths from cancer, and the prevention of the spread of tumor cells from the local tumor is of great clinical attention. Acquiring more accurate knowledge about the process of dissemination is thus of importance. With the results from genomic analysis of thousands of cancer genomes and advances in molecular techniques, the traditional model of breast carcinoma progression, saying that tumor cell dissemination occurs late in tumor development, has been challenged. Increasing evidence indicates that tumor cells may spread to distant sites much earlier than previously believed[57,58].

Cancer progression can be explained by two basic models, referred to as the linear progression model and the parallel progression model (figure 9). The first model is based on the stepwise progression of the tumor, where cancer cells in the primary tumor pass through successive rounds of mutation and selection. Only after an accumulation of a significant number of genomic and epigenetic changes, the cancer cells may achieve metastatic potential and be able to leave the primary site and start growing in a new environment. Once the disseminated tumor cells have adapted to a distant site and formed a microscopic metastasis, this tumor can then generate secondary metastasis. This model predicts that metastases will be genetically similar to the primary tumors from which they descend[58,59].

In the parallel progression model individual cells may confer metastatic potential early in cancer development, and the tumor cells may depart the primary lesions before they have acquired a completely malignant phenotype. The metastatic cells will still be evolving, and there is a parallel and independent accumulation of genetic and epigenetic alterations in the primary tumor and the metastasis leading to greater molecular divergence between the two[58,59].

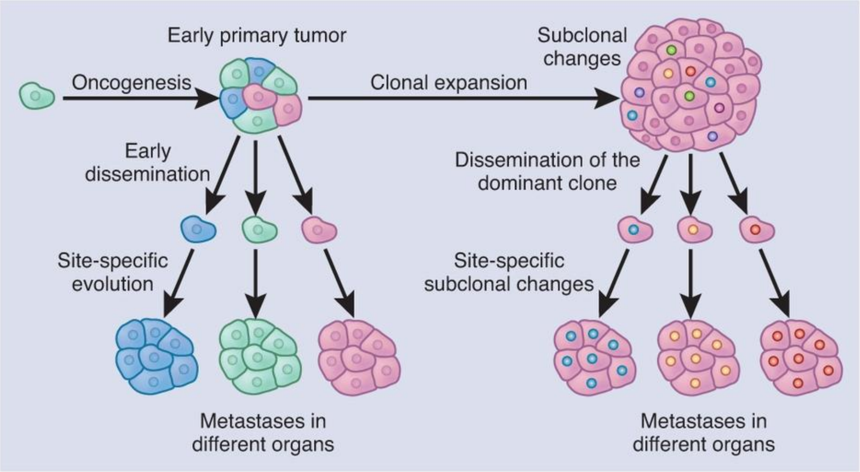

Early primary tumor
Oncogenesis
Clonal expansion
Subclonal changes
Early dissemination
Dissemination of the dominant clone
Site-specific evolution
Site-specific subclonal changes
Metastases in different organs
Metastases in different organs

**Figure 9:** The linear and parallel metastatic progression model. Top and to the left: In parallel progression, tumor cells may disseminate from the primary tumor at an early stage of tumor progression and a parallel and independent accumulation of genetic and epigenetic changes occur at the primary and metastatic site. Top and to the right: In linear progression, the cancer cells pass through successive rounds of mutation and selection in context of the primary tumor before metastatic potential is achieved and dissemination to distant organs occur. The metastases will be genetically similar to their corresponding primary tumor[4].

Neither of the two models is supported by direct and indisputably evidence and it might be that both exist either independently but also in combination. There is a well known association between tumor size and frequency of metastasis, which may support the concept that only tumor cells disseminated late in the tumor progression process have the possibility to form metastases. However, studies of breast cancer growth rates indicate that there is a correlation between the growth rates of primary tumors and metastases[59,60]. In cases where the patient has a metastasis present at the time for diagnosis, and in cancers with no detectable primary tumor, the metastatic growth rates would have to far exceed that of the primary tumor if linear progression has occurred. The finding of DTCs in the bone marrow of patients with ductal carcinoma in situ may also be an indication of an early dissemination of tumor cells[61].

Comparative data from primary tumors and metastases have been collected through several studies. Some demonstrate a close correlation between primary breast tumors and corresponding metastases[62–64], while others display genetic divergence[58,65,66]. However, the concept of intratumor heterogeneity complicates an interpretation of such differences, as a metastasis may be derived from a minor sub-clone of the primary tumor, not represented in the part of the primary tumor selected for analysis.

Today, risk allocations and treatment recommendations targeting breast cancer metastases and disseminated tumor cells are largely based on characteristics of the primary tumor. In order for the systemic treatments to be successful, metastases should be genetically similar to the primary carcinoma. If the dissemination of cancer cells is an early step in tumor progression and a parallel evolution occurs, adjuvant therapies targeting events present in the primary tumor, may have a low success rate in eradicating metastatic cancer cells[67].

## 1.5 Mutation detection by Next generation Sequencing

### 1.5.1 DNA sequencing

Throughout the last half-century several technologies have been developed to characterize the genomic abnormalities found in cancer cells (figure 10). The emergence of DNA sequencing revealed tremendous information about the structure of cancer genomes, and enabled the discovery of several genes involved in the process of tumorgenesis[40,41]. Genome-based diagnosis of cancer is also of increasing importance in the treatment of the disease[68]. Worldwide collaborative efforts are being made to register the genomic landscape of thousands of cancer genomes. The International Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) are both large scale studies cataloguing genomic alterations in cancers, and the genomic data are today publically available[68,69]. Various sequencing technologies have been developed, and before going into the sequencing method utilized in this thesis, a quick review of the history of DNA sequencing will be presented.

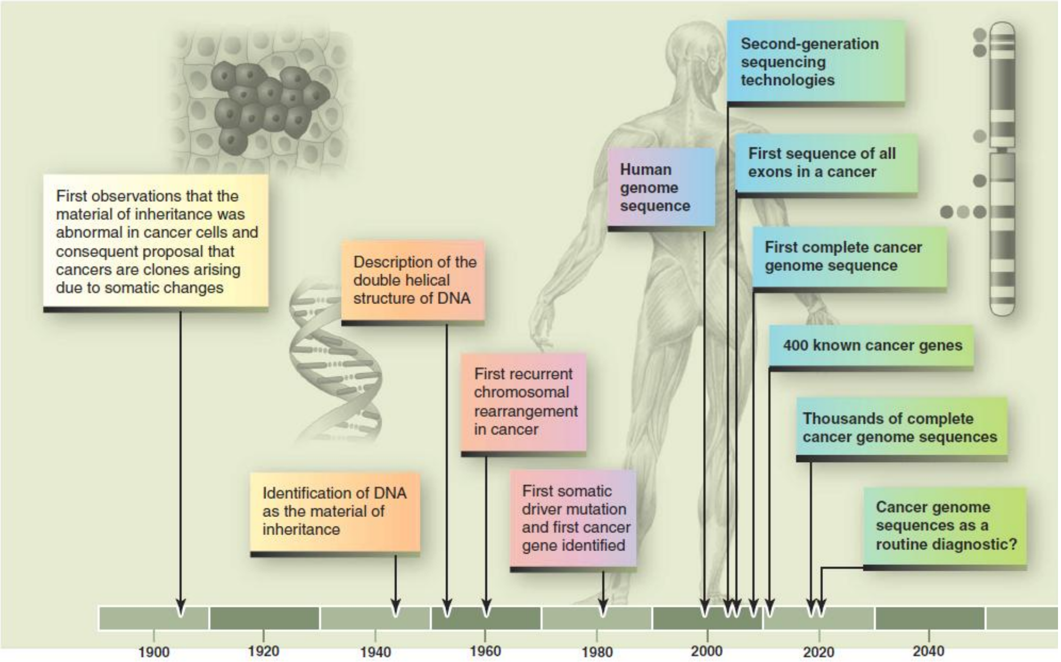

Second-generation sequencing technologies
First sequence of all exons in a cancer
Human genome sequence
First complete cancer genome sequence
First observations that the material of inheritance was abnormal in cancer cells and consequent proposal that cancers are clones arising due to somatic changes
Description of the double helical structure of DNA
400 known cancer genes
First recurrent chromosomal rearrangement in cancer
Thousands of complete cancer genome sequences
Identification of DNA as the material of inheritance
First somatic driver mutation and first cancer gene identified
Cancer genome sequences as a routine diagnostic?
1900
1920
1940
1960
1980
2000
2020
2040

**Figure 10**: Time line displaying key events in cancer genome research[41].

In 1977, Sanger and his colleagues developed one of the first methods to sequence DNA, commonly referred to as Sanger sequencing, which became the primary technology in the "first generation" of DNA sequencing[70]. The method utilizes dideoxynucleotides (ddNTPs) as chain terminators during DNA replication. Single-stranded DNA are split into four aliquots and mixed with primers, DNA polymerase, the four deoxyribonucleotide triphosphates (dNTPs) and a replication terminator. Nucleotide specific terminated fragments for each of the four nucleotides are constructed, and by separating the fragments by size on a gel, the DNA sequence may be read out. The method became more efficient in the following years, partly by the introduction of fluorescently labeled ddNTPs, and the automated capillary sequencing instrument introduced by Applied Biosystems in the 1990s[68,71,72]. The advances in the Sanger sequencing technique enabled the completion of the first human genome sequence in 2004. However, the Human Genome Project was extremely time consuming and resource intensive, and faster and cheaper technologies with higher throughput were required[73].

Beginning in 2005, massive parallel sequencing, or Next Generation sequencing (NGS), methods emerged[73]. A variety of different technologies have been developed, all being able to sequence thousands to many millions of DNA fragments simultaneously. The methods are based on sequencing by synthesis, as the DNA synthesis and detection of sequence is done at the same time. The DNA to be sequenced is used to construct a library of fragments, and synthetic DNA adapters are ligated to each end of the DNA strands. The fragments are amplified onto a solid surface, specific to each platform: a bead or a glass slide. Various methods are used to detect the sequence, all consisting of a stepwise reaction including a nucleotide addition step, a detection step and a wash step to remove unmatched bases[71].

The first NGS machine was introduced by 454 in 2005 (purchased by Roche in 2006) followed by the release of the Genome Analyzer by Solexa (purchased by Illumina in 2007) and SOLiD provided by Agencourt (purchased by Applied Biosystems in 2007). In recent years, the sequencing industry has been dominated by Illumina with the Illumina HiSeq instrument as the current market leader[70,74]. There are advantages and disadvantages associated with all of the NGS platforms, including different costs per Mb, read-length, sample preparation time and different overall error rate (table 1)[75].

**Table 1:** Overview of some of the major NGS-platforms. Modified from[76].

| Company | Platform | Amplification method | Sequencing method | Read length | Max output/time per run | Dominant error type | Overall error rate | Cost per run (in 2011) |
|---|---|---|---|---|---|---|---|---|
| Roche/454 Life Sciences | 454 GS FLX+ | emPCR | Pyrosequencing | 700 bp | 700 Mb/10-23 h | Indel | 0,50 % | $6200 |
| | 454 GS Junior | emPCR | Pyrosequencing | 400 | 35 Mb/10h | Indel | 0,50 % | $1100 |
| Illumina | Illumina HiSeq 2000 | Bridge PCR | Sequencing-by-synthesis with reversible terminator | 150 bp | ≤300 Gb/ 2,5-11 days | Substitution | 0,20 % | $20120 |
| | Illumina MiSeq | Bridge PCR | Sequencing-by-synthesis with reversible terminator | 250 bp | >1 Gb/4-27 h | Substitution | 0,20 % | $750 |
| Life Technologies/ Applied Bioscience | SOLiD 4 System | emPCR | Sequencing by ligation | 50 bp | 100 Gb/ 3,5-16 days | Substitution | 0,10 % | $8128 |
| Life Technologies/ Ion Torrent | Ion PGM Sequencer (318 chip) | emPCR | Ion semiconductor sequencing | 200 bp | 10 Gb/2-4 h | Indel | 1 % | $925 |
| | Ion Proton Sequencer (Proton I chip) | emPCR | Ion semiconductor sequencing | 200 bp | 1 Gb/ 0,9-4,5 h | Indel | 1 % | NA |
| Pacific Biosciences | PacBio RS | None | Single molecule sequencing | 10 kb | NA | Indel | 15 % | NA |

Together these new technologies have increased the capacity and affordability of sequencing. The NGS methods provide a much more comprehensive picture of the cancer genome than previously available methods. By aligning the reads to a reference genome, the major alterations found in the cancer genome can be detected, including point mutations, copy number alterations and chromosomal rearrangements (figure 11). To accurately call bases, the depth of coverage is important. This is a measure of the number of reads covering each base, i.e. how many times a single base is sequenced. For cancer samples, the coverage needs to be increased to account for the heterogeneity (including decreased purity due to normal cells) of a sample[68,69].

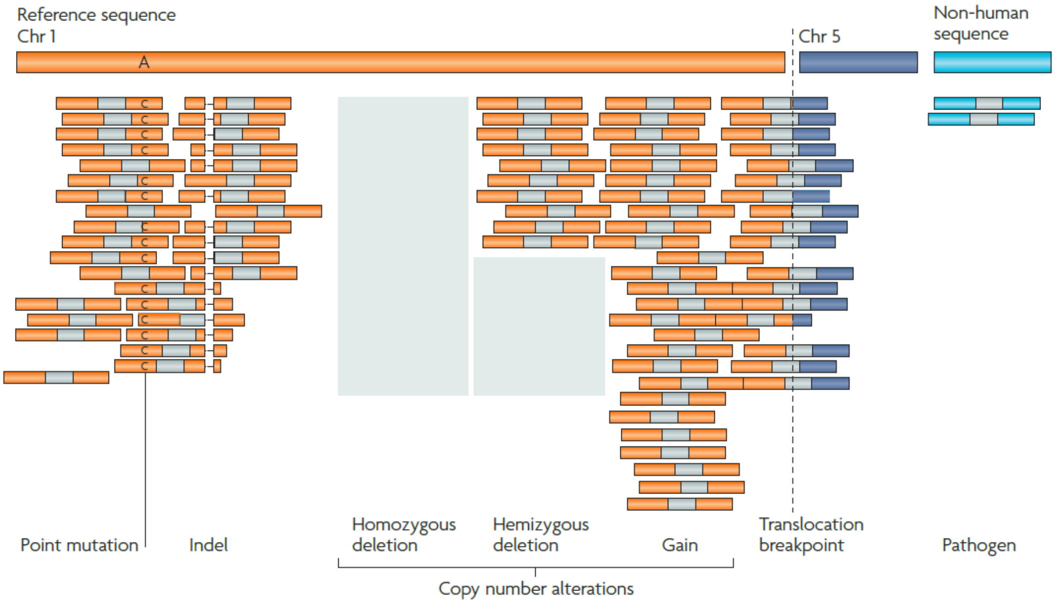
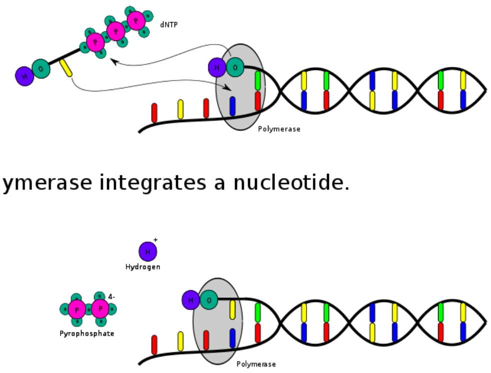


**Figure 11:** Genome alterations that can be detected with next generation sequencing. Sequenced fragments are illustrated as bars where the colored tips represent the sequenced ends and the grey center section represent the unsequenced part of the fragment. The color of the reads indicates which chromosome they align to. Different genomic alterations can be recognized: point mutations (A>C in this example), indels (a deletion is shown by a dashed line in the picture), copy number alterations (shaded boxes represent absent or decreased number of reads and another region has more reads than expected), rearrangements (the paired ends that align to different chromosomes), and the presence of genomic material from pathogens (fragments that map to non-human sequences)[68].

### 1.5.2 Ion Torrent Personal Genome Machine

In 2010 Ion Torrent released the small and compact benchtop sequencer, named the Ion Personal Genome Machine (PGM). It features a short run time but limited data throughput and are primarily made for clinical applications and small labs[70]. The Ion Torrent technology sequences the DNA by monitoring pH change: the hydrogen ions that are released during nucleotide incorporation are detected (figure 12)[70,71].



Polymerase integrates a nucleotide.

Hydrogen and pyrophosphate are released.

**Figure 12:** The incorporation of a dNTP into a growing DNA strand causes the release of a hydrogen ion and pyrophosphate[113].

The basic element of the PGM is a semiconducting chip, which consists of millions of wells with transistor sensors beneath[70]. The sequencing process begins with the construction of a DNA sequencing library and the ligation of an adapter sequence containing a barcode to each DNA fragment (figure 13). The adapter sequence makes it possible to attach the fragments to beads called ion sphere particles (ISPs). To be able to run multiple libraries on a single chip, each library must be assigned a unique barcode.

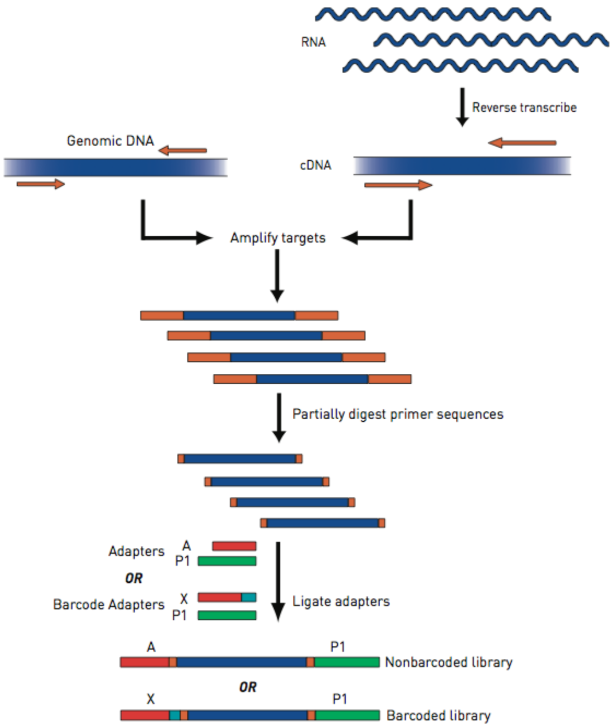


**Figure 13:** The steps to be performed when constructing DNA libraries for Ion Torrent sequencing. DNA targets of interest are amplified by the use of PCR, the primer sequence is partially digested and barcoded adapters are ligated to the ends of the fragments. Reprinted from the Ion Amplisec$^{TM}$ protocol for DNA and RNA library preparation.

By the use of emulsion PCR, the templates are clonally amplified and attached to the ISPs. The templated ISPs, together with DNA Polymerase, are then loaded on the chip, and the beads are deposited into different wells. The chip is flooded with one of the four nucleotides. If a nucleotide is incorporated into the DNA strand, a hydrogen ion is released. The hydrogen ion will change the pH of the solution in the well, and the transistor sensor beneath the well measures the change in pH and converts it to voltage. The process is repeated every 15 seconds with a different dNTP flooding the chip. If the nucleotide is not correct, no voltage will be found. If there are two identical bases next to each other, two nucleotides will be incorporated, and a double voltage will be detected (figure 14)[77].

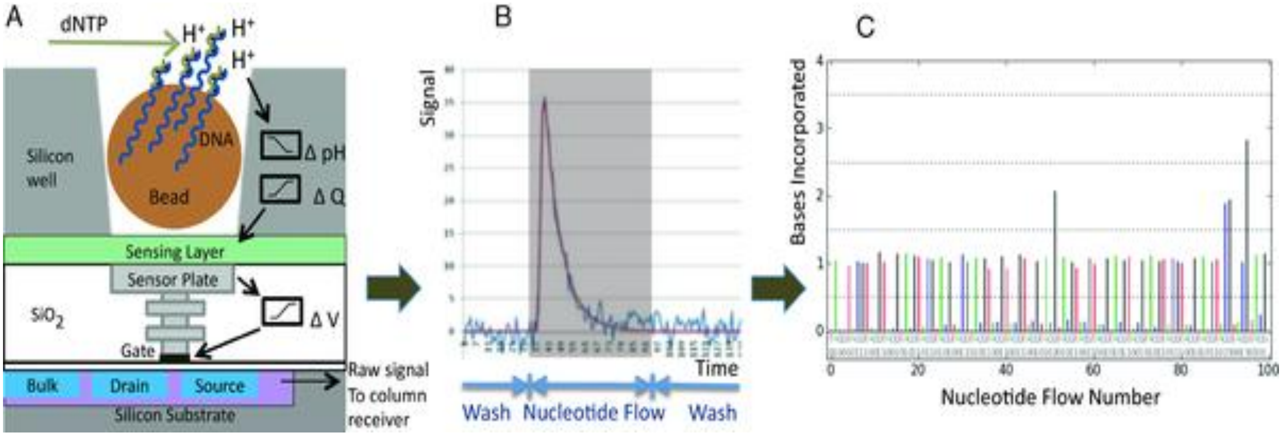


**Figure 14:** Overview of the progress in ion torrent sequencing. (**A**) The templated Ion Sphere Particle (ISP) is deposited in a well on the sequencing chip, above the pH sensor plate. The clonally amplified DNA is single-stranded and bound to a primer and polymerase. One of the four dNTPs flows into the well, and $H^+$ is released if the base is incorporated into the growing DNA strand. The $H^+$ release changes the pH in the well, and the charge build-up is transmitted as a voltage change at the transistor gate. (**B**) The pH signal from an individual sensor well, with the extracted net signal (red line) and background corrected data (blue line). (**C**) The first 100 flows from one well. Each colored bar indicates the number of incorporated bases during a nucleotide flow[77].

# 2 Aim of study

To be able to improve outcome of breast cancer and develop effective treatment strategies, increased knowledge about the biology underlying the metastatic process is necessary. The aim of this study was to compare genomic changes in primary tumors with changes in a corresponding lymph node metastasis to investigate patterns of evolution of breast cancer. A main objective was to establish targeted sequencing of selected genes and apply bioinformatical pipelines for variant annotation followed by comparison of results from individual patients paired samples (tumor and metastasis).

# 3 Material

## 3.1 OsloVal and PriMet

The patient material used in the study was part of the PriMet cohort, which is a subset of tumors from the OsloVal cohort. The study was approved by the regional ethical committee (REK Sør-Øst C, approval 2010/498).

The OsloVal cohort, consisting of fresh-frozen tumors from 184 breast cancer patients, was collected at the Norwegian Radium Hospital from 1981 to 1999. The majority of the tumor samples were excess material that was stored in a biobank (at -80°C) after routine hormone receptor analysis. The tumors have been characterized by CNA profiles and SNP genotypes, as well as clinical annotation including long term follow-up. The cohort was first utilized as a validation dataset in the Sage Bionetworks DREAM Breast Cancer Prognosis Challenge, an open challenge aimed to build computational models that predict breast cancer survival[78].

Among the 184 patients, 20 had in addition to tissue from the primary tumor, tissue from a matched lymph node metastasis available. All together, a total of 44 samples have been used for analysis in this thesis, consisting of:

- 20 primary tumors
- 20 lymph node metastasis
- 1 benign tumor
- 3 blood samples

The demographic data for the cohort is presented in table 2.

**Table 2:** Demographic data for the 20 patients in the PriMet cohort, each providing a primary tumor and a lymph node metastasis (LNM).

| | | **No. of patients** | **%** |
|---|---|---|---|
| **Age (mean, min-max)** | | 59,7 (40,1-81,2) | |
| | | | |
| **Location primary tumor** | | | |
| | Right | 5 | 25 % |
| | Left | 15 | 75 % |
| | *Not available* | 0 | 0 % |
| **Location LNM** | | | |
| | Primary – ipsilateral axillary metastasis | 17 | 85 % |
| | Recurrence – ipsilateral axillary metastasis | 1 | 5 % |
| | Contra lateral regional lymph node | 2 | 10 % |
| | *Not available* | 0 | 0 |
| **ER status primary tumor** | | | |
| | Positive | 9 | 45 % |
| | Negative | 9 | 45 % |
| | *Not available* | 2 | 10 % |
| **ER status metastasis** | | | |
| | Positive | 6 | 30 % |
| | Negative | 8 | 40 % |
| | *Not available* | 6 | 30 % |
| **PgR status primary tumor** | | | |
| | Positive | 10 | 50 % |
| | Negative | 8 | 40 % |
| | *Not available* | 2 | 10 % |
| **PgR status metastasis** | | | |
| | Positive | 6 | 30 % |
| | Negative | 7 | 35 % |
| | *Not available* | 7 | 35 % |
| **Tumor size** | | | |
| | T1 | 5 | 25 % |
| | T2 | 4 | 20 % |
| | T3 | 4 | 20 % |
| | T4 | 7 | 35 % |
| | *Not available* | 0 | 0 % |
| **Lymph node status** | | | |
| | N0 | 1 | 5 % |
| | N1 | 9 | 45 % |
| | N2 | 3 | 15 % |
| | N3 | 6 | 30 % |
| | *Not available* | 1 | 5 % |
| **Distant metastasis** | | | |
| | M0 | 18 | 90 % |
| | M1 | 2 | 10 % |
| | *Not available* | 0 | 0 % |
| **Treatment, surgery** | | | |
| | Mastectomy + axillary dissection | 18 | 90 % |
| | Mastectomy only | 1 | 5% |
| | Resection | 1 | 5 % |
| | *Not available* | 0 | 0 % |
| **Treatment, chemotherapy** | | | |
| | No chemotherapy | 7 | 35 % |
| | Adjuvant | 6 | 30 % |
| | Neo-adjuvant | 6 | 30 % |
| | *Not available* | 1 | 5 % |
| **Treatment, hormone therapy** | | | |
| | No hormonal therapy | 7 | 35 % |
| | Tamoxifen adjuvant | 6 | 30 % |
| | Castration | 1 | 5 % |
| | *Not available* | 1 | 5 % |

# 4 Methods

The protocol for the laboratory work is presented in the following sections. A flowchart displaying the main steps of the sequencing procedure is to be found in appendix A. All chemicals were purchased from and produced by Life Technologies unless otherwise is specified.

## 4.1 Tissue preparation and isolation of DNA from tumor

Prior to this study, tissue from the tumors was processed and DNA was isolated. Tumor samples were frozen at -80°C after sampling. The frozen specimens were cut into three pieces, and tissue sections (6 μm) from the two cutting regions were made for morphological examination (section 4.2). Tissue-Tek was removed from the specimens and tissue pieces from the same tumor sample were mixed, homogenized and divided into different fractions for analysis. DNA extraction was performed by the use of either the QIAsymphony SP robot together with the QIAsymphony DNA minikit, or by the use of the Allprep DNA/RNA Mini Kit automated with the QIAcube robot. A summary of the protocols is written out in appendix B.

