Studies on sequencing analyses of genetic and epigenetics features in melanoma and breast cancer

Deepak B Poduval

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN

Studies on sequencing analyses of genetic and epigenetics features in melanoma and breast cancer

Deepak B Poduval

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 15.12.2020

Year: 2020

Title: Studies on sequencing analyses of genetic and epigenetics features in melanoma and breast cancer

Name: Deepak B Poduval

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

The work presented in this thesis was carried out at the Department of Clinical Science, Faculty of Medicine, University of Bergen and in the Mohn Cancer Research Laboratory, Haukeland University Hospital, and was funded by a PhD grant from Norwegian Cancer Society. The work was carried out under the supervision of Professor Stian Knappskog and Professor Per Eystein Lønning.

Acknowledgements

First and foremost, I would like to express my sincere gratitude and great appreciation to Prof. Stian Knappskog, you have been a remarkable mentor for me. I can never thank you enough for your valuable and continuous support throughout the work for this thesis. Without your immense patience, valuable suggestions, motivation and encouragement during the course of my PhD studies, it would not have been possible. I am very lucky to have a friendly and knowledgeable mentor. I would like to offer my special thanks to Prof. Per Eystein Lønning, for your insightful comments and encouragement, which helped me to widen my research. I am grateful to both of you, for giving me this splendid opportunity.

I would like to thank the previous candidates in the team who initiated the studies on which my thesis is built. Thus, I thank Christian Busch (Paper I), Elisabet Ognedal (Paper II) Svein Inge Helle (Papers II and III) and Anne Hege Straume (Paper III) for their initial work.

I would like to extend my heartfelt appreciations to all wonderful colleagues from the Mohn Cancer laboratory, for all the fruitful discussion and contribution to my work and life, making it easier and fun, since the day I joined. Special thanks to Beryl Leirvaag, for your great assistance and guidance through wet lab experiments. I would like to extend my gratitude to Zuzana Sichmanova, for all your assistance and filling up missing pieces and information in completing projects. An immense gratitude to all co – authors, without your contributions to the projects, I would never be able to finish them. I would like to thank all my colleagues from Mohn Lab, for supporting and extending words of advice, in making each and every step of my life, easy and wonderful in Bergen.

I am indebted to my family for always believing in me and showering with never ending love and support for my adventure far away from home. My parents and my sister have been an invaluable support system throughout my life. Most of all, thanks to my wonderful friends for being my source of motivation and strength. I would also

like to thank my wife for being there through all my ups and downs. I would like to remember my uncle late. Mr. V. K. Madhu. For showing the way to do what you love and what you believe in, despite of all the obstacles on course. Last but not least, thank GOD, for giving me strength, will blessing to make this possible.

Deepak B. Poduval, June 2020

Abbreviations

AI	Aromatase Inhibitors
BCC	Basal Cell Carcinoma
BCS	Breast Conserving Surgery
BCT	Breast Conserving Therapy
BWA	Burrows Wheeler Aligner
CLL	Chronic Lymphocytic Leukemia
CMF	Cyclophosphamide, Methotrexate, and 5-Fluorouracil
CNA	Copy Number Analysis
CNV	Copy Number Variation
DLBCL	Diffuse Large B-Cell Lymphoma
DNMT	DNA Methyltransfareses
FFPE	Formalin-Fixed Paraffin-Embedded
GA	Genome Analyzer
HTS	High Throughput Sequencing
LOH	Loss of Heterozygosity
MPS	massively parallel sequencing
NGS	Next Generation Sequencing
PE	Paired End reads
RISC	RNA Induced Silencing Complex
RRBS	Reduced Representation Bisulfite Sequencing
SCC	Squamous Cell Carcinoma
SERM	Selective Estrogen-Receptor Modulator

SNV	Single Nucleotide Variant
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SV	Structural Variation
TCGA	The Cancer Genome Atlas
TRBP	Transactivation RNA Binding Protein
UTR	Untranslated Region
WES	Whole Exome Sequencing
WGBS	Whole Genome wise Bisulfite Sequencing
WGD	Whole Genome Duplication
WGS	Whole Genome Sequencing

List of publications

Paper I

Einar Birkeland, Shaojun Zhang, **Deepak Poduval**, Jurgen Geisler, Sigve Nakken, Daniel Vodak, Leonardo A. Meza-Zepeda, Eivind Hovig, Ola Myklebost, Stian Knappskog and Per E. Lønning. Patterns of genomic evolution in advanced melanoma. *Nature Communications*. 9. 2665. (2018).

Paper II

Deepak Poduval, Elisabet O. Berge, Zuzana Sichmanova, Eivind Valen, Per E. Lønning and Stian Knappskog. Assessment of tumor suppressor promoter methylation in healthy individuals. *Manuscript submitted*.

Paper III

Deepak Poduval, Zuzana Sichmanova, Anne Hege Straume, Per E. Lønning and Stian Knappskog. The novel microRNAs *hsa-miR-nov7* and *hsa-miR-nov3* are over-expressed in locally advanced breast cancer. PLOS ONE 15(4): e0225357. (2020).

Contents

1. Introduction

1.1. Cancer

With more than 100 different subtypes and accounting to 13% of all deaths around the globe (Ferlay et al., 2015), cancer is one of the most complicated and highly mortal diseases worldwide (Figure 1). Research in the vast field of cancer is comprehensive and in general aims to unravel mechanisms characterizing the different malignancies in such a way that it may help in prevention, identification, and treatment.

As per the latest world cancer statistics of 2012 (Ferlay J, 2013), there was 14.1 million new cancer cases, including 7.4 million men and 6.7 million women, around the world. The corresponding number is projected to be 24 million new cases in the coming 20 years (Ferlay et al., 2015). The most common cancer types for both sexes around the world are breast, prostate, lung, colorectal and stomach, and in case of Norway, common cancer types for both sexes, are prostate, breast, colorectal, lung, and melanoma of the skin. Among males, most frequently diagnosed and high mortality cancer is lung cancer. Prostate and colorectal cancer had higher incidence rates, while liver and stomach cancer had higher mortality rates followed by lung cancer in males. Among females, the most common cancer and leading cause of cancer death, is breast cancer. In females after breast cancer, colorectal and lung cancer had higher incidence rates while higher mortality was observed in lung cancer followed by colorectal cancer (Bray et al., 2013, Bray et al., 2018, Ferlay et al., 2018).

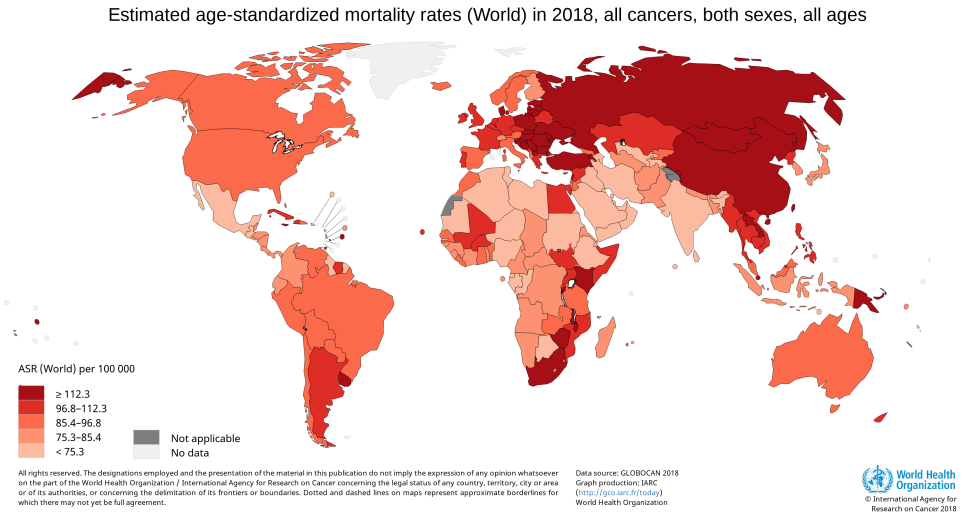


Figure 1. Age specific mortality rates from both sexes around the world. Created from http://gco.iarc.fr/today (Ferlay J, 2018).

While the underlying and evolving genetic disturbances for cancer may seem somewhat global, mostly varying between cancer types, the environmental factor influencing cancer development vary substantially between regions of the world. For example, in India the most prevalent cancer is cancers in the lip and oral cavity. This cancer ranks number one among men and third among women in terms of the incidence rates. This high incidence rates most likely accords to smoking and eating products related to tobacco as well as alcohol consumption (Byakodi et al., 2012). Lung cancer is the leading cause of cancer incidence and mortality for both genders combined. The highest incidence rates of lung cancer among males are found in Eastern Europe and the United States. For men, the most common occurring cancer is prostate cancer while in females it is breast cancer. Breast cancer is prevalent in highest rates in Western Europe and the United States and the lowest rates are found to be in Africa and Asia. Regarding melanoma, this is highly prevalent in Norway and Scandinavian countries as well as Australia, very likely due do fair skin and high sun-exposure (Torre et al., 2016, Bray et al., 2013, Bray et al., 2018, Ferlay et al., 2018).

Irregular growth of cells is generally known as neoplasm or tumor (which simply means mass). Tumors can broadly be divided into the two sub-groups benign and malignant. While benign tumor in general can be considered as a mass of cells that is localized (non-invasive) and malignant tumors (cancer), has a more invasive growth pattern and often the potential to spread (metastasize) in the body (Cooper, 1992). Cancers can be further classified based on the primary site of cancer and type of tissue in which the cancer (histological type) arises. The International Classification of Diseases for Oncology, Third Edition (ICD-O-3) classifies hundreds of different types of cancer into six major categories based on the histological type. They are carcinoma, sarcoma, myeloma, leukemia, lymphoma and mixed types. Carcinoma are class of cancer that arises in epithelial cells, internal or external lining of the body. Carcinoma are of two types – adenocarcinoma which occurs in a gland or an organ, and squamous cell carcinoma which arises in squamous epithelium. Sarcomas are neoplasms that originate in mesodermal cells (bones, muscles, tendons, cartilage, etc.). Myeloma instigates in plasma cells of bone marrow. Leukemias, commonly referred as blood cancers are cancer that arise in bone marrow, where blood cell production occurs and it is often observed with abnormal production of white blood cells. Lymphomas are cancer that arise in the nodes or glands in the lymphatic system. Lymphomas that are present in specific organs such as brain, stomach or breast are known as extranodal lymphomas. The other two categories of lymphomas are Hodgkin lymphoma and non-Hodgkin lymphoma, which are distinguished based on the presence of Reed-Sternberg cells. The last type of cancer are mixed as the name suggest cancer arises from different categories mentioned above (Fritz et al., 2000). As these malignant cells arise from different tissue types, the process of initiation of cancer (carcinogenesis) and the further advancement e.g. the spread from primary site (metastasis), differ in their molecular details as well (Pecorino, 2012).

In yester years, surgical removal of entire tumor along with lymph node were done to treat breast cancer (Halsted, 1894), until spreading of tumor through blood cells were reported (Paget, 1889). More recent advancement of technologies have made the surgery less complex, and through the last century, new methods were developed

combining surgery with chemotherapy and / or radiation therapy (Fidler, 2003). The success rates in treatment lead to further improvement in chemotherapy (Chabner and Roberts, 2005) and using drugs along with antibodies to target specific cancer cells (Parish, 2003) and chemo protective drugs to reduce the side effects (Summerhayes, 1995). Another treatment improved over time is hormonal therapy, which uses hormones directly or various strategies to control hormones. The latest major improvement in cancer treatment related to immune therapy by use of so-called checkpoint inhibitors or by reprogramming of immune cells (*CAR-T*) (Oldham and Dillman, 2008). While application of chemotherapy after surgery (adjuvant chemotherapy) is commonly used, downstaging of tumors by chemotherapy before surgery (neo-adjuvant chemotherapy) has also become a widely applied principle over the latest decades. (Notably, the latter strategy provides a good setting for research, since the growth or shrinkage of tumors may be directly assessed through the treatment (Sudhakar, 2009, DeVita and Chu, 2008, Lonning, 2003).)

All these enhancements of treatment came about by understanding tumorigenesis, growth of malignant tumor through multistage process of genotypic and phenotypic changes, which may occur over a period of time (Ashkenazi et al., 2008). To understand this compound and intricate disease, six hallmarks of cancer have been proposed as characteristic competencies that facilitate tumor development and metastatic propagation. These six hallmarks include "*Self-sufficiency in growth signals, Insensitivity to anti-growth signals, Evading apoptosis, Limitless replicative potential, Sustained angiogenesis and Tissue invasion & metastasis*" (Hanahan and Weinberg, 2000). As advancements in research moved forward, four more emerging hallmarks were later added on: "*Deregulating cellular energetics, Genome instability & mutation, Avoiding immune destruction and Tumor-promoting inflammations*" (Figure 2). Cancer treatments started improving and getting better by targeting and profiling molecular players of the hallmarks, and characteristics (Hanahan and Weinberg, 2011).

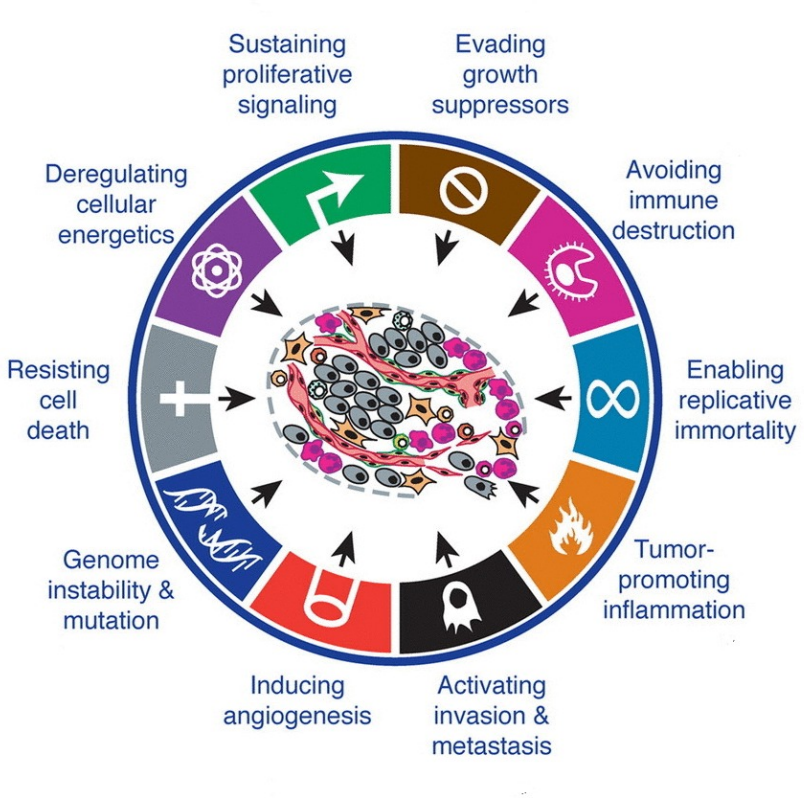


Figure 2: Hallmarks of cancer along with emerging and enabling characteristics proposed by Douglas Hanahan and Robert A. Weinberg. Modified. (Hanahan and Weinberg, 2011)

The work presented in this thesis is based on genetic and genomic analyses of samples from patients suffering from malignant melanoma and breast cancer. In the following sections, these two cancer forms are described in some more detail.

1.1.1 Melanoma

From previously being a rare kind of cancer, melanoma is now a cancer with high incidence all around the world (Ferlay et al., 2015, Erdmann et al., 2013, Tryggvadóttir et al., 2010, Azoury and Lange, 2014). The incidence of melanoma is

profoundly increased in developed countries and especially in people fairly skinned as compared to darker skinned (Tryggvadóttir et al., 2010, Erdmann et al., 2013). Increase in prevalence can be attributed to improved prognosis of disease along with added exposure to sun for recreation (Azoury and Lange, 2014). In the Nordic countries, the incidence were increased along with mortality, but improved treatments and prognosis has bettered survival rates among both men and women (Tryggvadóttir et al., 2010).

Melanoma is cancer arising from melanocytes, which produces melanosomes containing melanin in response to UV light from sun, giving skin different shades of color (Vijayasaradhi, 1995). Melanoma is likely to take two ways of developing into a malignant kind, it may originate from a normal nevus (mole) on the skin or it may appear on skin without a previous mole. The majority of melanomas occur de novo in sun exposed skin and it occasionally appears in eye and internal body mucosa (nasal, gastrointestinal, anorectal, etc.). However, there are also melanomas arising for other tissues: There rare cases of melanoma arising from the eye (ocular melanoma; about 5.2 % of total cases), cases from mucosal sites (mucosal melanoma; 1.3%) and around 2.2% cases from unknown sites (Chang et al., 1998). Notably, there are also other types of skin cancer defined as non-melanoma skin cancers. These are basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). Both of these cancers appear mostly at sun-exposed regions of skin and are less malignant than melanoma as they hardly metastasize (Xiang et al., 2014).

Melanoma of the skin can be divided in to subclasses - superficial spreading, lentigo and acral lentigo and then nodular – invasive form of melanoma (Clark et al., 1986). Superficial as name suggests grows along surface of skin. It is found as discolored patches or benign nevi along trunk, leg (mainly in women) and back (mainly in men) of the body. Lentigo melanoma is similar to superficial melanoma that starts as brownish patches and spreads into other parts of the body. It often develops in parts of skin that are often sun exposed (arms, face, ears etc.). Lentigo melanoma is often found in elderly population. Acral melanoma (acral lentiginous melanoma) starts as a black or brown at extremities of the body, like hand, feet, toes and finger, and it spreads faster than the superficial melanoma. This form is the

predominant form in darker skin populations, in Africa and Asia (Bradford et al., 2009) and it is unrelated to sun exposure. The most aggressive form of melanoma is nodular melanoma, which grows deeper into the skin and is often diagnosed as a bump on the skin at chest region or back (Clark et al., 1975, Chamberlain et al., 2003). Some other rare types of melanomas are mucosal lentiginous melanoma that occurs in mucosal membranes of internal organs (nasal passage, pharynx, mouth, vagina, anal canal and rectum) and it is not associated with sun exposure. Desmoplastic melanoma that occurs in thick inner layer of dermis or connective tissue surrounding mucosa.

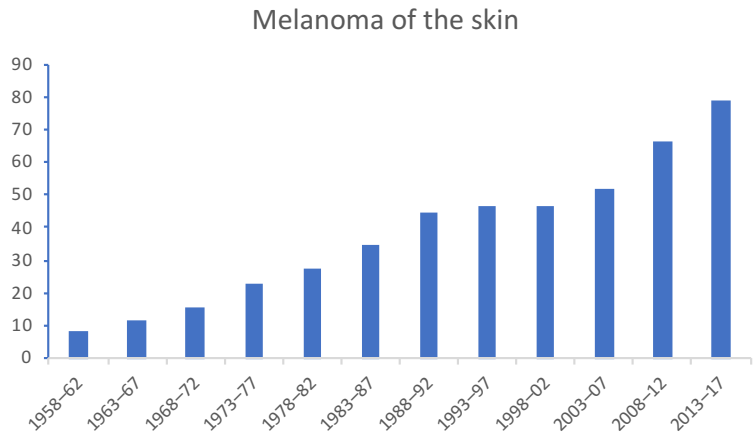


Figure 3: Age-standardised (Norwegian standard) incidence rates of melanoma per 100.000 person-years by primary site and five-year period, 1958–2017. (Source: https://www.kreftregisteret.no/en/The-Registries/data-and-statistics/the-statistics-bank/)

Risk factors involved with melanoma ranges from sun exposure to genetic disposition. The main risk factor of melanoma is exposure to UV-radiation from the sun as well as other sources (e.g., artificial tanning bed) and other underlying risk factors include history of sunburn, skin color, predisposition of nevi, family history and genetic factors.

The major environmental risk factor for developing melanoma is intermittent

exposure to sunlight, common source for UV radiation (Oliveria et al., 2006). People living at equator or higher altitudes are exposed to higher UV radiation and also people who use artificial sources of UV such as tanning light or tanning bed, that makes them highly susceptible to risk of melanoma (Boniol et al., 2012). People with history of sunburn as well as freckles are at increased risk of melanoma as it points to increased sun exposure. Fair skinned people are at increased risk of having melanoma as they have decreased levels of melanin, that helps in damage protection against UV. The risk of melanoma is also heavily dependent sensitivity to UV radiation but for people with darker skin such as Asians or Africans or Hispanics, the diagnosis of melanoma tends to be in late stage of melanoma and this is an obvious problem considering the survival of patients (Cormier et al., 2006).

Presence of nevi or multiple nevi are associated with increased risk of melanoma. Increased number of nevi, and large sized nevi are often correlated with higher risk of melanoma. A large fraction of non-familial melanoma are often diagnosed with atypical nevi or dysplastic nevi (Pampena et al., 2017). People with impaired immune system such as people who take immunosuppressants (e.g. after an organ transplant) or who have disease that immunodeficiency syndrome (e.g. AIDS) have shown greater risk of melanoma (Psaty et al., 2010). In history of melanoma disease, family history is one of the strong risk factors of melanoma. Risk of melanoma is correlated to whether first-degree relatives (parents, siblings, children) has melanoma and risk increase with number of first-degree relatives that have melanoma. In some of these families, it is observed that there are some common genetic abnormalities (Rastrelli et al., 2014).

In melanoma, the most common germline genetic abnormalities associated with high risk is mutations in the gene cyclin-dependent kinase inhibitor 2A (*CDKN2A*) encoding the *p16* protein, and in rarer cases, mutations in cyclin-dependent kinase 4 (*CDK4*). *p16* is an inhibitor of CDK4-activty, when mutated (inactivating mutations in *CDKN2A* and activating mutations in *CDK4*), this promotes cell cycle progression through increased phosphorylation of pRb (target for *CDK4* activity) and subsequent release of E2F1. For both genes, mutations act in a dominant manner, resulting in high penetrance of melanoma in affected families. However, mutations in these two

genes are only found in about 25% of families with apparent hereditary melanomas (Borg et al., 2000, Hussussian et al., 1994, Harland et al., 1997, Kamb et al., 1994). So there are most certainly other germline variants strongly associated with risk of the disease as well (Azoury and Lange, 2014, Hawryluk and Tsao, 2014, Rastrelli et al., 2014).

Regardless of underlying causes, the best way to prevent melanoma is, avoiding excess exposure to sun or UV rays. Then following “ABCDE”s signs of melanoma, which is A for finding asymmetric lesions or nevi, B for Borders of the lesions or nevi, which are irregular, or not well defined or being serrated, C for color, mainly ranging from black to brown to gray, D for diameter and E for those are evolving in any of the “ABCD” (Erdei and Torres, 2010).

Melanoma is staged using the TNM classification. TNM classification is based on the primary tumor (T), regional lymph node (N), and extent of metastasis (M). The classification is made by measuring dimensions of cancer mainly at T stage, tumor thickness and number of mitoses (Keohane et al., 2018). Further, melanomas are classified based on histology, into subtypes such as nodular, superficial, lentigo malignant and acral lentiginous melanoma (all cutaneous). Clark and Breslow micro staging is used to describe the spread of melanoma. Based on Breslow depth, T in TNM classification can be further classified into T1 - less than 1.0 mm, T2 - 1.01 to 2.0 mm, T3 – 2.01 to 4.0 mm and T4 - more than 4.0 mm. Survival rates of melanoma decreases with increase in Breslow depth (Breslow, 1970). Clark’s level describes spread of invasion and helps in prognostic distinction for tumor lesions. There are 5 levels of in Clark’s scale, Level 1 – melanoma is confined to outer layer of the skin (epidermis) and also called as “melanoma in-situ”, Level 2 – melanoma has crossed into outermost layer of the dermis (papillary dermis), Level 3 – melanoma has completely invaded papillary dermis and has touched the deeper layer dermis (reticular dermis), Level 4 – tumor has invaded the reticular dermis and level 5 – melanoma has reached into layer of fat under the skin (subcutaneous tissue) (Clark et al., 1969, Clark et al., 1975, Clark et al., 1986). Clark’s level has been no longer used in current staging systems as it is less prognostic and more subjective. In TNM classification, Breslow’s depth is used in measuring the thickness of tumor

(Cho and Chiang, 2010).

Different types of treatment are available for melanoma and it is decided based on several factors including melanoma characteristics and patient characteristics. Surgery is first and standard treatment. Surgery involves an operation that removes tumor and surrounding healthy tissue. Types of melanoma surgeries include wide excision (removal of primary melanoma), lymphatic mapping and sentinel mapping and lymph node biopsy, lymph node dissection. A second form of treatment available for treating melanoma is radiation therapy. It involves high energy radiation mainly x-rays or other form of rays to eliminate cancer cells or arrest their growth. Radiation therapy is generally carried out with an external radiation machine that sends high energy to target regions in the body. This form of therapy is mainly used when cancer is functionally inoperable. Radiation therapy may also be given after surgery to prevent cancer recurrence (adjuvant radiation therapy). Radiation therapy is also given to patients to alleviate symptoms and help in improved life quality, such radiation therapy is known as palliative radiation therapy. Chemotherapy, is also used. These are in general drugs to stop growth and spread of cancer, usually by stopping the cancer cells from growing or dividing. A chemotherapy regimen may include a single drug or drugs in combination, a set number of times over a specific time period. Some of the established drugs used in chemotherapy of melanoma are Dacarbazine (DTIC), Temozolomide (oral version of DTIC), cisplatin and taxanes (e.g., paclitaxel (taxol)). Chemotherapy with combination of the drugs tend to have more effectiveness along with increase in side effects. Side effects of chemotherapy depends on several factors such as dose of the drug used, health status of patients treated. Targeted therapy as the name suggests it uses drugs to target specific gene and proteins that enable cancer growth. This treatment causes less harm to normal cells as they are targeted to cancer cells. Two main types of targeted therapy are BRAF inhibitors (dabrafenib, vemurafenib, encorafenib), that blocks activity of proteins from mutated BRAF gene and second is MEK inhibitors (trametinib, cobimetinib, binimetinib), that blocks MEK1 and MEK2 proteins. Activity of these proteins are essential for the cancer cells to grow and survive. Combination of BRAF inhibitors and MEK inhibitors are also in use to treat melanoma(Perera et al., 2013,

Davis et al., 2019). Immunotherapy is treatment of melanoma using the patient's immune system. It is also known as biologic therapy as it uses medications that are made by body or in a laboratory to boost, target and restore body's natural immunity system to fight against cancer. A main form of immunotherapy is by using immune check point inhibitors, here the medication blocks immune check point protein, such as T cells that keeps are immune system response in check in turn increase immune system capability to killing cancer cells. Two main immune check point inhibitors are PD-1 inhibitors and CTLA-4 inhibitors. Some of the known PD-1 inhibitor are nivolumab and pembrolizumab, as the name suggest they block PD-1 protein on surface of T- cells and free T-cells to kill the cancer cells. A well known CTLA-4 inhibitor is Ipilimumab, it works by blocking CTLA-4, a protein on surface of T-cells and removes the check on T cells. Another immunotherapy is Interleukin-2 (IL-2), which boosts growth and activity of T-cells, thus enabling body to kill cancer cells. Interferon therapy is another form of immunotherapy that slows down tumor growth and arrest division of cancer cells.

In Norway, surgical removal of primary melanomas cure about >90% of the cases (Robsahm et al., 2018). For metastatic melanoma all modes of cancer therapy is used, including radiotherapy, chemotherapy and targeted therapies as well as immunotherapy (Erdei and Torres, 2010). Historically, dacarbazine (DTIC) chemotherapy remained sole treatment, with response rates of <20%. Improvements in therapy includes *BRAF* inhibitors and *MEK* inhibitors (Vennepureddy et al., 2016). A massive shift and improvement of treatment of advanced melanomas was brought about by the introduction of immunotherapy. Immunotherapy as a concept includes the use of vaccines and inflammatory cytokines helping in patient's immune system to recognize and eliminate cancer. However, the major breakthrough in terms of melanoma treatment over the recent years is related to the use of so-called checkpoint inhibitors, blocking negative signaling from cancer cells to T-cells. Successful treatment of melanoma based on this principle has in particular been related to use of ipilimumab which blocks the CTLA4-surface receptor and nivolumab blocking PD-1 (Seidel et al., 2018, Buchbinder and Desai, 2016).

1.1.2. Breast cancer

The second most common and fifth ranked in mortality amongst cancers both sexes combined, breast cancer is a most common cancer in women around the globe (Bray et al., 2013, Ferlay et al., 2015, Vennepureddy et al., 2016). Even though global incidence is on the rise, there is a slight decline in mortality in recent years, which can be accorded to better understanding and treatment of the disease as well as screening and early detection (Paap et al., 2014, Hofvind et al., 2013).

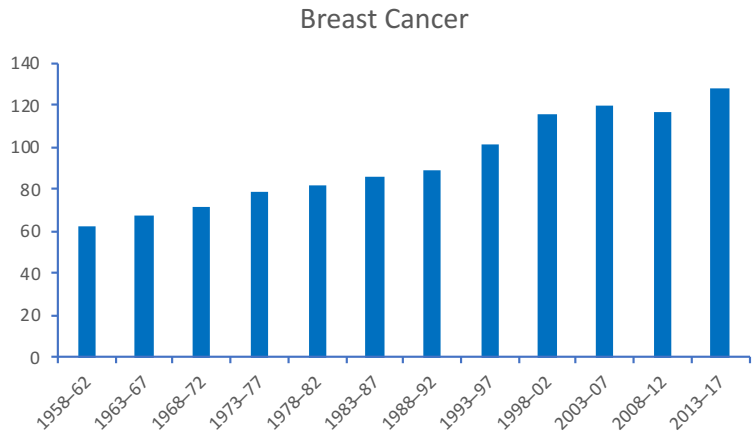


Figure 4: Age-standardised (Norwegian standard) incidence rates of breast cancer per 100.000 person-years by primary site and five-year period, 1958–2017. (Source: https://www.kreftregisteret.no/en/The-Registries/data-and-statistics/the-statistics-bank/)

Breast cancer is a heterogeneous disease with large differences in the biology of the tumors. While, traditionally, breast cancer can be stratified according to their expression of hormone receptors (estrogen- and progesterone receptor) as well as overexpression of Her2, they are in later years also often stratified according to molecular 'intrinsic' subtypes based on mRNA expression patterns (Perou et al., 2000). While several additional subtypes have been proposed, five major subtypes are commonly used for expression based subtyping: Luminal A, Luminal B, Her2 –

enriched, Basal-like, and Normal breast-like (Perou et al., 2000, Sørlie et al., 2001).

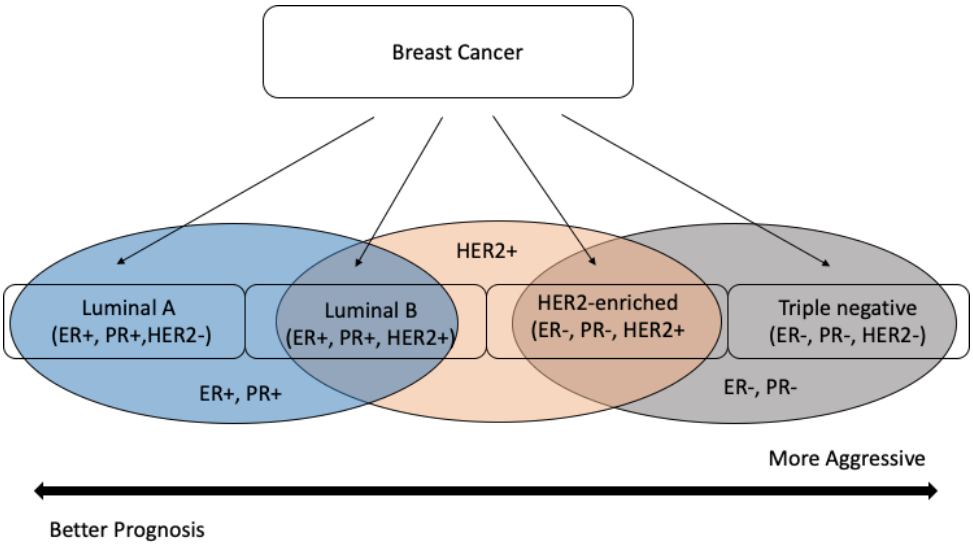


Figure 5: Schematic representation of breast cancer subtypes. Blue and brown oval group represents subtypes based on ER/PR expression. Orange oval subgroup represents Her2 amplification.

These subtypes differ in incidence (Millikan et al., 2008), survival (Cheang et al., 2009) and also to some extent in response to treatment (Nielsen et al., 2010, Hu et al., 2006, Ma et al., 2004, van 't Veer et al., 2002). Luminal A and Luminal B subtypes are in general ER+ (or PR+). and Luminal A is relatively slow growing cancers. *HER2* subtype usually has higher expression of the *HER2* protein and lack ER (ER-). This subtype has been linked to poor prognosis (Prat and Perou, 2011), a feature that has been reversed by the successes of anti-Her2 therapy. The basal-like category is largely overlapping with “triple negative” breast cancers (negative for *ER*, *PgR* and *Her2* overexpression) and in general has a poor prognosis (Sørlie et al., 2001, Sorlie et al., 1999). “Claudin-low” is a more recent subgroup of “triple negative” breast cancers, which has lower levels of claudin gene expression. Normal-like breast tumors group are poorly categorized and have been grouped with true normal as they show tumor cellularity as low as <50% in pathological examination. Their gene

characteristic lies between luminal-like and basal-like and do not respond well to neoadjuvant chemotherapy. They lack in expression of genes like *ER*, *PR*, *HER2* (similar to triple negative) as well as *CK5* and *EGFR* (non basal-like). (2012, Bastien et al., 2012, Prat et al., 2012, Perou, 2010, Perou et al., 2000, Sorlie et al., 2001, Prat and Perou, 2011, Nielsen et al., 2010). This subtyping of cancers has helped in improving the prognosis and treatment of the disease and have paved the way for development of prognostic expression profiles that may identify patients where chemotherapy is necessary versus those it may be omitted (Cardoso et al., 2016).

Breast cancer risk factors are diverse and mainly depend on the individual women, their age, personal history and family history, obesity, hormonal exposures and life style choices like drinking alcohol (Barnard et al., 2015, Dumalaon-Canaria et al., 2014). Aging is one of the most important breast cancer risk factors. Incidence of breast cancer increases with increase in age. Risk of breast cancer increase in women those have first degree or second-degree relatives with breast cancer. Factors associated with reproduction play a part in breast cancer risk. It is reported that women with late menopause as well as early age at menarche, low parity, and late age at first birth were associated with increased risk of breast cancer (Nagata et al., 1995). Breast cancer risk are increased in individuals with hormonal levels, including endogenous and exogenous estrogens as well as hormonal therapy. Lifestyle factors such as high dietary intake, sedentary lifestyle, smoking and alcohol consumption are associated with increase in risk of breast cancer (Iwasaki and Tsugane, 2011).

With respect to known genetic factors involved in breast cancer risk, these comprise about 30 genes that contribute only ~30% of familial risk (Collins and Politopoulos, 2011). The inherited genetic variants that influence breast cancer are a very variable with respect to penetrance. Some genes can harbor high penetrance variants. Among these are *BRCA1*, *BRCA2*, *PALB2*, *ATM* and *CHEK2* (Kamińska et al., 2015, Collins and Politopoulos, 2011, Michailidou et al., 2015). Additionally, variants in other genes only contribute to a low or moderate increase in risk and/or only in combination with other germline variants. In addition, amplification of genes such as HER2, EGFR, c-Myc, Ras are also documented with increased risk of breast cancer (Sun et al., 2017).

The treatment strategy for breast cancer most commonly includes surgical removal of the tumor followed by adjuvant chemotherapy, and/or adjuvant hormonal therapy (Rosenberg and Partridge, 2015, Perou, 2010). Surgical treatment can be performed with a breast conserving therapy (BCT) or by total mastectomy (Dutta et al., 2017). In early stage breast cancer patients, the widely used adjuvant treatment methods are endocrine treatment, chemotherapy and anti-HER2 treatment. Endocrine therapy has grown to be a preferred treatment option for the Luminal A and possibly for Luminal B subtype of breast cancer, and there are now several efforts to use this strategy only and to omit the use of chemotherapy from the treatment of hormone sensitive cancers. Adjuvant chemotherapy regimens for early and advanced stage breast cancers include active classes of cytotoxic agents such as anthracyclines like doxorubicin and epirubicin (Hortobagyi, 1997) and / or taxanes like paclitaxel and docetaxel (Rowinsky and Donehower, 1995). The first combination adjuvant therapy regimen was CMF (cyclophosphamide (C), methotrexate (M), and 5-fluorouracil(F)) followed by CAF (cyclophosphamide (C), doxorubicin (A), and 5-fluorouracil(F)) or CEF (cyclophosphamide (C), epirubicin (E), and 5-fluorouracil(F)). A common regimen of adjuvant therapy for breast cancer is AC-T (doxorubicin (A), cyclophosphamide and paclitaxel (T)) (Citron et al., 2003). In Norway, a common regimen has been “EC90” consisting of epirubicin and cyclophosphamide.

The most established adjuvant endocrine regimen is tamoxifen. This is a so-called selective estrogen-receptor modulators (SERMs), which acts by blocking the estrogen receptor for binding of the ligand (estrogen) and thereby blocks estrogen-related growth signaling to the cancer cells. Another well-established strategy relates to the use of Aromatase inhibitors (AIs). Tamoxifen started out as standard treatment for pre- and post-menopausal patients. While many pre-menopausal receive tamoxifen with a combination of LH-RH agonist. AIs are mainly used for postmenopausal patients and it lowers the concentration of serum estradiol by blocking the enzyme (aromatase) that converts androgens to estrogen. GnRH agonists, another SERMs, restrains ovarian function and it is used for premenopausal patients as it induces menopause like condition (Lumachi et al., 2011).

Anti-HER2 therapy is very effective but naturally restricted to the tumors that are

overexpressing the HER2 receptor (Peto et al., 2012, Sorlie et al., 2003, Davies et al., 2011). In HER2 positive tumors, humanized monoclonal antibody, trastuzumab used as anti-HER2 treatment (Anampa et al., 2015). Recent studies have also revealed that dual treatment (i.e. concomitant use of two different anti-bodies such as trastuzumab and pertuzumab) is superior to the use of trastuzumab alone (Swain et al., 2015, Loibl and Gianni, 2017).

Neoadjuvant therapy (medical treatment in the timespan before surgery) is often used in case of patients with larger tumors (so-called locally advanced breast cancers) and increasingly used for smaller tumors, in particular of the triple-negative and HER2+ types. In this group of tumors, where surgery followed by adjuvant chemotherapy proved to be a far-from-optimal treatment strategy leading to a high rate of relapses. Some decades ago, the strategy was changed to neoadjuvant chemotherapy, in order to reduce the tumor burden before surgery and thereby increasing the success rate of curative surgery and reduce the risk of relapse. Importantly, for the research in the field of chemoresistance, such treatment has proven to be an optimal study setting in as much as one can accurately monitor the effect of a given drug by measuring the growth/shrinkage of the tumor during treatment. This contrast the adjuvant setting where the effects of a drug can only be assessed by assessment of relapses, meaning that effects must be assessed over very long timespans. Further, a clinically effective method in the adjuvant and palliative setting, is radiotherapy. Radiotherapy is given to residual breast of patients following undergoing breast-conserving surgery (BCS). Irradiation after BCS reduces the chance of local recurrence, modestly improving survival rates of patients (Joshi et al., 2007).

Immunotherapy has been tested in breast cancer, but the results are so far somewhat disappointing (Schmid et al., 2018). The most promising effects of this treatments concept has been seen within the subtype of triple negative breast cancers, but also here the effects have been limited (Jia et al., 2017, Schmid et al., 2018). However there are hopes that, there may still be an important role for immunotherapy in breast cancer, pending identification of predictive biomarkers that can predict which of the patients will actually respond to the therapy, and of course pending development of

novel immuno-based therapies, other than the ones available today (Emens, 2018).

1.2 Molecular characteristics of cancer

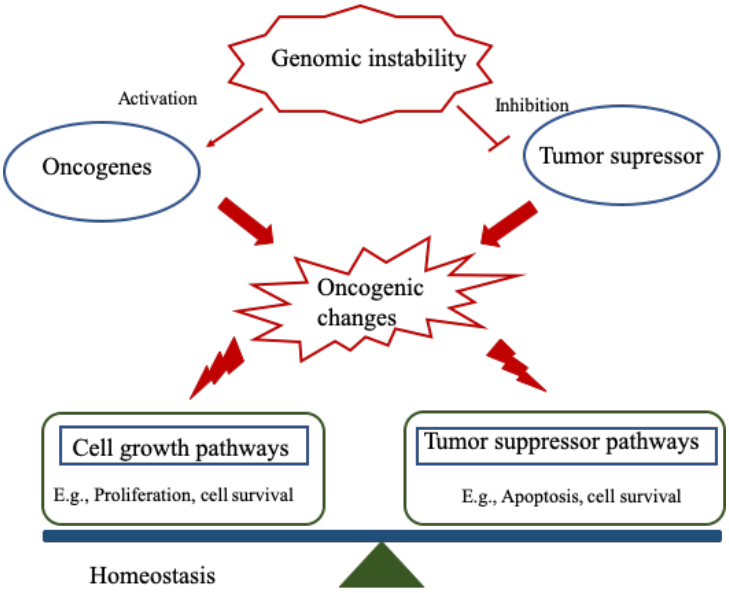


Figure 6: Overview of cellular homeostasis and factors affecting them.

Homeostasis refers to the network of complex interactions that help maintain and regulate internal condition of the organism under stable conditions (Figure 6). Any imbalance in regulation of homeostasis can in turn result in tumor formation and other diseases. In tumor formation is a complex process involving different factors, that is discussed in subsequent chapter.

1.2.1 Tumor suppressors and oncogenes

Distinct and stochastic forces drive whole mutational processes depending on each individual cancer type, stage and interaction with other important processes involved in tumorigenesis. In tumors, the uncontrolled growth of cells that hinders vital functions can be said, in general to be driven by mutations causing activation of

oncogenic pathways (e.g. mutations in *ras, src, abl*) and/or inactivation of tumor suppressing pathways (e.g. mutations in tumor suppressor genes such as *TP53*, *RB1*, *PTEN,* etc. (Weinberg, 1994, Lehman et al., 1991). Understanding specific elements of the cellular pathways in which the mutated oncogenes and tumor suppressor genes execute their function is important with respect to finding a way to improve treatment, reduce toxicity and overall survival of the patient. (Borges et al., 2005).

Oncogenes are genes where activating mutations in a proto-oncogene cause hyperactivity of the encoded protein, and where this hyperactivity gives the cancer cell a growth advantage or malignant phenotype. On the other hand, tumor suppressor genes work antagonistically and have functions keeping order in the cells' proliferation and repair. Rather than hyper-activating mutations, cancer related mutations in tumor suppressor genes are usually resulting reduced or abolished protein function. Regarding tumor suppressors, the first found was the retinoblastoma gene *RB1,* where germline defect are strongly linked to retinoblastoma in kids (Lee et al., 1987). Knudson proposed two-hit hypothesis in retinoblastoma, that postulates recessive nature of mutations in tumor-initiating gene and inheritance of familial cancer (Knudson, 1971). This hypothesis was confirmed the by demonstration of loss of heterozygosity at 13q14 in *RB1*, leading to no functional allele in cells with previous heterozygous mutation in the *RB1* gene. This laid foundation to concepts such as tumor suppressor genes and loss-of-heterozygosity (LOH) (Cavenee et al., 1985). In breast cancer, the *RB1* gene has been found inactivated in 30 % of cases through either mutations or deletions or allelic loss of the entire gene (loss of heterozygosity; LOH) (Burkhart and Sage, 2008).

In cancer, the p53 signaling pathway is vital for protection against stress and imbalance caused by oncogenic pressure. The tumor suppressor gene *TP53*, is known as the guardian of the genome (Lane, 1992) as it facilitates apoptosis or senescence and help in damage control and repair mechanisms making it an important component in maintaining genomic stability (Park et al., 2016). Germline mutations in *TP53* causes Li-Fraumeni syndrome and female carriers of this gene mutations carries a risk of 5% even before the age of 30 and approximately 50% of all cancers cases harbour somatic mutations. (Nichols et al., 2001, Gonzalez et al., 2009). In response

to DNA damage, the protein products of genes like *ATM*, *CHEK2*, *MDM2*, *MYC*, *RAS*, *CDKN2A* are some of the genes that play a major role in arresting growth or induce programmed cell death by activating p53 and counter oncogenic transformation and proliferation (Palmero et al., 1998, Lowe and Sherr, 2003, Zhang et al., 2011, Smith et al., 2010, Zindy et al., 1998).

Importantly, recent studies show that the same gene might act as an oncogene and as a tumor suppressor depending upon the cancer type and also depending on the status of the gene (whether the gene is wild-type or mutated may flip the role from oncogene to tumor suppressor and vice versa) (Borges et al., 2005, Zhang et al., 2010, Wang et al., 2009, Hutchinson et al., 2004, Lynch et al., 2008) (The Cancer Genome Atlas et al., 2012, de Jong et al., 2002). The *TP53* gene is generally considered to be an extremely important tumor suppressor. However, there are also more recent studies providing clear indications that some of the *TP53* mutants may be gain-of-functions mutants (Lozano, 2007). Thus, *TP53*, seemingly can be both a tumor suppressor and in some cases an oncogene. In breast cancer, mutations in *TP53* are a little less frequent than in many other cancer forms, with approximately 20% of breast cancers being affected. Other important tumor suppressor genes, strongly related to breast cancer are *BRCA1* and *BRCA2*. Somatic mutations of *BRCA1* is observed in more than 5% of breast cancer patients, while mutations of *BRCA2* are less frequent (Nik-Zainal et al., 2016). *BRCA1* and *BRCA2* are genes involved in double stranded break repair act as tumor suppressor genes. Germline mutation in these two genes impart more than 75% of lifetime breast cancer risk to female carriers (King et al., 2003). Similarly, genes like *PTEN* and *CHK2* mutation cause increased risk of breast cancer (Osborne et al., 2004, Lee and Muller, 2010). *PTEN* gene, is an important tumor suppressor gene, mainly known for Cowden syndrome (autosomal dominant disorder). Allelic loss of *PTEN* (Phospatase and Tensin) was present in ~20% of melanoma (Yin and Shen, 2008). *CHEK2* (Checkpoint kinase 2), vital player in DNA damage response and deletions in this gene causes increased risk of breast cancer as well as multiple risk associated with susceptibility to multiple other genes (2004). *PALB2*, Fanconi anemia assosciated gene (also called as *FANCN*) that act with interaction along with *BRCA2* gene (Rahman et al., 2007).

Oncogenes like *HER2*, *MYC*, *PI3KCA*, *BRAF*, etc., are found frequently deregulated in breast cancer. *HER2* oncogene activation is found in about 20 % of primary breast cancers (Guo et al., 2006). Similarly *MYC* gene is found to be overexpressed in around 15-20% of breast cancers (Steeg and Zhou, 1998). In breast cancer *PIK3CA* mutations are found in 20-30% patients. *BRAF* mutations especially, position V600 (Davies et al., 2002), has been reported in 10 % of all human cancers with oncogenic effects (Dhomen and Marais, 2007).

Table 1: Summary of some major somatic mutations and amino acid (AA) changes for oncogenes (in red) and tumor suppressor genes reported in COSMIC v89(Forbes et al., 2017).

Gene	Most Observed Substitution	Most Observed AA Mutation
AKT1 / AKT2 / AKT3	Missense	E17K
BRAF	Missense	V600E
BRCA1	Missense	P871L
BRCA2	Missense	NA
CDKN2A	Nonsense	R80*
CHEK2	Silent	A392A
HER2	Missense	S310Y
K-RAS	Missense	G21D
MYC	Missense	P59L
N-RAS	Missense	Q61R
PALB2	Nonsense	R753*
PIK3CA	Missense	NA
PTEN	Missense	NA
RB1	Nonsense	R251*
TERT	Silent	A305A
TP53	Missense	NA

Mutation in *BRAF* is prevalent in malignant melanoma around 30-70% and common mutation found is $BRAF^{V600E}$ which accounts to more than 90% of mutations (Pollock et al., 2003). In melanoma, isoforms of the RAS gene are N-RAS and K-RAS, with N-*RAS* being commonly mutated, while *K-RAS* happens to be a rarely mutated form (Jafari et al., 1995, Shukla et al., 1989). *AKT* has three isoforms *AKT1*, *AKT2* and

AKT3, out of which *AKT2* and *AKT3* are isoforms mainly found dysregulated in melanoma (Read et al., 2016, Wangari-Talbot and Chen, 2013, Stahl et al., 2004). Mutations in the tumor suppressor gene, *CDKN2A,* result in increased risk of familial melanoma ~20-60% (Goldstein et al., 2006). *TERT* (Telomerase RT), a reverse transcriptase subunit of telomerase complex, is important in regulation of telomere length. Somatic mutations in *TERT* promoters was observed in more than 29% melanoma (Vinagre et al., 2013).

In addition to mutations (either somatic or germline), epigenetic mechanisms, such as methylation and chromatin organisation, are important in modulation of genes that play important role in neoplasia. In particular, in cancer, hypermethylation of tumor suppressor genes, thereby inactivating them, is a key event. This is described in more detail in chapter 2.4.

1.2.2. Genome instability

A modern cancer hallmark, is genomic instability that arise from mutations and chromosomal rearrangements and drives tumorigenesis (Hanahan and Weinberg, 2011). Genomic instability occurs from somatic point mutations, copy number alterations and they also show structural variations in chromosomes (chromosomal instability) via change in number of chromosomes or structural changes as well as microsatellite instability brought about via increase in mutation burden and rate. Normal cells differ from tumor cells with acquirement of these genomic alterations giving rise to more aggressive tumor subclones resulting in tumor initiation and progression. Genomic instabilities provide malignant tumor cells to bypass cell cycle checkpoints and other important cell processes, making them a hallmark in better understanding of cancer (Nowell, 1976). Abnormal chromosomal structures and aberrant chromosome numbers are a main cause of genomic instability during mitosis. Another key source of genomic instability is presence of somatic copy number variations. A cancer cell can gain or lose a copy of chromosome during tumorigenesis and regain copies or again lose copies based on the cellular environment. The presence of alternations of copy number affects the integrity of

chromosomes causing disruptions in integrity of chromosomes leading to genomic instability (Andor et al., 2017).

DNA damage repair machinery of cells are vital in tackling genome instability and their shortcomings. DNA repair pathways are important to shun tumorigenesis by facilitating DNA repair or initiating apoptosis. It is maintained by DNA mismatch repair system which corrects mismatches along with correcting insertions and deletion on DNA. Mutations in DNA mismatch repair mechanism leads to increased mutation burden causing instability in microsatellites affecting the genomic integrity (Kunkel, 1995).

1.2.3. Tumor heterogeneity

Heterogeneity, both functional and phenotypical, is very important in context of tumor. Notably, there are several layers of heterogeneity that should be recognised: first, inter-individual heterogeneity represents the differences seen between tumor in different individuals. Secondly, inter-tumor heterogeneity can also represent differences across tumors in an individual patient (typically heterogeneity between different metastases). Finally, intra-tumor heterogeneity refers to the differences between different subclones (or even single cells) within a single tumor.

The levels of heterogeneity, is often a reason that make cancers a complicated disease to treat, gives rise to inaccurate diagnosis, different clinical responses as wells as outcome (Heppner, 1984, Illingworth et al., 2010). Rapid evolution in the field of genome-wide studies and high throughput sequencing brought about deeper insights into mechanisms involved in tumor heterogeneity. Tumor heterogeneity occurs in cellular and molecular levels that can be driven by clonal evolution caused by genomic instability as well as altered levels of genetic and epigenetic factors.

Heterogeneity can be found in non-heritable manner that arise form phenotypic plasticity and cancer stem cell differentiation and heritable manner of heterogeneity arises from clonal expansion in Darwinian tumor evolution. In tumors this plasticity affects the ability to form different forms tumor cells with influence of microenvironmental factors (Roeder and Loeffler, 2002). So, any difference that is

brought about from changes in genotypes and environmental conditions can give rise to heterogeneity in tumor cells (Park et al., 2000). Similarly stem cells in tumors possess ability similar to stem cells to replenish tumor cells as well as differentiate in to different tumor cells (Dick, 2008). Any changes in this differentiation of stem cells give rise to heterogeneity that is no-heritable in manner (Kern and Shibata, 2007). A large portion of tumor cell heterogeneity arises from these non-heritable manners of mechanisms. The other main way of tumor heterogeneity is clonal evolution, that is mostly accumulated from mutations in genes. Tumor progression is generally propagated by stochastic process in acquirement of mutational events in response to genomic instability and tumor cell proliferation (Parmigiani et al., 2009). Mutational events are selected in Darwinian way, that are advantageous to tumor progression and make way for clonal expansion. This process is influenced by changes in tumor microenvironment and selective pressures in tumor cells, thus providing tumor heterogeneity (Parmigiani et al., 2009). Clonal expansion can be linear as mutations drive linear succession of tumors and it can also be expansive as mixture of multiple co-existing and expanding linear clones. Clonal heterogeneity differ from tumor heterogeneity as former is caused by clonal expansion and latter is genetic differences in tumor cells (Marusyk and Polyak, 2010).

Genomic sequencing study of primary tumor and metastasis pairs has revealed sub clonal evolution in breast cancer. Variable degrees of genomic heterogeneity were found across breast cancers as well as subclonal heterogeneity across subgroups of breast cancer. It was also reported that subclonal mutations were found in only a small fraction of tumor cells but were variably distributed allele fractions. Clonal expansion occurs by accumulating mutations and happens until it reaches a mass, therefore represents a rate limiting step in development of breast cancer (Yates et al., 2015, Nik-Zainal et al., 2012).

1.2.4. Methylation and epimutations

Epigenetics is a broad field covering a multitude of different molecular mechanisms, including histone modifications etc. (Verma et al., 2014). However, the best studied

epigenetic mechanism is DNA methylation. Research has been expanding in the field of exploring different characteristics of DNA and methylation (Gayon, 2016). Methylation in genomic sites and patterns are known to play key roles in different stages of early development. In addition to turning on and off genes that are required to be active / inactive at different stages of development, methylation also plays a crucial role as a cancer protective mechanism through transcriptional silencing of regions harbouring transposons and viral elements (Reik, 2007, Weber and Schubeler, 2007, Cedar and Bergman, 2009, Schubeler, 2015).

In the eukaryotic cell nucleus, DNA is tightly packaged with the help of histones as chromatin, a highly ordered structure. Chromatin is a physiological center for all genetic information, encompassing DNA, histone and other protein forming nucleoprotein complex (Luger and Richmond, 1998). The chromatin is largely divided into 2 distinct classes based on transcriptional status – euchromatin, that is decondensed and transcriptionally active, and heterochromatin, that is condensed and transcriptionally shut down (Akey and Luger, 2003). Histone tails are target for post-transcriptional gene regulation mainly via modifications such as methylation, acetylation, phosphorylation, ubiquitination, sumolyation to name a few that are covalent in nature (Roth et al., 2001, Strahl and Allis, 2000, Paro, 1995, Hsu et al., 2000, Lachner et al., 2001). These modifications alter the accessibility of the DNA for transcription, and as such, can modulate gene expression. DNA methylation is stable inherited epigenetic modification that alters chromatin density and modulates gene regulation (Holliday and Pugh, 1975). It is a reversible process and it is affected by environmental factors, but it is stable throughout process of cell division.

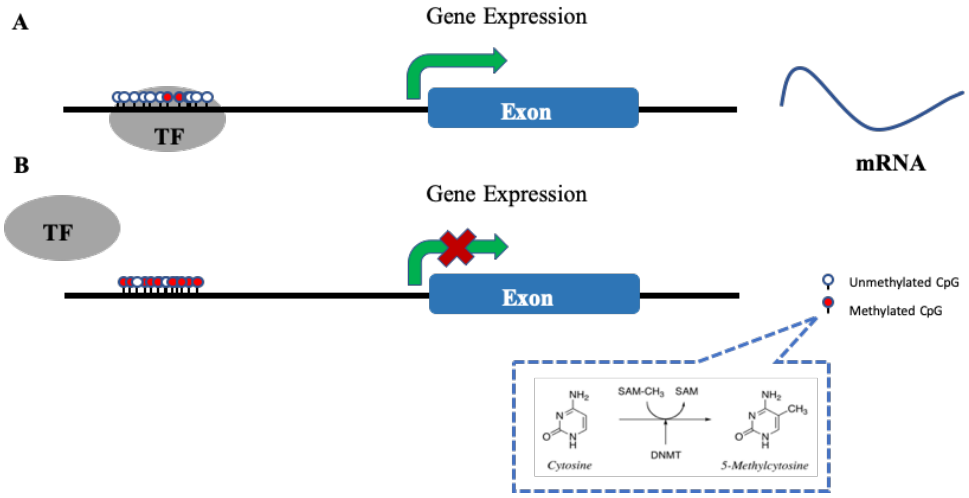


Figure 7: General model of DNA methylation and effects on gene transcription. A) Unmethylated CpG islands in gene promoters allow the binding of transcription factors (TF) that enables transcription and gene expression. B) Methylation of promoter CpG islands inhibit the binding of transcription factor, which in turn inhibits transcription and gene expression.

The process in vertebrates happens with addition of methyl or hydroxymethyl group at 5' end of cytosine by DNA methyltransferases (DNMTs) and it occurs at CG (CpG) nucleotide rich promoter regions in a symmetrical fashion. In plants and embryonic stem cells non CpG methylation is observed that occurs in context of CHH and CHG, where H being A, C or T and it happens in an asymmetrical fashion (Robertson, 2005, Lister et al., 2009, Bock et al., 2012). DNA methylation happens in a non-uniform manner across region. In genome, CpGs occur methylated in majority but a minority of CpGs occur unmethylated at in CpG dense regions known as CpG islands, which is rich in GC content and CpGs with an average length of 1000 nucleotides (Smith and Meissner, 2013, Bird et al., 1985). More than 70 % of curated gene promoters, including promoters for housekeeping genes and widely expressed genes are associated with CpG islands (Saxonov et al., 2006, Larsen et al., 1992). These promoters generally show distinct chromatin organization and transcriptional pattern which is affected by the methylation status of the CpGs present in the island

making it an important player in post transcriptional gene regulation (Deaton and Bird, 2011). This shows importance of CpG methylation levels as a biomarker to study gene regulation to help studying different diseases including tumor.

Epimutations, does not involve changes in DNA sequence but rather imply epigenetic changes such as changes in methylation status of DNA or other chromatin modifications (Holliday, 1987). In disease conditions aberrant chromatin states gives rise to aberrant epigenetic patterns which is identified as epimutations. Epimutation in general involves epigenetic process that causes repression of active genes without suppression of expression, or activation of gene expression (Oey and Whitelaw, 2014, Horsthemke, 2006). It reduces levels of gene products by preventing translation and it can affect either allele of gene or both.

Epimutations in cancer typically appears at somatic cells of non-cancerous tissues at the stage of tumorigenesis as well as later stages in tumor evolution sometimes even later stages of metastasis (Banno et al., 2012, Greger et al., 1989, Sakai et al., 1991). Inactivation of promoter activity by CpG hypermethylation was first found in human retinoblastoma (*RB1*) tumor-suppressor gene (Ohtani-Fujita et al., 1993). Similar inactivation of tumor suppressor genes such as *CDKN2A*, *MTSI*, *MDR1* in human cancer were found to associated with CpG island hypermethylation (Merlo et al., 1995, Herman et al., 1995, Kusaba et al., 1999, Herman et al., 1997). *BRCA1* gene promoter hyper methylation has been reported in breast cancer and ovarian cancers supporting in tumorigenesis (Esteller et al., 2000). In colorectal cancers, *MLH1* hypermethylation is associated with microsatellite instability (Herman et al., 1998). *GSTP1* appears to be silenced in prostate cancer in which the process identified is epigenetic silencing via hypermethylation (Millar et al., 2000). Profiling of CpG island methylation in different kinds of tumors can help in treatment and diagnosis of cancer (Melki et al., 1999, Esteller, 2002, Clark and Melki, 2002).

DNA methylation and chromatin remodelling events play a key repressive role at gene promoters of tumor related genes (Fahrner et al., 2002). Mechanisms involved in DNA methylation are different in malignant cells versus normal cells, as they may appear more abnormal in malignant cells. DNA methylation patterns can have repressive effects in tumors as opposed to that in normal cells. Hypo or hyper

methylation of genes have vital effects in controlling the behaviour cancer development and progression (Jones and Baylin, 2002, Esteller, 2006, El-Osta, 2004). Epigenetic modifications usually happens at early onset of carcinogenesis and these modifications are reversible in nature thus potentially making them a good target for cancer diagnosis and treatment (Herranz and Esteller, 2007). However, so far, the successful use drugs targeting epigenetic features in cancer cells has been limited.

Although hypermethylation of tumor suppressor genes has been recognised as an important gene inactivation event in cancer cells, little has been known about the role of normal cell tumor suppressor methylation as a cancer risk factor. Over recent years some few studies have addressed this question and the preliminary findings are intriguing. In colorectal cancer, mosaic methylation of the *MLH1* gene has been observed in leukocytes (Gazzoli et al., 2002), indicating that there is a background of normal cells with methylation in some of the cancer patients. Also, there are reports of specific families with high risk of colorectal cancer, where *MLH1* methylation of normal cells seem to be inherited (Hitchins et al., 2007). Further, promoter methylation of the *MGMT* gene, it epigenetically silences the DNA repair gene. High levels of promoter methylation of the *BRCA1* gene was observed in peripheral blood cells implicating predisposition of early onset of breast cancer in some patients (Iwamoto et al., 2011, Al-Moghrabi et al., 2014). Importantly, recently, our team performed a large case-control study identifying normal tissue *BRCA1* promoter methylation to confer a significantly increased risk of high-grade serous ovarian cancer. In this study the risk was also seemingly proportional to the level of methylation (Lonning et al., 2018, Lønning and Knappskog, 2018). Notably, *BRCA1* methylation was also detected in cord-blood from newborns, indicating that methylation is an event taking place very early, presumably in embryonic life.

Taken together, these recent findings have led to the novel hypothesis that a certain fraction of cancers are caused by early methylation of normal cells. As such it is highly interesting to assess the tumor suppressor methylation landscape in healthy individuals, in order to potentially find new genes where methylation varies between individuals and may be linked to cancer risk.

1.2.5. MicroRNAs

The discovery of microRNA (miRNA) two decades ago brought about a shift in small RNA molecular biology and changed the understanding of processes that involve post transcriptional regulation. miRNAs are single stranded 20-23 nt RNA molecules that play a pivotal role in modulation and stability of an array of molecular processes in physiological and pathological pathways. These pathways include embryonic development, metabolic, as well as pathways involved in tumor progression such as apoptosis, differentiation, stress response, homeostasis, inflammation, neoplastic progression, cell cycle process. All of these characteristics of the small RNA molecule family obviously make them an important class of molecules and also provides a new field of research for potential biomarkers and therapeutic targets with respect to cancer.

The first miRNA was discovered in *Caenorhabditis elegans*, a nematode model organism. The study found downregulation of protein *LIN-14* with accordance to transcription of *lin-4* gene without translation. It was also noted *lin-4* encoded two transcripts 22 nucleotide and 62 nucleotides that regulated *LIN-14* translation by binding to the 3'UTR regions with antisense RNA-RNA interactions (Lee et al., 1993). This first study was published in 1993(Wightman et al., 1993) and a second miRNA was discovered years later, in 2000, in the same organism, *C. elegans* (Reinhart et al., 2000, Slack et al., 2000). In this study, they found the miRNA let-7 21nt transcript, that works similarly to *lin-4*. Also another group found let-7 to negatively control genes important for larval development (Reinhart et al., 2000, Slack et al., 2000). Homologs of *let-7* miRNA was then found in different species in different phyla, including humans, with functions in developmental stages (Pasquinelli et al., 2000).

miRNA biogenesis can be split into two pathways, namely the canonical and non-canonical pathways. In the canonical pathway (Figure 3), miRNAs are transcribed from coding or non-coding regions of genome as primary miRNAs (pri-miRNAs) by RNA polymerase II. Pri-miRNAs are usually several hundred nucleotides long and variable lengths that have 5' capped guanosine and poly-adenylated (poly A) tail. The

pri-miRNA are then processed into a long precursor miRNA (pre-miRNA) in the nucleus by a large protein complex known as microprocessor. The pre-miRNA contains premature forms of miRNA that contains a hairpin structure created with the help of Drosha, a type of RNase III enzyme and *DGCR8* (DiGeorge syndrome Critical Region 8 or also known as Pasha), a type of RNA binding protein in Microprocessor (Lee et al., 2003, Han et al., 2004, Denli et al., 2004). Exportin, a Ran-dependent nuclear transport receptor protein (Yi et al., 2003), exports pre-miRNA, with a distinctive 3' nucleotide overhang (approximately 2 nt in length) and a 5' phosphate into cytoplasm, where the stem loop structure of pre-miRNA is cleaved by *Dicer*, RNAse type III enzyme.

Dicer form a protein complex with double stranded RNA (dsRNA) binding cellular protein, transactivation response RNA binding protein (*TRBP*) that finally process pre-miRNAs to mature miRNA (Chendrimada et al., 2005). The two strands are separated with respect to base pairing, duplex stability and thermodynamic factors to form a guide strand, i.e., miRNA and passenger strand, that eventually degrades. The Dicer-TRBP complex facilitates the formation of microribonuclear protein complex (miRNP) called RNA-induced silencing complex (RISC) that includes a dsRNA binding protein, trinucleotide repeat-containing gene 6A (*TNRC6A*) and Argonaute protein 2 (*Ago2*), a catalytic protein (Schwarz et al., 2003). The formation of this complex is followed by degradation of passenger strand and binding of guide strand to 3'UTR of target mRNA. The RISC along with guide RNA inhibits the transcription by degradation. The path involves formation cytoplasmic bodies called processing bodies (P-bodies) by localizing miRNA-mRNA incorporated *Ago2* proteins into them and results in either degradation or translational repression (Castilla-Llorente et al., 2012, Hammond, 2015, Bhaskaran and Mohan, 2014).

The non-canonical pathways of biogenesis of miRNAs usually that bypass splicing by Drosha or Dicer independent or independent of RISC complex (Ruby et al., 2007, Cheloufi et al., 2010, Yang and Lai, 2010, Janas et al., 2012).

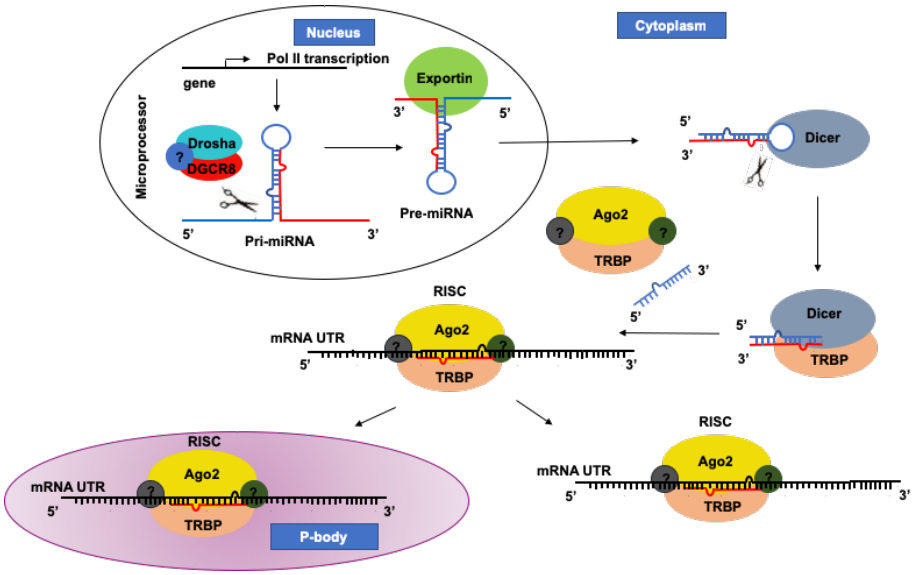


Figure 8: Representational figure of canonical miRNA pathway. Adapted.(Yeung et al., 2005)

The processed, mature miRNAs function by binding to their target genes with sequence complementarity in regions of mRNA, mostly UTR regions. The miRNA sequence from 2nd to 8th nucleotide (called as seed region) (Lewis et al., 2005) binds to UTR of mRNA and suppress the expression of the gene product in two different ways: either by degradation of the mRNA or by inhibition of the translation. The repression of target depends on the complementarity between miRNA seed and target mRNA. If the complementarity is partial, the target genes are in general repressed from translation, while when complementarity is perfect the target mRNAs is more likely to get degraded (Gregory et al., 2005). Notably, translational repression of mRNA is reversible while mRNA degradation is not. The level of gene regulation conferred by miRNAs is complex: a single target gene can be simultaneously translationally repressed by multiple miRNAs with sequence complementarity at different regions of mRNA. Similarly, a single miRNA can control multiple targets and regulate their expression (Li and Kowdley, 2012). A lot of the miRNA mediated gene silencing and mRNA decay takes place at P-bodies in the cytoplasm. P-bodies, along with gene silencing, is also the site for mRNA turnover, and these bodies

contain ingredients involved in mRNA decapping, deadynelation, RNA degradation (Cougot et al., 2004).

The miRNAs play a vital role in reproduction as studies suggest failure in expression of miRNA lead to abnormal reproduction as they affect tissue morphogenesis, apoptosis (Kloosterman and Plasterk, 2006). It has been studied that miRNAs play a key role in regulation of stem cell pluripotency and programming as well as differentiation and self-renewal capabilities of embryonic stem cells (Melton and Blelloch, 2010, Kanellopoulou et al., 2005, Li and He, 2012). miRNAs are known to play a vital role in development of organ systems, such as neuronal cardiac, lung and other major organs. It usually involves controlling expression patterns (temporal or spatial expression) of tissue specific miRNAs in tissue development and differentiation (Hobert, 2006, Zhao et al., 2005, Lu et al., 2007, Ebert and Sharp, 2012, Poy et al., 2004). Even if miRNA still is understudied compared to genetic alterations and regular epigenetic features (DNA methylation), given the total volume of evidence emerging, it is natural to assume that miRNAs are more or less ubiquitously involved in a multitude of cellular processes, similar to genetics and epigenetics methylation.

As miRNAs are found to regulate targets that are important in cellular functions such as differentiation, apoptosis, cell growth, it directly translates their importance into initiation and progression of human cancers (Schickel et al., 2008). A seminal study in 2002 showed that frequent deletion of *miR-15/16* cluster was associated with tumorigenesis in chronic lymphocytic leukemia (CLL) (Calin et al., 2002). It was observed that there is a downregulation of miRNAs which was associated with increase in levels of mRNA expression. Also it was more prominent in tumor tissues than in normal tissues (Lu et al., 2005). Expression profiling of miRNA in human tumors has revealed signatures of altered expression levels as compared to normal tissues. Such altered signatures has also been associated with prognosis and response to treatment (Calin and Croce, 2006, Lu et al., 2005).

In general, the concept of oncogenes and tumor suppressors is also applicable to miRNAs: Upregulation or downregulation of miRNAs as part of a malignant molecular profile of cancer cells, can be attributed to them being either oncogenic

miRNAs (OncomiRs) or tumor suppressor miRNAs (Kota et al., 2009, Trang et al., 2010, Costinean et al., 2006, Ji et al., 2009, Schickel et al., 2008, Calin et al., 2002).

In melanoma, a multitude of miRNAs are found to be differentially expressed and this is in general associated with genomic regions with frequent gains and losses in the tumors. Several miRNAs control and regulate *MITF*, an important gene in regulation of melanocyte development. E.g. the *miR-137* act as a trigger of *MITF* transcription (Bemis et al., 2008), while *miR-182* acts as a negative regulator of *MITF* expression (Segura et al., 2009). The miR-17-92 cluster (*miR-17, -18a, -19a, -20a, -19b-1, and -92a-1*), was found to be playing oncogenic roles with observed upregulation in primary tumor cell lines than in normal melanocyte (Mueller et al., 2009), while inhibition of expression of *miR-221* and *miR-222*) is often oncogenic and it usually involves downregulation of *c-KIT* receptor and *p27Kip* (Felicetti et al., 2008). Further, *MET*, an oncogene that mediates invasive growth, was negatively regulated in melanoma by miRNAs, *miR-34b, miR-34c*, and *miR-199a* (Migliore et al., 2008). In addition to the mere mechanistic observations, several miRNAs have also been linked to clinical features of melanoma. The miRNA *mir-15b* was identified as a potential biomarker for tumorigenesis of melanoma as expression of this miRNA was correlated with both poor recurrence-free survival and overall survival (Howell et al., 2010, Satzger et al., 2010), while low expression of *miR-191* and high expression of *miR-193b* were associated with poor survival rates in melanoma (Caramuta et al., 2010).

All of this points to the miRNAs being an important class of molecules, that should be subject to many future investigations focusing on identification of clinically applicable biomarkers.

In the present work (paper III) miRNA was studied in breast cancer samples. In breast cancers, *miR-10b*, *miR-125b*, and *miR-145* are previously found to be downregulated and upregulation of *miR-21* and *miR-155* was observed (Iorio et al., 2005). As such, association of clinical parameters with expression of miRNAs has proven important in understanding and improving breast cancer clinical outcomes (Nassar et al., 2017, Goh et al., 2016). A study found invasiveness of breast cancer as well as prognostic value were observed to be associated with altered expressions of a

group of miRNAs including *miR-210*, *miR-21*, *miR-106b**, *miR-197*, *miR-let-7i*, and *miR-210* (Volinia et al., 2012). In another study of breast cancer, metastasis and poor survival has been found to be related with expression of *miR-21* (Yan et al., 2008). In specific subtypes of breast cancer, miRNAs *let-7b-5p, let-7c-5p,* and *miR-30a-5p* where found to be downregulated in luminal A and basal-like subtypes. Also, *miR-130a-3p, miR-92a-1-5p, miR-211-5p,* and *miR-500a-3p* were found to be upregulated in those tumor subtypes (Oztemur Islakoglu et al., 2018). While, in breast cancer in general, the *mir-181* family of miRNAs are found to be upregulated, the *miR-181*c in particular is activated by the expression of *HER2* (Tashkandi et al., 2015), and thus closely linked to this *HER2* positive cancers. Further, in ERα-positive breast cancer cells, *mir-140* has been found suppressed by estrogen stimulation. This is most likely due to ER response elements in the flanking element of the *miR-140* promoter (Güllü, 2015). Despite the many emerging correlations between different miRs and specific subtypes of breast cancer, mechanistic information about if and how the miR influence the phenotype of the cancers is to a large extent lacking.

Some of the well-known oncomiRs and tumor suppressors in breast cancer are listed in Table 1.

Table 2: List of miRNAs that act as tumor suppressors (in blue), and oncomiRs (in red) in breast cancer along with their target genes and mode of action in tumor (Di Leva et al., 2014, Corcoran et al., 2011) (Sassen et al., 2008, Di Leva et al., 2014, Hammond, 2015, Bhaskaran and Mohan, 2014, Farazi et al., 2013).

miR	***Target Gene***	**Mode of Action**
miR-15/16	*Wip1*	Regulation of DNA damage response and tumorigenesis
Let 7 family	*Il6*	positive feedback loop control on epigenetic transformed state
miR-200	*ZEB1,*	Down regulation is a main step in tumor progression.

family	*ZEB2, PLCγ1, Suz-12, FN1, LEPR, NTRK2, ARHGAP19*	It sustains cancer stem cell growth as well as invasiveness. Down regulation of EGF-driven cell cycle progression, and invasion. Sensitization of cells to CD95 mediated apoptosis. Inhibition of cell motility and anokis resistance
miR-125	*HuR*	It reduces cell proliferation, migration and induces apoptosis.
miR-205	*HER3*	Inhibition of clonogenic potential and by removing HER3 mediated resistance improves response to tyrosine-kinase based inhibitors
miR-17-92	*HBP1*	Upregulates invasion activating Wnt/β-catenin.
miR-222/221	*FOXO3A, TRSP1, Dicer*	Suppression of the tumor suppressor and apoptosis promoter gene $p27^{Kip1}$. Promotion of epithelial to mesenchymal transition (EMT) In basal- like breast cancers Repression of Dicer in ERa negative breast cancers causing associated clinical aggressiveness
miR-21	*TPM1, PDCD4*	Increases tumor growth. Negative regulation of apoptosis.
miR-155	*WEE1, FOXO3A*	Decreases efficiency of DNA repair and mechanisms. Negatively affect cell survival and response to chemotherapy.
miR-27a	*Sp*	Promotes angiogenesis and proliferation
miR-96, and miR-182	*FOXO1*	Induces myogenic growth and differentiation.

1.3 Contemporary sequencing methods assessing biological factors in cancer

"Sequencing" means techniques used to identify the order of nucleotide bases in a strand of DNA. Watson and Crick solved 3-dimensional structure of DNA, with help of crystallographic data produced by Franklin and Wilkins (Watson and Crick, 1953, Franklin and Gosling, 1953), which lead to new era of finding exact order of how nucleotides are arranged in DNA, it's conformations, replication and translation. The first real sequencing of nucleotides came from ribosomal RNA from microbes and sequencing transfer RNAs (tRNA) (Holley et al., 1964, Holley et al., 1965). At the same time period, long DNA sequencing methods, using detection of radiolabeled and partial digested of ribonucleotides was also established (Sanger and Coulson, 1975). This technology, called a 'plus-minus' method, was used to sequence the complete genome of bacteriophage phi X174, that became the first genome to be sequenced (Sanger et al., 1977a).

A major development in the field of sequencing was Sanger's developed of the 'chain-termination' (dideoxy) technique (Sanger et al., 1977b). Sanger sequencing continued to be improved over the years, with improvements in labeling (Ansorge et al., 1986, Smith et al., 1985), and detection using capillary gel electrophoresis (Ansorge et al., 1987). These arrays of improvements lead to first set of automated Sanger sequencing and commercialization of sequencing (Hunkapiller et al., 1991). The first generation of sequencing was capable of sequencing DNA of lengths less than 1000 bases (1 Kb) and advances in the fields of recombinant DNA (rDNA) technology and polymerase chain reactions (PCR), lead to a range of dideoxy sequencers (Smith et al., 1986).

The next set of evolutions came from introduction of "sequence-by-synthesis", which used measuring of pyrophosphate production using luciferase activity, which emits light proportionate amount of pyrophosphate (Nyren and Lundin, 1985, Hyman, 1988). This method became popular as it can be observed real time and no modification of deoxy ribonucleotides (dNTPs), used in Sanger-sequencing was required (Ronaghi et al., 1998). This method also became known as pyrosequencing

and was licensed by 454 Life Sciences, which became major company in next-generation sequencing (NGS) technology. 454 sequencing machines (later acquired by Roche) innovated using emulsion PCR, which increased the amount of DNA sequenced (Tawfik and Griffiths, 1998, Margulies et al., 2005).

Sequencing by oligonucleotide ligation and detection (SOLiD), developed by Applied Biosystems (which is now Life Technologies), was another popular choice for sequencing (McKernan et al., 2009). The development in sequencing world is still ongoing, evolving and innovating, some of them are Ion Torrent (Life Technologies), which uses difference in pH caused by release of H+ ions (protons). Single molecule sequencing are being developed by companies like Pacific Biosciences, VisiGen Biotechnologies and various institutional collaboration with companies to name few (Heather and Chain, 2016, Ansorge, 2009, Barba et al., 2014). The Solexa method, which was purchased by Illumina, replaced emulsion PCR with complementary oligo nucleotides fastened to flow cell using solid phase PCR (Turcatti et al., 2008). The Solexa method was first implemented in the Genome Analyzer (GA), which could sequence DNA as paired end (PE) reads. These sequencers where followed by GAIIx, HiSeq and MiSeq systems, which made sequencing accessible, affordable, faster and efficient. This technology has over the recent years rapidly become the market leader and most sequencing facilities today applies illumina instruments with Solexa-based technology. This type of sequencing is also the basis for the work performed in the present thesis: all NGS data presented here is generated by Solexa/illumina technology. The specific approaches applied are described in some more detail in the sections below.

The recent technological explosion of developments in the field coined High Throughput Sequencing (HTS) or Next Generation Sequencing (NGS) or Deep Sequencing or Massively Parallel Sequencing (MPS), has revolutionized the field of genomics, and hereunder the field of cancer genomics.

In the following chapters, we will discuss the different sequencing strategies used in projects that are included in the present thesis.

1.3.1. Whole genome and whole exome sequencing

Mendelian disorders for selected genes are often investigated using conventional methods such as Sanger sequencing (capillary sequencing). The limitations of conventional methods required numerous experiments to uncover complex molecular and genetic pathways. The introduction of NGS opened the possibility of carrying out complex experiments in a cost effective and comprehensive manner. The high throughput MPS have brought about advancement in the field of genomics that opens a horizon of opportunities as well as challenges. Application of NGS in different forms include sequencing such as ChiP-seq (Chromatin immunoprecipitation on DNA microarray chip), RNA-seq, *de novo* assembly of gene and genomes, etc. Here we discuss the application whole genome and whole exome sequencing and in sections 3.2 and 3.3 we will discuss bisulfite sequencing and small RNA sequencing, respectively.

Whole genome sequencing (WGS), is a helpful tool to understand the detailed organization of the entire genomic landscape. WGS helps in identifying a plethora of genomic alterations including point mutations (SNVs), indels, copy number changes (CNVs) and structural rearrangements (SVs) to name a few. It also has a unique ability to identify exact breakpoints and the nature of SVs, i.e. whether they are inversions, translocations tandem duplications etc. (Tuna and Amos, 2013, Xuan et al., 2013). Given that WGS provides such a comprehensive overview of the genomic landscape of a sample, this strategy can also be used to detect patterns such as mutational signatures tumors, reflecting the mutational processes the cells have been subject to during the tumor evolution (Alexandrov et al., 2013).

The main advantages of WGS can be summarized as follows:

a) It can comprehensively sequence the whole genome and identify all types of genetic alterations.
b) Achievable coverage of the genome (percentage of the genome covered) is very high.
c) It can detect mitochondrial mutations and disorders in DNA repeats.

d) It can detect DNA not annotated to the human reference genome (e.g. viral elements)
d) Wet-lab procedures (library preparation) is relatively easy, since no selection of regions is required.

The main disadvantages of WGS can be summarized as:

a) Churns out gigabytes (Gb) of data, making it difficult to analyse (requires large computational infrastructure).
b) The sensitivity to identify mutations and subclones etc. is less as compared to other methods (e.g. whole exome sequencing) since the sequencing depth is typically lower (due to costs).
c) WGS is still expensive (Brittain et al., 2017).

While WGS covers the entire genome and comprehensively captures vast amounts of information per sample, there are some drawbacks, as described above. In many cases, including many cases of cancer research, there is a more narrow interest in somatic mutations affecting protein coding regions, and/or there is a need for higher sequencing depth, at a lower cost than what is feasible by application of WGS. The alternative solution to study protein coding regions in genome at a lower cost is Whole exome sequencing (WES). WES can capture protein coding regions of the genome (exons), and potentially other functional regions of specific interest (miRNA genes, and non-exonic untranslated regions (UTR) etc. For the purpose of assessing alteration within the protein coding regions of the genome, this approach is very cost effective (Meienberg et al., 2015).

The main advantages of WES can be summarized as follows:

a) It is cheaper and more cost effective than WGS, when assessing only the protein coding regions of the genome.
b) Data churned out is in more manageable for smaller computational environments.
c) It is sensitive with respect to mutation calling, as it typically has more read

depth in regions that get sequenced.

The main disadvantages of WES include

a) Only a minute portion (1-2%) of the genome is assessed (Brittain et al., 2017), thus many alteration affecting regulatory elements etc are lost.
b) Structural rearrangements cannot be assessed.
c) CNV calling is less accurate than for WGS
d) Wet-lab work and costs related to sample preparation are higher for WES than for WGS, since baits must be used to capture the regions of interest.

WGS and WES are effective in identifying somatic mutations and indels (small insertions / deletions of 1-10 nt) mainly in coding regions as well as in splicing sites in exon-intron junctions. But WGS can detect mutations in intronic sites and in 3' and 5' untranslated regions (UTR), that are important in post transcriptional regulation of RNA and translational processes. To detect copy number alteration (CNA), primarily DNA chip or array CGH was primarily used before advent of WES/WGS. CNA analysis using WES with higher depth, produces results that are lower resolution and more prone to bias than WGS. WGS helps in analysing CNA in a more unbiased way even at a lower depth. As WGS covers the genome more uniformly than WES, and covers potential breakpoints, it is suitable for detection of structural variations. Similarly, as WES does not capture mitochondrial genomes WGS is well suited in that scenario. WGS also helps in effectively identifying sequences from unknown pathogens making it an important method of studying pathogen and host human genomic interactions in diseases (Nakagawa and Fujita, 2018, Petersen et al., 2017, Stranneheim and Wedell, 2016).

WGS and WES differ mainly in the cost of experiment, the amount of time taken for experiment and the data generated per experiment. Cost difference between these technologies are getting closer as well as the time used to sequence is getting insignificant. However, in addition to costs a main issue and challenge is the handling and analysis of the data generated. While WES produces manageable amount of data but, WGS churns out a lot of data but gives comprehensive information covering all

aspects. So moving forward in the field, in the near future, most likely will comprise the use of a combination of both WES and WGS in an experimental setting as WES can provide relevant information effectively and elaborate study on findings can be carried forward using WGS (Nakagawa and Fujita, 2018, Petersen et al., 2017, Stranneheim and Wedell, 2016). At the same time, in a slightly longer perspective it is likely that WGS will become standard when sequencing costs and IT-cost becomes more affordable for most research teams.

1.3.2. DNA methylation detection - Bisulfite sequencing

Genome wide identification techniques to identify epigenetic alterations are being developed rapidly. To study DNA methylation at single nucleotide level, the popular and reliable way is bisulfite sequencing.

In general, there are mainly four categories to profile DNA methylation based on preparation 1) MRE-seq – DNA methylation that are sensitive to restriction enzymes, 2) MeDIP-seq – DNA methylation captured by immunoprecipitation using methylcytosine-specific antibodies, 3) MethylCap-seq – DNA methylation captured using methyl binding domain of MeCP2 (methyl CpG binding protein 2) and 4) Bisulfite-seq – DNA-methylation captured using bisulfite treatment of DNA. The bisulfite method has an edge over other methods as it can measure methylation both at a large number of CpGs and with a resolution of single nucleotides (Baubec and Akalin, 2016, Nagarajan et al., 2013, Wreczycka et al., 2017). Also, bisulfite sequencing has the advantage of capturing the majority of (all) CpGs, allowing identification of less methylated regions in the genome, including regulatory elements.