## 4.2 Calculations of tumor cell percentage

The tumor cell percentage of each sample was estimated by two different methods:

- **Morphologic evaluation**: The percentage of tumor cells in two tissue sections from the frozen tumor pieces were estimated by a pathologist on hematoxylin and eosin (HE)-stained slides. Hematoxylin give the nuclei of the cells a blue color and a following counterstaining with eosin make proteins, membranes and other acidophilic structures in the cell turn red/pink. One estimate per section was made, abbreviated HE1 and HE2 tumor cell percentage. If the tumor cell percentage was below 10 % for one of the two sections, the corresponding tumor piece was excluded. Otherwise the three pieces were mixed together, before DNA extraction was performed. An overall tumor percentage has been estimated, by calculating the average of the HE1 and HE2 values. Figure 15 displays an example of HE-stained tissue sections.

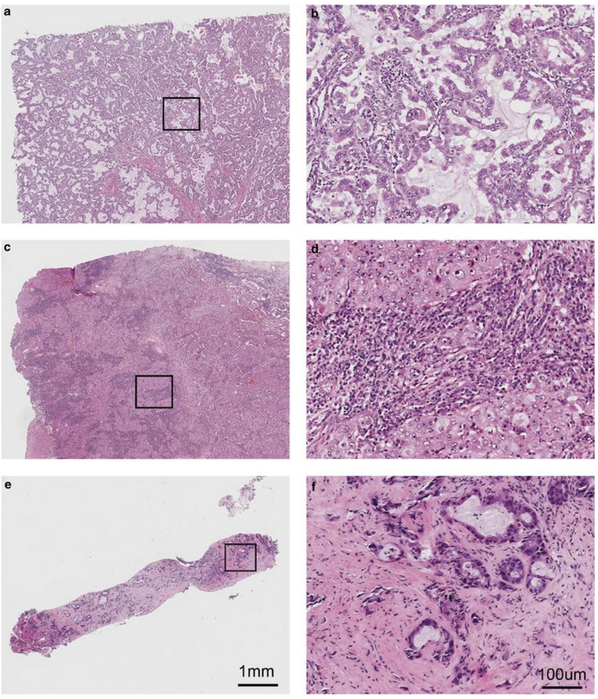


**Figure 15:** Example of hematoxylin and eosin (HE)-stained slides used for estimation of tumor cell percentage. Hematoxylin color the nuclei of the cells blue and eosin make proteins, membranes and other acidophilic structures in the cell turn red/pink[100].

- **Battenberg-algorithm:** The Battenberg algorithm[51] calculated the percentage of tumor cells present in each sample by the use of previously generated SNP 6.0 data.

In advance of this study, the copy number status of the OsloVal-PriMet cohort had been explored. Copy number data was generated using Affymetrix SNP 6.0 arrays, which are DNA microarrays containing probes corresponding to ~1,8 million unique positions in the genome. The probes were 25 base pairs long, and two probes for each SNP position were present (the probes being identical apart from the base corresponding to the given SNP-variant). This enabled the measurement of total DNA content at the SNP position, the log ratio (LogR), as well as the frequency of each SNP-variant, the A- and B-allele. These values were utilized to calculate a BAF value (B allele frequency), i.e the proportion contributed by one SNP allele (B) to the total copy number.

The LogR and BAF values were used as input for the Battenberg algorithm. The algorithm first assigned each SNP to the maternal or paternal allele (a process called phasing), and adjusted the BAF value thereafter. Next, a segmentation of the phased

BAF values was performed using a piecewise constant fitting (PCF) algorithm[79], seeking the best possible fit to the data using one or more constant plateaus. Breakpoints representing haplotype recombination hotspots were identified. Subsequently, the phased haplotype frequencies were segmented across whole chromosomes and the LogRs were segmented using the same breakpoints of the phased haplotype frequencies. Each phased haplotype segment represents a copy number imbalanced region. To calculate the copy number of each segment, the ploidy and purity of each sample were estimated. A Goodness of Fit score was calculated for all possible values for both parameters, and on the basis of these scores, the optimal solution for the genome-wide copy number was selected. The tumor cell percentage is the only output included in this thesis.

## 4.3 DNA integrity

The Genomic ScreenTape produced by Agilent Technologies was utilized to measure the integrity of a selection of the tumor DNA samples. The assay enables the electrophoretic separation and analysis of genomic DNA samples from 200 bp to over 60 000 bp, using 1 µL of sample. The system consists of the 2200 TapeStation System, the Genomic DNA ScreenTape box with Genomic DNA ScreenTape Reagents (Ladder and Sample Buffer) and the Agilent 2200 TapeStation Software. The reagents were purchased from Agilent Technologies and the Agilent 2200 TapeStation User Manual was followed.

***Procedur*e**

The DNA samples were prepared by mixing 1 µL of sample with 10 µL of Genomic DNA Sample Buffer in a tube strip. The solutions were vortexed for 5 seconds and spun down to collect the droplets. The strip was loaded into the Agilent 2200 TapeStation together with the Genomic DNA ScreenTape device and filtered loading tips. The TapeStation then performed loading, electrophoresis and imaging of the samples.

## 4.4 Isolation of DNA from whole blood

Blood samples were available from three of the patients in the cohort. The blood samples were frozen at -80°C after sampling, and thawed on the bench before DNA extraction. Extraction was performed by the use of the Maxwell® 16 instrument together with the

Maxwell®16 Blood DNA Purification Kit provided by Promega. The instrument purifies samples using MagneSilR Paramagnetic Particles (PMPs) to capture, wash and elute the DNA. All reagents were purchased from Promega.

***Procedure***

400 µL of blood from each patient were transferred into well number 1 of prefilled reagent cartridges. The cartridges contained a lysis buffer with 50-75% Guanidine thiocyanat in well number 1, PMPs in well number 2 and a wash buffer in well 3-7. A plunger was placed in well number 7, and the three cartridges, one for each patient, were placed into the instrument. The run was initiated and the plunger was automatically moved from well to well, capturing the DNA with the PMPs, washing and finally eluting the DNA in discrete elution tubes.

When purification was completed, the elution tubes were removed from the instrument to a magnetic elution rack and covered with parafilm to avoid contamination. The DNA samples were pipette from the corner of the elution tubes over to matrix tubes. If magnetic particles were left in the DNA stock, the stock was transferred to eppendorf tubes and spun at 13200 rpm for 2 minutes. The supernatant was then transferred to the matrix tubes.

## 4.5 DNA concentration and purity

The DNA concentration was measured using a Thermo Scientific NanoDrop$^{TM}$ 1000 Spectrophotometer.

***Procedure***

Before sample loading, the measurement pedestal was cleaned with 2 µl of nuclease free water. The NanoDrop software on the corresponding computer was initiated, and nucleic acid measurement was selected. The instrument was blanked with 1,5 µl nuclease free water, followed by loading of 1,5 µl of sample onto the lower measurement pedestal. The sampling arm was closed, and a spectral measurement was initiated using the operating software on the corresponding computer. The sample column was automatically drawn between the upper and lower measurement pedestal, and the spectral measurement was made.

## 4.6 Dilution of DNA samples

The DNA to be used for sequencing analysis was extracted from the stock solutions, and dilutions of 10 ng/µL were made for each sample. A control measurement of the concentrations was accomplished by the use of Qubit® 2.0 Fluorometer and the Qubit® dsDNA HS Assay Kit. The assay is highly selective for double-stranded DNA.

***Procedure***

A Qubit working solution was prepared by diluting the Qubit dsDNA HS Reagent 1:200 in Qubit dsDNA HS Buffer in a 1,5 mL LoBind tube. The amount of working solution created depended on the number of samples to be measured. Each DNA sample required 199 µL of working solution, and the two standards required 190 µL each. Fresh 0,5-mL Qubit tubes for standards and samples were labeled, and the correct volume of Qubit working solution were added to each tube. 1 µl of the samples and 10 µl of the standards were added their respective tubes, leaving the final volume at 200 µL. The tubes were vortexed for 2-3 seconds and incubated at room temperature for 2 minutes.

On the home screen of the Qubit Fluorometer, dsDNA High Sensitivity was selected as the assay type. To calibrate the instrument the standard tubes were read. When calibration was completed, the sample tubes were inserted into the sample chamber, one at a time, and measurements were performed.  If the concentrations were above 10 ng/µl, further dilutions of the samples were performed.

## 4.7 Construction of DNA libraries

To prepare DNA libraries the Ion AmpliSeq™Cancer Hotspot Panel v2 (CHP2) was utilized. This panel consists of a pool of 207 primer pairs that covers frequently mutated regions in 50 human cancer genes (genomic “hot spot” regions). Amplicons of 100-130 base pairs were amplified, covering approximately 2 800 potential somatic mutations found in COSMIC (Catalogue of somatic mutations in cancer)[80]. A list of the genes is to be found in appendix C. The protocol Ion AmpliSeq™ Library Preparation (MAN0006735, Rev B.0) was followed.

***Procedure***

For each DNA/primer pool combination, the components displayed in table 3 were added to a single tube of a PCR tube strip.

**Table 3:** Set up for PCR amplification of DNA targets.

| Component | Volume |
|---|---|
| 5X Ion AmpliSeq™ HiFi Mix | 4µL |
| 5X Ion AmpliSeq™ Primer Pool | 4 µL |
| DNA, 3000 copies (10 ng) | Y |
| Nuclease-free Water | 12 µL -Y |
| Total | 20 µL |

The tubes were vortexed, spun down and loaded in the Applied Biosystems Verti Dx Thermal Cycler. 17 amplification cycles of PCR was performed to amplify genomic DNA targets. The program is presented in table 4.

**Table 4:** PCR protocol to amplify target DNA.

| Stage | Step | Temperature | Time |
|---|---|---|---|
| Hold | Activate the enzyme | 99°C | 2 min |
| Cycle (17 cycles) | Denature | 99°C | 15 sec |
| | Anneal and extend | 60°C | 4 min |
| Hold | - | 10 °C | Hold |

2 µL of FuPa Reagent were added to each amplified sample, to partially digest the primers and phosphorylate the amplicons. The tubes were once again loaded in the thermal cycler and the program displayed in table 5 was completed.

**Table 5**: Thermal cycler protocol to partially digest primers.

| Temperature | Time |
|---|---|
| 50°C | 10 min |
| 55°C | 10 min |
| 60°C | 20 min |
| 10°C | Hold (for up to 1 hour) |

The amplicons were then to be ligated to barcode sequencing adapters, termed X and P1. For each barcode chosen, a mix of Ion P1 Adapter and Ion Express™ Barcode X was prepared, at a final dilution of 1:4 for each adapter. The components listed in table 6 were then added to each tube of the PCR strip containing the digested sample.

**Table 6:** Set up for adapter ligation reaction.

| **Component** | **Volume** |
|---|---|
| Switch Solution | 4 µL |
| Diluted barcode adapter mix | 2 µL |
| DNA Ligase | 2 µL |
| Total volume (includes 22 µL of digested amplicon) | 30 µL |

The solutions were mixed by pipetting up and down, the tubes were loaded in the thermal cycler and the program presented in table 7 was completed.

**Table 7:** Thermal cycler protocol to ligate DNA to barcode adapters.

| **Temperature** | **Time** |
|---|---|
| 22°C | 30 min |
| 72°C | 10 min |
| 10°C | Hold (for up to 1 hour) |

To purify the library, the bead suspension Agencourt® AMPure® XP Reagent was utilized. 45 µL (1,5X sample volume) were added to each tube containing DNA and the mixture was incubated for 5 minutes at room temperature. The tubes were placed in a magnetic rack followed by two minutes of incubation until the solution was clear and the beads formed a pellet. The supernatant was carefully removed and discarded, before 150 µL of 70% ethanol were added to wash the beads. The tubes were moved side-to-side in the two positions of the magnet, and the supernatant was removed and discarded. The washing step was repeated for a second wash.

Keeping the plate in the magnet, the beads were air-dried for 5 minutes at room temperature to allow the remaining ethanol to evaporate. To elute the DNA, the tubes were removed from the magnet and 50 µL of low TE were added. The tubes were once again placed in the magnetic rack, and after 2 minutes of incubation at room temperature the supernatant was removed and stored in new tubes.  2 µL were then taken out and combined with 198 µL of Nuclease-free water to create a 100-fold dilution of the library for quantitation.

## 4.8 qPCR quantitation

The 100-fold dilutions of the unamplified libraries were quantified by the use of qPCR and the Ion Library Quantitation Kit. The protocol Ion AmpliSeq™ Library Preparation User Guide (MAN0006735, Rev B.0) was followed.

When running qPCR, fluorescent dyes are used to label PCR products, and by measuring the accumulations of fluorescent signals during the exponential phase of the reaction, the amount of DNA is quantified. The TaqMan chemistry was utilized, containing oligonucleotide probes with a reporter fluorescent dye on the 5' end and a quencher dye on the 3' end (figure 16). The quencher will greatly reduce the fluorescence emitted from the dye when the probe is intact because of fluorescence resonance energy transfer (FRET). When the probe anneals to a target sequence, downstream for one of the primer sites, it will be cleaved by the nuclease activity of Taq DNA Polymerase when it extends the primer. The reporter dye will be separated from the quencher, and the dye signal is increased. The cleavage removes the probe from the target strand, allowing primer extension to continue to the end of the DNA strand. The fluorescence intensity will be proportional to the amount of DNA amplified, and the quantity of DNA may be calculated[81].

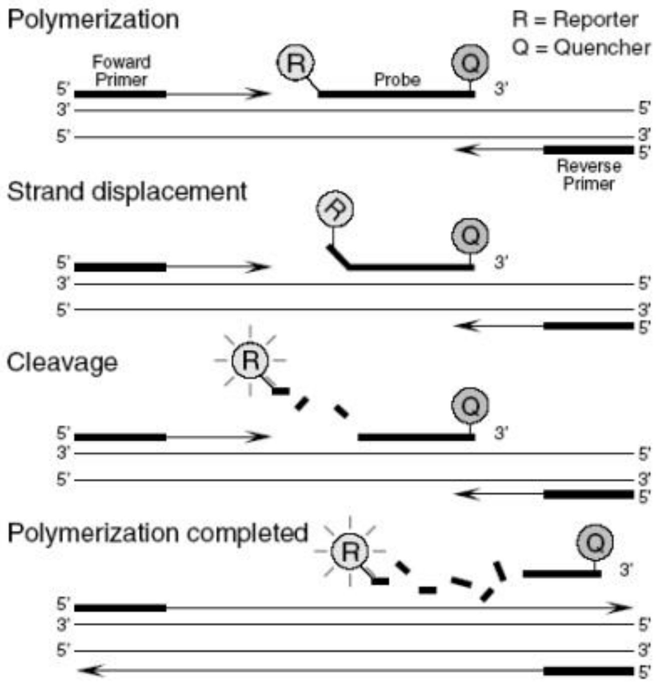


**Figure 16:** The cleavage of the reporter dye (R) from the quencher (Q) during the qPCR reaction. An ologionucleotide probe containing a reporter fluorescent dye on the 5’ end and a quencher dye on the 3’end anneals to the DNA target sequence downstream of the primer site. Due to fluorescence resonance energy transfer between the reporter and the quencher, the fluorescence emitted from the reporter is greatly reduced. When the Taq DNA polymerase extends the primer, the probe will be cleaved, and the reporter dye is separated from the quencher. The dye signal is increased, and the fluorescence intensity is measured[81].

### *Procedure*

For each sample, three standards and one negative control, 20 µL of 2X TaqMan® MasterMix was combined with 2 µL of 20X Ion TaqMan® Assay. 11 µL aliquots were added to the wells of an optical PCR plate. The standards were prepared as a three 10-fold serial dilution of the E.coli DH10B Ion Control Library to 6,8 pM, 0,68 pM and 0,068 pM. 9 µL of the diluted DNA library, standards and negative control were then added to different wells of the PCR plate, for a total reaction volume of 20 µL. The real-time instrument was set to run the program displayed in table 8.

**Table 8:** qPCR program to quantify DNA libraries.

| Stage | Temperature | Time |
|---|---|---|
| Hold | 50°C | 2 min |
| Hold | 95°C | 20 sec |
| Cycle (40 cycles) | 95°C | 1 sec |
|  | 60°C | 20 sec |

Based on the measured library concentration, dilutions of the libraries were made by the use of low TE buffer. Some libraries were diluted to 8 pM, and others to 10, 12 and 14 pM. The DNA concentration was altered to reach the optimal range of template-positive Ion Sphere Particles (ISPs), which is required to be 10-30% (section 4.10). The four DNA libraries to be combined on a sequencing chip were always diluted to the same concentration.

## 4.9 Preparation of Template

The OneTouch™ emulsion PCR system (figure 17) was utilized to clonally amplify the templates onto carrier beads, Ion Sphere™ Particles (ISPs), by the use of emulsion PCR. The protocol Ion PGM™ Template OT2 200 Kit User Guide (MAN0007220, Rev A.0) was followed.

The surface of the ISP is covered with oligonucleotide probes, with sequences complementary to the adapters ligated to the DNA library. The single-stranded DNA fragments are attached to the surface of ISPs by the use of the adapters, one bead to a single fragment. The beads are emulsified into separate water-oil droplets, together with an amplification mix including primers and polymerase, and multiple independent PCR reactions will proceed in parallel. Each of the droplets will ideally capture only one bead, and ISPs covered with amplified DNA fragments will be created[82]. Monoclonal beads are required; no two different fragments are to be attached to the same bead.

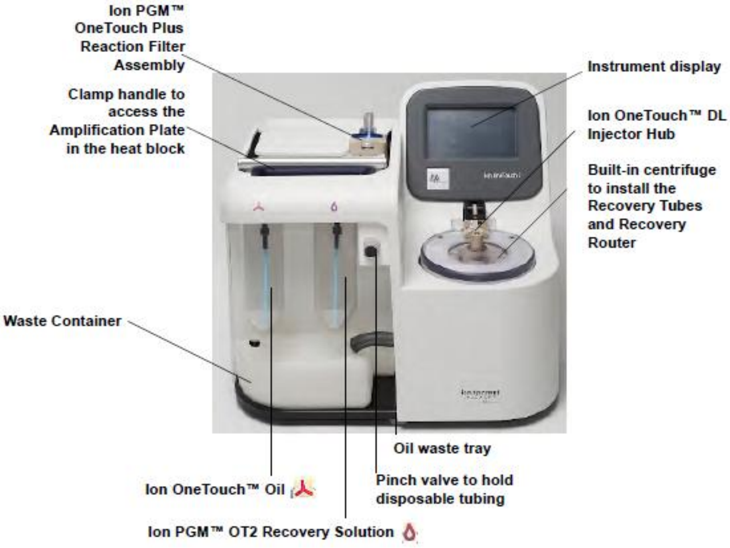

Ion PGM™ OneTouch Plus Reaction Filter Assembly
Clamp handle to access the Amplification Plate in the heat block
Waste Container
Instrument display
Ion OneTouch™ DL Injector Hub
Built-in centrifuge to install the Recovery Tubes and Recovery Router
Oil waste tray
Pinch valve to hold disposable tubing
Ion OneTouch™ Oil
Ion PGM™ OT2 Recovery Solution

**Figure 17:** The OneTouch$^{TM}$ emulsion PCR system. The DNA library is loaded into the reaction filter on top of the instrument together with Ion Sphere particles (ISPs), reaction oil and amplification solution. Single DNA fragments are attached to the ISPs and amplified in emulsions on the amplification plate. Reprinted from the Ion PGM$^{TM}$ Template OT2 200 Kit User Guide.

### *Procedure*

The diluted DNA libraries from two primary-metastasis pairs were combined in a 1,5 mL LoBind tube, 10 µl of each library. In another LoBind tube an amplification solution was made, as shown in table 9.

**Table 9**: Protocol for preparation of amplification solution for use with the Ion OneTouch 2 instrument.

| **Order** | **Reagent** | **Volume** |
|---|---|---|
| 1 | Nuclease-Free Water | 25 µL |
| 2 | Ion PGM™ Template OT2 200 Reagent Mix | 500 µL |
| 3 | Ion PGM™ Template OT2 200 PCR Reagent B | 300 µL |
| 4 | Ion PGM™ Template OT2 200 Enzyme mix | 50 µL |
| 5 | Diluted library | 25 µL |
| **-** | **Total** | **900 µL** |

The Ion Sphere Particles were vortexed for 1 minute to resuspend the particles, and 100 µL of the ISPs were added to the amplification solution. The final solution was vortexed for 5 seconds and quickly spun down before the total volume of 1000 µL was added to the sample port of the Ion PGM OneTouch Plus Reaction Filter Assembly. 1000 µL Ion OneTouch

Reaction Oil were then added, followed by another 500 µL. The reaction filter was carefully rotated until the three ports of the filter were faced down, and the filter was inserted on the top stage of the Ion OneTouch™ Instrument. The Ion OneTouch reaction was then initiated.

At the end of the run, the sample was centrifuged in the Ion OneTouch instrument for 9 minutes. The Recovery tubes, containing the sample, were removed from the instrument and placed in a tube rack. By the use of a pipette, all but 50 µL of recovery solution were carefully removed from each Recovery Tube without disturbing the pelleted ISPs. The remaining solution was resuspended, and transferred to a new tube together with 1000 µL of Ion OneTouch Wash solution. The ISPs could be stored at 2-8 °C for 3 days.

## 4.10 Quality control measurement by Qubit® 2.0 Fluorometer

Samples with a percentage of templated ISPs of 10-30 % generally produce the most data. To verify that the sample contained templated ISPs within the optimal range, a quality control measurement was accomplished by the use of Qubit® 2.0 Fluorometer. The protocol Ion PGM™ Template OT2 200 Kit User Guide (MAN0007220, Rev A.0) was followed.

The Ion Sphere™ Quality control assay labels the ISPs with two different fluorophores: Alexa Fluor® 488 and Alexa Fluor® 647. The two fluorophores anneals to primer B sites (all of the ISPs present) and primer A sites (only the ISPs with extended templates) respectively (figure 18). The ratio of the ISPs to the templated ISPs, or the ratio of the Alexa Fluor® 488 fluorescence to the Alexa Fluor® 647 fluorescence, yields the percent templated ISPs.

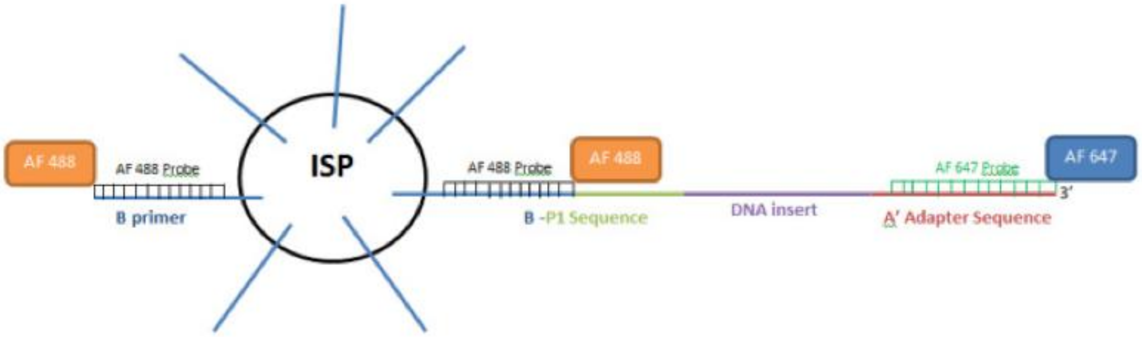


**Figure 18:** The Alexa fluorophores in the Sphere™ Quality control assay: AF 488 binds to primer B-site extending from the ISPs and AF647 binds to the primer A site at the other end of the DNA fragments. Reprinted from the Ion OneTouch™ 200 Template Kit v2 User Guide.

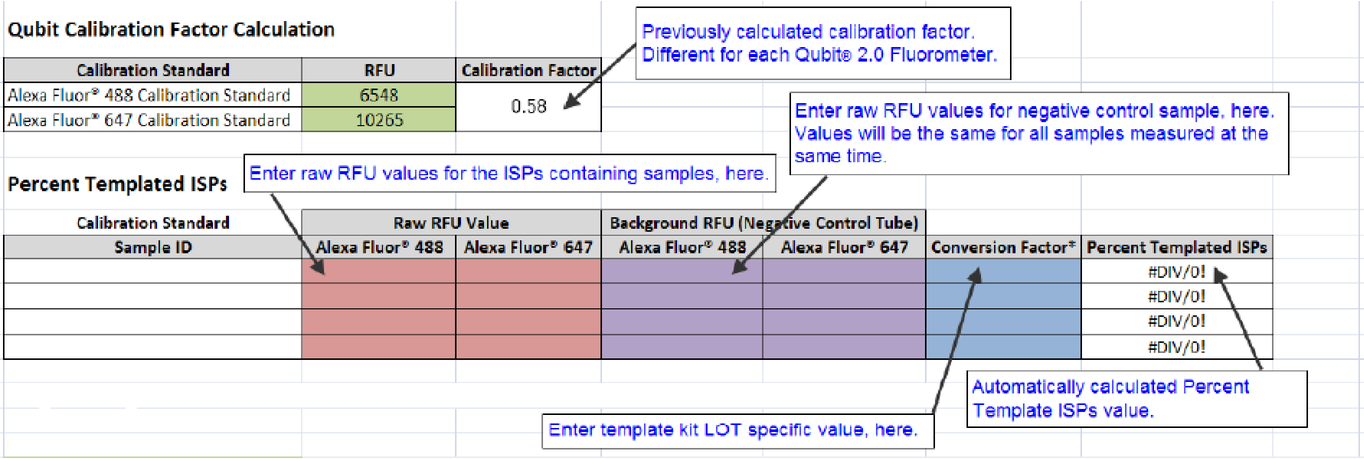
***Procedure***

The template-positive ISP suspension was centrifuged for 2,5 minutes at 15500 x g. All but 100 µL of the supernatant were removed, and the pellet was vortexed for 30 seconds to resuspend the ISPs. 2 µL were transferred to 0,2 mL PCR tube together with 19 µL Annealing Buffer and 1 µL Ion Probes. The tubes were loaded into a thermal cycler, and the protocol displayed in table 10 was performed to anneal the Ion Probes.