Bisulfite sequencing can be done whole genome wise (WGBS), reduced representation (RRBS) and targeted (applying a specific panel of genes / regions). Irrespective of the type, the steps involved in bisulfite sequencing are the same: Genomic DNA of interest is fragmented into smaller sizes in order to prepare sequencing library. The fragmented DNA is then ligated with adapters containing 5' methyl cytosines (5mC) that is taken for bisulfite treatment. In the bisulfite treatment

procedure, all unmethylated cytosines are converted to uracils. The bisulfite treated DNA is then amplified using PCR and any uracils are replaced by thymines. The methylation calling in the sequencing procedure depends on bisulfite conversion, library amplification and sequencing depth.

WGBS helps in obtaining global coverage of CpG methylation and considered as global standard. Since WGBS needs larger quantities of DNA as well as being one of the most expensive methods, in real life it is often a less favorable choice of analysis.

An alternative is to get methylation information with higher resolution is treating the genomic DNA with restriction enzymes thus obtaining only regions in genome with higher CpG content (RRBS). This step is a tradeoff for cost effectiveness as it only sequences at CpG dense regions, it tends to miss out on low methylated regions, enhancers, intergenic regions etc.

Targeted bisulfite sequencing is similar to WGBS, but instead of whole genome specific regions are selected based on interest. This method helps in sequencing targets from genomic regions of interest in more cost effective and reproducible manner. This provides an affordable solution to run large number of samples by multiplexing and sequencing without compromising on high resolution (Baubec and Akalin, 2016, Nagarajan et al., 2013, Wreczycka et al., 2017). A specific targeted bisulfite sequencing protocol used in the present work involves is the SeqCap Epi protocol, a proprietary protocol to targeted enrichment of bisulfite DNA. The SeqCap Epi, a procedure that helps on to focus smaller segments of genome for studying methylation analysis at a higher resolution, also has ability to multiplex and sequence multiple samples to study methylation data in an inexpensive manner (Wendt et al., 2018). For details on how this was performed in the present work, see materials and methods, Chapter 5.

1.3.3. Small RNA sequencing

To understand transcriptional regulation, there has been a need to study the small RNA transcriptome. This class of RNAs are not transcribed to protein but have their function in post transcriptional regulation. Small RNAs are generally less than 200

nucleotides in size and they span in diverse classes based on size and functions. Some of the small RNAs are transfer RNAs (tRNAs), micro RNAs (miRNAs), piwi-interacting RNAs (piRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), short interfering RNAs (siRNAs), etc (Mattick and Makunin, 2005, Wang et al., 2017). Evolution of high throughput sequencing and next generation sequencing platforms has vastly contributed to our novel understanding of the complexity of the small RNA transcriptome. In earlier stages, small RNA sequencing, was synonymous to miRNA sequencing, but nowadays one can isolate and sequence small RNA species including both miRNA and several other species (Wang et al., 2017).

Before deep sequencing, the general strategies used in profiling expression of miRNAs was done with large quantities of total RNA run on Northern blot, that required radioactive isotope enabled autoradiography. Sequences of miRNA were identified by cloning products and sequencing with Sanger sequencing method. Also, to some extent microarrays were used in unraveling expression of miRNA signatures in different tissue level expression in a variety of patients (Liu et al., 2004).

Developments in sequencing miRNAs, helped in creating database known as miRbase that hold information of 218 miRNA sequence loci from 5 species. In the latest release, the miRbase has more than 38,000 miRNA entries spanning 271 species (http://www.mirbase.org) (Griffiths-Jones et al., 2006, Griffiths-Jones et al., 2008, Kozomara and Griffiths-Jones, 2011, Kozomara and Griffiths-Jones, 2014). Bioinformatic tools exploit biogenesis of miRNA mechanisms to predict sequences and structures. It would unravel new miRNAs as well as known miRNAs that can shed light into different biological processes involved and their involvements. There are several databases that holds information on miRNA family annotation (Rfam, http://rfam.xfam.org) (Nawrocki et al., 2015), as well as intragenic and structural details of miRNAs (miRIAD, http://www.miriad-database.org) (Hinske et al., 2014). With respect to predictions based on biological features there are myriad number of algorithms that helps in discovery and quantification of miRNAs. miRdeep2 (Friedländer et al., 2012), a tool that predicts canonical and non-canonical miRNAs with quantification. It achieves that objective with incorporation of information from

ViennaRNA package (http://www.tbi.univie.ac.at/RNA), that holds information on RNA-RNA interactions as well as their evolutionary conservations (Lorenz et al., 2011). RNAstructure (https://rna.urmc.rochester.edu/RNAstructureWeb/) imparts information secondary structure prediction and stability of it with respect to the energy of binding (Bellaousov et al., 2013). These are few examples of tools used and there are plenty available that works in a similar lane of objectives that are well suited for next generation sequencing (Chen et al., 2018). Once prediction or quantification of miRNAs are carried out, a typical next step for researchers is to predict the target genes that these miRNA target. For this purpose there are established databases that hold target information from de novo predictions as well as experimental validations, e.g. miRanda (Betel et al., 2008, Betel et al., 2010), TargetScan (Lewis et al., 2003, Lewis et al., 2005), miRDB (Wang and El Naqa, 2008).

1.3.4. General approaches in analysis of deep sequencing data

For the type of work presented in the current thesis, it is essential to describe how transformation of an idea or hypothesis to results occur in a bioinformatics setting. It is important to understand knowledge gathering and information processing, with help bioinformatics tools and methods, to fully understand and resolve complex biological issues. Another important aspect is to integrate data from different sources and platforms to fully exploit bioinformatics approaches, that helps in unravelling complex biological puzzles. Advancements in NGS technologies and availability of cost effective ways to handle bulk amount of data have paved a way for availability wide variety of genomic data. It will be recommended to use combinations of multiple platforms and different types of NGS methods to yield results with higher accuracy and fidelity. I.e. even when applying NGS technologies one must use different approaches for different biological questions, e.g. mutational analyses vs. expression or methylation analyses.

As similar to a wet lab experiments, work with biological questions from a bioinformatics perspective, can roughly be split up into three main stages; design, experiment and analysis (Figure 8). Each of these stages involve essential steps for

proper tackling of a biological problem / research question as wells providing solution for processing information in the same objective. Subprocesses within these three stages will obviously vary depending of the biological problem / research question. In the following, we will describe some general features of the three stages, and also some features that are more specific for the work in this thesis; features that are of particular use and relevance for the field of cancer research.

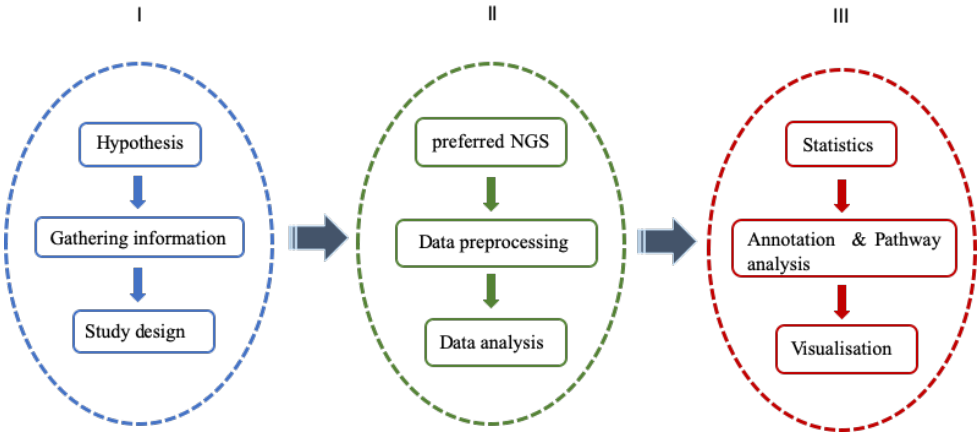


Figure 9: Overview of Bioinformatics experiment workflow. (I) Design, (II) Experiment, (III) Analysis.

The first stage can be described similarly to how any experiment is conceived and designed, regardless of whether it is a pure wet lab experiment or a project where the main work load will be on the bioinformatics side. This is the preliminary stage where an idea is evolved into a research question and consequently to a research project. The main steps involve gathering already known information about the concept and define the areas which is already known and defined. Once the hypothesis is ready after refining through already known information, the next step is gathering already available data and design a study to achieve the goal. With relevance to the present work, here, we discuss with a study design that involves wet lab experiments followed by bioinformatics experiment. As required wet lab experiments suitable to our study design are concluded, we then proceed to finalise what kind of sequencing (e.g., WGS, Bisulfite Seq, RNA-Seq, etc.) on what platform (e.g., Illumina, PacBio, etc.). At present, in addition to the scientific question, type of

sequencing to be used is, heavily dependent on costs. For the field of cancer research, this is also often a question of available material, not only from the tumor, but also other tissues, as normal tissue is still required to distinguish somatic mutations from germline variants in individual patients. Once one has finished wet lab experiments and carried out suitable sequencing, the second stage comes, where the main bioinformatics and computational approaches of the work take over.

The second stage is dependent on the type of sequencing set up in a desired / available platform and the experimental approach. Once sequencing is completed, data from the platform used for sequencing is carried to computational platform accordingly to available facilities. The sequencing of DNA occurs in two different ways, a) single-end, that is the fragment gets sequenced only from one end of DNA, b) paired-end where the DNA gets sequenced from both ends. The first step after sequencing is to ensure the quality of sequenced FASTQ files, a common output format for sequences. These file contains a numerical quality score called PHRED (Ewing et al., 1998, Ewing and Green, 1998) that describes probability of incorrect sequencing per base. This score gives information regarding quality of sequencing between lanes and cycles, GC content, sequence bias, artefacts and other contaminations. This step of quality control can be generally done in software on board platforms or openware tools like FASTQC. The next step involves filtering the sequencing data based on quality score, removal of adapters and other artefacts and make the sequenced reads, ready for analysis. The data analysis step in this stage involves making sense of reads where they come from. Based on experimental set up these reads are aligned to a reference genome of interest with help of tools like BWA (Li and Durbin, 2009), Bowtie (Langmead et al., 2009), etc. For the work presented in this thesis, we used three different version of human genome references such as build b37, GRCh38 and hg19. The choice of genome version was dependent on the nature of the work involved and other factors involved, such as how and when the library preparation protocols were designed. Aligned reads can then be used to identify regions of interest in different kinds of formats and study them both quantitative and qualitative. The analysis from here starts to become diverse based on type of sequencing. Each type of sequencing involves different analytical and

statistical methods to resolve biological problems as per design of experiment. This is where extensive use of different types of programming languages and software is required. For analyses of human cancer genomes, this step usually involves mutation calling and / or use of sequencing depth or germline polymorphisms to assess copy number changes. We used algorithms such as MuTect (Cibulskis et al., 2013a) and STRELKA (Saunders et al., 2012) to call somatic mutations and indels. The algorithm results for somatic variants were improved by filtering against variants detected in matching blood samples. Copy number variations (CNVs) are identified from NGS data using several methods. We used a program called as ASCAT (Figure 10) (Van Loo et al., 2010), to identify somatic CNVs, that calculates allele specific copy number profiles along with calculating tumor ploidy. ASCAT achieves this goal by calculating allele counts with help of read depth at SNPs. This allele counts are then used to calculate logR (normalized log transform of read depths from tumor/normal samples) and BAF (allele frequencies in tumor / normal samples). ASCAT algorithm uses BAF and logR along with GC correction to provide information regarding copy number estimates along with estimation of purity (Raine et al., 2016). Notably, common algorithms for CNV calling has a threshold for how low the tumor cell fraction (TCF) can be in a sample before CNV calling is hampered. E.g. for ASCAT, this is around 20-25%. Thus, for the present work in paper II, this problem led us to create an in house algorithm aiming to improve CNV calls and tumor purity estimations in samples with low TCF.

As a main step in bisulfite sequencing, mapping requires special programs since a crucial step in wet lab protocol changes unmethylated cytosines (C) to thymines (T). Programs that maps bisulfite sequencing reads tackles this problem as well as calculates methylation by taking in account of all C – T conversion in aligned reads and taking into account the Cs and Ts in genome. This step makes it difficult to distinguish C > T substitutions caused by the bisulfite treatment and C > T SNPs that were present in the genome of the samples in the first place. Solving this problem requires dual strand sequencing and special SNP calling programs for bisulfite sequencing reads (based on the fact that SNPs are located in single base pairs, while methylation occurs in CpGs and the C > T change is therefore shifted by one

nucleotide from one DNA strand to the complementary strand). Methylation calls can be carried out using several methods but our choice of method was BSMAP (Xi and Li, 2009) and SNP calling was done BisSNP (Liu et al., 2012).

After the data is condensed into variants of different types, these are then usually annotated to genes and effect on the gene level (e.g. amino acid change) and also usually used for higher level analyses, such as mutational signature analyses, subclone analyses, co-occurrence / mutual exclusivity – analyses, pathway analyses / gene ontology etc. Annotation of single variants in works in this thesis, were carried out mainly using Annovar (Wang et al., 2010), while higher level analyses were performed using tools such as DAVID (Huang da et al., 2009), KEGG (Kanehisa and Goto, 2000) and GATHER (Chang and Nevins, 2006). In identifying and quantifying miRNA, tools effectively exploit mechanisms involved in miRNA-mRNA interactions. We used Miranda (John et al., 2004) and mirDeep2 (Friedlander et al., 2008, Friedländer et al., 2012) algorithms in analysis of small RNA sequencing for miRNA analysis.

The last and final stage involves refinement of analysed data and linking the data to biology; i.e. what does the data tell in terms of the biological questions / hypotheses that was the starting point for analyses. As per sequencing type, we can collect and gather different types of statistical analysis on data attained from stage II. This stage also holds the important part of presenting and publishing the relevant findings. Although, sometimes thought of as trivial, making proper and readable presentations of complex genetic and genomic data is often very challenging and requires large efforts form an informatics side. Once desired statistics is obtained, it can be then visualised using different kind of tools and software such as SPSS, R, Matlab. Visualisation of data and results is most important as this stage directly communicates to the audience. It is easier way to represent our results in an easier and understandable ways.

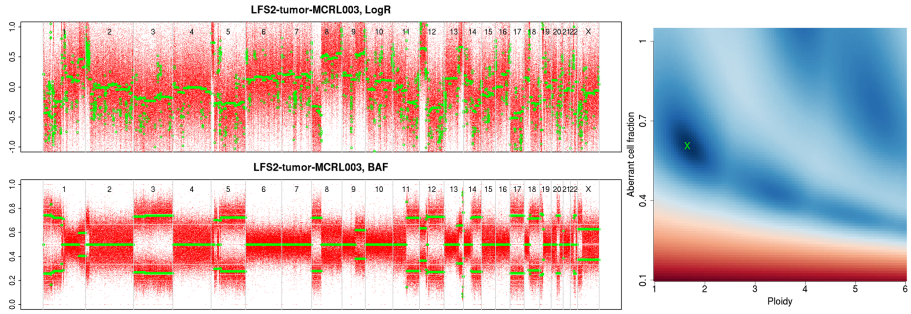


Figure 10: ASCAT profile containing allele specific copy number (left), tumor with ploidy close to 2n(Right).

2. Aims of the study

In the work presented in this thesis we aimed at applying different approaches to massive parallel sequencing and subsequent bioinformatics analyses to gain a better understanding of the molecular characteristics of human cancers. The sequencing approaches applied were whole exome sequencing (paper I), promoter methylation-specific sequencing (paper II) and microRNA sequencing (paper III). The biological characteristics assessed in the analysed samples were multiple and related to survival, response to therapy, tumor heterogeneity as well as cancer risk.

More specifically, the aims of each of the sub-projects (papers) included in the thesis were as follows:

Paper I

We aimed at performing a comprehensive characterization of the genomic evolution in advanced melanoma, through whole exome sequencing of metastatic deposits. This included several metastases from the same patients, and samples collected pre- and post- different types of therapy. Based on these data, we wanted to assess the intra-patient, inter-metastatic heterogeneity including the potential differences between early ("trunk") and late ("branch") mutations. Further, based on these data, we aimed at building a model for the timing of events (mutations and copy number changes) through the tumor evolution in advanced melanomas.

Paper II

Our aim was to try and establish a screening method for methylation patterns in tumor suppressor genes at base-pair revolution. Avoiding the cost of methylation specific whole genome sequencing, we established and applied a predesigned panel consisting of the promoter regions of 283 tumor suppressor genes. A sub-aim here was to establish a relatively fast and cost effective method that could be run on larger number of patients, and that could be run as massive parallel sequencing on

instruments with relatively quick turn-around-time (i.e. MiSeq). We also wanted to establish a data analysis work-flow for data generated from the sequencing, in order to identify potential differentially methylated regions. The back-drop of this work was the identification of *BRCA1* promoter normal cell methylation being a risk factor for ovarian cancer, thus, the overall aim was to establish a method for identification of other potential genes where normal cell methylation could be a cancer risk factor. We then aimed to assess the feasibility of our strategy and provide proof of concept by pilot analyses on normal tissue (blood) samples from 34 healthy postmenopausal women.

Paper III

We aimed to identify potential novel miRNAs in breast cancer tissue, and if any such were found, to explore whether these microRNA are specifically expressed in breast cancer (as opposed to non-cancerous breast tissue). Also, we aimed assessing any such microRNA's biological functions by predicting their mRNA targets as well as analyzing correlations between their expression levels and the expression levels of mRNA involved in different pathways / cellular functions. Further, we aimed at exploring any potential clinical roles of the identified microRNAs, as predictive and / or prognostic markers in breast cancer.

3. Materials and methods

3.1. Biobank material and previous work

In the work presented in this thesis, we have analyzed samples from several biobanks, based on studies designed for different purposes.

Paper I

The analysed samples were from patients included in a single arm study originally designed to assess response to dacarbazine therapy in metastatic melanoma (REK no: 020/00-109.99) (Jonsson et al., 2010, Busch et al., 2010). Tumor samples were collected and snap-frozen in the operating theatre and peripheral blood was collected as control/normal tissue. From a total of 85 patients, 114 samples from 60 patients were available along with matched normal tissue for whole exome sequencing analysis. From 23 out of these 60, two or more metastatic lesions were analysed and used for assessments of intra patient heterogeneity. The collection of samples and initial genetic analyses was initiated by Christian Busch during his PhD work in the team.

Paper II

We analyzed blood samples from 34 women drawn from a study including a total of 114 anonymized healthy postmenopausal women (Helle et al., 2002). These women were recruited from routine mammographic screening at Haukeland University Hospital, Bergen, Norway. At the time of selection, women with diabetes or with other endocrine disorders or individuals undergoing hormone replacement therapy were omitted. These 34 individuals were at BMI range of 19.4 – 39.6 and age range of 56 – 71 years. The collection of samples was performed during the PhD work of Svein Inge Helle and the use for the present project was initiated by Stian Knappskog and Elisabet Ognedal (Berge) during their PhD and post-doctoral work in the team.

Paper III

We included patients from two different studies of breast cancer. The first set of samples were incisional biopsies from a study 223 patients with locally advanced breast cancer aimed to study treatment responses to epirubicin and paclitaxel monotherapy (REK no: 273/96-82.96). Primary responses to therapy along with a follow up of more than 10 years or till death were recorded and available for our analysis (Chrisanthar et al., 2008). We first studied miRNA expression from 50 patients taken from this study and quantified expression of candidate miRNAs using qPCR in all 223 patients. The second study we took advantage of, included 46 anonymized breast cancer patients who were undergoing mastectomy. The study was originally designed to determine tissue estrogen levels in tumor tissue, normal breast tissue from the tumor bearing quadrant, and from non-tumor bearing quadrants (Lonning et al., 2009). The collection of samples was performed during the PhD work of Svein Inge Helle and the use for the present project was initiated by Anne Hege Straume, during her PhD work in the laboratory. In paper III we drew samples from 13 of these patients where RNA was available from both tumor and a non-tumor bearing quadrant.

3.2 Methods in brief

The methods used in the projects in this thesis follow a pattern of sequence library preparation for high throughput sequencing before the actual sequencing process. The data obtained from the sequencing is then analyzed different ways according to the aim of the projects. The results from the analysis is then in turn interpreted with respect to potential implications for clinical aspects.

Paper I

We carried out paired-end whole exome sequencing of samples, followed by quality control of sequencing data. Analysis ready sequencing reads were aligned using BWA algorithm (Li and Durbin, 2009) to the human genome reference build b37. The cleaned data was then called for somatic mutations and indels with the help of two algorithms, Mutect and STRELKA (Cibulskis et al., 2013b, Saunders et al., 2012). This mutation calling included filtering against variants detected in the patients' normal tissue (blood). The intersection of somatic variants from the two algorithms were considered high confidence variants and used for further downstream analysis. The somatic variants were then further annotated for functions (potential effect on gene / amino acid sequence). Annotations were performed using Annovar (Wang et al., 2010). The called and annotated mutations were then restricted to those affecting the protein coding regions of genome. Further, for some sub-analyses, somatic mutations were then classified into driver and passenger mutations (Hodis et al., 2012, Cancer Genome Atlas, 2015, Lawrence et al., 2014). We studied mutational signature patterns, using DeconstructSigs (Rosenthal et al., 2016), involved in these samples to study whether mutations are likely to be caused by any particular previously known mutational process (as reflected by a mutational signature) (Alexandrov et al., 2013). We also studied copy number profiles of the samples along with estimation of allelic tumor copy numbers and tumor purity. These latter analyses were performed using an in-house algorithm that was generated through the work on the paper. As part of an integrated analyses, applying the output from all the analyses

above, we assessed the clonality of mutations and relative timing of mutations, copy number changes and genome duplication events.

Paper II

DNA was extracted from samples taken from healthy individuals recruited form mammography screening and analysis was carried out using SeqCap Epi Enrichment system by Roche. We captured regions of interest for methylation events, with a custom probe design followed by solution-based bead capture of bisulfite DNA. The libraries obtained from capturing was then subjected to a paired end bisulfite sequencing. In collaboration with Roche, an in-house methylation calling workflow was designed and used to identify methylation levels on each single CpG within the regions of interest. In the workflow, the initial steps of preprocessing and quality control was carried out with help of tools in the package Trimmomatic (Bolger et al., 2014). After this step, we aligned processed read to human genome (GRCh38), with Enterobacteria phage lambda (NC_001416.1) complete genome (added for bisulfite conversion efficiency control) using bisulfite mapping algorithm BSMAP (Xi and Li, 2009). Methylation calling needed further processing of aligned reads with help of tools SAMtools (Li, 2011), BamTools (Barnett et al., 2011). This process involves gathering methylation information from both strands, as after bisulfite conversion the DNA strands are no longer complementary. Methylation analysis was carried out by calculating methylation percentage using methratio.py package in BSMAP and SNP calling using BisSNP (Liu et al., 2012). Differentially methylated regions were identified with help of z-score evaluation of average methylation across samples. Sets of identified regions (promoters) with differential methylation were functionally annotated with use of GATHER (Chang and Nevins, 2006) and KEGG (Kanehisa and Goto, 2000).

Paper III

We carried out single end small RNA sequencing of 50 selected patient samples from a study including a total of 223 patients. Sequenced reads were processed and aligned to human reference genome (hg19) as well as known miRNAs from humans and other hominids from miRbase 20 (Sasidharan et al., 2013), using the miRNA prediction algorithm mirdeep v2.0.0.5 (Friedlander et al., 2008). We predicted novel miRNAs from sequencing data after completion of quality control assessment. We quantified levels of selected miRNA in all samples from the study (n=223) with the help of quantitative PCR. We further validated the presence of miRNAs with help of cloning followed by capillary sequencing, to assess *in vitro* poly-adenylation and thus correct miR size. We also predicted target genes for newly identified miRNAs using the offline algorithm miRanda (Enright et al., 2003, John et al., 2004) and the online algorithms miRDB (Wong and Wang, 2015) as well as TargetScanHuman Custom (Release 5.2) (Lewis et al., 2005). Pathways involved and gene ontology for those genes were identified were assessed using GATHER (Chang and Nevins, 2006) and KEGG (Kanehisa and Goto, 2000). Different statistical analysis was carried out with help of the SPSS software v.19 and R, to study the potential clinical impact of the miRNA expression in patients.

4. Summary of results

Paper I

In this study, we carried out whole exome sequencing of metastatic lesions with matched normal tissue (blood) from 60 patients with advanced melanoma. Among these 60, multiple lesions were available from 23. Mutation spectrum and copy number landscape was determined for all samples and the 23 patients from whom multiple samples were available were used to assess intra-patient, inter-metastatic heterogeneity.

We found the genetic differences between metastatic lesions within the same patient to be relatively small. All identified driver mutations were shared between lesions within patients with multiple lesion, but a mutation in p.Y163C of TP53, was identified as heterogenous driver mutation. Mutational signature analysis of the detected mutations indicated that most of the mutations arise from influence / damage caused by UV radiation. Whole genome duplication events were identified in 40 % of the patients by copy number analysis and genomic complexity was observed to be higher in these patients than in patients with diploid tumors. The genome duplication is often followed by copy number losses and duplication events. This in turn leads to more copy number variations and increasing genomic complexity. From analysing clonal status of private mutations, it was seen that there was no evidence to support polyclonal seeding with an exception in the case of one patient. Assessing the influence of therapy, two patients showed peculiar mutation patterns as compared to the others. One of the patients treated with dacarbazine exhibited mutations most likely occurring from inactivation of DNA repair pathways, thereby leaving the cells vulnerable to the DNA damaging nature of the drug. Another patient showed mutations recognised as a mutational signature linked to damage from ionising radiation. Importantly, this patient had received radiation therapy towards a previous metastatic lesion and we found the signature in a subsequent lesion outside of the radiation field. As such, these results provided proof of concept of secondary metastatic seeding in melanoma. Importantly, patients with BRAF mutations showed a selective increase of the mutated allele, in cases with increased copy numbers at the

BRAF locus.

Paper II

In this study we carried out methylation specific sequencing of 565 capture regions representing 356 target regions from of 283 tumor suppressor genes in blood samples from 34 healthy woman (post-menopausal). We obtained 4.95 million reads per samples, on average, and subsequent quality filtering left 88% of reads for further analysis. These reads mapped to genome at an average of 3.6 million reads per sample, yielding a mean primary target coverage of 189.6x across samples. We found CpGs with read depths around 15 x 103 and 0.1 x 106 islands per sample. Bisulfite conversion rate in an internal lambda DNA control was found to be on average of 99.7 %. The reproducibility of the assay was tested and found to be very good, with a technical variability far below the biological variability found in the samples. Thus, overall, the technical quality of the data was found to be sufficient for proper methylation calling analyses.

We found variable levels of methylation across the panel of promoter regions for the 283 tumor suppressor genes included. We identified 149 genes and 206 regions within their promoter regions to be differentially methylated with a confidence level of more than 99%, within the sample set, based on a z-score matrix assessment of methylation ratios across 565 regions from all 34 samples. We found 115 regions with a positive z-score, indicating hyper-methylation in a minority of the samples compared to the majority. Out of these 115, 25 regions showed more than 10 percent points difference from highest to lowest methylated sample. There were 7 regions with more 30 percent points difference within promoter regions of genes: *CIITA*, *RASSF1*, *CHN1*, *PDCD1LG2*, *GSTP1*, *XPA*, and *ZNF668*. We also identified regions, in genes, having more than 2-fold relative difference, while lower than 20 percent points, in *AIP*, *RABEP1*, *RASSF1*. From extensive literature and database searches on the 7 identified genes, we found that *CHN1*, *PDCD1LG2*, *XPA*, *ZNF668*, *RABEP1* were not reported differentially somatically methylated in tumors. On the other hand, *CIITA* was reported as hypo-methylated in a very small fraction of tumors. *RASSFI* and *GSTP1* were reported to be hyper-methylated in cancers, indicating that these

genes could be potential candidates for normal cell methylation assessment, for identification of involvement cancer risk.

In our study we also used the methylation data to assess potential co-methylation patterns across genes and individuals. By hierarchal clustering, we identified two potentially major clusters and other sub clusters of altering methylation. However, given the limited number of samples in the present study, these clusters must be validated in larger sample sets. Assessing difference in methylation and relative difference of between region with minimum and maximum methylation across patients, methylation was found to be highly variable across tumor suppressor genes. It was also observed that some genes showed hyper methylation across regions for each patient as well as across patients for each region. Genes identified as hyper methylated have been previously reported to be hypermethylated in different cancers. This solidifies the impact of studying tumor suppressor genes as candidates that could be used to assess cancer risk on large case-control studies. Most importantly, we present a feasible method by which future studies in this field could be conducted.

Paper III

We investigated small RNA sequencing data from biopsies taken before treatment, from 50 patients with locally advanced breast cancer. Our initial analysis predicted 10 novel miRNAs within these samples. Eight out of these were found in single samples. Thus, based on the presence in two or more patient samples, we narrowed these findings down to two novel miRNAs. First, we validated the presence of these miRNAs by cloning them into a carrier plasmid and sequencing them. Further validating the presence of these two miRNAs, we applied a highly sensitive qPCR to the entire patient set from which the first 50 were drawn (n=223). The qPCR results showed presence of the two miRNAs (temporarily named as hsa-miR-nov3 and hsa-miR-nov7) in 206 and 214 patients, respectively. We then analysed a separate set of 13 breast cancer patients, where we had samples from both tumor tissue and normal tissue from a non-tumor quadrant of the same breast. In this paired sample set, both miRNAs hsa-miR-nov3 and hsa-miR-nov7 were found to be overexpressed in tumor

as compared to normal breast tissues, strongly suggesting overexpression of these miRNAs in locally advanced breast cancer. We used mRNA expression data available for 203 patients out of the 223 patients to narrow down predicted target genes for the identified miRNAs. Assuming the miRNAs to inhibit and/or degrade mRNAs, we took predicted target genes that were inversely correlated with the miRNAs and analysed gene ontology and pathways. It was seen from pathway analysis of target genes for these 2 miRNAs that they played roles in cell development, communication and homeostasis. Assessing the miRNAs' impact on prognosis, we found hsa-miR-nov3 to have no association to patient survival. hsa-miR-nov7, showed an association between high expression levels and poor survival, but this association did not reach statistical significance.

5. Discussion

Paper I

We believe that this is the first study that explored heterogeneity in multiple metastatic deposit in melanoma in a systematic fashion, thus assessing the genomic evolution in the late stages of the disease. While some previous studies have also dealt with genomic evolution in melanoma, these have dealt with alterations in the genome occurring during melanoma progression, along with locally advanced disease and thus provided little or no information about the late metastatic evolution (Shain et al., 2015, Ding et al., 2014, Sanborn et al., 2015, Harbst et al., 2014, Harbst et al., 2016).

In general, we found heterogenous mutations in low numbers and the vast majority of mutations to be truncal, in accordance with regional metastatic disease (Harbst et al., 2016, Harbst et al., 2014, Sanborn et al., 2015). Scarcity of private mutations in metastases was similar to previous findings in breast cancer (Yates et al., 2017), that shows metastasis to be a late event that occur during evolutionary divergence. The fact that we observed branch mutations of UV-related mutational signature (Alexandrov et al., 2013), is consistent with our proposed hypothesis, that different metastasis may arise from different subclones in the primary tumor. We found a low number of private mutations across individual lesions with a correspondingly high intra-patient consistency. A similar observation was described in metastatic breast cancer (Yates et al., 2017). As this phenomenon was found across two tumor forms with the presence of different mutational signature patterns, this may indicate it to be an intrinsic propensity related to several cancer forms. Along with this finding we observed that heterogeneity correlates to BRAF mutation status, similar study done in primary melanoma (Shain et al., 2015), that imply presence of genetic mechanisms associated with this process. From our data, most metastases tend to have a monoclonal origin with exception on one patient that had indications of re-seeding, contradictory to findings made by Sanborn and colleagues that suggests re-seeding to be a common phenomenon (Sanborn et al., 2015). Notably, it appears that many of the tumors, which they used to find shared sub clones, were loco-regional relapses

that were in close anatomical proximity to each other. Loco-regional relapses, in spite of having wide surgical excision margins (Leiter et al., 2004, Urist et al., 1985), are a characteristic of thick and large primary melanomas. We found signs of re-seeding between metastases in one patient (MM61). This patient showed an unusual clinical phenotype, having cutaneous metastatic deposits at more than 100 sites prior to death. These deposits were found at different anatomical locations like truncus, shoulders and head area and it was along with the 5 other lesions sampled. This suggests the development of cutaneous melanoma metastases to have an organ specific propensity in line with the studies from Nguyen and colleagues (Nguyen et al., 2009). Even though the patient showed distant metastases, tumor cells trafficking were similar to patterns of reseeding observed in disease that disseminated regionally (Sanborn et al., 2015). We observed whole genome duplication (WGD) events prior to metastatic divergence and it was associated with high copy number diversity relative to near-diploid tumors. This ongoing process of copy number alterations was similar to observations from other cancer forms (Yates et al., 2017, Jamal-Hanjani et al., 2017).

We found *BRAF* mutations with arm- or chromosome-spanning gains of 7q, that have previously been described in primary melanoma (Maldonado et al., 2003). Also, we found that low level gains of mutated BRAF allele occurred earlier than whole genome duplication. Thus, in general, the amplification of mutant *BRAF* alleles is a rather early event in BRAF-mutated tumors. While *BRAF* gains through focal amplification of smaller segments has been described as a mechanism for acquired resistance to BRAF-inhibitors occurred (Shi et al., 2012, Shi et al., 2014), our finding indicates *BRAF* amplification also to be progression driving in non BRAF-inhibitor treated cancers. It would be necessary to validate that selective advantage growth is boosted by the presence of low level gains of *BRAF* similar to low level gains of *KRAS* mutants in lung cancer (Kerr et al., 2016).

In one patient, we observed an alkylating chemotherapy mutational signature, in relation to DNA mismatch repair defects. The signature was previously reported in melanoma in relation with temozolomide treatment but not linked to any other genomic alteration (Alexandrov et al., 2013). In our study, we observed this signature in a patient with numerous *MSH6* mutations. Since none of the other patients

receiving the same therapy had a similar mutational signature, this finding may indicate that a DNA mismatch repair impairment is required as a background for the therapy induced mutational signature to occur.

In another patient, MM85, we found a metastatic deposit with a radiation damage mutational signature. Previously, a prominent mutational signature was observed in secondary cancer from areas that had undergone carcinogenic ionising radiation (Leuraud et al., 2015). This signature was characterised by accumulation of small deletions in secondary cancers (Behjati et al., 2016). In our patient, we found truncal 2-nt deletions and multiple private deletions that resembled the above-mentioned signature in two distant metastatic deposits after 5 and 6 months of radiotherapy for a regional lymph node metastasis. It is difficult to find direct evidence for secondary spread of tumor in melanoma and cancer forms and it remains a topic of controversy (Morton et al., 2006). However, our finding of a radiation signature in deposits far away from the previously irradiated region (that in turn was far away from the site of the primary tumor), strongly indicates secondary seeding of metastases in this patient. Unlike chemotherapy, where tumor cells and metastases are affected irrespective of their anatomical location, radiation therapy is more focused and applied at localised regions with limited radiation scattering outside the treatment field. We identified from radiation signature that in turn acts as a "cellular labelling", that indicated secondary seeding from radiation treated lymph node to chest wall and liver. Emergence of clonal 2-nt deletions in two deposit within 6 months of radiation may even raise some concerns regarding whether radiotherapy treatment could have acted as an enhancer of metastatic propensity and tumor aggressiveness. However, although we observed the mutational signature to occur, the biological effects of these radiation-induced deletions are still unknown, and it may well be that none of them were real driver-mutations contributing to the disease progression.

In general, for melanoma progression, this study provides evidence for common patterns of genomic alterations. It seems from our study that in most cases, metastatic deposits have a monoclonal origin with a possible exception of patients harbouring multiple cutaneous deposits. The secondary spread from metastatic deposits may have potential clinical implications and further studies are required to further

characterize this phenomenon in advanced melanoma.

Paper II

As described in chapter 2.4, previous findings have linked methylation of both *MLH1* and *MGMT* to cancer risk. Importantly, our team's recent findings have also firmly established that early life methylation of *BRCA1* confers an increased risk of ovarian cancer (Lonning et al., 2018, Lonning et al., 2019). We see it as likely that similar mechanisms may be at play for other tumor suppressor genes and other tumor forms. In other words that early life and/or inherited methylation may be a cause of a considerable fraction of human tumors.

In light of this hypothesis, it will be crucial to investigate additional tumor suppressor genes where differential methylation in the healthy population may be detected. We therefore set out to establish a wet-lab- and informatics pipeline for this type of investigation. In order to investigate such hypothesis, it is important to apply adequate tools. There are vast variety of epigenetic data based on single gene analyses by MSP, MLPA, pyrosequencing or global methylation-array analyses or other similar technologies. However, a broad approach, at a base pair resolution is required to pinpoint exactly what CpGs / regions of CpGs that are important for individual genes to screen for potential candidate genes as risk factors for cancer. The ideal case would be to perform whole genome methylation specific massive parallel sequencing, but this is currently limited by high costs.

We established a gene panel of 283 tumor suppressor genes, consisting of the main regions of interests with respect to transcriptional regulation, i.e. restricted to the promoters. Also, we established a massive parallel sequencing-based approach, enabling base pair resolution analyses methylation status in the gene promoters. Using the gene panel helped in massively reducing the cost/resources required for analyses and increasing the potential for high-throughput assessments. We successfully made DNA-libraries that could be run on any illumina instrument and to save time and availability, we ran it on MiSeq. In our runs, we obtained an average coverage of approximately 190x, but there were regions with lower coverage. We

managed to achieve a sensitivity 0.5%, on average. From our previous studies, we found methylation of *BRCA1* to confer a risk of ovarian cancer when >4% of alleles in WBC were methylated. As such, a sensitivity of 0.5% should be adequate, although in theory, there could be tumor suppressors that confer a significant cancer risk even when methylated in a lower fraction of alleles in an individual.

In our present proof-of-concept experiments, the limited number of samples from individuals precluded any form of formal assessment of potential risk factors related to genes that are methylated in a low percentage of the population. We know methylation is required in at least 11% of the population in order to detect at least one methylated individual among 34, with 95% CI. One of the best-characterized epimutations conferring cancer risk so far, for *BRCA1*-methylation, we previously observed methylation in approximately 4%, of healthy individuals (Lonning et al., 2018, Gazzoli et al., 2002). This states that our study was too limited to fully uncover all tumor suppressors that may be cancer risk factors when methylated. However, we put forward a proof-of-concept for a strategy to identify risk factors that should be extended to large cohorts.

Even within our limited cohort of 34, we were able to identify some genes that may be potential risk factors that could be explored in case-control studies in a larger cohort. We also identified the tumor suppressor promoters most hypermethylated in a minority of individuals, to be those of *CIITA*, *RASSF1*, *CHN1*, *PDCD1LG2*, *GSTP1*, *XPA*, *ZNF668* and *RABEP1*. Interestingly, in these tumor suppressor genes, three of them were reported with germline mutations and have been found to be epigenetically deregulated somatically in cancer. Hypo methylation was reported in *CIITA* across several cancer form, in the COSMIC database (Forbes et al., 2017), while importantly, both *RASSF1* and *GSTP1* have been found somatically hyper-methylated in cancers (Forbes et al., 2017). Normal cell methylation of *RASSF1* and *GSTP1* is not well studied, but the presence of somatically epigenetic deregulation of these genes in cancers, may indicate a potential role also in terms of cancer risk and thus potential roles as targets in larger scale case-control studies designed for assessment of cancer risk.

One potential caveat with our studies is the fact that different leukocyte fractions

may harbor different methylation patterns. This poses a scenario where differences in detected methylation between individuals might arise from differences in their distribution of leukocyte sub-fractions. From our previous studies we know that CpG-rich region of the *BRCA1* promoter, seems equally methylated across all leukocyte fractions (Lonning et al., 2018), indicating that this problem may be negligible for tumor suppressors. On the other hand, there is no guarantee that the findings in *BRCA1* could be extrapolated into all other tumor suppressor promoters. In our present study, we found certain regions in well-known cancer risk genes such as *MSH2* and *PALB2* to be relatively highly methylated in all individuals. However, the finding of e.g. *MSH2* methylation on limited, defined regions, rather than entire promoters, makes their effect on transcription uncertain.

In a heterogenous population, it may be a challenge to assess and identify methylation differences as potential cancer risk factors, since characteristics of DNA methylation is more dynamic than mutations, that are stable alterations. In our study design, one of the strengths in this respect, is that the individuals included are drawn from a relatively homogenous group. This homogeneity relates to several key factors that are well known to affect epigenetics, such as gender, age and hormonal influence. Also, the analysed individuals (Norwegians) were homogenous in terms of ethnicity.

We found 2 major clusters of genes in individuals based on the tumor suppressor methylation patterns. Since the individuals we analysed were limited in number, these clusters should be interpreted with caution and the results need confirmation in larger sample series. Nevertheless, one group of genes clearly differing between the two clusters (A), consisted of genes important in pathways like Wnt signalling and TGF-beta signalling that are important in developmental and regulation cellular processes. Another group of differing genes (B) were found to be involved in apoptotic pathways and leukocyte differentiation. This implicates that group A may reflect methylation patterns of relevance to cancer risk while group B could also be implicated in cancer risk but may well be a result of our tissue of choice for analysis being blood.

In conclusion, we here designed a proof of concept for a relatively fast and affordable strategy for detailed assessments of differential methylation of tumor suppressor promoters. This strategy is attractive in the warranted search for additional tumor suppressors that may be cancer risk factors when methylated in normal tissues. Here, we applied a MiSeq sequencer, but use of HiSeq or NovaSeq for high capacity analyses of many more individuals over short time would be feasible. This would also help to discover and identify regions and other specific features between individuals. We made some interesting findings indicating some potential genes that could be interesting to do larger case control studies are warranted.

Paper III

In this paper our aim was to identify novel miRNAs, potentially specifically expressed in breast cancer, in a cohort of locally advanced breast cancer samples. By use of massive parallel sequencing, we found 10 novel miRNAs out of which 2 were found present in more than one patient, and therefore considered as trustworthy findings (these were preliminary termed *hsa-miR-nov7* and *hsa-miR-nov3*). In spite of these two miRs only being predicted in 2 and 6 out of 50 initially sequenced biopsies, both were found expressed in the vast majority of patients with varying levels of expression, when assessed by highly sensitive qPCR. Our initial findings were also validated in vitro.

In a separate set of breast cancer patients, we assessed expression of these 2 miRs in tumor tissue versus matched normal breast tissue, from a non-tumor bearing quadrant. From the relative expression from matched samples, we suspected that these miRs may play a role in breast cancer as they were more highly expressed in tumor samples as compared to normal breast tissue. However, the absolute expression levels as well as the ratio of overexpression in tumors versus normal tissue were still rather low, so the biological roles of these miRs with respect to breast cancer have to be interpreted with caution.

The potential involvement of these miRs in cancer development and progression was also studied with help of in silico prediction of targets followed by validation using correlation study using mRNA array data, the KEGG and GO annotations.

While correlations to biological and cellular processes were detected, no of these were clearly linked to cancer, to such a degree that the miRNAs specific function can be said to be cancer related.

We found hsa-miR-nov3 to be significantly higher expressed in ER-positive breast cancers as compared to ER negative ones. We also found high expression levels of hsa-miR-nov3 in the expression based breast cancer subtypes like luminal and normal-like tumors. Whether the hsa-miR-nov3 has any role in the development in these subtypes of breast cancer or whether the correlations are mere co-variates of other molecular features in these cancer remains to be assessed.

We then carried out a search for potential targets for these two miRs. Realising the variability in different prediction algorithms, we chose a conservative approach and used three different target prediction algorithms followed by filtering of the results by only using the intersection of predicted target from all these 3 algorithms. In a subsequent step, we further refined the list of targets using the intersection with a predefined list of tumor suppressors. We could not observe any statistically significant inverse correlation of these final set of genes to miRNAs. But there were some interesting connections. We propose ATRX as a target *hsa-miR-nov3*. This is a gene involved in in chromatin remodelling. It is part of the SWI/SNF family, and it has been associated with LOH in breast cancer (Roy et al., 2008). This was in line with our finding where it was reported that mutations in the SWI/SNF family genes to be enriched in relapsed breast cancer as compared to primary cancers (Yates et al., 2017). Thus, this supports the hypothesis of a breast cancer promoting function for *hsa-miR-nov3*. Similarly, for hsa-miR-nov7, we propose APC, SFRP2, and CDH11 as potential targets. Interestingly both APC and SFRP2 are involved in regulation of the Wnt-signalling pathway (von Marschall and Fisher, 2010, Rattner et al., 1997, Hankey et al., 2018) and are reported targets for targets for several miRNAs in breast cancer (Isobe et al., 2014, Tan et al., 2016, Liu et al., 2018). This may imply a role for *hsa-miR-nov7* in Wnt signalling from our observations. Notably, *hsa-miR-nov7*, during our work with the present project, was identified by Lim and colleagues as miR-10393-3p (Lim et al., 2015). Their study reported this miRNA was associated with pathogenesis of Diffuse large B-cell lymphoma (DLBCL) by targeting genes

involved in chromatin modifications. While this differs from our present finding and study setting, this may be like due to tissue specific effects of the miRNA. As such, we need further investigations to fully identify the functions of these two miRs.

We examined the possibility of association of miRNAs *hsa-miR-nov7* and *hsa-miR-nov3* with clinical outcomes in 223 breast cancer patients based on finding that were overexpressed in the tumor tissue of breast cancer patients. These patients were enrolled to assess response to primary chemotherapy (monotherapy epirubicin or paclitaxel in the neoadjuvant setting) that enabled us to assess association of *hsa-miR-nov7* and *hsa-miR-nov3* levels with primary therapy response and also with long term survival (10-years). We found weak trends only and no significant impact on treatment response or survival, regarding any predictive or prognostic role for the two investigated miRNAs. We found no effect in epirubicin arm, but we found trends towards poor overall survival and relapse-free survival linked to *hsa-miR-nov7*, in the paclitaxel arm of the assessed clinical trial. No prominent association to clinical outcome was found for *hsa-miR-nov3*.

As number of samples included in the quadrant study was low, even though we observed overexpression in cancerous tissue versus normal breast, we cannot confirm the prognostic role of these miRNAs. This issue can only be sorted with further studies with a larger patient cohort. Alternatively, as the observed overexpression in tumor tissue compared to normal breast tissue may indicate signals from tumorigenesis and implies the miRNAs could play a role in tumorigenesis but not later tumor progression.

General discussion

The presented work, focus on molecular features of two of the most common cancer forms in Norway. Although both are well studied, there are still many important unanswered questions regarding the molecular features of both cancer forms. In the present work, paper I represent a large effort to reach new insight into the genomic evolution of metastatic melanoma. Here, we provide new biological information. Paper II is angles differently; here we perform a pilot study, paving the way for future

studies that may be much larger and answer the biological question that was underlying the pilot: are there any additional genes where early life methylation of normal cells may cause cancer. This question remains open, but we have provided a possible strategy by which future studies can address this question. The work for paper III was originally designed with the intention to identify novel miRNAs. Even if we did so, the number of detected novel miRNAs was limited and the question of the biological importance of low level expressed miRNA remains open.

From a methodological perspective, the work in the present thesis represent three different approaches for application of massive parallel sequencing in translational cancer research. While the small RNA sequencing to identify novel miRNAs is a "global" approach, without any pre-selection of regions, both papers I and II rely on selection of specific regions before sequencing. Although these strategies of pre-selection have their advantages as discussed in the sections above, it is clear that the way forward for translational cancer research will be full genome assessments, both for mutation calling, but also probably for methylation calling. As such, it is likely that the sequencing strategy used in paper III will be applied also in a longer time perspective than the strategies used in papers I and II.

6. Future perspectives

Paper I

This study provided insight into common patterns involved in genomic alterations at time of melanoma progression. We performed our study on metastatic lesions of melanoma and we could track back "historical events" in the evolution of the disease. In order to validate our findings and possibly also dig deeper in the earlier events during disease progression, it would be very interesting to perform in depth analysis of primary tumors as well as regional metastatic deposits. It is very hard to design a prospective study for both primary and metastatic disease because one would have to recruit patients at the time of primary and then "wait" for the metastases. The most realistic option would be to use retrospective material by gaining access FFPE-block of the primaries after patients have been recruited to the study due to metastatic disease. For the sample set used in Paper I, such collection of FFPE-material from the primary tumors is currently ongoing.

It should also be noted that we applied WES in our present study. This led to a very low number of data points for some patients in some of the analyses (both analyses with respect to subclonality and with respect to mutational signatures of branches versus trunks in the phylogenetic trees of each patient's disease). It is clear that, in the future, such analyses would benefit from full genome sequencing (WGS), a method that would also allow for analyses of other types of mutational signatures than SNV-signatures, such as rearrangement signatures. As such full WGS would provide us with richer data set to assess the timing of events during the metastatic process in melanoma.

In addition, heterogeneity between metastases is most likely not limited to alterations on the genetic / genomic level. There may be important differences with respect to epigenetic features. As such, it would be of interest to characterise the sample set used in paper I for epigenetic variations.

Paper II

In this study, we have effectively established a workflow to identify potential cancer risk factors in healthy individuals by the evaluation of patterns in promoter methylation of tumor suppressor genes. The main limitation to our study was the smaller selection of individuals (n = 34). Even though we provided proof-of-concept that our method could be used to identify hypermethylated tumor suppressors in normal tissues in a small minority of healthy population, it is clear that the real validity of the approach lies in finding a real risk factor in a much larger sample set. Our initial study was restricted to females following the findings in our previous paper identifying *BRCA1* promoter methylation as a risk factor for ovarian cancer. In next phase, where we plan to include more samples, we would also include male individuals since we expect similar mechanism of tumor suppressor methylation as risk for other tumors irrespective of gender. We have provided an analysis strategy to identify novel potential candidates as these epimutations may confer elevated risk of cancer in individuals. Provided we can identify likely important risk factors by larger screening of healthy individuals, an obvious second step would be scale up further, and perform larger case-control studies to assess the level of risk (ORs) related to methylation in our candidate genes. Since the ORs related to methylation might be limited (lower than what we see for mutations in the same genes), it is clear that large cohorts are needed. Such cohorts are perhaps difficult to get access to in Norway, but for our findings related to *BRCA1*, we have now an established collaboration with the American Women Health Initiative (WHI) biobank, and it may be possible to use this, or similar, case-control studies of other genes as well in the future.

Paper III

The main aim of our study was to identify novel breast cancer miRNAs. We detected two novel miRNAs in 2 and 6 patients, respectively, by initial screen using NGS analysis, and then in all most all patients by quantitative PCR. This indicates that the sensitivity was much higher for the qPCR method than NGS and underlines the requirement of higher depth of sequencing for future screens for novel miRNAs by

application of NGS.

In our screen, we found 8 novel miRNAs in one patient each. Due to the sensitivity issue discussed above, it may be that these miRNAs are specifically expressed in locally advanced breast cancers and/or that these are in fact expressed in many more patients, but that they were missed by our screening. If we had chosen to test these by ultra-high depth of sequencing or by qPCR assay, then the outcome could potentially have been different. However, even if detected in the patient samples, we are unsure of the biological impacts of these miRNAs on clinical features such as response to treatment and prognosis since the expression levels must be very low.

The next step, based on the findings for *hsa-miR-nov7* and *hsa-miR-nov3*, would be to further assess why they are overexpressed in breast cancer; whether they have any impact on features such as cell cycle progression, invasiveness, metabolic changes etc, or whether the observed overexpression is merely a side effect of some other deregulated mechanism.

Future of NGS-based molecular cancer research

Many of the issues discussed above as future perspectives, are related to the generation of more data, higher sequencing depth etc. Today, many of the analyses we would like to do are limited due to prohibitively high costs related to sequencing, availability of proper informatics tools, storage space etc. It is important to note that new and exciting technologies and instruments are already on their way, that would enable many of the analyses we cannot do today. In the time passed since the wet-lab work for this thesis was completed and up until now, the team has acquired and installed the very first NovaSeq instrument in Norway. This has led to large drop in sequencing costs and enabled a massive increase in the throughput of analyses. With time, in the not so far future, it is likely that further drops in sequencing prices will make WGS almost as cheap as WES. In a longer time, perspective, it is clear that most of the issues related to restrictions in data amounts will be solved by affordable high depth WGS.

Further, alternative sequencing technologies yielding longer reads are also being

developed. The nanopore technology is already established in many laboratories requiring long reads. For the field of cancer research, this technology is still hampered with a high error rate for single base calls, making it challenging to use for mutation calling. However, cancer researchers already use this type of sequencing for calling of structural rearrangements, since it gives high confidence in identifying break points. Also, it is an ideal technology for identification of splice variants on the mRNA level. With reduced error rates, it is reasonable to assume that this type of sequencing will become more used in cancer research in the future. Notably, it is also possible that this technology will be able to distinguish between methylated and unmethylated Cs in the DNA, and thereby merge genetic and epigenetic wet-lab analyses into one.

With the ongoing rapid technological development in DNA sequencing and the continuous generation of large data sets, handling the vast amounts of information will for sure be a major challenge in the future, but also a great source of knowledge that can be explored.

7. References

2004. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am J Hum Genet,* 74**,** 1175-82.

2012. Comprehensive molecular portraits of human breast tumors. *Nature,* 490**,** 61-70.

AKEY, C. W. & LUGER, K. 2003. Histone chaperones and nucleosome assembly. *Curr Opin Struct Biol,* 13**,** 6-14.

AL-MOGHRABI, N., NOFEL, A., AL-YOUSEF, N., MADKHALI, S., BIN AMER, S. M., ALAIYA, A., SHINWARI, Z., AL-TWEIGERI, T., KARAKAS, B., TULBAH, A. & ABOUSSEKHRA, A. 2014. The molecular significance of methylated BRCA1 promoter in white blood cells of cancer-free females. *BMC Cancer,* 14**,** 830.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A., BEHJATI, S., BIANKIN, A. V., BIGNELL, G. R., BOLLI, N., BORG, A., BORRESEN-DALE, A. L., BOYAULT, S., BURKHARDT, B., BUTLER, A. P., CALDAS, C., DAVIES, H. R., DESMEDT, C., EILS, R., EYFJORD, J. E., FOEKENS, J. A., GREAVES, M., HOSODA, F., HUTTER, B., ILICIC, T., IMBEAUD, S., IMIELINSKI, M., JAGER, N., JONES, D. T., JONES, D., KNAPPSKOG, S., KOOL, M., LAKHANI, S. R., LOPEZ-OTIN, C., MARTIN, S., MUNSHI, N. C., NAKAMURA, H., NORTHCOTT, P. A., PAJIC, M., PAPAEMMANUIL, E., PARADISO, A., PEARSON, J. V., PUENTE, X. S., RAINE, K., RAMAKRISHNA, M., RICHARDSON, A. L., RICHTER, J., ROSENSTIEL, P., SCHLESNER, M., SCHUMACHER, T. N., SPAN, P. N., TEAGUE, J. W., TOTOKI, Y., TUTT, A. N., VALDES-MAS, R., VAN BUUREN, M. M., VAN 'T VEER, L., VINCENT-SALOMON, A., WADDELL, N., YATES, L. R., ZUCMAN-ROSSI, J., FUTREAL, P. A., MCDERMOTT, U., LICHTER, P., MEYERSON, M., GRIMMOND, S. M., SIEBERT, R., CAMPO, E., SHIBATA, T., PFISTER, S. M., CAMPBELL, P. J. & STRATTON, M. R. 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415-21.

ANAMPA, J., MAKOWER, D. & SPARANO, J. A. 2015. Progress in adjuvant chemotherapy for breast cancer: an overview. *BMC medicine,* 13**,** 195-195.

ANDOR, N., MALEY, C. C. & JI, H. P. 2017. Genomic Instability in Cancer: Teetering on the Limit of Tolerance. *Cancer research,* 77**,** 2179-2185.

ANSORGE, W., SPROAT, B., STEGEMANN, J., SCHWAGER, C. & ZENKE, M. 1987. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res,* 15**,** 4593-602.

ANSORGE, W., SPROAT, B. S., STEGEMANN, J. & SCHWAGER, C. 1986. A non-radioactive automated method for DNA sequence determination. *J Biochem Biophys Methods,* 13**,** 315-23.

ANSORGE, W. J. 2009. Next-generation DNA sequencing techniques. *New Biotechnology,* 25**,** 195-203.

ASHKENAZI, R., GENTRY, S. N. & JACKSON, T. L. 2008. Pathways to Tumorigenesis—Modeling Mutation Acquisition in Stem Cells and Their Progeny. *Neoplasia,* 10**,** 1170-82.

AZOURY, S. C. & LANGE, J. R. 2014. Epidemiology, Risk Factors, Prevention, and Early Detection of Melanoma. *Surgical Clinics of North America,* 94**,** 945-962.

BANNO, K., KISU, I., YANOKURA, M., TSUJI, K., MASUDA, K., UEKI, A., KOBAYASHI, Y., YAMAGAMI, W., NOMURA, H., TOMINAGA, E., SUSUMU,

N. & AOKI, D. 2012. Epimutation and cancer: a new carcinogenic mechanism of Lynch syndrome (Review). *International journal of oncology,* 41**,** 793-797.

BARBA, M., CZOSNEK, H. & HADIDI, A. 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses,* 6**,** 106-36.

BARNARD, M. E., BOEKE, C. E. & TAMIMI, R. M. 2015. Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochim Biophys Acta,* 1856**,** 73-85.

BARNETT, D. W., GARRISON, E. K., QUINLAN, A. R., STROMBERG, M. P. & MARTH, G. T. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics,* 27**,** 1691-2.

BASTIEN, R. R. L., RODRÍGUEZ-LESCURE, Á., EBBERT, M. T. W., PRAT, A., MUNÁRRIZ, B., ROWE, L., MILLER, P., RUIZ-BORREGO, M., ANDERSON, D., LYONS, B., ÁLVAREZ, I., DOWELL, T., WALL, D., SEGUÍ, M., BARLEY, L., BOUCHER, K. M., ALBA, E., PAPPAS, L., DAVIS, C. A., ARANDA, I., FAURON, C., STIJLEMAN, I. J., PALACIOS, J., ANTÓN, A., CARRASCO, E., CABALLERO, R., ELLIS, M. J., NIELSEN, T. O., PEROU, C. M., ASTILL, M., BERNARD, P. S. & MARTÍN, M. 2012. PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. *BMC Med Genomics,* 5**,** 44.

BAUBEC, T. & AKALIN, A. 2016. Genome-Wide Analysis of DNA Methylation Patterns by High-Throughput Sequencing. *In:* ARANSAY, A. M. & LAVÍN TRUEBA, J. L. (eds.) *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing.* Cham: Springer International Publishing.

BELLAOUSOV, S., REUTER, J. S., SEETIN, M. G. & MATHEWS, D. H. 2013. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res,* 41**,** W471-4.

BEMIS, L. T., CHEN, R., AMATO, C. M., CLASSEN, E. H., ROBINSON, S. E., COFFEY, D. G., ERICKSON, P. F., SHELLMAN, Y. G. & ROBINSON, W. A. 2008. MicroRNA-137 targets microphthalmia-associated transcription factor in melanoma cell lines. *Cancer Res,* 68**,** 1362-8.

BETEL, D., KOPPAL, A., AGIUS, P., SANDER, C. & LESLIE, C. 2010. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol,* 11**,** R90.

BETEL, D., WILSON, M., GABOW, A., MARKS, D. S. & SANDER, C. 2008. The microRNA.org resource: targets and expression. *Nucleic Acids Res,* 36**,** D149-53.

BHASKARAN, M. & MOHAN, M. 2014. MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. *Veterinary pathology,* 51**,** 759-774.

BIRD, A., TAGGART, M., FROMMER, M., MILLER, O. J. & MACLEOD, D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell,* 40**,** 91-99.

BOCK, C., BEERMAN, I., LIEN, W.-H., SMITH, ZACHARY D., GU, H., BOYLE, P., GNIRKE, A., FUCHS, E., ROSSI, DERRICK J. & MEISSNER, A. 2012. DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Molecular Cell,* 47**,** 633-647.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

BONIOL, M., AUTIER, P., BOYLE, P. & GANDINI, S. 2012. Cutaneous melanoma attributable to sunbed use: systematic review and meta-analysis. *BMJ (Clinical research ed.),* 345**,** e4757-e4757.

BORG, A., SANDBERG, T., NILSSON, K., JOHANNSSON, O., KLINKER, M., MASBACK, A., WESTERDAHL, J., OLSSON, H. & INGVAR, C. 2000. High

frequency of multiple melanomas and breast and pancreas carcinomas in CDKN2A mutation-positive melanoma families. *J. Natl. Cancer Inst.,* 92**,** 1260-6.

BORGES, H. L., BIRD, J., WASSON, K., CARDIFF, R. D., VARKI, N., ECKMANN, L. & WANG, J. Y. J. 2005. Tumor promotion by caspase-resistant retinoblastoma protein. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 15587-15592.

BRADFORD, P. T., GOLDSTEIN, A. M., MCMASTER, M. L. & TUCKER, M. A. 2009. Acral Lentiginous Melanoma: Incidence and Survival Patterns in the United States, 1986-2005. *Arch Dermatol,* 145**,** 427-34.

BRAY, F., FERLAY, J., SOERJOMATARAM, I., SIEGEL, R. L., TORRE, L. A. & JEMAL, A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin,* 68**,** 394-424.

BRAY, F., REN, J. S., MASUYER, E. & FERLAY, J. 2013. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer,* 132**,** 1133-45.

BRESLOW, A. 1970. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Annals of surgery,* 172**,** 902-908.

BRITTAIN, H. K., SCOTT, R. & THOMAS, E. 2017. The rise of the genome and personalised medicine. *Clinical medicine (London, England),* 17**,** 545-551.

BUCHBINDER, E. I. & DESAI, A. 2016. CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. *American journal of clinical oncology,* 39**,** 98-106.

BURKHART, D. L. & SAGE, J. 2008. Cellular mechanisms of tumor suppression by the retinoblastoma gene. *Nature Reviews Cancer,* 8**,** 671.

BUSCH, C., GEISLER, J., LILLEHAUG, J. R. & LONNING, P. E. 2010. MGMT expression levels predict disease stabilisation, progression-free and overall survival in patients with advanced melanomas treated with DTIC. *Eur J Cancer,* 46**,** 2127-33.

BYAKODI, R., BYAKODI, S., HIREMATH, S., BYAKODI, J., ADAKI, S., MARATHE, K. & MAHIND, P. 2012. Oral cancer in India: an epidemiologic and clinical review. *J Community Health,* 37**,** 316-9.

CALIN, G. A. & CROCE, C. M. 2006. MicroRNA signatures in human cancers. *Nat Rev Cancer,* 6**,** 857-66.

CALIN, G. A., DUMITRU, C. D., SHIMIZU, M., BICHI, R., ZUPO, S., NOCH, E., ALDLER, H., RATTAN, S., KEATING, M., RAI, K., RASSENTI, L., KIPPS, T., NEGRINI, M., BULLRICH, F. & CROCE, C. M. 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A,* 99**,** 15524-9.

CANCER GENOME ATLAS, N. 2015. Genomic Classification of Cutaneous Melanoma. *Cell,* 161**,** 1681-96.

CARAMUTA, S., EGYHAZI, S., RODOLFO, M., WITTEN, D., HANSSON, J., LARSSON, C. & LUI, W. O. 2010. MicroRNA expression profiles associated with mutational status and survival in malignant melanoma. *J Invest Dermatol,* 130**,** 2062-70.

CARDOSO, F., VAN'T VEER, L. J., BOGAERTS, J., SLAETS, L., VIALE, G., DELALOGE, S., PIERGA, J.-Y., BRAIN, E., CAUSERET, S., DELORENZI, M., GLAS, A. M., GOLFINOPOULOS, V., GOULIOTI, T., KNOX, S., MATOS, E., MEULEMANS, B., NEIJENHUIS, P. A., NITZ, U., PASSALACQUA, R., RAVDIN, P., RUBIO, I. T., SAGHATCHIAN, M., SMILDE, T. J., SOTIRIOU, C., STORK, L., STRAEHLE, C., THOMAS, G., THOMPSON, A. M., VAN DER HOEVEN, J. M., VUYLSTEKE, P., BERNARDS, R., TRYFONIDIS, K.,

RUTGERS, E. & PICCART, M. 2016. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine,* 375**,** 717-729.

CASTILLA-LLORENTE, V., SPRAGGON, L., OKAMURA, M., NASEERUDDIN, S., ADAMOW, M., QAMAR, S. & LIU, J. 2012. Mammalian GW220/TNGW1 is essential for the formation of GW/P bodies containing miRISC. *J Cell Biol,* 198**,** 529-44.

CAVENEE, W. K., HANSEN, M. F., NORDENSKJOLD, M., KOCK, E., MAUMENEE, I., SQUIRE, J. A., PHILLIPS, R. A. & GALLIE, B. L. 1985. Genetic origin of mutations predisposing to retinoblastoma. *Science,* 228**,** 501.

CEDAR, H. & BERGMAN, Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet,* 10**,** 295-304.

CHABNER, B. A. & ROBERTS, T. G. 2005. Chemotherapy and the war on cancer. *Nat Rev Cancer,* 5**,** 65-72.

CHAMBERLAIN, A. J., FRITSCHI, L. & KELLY, J. W. 2003. Nodular melanoma: patients' perceptions of presenting features and implications for earlier detection. *J Am Acad Dermatol,* 48**,** 694-701.

CHANG, A. E., KARNELL, L. H. & MENCK, H. R. 1998. The National Cancer Data Base report on cutaneous and noncutaneous melanoma: a summary of 84,836 cases from the past decade. The American College of Surgeons Commission on Cancer and the American Cancer Society. *Cancer,* 83**,** 1664-78.

CHANG, J. T. & NEVINS, J. R. 2006. GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics,* 22**,** 2926-33.

CHEANG, M. C. U., CHIA, S. K., VODUC, D., GAO, D., LEUNG, S., SNIDER, J., WATSON, M., DAVIES, S., BERNARD, P. S., PARKER, J. S., PEROU, C. M., ELLIS, M. J. & NIELSEN, T. O. 2009. Ki67 Index, HER2 Status, and Prognosis of Patients With Luminal B Breast Cancer. *Journal of the National Cancer Institute,* 101**,** 736-750.

CHELOUFI, S., DOS SANTOS, C. O., CHONG, M. M. & HANNON, G. J. 2010. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature,* 465**,** 584-9.

CHEN, L., HEIKKINEN, L., WANG, C., YANG, Y., SUN, H. & WONG, G. 2018. Trends in the development of miRNA bioinformatics tools. *Briefings in Bioinformatics***,** bby054-bby054.

CHENDRIMADA, T. P., GREGORY, R. I., KUMARASWAMY, E., NORMAN, J., COOCH, N., NISHIKURA, K. & SHIEKHATTAR, R. 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature,* 436**,** 740-4.

CHO, Y. R. & CHIANG, M. P. 2010. Epidemiology, Staging (New System), and Prognosis of Cutaneous Melanoma. *Clinics in Plastic Surgery,* 37**,** 47-53.

CHRISANTHAR, R., KNAPPSKOG, S., LOKKEVIK, E., ANKER, G., OSTENSTAD, B., LUNDGREN, S., BERGE, E. O., RISBERG, T., MJAALAND, I., MAEHLE, L., ENGEBRETSEN, L. F., LILLEHAUG, J. R. & LONNING, P. E. 2008. CHEK2 mutations affecting kinase activity together with mutations in TP53 indicate a functional pathway associated with resistance to epirubicin in primary breast cancer. *PLoS One,* 3**,** e3062.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. & GETZ, G. 2013a. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology,* 31**,** 213.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. & GETZ, G. 2013b. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

CITRON, M. L., BERRY, D. A., CIRRINCIONE, C., HUDIS, C., WINER, E. P., GRADISHAR, W. J., DAVIDSON, N. E., MARTINO, S., LIVINGSTON, R., INGLE, J. N., PEREZ, E. A., CARPENTER, J., HURD, D., HOLLAND, J. F., SMITH, B. L., SARTOR, C. I., LEUNG, E. H., ABRAMS, J., SCHILSKY, R. L., MUSS, H. B. & NORTON, L. 2003. Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: first report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. *J Clin Oncol,* 21**,** 1431-9.

CLARK, S. J. & MELKI, J. 2002. DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene,* 21**,** 5380.

CLARK, W. H., JR., AINSWORTH, A. M., BERNARDINO, E. A., YANG, C. H., MIHM, C. M., JR. & REED, R. J. 1975. The developmental biology of primary human malignant melanomas. *Semin Oncol,* 2**,** 83-103.

CLARK, W. H., JR., ELDER, D. E. & VAN HORN, M. 1986. The biologic forms of malignant melanoma. *Hum Pathol,* 17**,** 443-50.

CLARK, W. H., JR., FROM, L., BERNARDINO, E. A. & MIHM, M. C. 1969. The histogenesis and biologic behavior of primary human malignant melanomas of the skin. *Cancer Res,* 29**,** 705-27.

COLLINS, A. & POLITOPOULOS, I. 2011. The genetics of breast cancer: risk factors for disease. *The application of clinical genetics,* 4**,** 11-19.

COOPER, G. 1992. *Elements of human cancer*, Boston: Jones and Bartlett Publishers.

CORCORAN, C., FRIEL, A. M., DUFFY, M. J., CROWN, J. & DRISCOLL, L. 2011. Intracellular and Extracellular MicroRNAs in Breast Cancer. *Clinical Chemistry,* 57**,** 18.

CORMIER, J. N., XING, Y., DING, M., LEE, J. E., MANSFIELD, P. F., GERSHENWALD, J. E., ROSS, M. I. & DU, X. L. 2006. Ethnic differences among patients with cutaneous melanoma. *Arch Intern Med,* 166**,** 1907-14.

COSTINEAN, S., ZANESI, N., PEKARSKY, Y., TILI, E., VOLINIA, S., HEEREMA, N. & CROCE, C. M. 2006. Pre-B cell proliferation and lymphoblastic leukemia/high-grade lymphoma in E(mu)-miR155 transgenic mice. *Proc Natl Acad Sci U S A,* 103**,** 7024-9.

COUGOT, N., BABAJKO, S. & SERAPHIN, B. 2004. Cytoplasmic foci are sites of mRNA decay in human cells. *J Cell Biol,* 165**,** 31-40.

DAVIES, C., GODWIN, J., GRAY, R., CLARKE, M., CUTTER, D., DARBY, S., MCGALE, P., PAN, H. C., TAYLOR, C., WANG, Y. C., DOWSETT, M., INGLE, J. & PETO, R. 2011. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet,* 378**,** 771-84.

DAVIES, H., BIGNELL, G. R., COX, C., STEPHENS, P., EDKINS, S., CLEGG, S., TEAGUE, J., WOFFENDIN, H., GARNETT, M. J., BOTTOMLEY, W., DAVIS, N., DICKS, E., EWING, R., FLOYD, Y., GRAY, K., HALL, S., HAWES, R., HUGHES, J., KOSMIDOU, V., MENZIES, A., MOULD, C., PARKER, A., STEVENS, C., WATT, S., HOOPER, S., WILSON, R., JAYATILAKE, H., GUSTERSON, B. A., COOPER, C., SHIPLEY, J., HARGRAVE, D., PRITCHARD-JONES, K., MAITLAND, N., CHENEVIX-TRENCH, G., RIGGINS, G. J.,

BIGNER, D. D., PALMIERI, G., COSSU, A., FLANAGAN, A., NICHOLSON, A., HO, J. W., LEUNG, S. Y., YUEN, S. T., WEBER, B. L., SEIGLER, H. F., DARROW, T. L., PATERSON, H., MARAIS, R., MARSHALL, C. J., WOOSTER, R., STRATTON, M. R. & FUTREAL, P. A. 2002. Mutations of the BRAF gene in human cancer. *Nature,* 417**,** 949-54.

DAVIS, L. E., SHALIN, S. C. & TACKETT, A. J. 2019. Current state of melanoma diagnosis and treatment. *Cancer biology & therapy,* 20**,** 1366-1379.

DE JONG, M. M., NOLTE, I. M., TE MEERMAN, G. J., VAN DER GRAAF, W. T. A., OOSTERWIJK, J. C., KLEIBEUKER, J. H., SCHAAPVELD, M. & DE VRIES, E. G. E. 2002. Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. *Journal of Medical Genetics,* 39**,** 225.

DEATON, A. M. & BIRD, A. 2011. CpG islands and the regulation of transcription. *Genes & Development,* 25**,** 1010-1022.

DENLI, A. M., TOPS, B. B., PLASTERK, R. H., KETTING, R. F. & HANNON, G. J. 2004. Processing of primary microRNAs by the Microprocessor complex. *Nature,* 432**,** 231-5.

DEVITA, V. T. & CHU, E. 2008. A History of Cancer Chemotherapy. *Cancer Research,* 68**,** 8643-8653.

DHOMEN, N. & MARAIS, R. 2007. New insight into BRAF mutations in cancer. *Curr Opin Genet Dev,* 17**,** 31-9.

DI LEVA, G., GAROFALO, M. & CROCE, C. M. 2014. MicroRNAs in cancer. *Annual review of pathology,* 9**,** 287-314.

DICK, J. E. 2008. Stem cell concepts renew cancer research. *Blood,* 112**,** 4793-807.

DUMALAON-CANARIA, J. A., HUTCHINSON, A. D., PRICHARD, I. & WILSON, C. 2014. What causes breast cancer? A systematic review of causal attributions among breast cancer survivors and how these compare to expert-endorsed risk factors. *Cancer Causes & Control,* 25**,** 771-785.

DUTTA, S. W., SHOWALTER, S. L., SHOWALTER, T. N., LIBBY, B. & TRIFILETTI, D. M. 2017. Intraoperative radiation therapy for breast cancer patients: current perspectives. *Breast Cancer (Dove Med Press),* 9**,** 257-263.

EBERT, M. S. & SHARP, P. A. 2012. Roles for microRNAs in conferring robustness to biological processes. *Cell,* 149**,** 515-24.

EL-OSTA, A. 2004. Understanding the consequences of epigenetic mechanisms and its effects on transcription in health and disease. *Cancer Biol Ther,* 3**,** 816-8.

EMENS, L. A. 2018. Breast Cancer Immunotherapy: Facts and Hopes. *Clin Cancer Res,* 24**,** 511-520.

ENRIGHT, A. J., JOHN, B., GAUL, U., TUSCHL, T., SANDER, C. & MARKS, D. S. 2003. MicroRNA targets in Drosophila. *Genome Biol,* 5**,** R1.

ERDEI, E. & TORRES, S. M. 2010. A new understanding in the epidemiology of melanoma. *Expert review of anticancer therapy,* 10**,** 1811-1823.

ERDMANN, F., LORTET-TIEULENT, J., SCHÜZ, J., ZEEB, H., GREINERT, R., BREITBART, E. W. & BRAY, F. 2013. International trends in the incidence of malignant melanoma 1953–2008—are recent generations at higher or lower risk? *International Journal of Cancer,* 132**,** 385-400.

ESTELLER, M. 2002. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene,* 21**,** 5427.

ESTELLER, M. 2006. Epigenetics provides a new generation of oncogenes and tumor-suppressor genes. *British journal of cancer,* 94**,** 179-183.

ESTELLER, M., SILVA, J. M., DOMINGUEZ, G., BONILLA, F., MATIAS-GUIU, X., LERMA, E., BUSSAGLIA, E., PRAT, J., HARKES, I. C., REPASKY, E. A., GABRIELSON, E., SCHUTTE, M., BAYLIN, S. B. & HERMAN, J. G. 2000. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst,* 92**,** 564-9.

EWING, B. & GREEN, P. 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research,* 8**,** 186-194.

EWING, B., HILLIER, L., WENDL, M. C. & GREEN, P. 1998. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research,* 8**,** 175-185.

FAHRNER, J. A., EGUCHI, S., HERMAN, J. G. & BAYLIN, S. B. 2002. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer Res,* 62**,** 7213-8.

FARAZI, T. A., HOELL, J. I., MOROZOV, P. & TUSCHL, T. 2013. MicroRNAs in human cancer. *Advances in experimental medicine and biology,* 774**,** 1-20.

FELICETTI, F., ERRICO, M. C., SEGNALINI, P., MATTIA, G. & CARE, A. 2008. MicroRNA-221 and -222 pathway controls melanoma progression. *Expert Rev Anticancer Ther,* 8**,** 1759-65.

FERLAY, J., COLOMBET, M., SOERJOMATARAM, I., MATHERS, C., PARKIN, D. M., PINEROS, M., ZNAOR, A. & BRAY, F. 2018. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*.

FERLAY J, E. M., LAM F, COLOMBET M, MERY L, PIÑEROS M, ZNAOR A, SOERJOMATARAM I, BRAY F. 2018. *Global Cancer Observatory: Cancer Today* [Online]. Lyon, France: International Agency for Research on Cancer. Available: https://gco.iarc.fr/today/ [Accessed Dec 26, 2018 2018].

FERLAY J, S. I., ERVIK M, DIKSHIT R, ESER S, MATHERS C, REBELO M, PARKIN DM, FORMAN D, BRAY, F 2013. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]*. [Online]. Lyon, France: International Agency for Research on Cancer. Available: http://globocan.iarc.fr [Accessed].

FERLAY, J., SOERJOMATARAM, I., DIKSHIT, R., ESER, S., MATHERS, C., REBELO, M., PARKIN, D. M., FORMAN, D. & BRAY, F. 2015. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer,* 136**,** E359-E386.

FIDLER, I. J. 2003. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat Rev Cancer,* 3**,** 453-458.

FORBES, S. A., BEARE, D., BOUTSELAKIS, H., BAMFORD, S., BINDAL, N., TATE, J., COLE, C. G., WARD, S., DAWSON, E., PONTING, L., STEFANCSIK, R., HARSHA, B., KOK, C. Y., JIA, M., JUBB, H., SONDKA, Z., THOMPSON, S., DE, T. & CAMPBELL, P. J. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research,* 45**,** D777-D783.

FRANKLIN, R. E. & GOSLING, R. G. 1953. Molecular Configuration in Sodium Thymonucleate. *Nature,* 171**,** 740-741.

FRIEDLANDER, M. R., CHEN, W., ADAMIDI, C., MAASKOLA, J., EINSPANIER, R., KNESPEL, S. & RAJEWSKY, N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol,* 26**,** 407-15.

FRIEDLÄNDER, M. R., MACKOWIAK, S. D., LI, N., CHEN, W. & RAJEWSKY, N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res,* 40**,** 37-52.

FRITZ, A., PERCY, C., JACK, A., SHANMUGARATNAM, K., SOBIN, L. H., PARKIN, D. M., WHELAN, S. L. & WORLD HEALTH, O. 2000. International classification of diseases for oncology / editors, April Fritz ... [et al.]. 3rd ed ed. Geneva: World Health Organization.

GAYON, J. 2016. From Mendel to epigenetics: History of genetics. *C R Biol,* 339**,** 225-30.

GAZZOLI, I., LODA, M., GARBER, J., SYNGAL, S. & KOLODNER, R. D. 2002. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. *Cancer Res,* 62**,** 3925-8.

GOH, J. N., LOO, S. Y., DATTA, A., SIVEEN, K. S., YAP, W. N., CAI, W., SHIN, E. M., WANG, C., KIM, J. E., CHAN, M., DHARMARAJAN, A. M., LEE, A. S. G., LOBIE, P. E., YAP, C. T. & KUMAR, A. P. 2016. microRNAs in breast cancer: regulatory roles governing the hallmarks of cancer. *Biological Reviews,* 91**,** 409-428.

GOLDSTEIN, A. M., CHAN, M., HARLAND, M., GILLANDERS, E. M., HAYWARD, N. K., AVRIL, M.-F., AZIZI, E., BIANCHI-SCARRA, G., BISHOP, D. T., BRESSAC-DE PAILLERETS, B., BRUNO, W., CALISTA, D., CANNON ALBRIGHT, L. A., DEMENAIS, F., ELDER, D. E., GHIORZO, P., GRUIS, N. A., HANSSON, J., HOGG, D., HOLLAND, E. A., KANETSKY, P. A., KEFFORD, R. F., LANDI, M. T., LANG, J., LEACHMAN, S. A., MACKIE, R. M., MAGNUSSON, V., MANN, G. J., NIENDORF, K., NEWTON BISHOP, J., PALMER, J. M., PUIG, S., PUIG-BUTILLE, J. A., DE SNOO, F. A., STARK, M., TSAO, H., TUCKER, M. A., WHITAKER, L. & YAKOBSON, E. 2006. High-risk Melanoma Susceptibility Genes and Pancreatic Cancer, Neural System Tumors, and Uveal Melanoma across GenoMEL. *Cancer Research,* 66**,** 9818-9828.

GONZALEZ, K. D., NOLTNER, K. A., BUZIN, C. H., GU, D., WEN-FONG, C. Y., NGUYEN, V. Q., HAN, J. H., LOWSTUTER, K., LONGMATE, J., SOMMER, S. S. & WEITZEL, J. N. 2009. Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *J Clin Oncol,* 27**,** 1250-6.

GREGER, V., PASSARGE, E., HOPPING, W., MESSMER, E. & HORSTHEMKE, B. 1989. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum Genet,* 83**,** 155-8.

GREGORY, R. I., CHENDRIMADA, T. P., COOCH, N. & SHIEKHATTAR, R. 2005. Human RISC Couples MicroRNA Biogenesis and Posttranscriptional Gene Silencing. *Cell,* 123**,** 631-640.

GRIFFITHS-JONES, S., GROCOCK, R. J., VAN DONGEN, S., BATEMAN, A. & ENRIGHT, A. J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res,* 34**,** D140-4.

GRIFFITHS-JONES, S., SAINI, H. K., VAN DONGEN, S. & ENRIGHT, A. J. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res,* 36**,** D154-8.

GÜLLÜ, G. 2015. Clinical significance of miR-140-5p and miR-193b expression in patients. 38**,** 21-9.

GUO, W., PYLAYEVA, Y., PEPE, A., YOSHIOKA, T., MULLER, W. J., INGHIRAMI, G. & GIANCOTTI, F. G. 2006. Beta 4 integrin amplifies ErbB2 signaling to promote mammary tumorigenesis. *Cell,* 126**,** 489-502.

HALSTED, W. S. 1894. I. The Results of Operations for the Cure of Cancer of the Breast Performed at the Johns Hopkins Hospital from June, 1889, to January, 1894. *Annals of Surgery,* 20**,** 497-555.

HAMMOND, S. M. 2015. An overview of microRNAs. *Advanced drug delivery reviews,* 87**,** 3-14.

HAN, J., LEE, Y., YEOM, K.-H., KIM, Y.-K., JIN, H. & KIM, V. N. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development,* 18**,** 3016-3027.

HANAHAN, D. & WEINBERG, R. A. 2000. The Hallmarks of Cancer. *Cell,* 100**,** 57-70.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

HARLAND, M., MELONI, R., GRUIS, N., PINNEY, E., BROOKES, S., SPURR, N. K., FRISCHAUF, A. M., BATAILLE, V., PETERS, G., CUZICK, J., SELBY, P., BISHOP, D. T. & BISHOP, J. N. 1997. Germline mutations of the CDKN2 gene in UK melanoma families. *Hum. Mol. Genet.,* 6**,** 2061-7.

HAWRYLUK, E. B. & TSAO, H. 2014. Melanoma: Clinical Features and Genomic Insights. *Cold Spring Harbor Perspectives in Medicine,* 4**,** a015388.

HEATHER, J. M. & CHAIN, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107**,** 1-8.

HELLE, S. I., EKSE, D., HOLLY, J. M. P. & LØNNING, P. E. 2002. The IGF-system in healthy pre- and postmenopausal women: relations to demographic variables and sex-steroids. *The Journal of Steroid Biochemistry and Molecular Biology,* 81**,** 95-102.

HEPPNER, G. H. 1984. Tumor Heterogeneity. *Cancer Research,* 44**,** 2259.

HERMAN, J. G., CIVIN, C. I., ISSA, J.-P. J., COLLECTOR, M. I., SHARKIS, S. J. & BAYLIN, S. B. 1997. Distinct Patterns of Inactivation of p15^{INK4B} and p16^{INK4A} Characterize the Major Types of Hematological Malignancies. *Cancer Research,* 57**,** 837.

HERMAN, J. G., MERLO, A., MAO, L., LAPIDUS, R. G., ISSA, J. P., DAVIDSON, N. E., SIDRANSKY, D. & BAYLIN, S. B. 1995. Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res,* 55**,** 4525-30.

HERMAN, J. G., UMAR, A., POLYAK, K., GRAFF, J. R., AHUJA, N., ISSA, J. P., MARKOWITZ, S., WILLSON, J. K., HAMILTON, S. R., KINZLER, K. W., KANE, M. F., KOLODNER, R. D., VOGELSTEIN, B., KUNKEL, T. A. & BAYLIN, S. B. 1998. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci U S A,* 95**,** 6870-5.

HERRANZ, M. & ESTELLER, M. 2007. DNA Methylation and Histone Modifications in Patients With Cancer. *In:* SIOUD, M. (ed.) *Target Discovery and Validation Reviews and Protocols: Volume 2: Emerging Molecular Targets and Treatment Options.* Totowa, NJ: Humana Press.

HINSKE, L. C., FRANCA, G. S., TORRES, H. A., OHARA, D. T., LOPES-RAMOS, C. M., HEYN, J., REIS, L. F., OHNO-MACHADO, L., KRETH, S. & GALANTE, P. A. 2014. miRIAD-integrating microRNA inter- and intragenic data. *Database (Oxford),* 2014.

HITCHINS, M. P., WONG, J. J., SUTHERS, G., SUTER, C. M., MARTIN, D. I., HAWKINS, N. J. & WARD, R. L. 2007. Inheritance of a cancer-associated MLH1 germ-line epimutation. *N Engl J Med,* 356**,** 697-705.

HOBERT, O. 2006. Architecture of a microRNA-controlled gene regulatory network that diversifies neuronal cell fates. *Cold Spring Harb Symp Quant Biol,* 71**,** 181-8.