**Table 10:** Thermal cycler protocol to anneal Ion Probes to the Ion sphere particles.

| **Temperature** | **Time** |
|---|---|
| 95°C | 2 min |
| 37°C | 2 min |

Unbound probes were removed by washing three times with Quality Control Wash Buffer. 200 µL of Quality Control Wash Buffer were added to the PCR tube, the tube was briefly vortexed and then centrifuged for 2 minutes at 15 600 rpm. All but 10 µL of the supernatant were removed for each wash (measured by visually comparing the supernatant to 10 µL of water in a separate tube). After the final wash, 190 µL of Quality Control Wash Buffer were added, and the entire sample was transferred to a Qubit® assay tube. 200 µL of Quality Control Wash Buffer were added to a fresh Qubit® assay tube to be used as a negative control. The samples were then read by the Qubit® 2.0 Fluorometer, and the AF 488 and AF647 raw values were registered. By the use of the Qubit 2.0 Easy Calculator Microsoft Excell spreadsheet file (figure 19), produced by Life Technologies, the percentage of template positive ISPs was calculated. The raw values from both fluorophores were entered in the appropriate fields for the ISPs and the negative control sample. The lot-specific conversion factor for each Ion PGM Template OT2 reagents kit was entered, as well as the calibration factor for the specific Quibit 2,0 Fluorometer. The percentage of templated ISPs was then calculated.



Qubit Calibration Factor Calculation

| Calibration Standard | RFU | Calibration Factor |
|---|---|---|
| Alexa Fluor® 488 Calibration Standard | 6548 | 0.58 |
| Alexa Fluor® 647 Calibration Standard | 10265 | |

Previously calculated calibration factor. Different for each Qubit® 2.0 Fluorometer.

Percent Templated ISPs

Enter raw RFU values for the ISPs containing samples, here.

Enter raw RFU values for negative control sample, here. Values will be the same for all samples measured at the same time.

| Calibration Standard | Raw RFU Value | | Background RFU (Negative Control Tube) | | | |
|---|---|---|---|---|---|---|
| Sample ID | Alexa Fluor® 488 | Alexa Fluor® 647 | Alexa Fluor® 488 | Alexa Fluor® 647 | Conversion Factor* | Percent Templated ISPs |
| | | | | | | #DIV/0! |
| | | | | | | #DIV/0! |
| | | | | | | #DIV/0! |
| | | | | | | #DIV/0! |

Automatically calculated Percent Template ISPs value.

Enter template kit LOT specific value, here.

**Figure 19:** The Qubit 2.0 Easy Calculator Microsoft Excell spreadsheet. The raw values from the AF488 and AF648 fluorophores (measured by the Qubit Fluorometer) are entered, together with the lot-specific conversion factor for the Ion PGM Template OT2 reagent kit and the calibration factor for the specific Qubit 2.0 fluorometer. The percent templated ISPs are calculated. Reprinted from the Ion OneTouch™ 200 Template Kit v2 User Guide.

## 4.11 Enrichment

An enrichment procedure was performed by the use of The OneTouch ES™ enrichment station and the protocol Ion PGM™ Template OT2 200 Kit User Guide (MAN0007220, Rev A.0). The principle is to separate the ISPs that do not have any DNA attached from those who do. This is done by the use of streptavidin coated magnetic beads and the biotinylated enrichment primers that will be bound to the templated ISPs after amplification. The streptavidin molecule will bind the biotinylated primer, and hence the DNA molecule and ISP to which the primer is attached (figure 20). By exposing the system to a magnet the templated ISPs will be pulled out of the solution, while any ISPs without amplified DNA will stay in the solution and be washed away. A Melt-Off Solution containing NaOH is used to denature the complementary strand of the target DNA, to be able to use the target strand as sequencing template.

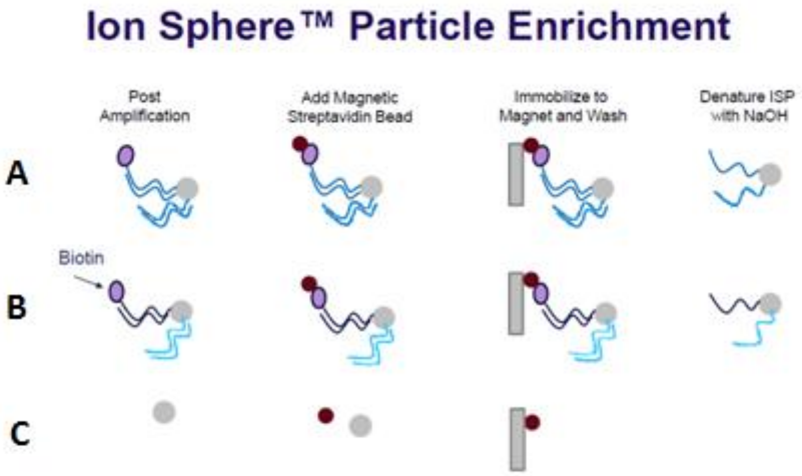

Ion Sphere™ Particle Enrichment
Post Amplification
Add Magnetic Streptavidin Bead
Immobilize to Magnet and Wash
Denature ISP with NaOH
A
B
C
Biotin

**Figure 20:** The basic principle of the Enrichment procedure. Streptavidin coated magnetic beads are added to the solution containing templated ISPs. The beads will bind to the biotinylated primers that were attached to the DNA fragments during the emulsion PCR. By exposing the system to a magnet, the templated ISPs will be pulled out of the solution, while any ISPs without amplified DNA will stay in the solution. Both monoclonal (A) and polyclonal (B) ISPs will be captured, while the non-templated beads (C) remain. Reprinted with permission from Life Technologies.

***Procedure:***

A fresh Melt-Off Solution was made by combining the components displayed in table 11.

**Table 11:** Protocol for preparation of Melt-Off Solution to be used for enrichment of template positive Ion Sphere Particles.

| **Order** | **Component** | **Volume** |
|---|---|---|
| 1 | Tween Solution | 280 µL |
| 2 | 1 M NaOH | 40 µL |
|  | Total | 320 µL |

The Dynabeads® MyOne™ Strepdavidin C1 Beads were then washed and prepared. The reagent tube was vortexed for 30 seconds to thoroughly resuspend the beads, then centrifuged for 2 seconds. 13 µL of the beads were added to a new 1,5 mL Eppendorf LoBind Tube, and the tube was placed in a magnetic rack for 2 minutes. The supernatant was removed and discarded. 130 µL of MyOne™ Beads Wash Solution were added to the Dynabeads, and the tube was removed from the magnet, vortexed and finally centrifuged for 2 seconds. An 8-well strip was then placed in the slot of the enrichment statio, with the square-shaped tab to the right. Table 12 displays the solutions filled in the different wells.

**Table 12:** The reagents to be filled in the 8-well strip for enrichment.

| **Well number** | **Reagent to dispense in well** |
|---|---|
| Well 1 | Entire template-positive ISP sample (100 µL) |
| Well 2 | 130 µL of Dynabeads® MyOne™ Strepdavidin C1 Beads resuspended in MyOne™ Beads Wash Solution |
| Well 3 | 300 µL of Ion OneTouch™ Wash Solution |
| Well 4 | 300 µL of Ion OneTouch™ Wash Solution |
| Well 5 | 300 µL of Ion OneTouch™ Wash Solution |
| Well 6 | Empty |
| Well 7 | 300 µL of Melt-Off solution |
| Well 8 | Empty |

A new pipette tip was loaded in the tip arm of the enrichment station and 10 µL of Neutralization Solution were added to a new 0,2 mL PCR tube and inserted into the hole in the base of the Tip Loader. The run was then initiated. After enrichment the PCR tube containing enriched ISPs may be stored at 2-4°C for 3 days.

## 4.12 DNA sequencing

The sample with enriched, templated ISPs was then loaded to an ion 318 v2 chip and the sequencing reaction was performed by the use of the Ion Personal Genome Machine (PGM) (figure 21). The protocol Ion PGM™ Sequencing 200 Kit v2 User Guide (MAN0007273, Rev 3.0) was followed.

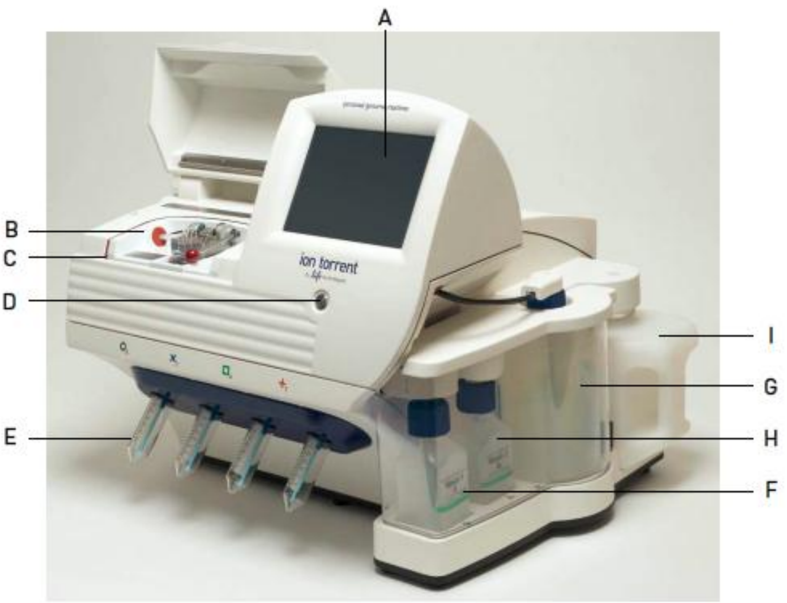


**Figure 21:** The Ion Personal Genome Maschine$^{TM}$ System with a touchscreen (**A**), chip clamp (**B**), grounding plate (**C**), power button (**D**), reagent bottles (**E**), “wash 1” bottle (**F**), “wash 2” bottle (**G**), “wash 3” bottle (**H**) and “waste” bottle (**I**). Reprinted from the Ion PGM Sequencing 200 Kit v2 User Guide.

### *Procedure*

First, an initialization of the Ion PGM$^{TM}$ System was performed. The wash bottles to be attached to the PGM were rinsed with 18 MΩ water. The “wash 2” bottle was filled with 2 liters of water, followed by the whole volume of an Ion PGM$^{TM}$ Sequencing 200 v2 W2 Solution bottle and 70 µL of freshly prepared 100 mM NaOH. The bottle was capped and inverted five times to mix the content. 350 µL of 100 mM NaOH were transferred to the “wash 1” bottle and 50 mL of Ion PGM$^{TM}$ Sequencing 200 v2 1X W3 Solution to the “wash 3” bottle. New sipper tubes were attached to the caps on the instrument, and the wash bottles were secured. The procedure was initiated by pressing initialization on the main menu. The Ion PGM$^{TM}$ System tested the bottles for leaks, filled the “wash 1” bottle and adjusted the pH of the W2 solution.

The dNTP stock solutions were thawed, vortexed and centrifuged to collect the droplets. Four reagent bottles were labeled with the four nucleotide names and 20 µL of each dNTP stock solution were carefully transferred to its respective reagent bottle. Clean cloves were used for each dNTP to avoid cross-contamination. After the wash solutions had initialized, new sipper tubes were inserted into each dNTP port on the PGM. The reagent bottles were attached to the correct dNTP ports, and the touch screen prompts on the instrument was followed to complete initialization. The PGM checked the pressure of the reagent bottles and measured the pH of the reagents. For each initialization the first run was started within 1 hour, and the second run within 27 hours.

The sample was then prepared and loaded on the chip. 5 µL of Control Ions Sphere™ particles were transferred to the PCR tube containing enriched, template-positive ISPs. The solution was mixed by pipetting up and down and then centrifuged for 2 minutes at 15,500 x g. The supernatant was carefully removed, leaving 15 µL behind in the tube (measured by visually comparing to 15 µL of water in a separate tube). 12 µL of Sequencing Primer were then added. In addition, Annealing Buffer was added if the resulting total volume was less than 27 µL. The sample was pipetted up and down to disrupt the pellet before it was loaded in the thermal cycler and the program displayed in table 13 was completed.

**Table 13:** Thermal cycler protocol to anneal sequencing primers.

| Temperature | Time |
|---|---|
| 95°C | 2 minutes |
| 37 °C | 2 minutes |

A chip check was performed to ensure that the chip to be used was functioning properly. The chip was not handled with gloves, and fingers were grounded by touching the grounding pad next to chip clamp on the instrument. The old chip in the chip socked was replaced with a new Ion 318™ chip v2, and the procedure was initiated by pressing chip check on the touchscreen.

Following a successful chip check, sequencing polymerase was bound to the ISPs. 3µL of Ion PGM™ Sequencing 200 v2 Polymerase were added to the sample. The sample was pipetted up and down to mix, and incubated at room temperature for 5 minutes.

The sample was then ready to be loaded on the Ion 318™ chip v2. The chip was tilted 45 degrees, keeping the loading port as the lower port (figure 22). A pipette tip was inserted into the loading port, and the liquid inside was removed and discarded. The chip was placed upside-down in the centrifuge adapter bucket in the MiniFuge. The centrifuge adapter was balanced with a used chip of the same type and orientation. A 5 seconds spin was performed to completely empty the chip. The entire sample (~30 µL) of ISPs was then collected into a Rainin® SR-L200F pipette tip. The tip was inserted into the loading port, and the pipette was dialed down to slowly deposit the ISPs into the chip (~1 µL per second). A 30 seconds spin was performed, with the chip tap pointing in, and the sample was mixed by pipetting the sample in and out of the chip three times. Another round of centrifuging and mixing followed, with the chip tap pointing out this time. Following one last 30 seconds spin, the liquid was carefully removed from the chip by dialing the pipette, leaving only the ISPs behind. The chip was then inserted in the PGM and the run was initiated.

After the sequencing reaction was completed, a sequencing report was available at the Ion Torrent server, displaying the achieved loading density on the chip, the amount of polyclonal beads, the amount of usable reads generated, and the mean sequencing depth amongst other. The chip loading should be above 50 % to achieve a sufficient amount of data.

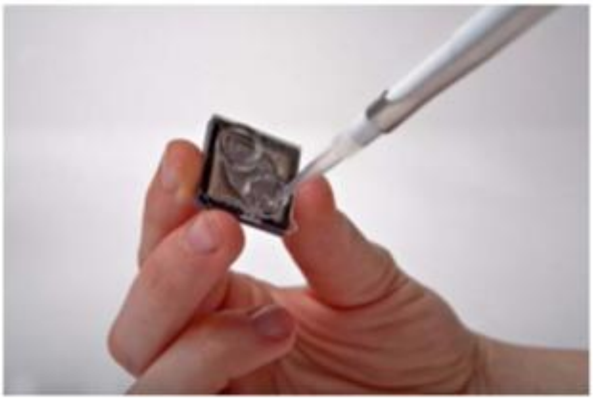


**Figure 22**: Loading of sample into the loading port of the Ion 318™ chip v2. Reprinted from the Ion PGM Sequencing 200 Kit v2 User Guide.

## 4.13 Data filtering and variant calling

The sequencing data was stored at the Ion Torrent server, which is the required computing hardware to support the PGM. Signal processing and base calling algorithms were used at the server to generate the DNA sequences associated with the different reads. The system also performed a quality check of the data. Reads were tested to see if they were generated from mixed DNA templates on an ISP or if they were of low signal quality, and these reads were filtered out. A removal of the adapter sequence was also performed[83].

The Torrent Variant Caller (TVC) plugin was utilized to identify variants that differed from the reference genome hg19. TVC analyzed the mapped reads and decided whether there was sufficient statistical evidence to call base changes or indels at individual base positions. To be included in the final dataset, the variants had to meet the criteria displayed in table 14.

Ion Torrent uses a quality score system with a Phred-like method to predict the probability that a base call is correct. The minimum quality score is a Phred-scaled number, which is a measure of the quality of the identification of bases. The phred quality score Q is logarithmically related to the probability of a base-calling error (P), and is calculated as follows:

$$Q = -10 \log P$$

If a base is assigned a phred quality score of 6, there is thus a 25% chance that this base is incorrect[84,85].

**Table 14:** Restrictions for variant calling. The variants that did not meet the following criteria were excluded from the sequencing report. Modified from[86].

| Parameter | Explanation | SNP | INDEL |
|---|---|---|---|
| Minimum allele frequency | Do not call variants if the observed allele frequency is below this value | 0,02 | 0,05 |
| Minimum quality | Do not call variants if the Phred-scaled call quality is below this value | 6 | 6 |
| Minimum coverage | Do not call variants if the total coverage on both strands is below this value | 6 | 15 |
| Minimum coverage on either strand | Do not call variants if the coverage on either strand is below this value | 0 | 2 |
| Maximum strand bias | Do not call variants if the proportion of variant alleles coming from one strand only, exceeds this value. | 0,95 | 0,9 |
| Maximum relative read quality | Do not call variants if the relative Phred-scaled call quality is below this value. | 6,5 | 6,5 |
| Maximum common signal shift | Do not call variants if the distance between the predicted and observed signal at the allele locus exceeds this value (0,3 = 30% of variant change size). | 0,3 | 0,3 |
| Maximum reference/variant signal shift (insertions) | Do not call insertions if the distance between predicted and observed signal in the reference allele/variant allele exceeds this value. | 0,2 | 0,2 |
| Maximum reference/variant signal shift (deletions) | Do not call insertions if the distance between predicted and observed signal in the reference allele/variant allele exceeds this value. | 0,2 | 0,2 |

The filtrated data from each sample were stored in separate VCF (variant call format) files. The files were exported from the Ion Torrent Server, and annotation of the variants were performed by running a script created by bioinformaticians at the department of Cancer Genetics at Oslo University Hospital (Radium hospitalet). The VCF files with data from the primary tumor and metastasis (and the normal sample) from the same patients were combined to one file using ‘IonVcfCombine.pl’ from the ‘VcfProcess’ package. The software tool ANNOVAR[87] and the program ’table_annoval.pl’ was then utilized to perform an annotation of the genomic variants. ANNOVAR is a command-line driven software tool that compares genetic variants found in a text-based input file with data from genomic databases such as 1000 genomes[88], COSMIC[89] and dbSNP[90]. Information about the variant-type, whether the variant is a part of an exon or intron and the functional consequence was achieved, and a new VCF output file with ANNOVAR annotations was created.

PHP-based variant reports were then produced by running ‘IonVcfReport.pl’ on the annotated VFC file. The report contained PHP codes and functions to structure the variant information, from which a HTML document finally was generated. The HTML-reports are web browser documents, with a diagram displaying the different variants annotated in each sample, as well as the percentage of reads with the different variants. The diagram made it possible to visually compare the variants detected in primary tumor and lymph node metastasis from each patient. If low frequent variants were present in a report, the output BAM-file (Binary Alignment Map) was visualized in IGV (integrative genomics viewer). The Ion Torrent server uses BAM-files to store flow-signal and base calling information. If technical errors seemed to have occurred, the variants were manually excluded from the sequencing reports.

# 5 Results

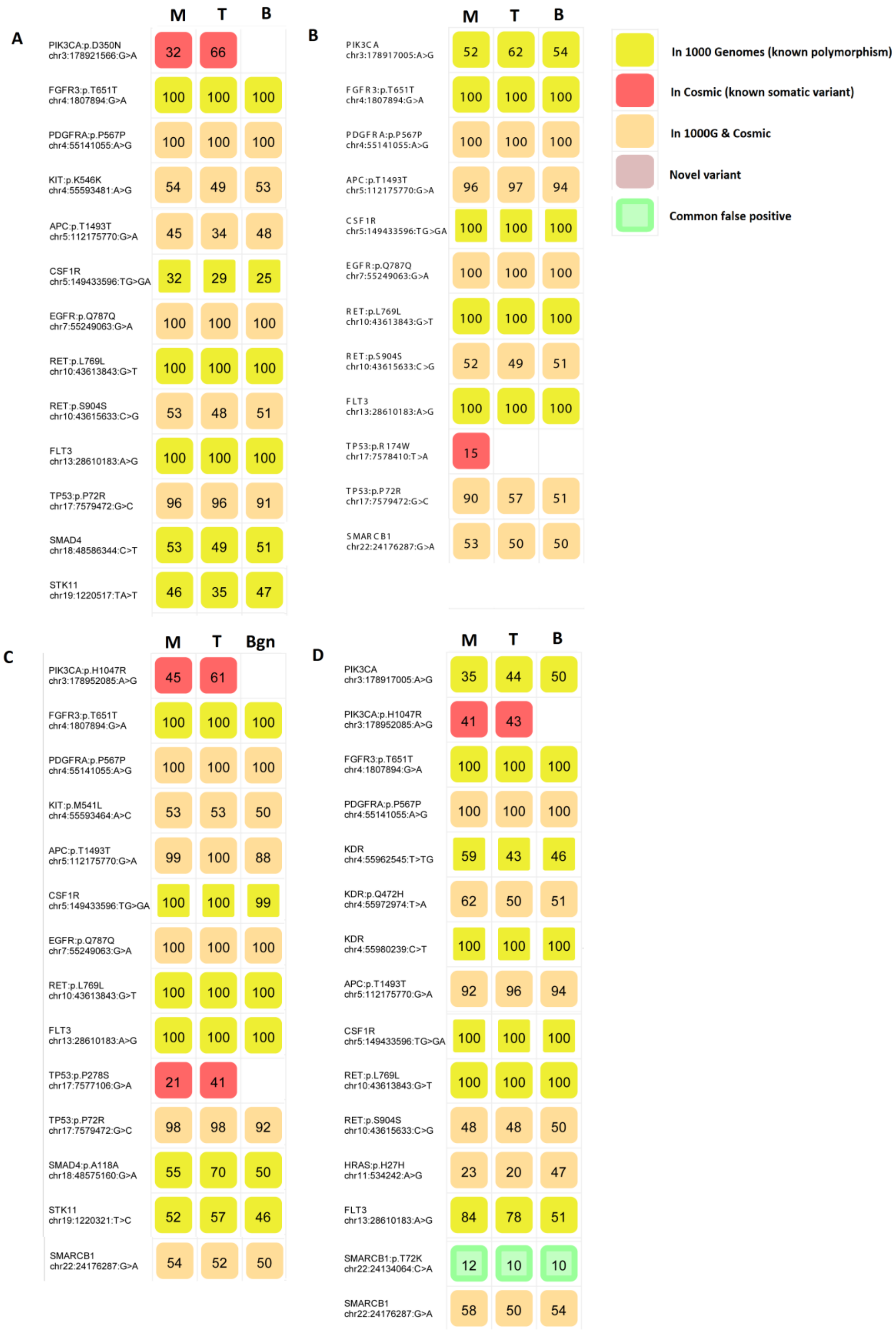
In this study targeted sequencing was performed to compare the genomes of 20 primary tumors and their matched lymph node metastasis. Variant calling and quality-based filtering of the results were performed, and sequencing reports were generated identifying and visualizing the genomic variants present in the samples. In the following section the mutations detected in the different tumors will be presented, followed by a presentation of the results concerning the general performance of the laboratory work. The complete collection of sequencing reports is to be found in appendix D.

## 5.1 Interpretation of variation

### 5.1.1 Annotation of DNA variants

By the use of the Ion Torrent Variant Caller the variants differing from the reference genome hg19 was called. The type of variant was decided by comparing them with genomic data in COSMIC, 1000 genomes and dbSNP. A normal sample was available from four of the patients (three blood samples and one benign tumor). The reports from these patients were useful to validate the variant annotation, and to indicate which variants that was likely to be somatic mutations or germline variants in the other samples. These patient reports are depicted in figure 23. The numbers displayed inside the colored boxes indicate the percentage of reads with the variant in the metastasis and the primary tumor. The variants displayed in yellow are known polymorphisms in the 1000 genomes database and the variants displayed in a red color are known somatic mutations in cancer in the COSMIC database. The pink colored variants are present in both 1000 genomes and in COSMIC.

All the variants annotated as known polymorphisms due to their presence in the 1000 genomes database, was found in the germ line of the patients. The same was true for the variants present in both 1000 genomes and COSMIC. The only genomic variants not to be found in the normal samples were the variants registered in COSMIC only. In addition, the read frequencies of the germ line variants were (with some exceptions) always close to 50 or 100%, while many of the somatic variants clearly deviated from these two values.



A

| | M | T | B |
|---|---|---|---|
| PIK3CA:p.D350N<br>chr3:178921566:G>A | 32 | 66 | |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| KIT:p.K546K<br>chr4:55593481:A>G | 54 | 49 | 53 |
| APC:p.T1493T<br>chr5:112175770:G>A | 45 | 34 | 48 |
| CSF1R<br>chr5:149433596:TG>GA | 32 | 29 | 25 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 53 | 48 | 51 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 | 100 |
| TP53:p.P72R<br>chr17:7579472:G>C | 96 | 96 | 91 |
| SMAD4<br>chr18:48586344:C>T | 53 | 49 | 51 |
| STK11<br>chr19:1220517:TA>T | 46 | 35 | 47 |

B

| | M | T | B |
|---|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 52 | 62 | 54 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 96 | 97 | 94 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 52 | 49 | 51 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 | 100 |
| TP53:p.R174W<br>chr17:7578410:T>A | 15 | | |
| TP53:p.P72R<br>chr17:7579472:G>C | 90 | 57 | 51 |
| SMARCB1<br>chr22:24176287:G>A | 53 | 50 | 50 |

- 🟨 In 1000 Genomes (known polymorphism)
- 🟥 In Cosmic (known somatic variant)
- 🟧 In 1000G & Cosmic
- 🟫 Novel variant
- 🟩 Common false positive

C

| | M | T | Bgn |
|---|---|---|---|
| PIK3CA:p.H1047R<br>chr3:178952085:A>G | 45 | 61 | |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| KIT:p.M541L<br>chr4:55593464:A>C | 53 | 53 | 50 |
| APC:p.T1493T<br>chr5:112175770:G>A | 99 | 100 | 88 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 | 99 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 | 100 |
| TP53:p.P278S<br>chr17:7577106:G>A | 21 | 41 | |
| TP53:p.P72R<br>chr17:7579472:G>C | 98 | 98 | 92 |
| SMAD4:p.A118A<br>chr18:48575160:G>A | 55 | 70 | 50 |
| STK11<br>chr19:1220321:T>C | 52 | 57 | 46 |
| SMARCB1<br>chr22:24176287:G>A | 54 | 52 | 50 |

D

| | M | T | B |
|---|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 35 | 44 | 50 |
| PIK3CA:p.H1047R<br>chr3:178952085:A>G | 41 | 43 | |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| KDR<br>chr4:55962545:T>TG | 59 | 43 | 46 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 62 | 50 | 51 |
| KDR<br>chr4:55980239:C>T | 100 | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 92 | 96 | 94 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 48 | 48 | 50 |
| HRAS:p.H27H<br>chr11:534242:A>G | 23 | 20 | 47 |
| FLT3<br>chr13:28610183:A>G | 84 | 78 | 51 |
| SMARCB1:p.T72K<br>chr22:24134064:C>A | 12 | 10 | 10 |
| SMARCB1<br>chr22:24176287:G>A | 58 | 50 | 54 |

**Figure 23**: Sequencing reports from the four samples with a matched normal sample available (**A**-**D**). The metastasis is abbreviated M, the primary tumor T, blood B and the benign tumor Bgn. The numbers inside the colored boxes indicate the percentage of reads with the different variants.