HODIS, E., WATSON, I. R., KRYUKOV, G. V., AROLD, S. T., IMIELINSKI, M., THEURILLAT, J. P., NICKERSON, E., AUCLAIR, D., LI, L., PLACE, C., DICARA, D., RAMOS, A. H., LAWRENCE, M. S., CIBULSKIS, K., SIVACHENKO, A., VOET, D., SAKSENA, G., STRANSKY, N., ONOFRIO, R. C., WINCKLER, W., ARDLIE, K., WAGLE, N., WARGO, J., CHONG, K., MORTON,

D. L., STEMKE-HALE, K., CHEN, G., NOBLE, M., MEYERSON, M., LADBURY, J. E., DAVIES, M. A., GERSHENWALD, J. E., WAGNER, S. N., HOON, D. S., SCHADENDORF, D., LANDER, E. S., GABRIEL, S. B., GETZ, G., GARRAWAY, L. A. & CHIN, L. 2012. A landscape of driver mutations in melanoma. *Cell,* 150**,** 251-63.

HOFVIND, S., URSIN, G., TRETLI, S., SEBUØDEGÅRD, S. & MØLLER, B. 2013. Breast cancer mortality in participants of the Norwegian Breast Cancer Screening Program. *Cancer,* 119**,** 3106-3112.

HOLLEY, R. W., APGAR, J., EVERETT, G. A., MADISON, J. T., MARQUISEE, M., MERRILL, S. H., PENSWICK, J. R. & ZAMIR, A. 1965. STRUCTURE OF A RIBONUCLEIC ACID. *Science,* 147**,** 1462-5.

HOLLEY, R. W., MADISON, J. T. & ZAMIR, A. 1964. A new method for sequence determination of large oligonucleotides. *Biochemical and Biophysical Research Communications,* 17**,** 389-394.

HOLLIDAY, R. 1987. The inheritance of epigenetic defects. *Science,* 238**,** 163-70.

HOLLIDAY, R. & PUGH, J. E. 1975. DNA modification mechanisms and gene activity during development. *Science,* 187**,** 226-32.

HORSTHEMKE, B. 2006. Epimutations in human disease. *Curr Top Microbiol Immunol,* 310**,** 45-59.

HORTOBAGYI, G. N. 1997. Anthracyclines in the treatment of cancer. An overview. *Drugs,* 54 Suppl 4**,** 1-7.

HOWELL, P. M., JR., LI, X., RIKER, A. I. & XI, Y. 2010. MicroRNA in Melanoma. *The Ochsner journal,* 10**,** 83-92.

HSU, J. Y., SUN, Z. W., LI, X., REUBEN, M., TATCHELL, K., BISHOP, D. K., GRUSHCOW, J. M., BRAME, C. J., CALDWELL, J. A., HUNT, D. F., LIN, R., SMITH, M. M. & ALLIS, C. D. 2000. Mitotic phosphorylation of histone H3 is governed by Ipl1/aurora kinase and Glc7/PP1 phosphatase in budding yeast and nematodes. *Cell,* 102**,** 279-91.

HU, Z., FAN, C., OH, D. S., MARRON, J., HE, X., QAQISH, B. F., LIVASY, C., CAREY, L. A., REYNOLDS, E., DRESSLER, L., NOBEL, A., PARKER, J., EWEND, M. G., SAWYER, L. R., WU, J., LIU, Y., NANDA, R., TRETIAKOVA, M., ORRICO, A. R., DREHER, D., PALAZZO, J. P., PERREARD, L., NELSON, E., MONE, M., HANSEN, H., MULLINS, M., QUACKENBUSH, J. F., ELLIS, M. J., OLOPADE, O. I., BERNARD, P. S. & PEROU, C. M. 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics,* 7**,** 96.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

HUNKAPILLER, T., KAISER, R. J., KOOP, B. F. & HOOD, L. 1991. Large-scale and automated DNA sequence determination. *Science,* 254**,** 59-67.

HUSSUSSIAN, C. J., STRUEWING, J. P., GOLDSTEIN, A. M., HIGGINS, P. A., ALLY, D. S., SHEAHAN, M. D., CLARK, W. H., JR., TUCKER, M. A. & DRACOPOLI, N. C. 1994. Germline p16 mutations in familial melanoma. *Nat. Genet.,* 8**,** 15-21.

HUTCHINSON, J. N., JIN, J., CARDIFF, R. D., WOODGETT, J. R. & MULLER, W. J. 2004. Activation of Akt-1 (PKB-alpha) can accelerate ErbB-2-mediated mammary tumorigenesis but suppresses tumor invasion. *Cancer Res,* 64**,** 3171-8.

HYMAN, E. D. 1988. A new method of sequencing DNA. *Anal Biochem,* 174**,** 423-36.

ILLINGWORTH, R. S., GRUENEWALD-SCHNEIDER, U., WEBB, S., KERR, A. R. W., JAMES, K. D., TURNER, D. J., SMITH, C., HARRISON, D. J., ANDREWS, R. &

BIRD, A. P. 2010. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLOS Genetics,* 6**,** e1001134.

IORIO, M. V., FERRACIN, M., LIU, C. G., VERONESE, A., SPIZZO, R., SABBIONI, S., MAGRI, E., PEDRIALI, M., FABBRI, M., CAMPIGLIO, M., MENARD, S., PALAZZO, J. P., ROSENBERG, A., MUSIANI, P., VOLINIA, S., NENCI, I., CALIN, G. A., QUERZOLI, P., NEGRINI, M. & CROCE, C. M. 2005. MicroRNA gene expression deregulation in human breast cancer. *Cancer Res,* 65**,** 7065-70.

IWAMOTO, T., YAMAMOTO, N., TAGUCHI, T., TAMAKI, Y. & NOGUCHI, S. 2011. BRCA1 promoter methylation in peripheral blood cells is associated with increased risk of breast cancer with BRCA1 promoter methylation. *Breast Cancer Res Treat,* 129**,** 69-77.

IWASAKI, M. & TSUGANE, S. 2011. Risk factors for breast cancer: epidemiological evidence from Japanese studies. *Cancer Science,* 102**,** 1607-1614.

JAFARI, M., PAPP, T., KIRCHNER, S., DIENER, U., HENSCHLER, D., BURG, G. & SCHIFFMANN, D. 1995. Analysis of ras mutations in human melanocytic lesions: activation of the ras gene seems to be associated with the nodular type of human malignant melanoma. *J Cancer Res Clin Oncol,* 121**,** 23-30.

JANAS, M. M., WANG, B., HARRIS, A. S., AGUIAR, M., SHAFFER, J. M., SUBRAHMANYAM, Y. V., BEHLKE, M. A., WUCHERPFENNIG, K. W., GYGI, S. P., GAGNON, E. & NOVINA, C. D. 2012. Alternative RISC assembly: binding and repression of microRNA-mRNA duplexes by human Ago proteins. *Rna,* 18**,** 2041-55.

JI, Q., HAO, X., ZHANG, M., TANG, W., YANG, M., LI, L., XIANG, D., DESANO, J. T., BOMMER, G. T., FAN, D., FEARON, E. R., LAWRENCE, T. S. & XU, L. 2009. MicroRNA miR-34 inhibits human pancreatic cancer tumor-initiating cells. *PLoS One,* 4**,** e6816.

JIA, H., TRUICA, C. I., WANG, B., WANG, Y., REN, X., HARVEY, H. A., SONG, J. & YANG, J. M. 2017. Immunotherapy for triple-negative breast cancer: Existing challenges and exciting prospects. *Drug Resist Updat,* 32**,** 1-15.

JOHN, B., ENRIGHT, A. J., ARAVIN, A., TUSCHL, T., SANDER, C. & MARKS, D. S. 2004. Human MicroRNA targets. *PLoS Biol,* 2**,** e363.

JONES, P. A. & BAYLIN, S. B. 2002. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics,* 3**,** 415.

JONSSON, G., BUSCH, C., KNAPPSKOG, S., GEISLER, J., MILETIC, H., RINGNER, M., LILLEHAUG, J. R., BORG, A. & LONNING, P. E. 2010. Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin Cancer Res,* 16**,** 3356-67.

JOSHI, S. C., KHAN, F. A., PANT, I. & SHUKLA, A. 2007. Role of radiotherapy in early breast cancer: an overview. *International journal of health sciences,* 1**,** 259-264.

KAMB, A., SHATTUCK-EIDENS, D., EELES, R., LIU, Q., GRUIS, N. A., DING, W., HUSSEY, C., TRAN, T., MIKI, Y., WEAVER-FELDHAUS, J. & ET AL. 1994. Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat Genet,* 8**,** 23-6.

KAMIŃSKA, M., CISZEWSKI, T., ŁOPACKA-SZATAN, K., MIOTŁA, P. & STAROSŁAWSKA, E. 2015. Breast cancer risk factors. *Menopause Review/Przegląd Menopauzalny,* 14**,** 196-202.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KANELLOPOULOU, C., MULJO, S. A., KUNG, A. L., GANESAN, S., DRAPKIN, R., JENUWEIN, T., LIVINGSTON, D. M. & RAJEWSKY, K. 2005. Dicer-deficient

mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev,* 19**,** 489-501.

KEOHANE, S. G., PROBY, C. M., NEWLANDS, C., MOTLEY, R. J., NASR, I., MOHD MUSTAPA, M. F. & SLATER, D. N. 2018. The new 8th edition of TNM staging and its implications for skin cancer: a review by the British Association of Dermatologists and the Royal College of Pathologists, U.K. *Br J Dermatol,* 179**,** 824-828.

KERN, S. E. & SHIBATA, D. 2007. The fuzzy math of solid tumor stem cells: a perspective. *Cancer Res,* 67**,** 8985-8.

KING, M. C., MARKS, J. H. & MANDELL, J. B. 2003. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science,* 302**,** 643-6.

KLOOSTERMAN, W. P. & PLASTERK, R. H. 2006. The diverse functions of microRNAs in animal development and disease. *Dev Cell,* 11**,** 441-50.

KNUDSON, A. G., JR. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

KOTA, J., CHIVUKULA, R. R., O'DONNELL, K. A., WENTZEL, E. A., MONTGOMERY, C. L., HWANG, H. W., CHANG, T. C., VIVEKANANDAN, P., TORBENSON, M., CLARK, K. R., MENDELL, J. R. & MENDELL, J. T. 2009. Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. *Cell,* 137**,** 1005-17.

KOZOMARA, A. & GRIFFITHS-JONES, S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res,* 39**,** D152-7.

KOZOMARA, A. & GRIFFITHS-JONES, S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res,* 42**,** D68-73.

KUNKEL, T. A. 1995. DNA-mismatch repair. The intricacies of eukaryotic spell-checking. *Curr Biol,* 5**,** 1091-4.

KUSABA, H., NAKAYAMA, M., HARADA, T., NOMOTO, M., KOHNO, K., KUWANO, M. & WADA, M. 1999. Association of 5' CpG demethylation and altered chromatin structure in the promoter region with transcriptional activation of the multidrug resistance 1 gene in human cancer cells. *Eur J Biochem,* 262**,** 924-32.

LACHNER, M., O'CARROLL, D., REA, S., MECHTLER, K. & JENUWEIN, T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature,* 410**,** 116-20.

LANE, D. P. 1992. Cancer. p53, guardian of the genome. *Nature,* 358**,** 15-6.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10**,** R25.

LARSEN, F., GUNDERSEN, G., LOPEZ, R. & PRYDZ, H. 1992. CpG islands as gene markers in the human genome. *Genomics,* 13**,** 1095-1107.

LAWRENCE, M. S., STOJANOV, P., MERMEL, C. H., ROBINSON, J. T., GARRAWAY, L. A., GOLUB, T. R., MEYERSON, M., GABRIEL, S. B., LANDER, E. S. & GETZ, G. 2014. Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature,* 505**,** 495-501.

LEE, E. Y. H. P. & MULLER, W. J. 2010. Oncogenes and tumor suppressor genes. *Cold Spring Harbor perspectives in biology,* 2**,** a003236-a003236.

LEE, R. C., FEINBAUM, R. L. & AMBROS, V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75**,** 843-54.

LEE, W. H., BOOKSTEIN, R., HONG, F., YOUNG, L. J., SHEW, J. Y. & LEE, E. Y. 1987. Human retinoblastoma susceptibility gene: cloning, identification, and sequence. *Science,* 235**,** 1394.

LEE, Y., AHN, C., HAN, J., CHOI, H., KIM, J., YIM, J., LEE, J., PROVOST, P., RADMARK, O., KIM, S. & KIM, V. N. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature,* 425**,** 415-9.

LEHMAN, T. A., REDDEL, R., PEIIFER, A. M., SPILLARE, E., KAIGHN, M. E., WESTON, A., GERWIN, B. I. & HARRIS, C. C. 1991. Oncogenes and tumor-suppressor genes. *Environmental health perspectives,* 93**,** 133-144.

LEWIS, B. P., BURGE, C. B. & BARTEL, D. P. 2005. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell,* 120**,** 15-20.

LEWIS, B. P., SHIH, I. H., JONES-RHOADES, M. W., BARTEL, D. P. & BURGE, C. B. 2003. Prediction of mammalian microRNA targets. *Cell,* 115**,** 787-98.

LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-93.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, M. A. & HE, L. 2012. microRNAs as novel regulators of stem cell pluripotency and somatic cell reprogramming. *Bioessays,* 34**,** 670-80.

LI, Y. & KOWDLEY, K. V. 2012. MicroRNAs in Common Human Diseases. *Genomics, Proteomics & Bioinformatics,* 10**,** 246-253.

LISTER, R., PELIZZOLA, M., DOWEN, R. H., HAWKINS, R. D., HON, G., TONTI-FILIPPINI, J., NERY, J. R., LEE, L., YE, Z., NGO, Q. M., EDSALL, L., ANTOSIEWICZ-BOURGET, J., STEWART, R., RUOTTI, V., MILLAR, A. H., THOMSON, J. A., REN, B. & ECKER, J. R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature,* 462**,** 315-22.

LIU, C. G., CALIN, G. A., MELOON, B., GAMLIEL, N., SEVIGNANI, C., FERRACIN, M., DUMITRU, C. D., SHIMIZU, M., ZUPO, S., DONO, M., ALDER, H., BULLRICH, F., NEGRINI, M. & CROCE, C. M. 2004. An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc Natl Acad Sci U S A,* 101**,** 9740-4.

LIU, Y., SIEGMUND, K. D., LAIRD, P. W. & BERMAN, B. P. 2012. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology,* 13**,** 1-14.

LOIBL, S. & GIANNI, L. 2017. HER2-positive breast cancer. *The Lancet,* 389**,** 2415-2429.

LONNING, P. E. 2003. Study of suboptimum treatment response: lessons from breast cancer. *Lancet Oncol,* 4**,** 177-85.

LONNING, P. E., BERGE, E. O., BJORNSLETT, M., MINSAAS, L., CHRISANTHAR, R., HOBERG-VETTI, H., DULARY, C., BUSATO, F., BJORNEKLETT, S., ERIKSEN, C., KOPPERUD, R., AXCRONA, U., DAVIDSON, B., BJORGE, L., EVANS, G., HOWELL, A., SALVESEN, H. B., JANSZKY, I., HVEEM, K., ROMUNDSTAD, P. R., VATTEN, L. J., TOST, J., DORUM, A. & KNAPPSKOG, S. 2018. White Blood Cell BRCA1 Promoter Methylation Status and Ovarian Cancer Risk. *Ann Intern Med,* 168**,** 326-334.

LONNING, P. E., HELLE, H., DUONG, N. K., EKSE, D., AAS, T. & GEISLER, J. 2009. Tissue estradiol is selectively elevated in receptor positive breast cancers while tumor estrone is reduced independent of receptor status. *J Steroid Biochem Mol Biol,* 117**,** 31-41.

LØNNING, P. E. & KNAPPSKOG, S. 2018. BRCA1 methylation in newborns: genetic disposition, maternal transfer, environmental influence, or by chance only? *Clinical epigenetics,* 10**,** 128-128.

LORENZ, R., BERNHART, S. H., HONER ZU SIEDERDISSEN, C., TAFER, H., FLAMM, C., STADLER, P. F. & HOFACKER, I. L. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol,* 6**,** 26.

LOWE, S. W. & SHERR, C. J. 2003. Tumor suppression by Ink4a-Arf: progress and puzzles. *Curr Opin Genet Dev,* 13**,** 77-83.

LOZANO, G. 2007. The oncogenic roles of p53 mutants in mouse models. *Curr Opin Genet Dev,* 17**,** 66-70.

LU, J., GETZ, G., MISKA, E. A., ALVAREZ-SAAVEDRA, E., LAMB, J., PECK, D., SWEET-CORDERO, A., EBERT, B. L., MAK, R. H., FERRANDO, A. A., DOWNING, J. R., JACKS, T., HORVITZ, H. R. & GOLUB, T. R. 2005. MicroRNA expression profiles classify human cancers. *Nature,* 435**,** 834-8.

LU, Y., THOMSON, J. M., WONG, H. Y., HAMMOND, S. M. & HOGAN, B. L. 2007. Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells. *Dev Biol,* 310**,** 442-53.

LUGER, K. & RICHMOND, T. J. 1998. DNA binding within the nucleosome core. *Current Opinion in Structural Biology,* 8**,** 33-40.

LUMACHI, F., LUISETTO, G., BASSO, S. M. M., BASSO, U., BRUNELLO, A. & CAMOZZI, V. 2011. Endocrine Therapy of Breast Cancer. *Current Medicinal Chemistry,* 18**,** 513-522.

LYNCH, H. T., MARCUS, J. N. & RUBINSTEIN, W. S. 2008. Stemming the tide of cancer for BRCA1/2 mutation carriers. *J Clin Oncol,* 26**,** 4239-43.

MA, X. J., WANG, Z., RYAN, P. D., ISAKOFF, S. J., BARMETTLER, A., FULLER, A., MUIR, B., MOHAPATRA, G., SALUNGA, R., TUGGLE, J. T., TRAN, Y., TRAN, D., TASSIN, A., AMON, P., WANG, W., WANG, W., ENRIGHT, E., STECKER, K., ESTEPA-SABAL, E., SMITH, B., YOUNGER, J., BALIS, U., MICHAELSON, J., BHAN, A., HABIN, K., BAER, T. M., BRUGGE, J., HABER, D. A., ERLANDER, M. G. & SGROI, D. C. 2004. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell,* 5**,** 607-16.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y. J., CHEN, Z., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L., JARVIE, T. P., JIRAGE, K. B., KIM, J. B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU, H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E., NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-80.

MARUSYK, A. & POLYAK, K. 2010. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta,* 1805**,** 105-17.

MATTICK, J. S. & MAKUNIN, I. V. 2005. Small regulatory RNAs in mammals. *Hum Mol Genet,* 14 Spec No 1**,** R121-32.

MCKERNAN, K. J., PECKHAM, H. E., COSTA, G. L., MCLAUGHLIN, S. F., FU, Y., TSUNG, E. F., CLOUSER, C. R., DUNCAN, C., ICHIKAWA, J. K., LEE, C. C., ZHANG, Z., RANADE, S. S., DIMALANTA, E. T., HYLAND, F. C., SOKOLSKY, T. D., ZHANG, L., SHERIDAN, A., FU, H., HENDRICKSON, C. L., LI, B., KOTLER, L., STUART, J. R., MALEK, J. A., MANNING, J. M., ANTIPOVA, A. A., PEREZ, D. S., MOORE, M. P., HAYASHIBARA, K. C., LYONS, M. R., BEAUDOIN, R. E., COLEMAN, B. E., LAPTEWICZ, M. W., SANNICANDRO, A. E., RHODES, M. D., GOTTIMUKKALA, R. K., YANG, S., BAFNA, V., BASHIR, A., MACBRIDE, A., ALKAN, C., KIDD, J. M., EICHLER, E. E., REESE, M. G., DE LA VEGA, F. M. & BLANCHARD, A. P. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res,* 19**,** 1527-41.

MEIENBERG, J., ZERJAVIC, K., KELLER, I., OKONIEWSKI, M., PATRIGNANI, A., LUDIN, K., XU, Z., STEINMANN, B., CARREL, T., RÖTHLISBERGER, B., SCHLAPBACH, R., BRUGGMANN, R. & MATYAS, G. 2015. New insights into the performance of human whole-exome capture platforms. *Nucleic acids research,* 43**,** e76-e76.

MELKI, J. R., VINCENT, P. C. & CLARK, S. J. 1999. Concurrent DNA Hypermethylation of Multiple Genes in Acute Myeloid Leukemia. *Cancer Research,* 59**,** 3730.

MELTON, C. & BLELLOCH, R. 2010. MicroRNA Regulation of Embryonic Stem Cell Self-Renewal and Differentiation. *Adv Exp Med Biol,* 695**,** 105-17.

MERLO, A., HERMAN, J. G., MAO, L., LEE, D. J., GABRIELSON, E., BURGER, P. C., BAYLIN, S. B. & SIDRANSKY, D. 1995. 5' CpG island methylation is associated with transcriptional silencing of the tumor suppressor p16/CDKN2/MTS1 in human cancers. *Nat Med,* 1**,** 686-92.

MICHAILIDOU, K., BEESLEY, J., LINDSTROM, S., CANISIUS, S., DENNIS, J., LUSH, M. J., MARANIAN, M. J., BOLLA, M. K., WANG, Q., SHAH, M., PERKINS, B. J., CZENE, K., ERIKSSON, M., DARABI, H., BRAND, J. S., BOJESEN, S. E., NORDESTGAARD, B. G., FLYGER, H., NIELSEN, S. F., RAHMAN, N., TURNBULL, C., BOCS, FLETCHER, O., PETO, J., GIBSON, L., DOS-SANTOS-SILVA, I., CHANG-CLAUDE, J., FLESCH-JANYS, D., RUDOLPH, A., EILBER, U., BEHRENS, S., NEVANLINNA, H., MURANEN, T. A., AITTOMÄKI, K., BLOMQVIST, C., KHAN, S., AALTONEN, K., AHSAN, H., KIBRIYA, M. G., WHITTEMORE, A. S., JOHN, E. M., MALONE, K. E., GAMMON, M. D., SANTELLA, R. M., URSIN, G., MAKALIC, E., SCHMIDT, D. F., CASEY, G., HUNTER, D. J., GAPSTUR, S. M., GAUDET, M. M., DIVER, W. R., HAIMAN, C. A., SCHUMACHER, F., HENDERSON, B. E., LE MARCHAND, L., BERG, C. D., CHANOCK, S. J., FIGUEROA, J., HOOVER, R. N., LAMBRECHTS, D., NEVEN, P., WILDIERS, H., VAN LIMBERGEN, E., SCHMIDT, M. K., BROEKS, A., VERHOEF, S., CORNELISSEN, S., COUCH, F. J., OLSON, J. E., HALLBERG, E., VACHON, C., WAISFISZ, Q., MEIJERS-HEIJBOER, H., ADANK, M. A., VAN DER LUIJT, R. B., LI, J., LIU, J., HUMPHREYS, K., KANG, D., CHOI, J.-Y., PARK, S. K., YOO, K.-Y., MATSUO, K., ITO, H., IWATA, H., TAJIMA, K., GUÉNEL, P., TRUONG, T., MULOT, C., SANCHEZ, M., BURWINKEL, B., MARME, F., SUROWY, H., SOHN, C., WU, A. H., TSENG, C.-C., VAN DEN BERG, D., STRAM, D. O., GONZÁLEZ-NEIRA, A., et al. 2015. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature Genetics,* 47**,** 373.

MIGLIORE, C., PETRELLI, A., GHISO, E., CORSO, S., CAPPARUCCIA, L., ERAMO, A., COMOGLIO, P. M. & GIORDANO, S. 2008. MicroRNAs impair MET-mediated invasive growth. *Cancer Res,* 68**,** 10128-36.

MILLAR, D. S., PAUL, C. L., MOLLOY, P. L. & CLARK, S. J. 2000. A distinct sequence (ATAAA)n separates methylated and unmethylated domains at the 5'-end of the GSTP1 CpG island. *J Biol Chem,* 275**,** 24893-9.

MILLIKAN, R. C., NEWMAN, B., TSE, C. K., MOORMAN, P. G., CONWAY, K., DRESSLER, L. G., SMITH, L. V., LABBOK, M. H., GERADTS, J., BENSEN, J. T., JACKSON, S., NYANTE, S., LIVASY, C., CAREY, L., EARP, H. S. & PEROU, C. M. 2008. Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat,* 109**,** 123-39.

MUELLER, D. W., REHLI, M. & BOSSERHOFF, A. K. 2009. miRNA expression profiling in melanocytes and melanoma cell lines reveals miRNAs associated with formation and progression of malignant melanoma. *J Invest Dermatol,* 129**,** 1740-51.

NAGARAJAN, R. P., FOUSE, S. D., BELL, R. J. A. & COSTELLO, J. F. 2013. Methods for cancer epigenome analysis. *Advances in experimental medicine and biology,* 754**,** 313-338.

NAGATA, C., HU, Y. H. & SHIMIZU, H. 1995. Effects of menstrual and reproductive factors on the risk of breast cancer: meta-analysis of the case-control studies in Japan. *Jpn J Cancer Res,* 86**,** 910-5.

NAKAGAWA, H. & FUJITA, M. 2018. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer science,* 109**,** 513-522.

NASSAR, F. J., NASR, R. & TALHOUK, R. 2017. MicroRNAs as biomarkers for early breast cancer diagnosis, prognosis and therapy prediction. *Pharmacology & Therapeutics,* 172**,** 34-49.

NAWROCKI, E. P., BURGE, S. W., BATEMAN, A., DAUB, J., EBERHARDT, R. Y., EDDY, S. R., FLODEN, E. W., GARDNER, P. P., JONES, T. A., TATE, J. & FINN, R. D. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res,* 43**,** D130-7.

NICHOLS, K. E., MALKIN, D., GARBER, J. E., FRAUMENI, J. F., JR. & LI, F. P. 2001. Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. *Cancer Epidemiol Biomarkers Prev,* 10**,** 83-7.

NIELSEN, T. O., PARKER, J. S., LEUNG, S., VODUC, D., EBBERT, M., VICKERY, T., DAVIES, S. R., SNIDER, J., STIJLEMAN, I. J., REED, J., CHEANG, M. C., MARDIS, E. R., PEROU, C. M., BERNARD, P. S. & ELLIS, M. J. 2010. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin Cancer Res,* 16**,** 5222-32.

NIK-ZAINAL, S., ALEXANDROV, L. B., WEDGE, D. C., VAN LOO, P., GREENMAN, C. D., RAINE, K., JONES, D., HINTON, J., MARSHALL, J., STEBBINGS, L. A., MENZIES, A., MARTIN, S., LEUNG, K., CHEN, L., LEROY, C., RAMAKRISHNA, M., RANCE, R., LAU, K. W., MUDIE, L. J., VARELA, I., MCBRIDE, D. J., BIGNELL, G. R., COOKE, S. L., SHLIEN, A., GAMBLE, J., WHITMORE, I., MADDISON, M., TARPEY, P. S., DAVIES, H. R., PAPAEMMANUIL, E., STEPHENS, P. J., MCLAREN, S., BUTLER, A. P., TEAGUE, J. W., JÖNSSON, G., GARBER, J. E., SILVER, D., MIRON, P., FATIMA, A., BOYAULT, S., LANGERØD, A., TUTT, A., MARTENS, J. W. M., APARICIO, S. A. J. R., BORG, Å., SALOMON, A. V., THOMAS, G., BØRRESEN-DALE, A.-L., RICHARDSON, A. L., NEUBERGER, M. S., FUTREAL, P. A., CAMPBELL, P. J., STRATTON, M. R. & BREAST CANCER

WORKING GROUP OF THE INTERNATIONAL CANCER GENOME, C. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell,* 149**,** 979-993.

NIK-ZAINAL, S., DAVIES, H., STAAF, J., RAMAKRISHNA, M., GLODZIK, D., ZOU, X., MARTINCORENA, I., ALEXANDROV, L. B., MARTIN, S., WEDGE, D. C., VAN LOO, P., JU, Y. S., SMID, M., BRINKMAN, A. B., MORGANELLA, S., AURE, M. R., LINGJÆRDE, O. C., LANGERØD, A., RINGNÉR, M., AHN, S.-M., BOYAULT, S., BROCK, J. E., BROEKS, A., BUTLER, A., DESMEDT, C., DIRIX, L., DRONOV, S., FATIMA, A., FOEKENS, J. A., GERSTUNG, M., HOOIJER, G. K. J., JANG, S. J., JONES, D. R., KIM, H.-Y., KING, T. A., KRISHNAMURTHY, S., LEE, H. J., LEE, J.-Y., LI, Y., MCLAREN, S., MENZIES, A., MUSTONEN, V., O'MEARA, S., PAUPORTÉ, I., PIVOT, X., PURDIE, C. A., RAINE, K., RAMAKRISHNAN, K., RODRÍGUEZ-GONZÁLEZ, F. G., ROMIEU, G., SIEUWERTS, A. M., SIMPSON, P. T., SHEPHERD, R., STEBBINGS, L., STEFANSSON, O. A., TEAGUE, J., TOMMASI, S., TREILLEUX, I., VAN DEN EYNDEN, G. G., VERMEULEN, P., VINCENT-SALOMON, A., YATES, L., CALDAS, C., VEER, L. V. T., TUTT, A., KNAPPSKOG, S., TAN, B. K. T., JONKERS, J., BORG, Å., UENO, N. T., SOTIRIOU, C., VIARI, A., FUTREAL, P. A., CAMPBELL, P. J., SPAN, P. N., VAN LAERE, S., LAKHANI, S. R., EYFJORD, J. E., THOMPSON, A. M., BIRNEY, E., STUNNENBERG, H. G., VAN DE VIJVER, M. J., MARTENS, J. W. M., BØRRESEN-DALE, A.-L., RICHARDSON, A. L., KONG, G., THOMAS, G. & STRATTON, M. R. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature,* 534**,** 47-54.

NOWELL, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-8.

NYREN, P. & LUNDIN, A. 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal Biochem,* 151**,** 504-9.

OEY, H. & WHITELAW, E. 2014. On the meaning of the word 'epimutation'. *Trends in Genetics,* 30**,** 519-520.

OHTANI-FUJITA, N., FUJITA, T., AOIKE, A., OSIFCHIN, N. E., ROBBINS, P. D. & SAKAI, T. 1993. CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene. *Oncogene,* 8**,** 1063-7.

OLDHAM, R. K. & DILLMAN, R. O. 2008. Monoclonal Antibodies in Cancer Therapy: 25 Years of Progress. *Journal of Clinical Oncology,* 26**,** 1774-1777.

OLIVERIA, S. A., SARAIYA, M., GELLER, A. C., HENEGHAN, M. K. & JORGENSEN, C. 2006. Sun exposure and risk of melanoma. *Archives of disease in childhood,* 91**,** 131-138.

OSBORNE, C., WILSON, P. & TRIPATHY, D. 2004. Oncogenes and Tumor Suppressor Genes in Breast Cancer: Potential Diagnostic and Therapeutic Applications. *The Oncologist,* 9**,** 361-377.

OZTEMUR ISLAKOGLU, Y., NOYAN, S., AYDOS, A. & GUR DEDEOGLU, B. 2018. Meta-microRNA Biomarker Signatures to Classify Breast Cancer Subtypes. *OMICS: A Journal of Integrative Biology,* 22**,** 709-716.

PAAP, E., VERBEEK, A. L., BOTTERWECK, A. A., VAN DOORNE-NAGTEGAAL, H. J., IMHOF-TAS, M., DE KONING, H. J., OTTO, S. J., DE MUNCK, L., VAN DER STEEN, A., HOLLAND, R., DEN HEETEN, G. J. & BROEDERS, M. J. 2014. Breast cancer screening halves the risk of breast cancer death: a case-referent study. *Breast,* 23**,** 439-44.

PAGET, S. 1889. THE DISTRIBUTION OF SECONDARY GROWTHS IN CANCER OF THE BREAST. *The Lancet,* 133**,** 571-573.

PALMERO, I., PANTOJA, C. & SERRANO, M. 1998. p19ARF links the tumor suppressor p53 to Ras. *Nature,* 395**,** 125-6.

PAMPENA, R., KYRGIDIS, A., LALLAS, A., MOSCARELLA, E., ARGENZIANO, G. & LONGO, C. 2017. A meta-analysis of nevus-associated melanoma: Prevalence and practical implications. *J Am Acad Dermatol,* 77**,** 938-945.e4.

PARISH, C. R. 2003. Cancer immunotherapy: The past, the present and the future[ast]. *Immunol Cell Biol,* 81**,** 106-113.

PARK, C. C., BISSELL, M. J. & BARCELLOS-HOFF, M. H. 2000. The influence of the microenvironment on the malignant phenotype. *Mol Med Today,* 6**,** 324-9.

PARK, J. H., ZHUANG, J., LI, J. & HWANG, P. M. 2016. p53 as guardian of the mitochondrial genome. *FEBS Lett,* 590**,** 924-34.

PARMIGIANI, G., BOCA, S., LIN, J., KINZLER, K. W., VELCULESCU, V. & VOGELSTEIN, B. 2009. Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics,* 93**,** 17-21.

PARO, R. 1995. Propagating memory of transcriptional states. *Trends Genet,* 11**,** 295-7.

PASQUINELLI, A. E., REINHART, B. J., SLACK, F., MARTINDALE, M. Q., KURODA, M. I., MALLER, B., HAYWARD, D. C., BALL, E. E., DEGNAN, B., MULLER, P., SPRING, J., SRINIVASAN, A., FISHMAN, M., FINNERTY, J., CORBO, J., LEVINE, M., LEAHY, P., DAVIDSON, E. & RUVKUN, G. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature,* 408**,** 86-9.

PECORINO, L. 2012. *Molecular Biology of Cancer: Mechanisms, Targets, and Therapeutics*, OUP Oxford.

PERERA, E., GNANESWARAN, N., JENNENS, R. & SINCLAIR, R. 2013. Malignant Melanoma. *Healthcare (Basel, Switzerland),* 2**,** 1-19.

PEROU, C. M. 2010. Molecular Stratification of Triple-Negative Breast Cancers. *The Oncologist,* 15**,** 39-48.

PEROU, C. M., SORLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LONNING, P. E., BORRESEN-DALE, A.-L., BROWN, P. O. & BOTSTEIN, D. 2000. Molecular portraits of human breast tumors. *Nature,* 406**,** 747-752.

PETERSEN, B.-S., FREDRICH, B., HOEPPNER, M. P., ELLINGHAUS, D. & FRANKE, A. 2017. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genetics,* 18**,** 14.

PETO, R., DAVIES, C., GODWIN, J., GRAY, R., PAN, H. C., CLARKE, M., CUTTER, D., DARBY, S., MCGALE, P., TAYLOR, C., WANG, Y. C., BERGH, J., DI LEO, A., ALBAIN, K., SWAIN, S., PICCART, M. & PRITCHARD, K. 2012. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet,* 379**,** 432-44.

POLLOCK, P. M., HARPER, U. L., HANSEN, K. S., YUDT, L. M., STARK, M., ROBBINS, C. M., MOSES, T. Y., HOSTETTER, G., WAGNER, U., KAKAREKA, J., SALEM, G., POHIDA, T., HEENAN, P., DURAY, P., KALLIONIEMI, O., HAYWARD, N. K., TRENT, J. M. & MELTZER, P. S. 2003. High frequency of BRAF mutations in nevi. *Nat Genet,* 33**,** 19-20.

POY, M. N., ELIASSON, L., KRUTZFELDT, J., KUWAJIMA, S., MA, X., MACDONALD, P. E., PFEFFER, S., TUSCHL, T., RAJEWSKY, N., RORSMAN, P. & STOFFEL, M. 2004. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature,* 432**,** 226-30.

PRAT, A., PARKER, J. S., FAN, C. & PEROU, C. M. 2012. PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res Treat,* 135**,** 301-6.

PRAT, A. & PEROU, C. M. 2011. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology,* 5**,** 5-23.

PSATY, E. L., SCOPE, A., HALPERN, A. C. & MARGHOOB, A. A. 2010. Defining the patient at high risk for melanoma. *Int J Dermatol,* 49**,** 362-76.

RAHMAN, N., SEAL, S., THOMPSON, D., KELLY, P., RENWICK, A., ELLIOTT, A., REID, S., SPANOVA, K., BARFOOT, R., CHAGTAI, T., JAYATILAKE, H., MCGUFFOG, L., HANKS, S., EVANS, D. G., ECCLES, D., EASTON, D. F. & STRATTON, M. R. 2007. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet,* 39**,** 165-7.

RAINE, K. M., VAN LOO, P., WEDGE, D. C., JONES, D., MENZIES, A., BUTLER, A. P., TEAGUE, J. W., TARPEY, P., NIK-ZAINAL, S. & CAMPBELL, P. J. 2016. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics,* 56**,** 15.9.1-15.9.17.

RASTRELLI, M., TROPEA, S., ROSSI, C. R. & ALAIBAC, M. 2014. Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification. *In Vivo,* 28**,** 1005-11.

READ, J., WADT, K. A. W. & HAYWARD, N. K. 2016. Melanoma genetics. *Journal of Medical Genetics,* 53**,** 1.

REIK, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature,* 447**,** 425-32.

REINHART, B. J., SLACK, F. J., BASSON, M., PASQUINELLI, A. E., BETTINGER, J. C., ROUGVIE, A. E., HORVITZ, H. R. & RUVKUN, G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature,* 403**,** 901-6.

ROBERTSON, K. D. 2005. DNA methylation and human disease. *Nat Rev Genet,* 6**,** 597-610.

ROBSAHM, T. E., HELSING, P., NILSSEN, Y., VOS, L., RIZVI, S. M. H., AKSLEN, L. A. & VEIERØD, M. B. 2018. High mortality due to cutaneous melanoma in Norway: a study of prognostic factors in a nationwide cancer registry. *Clinical epidemiology,* 10**,** 537-548.

ROEDER, I. & LOEFFLER, M. 2002. A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity. *Exp Hematol,* 30**,** 853-61.

RONAGHI, M., UHLEN, M. & NYREN, P. 1998. A sequencing method based on real-time pyrophosphate. *Science,* 281**,** 363, 365.

ROSENBERG, S. M. & PARTRIDGE, A. H. 2015. Management of breast cancer in very young women. *The Breast,* 24, Supplement 2**,** S154-S158.

ROSENTHAL, R., MCGRANAHAN, N., HERRERO, J., TAYLOR, B. S. & SWANTON, C. 2016. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology,* 17**,** 31.

ROTH, S. Y., DENU, J. M. & ALLIS, C. D. 2001. Histone acetyltransferases. *Annu Rev Biochem,* 70**,** 81-120.

ROWINSKY, E. K. & DONEHOWER, R. C. 1995. Paclitaxel (taxol). *N Engl J Med,* 332**,** 1004-14.

RUBY, J. G., JAN, C. H. & BARTEL, D. P. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature,* 448**,** 83-6.

SAKAI, T., TOGUCHIDA, J., OHTANI, N., YANDELL, D. W., RAPAPORT, J. M. & DRYJA, T. P. 1991. Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *American journal of human genetics,* 48**,** 880-888.

SANGER, F., AIR, G. M., BARRELL, B. G., BROWN, N. L., COULSON, A. R., FIDDES, C. A., HUTCHISON, C. A., SLOCOMBE, P. M. & SMITH, M. 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature,* 265**,** 687-95.

SANGER, F. & COULSON, A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol,* 94**,** 441-8.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977b. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences,* 74**,** 5463-5467.

SASIDHARAN, V., LU, Y. C., BANSAL, D., DASARI, P., PODUVAL, D., SESHASAYEE, A., RESCH, A. M., GRAVELEY, B. R. & PALAKODETI, D. 2013. Identification of neoblast- and regeneration-specific miRNAs in the planarian Schmidtea mediterranea. *Rna,* 19**,** 1394-404.

SASSEN, S., MISKA, E. A. & CALDAS, C. 2008. MicroRNA: implications for cancer. *Virchows Archiv : an international journal of pathology,* 452**,** 1-10.

SATZGER, I., MATTERN, A., KUETTLER, U., WEINSPACH, D., VOELKER, B., KAPP, A. & GUTZMER, R. 2010. MicroRNA-15b represents an independent prognostic parameter and is correlated with tumor cell proliferation and apoptosis in malignant melanoma. *Int J Cancer,* 126**,** 2553-62.

SAUNDERS, C. T., WONG, W. S., SWAMY, S., BECQ, J., MURRAY, L. J. & CHEETHAM, R. K. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics,* 28**,** 1811-7.

SAXONOV, S., BERG, P. & BRUTLAG, D. L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences,* 103**,** 1412.

SCHICKEL, R., BOYERINAS, B., PARK, S. M. & PETER, M. E. 2008. MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death. *Oncogene,* 27**,** 5959-74.

SCHMID, P., ADAMS, S., RUGO, H. S., SCHNEEWEISS, A., BARRIOS, C. H., IWATA, H., DIERAS, V., HEGG, R., IM, S. A., SHAW WRIGHT, G., HENSCHEL, V., MOLINERO, L., CHUI, S. Y., FUNKE, R., HUSAIN, A., WINER, E. P., LOI, S. & EMENS, L. A. 2018. Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. *N Engl J Med,* 379**,** 2108-2121.

SCHUBELER, D. 2015. Function and information content of DNA methylation. *Nature,* 517**,** 321-6.

SCHWARZ, D. S., HUTVAGNER, G., DU, T., XU, Z., ARONIN, N. & ZAMORE, P. D. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell,* 115**,** 199-208.

SEGURA, M. F., HANNIFORD, D., MENENDEZ, S., REAVIE, L., ZOU, X., ALVAREZ-DIAZ, S., ZAKRZEWSKI, J., BLOCHIN, E., ROSE, A., BOGUNOVIC, D., POLSKY, D., WEI, J., LEE, P., BELITSKAYA-LEVY, I., BHARDWAJ, N., OSMAN, I. & HERNANDO, E. 2009. Aberrant miR-182 expression promotes melanoma metastasis by repressing FOXO3 and microphthalmia-associated transcription factor. *Proc Natl Acad Sci U S A,* 106**,** 1814-9.

SEIDEL, J. A., OTSUKA, A. & KABASHIMA, K. 2018. Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations. *Frontiers in oncology,* 8**,** 86-86.

SHUKLA, V. K., HUGHES, D. C., HUGHES, L. E., MCCORMICK, F. & PADUA, R. A. 1989. ras mutations in human melanotic lesions: K-ras activation is a frequent and early event in melanoma development. *Oncogene Res,* 5**,** 121-7.

SLACK, F. J., BASSON, M., LIU, Z., AMBROS, V., HORVITZ, H. R. & RUVKUN, G. 2000. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Mol Cell,* 5**,** 659-69.

SMITH, J., THO, L. M., XU, N. & GILLESPIE, D. A. 2010. The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Adv Cancer Res,* 108**,** 73-112.

SMITH, L. M., FUNG, S., HUNKAPILLER, M. W., HUNKAPILLER, T. J. & HOOD, L. E. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res,* 13**,** 2399-412.

SMITH, L. M., SANDERS, J. Z., KAISER, R. J., HUGHES, P., DODD, C., CONNELL, C. R., HEINER, C., KENT, S. B. & HOOD, L. E. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature,* 321**,** 674-9.

SMITH, Z. D. & MEISSNER, A. 2013. DNA methylation: roles in mammalian development. *Nature Reviews Genetics,* 14**,** 204.

SORLIE, P. D., THOM, T. J., MANOLIO, T., ROSENBERG, H. M., ANDERSON, R. N. & BURKE, G. L. 1999. Age-adjusted death rates: consequences of the Year 2000 standard. *Ann Epidemiol,* 9**,** 93-100.

SORLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M. & JEFFREY, S. S. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98.