### 5.1.2 Recurrent mutations

Among the 20 patients, 12 patients were found to hold one or more somatic mutations in the primary tumor and/or in the lymph node metastasis (60%) (figure 24). The mean mutation frequency was 0,8 and the range 0-3. In total 32 mutations were registered across the cohort, 16 in the primary tumors and 16 in the metastases. Mutations were found in the genes *TP53* (45%), *PIK3CA* (25%), *AKT1* (5%) and *GNAS* (5%). The distribution of the mutations in the primary tumors and the metastases are displayed in figure 25.

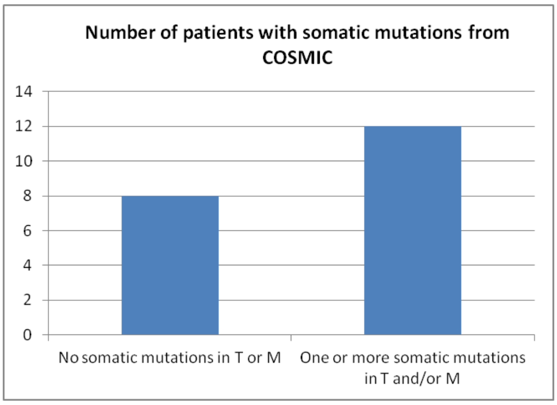


**Figure 24:** Bar graph displaying the number of patients found to hold one or more somatic mutations in the primary tumor (T) and/or metastasis (M) and the number of patients where no somatic mutations were revealed.

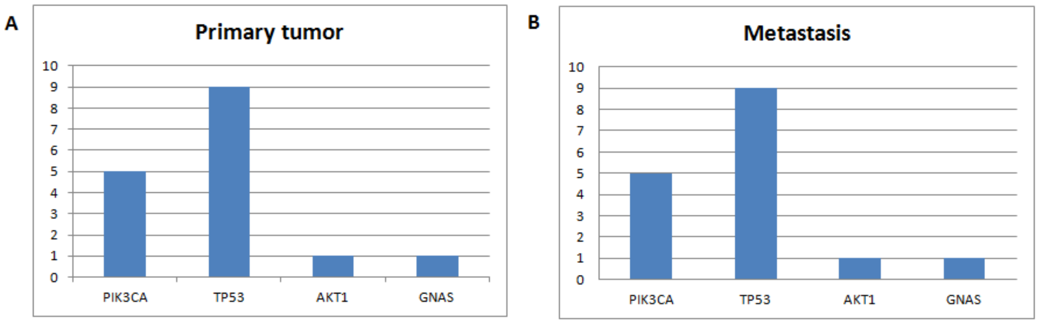


**Figure 25:** Bar graphs displaying the distribution of the different mutations in the 20 primary tumors (**A**) and the 20 metastases (**B**). Mutations were found in the genes *PIK3CA*, *TP53*, *AKT1* and *GNAS*.

The site of the mutations within each of the five genes differed between most of the samples. All together, 15 different missense mutations were detected, all caused by single base substitutions in the one of exons of the genes, as well as one frameshift mutation, caused by a single base deletion. Figure 26 displays the location of the various mutations.

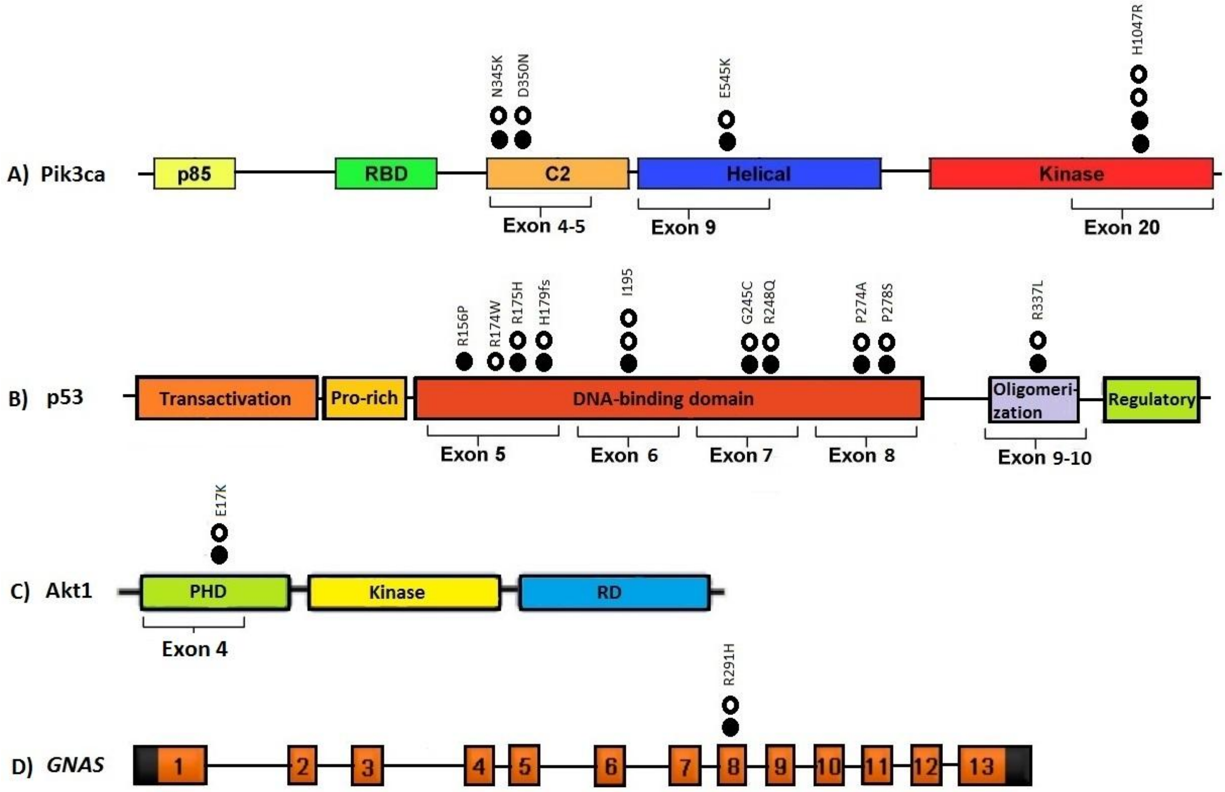


**Figure 26:** The protein domain structures of the Pik3ca (**A**), p53 (**B**) and Akt1(**C**) proteins and the *GNAS* gene locus (**D**). The frequencies of the different mutations are indicated on each figure; black spots are mutations in the primary tumor, and hollow circles are mutations in the metastasis. The amino acid changes are written above. The *GNAS* gene has a highly complex expression pattern, encoding different transcripts, and the protein structure is therefore not included in the figure. Modified from [91] (**A**), [45] (**B**),[92] (**C**) and [93] (**D**).

### 5.1.3 Tumor cell percentage and histopatological features

The percentages of reads with the somatic variants are presented in table 15. The read-frequencies have to be interpreted relative to the purity of the samples. The estimated tumor cell percentages for the different samples are displayed in column 6-7, the average of HE1 and HE2 and the Battenberg calculations respectively (se Methods section 4.2 for details).

The hormone receptor statuses of each of the tumors are presented in column 8-9. Among the primary tumors there were an equal distribution of ER-negative (ER-, 45%) and ER-positive tumors (ER+, 45%). The ER-negative tumors had a higher frequency of mutations, with 67% of the *TP53* mutations being found in ER-negative tumors and only 11% in ER-positive tumors. The remaining 22% were found in tumors with unknown hormone receptor status.

**Table 15:** Overview of the percentage of reads with the different mutations in the primary tumors (T) and metastases (M), together with the estimated tumor cell percentages, ER-status and PgR-status. Somatic mutations were detected in the genes *PIK3CA*, *TP53*, *AKT1* and *GNAS* (the red colored areas), and the numbers refer to the percentage of reads with the mutation.

| Sample ID | PIK3CA | TP53 | AKT1 | GNAS | Tumor% (Average HE1+HE2) | Battenberg Tumor % | ER-status | PgR-status |
|---|---|---|---|---|---|---|---|---|
| 4T | | | | | 60 % | 49 % | Pos | Pos |
| 4M | | | | | 50 % | 43 % | Pos | Pos |
| 7T | | 63 | | | 80 % | 53 % | Pos | Pos |
| 7M | | 31 | | | 50 % | 38 % | Neg | Neg |
| 9T | 66 | | | | 45 % | 65 % | Pos | Pos |
| 9M | 32 | | | | 90 % | 59 % | NA | NA |
| 10T | 43 | | | | 30 % | 37 % | Pos | Pos |
| 10M | 41 | | | | 30 % | 36 % | Pos | Pos |
| 11T | | | | | 45 % | 59 % | Pos | Pos |
| 11M | | | | | 80 % | 59 % | Pos | Pos |
| 14T | | | | | 75 % | 61 % | Pos | Pos |
| 14M | | | | | 90 % | 59 % | Pos | Pos |
| 18T | | | | | 60 % | 49 % | Pos | Pos |
| 18M | | | | | 85 % | 52 % | Pos | Pos |
| 16T | | 72 | | | 60 % | 62 % | Pos | Neg |
| 16M | | | | | 90 % | 65 % | NA | NA |
| 20T | | | | | 80 % | 53 % | Pos | Neg |
| 20M | | | | | 90 % | 73 % | Pos | Neg |
| 6T | 22 | 28 | | | 70 % | 25 % | Neg | Pos |
| 6M | 43 | 43 | | | 60 % | 34 % | Neg | Pos |
| 15T | | | | | 65 % | 55 % | Neg | Pos |
| 15M | | 15 | | | 80 % | 66 % | NA | NA |
| 1T | 61 | 41 | | | 70 % | 56 % | Neg | Neg |
| 1M | 45 | 21 | | | 55 % | 18 % | Neg | Neg |
| 3T | | 36 | | | 50 % | 36 % | Neg | Neg |
| 3M | | 35 | | | 65 % | 44 % | Neg | Neg |
| 5T | 64 | 45 | | | 80 % | 44 % | Neg | Neg |
| 5M | 69 | 56 | | | 60 % | 45 % | Neg | NA |
| 8T | | 40 | 44 | 38 | 40 % | 40 % | Neg | Neg |
| 8M | | 42 | 49 | 39 | 65 % | 38 % | NA | NA |
| 12T | | 27 | | | 60 % | 27 % | Neg | Neg |
| 12M | | 66 | | | 70 % | 50 % | Neg | Neg |
| 17T | | | | | 75 % | 34 % | Neg | Neg |
| 17M | | | | | 60 % | 27 % | Neg | Neg |
| 19T | | | | | 80 % | 29 % | Neg | Neg |
| 19M | | | | | 60 % | 40 % | Neg | Neg |
| 2T | | 63 | | | 50 % | 50 % | NA | NA |
| 2M | | 28 | | | 65 % | 23 % | NA | NA |
| 13T | | | | | 50 % | 25 % | NA | NA |
| 13M | | | | | 65 % | 31 % | NA | NA |

### 5.1.4 Comparison of mutations in primary tumor and metastasis

By comparing the somatic mutations in the primary tumor with those found in the metastasis from the 12 patients, two different situations were identified:

- *Concordant pairs*: The primary tumor and the metastasis had the same somatic mutations (9 patients)
- *Discordant pairs*: The primary tumor and the metastasis had different somatic mutations (3 patients)

#### 5.1.4.1 Concordant pairs

The same mutations were present in the primary tumor and the metastasis in 9 of the patients. The allele frequencies were however not identical in several of the pairs. As tumor samples of mixed purity were sequenced, the allele frequency would depend on the amount of wild-type DNA present in the sample (i.e normal cells), as well as copy number alterations and intratumor heterogeneity. To investigate this, we calculated the ratio between the allele frequency of a somatic variant in the metastasis and the allele frequency of the same variant in the primary tumor. Further, the ratio between tumor cell percentage of the metastasis and the primary tumor (calculated by the Battenberg algorithm) was computed. The relationship between these two ratios is displayed in the scatter plot in figure 27. Some sample pairs did not differ significantly in any of the values, and had both ratios around 1 (patient 8 and 10). Another patient had samples were the allele frequencies were different, but the ratios for allele frequencies and tumor cell percentage were equal, indicating that the observed differences in allele frequencies were due to different amounts of normal cells in the samples (patient 2). Discrepancies between the two ratios were observed for the rest of the mutations.

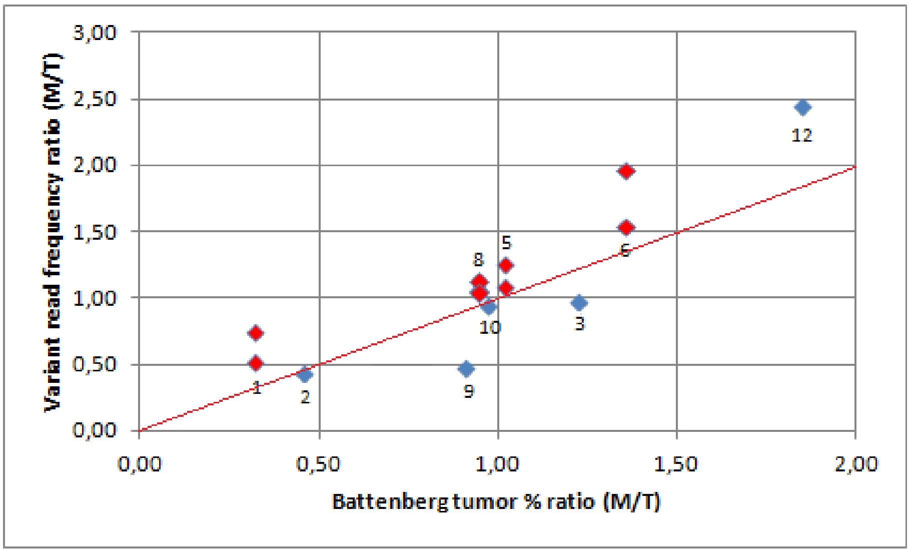
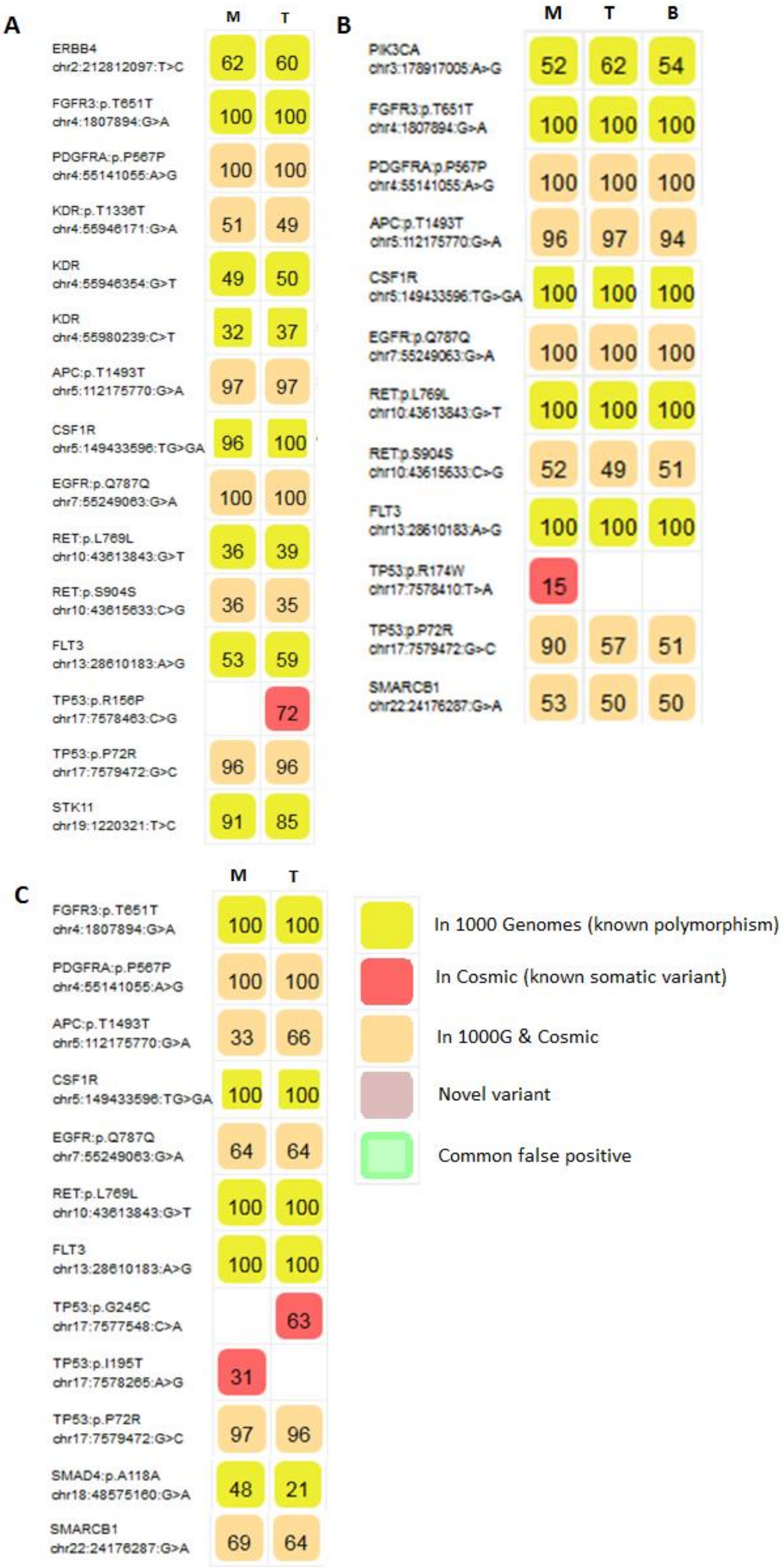


**Figure 27:** Concordant pairs: The calculated ratios between the variant read frequencies in the metastasis (M) and the primary tumor (T), and between the tumor cell percentage in the metastasis and the primary tumor. The red colored points on the same vertical axis are different variants from the same patient. The blue colored points are from patients were only one mutation was detected. The patient numbers are marked below or above the corresponding points. The red line indicates the area where the two ratios are equal.

### 5.1.4.2 Discordant pairs

To look further into the differences between primary tumor and the lymph node metastasis, the sequencing reports from the patients in the discordant group are presented in figure 28. In patient 16, the missense mutation *TP53*:p.R156P was found only in the primary tumor and not in the metastasis. In patient 15, the missense mutation *TP53*:p.R174W was found only in the metastasis and not in the primary tumor. Both of the mutations occurred in exon 5, which is a part of the DNA binding domain of the protein. The rest of the called variants were found in both tumors.

In patient 7, different *TP53* mutations were identified in the primary tumor and in the metastasis. In the metastasis the missense mutation *TP53*:p.I195T (in exon 6) was detected in 31 percent of the reads. In the primary tumor another missense mutation was detected in 63 percent of the reads; *TP53*:p.G245C (in exon 7). The rest of the variants were found in both tumors. The raw data from the samples were investigated using IGV, to confirm that the mutations truly were absent from all the reads in one of the two tumors.



A

| | M | T |
|---|---|---|
| ERBB4<br>chr2:212812097:T>C | 62 | 60 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR:p.T1336T<br>chr4:55946171:G>A | 51 | 49 |
| KDR<br>chr4:55946354:G>T | 49 | 50 |
| KDR<br>chr4:55980239:C>T | 32 | 37 |
| APC:p.T1493T<br>chr5:112175770:G>A | 97 | 97 |
| CSF1R<br>chr5:149433596:TG>GA | 96 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 36 | 39 |
| RET:p.S904S<br>chr10:43615633:C>G | 36 | 35 |
| FLT3<br>chr13:28610183:A>G | 53 | 59 |
| TP53:p.R156P<br>chr17:7578463:C>G | | 72 |
| TP53:p.P72R<br>chr17:7579472:G>C | 96 | 96 |
| STK11<br>chr19:1220321:T>C | 91 | 85 |

B

| | M | T | B |
|---|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 52 | 62 | 54 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 96 | 97 | 94 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 52 | 49 | 51 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 | 100 |
| TP53:p.R174W<br>chr17:7578410:T>A | 15 | | |
| TP53:p.P72R<br>chr17:7579472:G>C | 90 | 57 | 51 |
| SMARCB1<br>chr22:24176287:G>A | 53 | 50 | 50 |

C

| | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 33 | 66 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 64 | 64 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.G245C<br>chr17:7577548:C>A | | 63 |
| TP53:p.I195T<br>chr17:7578265:A>G | 31 | |
| TP53:p.P72R<br>chr17:7579472:G>C | 97 | 96 |
| SMAD4:p.A118A<br>chr18:48575160:G>A | 48 | 21 |
| SMARCB1<br>chr22:24176287:G>A | 69 | 64 |

| | |
|---|---|
| 🟨 | In 1000 Genomes (known polymorphism) |
| 🟥 | In Cosmic (known somatic variant) |
| 🟧 | In 1000G & Cosmic |
| 🟫 | Novel variant |
| 🟩 | Common false positive |

**Figure 28:** Sequencing reports from the three discordant pairs: Patient 16 had a *TP53* mutation present in the primary tumor only (**A**), patient 15 had a *TP53* mutation present in the metastasis only (**B**) and patient 7 had two different *TP53* mutations present, one in the primary tumor and the other in the metastasis (**C**). The metastasis is abbreviated M, the primary tumor T and blood B. The numbers inside the colored boxes indicate the percentage of reads with the different variants.

## 5.2 General Performance

### 5.2.1 Sequencing depth

The 20 pairs of tumor DNA samples were sequenced on 318 v2 chips, two pairs at a time. The resulting mean sequencing depth was 5467 (the mean number of reads covering each base), with a range from 3670-7782 (figure 29). The three blood samples, as well as the benign tumor, were sequenced with lower depth of coverage, with a mean of 2060 (range 1558-2820).

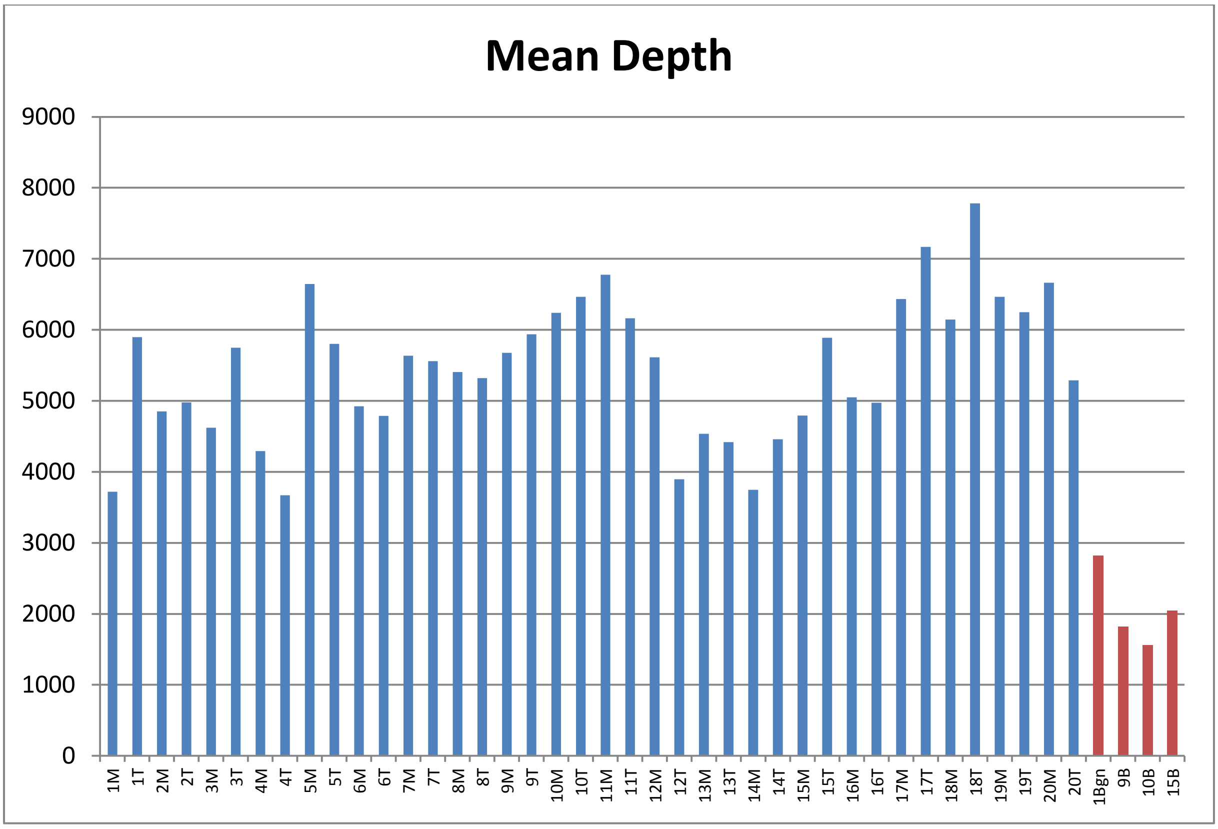
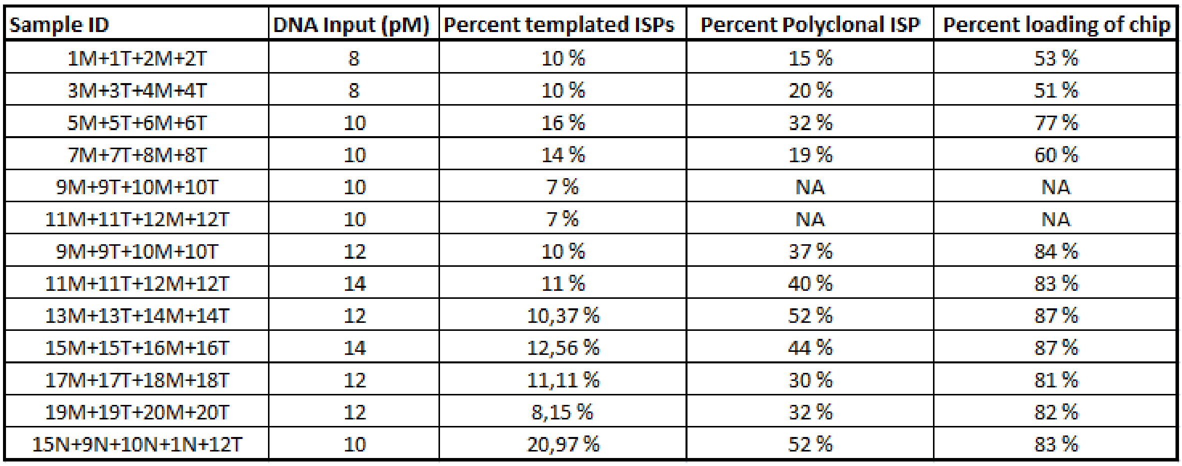


**Figure 29**: A bar graph displaying the mean sequencing depth of the different DNA libraries. The tumor samples are colored in blue and the normal samples (the benign tumor followed by the three blood samples) are colored in red.

### 5.2.2 Chip loading and percentage templated Ion sphere particles

The number of templated ISPs and the number of polyclonal ISPs are important for the sequencing result, as they will affect the number of useable reads generated. The acceptance criteria for un-enriched templated ISPs are 10-30%, as this range generally produces the most data. If the percentage is below 10, then the sample is said to have an insufficient number of templated ISPs to achieve an optimal loading on the Ion Chip. A percentage of templated ISPs above 30 will yield a large amount of polyclonal ISPs with unusable reads. To be within the optimal range, the recommended input of DNA library was 8 pM. In the two first reactions, the percent templated ISPs were calculated to be only 10% (measured by the Qubit Fluorometer). Aiming to increase the percentage of templated ISPs, we chose to vary the concentration of the DNA library included in the amplification mix (10-14 pM). However, for most of the samples the percentage did not exceed 14% (average 12%, range 7-20.97%). By contrast, an expansion of the number of polyclonal ISPs was observed for many of the samples, as well as an increased loading density of the chip. The results are presented in table 16.

**Table 16:** The amount of DNA library (pM) added to the emulsion PCR, together with the calculated percentage of templated ISPs, percentage of polyclonal ISPs and the achieved loading of the chip. The DNA libraries from two primary tumor-metastasis pairs (T and M) were combined in each reaction.



| Sample ID | DNA Input (pM) | Percent templated ISPs | Percent Polyclonal ISP | Percent loading of chip |
|---|---|---|---|---|
| 1M+1T+2M+2T | 8 | 10 % | 15 % | 53 % |
| 3M+3T+4M+4T | 8 | 10 % | 20 % | 51 % |
| 5M+5T+6M+6T | 10 | 16 % | 32 % | 77 % |
| 7M+7T+8M+8T | 10 | 14 % | 19 % | 60 % |
| 9M+9T+10M+10T | 10 | 7 % | NA | NA |
| 11M+11T+12M+12T | 10 | 7 % | NA | NA |
| 9M+9T+10M+10T | 12 | 10 % | 37 % | 84 % |
| 11M+11T+12M+12T | 14 | 11 % | 40 % | 83 % |
| 13M+13T+14M+14T | 12 | 10,37 % | 52 % | 87 % |
| 15M+15T+16M+16T | 14 | 12,56 % | 44 % | 87 % |
| 17M+17T+18M+18T | 12 | 11,11 % | 30 % | 81 % |
| 19M+19T+20M+20T | 12 | 8,15 % | 32 % | 82 % |
| 15N+9N+10N+1N+12T | 10 | 20,97 % | 52 % | 83 % |

As we had to adjust the amount of DNA analyzed, it was important to investigate whether the variation of DNA input had an impact on the downstream analyses. A Pearson correlation test was performed to examine the pair-wise correlation between DNA input, percentage of templated ISPs, percentage of polyclonal ISPs, chip loading density and the percentage usable reads generated. A scatterplot of the data is displayed in figure 30 together with the calculated p-values. There was no correlation between the DNA input and the percentage of templated beads (p-value=0.48) or between templated beads and the percentage of polyclonal beads (p-value=0.4). In contrast, a correlation between the input of DNA library and polyclonal beads (p-value=0.04) seemed to be present, and an increased number of polyclonal beads was clearly reducing the number of usable reads generated (p-value<<0.01). Further, increased DNA input lead to an increased loading density on the chip (p-value<<0.01). However, there was no correlation between the percentage of templated ISPs and loading density (p-value=0,81), in fact a loading density above 80% was achieved even when only 8,15 % of the ISPs were measured to be templated.

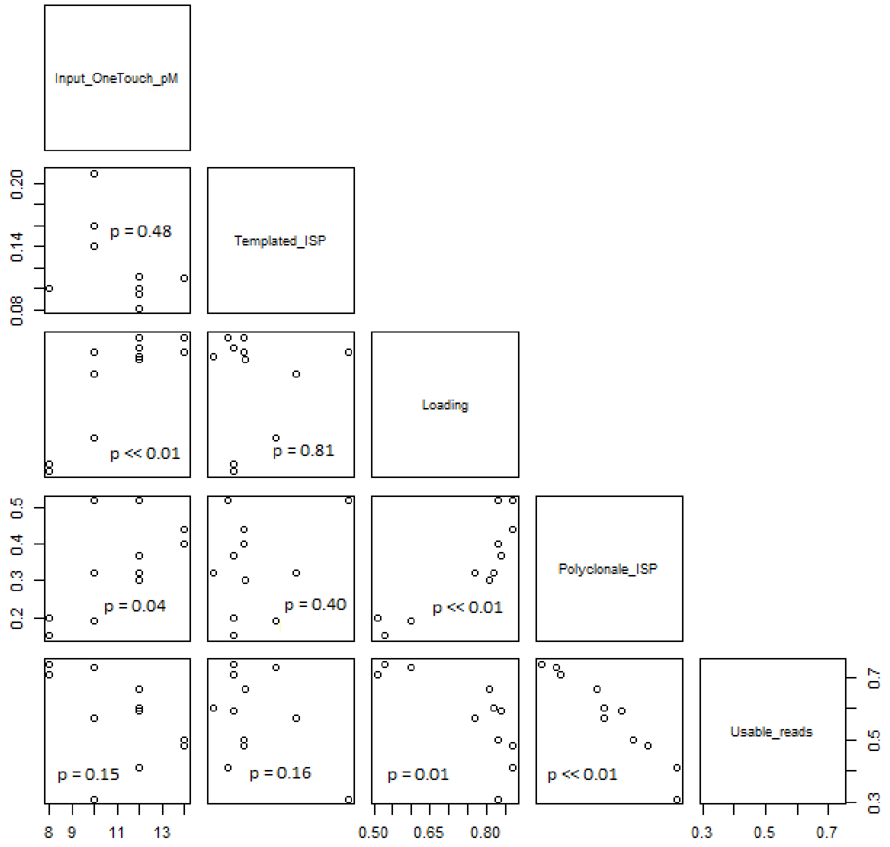
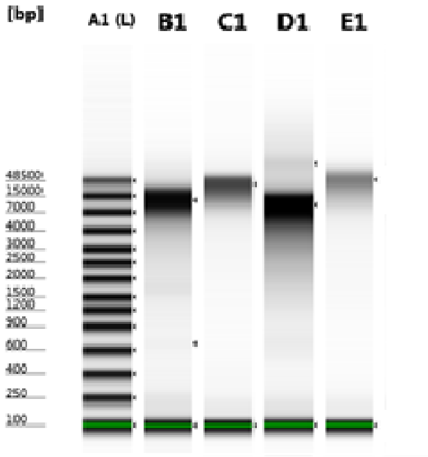


**Figure 30:** Scatterplot displaying the possible pair-wise interaction between the input of DNA library (Input_OneTouch_pM), the percentage of templated ISPs (Templated_ISP), the percentage of the chip to be loaded with ISPs (Loading), the percentage of polyclonal ISPs (Polyclonal_ISP) and the percentage of usable reads generated (Usable_reads). The associated p-values are shown in each individual plot.

### 5.2.3 DNA quality

Due to rarely having templated ISPs above 15 %, the quality of the DNA samples was questioned. Even though fragments of only 100-130 bp can be amplified when the CHP2 primer pool is utilized, a sample with significantly fragmented DNA will result in a poor DNA library. If the DNA strands are broken in between the two primer sites covering the ends of an amplicon, difficulties ligating an adapter region to both ends of the fragments will be experienced. Such fragments will not be attached to Ion Sphere Particles, and the fluorescent probes will not be able to hybridize to the DNA in advance of the Qubit measurements. To measure the integrity of the DNA the Genomic ScreenTape produced by Agilent

Technologies was utilized. Four of the samples were examined, and the results are presented in figure 31-32. The test did not indicate any significant fragmentation of DNA, as all the DNA fragments were gathered at the top of the gel image in figure 31 and a well-defined peak was observed in the electropherogram displayed in figure 32. Highly degraded DNA would have appeared as a smear in the gel image and as a low, broad peak in the electropherogram.



| Well | Sample Description |
|---|---|
| A1 | Ladder |
| B1 | 13M |
| C1 | 9M |
| D1 | 5M |
| E1 | 15T |

**Figure 31:** Gel image after performing a fragmentation test with Genomic ScreenTape. The four DNA samples (well B-E) do not seem to be fragmented, as they are gathered in the top of the gel image. The ladder in lane A1 indicates the lengths of the different fragments.

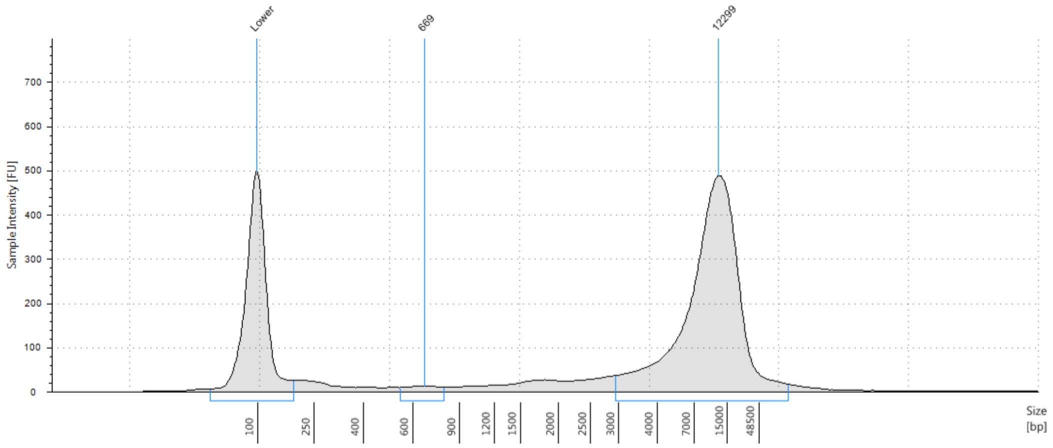


**Figure 32:** Electropherogram displaying the result of the genomic ScreenTape fragmentation test of sample 13M. A well defined peak is observed at 12 299 bp, indicating that the DNA to be tested was mainly long fragments. The lower peak at 100 bp is the lower marker. The electropherograms for the other samples displayed an almost identical peak.

# 6 Discussion

The aim of this thesis has been to study somatic mutations in primary breast tumors and corresponding lymph node metastases to see if they indicate how the metastatic process has progressed for each patient. The following section consists of two main parts. First, the biological considerations regarding the results will be discussed, and then a discussion of the general performance of the laboratory work and methodological considerations will follow.

## 6.1 Interpretation of variation

### 6.1.1 Matched normal sample

To identify somatic mutations in cancer, it is important to compare the DNA sequence from the tumor samples with the normal DNA sequence from the same individual, to prevent germ line variations being falsely considered as somatic mutations in cancer[68]. However, tumor samples collected for research do not always have normal cells available and there is a need to have other options to identify somatic mutations.

There is a vast amount of information about both normal and cancer genomes available in published scientific literature and public databases such as 1000 genome, COSMIC and dbSNP. The 1000 genomes project was an international research project aiming to create a detailed catalogue of human genetic variations. Beginning with the sequencing of the genome of 1000 anonymous participants[94], the genomes of over 2500 individuals sampled from different ethnic groups are today available[88]. The presence of a variant from the 1000 genomes database in a tumor sample is thus an indication that it is a common germ line variant, and that the variant is tolerated. Looking at the reports from the four patients with a matched normal sample available, this seems to be correct interpretations. However, as the 1000 genome database holds SNP information from a limited number of individuals, there might be a vast amount of SNPs not present in the database. Hence, any variants that were not matched with a SNP in 1000 genomes could not consequently be annotated as somatic mutations.

The catalogue of Somatic Mutations in Cancer (COSMIC) is a comprehensive database combining cancer mutation data manually selected from the scientific literature, with datasets from TCGA and ICGC. COSMIC contains more than 2 000 0000 coding mutations frequently found in different cancers[89,95]. The variants may have a possible impact on cancer development but not all of them will be driver mutations, the majority probably represents passenger mutations. It is also important to acknowledge that the COSMIC database is not filtered for SNPs[89], and some of the registered variants may be common polymorphisms in the population rather than somatic mutations in cancer. This is evident when studying the sequencing reports from the patients with matched normal samples. Quite a few of the variants present in the tumors are matched with both COSMIC and 1000 genomes. As all of these variants also were found in the normal sample, they are germ line variants rather than somatic mutations.

The tumor variants found in COSMIC only, could be germ line variants that is absent from the SNP databases due to rarity. To confirm these variants as somatic mutations, a matched normal sample is necessary. Even though a lot of genomic information is available, we do not have a complete knowledge of all the variations in the human genomes. Meyerson et al. point out that so far, each matched normal cancer genome to be sequenced has identified significant numbers of variants in the germ line that has not previously been described[68]. However, the percentage of reads with a variant will give an indication of whether it is a somatic mutation or a polymorphism. If the read percentage is 100, i.e. all reads show the variant, it is very likely a germline variant. The same is true for variants covered by 50% of the reads. The percentage of reads with somatic mutations will clearly deviate from these two values due to the heterogeneity of the tumor samples (discussed in more detail in section 6.1.3).

Studying the reports from the patients with a matched normal sample, we found that COSMIC and 1000 genomes complement each other reasonably well. None of the variants present in the tumor that were found in COSMIC only, were in fact present in the normal samples. However, this study analyzed only a small part of the genome, representing some of the most characterized cancer associated genes. The limited number of genes made it possible to check existing knowledge in the literature about every variant that were found in COSMIC, assuring us that the variants only represented somatic mutations. If a larger set of genes were studied (for instance exome sequencing or whole genome sequencing), it would be very important to have matched germline DNA sample for each case as literature search of each variant would be impossible.

### 6.1.2 Recurrent mutations

*PIK3CA* and *TP53* are the most frequently mutated genes in breast cancer, and were also the genes harboring most of the mutations in this cohort. All but one of the *TP53* mutations were missense mutations in the coding regions of the gene, which is consistent with the current release of the International Agency for Research on Cancer (IARC) *TP53* database (http://www-p53.iarc.fr/), saying that ~70% of the breast cancer alterations in *TP53* are missense mutations[16]. However, while about 30% of breast cancers have been reported to have *TP53* mutations[96], 45% of the primary tumors in this study harbored mutations in the gene. The increased mutation frequency may be caused by the selection bias present in the cohort. The cohort represents a non-consecutive collection of tumors, which do not give a representative picture of breast carcinomas in general. Patients were chosen based on sample availability, and only a selection of the breast cancer patients identified during the collection period were included. Locally advanced tumors are frequently found in the cohort, as excess material from small tumors is difficult to obtain for research. Today, ER-positive tumors are the most common type of breast carcinomas diagnosed, constituting 86% of breast cancers in Norway in 2013[97]. As the PriMet cohort consists of equal amounts of ER-positive and negative breast carcinomas, ER-negative tumors are over represented.

Substantial differences in mutation patterns and activated pathways have been revealed between ER-positive and ER-negative carcinomas, and to some extent the two tumor types can be viewed as separate diseases[31,43]. ER status is tightly linked to the molecular subtypes and a significantly higher TP53 mutation rate has been demonstrated in basal-like (mainly ER-negative) and HER2-enriched (both ER negative and positive) tumors than in the luminal types (mainly ER-positive)[43,96]. The high *TP53* mutation rate in the cohort can thus be explained by the high number of ER-negative cases as 67% of the observed *TP53* mutations occurred in the ER-negative tumors.

*PIK3CA* mutations are reported to occur in 20%-40% of breast carcinomas[46,98,99], and the gene was found to be mutated in 25% of the primary tumors in the cohort. Several studies of breast cancer suggest that *PIK3CA* mutations are more frequent in ER positive and HER2 positive cancers[46,98], and again the low mutation rate can be explained by the bias in subtype distribution of the cohort. We found that 2 out of 5 (40%) of the *PIK3CA* mutations occurred in ER-positive primary tumors. The small size of the cohort may have influenced this distribution.

The mutations *AKT1*:p.E17K and *GNAS*:p.R201H were revealed in one of the patients in the cohort (5%). *AKT1* mutations are reported to be present in 3% of breast cancers in COSMIC (cancer.sanger.ac.uk), and the frequency of E17K among the *AKT1*-mutated breast cancers are 90%[89]. *GNAS* is reported to be mutated in only 0,4% of breast cancers in COSMIC.

In this cohort, we did not detect any somatic mutations in 40 % of the primary tumors. It is important to acknowledge that we only analyzed a subset of genes, and it will be necessary to make a much more detailed sequencing of the samples. The hotspot panel is not designed for breast cancer in particular and includes regions and genes that are more frequently mutated in other cancer types. As the aim of this thesis was to investigate the evolution of the carcinomas from primary tumor to metastasis, these cases are so far uninformative.

### 6.1.3 Tumor cell percentage and allele frequency

Tumor samples contain a mixture of normal and malignant cells, and the DNA extracted from the samples will consequently be a mixture of normal and cancer genomes[68]. Further, several tumors consist of multiple subclones recognized by distinct mutations, ploidy and/or copy number alterations[51,52,65]. Due to such issues, somatic mutation calling is more complex than germ line variant calling.

The purity of the samples must be taken into account when the allele frequencies of mutations are analyzed. Tumor cell percentage can be determined by different approaches, in this study we had it estimated both by visual counting by a pathologist and by using genome wide allele frequencies based on SNP array analyses. There were certain discrepancies between the two tumor cell percent calculations. The estimation by pathologists on HE-stained slides is visual and subjective, and is generally not thought to be that accurate[100]. This estimate was used to eliminate samples that did not hold a significant amount of cancer cells in advance of molecular analysis. In contrast, the Battenberg algorithm using genome wide SNP information performs a calculation based on the phased LogR- and BAF-values, which give a more precise estimation. However, the algorithm is not optimal for cases of low purity (few tumor cells), and inaccurate calculations can occur.

Many of the mutations identified were present at higher or lower proportion of reads than would have been expected knowing the level of normal cells in the samples. A heterozygous mutation should have an allele frequency approximately half the calculated tumor percentage. Some of the variation may be due to aneuploidy or copy number alterations. For example, if only one of the two parental chromosomes has been duplicated, a mutation on the duplicated chromosome will contribute twice the number of sequenced reads than a mutations on the un-duplicated chromosome[101]. Loss of heterozygosity (LOH) is frequently found in cancer. A deletion of the remaining normal allele of a tumor suppressor gene is a classic mechanism of carcinogenesis, initially hypothesized by Knudson in his two-hit theory[102]. *TP53* mutations are often accompanied by loss of the wild-type allele in breast cancer[96], and if this occurs a mutation on the remaining allele will have a higher than expected read frequency.

Mutations that were present at a lower proportion of reads than expected may also be subclonal mutations, found only in a fraction of the cells. To investigate this, algorithms identifying tumor subclonality could be useful. Battenberg is one such algorithm, which utilizes differences in allele frequencies to recognize alterations that are present in a subset of tumor cells in a sample. The output returns the allele specific copy number of different chromosomal segments in two major subclonal populations, together with the calculated ratios of the two populations. However, the method assumes no more than two subclones to be present in a sample and falls short in cases where the tumor consists of multiple subclones. Also, the algorithm only demonstrates the presence of subclonal copy number alterations, and the relative order of occurrence between a somatic mutation and a CNA are not known. Li and Li describes three main scenarios that display possible mutation-CNA-combinations when a mutation occurs in a copy number increased region (Figure 33)[101]. Knowing the copy number of a chromosomal segment will not be sufficient for calculating the somatic variant allele frequency, as any mutation in the region may have occurred before (figure 33A) or after (figure 33B) the copy number alteration took place, as well as in a different cell lineage (figure 33C). In either case, different sub-populations with different variant allele fractions may coexist.

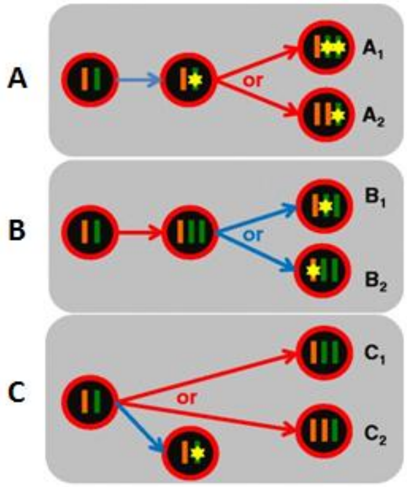


**Figure 33:** Different scenarios for a mutational event (yellow star) occurring in a region of heterozygous amplification. Scenario **A**: The mutation occurred first, followed by a CNA that doubled the mutation bearing chromosome (**$A_1$**) or the un-mutated chromosome (**$A_2$**). Scenario **B**: The amplification occurred first, and the following mutation occurred either on the amplified chromosome (**$B_1$**) or the un-amplified chromosome (**$B_2$**). Scenario **C**: The mutation and the CNA occurred independently in different cells, and the amplification affected one (**$C_1$**) or the other (**$C_2$**) chromosome. Blue arrow indicates mutational event, red arrow indicates CNA occurrence. Modified from[101].

All together, the scenarios display some of the complexity involved in an interpretation of the read frequencies. In addition, the percentage of reads with the different variants does seldom reflect the true allelic distribution. PCR amplification biases may occur during library preparation leading to an unbalanced sampling of parental alleles, or there might be an uneven distribution of DNA fragments attached to the ISPs. Some fragments may also be filtered out during base calling due to low quality. A correct interpretation of the read frequencies of the mutations will therefore be difficult without further analysis.

### 6.1.4 Comparison of primary tumor and metastases

#### 6.1.4.1 Concordant pairs

The majority of the patients had the same variants present in the primary tumor and in the metastasis. Among the 12 tumor pairs where somatic mutations were revealed, 9 had identical mutations in the primary tumor and the metastasis. Three of these patients seemed to have approximately equal allele frequencies in the two tumors (figure 27). Minor differences were evident for the other pairs. For some of the paired samples the variant read frequency ratio was above one, and higher than the tumor percent ratio, indicating that the allele frequency was slightly higher in the metastasis than the primary tumor (patient 1, 6, 5 and 12). In other samples the variant read frequency ratio was below 1, and lower than the tumor percent ratio, indicating that the allele frequency was slightly higher in the primary tumor than in the metastasis (patient 9 and 3). The allele frequencies were in many cases approximately equal to, or higher than, the tumor cell percentage in tumors, suggesting that the mutations were shared by the majority of the cancer cells. The discrepancies indicate that these pairs probably have differences in copy number for the genes with the mutations. However, as we have only analyzed the DNA from a single biopsy, representing a fraction of the tumor cells, some variations in allele frequencies are to be expected. Also, as only minor differences are detected for most of the patients, it is important to acknowledge the fact that technical artifacts and noise in the data may have generated deviations in allele frequencies (discussed in section 6.1.3). It will not be possible to draw any conclusions about the differences before copy number data have been analyzed.

The resemblance between the tumor pairs provides support to the idea that some carcinomas have a linear progression. The findings are consistent with a study of copy number alterations in primary breast tumors and distant metastasis, that found the vast majority of copy number alterations detected in the primary tumors to be retained in their corresponding metastasis[63]. Two recent studies performing whole genome sequencing of a primary tumor and a matched metastasis, demonstrated largely overlapping gene alterations in the tumor, but also some mutations and structural variants unique to the primary tumor and the involved lymph node[64,103]. It is important to have in mind that this study only investigated a few genes, and the probability of detecting minor differences between the pairs is therefore low.

#### 6.1.4.2 Discordant pairs

In three of the patients in the cohort (patient 7, 15 and 16), *TP53* mutations unique to the primary tumor and/or the metastasis were revealed. As they are all known mutations in breast cancer from the IARC TP53 database[16], there is reason to believe that they are of importance for the progression of the disease. Various mechanisms may have caused these differences. Figure 34 presents four hypothetical models explaining the presence of the different *TP53* mutations in the primary tumor and metastasis of patient 7. The concepts are also applicable for patient 15 and 16, and will be discussed in the following sections.

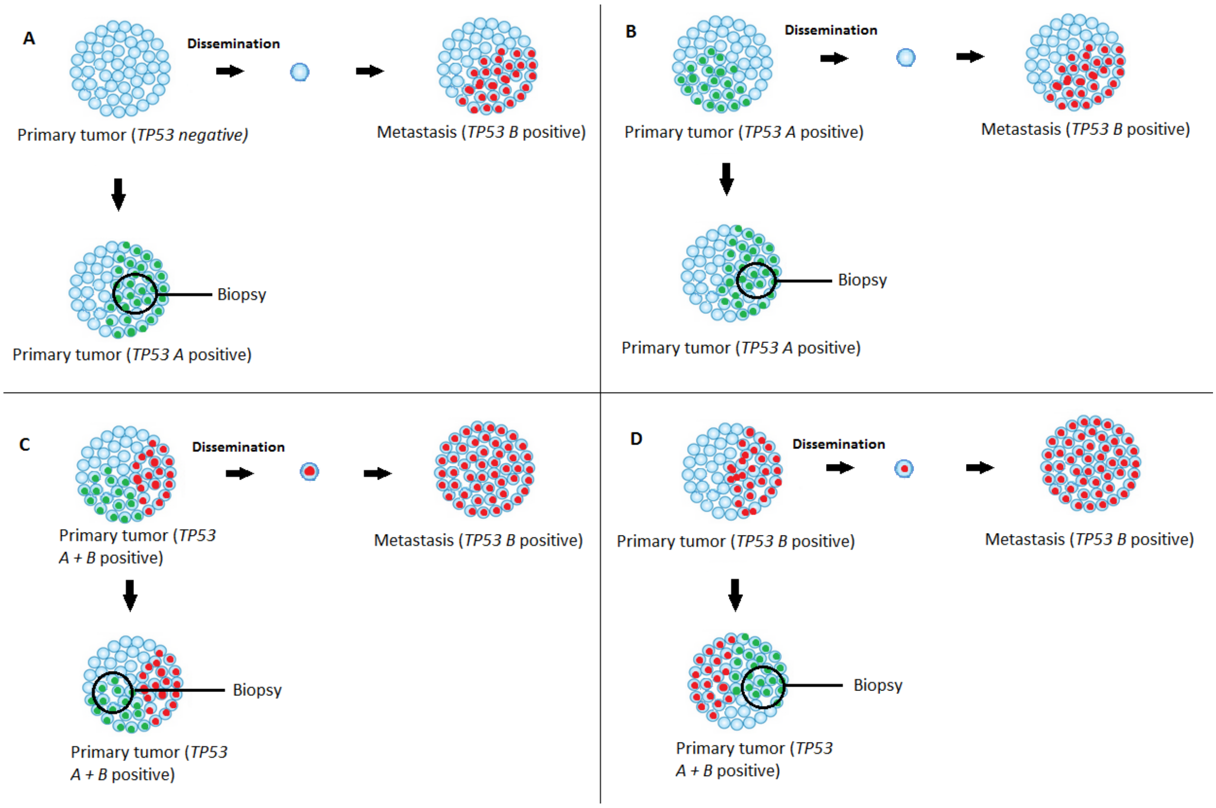

A
Dissemination
Primary tumor (*TP53 negative)*
Metastasis (*TP53 B* positive)
Biopsy
Primary tumor (*TP53 A* positive)
B
Dissemination
Primary tumor (*TP53 A* positive)
Metastasis (*TP53 B* positive)
Biopsy
Primary tumor (*TP53 A* positive)
C
Dissemination
Primary tumor (*TP53 A + B* positive)
Metastasis (*TP53 B* positive)
Biopsy
Primary tumor (*TP53 A + B* positive)
D
Dissemination
Primary tumor (*TP53 B* positive)
Metastasis (*TP53 B* positive)
Biopsy
Primary tumor (*TP53 A + B* positive)

**Figure 34:** Four hypothetical models describing the evolution of the primary tumor and metastasis of patient 7. The two different *TP53* mutations, named A and B, are indicated by green and red cells respectively. **A**: Mutation A and B occurred independently in the primary tumor and the metastasis, after the metastatic cell left the primary site. **B**: Mutation A was present in a subclone of the primary tumor that did not give rise to the metastasis. Mutation B occurred in the metastasis after dissemination. **C**: Both mutations were present in different subclones of the primary tumor, but only one was detected in the biopsy. The subclone holding mutation B gave rise to the metastasis. **D**: The primary tumor only held mutation B at the time of dissemination. At the time of diagnosis both mutations were present, but only one was detected in the biopsy.