SØRLIE, T., PEROU, C. M., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H., HASTIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., THORSEN, T., QUIST, H., MATESE, J. C., BROWN, P. O., BOTSTEIN, D., LØNNING, P. E. & BØRRESEN-DALE, A.-L. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences,* 98**,** 10869-10874.

SORLIE, T., TIBSHIRANI, R., PARKER, J., HASTIE, T., MARRON, J. S., NOBEL, A., DENG, S., JOHNSEN, H., PESICH, R., GEISLER, S., DEMETER, J., PEROU, C. M., LONNING, P. E., BROWN, P. O., BORRESEN-DALE, A. L. & BOTSTEIN, D. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A,* 100**,** 8418-23.

STAHL, J. M., SHARMA, A., CHEUNG, M., ZIMMERMAN, M., CHENG, J. Q., BOSENBERG, M. W., KESTER, M., SANDIRASEGARANE, L. & ROBERTSON, G. P. 2004. Deregulated Akt3 activity promotes development of malignant melanoma. *Cancer Res,* 64**,** 7002-10.

STEEG, P. S. & ZHOU, Q. 1998. Cyclins and breast cancer. *Breast Cancer Research and Treatment,* 52**,** 17-28.

STRAHL, B. D. & ALLIS, C. D. 2000. The language of covalent histone modifications. *Nature,* 403**,** 41-5.

STRANNEHEIM, H. & WEDELL, A. 2016. Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. *Journal of Internal Medicine,* 279**,** 3-15.

SUDHAKAR, A. 2009. History of Cancer, Ancient and Modern Treatment Methods. *J Cancer Sci Ther,* 1**,** 1-4.

SUMMERHAYES, M. 1995. Review : Amifostine and other chemoprotective agents in cancer chemotherapy: A brief review. *Journal of Oncology Pharmacy Practice,* 1**,** 21-31.

SUN, Y. S., ZHAO, Z., YANG, Z. N., XU, F., LU, H. J., ZHU, Z. Y., SHI, W., JIANG, J., YAO, P. P. & ZHU, H. P. 2017. Risk Factors and Preventions of Breast Cancer. *Int J Biol Sci,* 13**,** 1387-1397.

SWAIN, S. M., BASELGA, J., KIM, S.-B., RO, J., SEMIGLAZOV, V., CAMPONE, M., CIRUELOS, E., FERRERO, J.-M., SCHNEEWEISS, A., HEESON, S., CLARK, E., ROSS, G., BENYUNES, M. C. & CORTÉS, J. 2015. Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer. *New England Journal of Medicine,* 372**,** 724-734.

TASHKANDI, H., SHAH, N., PATEL, Y. & CHEN, H. 2015. Identification of new miRNA biomarkers associated with HER2-positive breast cancers. *Oncoscience,* 2**,** 924-9.

TAWFIK, D. S. & GRIFFITHS, A. D. 1998. Man-made cell-like compartments for molecular evolution. *Nat Biotechnol,* 16**,** 652-6.

THE CANCER GENOME ATLAS, N., KOBOLDT, D. C., FULTON, R. S., MCLELLAN, M. D., SCHMIDT, H., KALICKI-VEIZER, J., MCMICHAEL, J. F., FULTON, L. L., DOOLING, D. J., DING, L., MARDIS, E. R., WILSON, R. K., ALLY, A., BALASUNDARAM, M., BUTTERFIELD, Y. S. N., CARLSEN, R., CARTER, C., CHU, A., CHUAH, E., CHUN, H.-J. E., COOPE, R. J. N., DHALLA, N., GUIN, R., HIRST, C., HIRST, M., HOLT, R. A., LEE, D., LI, H. I., MAYO, M., MOORE, R. A., MUNGALL, A. J., PLEASANCE, E., GORDON ROBERTSON, A., SCHEIN, J. E., SHAFIEI, A., SIPAHIMALANI, P., SLOBODAN, J. R., STOLL, D., TAM, A., THIESSEN, N., VARHOL, R. J., WYE, N., ZENG, T., ZHAO, Y., BIROL, I., JONES, S. J. M., MARRA, M. A., CHERNIACK, A. D., SAKSENA, G., ONOFRIO, R. C., PHO, N. H., CARTER, S. L., SCHUMACHER, S. E., TABAK, B., HERNANDEZ, B., GENTRY, J., NGUYEN, H., CRENSHAW, A., ARDLIE, K., BEROUKHIM, R., WINCKLER, W., GETZ, G., GABRIEL, S. B., MEYERSON, M., CHIN, L., PARK, P. J., KUCHERLAPATI, R., HOADLEY, K. A., TODD AUMAN, J., FAN, C., TURMAN, Y. J., SHI, Y., LI, L., TOPAL, M. D., HE, X., CHAO, H.-H., PRAT, A., SILVA, G. O., IGLESIA, M. D., ZHAO, W., USARY, J., BERG, J. S., ADAMS, M., BOOKER, J., WU, J., GULABANI, A., BODENHEIMER, T., HOYLE, A. P., SIMONS, J. V., SOLOWAY, M. G., MOSE, L. E., JEFFERYS, S. R., BALU, S., PARKER, J. S., NEIL HAYES, D., PEROU, C. M., MALIK, S., MAHURKAR, S., SHEN, H., WEISENBERGER, D. J., et al. 2012. Comprehensive molecular portraits of human breast tumors. *Nature,* 490**,** 61.

TORRE, L. A., SIEGEL, R. L., WARD, E. M. & JEMAL, A. 2016. Global Cancer Incidence and Mortality Rates and Trends—An Update. *Cancer Epidemiology Biomarkers & Prevention,* 25**,** 16.

TRANG, P., MEDINA, P. P., WIGGINS, J. F., RUFFINO, L., KELNAR, K., OMOTOLA, M., HOMER, R., BROWN, D., BADER, A. G., WEIDHAAS, J. B. & SLACK, F. J. 2010. Regression of murine lung tumors by the let-7 microRNA. *Oncogene,* 29**,** 1580-7.

TRYGGVADÓTTIR, L., GISLUM, M., HAKULINEN, T., KLINT, Å., ENGHOLM, G., STORM, H. H. & BRAY, F. 2010. Trends in the survival of patients diagnosed with malignant melanoma of the skin in the Nordic countries 1964–2003 followed up to the end of 2006. *Acta Oncologica,* 49**,** 665-672.

TUNA, M. & AMOS, C. I. 2013. Genomic sequencing in cancer. *Cancer Lett,* 340**,** 161-70.

TURCATTI, G., ROMIEU, A., FEDURCO, M. & TAIRI, A. P. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res,* 36**,** e25.

VAN 'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J., WITTEVEEN, A. T., SCHREIBER, G. J., KERKHOVEN, R. M., ROBERTS, C., LINSLEY, P. S., BERNARDS, R. & FRIEND, S. H. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature,* 415**,** 530-6.

VAN LOO, P., NORDGARD, S. H., LINGJÆRDE, O. C., RUSSNES, H. G., RYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B., PEROU, C. M., BØRRESEN-DALE, A.-L. & KRISTENSEN, V. N. 2010. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences,* 107**,** 16910-16915.

VENNEPUREDDY, A., THUMALLAPALLY, N., MOTILAL NEHRU, V., ATALLAH, J.-P. & TERJANIAN, T. 2016. Novel Drugs and Combination Therapies for the Treatment of Metastatic Melanoma. *Journal of clinical medicine research,* 8**,** 63-75.

VERMA, M., ROGERS, S., DIVI, R. L., SCHULLY, S. D., NELSON, S., SU, L. J., ROSS, S., PILCH, S., WINN, D. M. & KHOURY, M. J. 2014. Epigenetic Research in Cancer Epidemiology: Trends, Opportunities, and Challenges. *Cancer Epidemiol Biomarkers Prev,* 23**,** 223-33.

VIJAYASARADHI, S. 1995. Intracellular sorting and targeting of melanosomal membrane proteins: identification of signals for sorting of the human brown locus protein, gp75. *J Cell Biol,* 130**,** 807-20.

VINAGRE, J., ALMEIDA, A., POPULO, H., BATISTA, R., LYRA, J., PINTO, V., COELHO, R., CELESTINO, R., PRAZERES, H., LIMA, L., MELO, M., DA ROCHA, A. G., PRETO, A., CASTRO, P., CASTRO, L., PARDAL, F., LOPES, J. M., SANTOS, L. L., REIS, R. M., CAMESELLE-TEIJEIRO, J., SOBRINHO-SIMOES, M., LIMA, J., MAXIMO, V. & SOARES, P. 2013. Frequency of TERT promoter mutations in human cancers. *Nat Commun,* 4**,** 2185.

VOLINIA, S., GALASSO, M., SANA, M. E., WISE, T. F., PALATINI, J., HUEBNER, K. & CROCE, C. M. 2012. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. *Proc Natl Acad Sci U S A,* 109**,** 3024-9.

WANG, J., SAMUELS, D. C., ZHAO, S., XIANG, Y., ZHAO, Y. Y. & GUO, Y. 2017. Current Research on Non-Coding Ribonucleic Acid (RNA). *Genes (Basel),* 8.

WANG, K., LI, M. & HAKONARSON, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38**,** e164.

WANG, S. E., XIANG, B., ZENT, R., QUARANTA, V., POZZI, A. & ARTEAGA, C. L. 2009. Transforming growth factor beta induces clustering of HER2 and integrins by activating Src-focal adhesion kinase and receptor association to the cytoskeleton. *Cancer Res,* 69**,** 475-82.

WANG, X. & EL NAQA, I. M. 2008. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics,* 24**,** 325-32.

WANGARI-TALBOT, J. & CHEN, S. 2013. Genetics of melanoma. *Frontiers in genetics,* 3**,** 330-330.

WATSON, J. D. & CRICK, F. H. C. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature,* 171**,** 737-738.

WEBER, M. & SCHUBELER, D. 2007. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr Opin Cell Biol,* 19**,** 273-80.

WEINBERG, R. A. 1994. Oncogenes and tumor suppressor genes. *CA: A Cancer Journal for Clinicians,* 44**,** 160-170.

WENDT, J., ROSENBAUM, H., RICHMOND, T. A., JEDDELOH, J. A. & BURGESS, D. L. 2018. Targeted Bisulfite Sequencing Using the SeqCap Epi Enrichment System. *In:* TOST, J. (ed.) *DNA Methylation Protocols.* New York, NY: Springer New York.

WIGHTMAN, B., HA, I. & RUVKUN, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* 75**,** 855-62.

WONG, N. & WANG, X. 2015. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res,* 43**,** D146-52.

WRECZYCKA, K., GOSDSCHAN, A., YUSUF, D., GRÜNING, B., ASSENOV, Y. & AKALIN, A. 2017. Strategies for analyzing bisulfite sequencing data. *Journal of Biotechnology,* 261**,** 105-115.

XI, Y. & LI, W. 2009. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics,* 10**,** 232.

XIANG, F., LUCAS, R., HALES, S. & NEALE, R. 2014. Incidence of nonmelanoma skin cancer in relation to ambient UV radiation in white populations, 1978-2012: empirical relationships. *JAMA Dermatol,* 150**,** 1063-71.

XUAN, J., YU, Y., QING, T., GUO, L. & SHI, L. 2013. Next-generation sequencing in the clinic: promises and challenges. *Cancer letters,* 340**,** 284-295.

YAN, L. X., HUANG, X. F., SHAO, Q., HUANG, M. A. Y., DENG, L., WU, Q. L., ZENG, Y. X. & SHAO, J. Y. 2008. MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. *Rna,* 14**,** 2348-60.

YANG, J. S. & LAI, E. C. 2010. Dicer-independent, Ago2-mediated microRNA biogenesis in vertebrates. *Cell Cycle,* 9**,** 4455-60.

YATES, L. R., GERSTUNG, M., KNAPPSKOG, S., DESMEDT, C., GUNDEM, G., VAN LOO, P., AAS, T., ALEXANDROV, L. B., LARSIMONT, D., DAVIES, H., LI, Y., JU, Y. S., RAMAKRISHNA, M., HAUGLAND, H. K., LILLENG, P. K., NIK-ZAINAL, S., MCLAREN, S., BUTLER, A., MARTIN, S., GLODZIK, D., MENZIES, A., RAINE, K., HINTON, J., JONES, D., MUDIE, L. J., JIANG, B., VINCENT, D., GREENE-COLOZZI, A., ADNET, P. Y., FATIMA, A., MAETENS, M., IGNATIADIS, M., STRATTON, M. R., SOTIRIOU, C., RICHARDSON, A. L., LONNING, P. E., WEDGE, D. C. & CAMPBELL, P. J. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med,* 21**,** 751-9.

YEUNG, M. L., BENNASSER, Y., LE, S. Y. & JEANG, K. T. 2005. siRNA, miRNA and HIV: promises and challenges. *Cell Research,* 15**,** 935.

YI, R., QIN, Y., MACARA, I. G. & CULLEN, B. R. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev,* 17**,** 3011-6.

YIN, Y. & SHEN, W. H. 2008. PTEN: a new guardian of the genome. *Oncogene,* 27**,** 5443-53.

ZHANG, J., CHEN, Y.-H. & LU, Q. 2010. Pro-oncogenic and anti-oncogenic pathways: opportunities and challenges of cancer therapy. *Future Oncology,* 6**,** 587-603.

ZHANG, X. P., LIU, F. & WANG, W. 2011. Two-phase dynamics of p53 in the DNA damage response. *Proc Natl Acad Sci U S A,* 108**,** 8990-5.

ZHAO, Y., SAMAL, E. & SRIVASTAVA, D. 2005. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature,* 436**,** 214-20.

ZINDY, F., EISCHEN, C. M., RANDLE, D. H., KAMIJO, T., CLEVELAND, J. L., SHERR, C. J. & ROUSSEL, M. F. 1998. Myc signaling via the ARF tumor

suppressor regulates p53-dependent apoptosis and immortalization. *Genes Dev,* 12**,** 2424-33.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05063-1 OPEN

Patterns of genomic evolution in advanced melanoma

E. Birkeland[1,2], S. Zhang[1,2], D. Poduval[1,2], J. Geisler[3,4], S. Nakken[5,6], D. Vodak[5,6], L.A. Meza-Zepeda[5,6,7], E. Hovig[5,6,8,9], O. Myklebost[5,6], S. Knappskog[1,2] & P.E. Lønning[1,2]

Genomic alterations occurring during melanoma progression and the resulting genomic heterogeneity between metastatic deposits remain incompletely understood. Analyzing 86 metastatic melanoma deposits from 53 patients with whole-exome sequencing (WES), we show a low branch to trunk mutation ratio and little intermetastatic heterogeneity, with driver mutations almost completely shared between lesions. Branch mutations consistent with UV damage indicate that metastases may arise from different subclones in the primary tumor. Selective gain of mutated *BRAF* alleles occurs as an early event, contrasting whole-genome duplication (WGD) occurring as a late truncal event in about 40% of cases. One patient revealed elevated mutational diversity, probably related to previous chemotherapy and DNA repair defects. In another patient having received radiotherapy toward a lymph node metastasis, we detected a radiotherapy-related mutational signature in two subsequent distant relapses, consistent with secondary metastatic seeding. Our findings add to the understanding of genomic evolution in metastatic melanomas.

[1] Section of Oncology, Department of Clinical Science, University of Bergen, 5020 Bergen, Norway. [2] Department of Oncology, Haukeland University Hospital, 5021 Bergen, Norway. [3] Institute of Clinical Medicine, University of Oslo, Campus Akershus University Hospital, 1478 Lørenskog, Oslo, Norway. [4] Department of Oncology, Akershus University Hospital, 1478 Lørenskog, Norway. [5] Department of Tumor Biology, Institute of Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, 0310 Oslo, Norway. [6] Norwegian Cancer Genomics Consortium, Institute for Cancer Research, Oslo University Hospital –Radium Hospital, 0310 Oslo, Norway. [7] Genomics Core Facility, Department of Core Facilities, Institute of Cancer Research, the Norwegian Radium Hospital, 0310 Oslo, Norway. [8] Department of Informatics, University of Oslo, 0316 Oslo, Norway. [9] Institute of Cancer Genetics and Informatics, The Norwegian Radium Hospital, Oslo University Hospital, 0310 Oslo, Norway. Correspondence and requests for materials should be addressed to P.E.Løn. (email: per.lonning@helse-bergen.no)

The incidence of melanoma is rapidly increasing among light-skinned people[1], where both epidemiological[2] and genomic evidence have established the link between melanoma etiology and UV radiation[3–5]. Many melanomas reveal an indolent course characterized by locoregional relapses followed by a rapid emergence of metastatic disease, and there is evidence suggesting that systemic dissemination may bypass intermediary stages of lymph node involvement[6,7].

Somatic mutations found in a cancer mirror its initiation and evolution, and genomic sequencing may thus map the progression of melanomas from earlier stages of development, enabling inferences that are empowered by comparisons of multiple lesions. While a few studies have used comparative lesion sequencing to assess genomic events during the process from benign lesions to primary melanomas[8] and progression from primary to regional disease[9], most studies of metastatic melanoma have explored genome evolution in response to targeted therapy[10–12]. A picture is emerging where most UV-associated mutations arise in the primary tumor prior to malignant transformation, followed by an increased frequency of copy number alterations[8]. The genomic events driving tumor progression toward advanced disease, however, remain incompletely understood.

Melanomas have low sensitivity to chemotherapy[13]. While recent developments including immune checkpoint inhibitors and BRAF/MEK targeting agents have improved the outcomes significantly, many patients do not achieve durable remissions[14,15]. Thus, improvements in therapy are needed. This may be facilitated by an improved understanding of genomic events associated with accelerated growth and dissemination.

Here we performed whole-exome sequencing (WES) of single or multiple metastases from a cohort of patients with advanced melanoma. Our findings add novel data to the understanding of the chronological sequence of genomic alterations. This includes early copy number gain of the mutated *BRAF* allele and the finding that whole-genome duplication (WGD) in general occurs as a late truncal event. While we found evidence indicating polyclonal seeding in one patient, this seems to be a rare event. Among four patients exposed to dacarbazine, we observed a "mutational signature" in one, probably related to several *MSH6* mutations in her tumor. Moreover, the finding that radiotherapy toward a lymph node metastasis may influence mutation signatures in subsequent deposits in organs distant from the treatment site supports the hypothesis that cancers may progress also through secondary spread from metastatic deposits.

Table 1 Patient characteristics

Baseline characteristics	Patients
Sex	
Female	22
Male	31
Disease origin	
Cutaneous (non-glabrous skin)	
Head	5
Upper extremities	5
Trunk	20
Lower extremities	7
Acral	3
Uveal	2
Mucosal	2
Primary unknown	9
Number of samples	
1	30
2	16
≥3	7
Molecular characteristics	
Mutational subtype	
BRAF	27
NRAS	17
NF1	2
Triple wild type	7
Genome duplication	
Near-diploid	32
Genome duplicated	21
Total	53

Results

Single-base substitutions and indels. We analyzed 114 metastatic lesions with matched normal tissue from 60 patients diagnosed with advanced melanoma by WES. All patients were from a prospective study assessing dacarbazine therapy for metastatic melanoma[16,17]. Eighty-six lesions from 53 patients consisting of at least 20% tumor cells (threshold for copy number profiling) were selected for further analysis (identified mutations in these samples are presented in Supplementary Table 1). Multiple lesions were available from 23 out of the 53 patients, and single-metastatic lesions were available from the remaining 30 (Table 1, and Supplementary Tables 2 and 3).

The number of somatic variants identified in coding regions per patient (average across samples for patients with multiple biopsies) varied substantially, with between 17 and 4089 mutations identified (range: 0.34–81.8 mutations per megabase, median: 9.6; Fig. 1a). With few exceptions, tumors with primary origins at sun-exposed sites all displayed mutational patterns characterized by C>T transitions at dipyrimidine sites, in contrast to tumors derived from areas not exposed to UV radiation (Fig. 1b), consistent with UV-induced DNA damage (Fig. 1c). One acral melanoma had a UV-associated mutational signature, as has also been observed by others[18,19]. Overall though, patients with sun-exposed primary tumors had a higher mutational load than patients with primary lesions at sites with little or no such exposure ($p < 0.001$, Mann–Whitney *U*-test [MW]; Supplementary Figure 1a). No difference in mutation load between the lymph node and subcutaneous or visceral organ metastases was recorded (Supplementary Figure 1b).

Among nine patients diagnosed with metastatic melanoma without known primary lesions, the types and numbers of mutations resembled those observed in metastases from sun-exposed primary lesions, strongly suggesting cutaneous origins (Fig. 1, Supplementary Figure 1a) as previously reported by others[20].

Driver mutations and genomic complexity. Using a conservative approach to identify driver mutations, we considered mutations in a set of predefined genes based on recently published studies[3,21,22]. Mutations in these genes were manually assessed to determine their status as drivers or passengers (Methods section). The complete list of mutations in these genes is presented in Supplementary Table 4. Driver mutations in *BRAF* and *NRAS* were detected among 28 (53%) and 17 (32%) patients, respectively (Fig. 1d), with one patient carrying a non-canonical driver mutation in *BRAF* (p.E586K) in combination with a driver mutation in *NRAS* (p.Q61L). While protein-altering mutations in *NF1* were identified in five patients, only two of these fulfilled our criteria for definition as driver mutation. Driver mutations in *GNAQ* and *GNA11* were identified in two uveal melanomas, and a driver mutation in *KIT* was found in mucosal melanoma.

Considering patients with multiple sampled lesions, all driver mutations identified were shared between metastatic deposits,

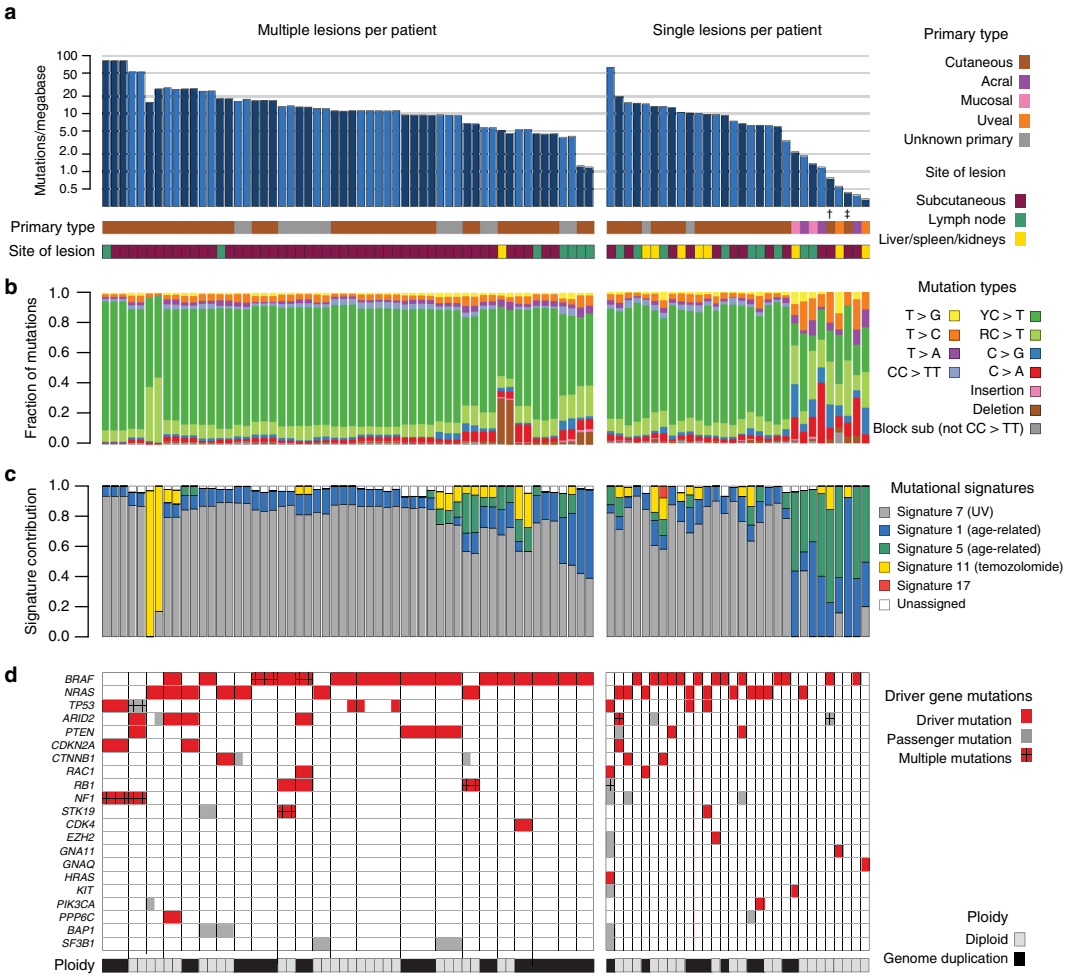


Fig. 1 Overview of mutations. **a**–**d** Left: patients from whom multiple lesions were analyzed. Right: patients from whom single lesions were sampled. Patients are ordered by the number of mutations identified per patient, and lesions are further ordered according to time of sampling. **a** Number of mutations per megabase in each individual sample. Samples from different patients are indicated by alternating shades of blue. † One patient had a borderline acral primary tumor situated at a toe. ‡ One patient had a perianal cutaneous primary tumor, likely not exposed to UV radiation. **b** Fraction of mutation types per lesion. **c** Estimated contribution of mutational processes by fraction of mutations explained by each mutational signature[26], according to the classification of Alexandrov et al.[32]. Only signatures explaining >5% of mutations are shown, and only signatures 7, 1, 5, 11, and 17 were assessed. **d** Mutations identified per lesion in established melanoma driver genes are color-coded: red boxes indicate driver mutations and gray boxes indicate passenger mutations. Multiple mutations per gene are indicated with "+". Genome duplication events are shown per sample in gray (diploid) or black (genome duplication) for each sample

except for two patients, both revealing heterogeneous and subclonal distribution of the p.Y163C *TP53* mutation.

In accordance with previous reports[22,23], we found the number of mutations to vary according to driver mutation status in *BRAF*, *NRAS*, and *NF1* ($p = 0.002$, Kruskal–Wallis rank-sum test [KW]; Supplementary Figure 1c). Based on copy number profiling (Supplementary Figure 2a), we inferred whole-genome duplication (WGD) events to have occurred in about 40% of patients (Supplementary Figure 2b), with no difference between tumors harboring *BRAF* (11/27) or *NRAS* (7/17) mutations. The duplication events likely predated evolutionary divergence of metastases, as they were identified across all lesions obtained from these patients. Notably, the genomic complexity (defined as the fraction of the genome in an aberrant state, i.e., deviation from a balanced copy number of two for diploid tumors and four for WGD) was substantially higher in samples with WGD, with a mean of 69% for WGD and 30% for diploid tumors ($p < 0.001$, MW test; Supplementary Figure 2c). A difference in genomic complexity of this magnitude indicates a greater propensity for genomic alterations following genome duplication, as previously reported in other cancer forms[24,25].

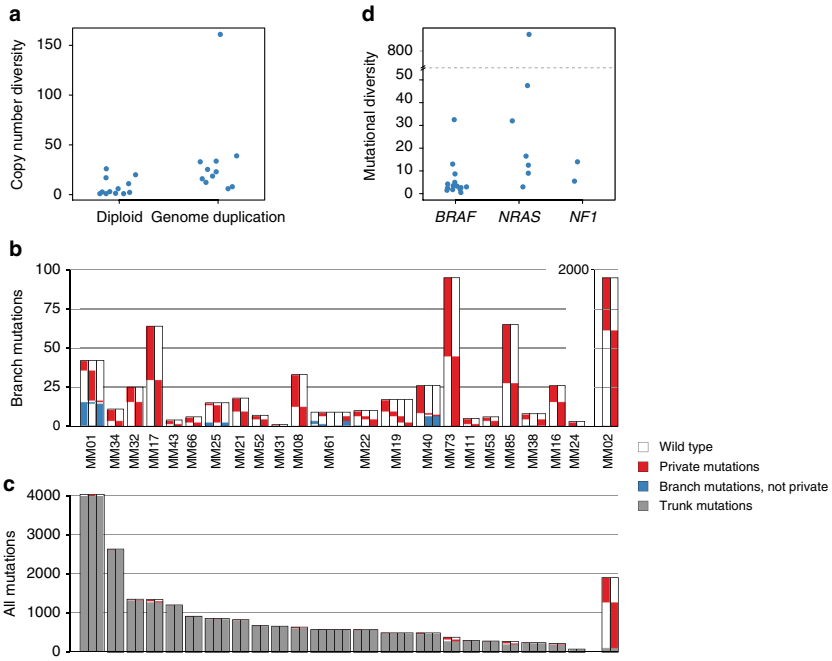


Fig. 2 Mutational heterogeneity. **a** Copy number diversity according to whether genome duplication was identified in samples from each patient. **b** Branch mutations (found in more than one, but not all samples) and private mutations (found exclusively in one sample) per lesion. For MM02, a separate axis is used to capture the large number of private mutations. **c** All mutations (private, branch, and trunk mutations) presented together for each lesion. **d** Mutational diversity (average number of branch mutations per sample) in patients according to driver mutation status of *BRAF* and *NRAS*. The *Y*-axis is broken for clarity due to the high mutational diversity in MM02

Heterogeneity across intraindividual lesions. We also observed larger copy number diversity (defined as the mean number of copy number alterations separating samples from individual patients) in patients if WGD was present compared to patients with diploid tumors, where the median copy number diversity of patients with diploid and WGD cancers was 2.8 (range 1–26) and 23 (range 6–161), respectively ($p = 0.004$, Fig. 2a). This suggests the copy number evolution to be an ongoing process occurring at a higher rate in melanomas with WGD. Diversity in copy numbers was observed across the genome, with no chromosome being overrepresented ($p = 0.3$, KW test; Supplementary Figure 3).

In order to investigate the mutational heterogeneity between melanoma metastases, we identified trunk and branch mutations for each of the 23 patients having multiple lesions examined. Mutations were classified as trunk mutations when found in all lesions examined from a particular patient, or when the absence of a mutation could be explained by a copy number loss or lack of sequencing depth in a sample without this mutation. Branch mutations were accordingly defined as those mutations whose absence could not be explained by the same features. Branch mutations were further defined as private when exclusively identified in a single sample. Thus, we defined mutational diversity for each patient as the average number of branch mutations across lesions.

Patients generally displayed a low degree of mutational diversity (range: 0.5–893, median: 5) when compared to the number of trunk mutations (range: 17–3966, median: 465; Fig. 2b, c). Thus, with the exception of a single patient (MM02) whose metastatic deposits contained 89% branch mutations (probably related to chemotherapy exposure; see below), the branch mutations for each individual patient accounted for only 0.08–14.9% of the mutation load. Notably, across patients, no correlation between the number of trunk mutations and mutational diversity was observed ($r_s = 0.01$, $p = 0.95$, Spearman's rank correlation).

While the number of mutations private to any lesion varied substantially (range 0–1156), the number of private mutations revealed a remarkable within-patient consistency, indicating an intrinsic propensity for mutational accumulation (Supplementary Figure. 4). Excluding patient MM02, who had an extremely high number of private mutations in both lesions sampled, from statistical comparison, we found the degree of intraindividual variation across the sample set to be significantly lower as compared to interindividual variation ($p = 0.003$, Levene's test for homogeneity between groups). Assessing within-patient differences in types of branch mutations, we found small variations in mutation types related to private mutations across samples, as well as branch mutation types according to clonal status (Supplementary Figure. 5a and b), supporting mutational accumulation to be related to tumor intrinsic phenotypes.

Mutational diversity was significantly lower in tumors harboring a *BRAF* versus an *NRAS* or *NF1* mutation ($p = 0.01$, KW test; Fig. 2d). While this mirrored the difference in mutational load in general, the lack of correspondence between the number of trunk mutations and mutational diversity between tumors suggests these observations to be independent. No correlation between mutational and copy number diversity across patients was observed ($r_s = -0.07$, $p = 0.8$, Spearman), and copy number diversity was unrelated to *BRAF*, *NRAS*, or *NF1* mutational status ($p = 0.8$, KW test).

Categorizing patients into four groups based on the largest anatomical distance between sampled lesions (same site; different site, but same region; different regions; or different organ system), we observed no difference in either mutational or copy number diversity related to anatomic distance between the samples ($p = 0.3$ and $p = 0.7$, KW test; Supplementary Figure 6). Also, there was no difference in diversity between synchronous metastases and those collected with an intervening time period ($p = 0.5$ and 0.7, KW test, for mutational and copy number diversity, respectively).

Shift in mutational processes. Comparing trunk to branch mutations, there was a clear shift in the types of mutations between the two groups, with branch mutations being drawn from a much more widely distributed repertoire of mutation types (Fig. 3). All of the patients with multiple sampled lesions had primary lesions in sun-exposed locations (or unknown primaries) and, consistent with a history of sun-exposure, mutational signature analysis[26] revealed 42–93% (median 84%) of trunk mutations to belong to the UV signature (Supplementary Figure. 7a). The limited number of branch mutations made any signature derivation uncertain. However, we observed a mutation pattern consistent with an UV signature in a total of six out of 14 patients (Fig. 3 and Supplementary Figure 7b), and in one patient (MM01), UV-related mutations was the major mutation type in the branches. In contrast, we observed no enrichment of UVA-associated T>G transversions[27,28].

Evaluation of polyclonal seeding. Studies of metastatic cancers including melanoma have unveiled polyclonal seeding and complex patterns of metastatic dissemination[9,29]. Applying the pigeonhole principle[30], the cellular prevalence of mutations can be used to infer the order of mutational accumulation and selective sweeps in populations of cancer cells. When comparing the cellular prevalence of mutations in two different samples of common ancestry, subclonal mutations shared across lesions may indicate polyclonal seeding, while the presence of lesion-private and clonal (defined as a mutation occurring in all tumor cells in that lesion and not in others) mutations would preclude such an interpretation and likely indicate a monoclonal origin.

We compared the relative variant allele frequency (rVAF; reflecting cellular prevalence) of private mutations in each lesion to that of trunk mutations (Fig. 4a, Supplementary Figure. 8). The rVAF distribution of trunk mutations was used to infer the likely clonal status of private mutations in each sample. Although many private mutations were clearly subclonal (e.g., MM17; Fig. 4a), 41 out of 53 samples revealed at least one clonal private mutation (Fig. 4b), implying an absence of polyclonal seeding. Only two patients (MM24 and MM31) lacked clonal private mutations altogether. Except for two mutations in MM31 having low rVAFs in both sampled lesions, these patients did not have shared subclonal mutations. Thus, we concluded that there was no strong evidence supporting polyclonal seeding in these patients either. Cross-sample mutation clustering, applying PyClone[31], corroborated these observations (Supplementary Figure 9). Yet, in

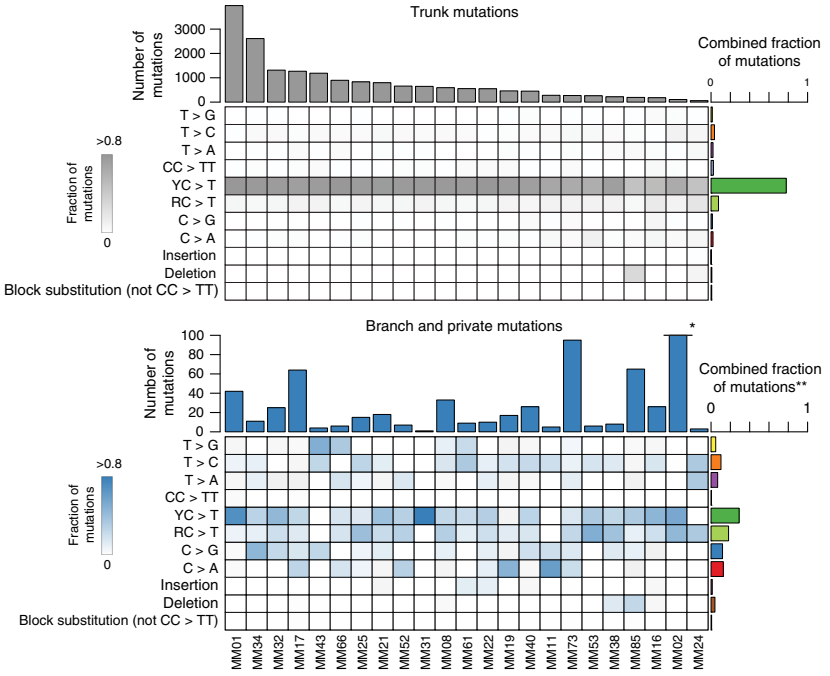


Fig. 3 Comparison of trunk and branch mutations. Heatmaps show the relative frequency of mutations among branch mutations (top panel; gray) and branch mutations (bottom panel; blue) for each patient. C>T transitions are categorized as occurring downstream of pyrimidines (Y) or purines (R). The combined fractions of mutations represent the sum of mutations for each type relative to the total number of mutations for either trunk or branch mutations. * Due to the high number of branch mutations in MM02 ($n = 1786$), the bar is truncated for this patient. ** In the summary of branch mutation types, mutations in MM02 are omitted for clarity (branch mutations in MM02 displayed a particular mutational signature; see Supplementary Fig. 7 and 11 for details)

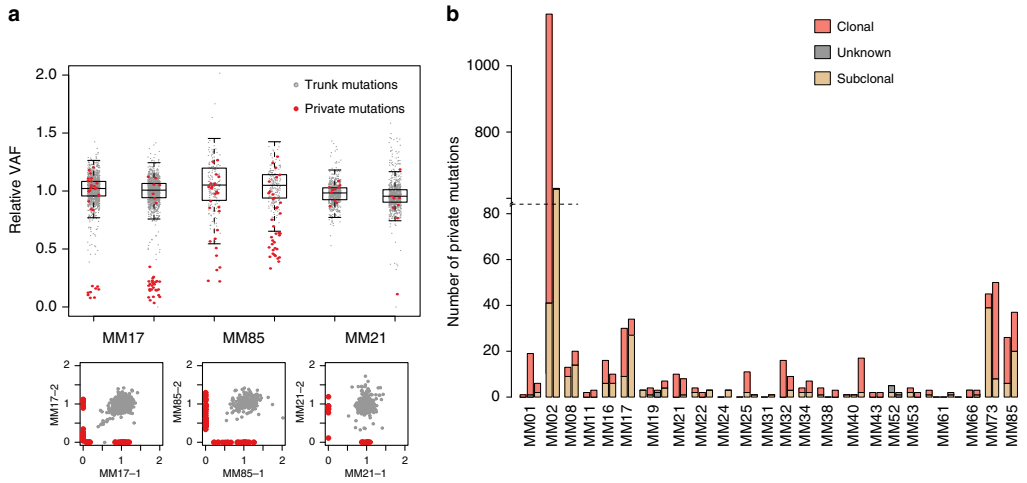


Fig. 4 Cellular prevalence of mutations. **a** Relative variant allele frequency (rVAF); that is observed variant allele frequency corrected by tumor purity, local total copy number and estimated number of mutated alleles, for mutations in six representative samples from three patients. In theory, relative VAF is equivalent to cellular prevalence of the mutations. Mutations are colored according to presence in other lesions; gray, trunk mutations; red, private mutations. Boxes with whiskers are based solely on the trunk mutation relative VAFs and span the interquartile range (IQR), with whiskers extending to 1.5 times the IQR of the trunk relative VAF from the upper and lower bounds of the boxes. The three lower panels show examples of pairwise comparisons. **b** Private mutations were classified according to status as clonal or subclonal, where subclonal mutations are those whose relative VAF is below the whiskers in **a**. Mutations below half the median relative VAF and above the subclonality threshold are defined as unknown

one patient (MM61; Supplementary Figure 10), from whom five lesions were sampled, three were without clonal private mutations, and a number of shared mutations with low rVAFs were detected in multiple samples, possibly indicating a population of cells shared subclonally between lesions[31]. These findings may indicate polyclonal origins of, or reseeding between, lesions in this patient.

The common finding of private clonal mutations is consistent with a monoclonal origin of most metastatic lesions and indicates branching evolution. Furthermore, the observation of a UV-related mutational signature in a fraction of branch mutations (Supplementary Figure. 7b) could indicate that different metastases may originate from different subclones in the primary tumor.

Potential influence of therapy. Two patients revealed atypical mutational patterns probably caused by prior therapy. One patient (MM02) had received two cycles of dacarbazine after mistakenly being diagnosed with metastatic disease. Eight months later she was correctly diagnosed with a distant subcutaneous metastasis to the abdominal region and a locoregional relapse, both of which were sampled. Nearly all private mutations were observed to occur clonally (within all cells) in the distant metastasis, but in a minor subpopulation of cells (~15%) in the locoregional relapse (Figure. 11a) and were further attributed to a mutational process previously ascribed to temozolomide treatment in glioblastoma and melanoma[32,33]. Emergence of this signature has been found to depend on concomitant inactivation of DNA mismatch repair and, potentially, DNA methyltransferase *MGMT* in glioblastoma[33,34]. Here we identified three private *MSH6* mutations in two lesions sampled from this patient, all of which coincided with the (sub-)clonal populations of hypermutated cells (Supplementary Figure. 11a). Further, reassessing previously published data[16], we identified transcriptional loss of *MGMT* in one, while the second sample revealed an *MGMT* expression level close to the median across the sample set (Supplementary Figure. 11b). Notably, neither this signature nor mutations affecting *MSH6* was detected in tumors from any other of the three patients with at least one sample collected ≥6 months after dacarbazine therapy.

The second patient (MM85) received regional radiotherapy following surgical removal of a submandibular lymph node metastasis, with subsequent sampling of two metastatic lesions: a liver deposit (5 months later) and a subcutaneous lesion on the chest wall (6 months later; Supplementary Figure. 12a). Here, a large fraction of both trunk and private mutations constituted a unique mutational signature of small deletions, typically two nucleotides in length (Supplementary Figure. 12b), akin to a recently described pattern of mutations in radiation-induced secondary malignancies[35]. To the best of our knowledge, such a signature has not been described in melanomas. Strikingly, all private deletions were clonal, contrasting other private mutations in these samples (Supplementary Figure 12c). The finding of this signature in both subsequent samples located well outside the radiation field strongly favors the hypothesis of secondary spread, indicating the cells from the radiated submandibular area, and not the calvarian primary lesion (Supplementary Figure. 12a), to be the most recent common ancestor. However, in another five patients having tumor samples collected ≥6 months after initiation of radiation therapy, we did not observe a similar mutational signature. While we could not detect any mutations in DNA repair genes in the tumor tissues of patient MM85, it remains likely that this tumor may harbor particular defects conducive to the development of signature mutations in response to ionizing radiation.