As illustrated in figure 34A, a parallel evolution of genetic alterations in the primary tumor and the metastasis may have caused mutational differences. The time of dissemination of cancer cells from the primary tumor and into the circulations will be decisive for which mutations are carried on to the metastasis. A mutation that is only present in the primary tumor may have occurred after the metastatic “founder” cell disseminated from the primary tumor. By assuming that mutational complexity increases with time, phylogenetic methods can be used to reconstruct the relative order of occurrence of mutations. Studies of tumor heterogeneity assume that mutations shared by several subpopulations in a tumor are early events, that occurred before their divergence[55]. Several studies have attempted to identify the stage of breast tumorgenesis at which a somatic *TP53* mutation occurs, and *TP53* seem to be altered early in cancer development[104]. These mutations are thus likely to be shared by the majority of cancer cells in the tumor. The presence of a metastasis without the *TP53* mutation could therefore indicate that the metastatic founder cell left the primary tumor at an early stage of cancer development, before the gene was mutated and was a driver mutation leading to a dominating clone in the primary tumor at time of diagnosis.

The research of Schmidt-Kittler et al.[67] provides support to this observation. By performing single-cell hybridization analysis, they compared alterations in tumor cells disseminated in the bone marrow with alterations in areas of matched primary breast tumors, and found DTCs to hold fewer and different aberrations than the cells from the primary tumor. Their findings suggest that tumor cells can disseminate from primary tumors in a much less progressed state than previously thought. Another study, performing immunostaining of p53 protein on micrometastatic tumors in the bone marrow, indicated that a minority of the cancer patients had *TP53* mutations in their DTCs[105]. *TP53* sequence analyses of DTCs supported this finding[106], suggesting that *TP53* mutations are not required for early tumor cell dissemination.

If an early dissemination occur and the DTCs develop in parallel to the primary tumor over prolonged time; molecular targets on the DTCs should be identified to prevent metastases from arising. The high frequency of *TP53* mutations in human cancers makes it a potential target of cancer therapies. However, the differences observed in *TP53* mutation status between the primary tumor and the metastasis suggest that it might not be a successful strategy trying to eliminate metastatic cancer cells based on the genotype of the primary tumor. Whether all tumors disseminate cells at an early stage, and what exact link there is between these DTCs and the metastases, is not yet clarified[4].

Evolution at the metastatic site will also lead to differences between the paired tumors (figure 34 A-B). The *TP53* mutations revealed in the metastasis only (patient 15 and 7) may have occurred in the metastatic cell population after dissemination. Similar findings were revealed by Shah et al., who performed whole genome NGS of the genome of a metastatic breast tumor and the corresponding primary tumor surgically removed nine years earlier[66]. Only 11 of the 30 evaluated mutations were detected in the DNA of the primary tumor, suggesting that the remaining 19 mutations were acquired after dissemination. However, it is not known whether these mutations were a consequence of radiation therapy or intrinsic tumor progression.

The *TP53* mutation found in the metastasis of patient 15 had a read frequency of only 15 percent. With a calculated tumor percent above 60%, it might seem as though it is a subclonal mutation. Even though a metastatic tumor may have a clonal origin, genetic instability and environmental influences may lead to heterogeneous subpopulations of cells[63]. If substantial genetic evolution occurs in the metastatic process, the primary tumor alone will not reveal the genotype of the metastasis, and metastatic diseases that are resistant to cancer therapies may evolve.

Subclonality at the primary site may also explain the differences between the tumors. Various subclonal compositions of the primary tumor could yield different metastatic founder cells, some carrying the mutation from the primary site and others that do not (figure 34 B, C and D). Sequencing-based studies rely on representative sampling of tumors. As only one biopsy was taken from each tumor in the cohort, certain subclonal mutations may have been impossible to detect if they dominated cells in different regions of the tumors. The mutations revealed in only one of the two tumors, could in fact be present in a subclone not included in the biopsy taken. This assumption is supported by the findings of Yates et al. who demonstrated a significant heterogeneous spatial distribution of point mutations in breast tumors[52]. Even though known driver mutations such as *TP53* and *PIK3CA* seemed to be clonally dominant compared to other genes, their study suggested that the intratumor genetic heterogeneity could affect also these genes. In an ER-positive/HER2-negative carcinoma, three separate subclonal lineages with different *TP53* mutations were revealed. The mutations were not the same as the ones detected in patient 7, but it might suggest that a minor subclone, holding the *TP53*:p.I195T missense mutation have given rise to the metastasis, while the *TP53*:p.G245C mutation was to be found in the major clone of the primary tumor, and thus the one detected in the biopsy.

Not included in the figure is the possibility that a cluster of cancer cells have given rise to the metastasis. Even though the majority of cancer cells in the circulation are single cells, circulating tumor cell clusters have been found in the blood of breast cancer patients[107]. The contribution of these clusters to metastases is not well known, but modeling experiments indicate that CTC clusters arise because a group of cells are released from a tumor into the circulation and that they seem to have greater metastatic potential than single CTCs[108]. The fact that the metastasis of patient 7 was found in the contra lateral lymph node two years after removal of the primary tumor should be also taken into account. The tumor cells may have travelled through the circulatory system and had more time to evolve than the metastatic cells found in the axillary lymph node at the time of diagnosis.

Based on the limited number of mutations revealed in the cohort, it might seem as though some tumor pairs are more similar than others. This assumption is consistent with the analysis of allelic imbalance (AI) in primary breast carcinomas and matched axillary lymph node metastasis performed by Becker et al[58]. Hierarchical cluster analysis demonstrated that some metastases were genetically similar to primary tumor and that they shared a recent common ancestry, while others were genetically different from the primary tumor and appeared to have diverged from the primary carcinoma early in disease progression. These observations suggest that the process of breast cancer metastasis may be driven by several molecular mechanisms and that some cancer cells acquire metastatic potential early in tumorgenesis, while other metastatic tumors evolve later.

## 6.2 General performance

Next generation sequencing (NGS) is a comprehensive and sensitive sequencing method, and technical variability during the lab procedure can affect the results. In this study we used Ion Torrent sequencing technology. Library preparation is a key bottleneck in the process, together with the ligation of DNA fragments to the Ion Sphere Particles and the loading of the Ion chip. In addition, accuracy in the performance of variant calling is essential for credible results.

### 6.2.1 Tumor samples and DNA quality

The quality, purity and concentration of the DNA to be sequenced are important factors for the construction of a successful DNA library. Some tumors have only limited amounts of material available, and cancer specimens often include necrotic or apoptotic cells that will reduce the quality of the nucleic acids[68]. The tumor samples utilized in the study were collected several years ago as a dataset to be used in a validation project. The DNA was isolated using either the QIAcube robot and the AllPrep DNA spin columns or the QIAsymphony robot and the silica magnetic beads in the DNA mini Kit from Qiagen. High-purity nucleic acids for use in most downstream applications are said to be delivered, and there should not be any significant differences in the resulting DNA quality when using the two different methods[109].

Information about the DNA concentration and purity existed from previous measurements for most of the samples. The 260/280 and 230/260 values obtained by the nanodrop spectrophotometer were within the required range for a pure DNA sample. Based on the results from the fragmentation test performed of four of the samples, the DNA was assumed to be of high integrity for all samples.

### 6.2.2 Percent templated Ion Sphere Particles

The number of templated ISPs and the number of polyclonal ISPs are important variables affecting the amount of usable reads generated in the sequencing reaction. However, the assertion that a percentage of templated ISPs below 10 are insufficient to yield an optimal chip loading, was not supported by the results. Neither was the assertion that only a percentage of templated ISPs above 30 will yield a large amount of polyclonal ISPs. The percent templated ISPs was lower than optimal in most of the reactions (~10 %), but high density chip loadings were still achieved, as well as high numbers of polyclonal ISPs.

The results indicate that the optimal range provided by Life Technologies, should not be interpreted as a pass/fail criteria, but rather as guidance for the quality of the samples. The lack of correlation between the percentage of templated ISPs and polyclonal ISPs, as well as the apparent correlation between the chip loading density and percentage of polyclonal ISPs, indicate that the Qubit measurements might underestimate the percentage of templated ISPs. There are not supposed to be any free DNA in the solution to interfere with the measurements, as the pellet is washed several times before the sample is read by the Qubit Fluorometer. There might be some fluorophores binding non-specifically to the ISPs or getting caught in the pellet, however this should not affect the result significantly. In the end, the sequencing report seemed to be the most important guideline rather than the Qubit Fluorometer, a fact that was supported by the person in charge of technical issues from Life Technologies. If the percentage of polyclonal ISPs is as high as 50%, then the DNA input should be decreased in spite of low percent templated ISPs calculations, as a high amount of polyclonal beads will greatly reduce the number of usable reads.

### 6.2.3 Sequencing depth

The depth of coverage necessary to make accurate mutations calls will depend on the expected frequency of the mutation in the sample. In analyses of germ-line variants that follow Mendelian patterns of inheritance, an average depth of coverage of 100-200x would be sufficient. Somatic mutations present in heterogeneous cancer samples however, are typically present at low frequencies, and a higher coverage is thus necessary. A depth of coverage of 1000-2000x is recommended for Ion Torrent Amplicon sequencing of cancer samples[68,110].

To make sure a sufficient sequencing depth was achieved, only four samples were sequenced per 318 v2 chip. The resulting depth of coverage was above 3000x for all of the tumor samples. The variation between the samples may be caused by an uneven distribution of libraries on the chip. Some of the libraries may not have been as successfully attached to ISPs as others, or a higher amount of polyclonal beads may have occurred. There is only one calculation provided per chip, in other words the calculated percentage of templated and polyclonal ISPs is not specific for each library. In addition, variation between the individual amplicons may be caused by differences in primer efficiency.

### 6.2.4 Sequencing limitations and variant calling

Targeted sequencing provides a high coverage of regions of interest, at a significantly lower cost and with a higher throughput, than whole genome sequencing[68]. This makes targeted sequencing a suitable candidate when detecting mutations in cancer samples of mixed purity. Several studies have compared the different bench top sequencers that are available today[74,111]. The Ion Torrent PGM accurately detects single nucleotide variants, but is less useful for finding insertions and deletions and not reliable for studying copy number alterations or structural changes[112]. To get a more comprehensive overview of the cancer genomes in the cohort, additional analyses are necessary, such as exome or full genome sequencing.

The PGM has difficulties sequencing regions of highly repetitive base sequences, as well as the sequence at the very end of the amplicons[69], and incorrect registrations of base substitutions and deletions may occur. In the sequencing reports some of the variants were therefore annotated as common false positives, due to their position in a homopolymer region. During the conversion of ions to bases at the Ion Torrent server, a signal intensity equivalent to 3,5 identical bases may be detected, suggesting that the read sequence in this region could be “GGG” or “GGGG”. During the read alignment, if the reference sequence has three “G” bases in this region, an insertion could be called by the reads with four “G” bases[112]. This is probably the reason for some of the low frequent variants that were detected, and points to one of the main limitations of PGM sequencing.

Some variants, such as *KDR* chr4:55980239:C>T and *CSF1R* chr5:149433596: TG>GA, were annotated as false negatives during variant calling. They are known polymorphism in 1000 genomes, but situated at the very end of the amplicon, some sequencing difficulties are experienced. Base calling accuracy reduces as read length increases; partly because the signals become weak. The variants were either falsely absent from the sequencing reports, or present in a lower percentage of reads than to be expected at homozygous or heterozygous loci.

A strand bias (a disproportional number of plus and minus strands) was annotated in several of the variants. As the two DNA strands are complementary, true mutations are found at the same position in both strands and should occur on both + and – strands with equal frequencies. If a high fraction of the reads covering a specific variant is reported to have a strand bias, it is an indication of PCR or sequencing errors. The minimum strand bias accepted for variant calling was set to 0,05. However, low frequent variants will always need specific consideration. With a high depth of coverage and sufficiently statistical thresholds for mutation calling, inaccurate detection of variants will become limited.

# 7 Conclusion and future perspectives

Breast cancer is the most common malignancy among women. For early stage disease, treatment decisions are made based on clinical information and morphological and molecular analyses of the primary tumor. When adjuvant treatment is given, the main goal is to prevent eventual disseminated cells to form metastases. It is becoming more evident that tumor progression might not always be linear, meaning that metastases can evolve independently of the primary tumor. If so, analysis of metastatic tumors may provide additional insight affecting treatment decisions. To increase knowledge of the metastatic process this study investigated 20 pairs of primary tumors and lymph node metastases.

The extent of genomic heterogeneity between primary breast tumors and corresponding lymph node metastasis seem to differ among individual patients. The same somatic mutations were detected in the vast majority of the paired tumors. Some of these mutations were present at different frequencies in the primary tumor and the metastasis indicating subclonal variations or copy number alterations but the differences could also be due to technical artifacts during sequencing. For three patients we found a mutation to be unique to the primary tumor and/or the involved lymph node. Such differences may be explained by intratumor heterogeneity at the primary site (subclonal mutations may not have been included in the biopsy) or by a parallel evolution of genetic alterations in the primary tumor and the metastasis. However, as only specific regions of the genome were targeted for sequencing, the results give a very limited and simplified picture of the biology of the tumors. As more data now are available from other analyses of these tumors, the sequencing results are to be compared with copy number analyses. This was beyond the scope of this thesis, but will provide information necessary for a further interpretation of the variant allele fractions in the samples. The comparison will also reveal whether the pairs identified as discordant or concordant by targeted sequencing, display different or similar copy number profiles.

Based on the findings in this study, some samples are to be selected for further and more comprehensive analysis. Whole genome sequencing is to be performed to get a broader understanding of the genomes of the tumors. If tumor cells disseminate at an early stage of cancer development, the complexity of the primary tumor will not be reflected in the genome of the metastasis. The three patients in the discordant group will be interesting cases, to reveal if the genomes of the primary tumor and the metastasis truly are significantly different. It will also be of interest to select some of the other cases, to see if differences are detected when we study a larger number of genomic regions.

The results of the published studies referred to in the thesis are inconsistent, as some report of a close resemblance between the primary tumor and the metastasis, while others report of differences. The conflicting results reflect partly limitations of methods used but also the complexity involved in tumor progression. It is hard to grasp the complete picture of the metastatic process by studying only a selection of samples or specific genes, as multiple molecular mechanisms are probably involved. In addition, the metastatic process may not be uniform for all carcinomas. It is possible that some tumors disseminate at an early stage and some at a later stage, and that the degree of evolution differ. The resemblance between a primary tumor and matched metastasis may thus vary, and it will be important in a longer perspective to analyze hundreds of cases to fully understand this. The analysis within this thesis as well as the planned next steps of the project may be an important contribution to improve the current understanding of the metastatic process. It may also be a basis for planning of larger studies aiming at predicting risk of metastases as well as increasing tailoring of treatment regimens for individual patients.

# 8 References

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Genome Biol.* **458,** 719–724 (2008).

2. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, Taylor & Francis Group, 2008).

3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).

4. Wan, L., Pantel, K. & Kang, Y. Tumor metastasis: moving new biological insights into the clinic. *Nat. Med.* **19,** 1450–64 (2013).

5. Jin, X. & Mu, P. Targeting Breast Cancer Metastasis. *Lib. Acad.* **9,** 23–34 (2015).

6. Braun, S. & Naume, B. Circulating and disseminated tumor cells. *J. Clin. Oncol.* **23,** 1623–1626 (2005).

7. IARC. GLOBOCAN 2012: Estimated Incidence, Mortality and Prevalence Worldwide in 2012. (2012). at <http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx>

8. Cancer Registry of Norway. *Cancer in Norway 2013 - Cancer incidence, mortality, survival and prevalence in Norway*. (2013). doi:10.1136/bmj.1.5178.1031-a

9. Cancer Registry of Norway. *Cancer in Norway 2012 - Cancer incidence, mortality, survival and prevalence in Norway*. (2012). doi:10.1136/bmj.1.5178.1031-a

10. Mavaddat, N., Antoniou, A. C., Easton, D. F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. *Mol. Oncol.* **4,** 174–191 (2010).

11. He, Q. *et al.* Genome-wide prediction of cancer driver genes based on SNP and cancer SNV data. *Am. J. Cancer Res.* **4,** 394–410 (2014).

12. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45,** 353–61, 361e1–2 (2013).

13. Antoniou, a C. & Easton, D. F. Models of genetic susceptibility to breast cancer. *Oncogene* **25,** 5898–5905 (2006).

14. Conzen, S. D., Grushko. Tatyana A. & Olopade, O. I. in *CANCER - Principles & Practice of Oncology* (eds. DeVita Jr., V. T., Lawrence, T. S. & Rosenberg, S. A.) 1595–1604 (Lippincott Williams & Wilkins, 2008).

15. Foulkes, W. D. Inherited Susceptibility to Common Cancers. *N. Engl. J. Med.* **359,** 2143–2153 (2008).

16. Petitjean, A. *et al.* Impact of Mutant p53 Functional Properties on TP53 Mutation Patterns and Tumor Phenotype: Lessons from Recent Developments in the IARC TP53 Database. *Hum. Mutat.* **26,** 622–9 (2007).

17. Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nat. Genet.* **33 Suppl,** 238–244 (2003).

18. Trichopoulos, D., Adami, H.-O., Ekbom, A., Hsieh, C.-C. & Lagiou, P. Early life events and conditions and breast cancer risk: from epidemiology to etiology. *Int. J. Cancer* **122,** 481–485 (2008).

19. Russo, J. & Russo, I. H. Development of the human breast. *Maturitas* **49,** 2–15 (2004).

20. Allred, D. C., Mohsin, S. K. & Fuqua, S. a W. Histological and biological evolution of human premalignant breast disease. *Endocr. Relat. Cancer* **8,** 47–61 (2001).

21. Britt, K., Ashworth, A. & Smalley, M. Pregnancy and the risk of breast cancer. *Endocr. Relat. Cancer* **14,** 907–933 (2007).

22. Barnard, M. E., Boeke, C. E. & Tamimi, R. M. Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochim. Biophys. Acta - Rev. Cancer* **1856,** 73–85 (2015).

23. Yoder, B. J., Wilkinson, E. J. & Massoll, N. a. Molecular and morphologic distinctions between infiltrating ductal and lobular carcinoma of the breast. *Breast J.* **13,** 172–179 (2007).

24. Breast Cancer Basics and You: Introduction. *NIH Medlin. Plus* **5,** 17 (2010).

25. Polyak, K. Science in medicine Breast cancer : origins and evolution. *Cell* **117,** 3155–63 (2007).

26. Vuong, D., Simpson, P., Green, B., Cummings, M. & Lakhani, S. Molecular classification of breast cancer. *Virchows Arch* **465,** 1–14 (2014).

27. Vogelstein, B. & Kinzler, K. *The Genetic Basis of Human Cancer*. (McGraw-Hill Companies, 1998).

28. Breastcancer.org. Diagnosis of DCIS. (2015). at <http://www.breastcancer.org/symptoms/types/dcis/diagnosis>

29. Gannon, L., Cotter, M. & Quinn, C. The classification of invasive carcinoma of the breast. *Expert Rev. Anticancer Ther.* **13,** 941–954 (2013).

30. Le Romancer, M. *et al.* Cracking the estrogen receptor's posttranslational code in breast tumors. *Endocr. Rev.* **32,** 597–622 (2011).

31. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

32. Sørlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 8418–8423 (2003).

33. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10869–10874 (2001).

34. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J. Clin. Oncol.* **5,** 412–24 (2014).

35. Bernard, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

36. NBCG. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft. (2015). at <http://www.helsebiblioteket.no/retningslinjer/brystkreft/forord>

37. Thuerlimann, B., Koeberle, D. & Senn, H.-J. Guidelines for the adjuvant treatment of postmenopausal women with endocrine-responsive breast cancer: past, present and future recommendations. *Eur. J. Cancer* **43,** 46–52 (2007).

38. Caley, A. & Jones, R. The principles of cancer treatment by chemotherapy. *Surgery* **30,** 186–190 (2012).

39. Carpenter, R. & Miller, W. The role of aromatase inhibitors in breast cancer. *Britishh J. cancer* **93,** 51–55 (2005).

40. Chmielecki, J. & Meyerson, M. DNA sequencing of cancer: what have we learned? *Annu. Rev. Med.* **65,** 63–79 (2014).

41. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331,** 1553–1558 (2011).

42. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

43. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

44. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149,** 979–993 (2012).

45. Walerych, D., Napoli, M., Collavin, L. & Del Sal, G. The rebel angel: mutant p53 as the driving oncogene in breast cancer. *Carcinogenesis* **33,** 2007–2017 (2012).

46. Cizkova, M. *et al.* PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups. *Breast Cancer Res.* **14,** R28 (2012).

47. Kops, G. J. P. L., Weaver, B. a a & Cleveland, D. W. On the road to cancer: aneuploidy and the mitotic checkpoint. *Nat. Rev. Cancer* **5,** 773–785 (2005).

48. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107,** 16910–16915 (2010).

49. Bergamaschi, A. *et al.* Distinct patterns of DNA Copy Number Alterations Are Associated with Different Clinopathological Features and Gene-Expression Subtypes of Breast Cancer. *Genes. Chromosomes Cancer* **45,** 1033–1040 (2006).

50. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–64 (2009).

51. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** 994–1007 (2012).

52. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21,** (2015).

53. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).

54. Martinez, P. *et al.* Parallel evolution of tumour subclones mimics diversity between tumours. *J. Pathol.* **230,** 356–364 (2013).

55. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).

56. Russnes, H. G., Navin, N., Hicks, J. & Borresen-Dale, A. L. Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest.* **121,** 3810–3818 (2011).

57. Röcken, M. Early tumor dissemination, but late metastasis: Insights into tumor dormancy. *J. Clin. Invest.* **120,** 1800–1803 (2010).

58. Becker, T. E. *et al.* The genomic heritage of lymph node metastases: implications for clinical management of patients with breast cancer. *Ann. Surg. Oncol.* **15,** 1056–1063 (2008).

59. Klein, C. A. Parallel progression of tumour and metastases. *Nat. Rev. Cancer* **9,** 301–312 (2009).

60. Engel, J. *et al.* The process of metastasisation for breast cancer. *Eur. J. Cancer* **39,** 1794–1806 (2003).

61. Sänger, N. *et al.* Disseminated tumor cells in the bone marrow of patients with ductal carcinoma in situ. *Int. J. Cancer* **129,** 2522–2526 (2011).

62. D’Andrea, M. R. *et al.* Correlation between genetic and biological aspects in primary non-metastatic breast cancers and corresponding synchronous axillary lymph node metastasis. *Breast Cancer Res. Treat.* **101,** 279–284 (2007).

63. Moelans, C. B. *et al.* Genomic evolution from primary breast carcinoma to distant metastasis: Few copy number changes of breast cancer related genes. *Cancer Lett.* **344,** 138–46 (2014).

64. Blighe, K. *et al.* Whole Genome Sequence Analysis Suggests Intratumoral Heterogeneity in Dissemination of Breast Cancer to Lymph Nodes. *PLoS One* **9,** e115346 (2014).

65. Torres, L. *et al.* Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res. Treat.* **102,** 143–155 (2007).

66. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461,** 809–813 (2009).

67. Schmidt-Kittler, O. *et al.* From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 7737–7742 (2003).

68. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11,** 685–696 (2010).

69. Meldrum, C., Doyle, M. a & Tothill, R. W. Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective. *Clin. Biochem. Rev.* **32,** 177–195 (2011).

70. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** (2012).

71. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **6,** 287–303 (2013).

72. Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G. & Stahl, F. Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.* 22–30 (2012). doi:10.1016/j.copbio.2012.09.004

73. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** (2014).

74. Quail, M. *et al.* A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13,** 1 (2012).

75. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11,** 759–769 (2011).

76. Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Lett.* **340,** 284–295 (2013).

77. Merriman, B., Torrent, I. & Rothberg, J. M. Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis* **33,** 3397–3417 (2012).

78. Cheng, W.-Y., Ou Yang, T.-H. & Anastassiou, D. Development of a prognostic model for breast cancer survival in an open challenge environment. *Sci. Transl. Med.* **5,** 181ra50 (2013).

79. Baumbusch, L. O. *et al.* Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* **9,** 379 (2008).

80. Thermo Fischer Scientific. Ion AmpliSeq$^{TM}$ Cancer Hotspot Panel v2. (2015). at <https://www.thermofisher.com/order/catalog/product/4475346>

81. ThermoFischer Scientific. Essensials of Real-Time PCR. (2015). at <http://www.thermofisher.com/us/en/home/life-science/pcr/real-time-pcr/qpcr-education/essentials-of-real-time-pcr.html>

82. Chee-seng, K., Yun, L. E., Yudy, P. & Kee-Seng, C. in *Encyclopedia of life sciences* 1–12 (John Wiley & Sons, 2010). doi:10.1002/9780470015902.a0022548

83. Thermo Fischer Scientific. Technical Note - Trimming and Filtering. (2015). at <http://mendel.iontorrent.com/ion-docs/Technical-Note---Filtering-and-Trimming_6455370.html>

84. Ewing, B. *et al.* Base-Calling of Automated Sequencer Traces Using. *Genome Res.* 186–194 (1998). doi:10.1101/gr.8.3.175

85. Thermo Fischer Scientific. Technical note - The Per-Base Quality Score System. (2015). at <http://mendel.iontorrent.com/ion-docs/Technical-Note---Quality-Score_6128102.html>

86. Thermo Fischer Scientific. Torrent Browser Analysis Report Guide. (2015). at <http://mendel.iontorrent.com/ion-docs/Torrent-Variant-Caller-Plugin.html>

87. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).

88. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

89. Forbes, S. a. *et al.* COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2014).

90. Kitts, A., Phan, L., Minghong, W. & Holmes, J. B. in *The NCBI Handbook* (Bethesda (MD): National Center for Biotechnology Information (US), 2013). at <http://www.ncbi.nlm.nih.gov/books/NBK174586/>

91. Dam, V., Morgan, B. T., Mazanek, P. & Hogarty, M. D. Mutations in PIK3CA are infrequent in neuroblastoma. *BMC Cancer* **6,** 177 (2006).

92. Lovly, C. M., Horn, L. & Pao, W. AKT1 c.49G>A (E17K) Mutations in Non-Small Call Lung Cancer (NSCLC). *My Cancer Genome* (2015). at <http://www.mycancergenome.org/content/disease/lung-cancer/akt1/23/>

93. Pérez, D. N. G., Mantovani, G. & Fernandez-Rebollo, E. GNAS (GNAS complex locus). *Atlas Genet. Cytogenet. Oncol. Haematol.* **17,** 178–187 (2013).

94. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

95. Forbes, S. a *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–50 (2011).

96. Silwal-Pandit, L. *et al.* TP53 mutation spectrum in breast cancer is subtype specific and has distinct prognostic relevance. *Clin. Cancer Res.* **20,** 3569–3580 (2014).

97. The Cancer Registry of Norway. *Årsapport 2013-2014 Brystkreft*. (2015). at <http://kreftregisteret.no/Global/Publikasjoner og rapporter/%C3%85rsrapporter/2015/aarsrapport_2015_brystkreft.pdf>

98. Saal, L. H. PIK3CA Mutations Correlate with Hormone Receptors, Node Metastasis, and ERBB2, and Are Mutually Exclusive with PTEN Loss in Human Breast Carcinoma. *Cancer Res.* **65,** 2554–2559 (2005).

99. Stemke-Hale, K. *et al.* An Integrative Genomic and Proteomic Analysis of PIK3CA, PTEN, and AKT Mutations in Breast Cancer. *Cancer Res.* **68,** 6084–6091 (2008).

100. Smits, A. J. J. *et al.* The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Mod. Pathol.* **27,** 168–174 (2014).

101. Li, B. & Li, J. Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.* **15,** 473 (2014).

102. Knudson, a G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* **1,** 157–162 (2001).

103. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010).

104. Børresen-Dale, A. L. TP53 and breast cancer. *Hum. Mutat.* **21,** 292–300 (2003).

105. Offner, S. *et al.* P53 Gene Mutations Are Not Required for Early Dissemination of Cancer Cells. *Pnas* **96,** 6942–6946 (1999).

106. Klein, C. *et al.* Genetic heterogeneity of single disseminated tumour cells in minimal residual cancer. *Lancet* **360,** 683–689 (2002).

107. Yu, M. *et al.* Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition. *Science (80-. ).* **339,** 580–584 (2013).

108. Aceto, N. *et al.* Circulating Tumor Cell Clusters Are Oligoclonal Precursors of Breast Cancer Metastasis. *Cell* **158,** 1110–1122 (2014).

109. Schnibbe, T., Scherer, M., Scholle, N. & Lubenow, H. *Development of a new platform for fully automated purification of nucleic acids from a broad spectrum of forensic specimens*. (2008). at <https://www.qiagen.com/fr/resources/resourcedetail?id=579494ab-b903-4d51-8bb8-0db230dd80a4&lang=en>

110. Life Technologies. Ion Torrent Amplicon Sequencing. (2011). doi:10.7171/jbt.12

111. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30,** 434–526 (2012).

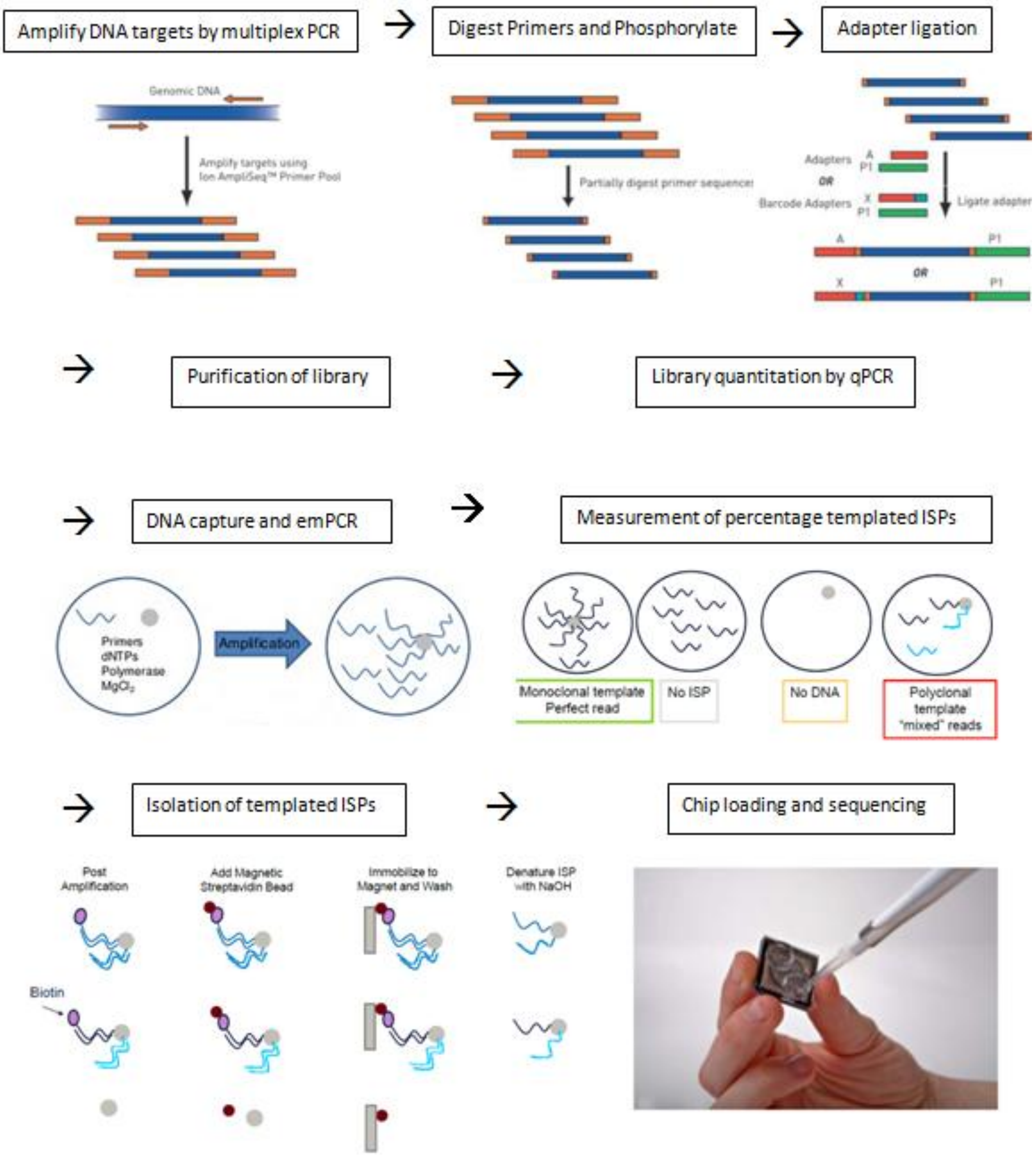
112. Yeo, Z., Wong, J. C., Rozen, S. G. & Lee, A. S. Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes. *BMC Genomics* **15,** 516 (2014).

113. Tack, D. Ion Semiconductor Sequencing. (2011). at <https://en.wikipedia.org/wiki/Ion_semiconductor_sequencing>

# Appendix

## Appendix A: Flowchart of the sequencing procedure

The main steps of the sequencing procedure are listed chronologically, from the amplification of DNA targets and library construction to chip loading and sequencing.



Amplify DNA targets by multiplex PCR
Genomic DNA
Amplify targets using Ion AmpliSeq™ Primer Pool
Digest Primers and Phosphorylate
Partially digest primer sequences
Adapter ligation
Adapters
A
P1
OR
Barcode Adapters
X
P1
Ligate adapter
A
P1
OR
X
P1
Purification of library
Library quantitation by qPCR
DNA capture and emPCR
Primers
dNTPs
Polymerase
$MgCl_2$
Amplification
Measurement of percentage templated ISPs
Monoclonal template Perfect read
No ISP
No DNA
Polyclonal template "mixed" reads
Isolation of templated ISPs
Post Amplification
Add Magnetic Streptavidin Bead
Immobilize to Magnet and Wash
Denature ISP with NaOH
Biotin
Chip loading and sequencing

Pictures reprinted with permission from Life Technologies.

## Appendix B: Protocol for DNA extraction

Prior to this study, tissue from the tumors was processed and DNA was extracted by the use of one of the following procedures:

***Procedure 1***

The QIAsymphony SP robot from Qiagen was used for DNA extractions and the QS DNA Tissue LC 200 protocol was followed. A fraction of approximately 15 mg from each sample was transferred to a 2 ml microcentrifuge tube. 20 µL of Proteinase K and 180 µL of ATL buffer were added, and the tube was placed in a thermomixer. The sample was incubated at 56 degrees with shaking at 900 rpm until the tissue was completely lysed (~ 3 hours). To generate RNA-free DNA, 4 µl RNase A were added followed by incubation at room temperature for 2 minutes. The sample was homogenized by pipetting up and down, and the supernatant was transferred to sample tubes to be loaded into the QIAsymphony SP. The extraction kit used was the QIAsymphony DNA mini Kit cat#931236 from Qiagen.

***Procedure 2***

A 5 mm Stainless Steel Bead was washed in 1 mL Buffer RLT Plus (lysis buffer) and transferred to the tube with an aliquote of approximately 15 mg tissue on dry ice. Half the amount (174,6 µL) of lysis buffer mix, containing 342,1 µL Buffer RLT Plus and 7 µL Reducing agent DTT [2M], was added to the biopsy on wet ice. 0,9 µL Reagent DX was added before the biopsy was homogenized (30Hz, 2 x 4 min) using a precooled (4˚C) TissueLyzer LT adapter on TissueLyser LT. The remaining lysis buffer mix (174,6 µL) were added, and the lysate was further homogenized by centrifugation (13200 rpm, 3 min) trough a QIAshredder column at room temperature. The lysate was transferred to a new tube, from which DNA and RNA >200 bp were extracted using the AllPrep DNA/RNA Mini Kit automated with use of the QIAcube robot. The extraction was performed following a custom protocol: "RNA_AllprepDNARNA_AnimalCells_AllPrep350_ID2481" for 350 µL input. The protocol makes use of an AllPrep DNA spin column, and DNA elution with 40 µL buffer EB gives a DNA fraction of 37 µL in buffer EB.

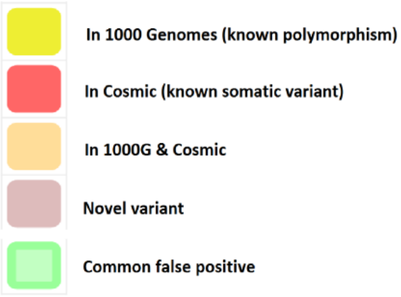
## Appendix C: The genes targeted by the Ion AmliSeq Cancer Hotspot Panel v2

- ABL Proto-Oncogene 1, non-receptor tyrosine kinase (*ABL1*)
- V-akt murine thymoma viral oncogene homolog 1 *(AKT1)*
- Anaplastic lymphoma tyrosinw kinse *(ALK)*
- Adenomatous polypsis coli (*APC)*
- Ataxia telangectasia mutated (*ATM* *)*
- B-Raf proto-oncogene, serine/threonine kinase *(BRAF)*
- Cadherin 1, type 1 (*CDH1)*
- Cyclin-dependent kinase inhibitor 2A (*CDKN2A)*
- Colony stimulating factor 1 receptor *(CSF1R)*
- Catenin beta 1 *(CTNNB1)*
- Epidermal growth factor receptor (*EGFR)*
- Erb-b2 receptor tyrosine kinase 2 *(ERBB2)*
- Erb-b2 receptor tyrosine kinase 4 *ERBB4*
- Enhancer of zeste 2 polycomb repressive complex 2 subunit (*EZH2)*
- F-box and WD repeat domain containing 7 (*FBXW7)*
- Fibroblast growth factor receptor 1 (*FGFR1)*
- Fibroblast growth factor receptor 2 (*FGFR2)*
- Fibroblast growth factor receptor 3 (*FGFR3)*
- Fms-related tyrosine kinase 3 (*FLT3)*
- Guanine nucleotide binding protein, alpha 11 (*GNA11)*
- GNAS complex locus (*GNAS)*
- Guanine nucleotide binding protein, q polypeptide (*GNAQ)*
- HNF1 homeobox A (*HNF1A)*
- Harvey rat sarcoma viral oncogene (*HRAS)*
- Isocitrate dehydrogenase 1 (*IDH1)*
- Janus kinase 2 (*JAK2)*
- Janus kinase 3 (*JAK3)*
- Isocitrate dehydrogenase 2 (*IDH2)*
- Kinase insert domain receptor *(KDR)*
- V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (*KIT)*
- Kirsten rat sarcoma viral oncogene homolog (K*RAS)*
- MET proto-oncogene receptor tyrosine kinase (*MET)*

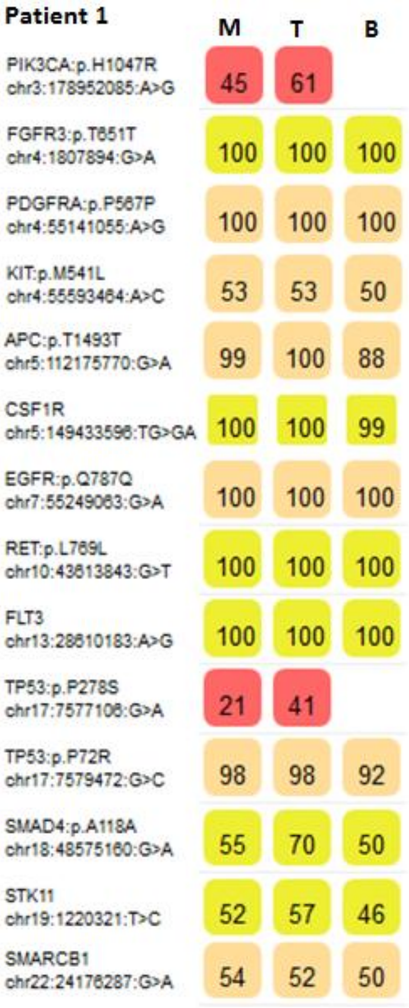
- MutL homolog 1 (*MLH1)*
- MPL proto-oncogene thrombopoietin receptor (*MPL)*
- Notch 1 (*NOTCH1)*
- Nucleophosmin (nucleolar phosphoprotein B23, numatrin) (*NPM1)*
- Neuroblastoma RAS viral (v-ras) oncogene homolog (*NRAS)*
- Platelet-derived growth factor receptor, alpha polypeptide (*PDGFRA)*
- Phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit alpha (*PIK3CA)*
- Phosphatase and tensin homolog (*PTEN)*
- Protein tyrosine phosphatase non-receptor type 11 (*PTPN11*
- Retinoblastoma 1 (*RB1)*
- Ret proto-oncogene (*RET)*
- SMAD family member 4 (*SMAD4)*
- SWI/SNF related, matrix associated, actin dependent regulator of subfamily b, member 1 (*SMARCB1)*
- Smoothened, frizzled class receptor *(SMO)*
- SRC proto-oncogene, non-receptor tyrosine kinase (*SRC)*
- Serine/threonine kinase 11 (*STK11)*
- Tumor protein p53 (*TP53)*
- Von Hippel Lindau tumor suppressor, E3 ubiquitin protein ligase (*VHL)*

## Appendix D: The complete collection of sequencing reports

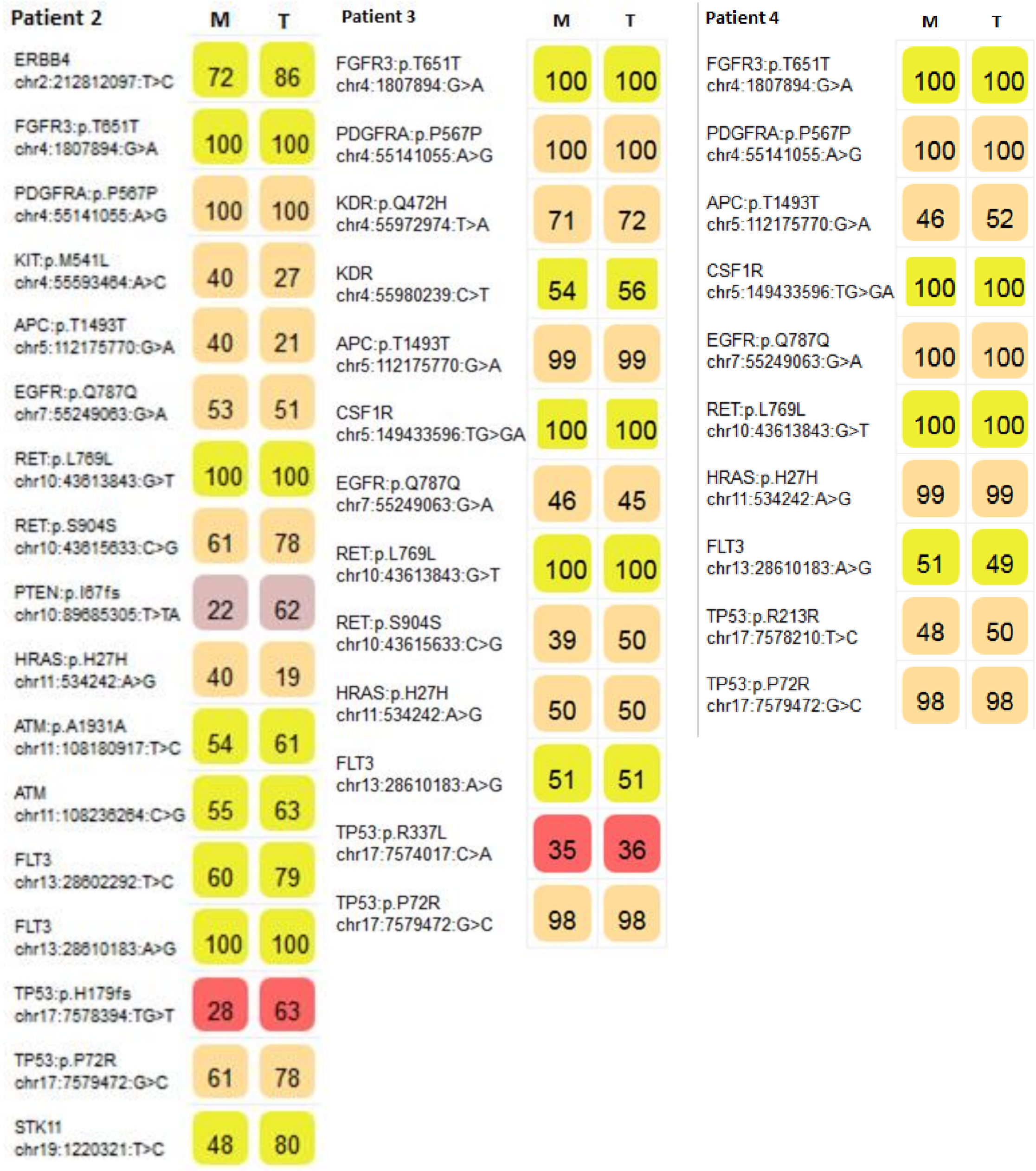
The sequencing reports from the 20 patients in the cohort are presented chronologically. The palette of colors on the left describes the color coding of the genomic variants, and the number inside the colored boxes indicate the percentage of reads with the variants.



| | |
|---|---|
| 🟨 | In 1000 Genomes (known polymorphism) |
| 🟥 | In Cosmic (known somatic variant) |
| 🟧 | In 1000G & Cosmic |
| 🟫 | Novel variant |
| 🟩 | Common false positive |



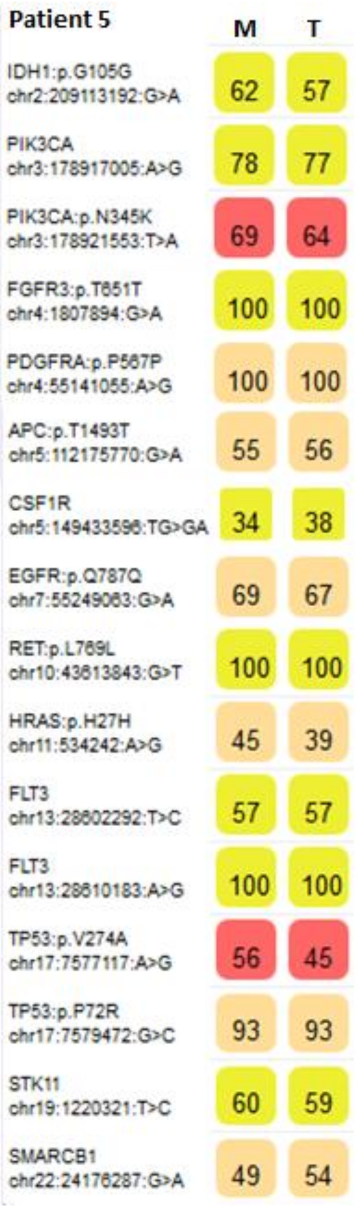
| Patient 1 | M | T | B |
|---|---|---|---|
| PIK3CA:p.H1047R<br>chr3:178952085:A>G | 🟥 45 | 🟥 61 | |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 🟨 100 | 🟨 100 | 🟨 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 🟧 100 | 🟧 100 | 🟧 100 |
| KIT:p.M541L<br>chr4:55593464:A>C | 🟧 53 | 🟧 53 | 🟧 50 |
| APC:p.T1493T<br>chr5:112175770:G>A | 🟧 99 | 🟧 100 | 🟧 88 |
| CSF1R<br>chr5:149433596:TG>GA | 🟨 100 | 🟨 100 | 🟨 99 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 🟧 100 | 🟧 100 | 🟧 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 🟨 100 | 🟨 100 | 🟨 100 |
| FLT3<br>chr13:28610183:A>G | 🟨 100 | 🟨 100 | 🟨 100 |
| TP53:p.P278S<br>chr17:7577106:G>A | 🟥 21 | 🟥 41 | |
| TP53:p.P72R<br>chr17:7579472:G>C | 🟧 98 | 🟧 98 | 🟧 92 |
| SMAD4:p.A118A<br>chr18:48575160:G>A | 🟨 55 | 🟨 70 | 🟨 50 |
| STK11<br>chr19:1220321:T>C | 🟨 52 | 🟨 57 | 🟨 46 |
| SMARCB1<br>chr22:24176287:G>A | 🟧 54 | 🟧 52 | 🟧 50 |



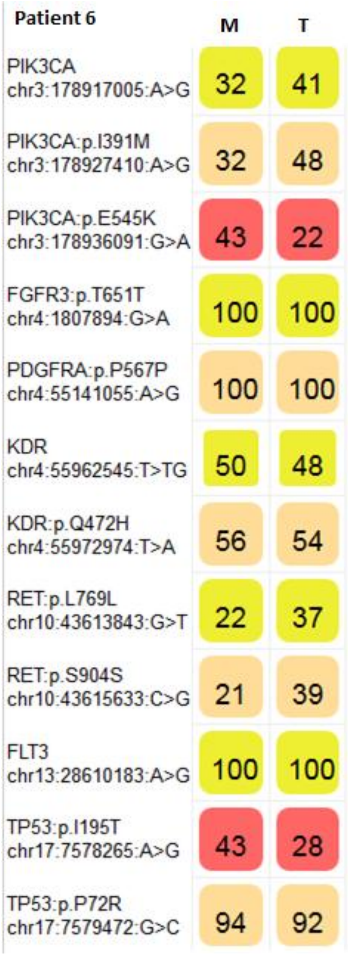
| Patient 2 | M | T |
|---|---|---|
| ERBB4<br>chr2:212812097:T>C | 72 | 86 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KIT:p.M541L<br>chr4:55593464:A>C | 40 | 27 |
| APC:p.T1493T<br>chr5:112175770:G>A | 40 | 21 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 53 | 51 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 61 | 78 |
| PTEN:p.I67fs<br>chr10:89685305:T>TA | 22 | 62 |
| HRAS:p.H27H<br>chr11:534242:A>G | 40 | 19 |
| ATM:p.A1931A<br>chr11:108180917:T>C | 54 | 61 |
| ATM<br>chr11:108236264:C>G | 55 | 63 |
| FLT3<br>chr13:28602292:T>C | 60 | 79 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.H179fs<br>chr17:7578394:TG>T | 28 | 63 |
| TP53:p.P72R<br>chr17:7579472:G>C | 61 | 78 |
| STK11<br>chr19:1220321:T>C | 48 | 80 |

| Patient 3 | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 71 | 72 |
| KDR<br>chr4:55980239:C>T | 54 | 56 |
| APC:p.T1493T<br>chr5:112175770:G>A | 99 | 99 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 46 | 45 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 39 | 50 |
| HRAS:p.H27H<br>chr11:534242:A>G | 50 | 50 |
| FLT3<br>chr13:28610183:A>G | 51 | 51 |
| TP53:p.R337L<br>chr17:7574017:C>A | 35 | 36 |
| TP53:p.P72R<br>chr17:7579472:G>C | 98 | 98 |