Sequence of genetic alterations during melanoma development. The relative timing of genomic events occurring throughout cancer progression may be inferred by integrating information about copy number alterations and somatic VAFs[36,37]. The

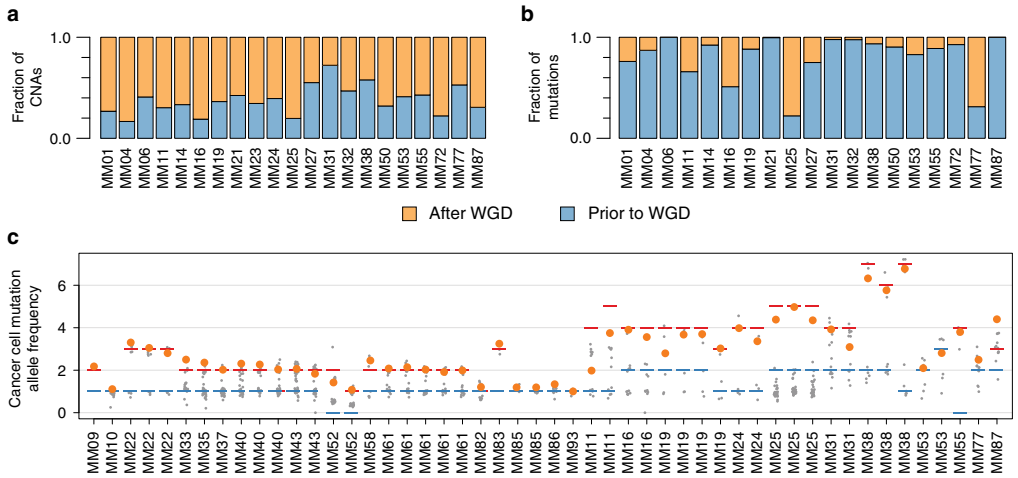


Fig. 5 Timing of genome duplication and gain of mutated *BRAF*. **a**, **b** Estimated fraction of copy number events (**a**) and mutations (**b**) in assessable regions of the genome that occurred prior to and after genome duplication. **c** The relative allelic frequency (corrected to show allelic status of mutation) for mutations in the genomic segment harboring the *BRAF* gene in samples with a *BRAF* mutation. Allelic states are shown as red (major allele) and blue (minor allele) lines; mutations are shown as points, where the *BRAF* mutation is colored orange and all others are colored gray

finding of a higher genomic complexity (Supplementary Figure. 2c) and a higher copy number diversity (Fig. 2b) among patients with WGD is consistent with ongoing genomic evolution following WGD[38]. Indeed, the majority of copy number events in patients with WGD was estimated to have occurred after genome duplication ($n = 21$, median: 63%, range: 27–83%; Fig. 5a). Contrasting copy number alterations, most SNVs and indels appeared prior to WGD in most patients (median: 89%; range: 22–100%; Fig. 5b).

BRAF mutations are known to be early events in melanoma[39] and have been associated with an increase in *BRAF* copy number[5,22,40,41]. We observed low-level copy number gains of at least one *BRAF*-containing allele in 21/27 tumors with *BRAF* mutations, compared to four out of 26 in tumors wild type for *BRAF* ($p < 0.001$, Fisher exact test). The copy number gains all comprised broad regions of chromosome 7, except for a single patient harboring a focal (although still low level) gain of the *BRAF* gene. Strikingly, out of the 21 patients with concurrent mutation and copy number increase of *BRAF*, the mutated allele was the one gained in 20 patients (Fig. 5c). We did not observe associations between copy number elevations and driver mutations for any other oncogene, including *NRAS* (Supplementary Figure. 13). Interestingly, when assessing the allele-specific copy numbers of segments carrying *BRAF*, the most parsimonious solution indicated that *BRAF* gains are most likely to occur prior to WGD in eight out of nine informative patients.

Based on the evidence presented, we may postulate a general model for the order of events in the evolution of metastatic melanoma (Fig. 6). This model is characterized by early acquisition of driver mutations in key genes such as *BRAF* and *NRAS* which, in the case of *BRAF*, is usually followed by a gain of the mutated allele. Whole-genome duplication in general occurs as a later event, taking place after most UV-induced mutations, but prior to most copy number alterations. Following divergence of metastases, mutational accumulation is low and shifts away from UV-induced mutations to others, with a fairly consistent mutational rate within each patient.

Discussion

While previous studies have described genomic alterations occurring in melanoma progression[8,42], including regionally advanced disease[9,43,44], limited knowledge exists in regard to distant metastases. To the best of our knowledge, this is the first study systematically exploring genomic heterogeneity in melanoma across multiple distant metastatic deposits.

We found most mutations to be truncal events. This is of relevance to driver mutations in particular, as we found a very low number of these to be heterogeneous in line with observations in regional metastatic disease[9,43,44]. The low number of heterogeneous mutations indicate metastatic divergence to be a late event, resembling recent findings in breast cancer[45]. Taking the observation of a UV-related mutational signature among some branch mutations into account, these findings are consistent with the hypothesis that different metastases may arise from separate late-developing subclones in the primary tumor, although other explanations may not be excluded.

We observed a surprisingly high intrapatient consistency regarding the number of private mutations across individual lesions. Interestingly, a similar phenomenon was recently described in metastatic breast cancer[45]. The finding of this phenomenon across two tumor forms with quite different mutational patterns[32] indicates this to be an intrinsic propensity related to several cancer forms. Moreover, the observation that heterogeneity correlates to *BRAF* mutation status, as was also made by others in primary melanoma[8], further supports the underlying genetic mechanisms associated with this process.

Our data indicate most metastases to have a monoclonal origin, even though we found indications of reseeding in one patient. This somewhat contrasts the findings of Sanborn and colleagues[9], who described reseeding as a more common phenomenon. Notably, many of the tumors from which they uncover shared subclones were locoregional relapses located in close anatomical proximity. Thick and large primary cutaneous melanomas are known to be associated with a substantial risk of locoregional relapse, despite wide margins in surgical excisions[46,47], consistent with local invasion, and it is reasonable to postulate that similar

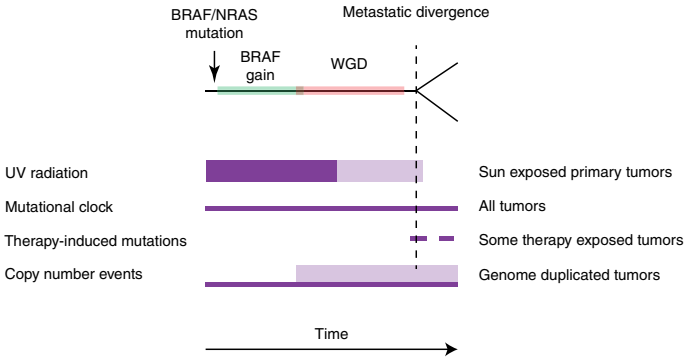


Fig. 6 Model of progression for metastatic melanoma. Purple fields portray the timing of mutational processes, with increased thickness indicating higher mutational activity. Lower opacity indicates variability in timing of processes in relation to each other; e.g., timing of UV radiation in relation to the timing of genome duplication events

processes may regulate the development of locoregional relapses as well.

The patient (MM61) in whom we found indications of reseeding between metastases had an unusual clinical phenotype with numerous (>100 prior to death) cutaneous metastatic deposits on the truncus, shoulders and head area in addition to the 5 lesions sampled (Supplementary Figure. 10c-d). This suggests this cancer to have an organ-specific propensity for the development of cutaneous metastases[48], potentially, including a high migratory potential for metastatic cancer cells within the skin. Thus, while this patient presented with distant metastatic disease, the trafficking of tumor cells might be more akin to the pattern of reseeding observed in regionally disseminated disease[9].

An important topic relates to the sequence of genomic events during cancer progression. We found WGD to occur prior to metastatic divergence, and the high copy number diversity associated with WGD relative to near-diploid tumors suggests an ongoing process of copy number alterations, resembling findings in other tumor forms[45,49]. While genomic complexity is a classic prognostic marker in many tumor forms, how WGD relates to melanoma prognosis remains to be elucidated in larger series.

We found selective low-level gains of the mutated *BRAF* allele as a remarkably common early event in *BRAF*-mutated tumors, generally pre-dating WGD. *BRAF* mutations have previously been described in primary melanoma to be associated with the frequently observed arm- or chromosome-spanning gains of 7q[40], which is consistent with our current results. This likely contrasts *BRAF* gains associated with acquired resistance to BRAF-inhibitors which, when reported, has occurred through focal amplification of smaller segments[10,50]. While it seems reasonable to postulate low-level gain of *BRAF* to provide a selective growth advantage analogous to the fitness-gains associated with low-level gains of mutant *KRAS* in lung cancer[51], this issue warrants more research.

Emergence of the alkylating chemotherapy signature we observed in one patient has been related to DNA mismatch repair defects, with less evidence implicating inactivation of *MGMT* in glioblastoma[33,34,52]. While the signature has been described in melanomas subsequent to temozolomide treatment[32], so far it has not been related to any genomic alterations. Our findings of this signature in a patient harboring several *MSH6* mutations, but not among dacarbazine-exposed patients without mutations, may indicate DNA mismatch repair defects to play a role in melanoma as well.

Ionizing radiation is a well-known carcinogen[53], and secondary cancers arising in areas of previous radiation have been described to reveal a distinct radiation-related mutational signature characterized by an accumulation of small deletions[35]. We found multiple private and truncal 2-nt deletions resembling this pattern of mutations in two distant metastatic deposits 5 and 6 months after radiotherapy for a regional lymph node metastasis. The issue of secondary metastatic spread remains controversial in melanoma[7], as well as in other tumor forms, much due to the fact that it is difficult to find direct evidence for this phenomenon. Chemotherapy exposure should affect tumor cells, including micrometastases, independent of anatomical location; in contrast, radiotherapy is applied to a localized area, with limited radiation scattering outside the treatment field. In this case, we found the radiation signature to constitute a form of "cellular labeling", strongly indicating secondary seeding from the radiation-treated lymph node to the chest wall and liver. While the biological effects of these radiation-induced deletions are unknown, the rapid emergence of two novel deposits <6 months after radiation both characterized by clonal 2-nt deletions should raise concerns that radiation therapy in some cases may enhance metastatic propensity and tumor aggressiveness.

In conclusion, this study provides evidence for common patterns of genomic alterations in melanoma progression. In most cases metastatic deposits seems to have a monoclonal origin with the possible exception of patients harboring multiple cutaneous deposits. The issue of potential secondary spread from metastatic deposits may have significant clinical implications; thus, further studies characterizing melanoma as well as other cancer metastases should seek to identify radiation-induced mutation signatures in all patients having previous exposure to radiotherapy.

Methods

Patients and sample collection. The patients analyzed in this study were part of a single-arm prospective study assessing the response to dacarbazine therapy for metastatic melanoma[16,17]. Out of a total study population of 85 patients, 114 samples from 60 patients and corresponding benign tissue material (blood) were available for analysis by whole-exome sequencing. Samples from all biopsies were examined by a pathologist to ensure representative tissue. Data from 53 individuals (86 samples) are presented; the remaining samples were excluded due to low tumor cell content (<20%). Patient- and sample-level characteristics are detailed in Supplementary Tables 2 and 3, respectively.

All tumor samples were snap-frozen in the operating theater. Peripheral blood was collected at initial biopsy collection.

Ethical approval. The clinical study as well as the genomic analysis was approved by the Regional Ethics Committee of Western Norway (REK Vest; reference

numbers 020/00-109.99, 030/06-06/5520, and 2012/1740). All patients provided written informed consent.

DNA sequencing. Approximately 1 mg of genomic DNA from tumor and matched normal tissue were used for library construction using the Agilent SureSelectXT Human All Exon V5 kit (covering 50 mega-bases of exonic sequence). Libraries were paired-end sequenced using Illumina's TruSeq SBS chemistry v3 on a HiSeq2500, resulting in a median depth of coverage in the targeted regions ranging from 140 to 422 for tumor samples (median across samples: 271), and 43–233 for normal samples (median across patients: 87).

Somatic variant calling pipeline. Reads of each sample were mapped (lane-wise) with BWA mem[54] to the human reference genome (build b37 with an added decoy contig, obtained from the GATK resource bundle). Sample-wise sorting and duplicate marking was performed on the initial alignments with Picard tools (http://broadinstitute.github.io/picard). GATK tools[55] were subsequently used for two-step local realignment around indels, with matching samples (i.e., tumor and its corresponding normal) being processed together. Each sample's pair-end read information was then checked for inconsistencies with Picard and base-quality recalibration was performed by GATK. Somatic variant calling on the matching paired samples was done by using the intersection of MuTect[56] (somatic SNV detection) and Strelka[57] (somatic SNV and indel detection). Block substitutions were defined as somatic mutations at consecutive positions where the variant allelic frequency of each was within 5% of the average allelic frequency of the two variants. The program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control of analysis input data. GATK tools were used for computing coverage statistics based on the recalibrated alignment files. Functional annotation of SNVs and InDels was performed with ANNOVAR (release 2015Mar22), using RefSeq as the gene transcript reference.

Most of the analysis (starting with the local realignment step) was limited to exome regions (the "exome" was in this context defined by Agilent exome v. 5 sequencing probe targets).

Driver mutation definitions. Mutations in a set of genes previously identified as drivers in melanoma[3,21,22] were manually assessed for likely status as drivers. For all considered genes, driver mutations were defined as drivers if they (1) were canonical melanoma-associated mutations; (2) as likely drivers based on evidence of gain or loss of function in the published literature, or if the positions were recurrently mutated in other forms of cancer; or as (3) inactivating if they occurred in tumor suppressors and disrupted the protein reading frame (i.e., nonsense, frameshift, or splice site mutations). Otherwise, mutations were deemed to be passengers (Supplementary Table 4). Patients were categorized according to driver mutation status in *BRAF*, *NRAS*, and *NF1*, where mutations at the canonical mutational hot-spots for *BRAF* (p.V600 or p.K601) and *NRAS* were prioritized in the case of driver mutations in more than one of these genes.

Mutational signature analysis. DeconstructSigs[26] was used to estimate the contribution of mutational processes to the observed patterns of mutations. Contributions from 5 mutational processes that have been described in melanoma were assessed (signatures 1, 5, 7, 11, and 17)[32,58]. Observed mutational patterns were corrected for the 3-base composition of exonic regions in the genome. Signatures reported in the COSMIC database (v79) were used as reference for the mutational pattern associated with each process[58].

For signature analysis of branch mutations we used a lower threshold of $n = 10$ mutations. Given that this number of mutations is too low for precise estimates of percentage contribution to individual signatures, we also performed manual assessment of mutations, focusing on typically UV-related mutations (such as YC>T transitions).

Copy number profiling. Copy number profiling was performed using an in-house algorithm optimized for the present dataset. Our algorithm was established to take advantage of two features in the data:

1. To optimize CNA and tumor purity estimates by use of the observed variant allele frequency of somatic mutations (i.e., to fit CNA estimates on to VAF of SNVs).
2. In the cases with multiple samples per patient, to take advantage of samples with high tumor purity to optimize allele-specific copy numbers across samples within the same patient.

In brief, copy number determination was carried out in three stages: First, segmentation was performed based on shifts in observed allele frequencies of heterozygous SNPs between genomic regions with differences in copy numbers. Second, allele-specific copy numbers across the genome, as well as tumor cell content, were estimated based on the magnitude of shifts in allele frequency of heterozygous SNPs relative to regions with a loss of the minor allele, or based on allele frequency of somatic mutations in the absence of copy number alterations. Third, in patients with multiple samples, cross-sample corrections were made for breakpoint identification and copy number determination based on a combination of germ line and somatic variant allele frequencies. False discovery rates were estimated by simulation, rather than SNP-based benchmarking tools[59], since the current dataset was restricted to WES.

The algorithm was based on the allelic frequency of germ-line variants in tumor and normal samples. Based on the ratio of sequencing depth between tumor and normal, tumor allelic copy numbers uncorrected for normal cell content, the relative copy number (RCN), can be observed. In theory, the interval between RCNs is directly proportional to the difference in number of alleles between adjacent copy number segments. Therefore, the absolute tumor copy number (TCNs) can be determined through inferring the interval of a RCN and the lowest observed RCN value, which normally corresponds to a copy number of zero, or loss of one allele. Based on this, we performed copy number profiling, as follows:

Segmentation. Identification of potential breakpoints: Potential break points were identified based on shifts in allelic frequency of heterozygous SNPs in each tumor relative to the corresponding normal sample across chromosomes. Here, a sliding window approach was used, where the genome was split into bins of 4 Mb, with a step size of 1 Mb. If the number of SNPs in a given bin was <40, the bin was merged with the nearest neighboring bin. For the i-th bin, which included k_i SNPs, we compared the standard deviation of major allele frequency between tumor and normal sample. If there was no difference, the B allele frequency (BAF) of the bin was regarded as 0.5 ($b_i = 0.5$). Otherwise, b_i was defined as the median value of allele frequencies m_i. A potential break point containing region was defined by a difference in BAF exceeding 0.015 between adjacent bins. This cutoff at 0.015 was determined by simulation of randomly generated break points:

At each BAF ranging from 0 to 1, with increments of 0.01, we generated 1000 simulation datasets, each including 40 segments. A randomly assigned number of SNPs was assigned to each segment, ranging from 40 to 1200, and coverage of each SNP followed the distribution of SNPs in the current exome sequencing dataset (geometric distribution; $p = 0.01$). Allelic read counts were modeled using the binomial distribution B(N, BAF), where N was the total sequencing depth of the SNP, and BAF ranged from 0 to 1. For the simulation data corresponding to each BAF we calculated the absolute differences of average BAF from all SNPs between two adjacent segments. In order to determine the significance of difference, we defined an empirical p-value for the likelihood that two segments corresponding to the same theoretical BAF were randomly separated. We estimated the empirical p-value based on the simulation data corresponding to each BAF, and found that a difference of 0.015 corresponded to an empirical p-value of 0.05.

$$p = \frac{\#(\Delta \geq \text{cutoff})_{\text{BAF}}}{(\#(\text{segments}) - 1) \cdot \#(\text{simulation})},$$

where # represented the counts and Δ represented the difference between adjacent segments.

Determining the precise breakpoint and merging of segments: In order to determine a more precise breakpoint between bins, regions flanking the potential breakpoint (±4 Mb) were split into smaller windows, each including 3 SNPs. For each SNP, we used the major allele frequency (m) in the following analysis. The average m value of the first window was compared to the rest of the flanking region. If the difference was more than 0.018, the midpoint of these two sub-regions was regarded as the final breakpoint. The cutoff at 0.018 was determined by estimation of simulation data using the same parameters as above. Based on the simulation data, we found the maximum random error between adjacent segments with the same theoretical BAF was no more than 0.018; although, the random error increased when BAF was closer to 0.5 (Supplementary Figure 14). If a difference of more than 0.018 was not identified by this initial assessment, the window was extended to encompass the second window and compared to the rest of the flaking region. This procedure was repeated until a break point was found, or until the end of the flanking region was reached. If no break points exceeding 0.018 were found, the potential break point was discarded. The genome was thus split into multiple segments according to these final break points, and m values corresponding to each segment were estimated (m^i and $\overline{m_i}$ representing the major allele frequency of segment i and maximum allele frequency of all SNPs in segment i, respectively).

Estimating allelic tumor copy numbers and tumor purity. The relative copy number (RCN) of each allele for each segment can be obtained based on the following formula.

$$\text{CNAH}_i = m^i \times \text{ratio}_i,$$

$$\text{CNAL}_i = \left(1 - m^i\right) \times \text{ratio}_i,$$

where i was the i-th segment. CNAH represents the relative allelic copy number of the major allele, and CNAL represents the relative allelic copy number of the minor allele. Ratio represents the ratio of sequencing depth between tumor and control. For each segment, we estimated the allelic RCNs based on the CNAH and CNAL.

CNALs of all segments were integrated by multiplying it with the number of SNPs in each segment. If the resulting distribution of weighted relative minor allele copy numbers had at least two peaks, we considered the minimal peak value as CN

= 0, and the second as CN = 1 or CN = 2. The distance between copy numbers, DIS, can be calculated based on CN = 0 and CN = 1 or CN = 2.

$$\mathrm{DIS} = \mathrm{CN1} - \mathrm{CN0}$$
$$\mathrm{DIS} = (\mathrm{CN2} - \mathrm{CN0})/2.$$

According to the CN0 and DIS values, allelic TCN of each segment corresponding to each relative copy number state was determined as follows:

$$\mathrm{dCNA}_i = \mathrm{round}\left(\frac{\mathrm{CN}_i - \mathrm{CN0}}{\mathrm{DIS}}\right).$$

Where i represented the i-th segment and round means that values are rounded to the nearest nonnegative whole number. Major (CNH) and minor (CNL) allelic copy numbers were thus calculated.

A tumor sample consists of a mixture of tumor cells and normal cells. For each locus in a chromosome, the expected major allele frequency of SNPs can be calculated as follows:

$$f_{\mathrm{major}} = \frac{\alpha \times \mathrm{CNH} + (1-\alpha)}{\alpha \times \mathrm{CN}_{\mathrm{total}} + 2(1-\alpha)},$$

where α is the tumor cell fraction. Thus, the tumor purity for each imbalanced segment was estimated through the following formula.

$$\alpha = \frac{2 \times f_{\mathrm{major}} - 1}{\mathrm{CNH} - 1 + f_{\mathrm{major}}(2 - \mathrm{CN}_{\mathrm{total}})}.$$

Weighting genomic segments as above, the density peak of purities calculated across segments was used as the tumor purity for each sample. Without sufficient imbalanced segments to obtain tumor purity, we used mutation allelic frequencies in balanced segments (TCN = 2; i.e., mutations on one allele) to estimate tumor purity. The allelic frequency of mutations is given by

$$f_{\mathrm{mut}} = \frac{\alpha \times \mathrm{CN}_{\mathrm{mut}}}{\alpha \times \mathrm{CN}_{\mathrm{total}} + 2(1-\alpha)}.$$

Thus, the tumor purity would be the local peak of mutations density across genome segments with balanced copy number 2. The tumor purity was then inferred from the mutation allele frequency in these segments:

$$\alpha = 2 \times f_{\mathrm{mut}}.$$

Inferring allelic tumor copy number and tumor purity of non-reference sample. For patients from whom multiple samples were analyzed, data from the different samples was used to adjust each other, adding strength to the estimates. In these cases, the sample with the highest tumor purity was coined the "reference sample" while the others were termed "non-reference" samples. Segments in non-reference samples with copy number 0 or copy number 1 were inferred from the corresponding segments in the reference sample. The difference in relative copy numbers (DIS) was estimated based on these segments with copy number 0 and copy number 1. Allele-specific copy numbers were re-evaluated based on the CN0 and DIS estimates, and tumor purity was calculated as shown above. For each non-reference sample, if TCNs of 50% segments differed from the reference sample, we would re-infer the TCN for this non-reference sample in case of genome doubling or tripling.

Estimation of multi-sample tumor allelic copy numbers by clustering of somatic mutations. The estimation of TCNs based on frequencies of mutations can be used to tune the accuracy of copy number calls estimated from SNPs. This approach can be strengthened by use of multiple samples from the same patient. In the present study, such additional tuning was performed for patient MM01, due to the combination of low tumor cell fraction and high ploidy. Thus, we submitted mutations shared between different samples from this patient to K-mean clustering based on variant allele frequencies of mutations in all combinations of the patient´s samples. The number of clusters, k, was defined to select the optimal clustering. Here, the number of clusters resulting in the minimum average sum of squared errors $E(C)$ for k in the range of 2–5 was selected, where $E(C)$ was defined as:

$$E(C) = \frac{\sum_{s=1}^{\binom{n}{2}} \sum_{t=1}^{k} \sum_{o \in C_{ts}} d(o, \mathrm{cen}_{ts})}{\binom{n}{2}},$$

where n was the number of samples, s was the combination of two sample, cen_{ts} was the centroids of cluster t in combination s, and o represented the mutation in cluster t of combination s. The distance d was calculated as Euclidean distance.

The optimal combination of pairwise comparisons of samples based on clustering was regarded as a standard to infer TCNs of each sample from the same patient. For each previously identified segment of the samples, the median value of mutation allele frequencies, corrected for copy number and tumor cell content, mapping into each standard cluster was regarded as the value of the cluster. We determined the optimal combination of two samples based on maximization of inter-cluster distances and minimization of intra-cluster distances. First, the distance between clusters from a combination of two samples was calculated. The distance d between two clusters C_i and C_j was defined as the Euclidean distance between the cluster centroids cen_i and cen_j.

$$d = \sum_{i=j} d\left(C_i, C_j\right) = \sum_{i=j} d\left(\mathrm{cen}_i, \mathrm{cen}_j\right).$$

The combination with maximum clustering distance was retained. In cases with more than one possible combination, the optimal combination of two samples was derived from the minimum average intra-cluster distances between centroids; the intra-cluster distance being defined as:

$$d = \frac{\sum_{t=1}^{k} \sum_{o_i \neq o_j \in C_t} d\left(o_i, o_j\right)}{\sum_{t=1}^{k} \binom{|C_t|}{2}},$$

where $|C_t|$ was the number of mutation cluster C_t. The combination with the minimum intra-cluster distance was regarded as the optimal combination of two samples.

Mutation frequencies of all standard clusters from all segments in the sample were integrated to estimate their probability densities. For any tumor copy number (TCN) state, F, local peak values of mutation frequency distributions were regarded to correspond to specific copy number states, f.

$$F = (f_1, f_2, \ldots, f_n)$$

$$f_1 < f_2 < \ldots < f_n,$$

where f_i was the i-th local peak in mutation frequency distribution. The minimum mutation frequency (f_1) in F was defined as corresponding to copy number of 1. We calculated the interval of TCN as the difference between each f_i and f_{i+1}. Further, based on f_1 and interval of TCN, CNH, and CNL of each segment in the sample were obtained.

Estimation of false discovery rates. To estimate the false positive CNA calls corresponding to the applied cutoff (a difference in BAF of 0.018 between segments), we assumed scenarios where the total copy number in tumor cells ranged from 1 to 8 following a uniform distribution. We simulated 1000 segments (similar with previous simulation process) under different tumor purities ranging from 1 to 100%, with the different total copy numbers (1–8, respectively; Supplementary Figure 15). Based on the segments with the same BAF, combining all tumor purity and total copy numbers, we found the global average false positive rate (FPR) to be 9.88% and the global average false negative rate (FNR) to be 8.44%. The FPR and FNR decreased with the increasing of tumor purity. At tumor purities below 20%, FPR and FNR increased rapidly. Importantly, when the tumor purity was higher than 20%, FPR and FNR was always <10% (Supplementary Figure 15).

Exclusion of samples from analysis. Simulations (see above) introducing different percentages of reads from normal DNA into samples of data from tumor DNA, indicated that aberrant cell fractions higher than 20% would be sufficient for accurately calling copy number alterations. Out of the 114 tumor samples that underwent sequencing, 86 fulfilled this criterion and were used in subsequent analyses.

Inference of whole-genome duplication. For each sample, to infer whether a whole-genome duplication event had taken place, we enumerated the fraction of the genome with a minor allele at copy number 2 and the estimated ploidy. A manual assignment was then performed, based on the assumptions that (1) the overall ploidy of a sample having undergone genome duplication would generally be higher than those of diploid samples, and (2) that the minor allele should be at copy number 2 in at least some fraction of the genome after a whole-genome duplication event (Supplementary Figure 2b).

Mutational heterogeneity between samples. For the analysis of inter-lesional mutational heterogeneity, we considered only mutations whose heterogeneity could not be reasonably be explained by copy number alterations or lack of sequencing depth. Thus, mutations were considered to be potentially heterogeneous if (1) in a sample without a particular mutation, there was no evidence of copy number loss relative to samples carrying the mutation; and (2) the sequencing depth at the position was high enough to have a 95% chance of detecting the mutation given an

allelic fraction of 1 allele out of 4 and the sample-specific tumor cell fraction, assuming a binomial distribution of variant reads. This resulted in a sample-wise depth threshold ranging from 56 for samples with a low aberrant cell fraction, to 18 for samples with a high aberrant cell fraction. In addition, a mutation that was not called by the somatic variant calling pipeline was deemed to be present if the number of reads supporting the mutation was over 1 and higher than what would be expected with an error rate of 1/200, assuming a binomial distribution of supporting reads, with a binomial test p-value of under 0.05. One patient (MM43) exhibited parallel loss of chromosomes 11q and 14 in each of the sampled lesions. Heterogeneous mutations on these chromosomes were considered to have been lost due to copy number alterations.

Calculation of relative VAF and assessment of clonality. As a measure of the cellular prevalence of each mutation, we calculated the relative variant allele frequency (rVAF) of each mutation as the ratio of observed to expected VAF, given local copy number state, tumor cell content and estimated number of mutated alleles:[37]

$$\mathrm{rVAF} = \frac{\mathrm{VAF_{obs}}}{\mathrm{VAF_{exp}}} = \frac{\mathrm{VAF_{obs}}}{\left(\frac{n_{\mathrm{mut}} \times \rho}{2 \times (1-\rho) + n_{\mathrm{tot}} \times \rho}\right)},$$

where n_{mut} refers to the number of mutated alleles, n_{tot} refers to the total copy number at the mutated locus, and ρ refers to the tumor cell content.

Relying on the accuracy of the determination of inter-lesional mutational heterogeneity, we evaluated the clonality of mutations by comparing the rVAF of trunk mutations to that of private mutations to infer likely clonal relationships, using clustering of mutations across samples to validate our findings[31]. Evaluations of mutation clonality were based on the interquartile range (IQR) of rVAF values of trunk mutations only. Thus, mutations were categorized as being subclonal if their rVAF were below the 25th percentile by 1.5 times the IQR, and otherwise as clonal if their rVAFs were above 0.5 times the median rVAF. Mutations not specified as subclonal, and with rVAFs below 0.5 times the median rVAF were considered to be of unknown clonality.

Relative timing of whole-genome duplication. To determine the fraction of copy number events that preceded or followed genome duplication, the shortest route to obtain the observed copy number state for each segment was determined. Here, a copy number change before duplication would lead to a change in observed copy number of two copies from the "unaltered" state of two copies, and a copy number change after genome duplication would lead to a change of one copy. Solving the resulting equation for the minimum number of events, the sum of events occurring prior to and following genome duplication was estimated for each allele in each segment. For each patient, the average number of events across samples was used as a measure of copy number changes prior to and following duplication. To estimate the number of mutations that occurred prior to and following genome duplication, mutations at each allelic state in informative regions of the genome (those with major:minor allele states of 2:2, 2:1 or 2:0) were enumerated. The fraction, m_1, of mutations preceding duplication was estimated as $m_1 = \frac{3n_2}{n_1 - n_2}$ for copy number 2:1, or $m_1 = \frac{n_2}{2n_1}$ for copy number 2:2 and 2:0, where n_1, and n_2 were the number of mutations with allele status 1 and 2, respectively.

Statistical analyses. All statistical analyses were performed in the statistical programming language R (v3.4.1)[60]. Ranked tests were used for comparisons of continuous variables across groups (Mann–Whitney U-tests or Kruskal–Wallis rank-sum tests), or when assessing correlations between continuous variables (Spearman's rank correlation), except if otherwise specified. All significance tests were two-sided, and statistical significance was considered for $p < 0.05$.

Data availability. Raw sequencing data are not publicly available due to national regulations regarding privacy concerns of study participants. Data on somatic mutations are presented in Supplementary Table 1.

Received: 11 October 2017 Accepted: 7 June 2018
Published online: 10 July 2018

References

1. Whiteman, D. C., Green, A. C. & Olsen, C. M. The growing burden of invasive melanoma: projections of incidence rates and numbers of new cases in six susceptible populations through 2031. *J. Invest. Dermatol.* **136**, 1161–1171 (2016).
2. Whiteman, D. C. et al. Anatomic site, sun exposure, and risk of cutaneous melanoma. *J. Clin. Oncol.* **24**, 3172–3177 (2006).
3. Hodis, E. et al. A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
4. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
5. Hayward, N. K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
6. Meier, F. et al. Metastatic pathways and time courses in the orderly progression of cutaneous melanoma. *Br. J. Dermatol.* **147**, 62–70 (2002).
7. Morton, D. L. et al. Sentinel-node biopsy or nodal observation in melanoma. *N. Engl. J. Med* **355**, 1307–1317 (2006).
8. Shain, A. H. et al. The genetic evolution of melanoma from precursor lesions. *N. Engl. J. Med* **373**, 1926–1936 (2015).
9. Sanborn, J. Z. et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc. Natl Acad. Sci. USA* **112**, 10995–11000 (2015).
10. Shi, H. et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. *Cancer Discov.* **4**, 80–93 (2014).
11. Van Allen, E. M. et al. The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov.* **4**, 94–109 (2014).
12. Wagle, N. et al. MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. *Cancer Discov.* **4**, 61–68 (2014).
13. Eigentler, T. K., Caroli, U. M., Radny, P. & Garbe, C. Palliative therapy of disseminated malignant melanoma: a systematic review of 41 randomised clinical trials. *Lancet Oncol.* **4**, 748–759 (2003).
14. Robert, C. et al. Pembrolizumab versus Ipilimumab in advanced melanoma. *N. Engl. J. Med* **372**, 2521–2532 (2015).
15. Chapman, P. B. et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med* **364**, 2507–2516 (2011).
16. Busch, C., Geisler, J., Lillehaug, J. R. & Lonning, P. E. MGMT expression levels predict disease stabilisation, progression-free and overall survival in patients with advanced melanomas treated with DTIC. *Eur. J. Cancer* **46**, 2127–2133 (2010).
17. Jonsson, G. et al. Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin. Cancer Res.* **16**, 3356–3367 (2010).
18. Turajlic, S. et al. Whole genome sequencing of matched primary and metastatic acral melanomas. *Genome Res.* **22**, 196–207 (2012).
19. Rawson, R. V. et al. Unexpected UVR and non-UVR mutation burden in some acral and cutaneous melanomas. *Lab. Invest.* **97**, 130–145 (2017).
20. Dutton-Regester, K. et al. Melanomas of unknown primary have a mutation profile consistent with cutaneous sun-exposed melanoma. *Pigment Cell Melanoma Res.* **26**, 852–860 (2013).
21. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495 (2014).
22. Cancer Genome Atlas, N. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
23. Krauthammer, M. et al. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat. Genet* **47**, 996–1002 (2015).
24. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet* **45**, 1134–1140 (2013).
25. Fujiwara, T. et al. Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature* **437**, 1043–1047 (2005).
26. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
27. Drobetsky, E. A., Turcotte, J. & Chateauneuf, A. A role for ultraviolet A in solar mutagenesis. *Proc. Natl Acad. Sci. USA* **92**, 2350–2354 (1995).
28. Mouret, S. et al. Cyclobutane pyrimidine dimers are predominant DNA lesions in whole human skin exposed to UVA radiation. *Proc. Natl Acad. Sci. USA* **103**, 13765–13770 (2006).
29. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
30. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
31. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
32. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
33. Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
34. Hunter, C. et al. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66**, 3987–3991 (2006).
35. Behjati, S. et al. Mutational signatures of ionizing radiation in second malignancies. *Nat. Commun.* **7**, 12605 (2016).
36. Durinck, S. et al. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* **1**, 137–143 (2011).

37. Purdom, E. et al. Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* **29**, 3113–3120 (2013).
38. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
39. Pollock, P. M. et al. High frequency of BRAF mutations in nevi. *Nat. Genet.* **33**, 19–20 (2003).
40. Maldonado, J. L. et al. Determinants of BRAF mutations in primary melanomas. *J. Natl Cancer Inst.* **95**, 1878–1890 (2003).
41. Modrek, B. et al. Oncogenic activating mutations are associated with local copy gain. *Mol. Cancer Res.* **7**, 1244–1252 (2009).
42. Ding, L. et al. Clonal architectures and driver mutations in metastatic melanomas. *PLoS ONE* **9**, e111153 (2014).
43. Harbst, K. et al. Molecular and genetic diversity in the metastatic process of melanoma. *J. Pathol.* **233**, 39–50 (2014).
44. Harbst, K. et al. Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma. *Cancer Res.* **76**, 4765–4774 (2016).
45. Yates, L. R. et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169 (2017).
46. Leiter, U., Meier, F., Schittek, B. & Garbe, C. The natural course of cutaneous melanoma. *J. Surg. Oncol.* **86**, 172–178 (2004).
47. Urist, M. M. et al. The influence of surgical margins and prognostic factors predicting the risk of local recurrence in 3445 patients with primary cutaneous melanoma. *Cancer* **55**, 1398–1402 (1985).
48. Nguyen, D. X., Bos, P. D. & Massague, J. Metastasis: from dissemination to organ-specific colonization. *Nat. Rev. Cancer* **9**, 274–284 (2009).
49. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
50. Shi, H. et al. Melanoma whole-exome sequencing identifies (V600E)B-RAF amplification-mediated acquired B-RAF inhibitor resistance. *Nat. Commun.* **3**, 724 (2012).
51. Kerr, E. M., Gaude, E., Turrell, F. K., Frezza, C. & Martins, C. P. Mutant Kras copy number defines metabolic reprogramming and therapeutic susceptibilities. *Nature* **531**, 110–113 (2016).
52. Johnson, B. E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
53. Leuraud, K. et al. Ionising radiation and risk of death from leukaemia and lymphoma in radiation-monitored workers (INWORKS): an international cohort study. *Lancet Haematol.* **2**, e276–e281 (2015).
54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
55. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
56. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
57. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
58. Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
59. Zare, F., Dow, M., Monteleone, N., Hosny, A. & Nabavi, S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinform.* **18**, 286 (2017).
60. R Core Team. (R Foundation for Statistical Computing, Vienna, Austria, 2016).

Acknowledgements

Parts of this work were performed in the Mohn Cancer Research Laboratory. The work was funded by the Norwegian Research Council through the Norwegian Cancer Genomics Consortium (NCGC; grant numbers 218241 and 221580), The Norwegian Health Region West, The Norwegian Cancer Society and The Bergen Research Foundation. We thank Dagfinn Ekse for technical assistance.

Author contributions

Patient recruitment: J.G. and P.E.L. Curation of clinical data: E.B., J.G., and P.E.L. Data analyses: E.B., S.Z., D.P., S.N., D.V., L.A.M.-Z., and E.H. Provided research infrastructure: O.M. and P.E.L. Study design: E.B., S.Z., J.G., S.K., and P.E.L. Manuscript writing: E.B., S. K., and P.E.L.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05063-1.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

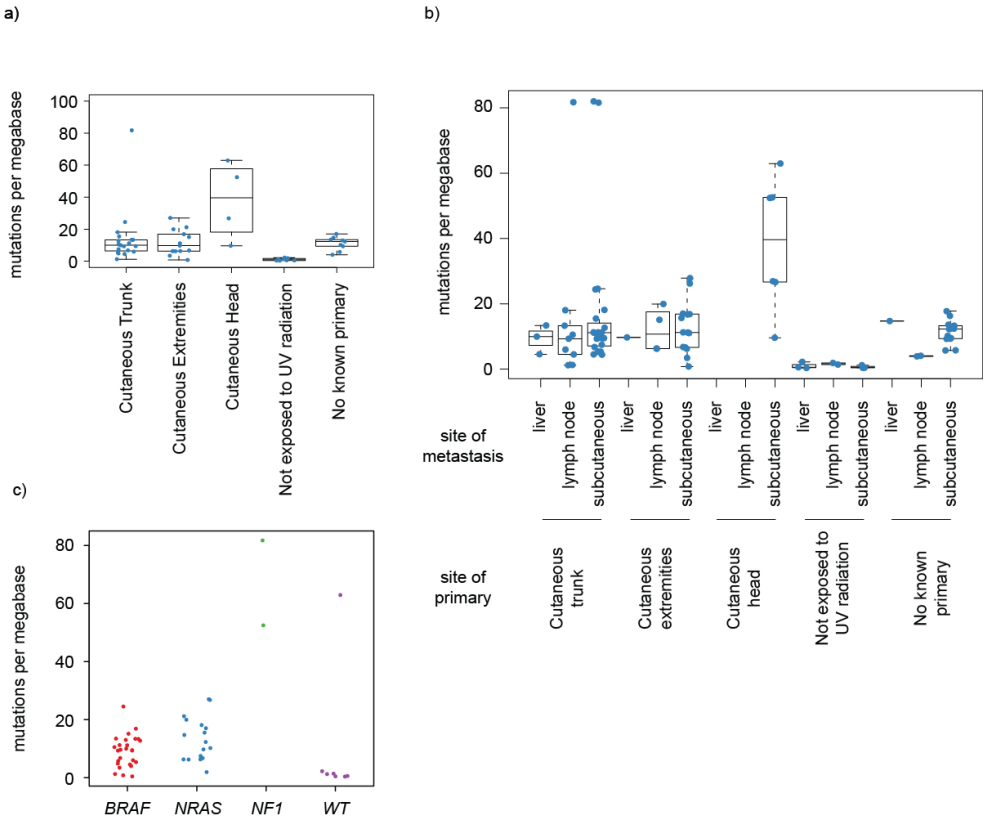
Supplementary Information

Patterns of genomic evolution in advanced melanoma

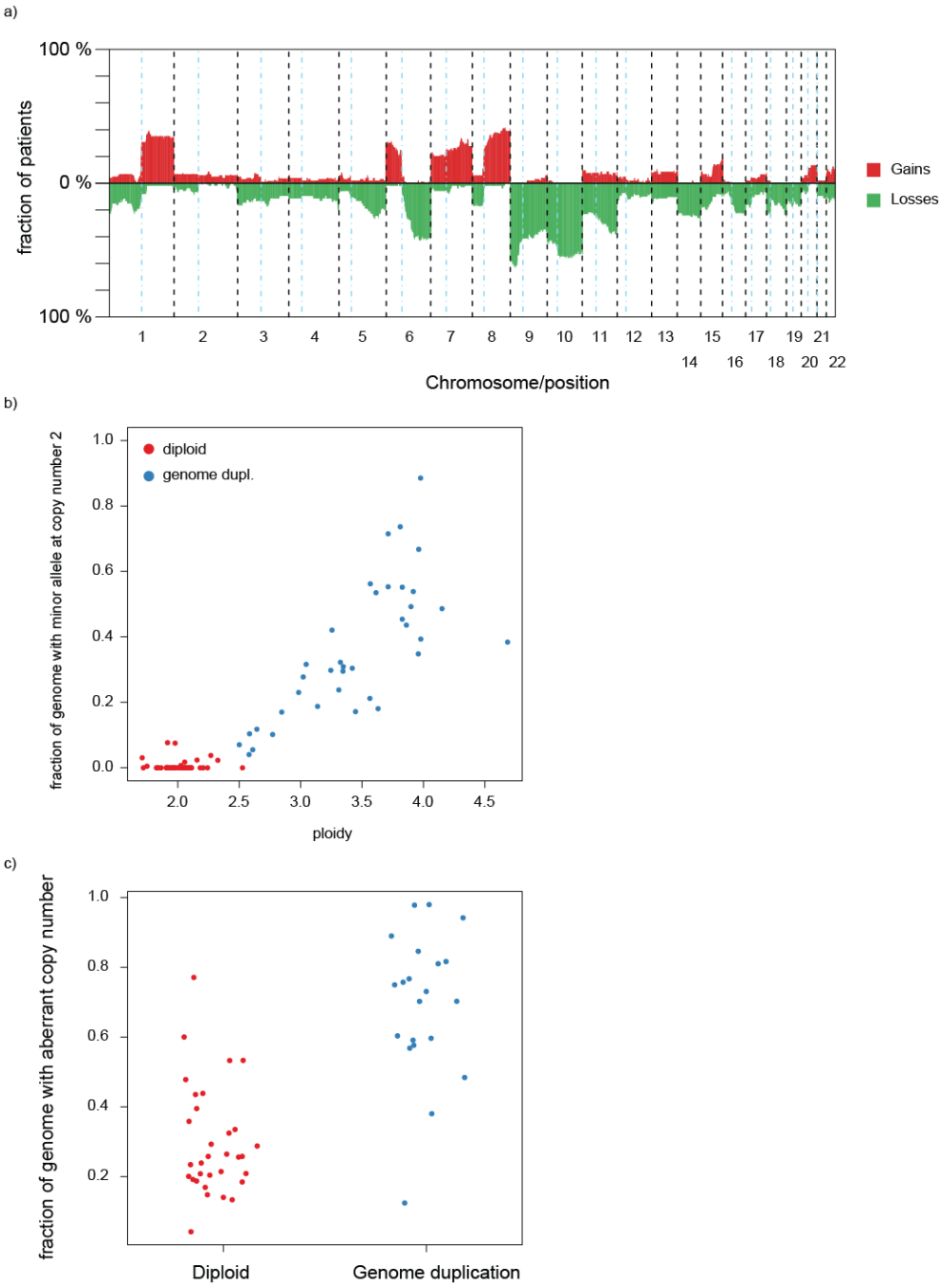
Birkeland et al.