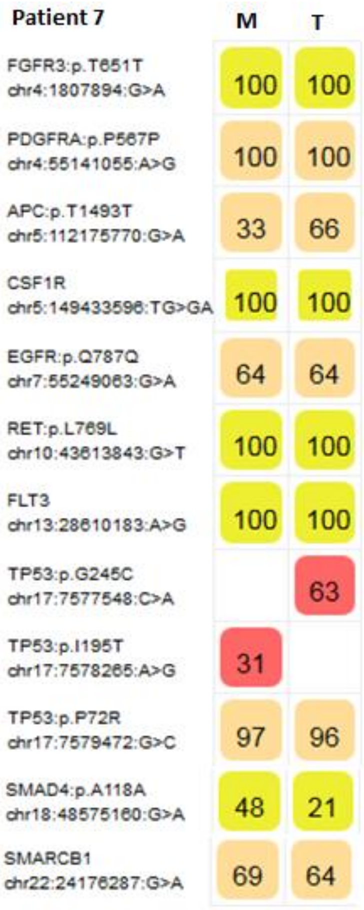
| Patient 4 | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 46 | 52 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| HRAS:p.H27H<br>chr11:534242:A>G | 99 | 99 |
| FLT3<br>chr13:28610183:A>G | 51 | 49 |
| TP53:p.R213R<br>chr17:7578210:T>C | 48 | 50 |
| TP53:p.P72R<br>chr17:7579472:G>C | 98 | 98 |



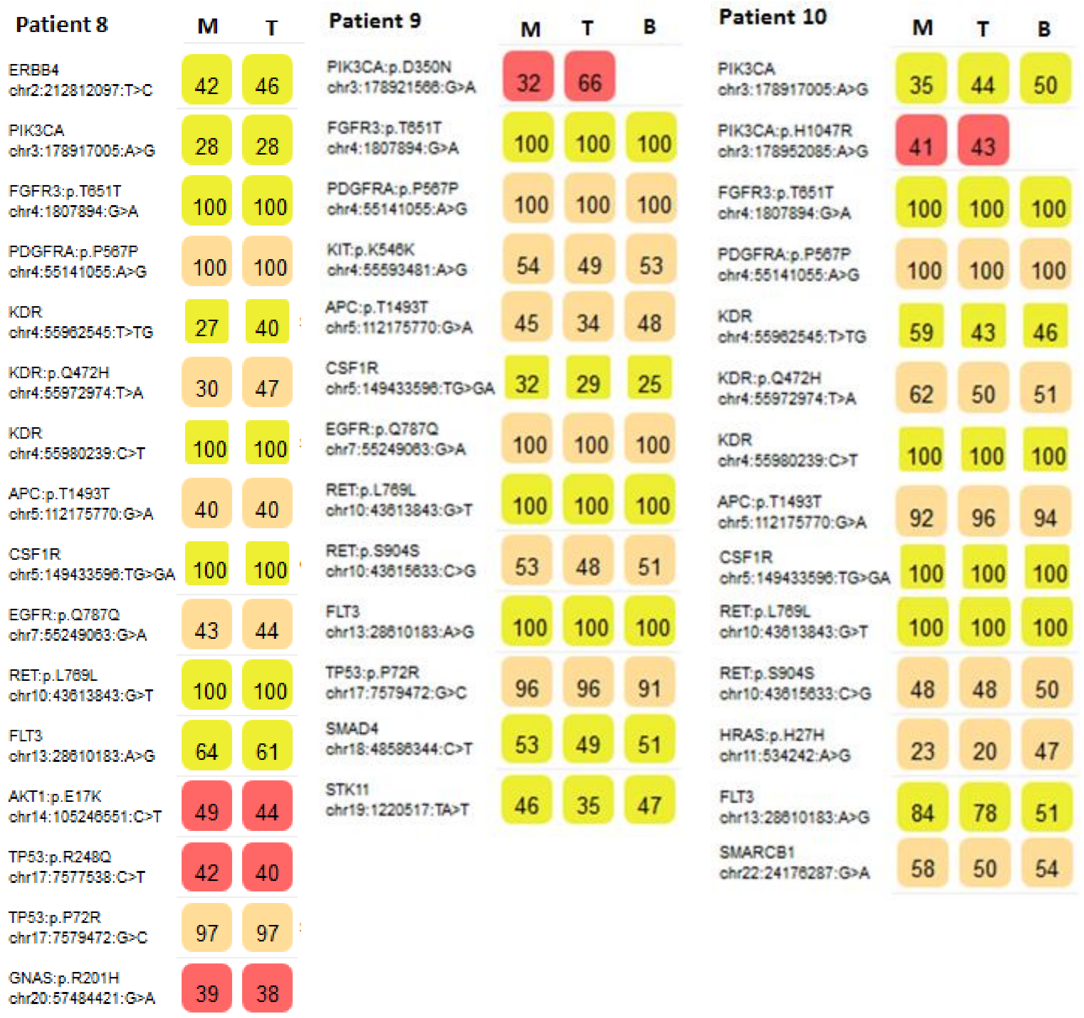
| Patient 5 | M | T |
|---|---|---|
| IDH1:p.G105G<br>chr2:209113192:G>A | 62 | 57 |
| PIK3CA<br>chr3:178917005:A>G | 78 | 77 |
| PIK3CA:p.N345K<br>chr3:178921553:T>A | 69 | 64 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 55 | 56 |
| CSF1R<br>chr5:149433596:TG>GA | 34 | 38 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 69 | 67 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| HRAS:p.H27H<br>chr11:534242:A>G | 45 | 39 |
| FLT3<br>chr13:28602292:T>C | 57 | 57 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.V274A<br>chr17:7577117:A>G | 56 | 45 |
| TP53:p.P72R<br>chr17:7579472:G>C | 93 | 93 |
| STK11<br>chr19:1220321:T>C | 60 | 59 |
| SMARCB1<br>chr22:24176287:G>A | 49 | 54 |



| Patient 6 | M | T |
|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 32 | 41 |
| PIK3CA:p.I391M<br>chr3:178927410:A>G | 32 | 48 |
| PIK3CA:p.E545K<br>chr3:178936091:G>A | 43 | 22 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR<br>chr4:55962545:T>TG | 50 | 48 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 56 | 54 |
| RET:p.L769L<br>chr10:43613843:G>T | 22 | 37 |
| RET:p.S904S<br>chr10:43615633:C>G | 21 | 39 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.I195T<br>chr17:7578265:A>G | 43 | 28 |
| TP53:p.P72R<br>chr17:7579472:G>C | 94 | 92 |



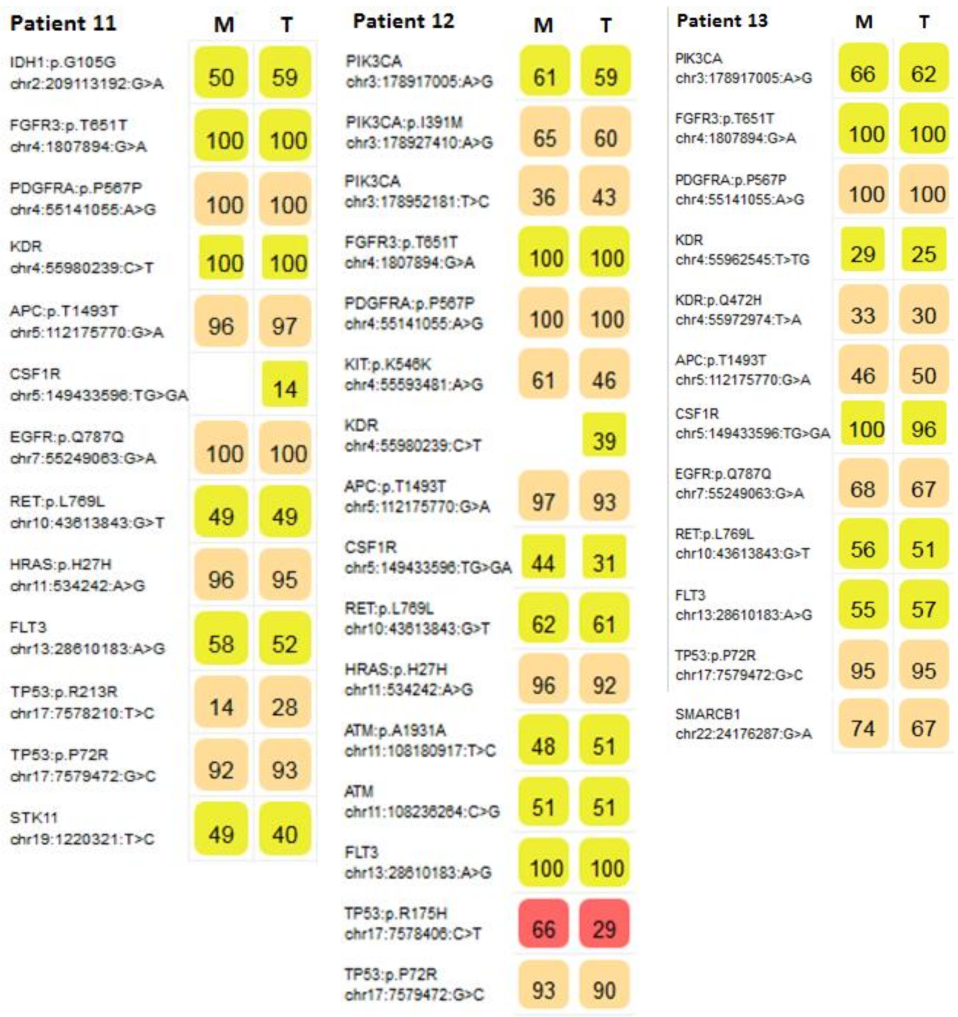
| Patient 7 | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 33 | 66 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 64 | 64 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.G245C<br>chr17:7577548:C>A |  | 63 |
| TP53:p.I195T<br>chr17:7578265:A>G | 31 |  |
| TP53:p.P72R<br>chr17:7579472:G>C | 97 | 96 |
| SMAD4:p.A118A<br>chr18:48575160:G>A | 48 | 21 |
| SMARCB1<br>chr22:24176287:G>A | 69 | 64 |



| Patient 8 | M | T |
|---|---|---|
| ERBB4<br>chr2:212812097:T>C | 42 | 46 |
| PIK3CA<br>chr3:178917005:A>G | 28 | 28 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR<br>chr4:55962545:T>TG | 27 | 40 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 30 | 47 |
| KDR<br>chr4:55980239:C>T | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 40 | 40 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 43 | 44 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| FLT3<br>chr13:28610183:A>G | 64 | 61 |
| AKT1:p.E17K<br>chr14:105246551:C>T | 49 | 44 |
| TP53:p.R248Q<br>chr17:7577538:C>T | 42 | 40 |
| TP53:p.P72R<br>chr17:7579472:G>C | 97 | 97 |
| GNAS:p.R201H<br>chr20:57484421:G>A | 39 | 38 |

| Patient 9 | M | T | B |
|---|---|---|---|
| PIK3CA:p.D350N<br>chr3:178921566:G>A | 32 | 66 | |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| KIT:p.K546K<br>chr4:55593481:A>G | 54 | 49 | 53 |
| APC:p.T1493T<br>chr5:112175770:G>A | 45 | 34 | 48 |
| CSF1R<br>chr5:149433596:TG>GA | 32 | 29 | 25 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 53 | 48 | 51 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 | 100 |
| TP53:p.P72R<br>chr17:7579472:G>C | 96 | 96 | 91 |
| SMAD4<br>chr18:48586344:C>T | 53 | 49 | 51 |
| STK11<br>chr19:1220517:TA>T | 46 | 35 | 47 |

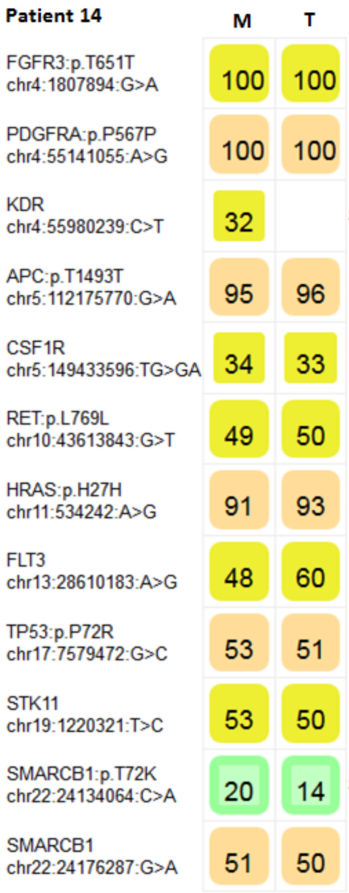
| Patient 10 | M | T | B |
|---|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 35 | 44 | 50 |
| PIK3CA:p.H1047R<br>chr3:178952085:A>G | 41 | 43 | |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| KDR<br>chr4:55962545:T>TG | 59 | 43 | 46 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 62 | 50 | 51 |
| KDR<br>chr4:55980239:C>T | 100 | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 92 | 96 | 94 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 48 | 48 | 50 |
| HRAS:p.H27H<br>chr11:534242:A>G | 23 | 20 | 47 |
| FLT3<br>chr13:28610183:A>G | 84 | 78 | 51 |
| SMARCB1<br>chr22:24176287:G>A | 58 | 50 | 54 |



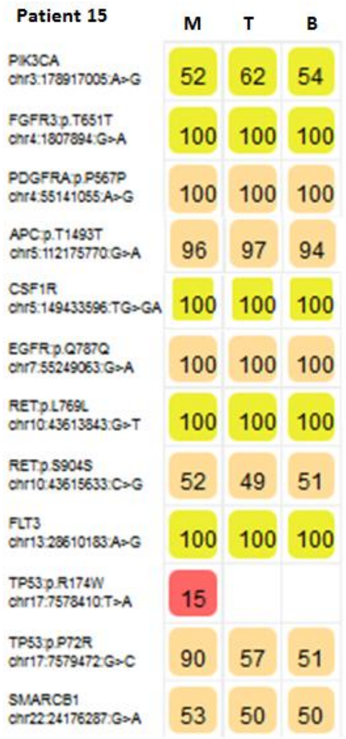
| Patient 11 | M | T |
|---|---|---|
| IDH1:p.G105G<br>chr2:209113192:G>A | 50 | 59 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR<br>chr4:55980239:C>T | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 96 | 97 |
| CSF1R<br>chr5:149433596:TG>GA |  | 14 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 49 | 49 |
| HRAS:p.H27H<br>chr11:534242:A>G | 96 | 95 |
| FLT3<br>chr13:28610183:A>G | 58 | 52 |
| TP53:p.R213R<br>chr17:7578210:T>C | 14 | 28 |
| TP53:p.P72R<br>chr17:7579472:G>C | 92 | 93 |
| STK11<br>chr19:1220321:T>C | 49 | 40 |

| Patient 12 | M | T |
|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 61 | 59 |
| PIK3CA:p.I391M<br>chr3:178927410:A>G | 65 | 60 |
| PIK3CA<br>chr3:178952181:T>C | 36 | 43 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KIT:p.K546K<br>chr4:55593481:A>G | 61 | 46 |
| KDR<br>chr4:55980239:C>T |  | 39 |
| APC:p.T1493T<br>chr5:112175770:G>A | 97 | 93 |
| CSF1R<br>chr5:149433596:TG>GA | 44 | 31 |
| RET:p.L769L<br>chr10:43613843:G>T | 62 | 61 |
| HRAS:p.H27H<br>chr11:534242:A>G | 96 | 92 |
| ATM:p.A1931A<br>chr11:108180917:T>C | 48 | 51 |
| ATM<br>chr11:108236264:C>G | 51 | 51 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.R175H<br>chr17:7578406:C>T | 66 | 29 |
| TP53:p.P72R<br>chr17:7579472:G>C | 93 | 90 |

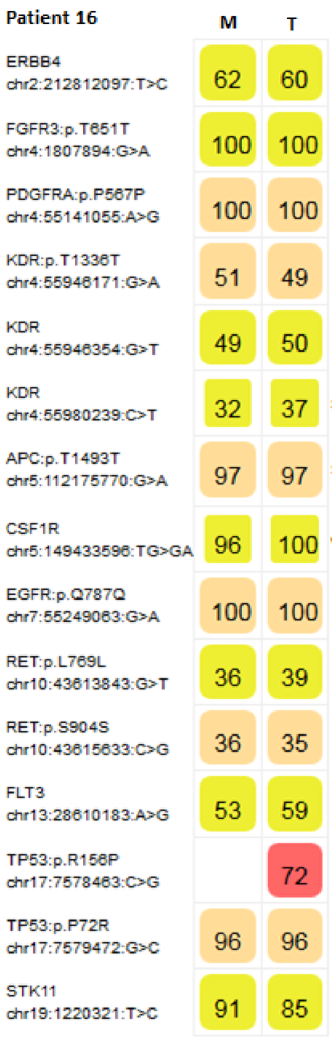
| Patient 13 | M | T |
|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 66 | 62 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR<br>chr4:55962545:T>TG | 29 | 25 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 33 | 30 |
| APC:p.T1493T<br>chr5:112175770:G>A | 46 | 50 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 96 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 68 | 67 |
| RET:p.L769L<br>chr10:43613843:G>T | 56 | 51 |
| FLT3<br>chr13:28610183:A>G | 55 | 57 |
| TP53:p.P72R<br>chr17:7579472:G>C | 95 | 95 |
| SMARCB1<br>chr22:24176287:G>A | 74 | 67 |



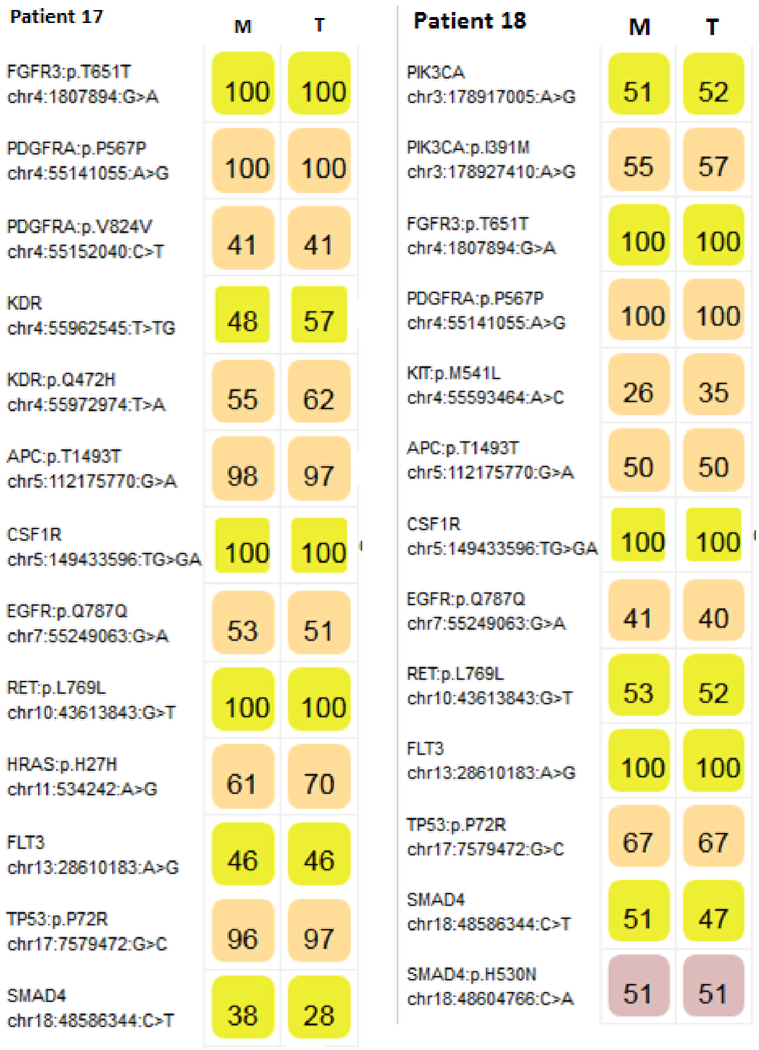
| Patient 14 | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR<br>chr4:55980239:C>T | 32 | |
| APC:p.T1493T<br>chr5:112175770:G>A | 95 | 96 |
| CSF1R<br>chr5:149433596:TG>GA | 34 | 33 |
| RET:p.L769L<br>chr10:43613843:G>T | 49 | 50 |
| HRAS:p.H27H<br>chr11:534242:A>G | 91 | 93 |
| FLT3<br>chr13:28610183:A>G | 48 | 60 |
| TP53:p.P72R<br>chr17:7579472:G>C | 53 | 51 |
| STK11<br>chr19:1220321:T>C | 53 | 50 |
| SMARCB1:p.T72K<br>chr22:24134064:C>A | 20 | 14 |
| SMARCB1<br>chr22:24176287:G>A | 51 | 50 |



| Patient 15 | M | T | B |
|---|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 52 | 62 | 54 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 | 100 |
| APC:p.T1493T<br>chr5:112175770:G>A | 96 | 97 | 94 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 52 | 49 | 51 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 | 100 |
| TP53:p.R174W<br>chr17:7578410:T>A | 15 | | |
| TP53:p.P72R<br>chr17:7579472:G>C | 90 | 57 | 51 |
| SMARCB1<br>chr22:24176287:G>A | 53 | 50 | 50 |

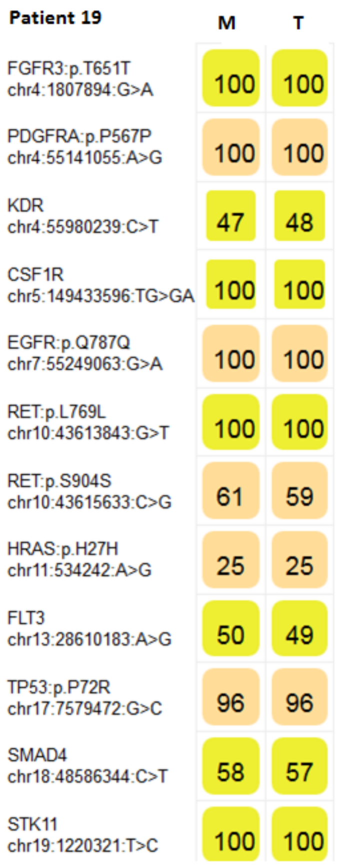


| Patient 16 | M | T |
|---|---|---|
| ERBB4<br>chr2:212812097:T>C | 62 | 60 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR:p.T1336T<br>chr4:55946171:G>A | 51 | 49 |
| KDR<br>chr4:55946354:G>T | 49 | 50 |
| KDR<br>chr4:55980239:C>T | 32 | 37 |
| APC:p.T1493T<br>chr5:112175770:G>A | 97 | 97 |
| CSF1R<br>chr5:149433596:TG>GA | 96 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 36 | 39 |
| RET:p.S904S<br>chr10:43615633:C>G | 36 | 35 |
| FLT3<br>chr13:28610183:A>G | 53 | 59 |
| TP53:p.R156P<br>chr17:7578463:C>G | | 72 |
| TP53:p.P72R<br>chr17:7579472:G>C | 96 | 96 |
| STK11<br>chr19:1220321:T>C | 91 | 85 |

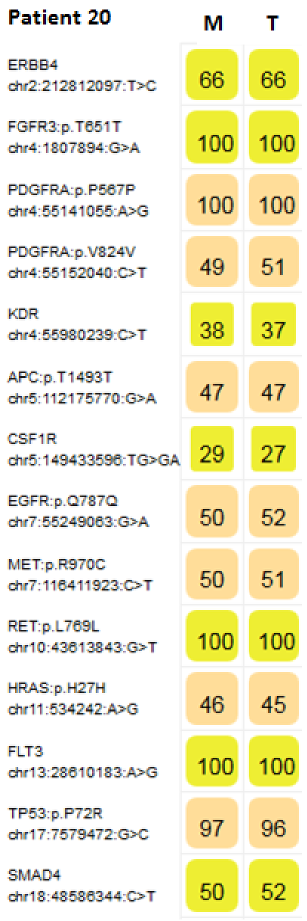


| Patient 17 | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| PDGFRA:p.V824V<br>chr4:55152040:C>T | 41 | 41 |
| KDR<br>chr4:55962545:T>TG | 48 | 57 |
| KDR:p.Q472H<br>chr4:55972974:T>A | 55 | 62 |
| APC:p.T1493T<br>chr5:112175770:G>A | 98 | 97 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 53 | 51 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| HRAS:p.H27H<br>chr11:534242:A>G | 61 | 70 |
| FLT3<br>chr13:28610183:A>G | 46 | 46 |
| TP53:p.P72R<br>chr17:7579472:G>C | 96 | 97 |
| SMAD4<br>chr18:48586344:C>T | 38 | 28 |

| Patient 18 | M | T |
|---|---|---|
| PIK3CA<br>chr3:178917005:A>G | 51 | 52 |
| PIK3CA:p.I391M<br>chr3:178927410:A>G | 55 | 57 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KIT:p.M541L<br>chr4:55593464:A>C | 26 | 35 |
| APC:p.T1493T<br>chr5:112175770:G>A | 50 | 50 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 41 | 40 |
| RET:p.L769L<br>chr10:43613843:G>T | 53 | 52 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.P72R<br>chr17:7579472:G>C | 67 | 67 |
| SMAD4<br>chr18:48586344:C>T | 51 | 47 |
| SMAD4:p.H530N<br>chr18:48604766:C>A | 51 | 51 |



| Patient 19 | M | T |
|---|---|---|
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| KDR<br>chr4:55980239:C>T | 47 | 48 |
| CSF1R<br>chr5:149433596:TG>GA | 100 | 100 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 100 | 100 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| RET:p.S904S<br>chr10:43615633:C>G | 61 | 59 |
| HRAS:p.H27H<br>chr11:534242:A>G | 25 | 25 |
| FLT3<br>chr13:28610183:A>G | 50 | 49 |
| TP53:p.P72R<br>chr17:7579472:G>C | 96 | 96 |
| SMAD4<br>chr18:48586344:C>T | 58 | 57 |
| STK11<br>chr19:1220321:T>C | 100 | 100 |



| Patient 20 | M | T |
|---|---|---|
| ERBB4<br>chr2:212812097:T>C | 66 | 66 |
| FGFR3:p.T651T<br>chr4:1807894:G>A | 100 | 100 |
| PDGFRA:p.P567P<br>chr4:55141055:A>G | 100 | 100 |
| PDGFRA:p.V824V<br>chr4:55152040:C>T | 49 | 51 |
| KDR<br>chr4:55980239:C>T | 38 | 37 |
| APC:p.T1493T<br>chr5:112175770:G>A | 47 | 47 |
| CSF1R<br>chr5:149433596:TG>GA | 29 | 27 |
| EGFR:p.Q787Q<br>chr7:55249063:G>A | 50 | 52 |
| MET:p.R970C<br>chr7:116411923:C>T | 50 | 51 |
| RET:p.L769L<br>chr10:43613843:G>T | 100 | 100 |
| HRAS:p.H27H<br>chr11:534242:A>G | 46 | 45 |
| FLT3<br>chr13:28610183:A>G | 100 | 100 |
| TP53:p.P72R<br>chr17:7579472:G>C | 97 | 96 |
| SMAD4<br>chr18:48586344:C>T | 50 | 52 |