Supplementary Figures 1-15

Supplementary Figures

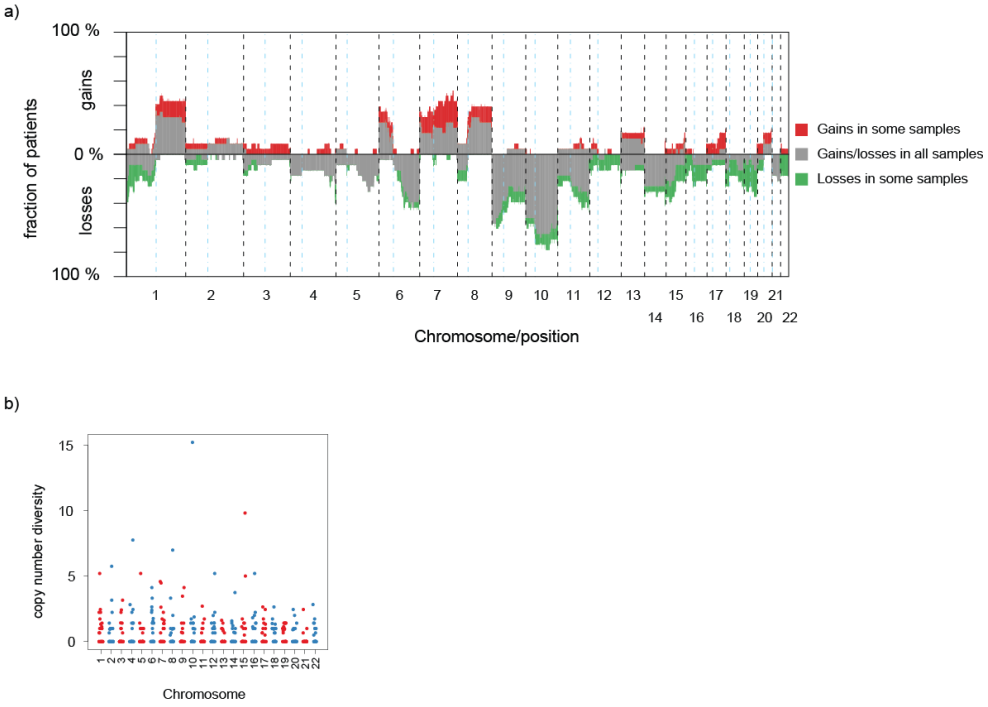


Supplementary Figure 1: Unique mutation per patient: **a**) number of mutations per patient (average of samples) in coding regions per megabase according to site of the primary lesion. The category "Not exposed" includes mucosal (n=2), acral (n=3), uveal (n=2), and one patient with a skin lesion that was situated perianally. **b**) The number of mutations per lesion according to the site of the lesion the patient's primary tumor. Boxes with whiskers span the interquartile range (IQR), with whiskers extending to 1.5 times the IQR from the upper and lower bounds of the boxes. **c**) The number of unique mutations per patient according to driver mutation status of *BRAF*, *NRAS* and *NF1*.

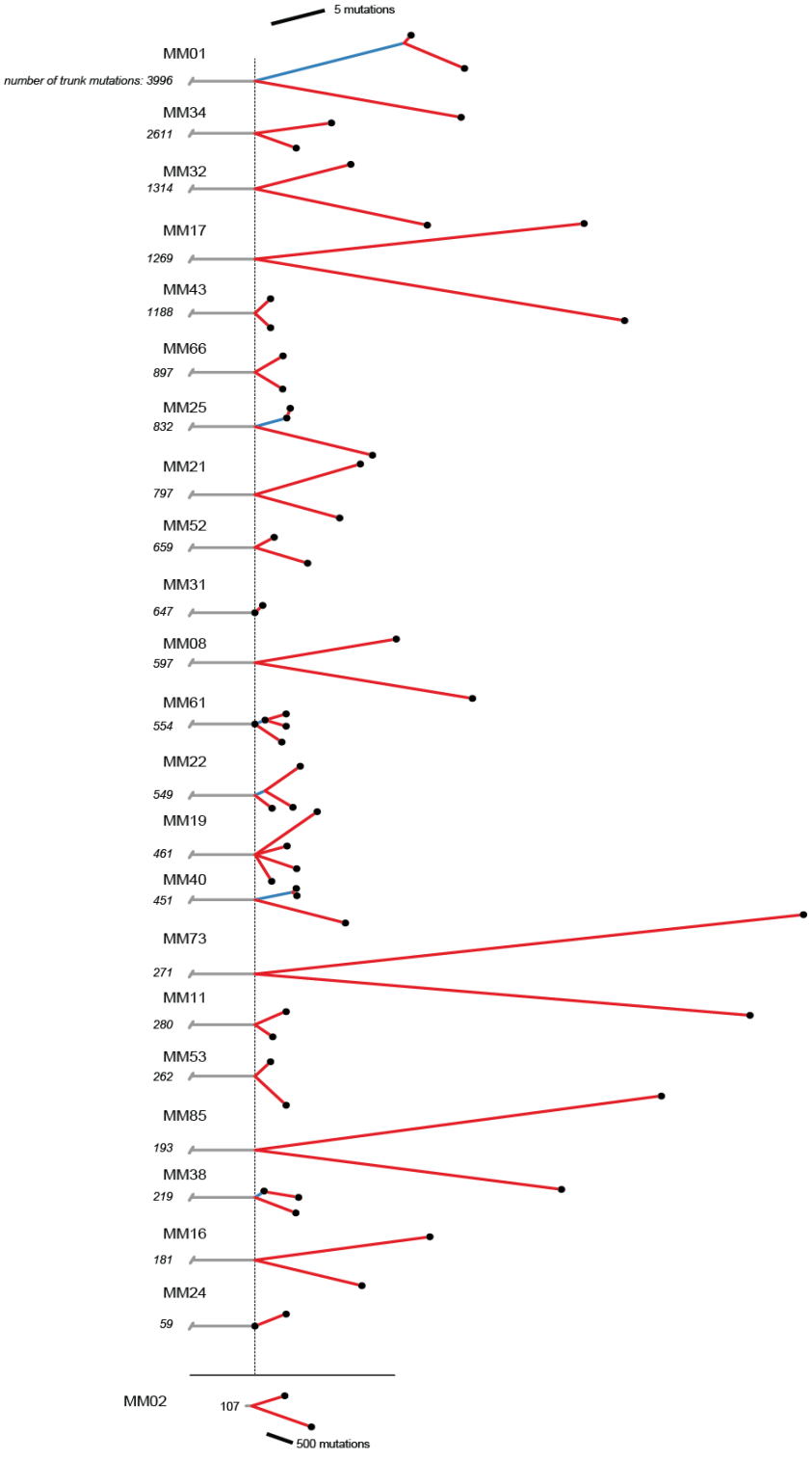


Supplementary Figure 2: Genomic complexity and genome duplication. **a**) Prevalence of copy number gains and losses across the genome. Gain or loss for any region was defined as ≥3 copies in total, or ≤1 copy in total, respectively. For patients with genome duplication, the respective thresholds were ≥6 and ≤2. For patients with more than one analyzed sample, the

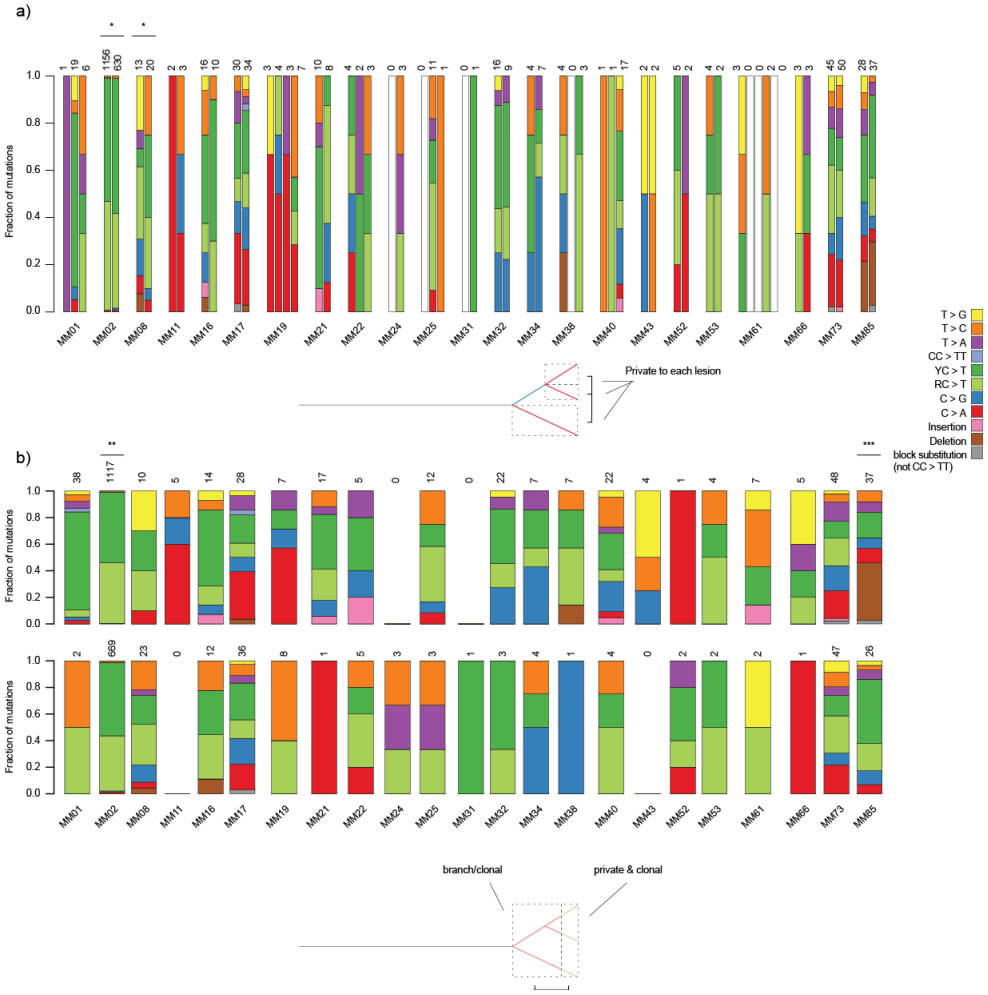
fraction of samples with gain or loss was recorded. **b**) Using the fraction of the genome where the minor allele was at copy number 2 in combination with ploidy, we categorized patients as having undergone a genome duplication event (blue) or not (red). **c**) Measuring genomic complexity as the fraction of the genome not at a balanced copy number of 2 (diploid) or 4 (genome duplicated), we compared patients with diploid tumors to patients with tumors having undergone genome duplication. For patients with multiple samples, we used the average value of the patient's samples.



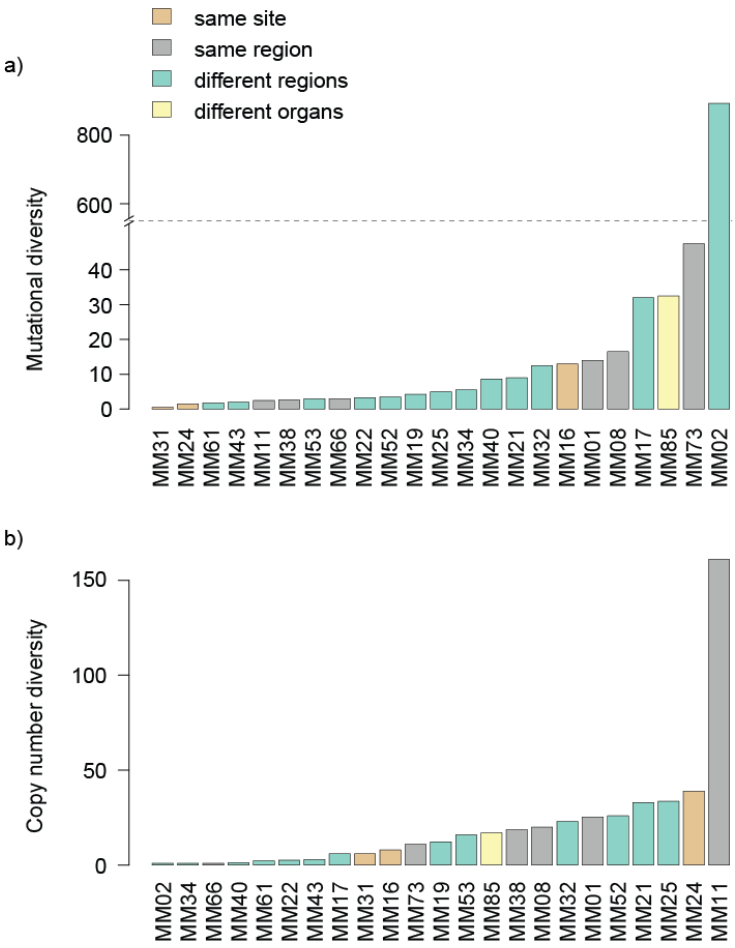
Supplementary Figure 3: Copy number diversity. **a**) Prevalence of heterogeneous copy number gains and losses across the genome. Gains and losses are defined as in figure S2. Grey fields represent the fraction of patients with gains or losses in all individual samples. Red and green fields represent the fraction of patients with heterogeneous copy number gains and losses, respectively. Only patients with multiple sampled lesions are included in this figure. **b**) The copy number diversity according to chromosome; each patient is represented by a point per chromosome.


5 mutations
MM01
number of trunk mutations: 3996
MM34
2611
MM32
1314
MM17
1269
MM43
1188
MM66
897
MM25
832
MM21
797
MM52
659
MM31
647
MM08
597
MM61
554
MM22
549
MM19
461
MM40
451
MM73
271
MM11
280
MM53
262
MM85
193
MM38
219
MM16
181
MM24
59
MM02
107
500 mutations

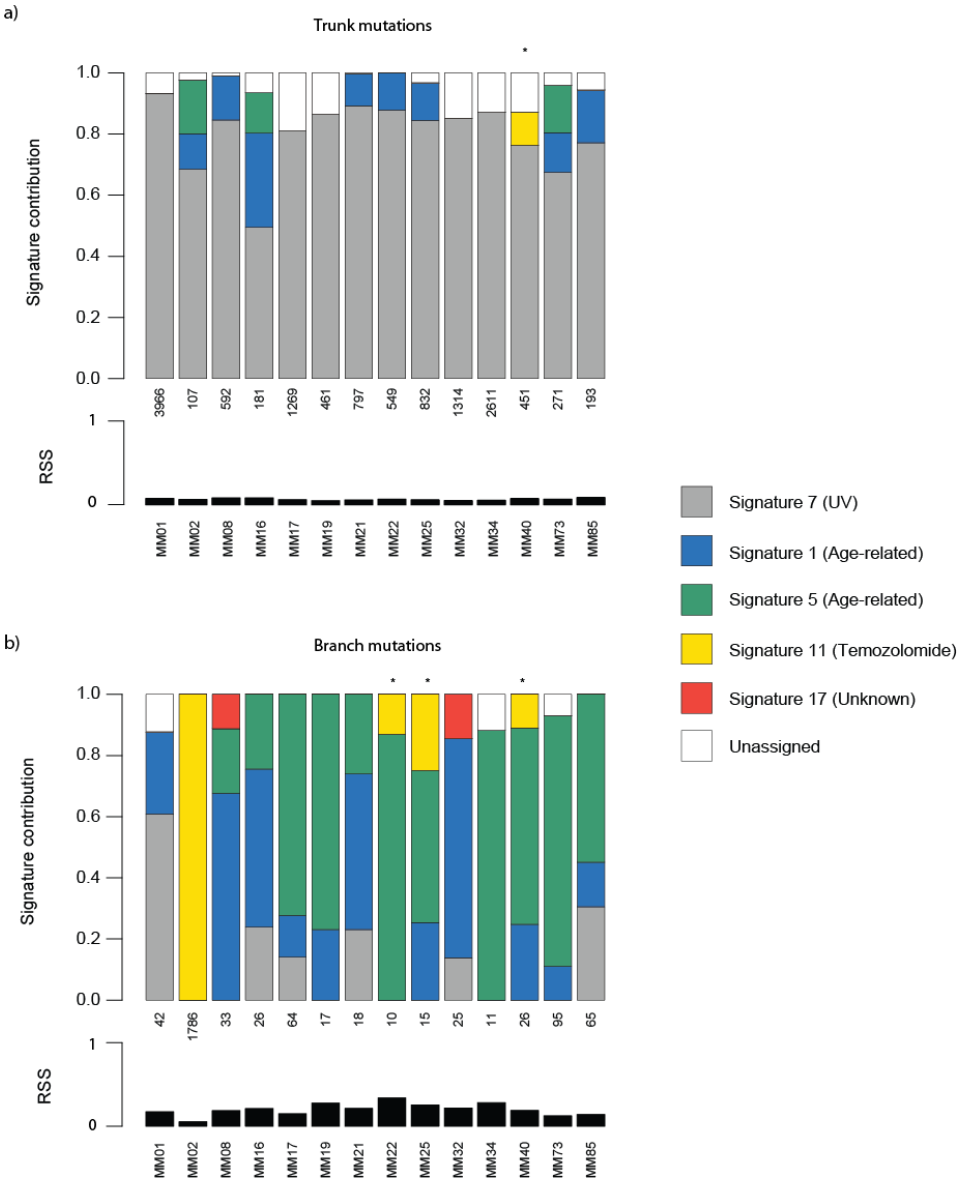
Supplementary Figure 4 (previous page): Phylogenetic trees showing the relationships between samples collected from the same patients. The phylogenetic trees are derived from the presence or absence of mutations across samples as depicted in figure 2b and c in the main text. Sampled lesions are indicated by black dots, and the color of branches match the color code in figure 2b and c: grey = trunk; blue = branch; red = private mutations. The trunks of the phylogenetic trees have been truncated, with the total number of mutations indicated next to the base of each tree. Branch lengths are proportional to the numbers of mutations specific to each branch, with the number of mutations indicated by a scale bar. Patient MM02 is depicted separately due to the high number of branch mutations in the samples from this patient.



Supplementary Figure 5: Mutation type distribution of branch mutations for each sample. **a**) Private mutations per lesion in each patient, or (**b**) branch mutations per patient according to status as subclonal are compared as portrayed in diagrams below barplots. Blank columns represent samples without mutations in the relevant category. The numbers of mutations are shown above each column. Asterisks indicate significant differences according to fisher exact test (or chi-square tests in the case of high numbers of mutations).

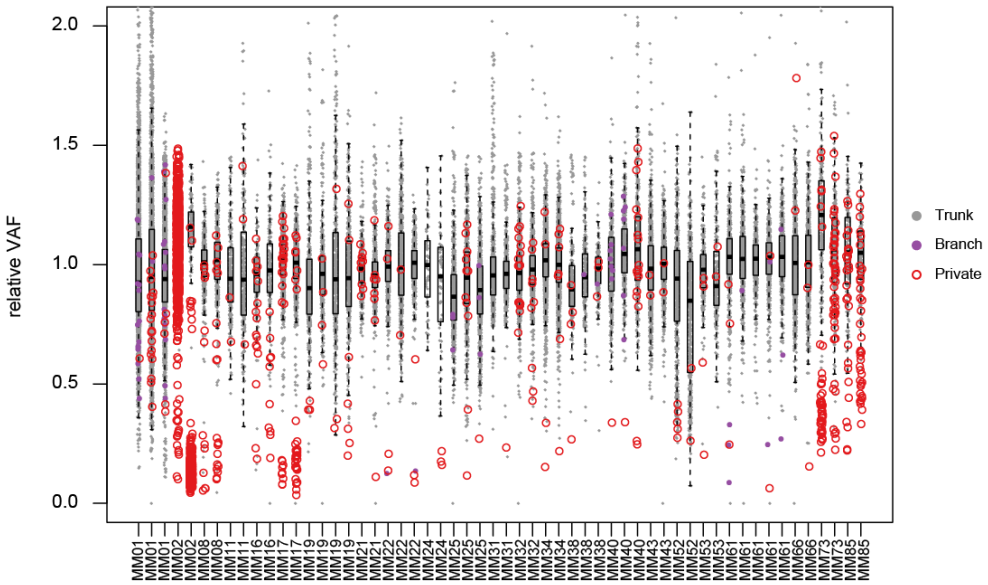


Supplementary Figure 6: mutational (**a**) and copy number (**b**) diversity according to the anatomical diversity of sampled lesions. Patients were categorized according to the anatomical distance between biopsy sites. For patients with more than two samples, the largest distance was used. Anatomical distance was categorized as same site, with samples taken from the same lesion at different time points; same region, defined as lesions in areas draining to the same lymph nodes; different regions; or different organs. Subcutaneous and lymph node deposits were not considered as separate organs by this classification.

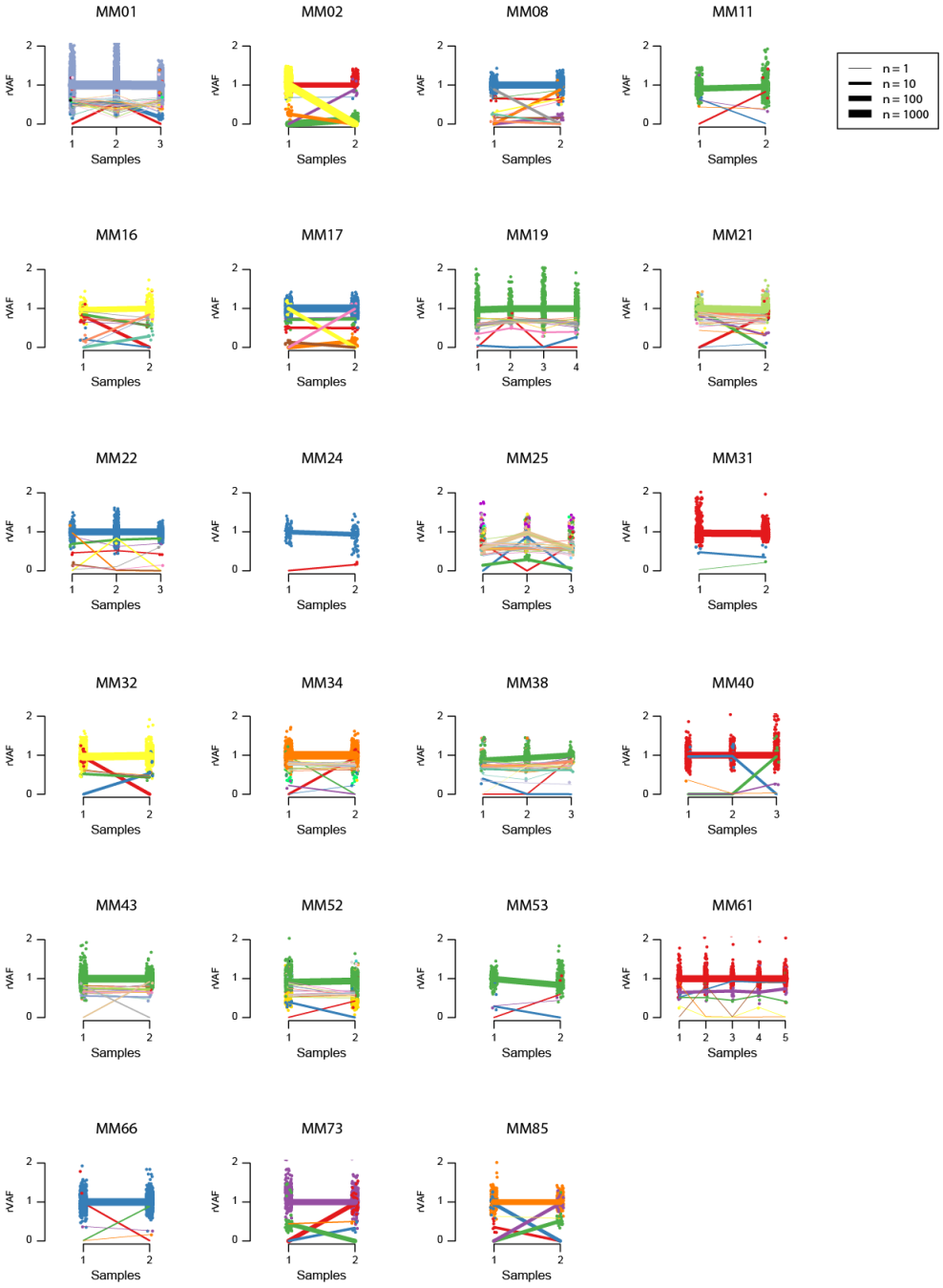


Supplementary Figure 7: Estimated contributions of mutational processes to (**a**) trunk and (**b**) branch mutations for each patient in which the total number of branch mutations was over 10 (the number of trunk or branch mutations is shown under each bar). The residual sum of squares (RSS) is shown below, which is a measure of how closely the observed mutations match the estimated process contributions. Only contributions of 5 mutational processes were assessed, and mutational signatures corresponding to less than 10% of mutations were not considered; thus, for some patients, the mutational signature contributions do not sum to 1.0. *Although some mutations in these patients were predicted

to be caused by alkylating agent exposure, only two of these samples had been exposed to such therapy, each predicted to have less than five mutations assigned to this source. We therefore consider this attribution to be by chance.

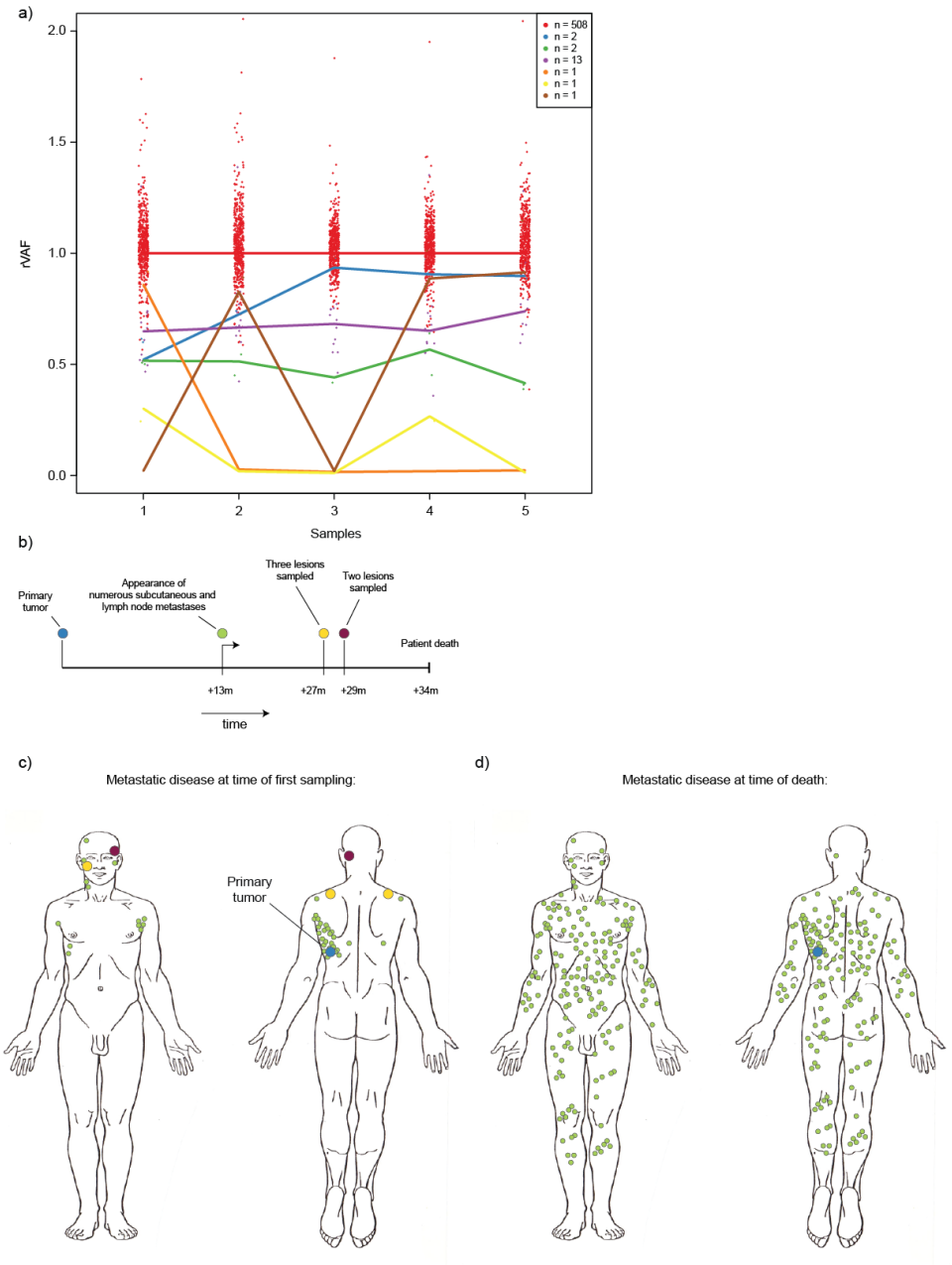


Supplementary Figure 8: Relative variant allele frequency (VAF) of mutations according to heterogeneity and subclonality status. Relative VAF represents the variant allele frequency corrected by tumor purity and local copy number, and is thus analogous to cellular prevalence. Mutations are color-coded according to status as trunk (found in all lesions; gray), branch (found in some, but not all lesions; purple), and private (found uniquely in one lesion; red). Boxes span the interquartile range and whiskers extend to 1.5 times the interquartile range from the bounds of the boxes.



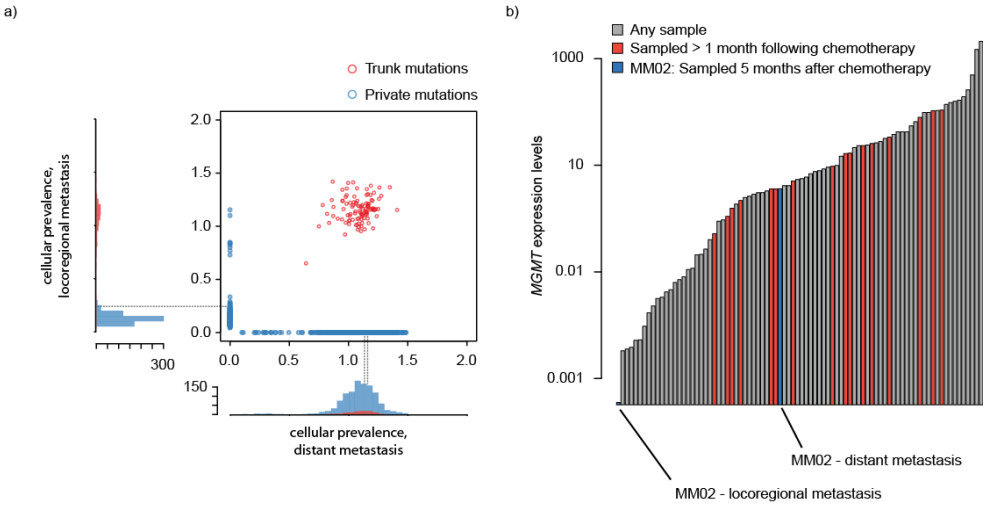
Supplementary Figure 9: Cross sample clustering of mutations according to cellular prevalence using PyClone. Mutations and their relative variant allele frequencies (rVAF) are indicated for each sample with dots, colored according to the cellular population to which

they were predicted to belong. The predicted cellular prevalence of each population of cells is indicated by lines, the weight of which correspond to the number of mutations belonging to each population.

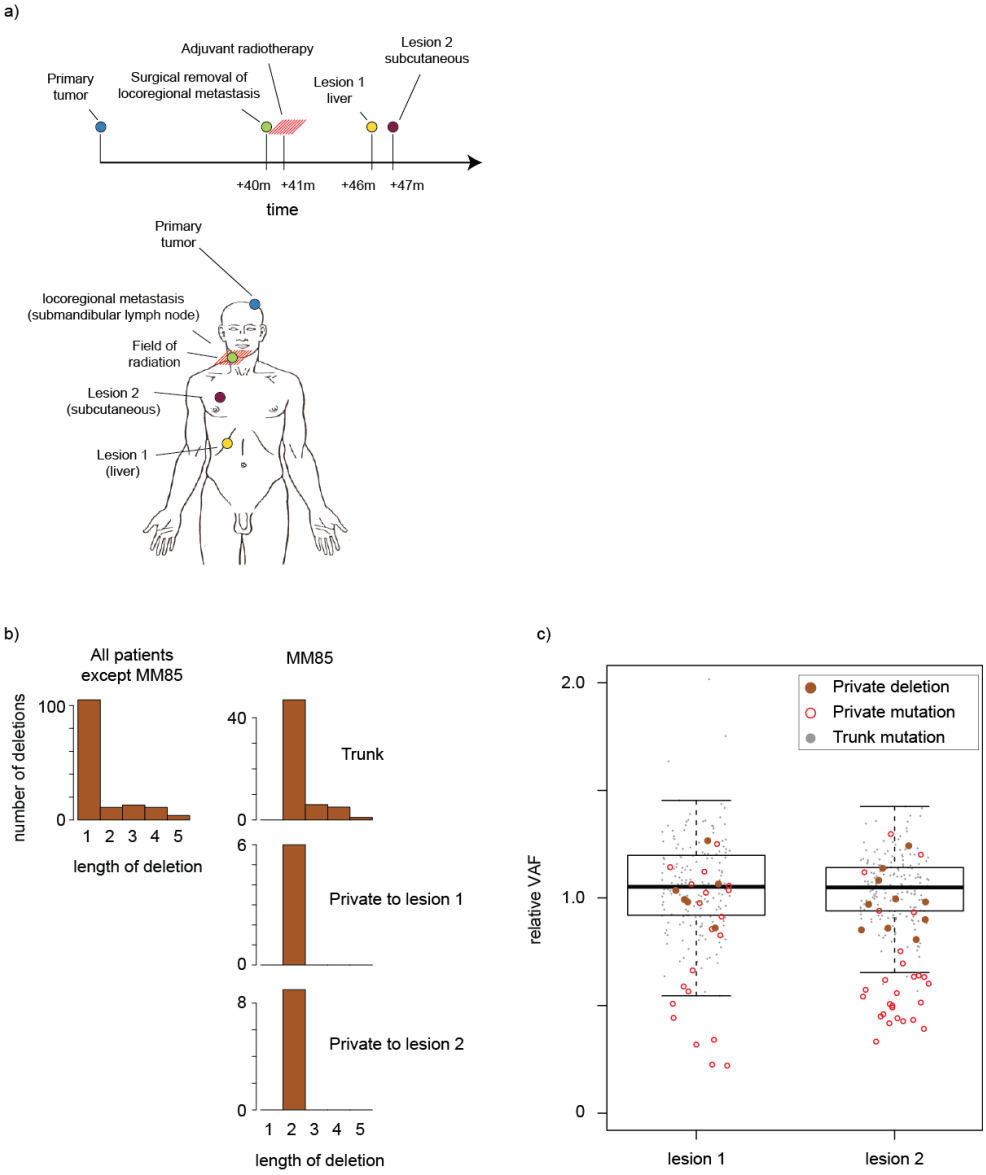


Supplementary Figure 10 (previous page): Mutations with a recurrent subclonal rVAF could indicate reseeding between lesions. **a**) Cross sample clustering of mutations according to cellular prevalence[1] as in Supplementary Fig. 9. The number of mutations on which the inference of each cell population is based, are shown in a panel in the top right corner. **b-d**) Patient MM61 had an unusual disease course, characterized by an extensive and rapid spread of cutaneous metastases, first described 13 months following surgical excision of the primary lesion. At the time point of sampling (c) of the first metastatic lesions (sample 1-3)

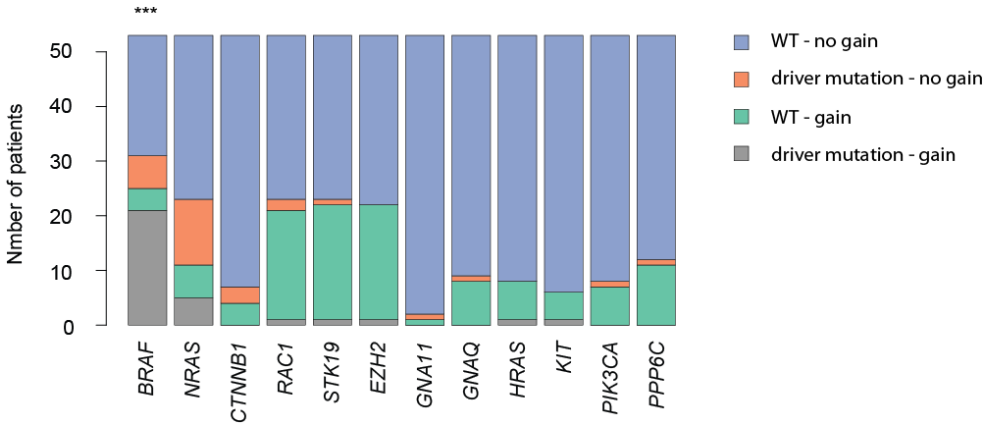
for the current study, regional and distant skin metastases were present on the truncus, neck and upper extremity. Wide-ranging cutaneous progression occurred in the months leading up to the patient's death, 7 months later (d).



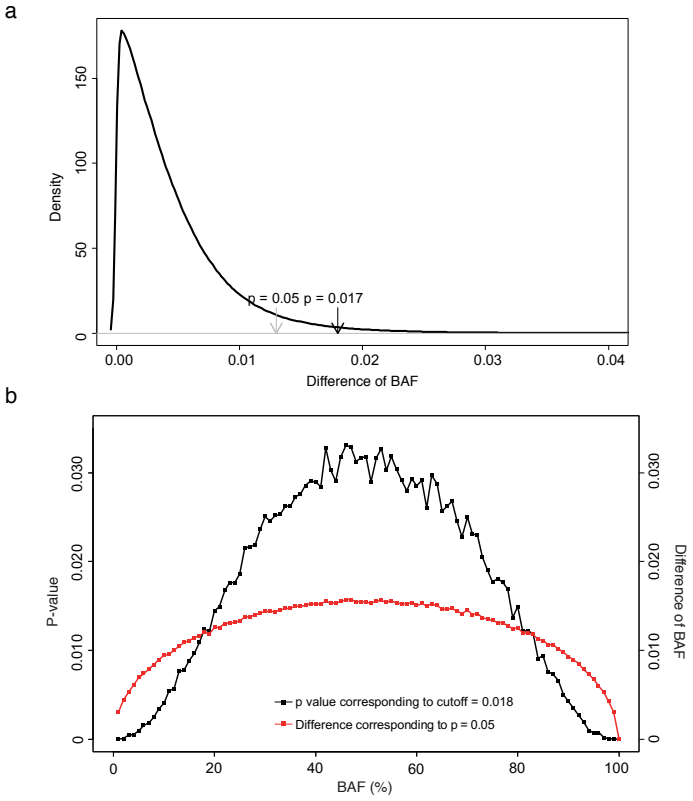
Supplementary Figure 11: *MSH6* and *MGMT* deficiency in patient MM02. **a**) Cellular prevalence of mutations in each sample from MM02. The lower and left panels show the frequency distributions of private (blue) and trunk (red) mutations in each sample. These correspond to clonal populations and each major peak of mutations coincides with private *MSH6* mutations, which are indicated with stapled lines. Private mutations in both samples conformed to the dacarbazine signature. **b**) Expression levels of *MGMT* (O-6-methylguanine-DNA methyltransferase) mRNA relative to those of *B2M* (beta-2-microglobulin) were measured for each sample, and compared to the relative *MGMT* expression levels in the melanoma cell line Sk-Mel-28. Samples collected in excess of one month following dacarbazine exposure are shown in red and samples from MM02 in blue. All other samples are shaded gray. Both samples from MM02 displayed a mutational signature consistent with dacarbazine treatment. One sample from MM02 displayed a loss of *MGMT* expression, whereas the other was found to have hypermethylation of the *MGMT* promoter (results not shown). Expression levels and promoter methylation status of *MGMT* for this patient cohort have been published previously [2].



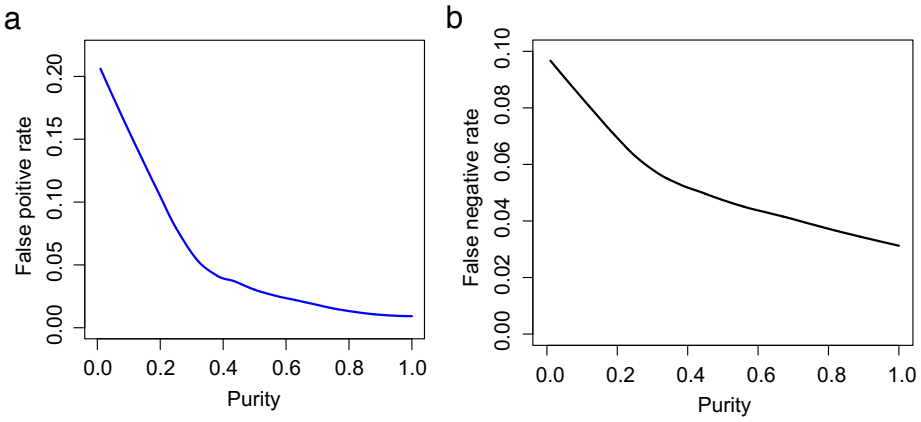
Supplementary Figure 12: Mutational pattern of 2-nt deletions in patient MM85. **a**) The upper panel shows the timing of disease progression and radiation treatment (m = months). The lower panel shows the localization of each lesion and the field of radiation. Lesion 1 and lesion 2 were sampled for the current study. **b**) The frequency of deletions according to the length of deletions in all patients except MM85 (upper left), and deletions in MM85; trunk deletions, or deletions private to each lesion (top to bottom, right). **c**) Relative VAF of trunk mutations (gray), private mutations other than deletions (red circles), and private deletions (brown). Boxes are based on the trunk mutations only (as in Supplementary Figure 8).



Supplementary Figure 13: Frequency of gains of activating driver mutations. Patients are categorized according to whether there is evidence of an increase in copy number of one allele relative to a balanced state of each gene, and whether a driver mutation is identified in each gene. Asterisks indicate a significant difference according to a fisher exact test ($p < 0.001$).



Supplementary Figure 14: Estimation of empirical p-values for random difference in *f* values between segments exceeding the chosen threshold (0.018). (a) Probability density plot of global random differences (x-axis). The gray arrow indicates the *f* value difference corresponding to the global empirical p value 0.05. The black arrow indicates the BAF difference of 0.018 corresponding to the global empirical p value 0.017. (b) Plot of local empirical p values for the cutoff at each specific *f* value. The x-axis shows *f* values ranging from 0.01 to 1, with with an increment of 0.01. For the black line, the y- axis is the local empirical p value corresponding to the difference 0.018 in each specific *f* value. For the red line, the y-axis is the difference of *f* value corresponding to the local empirical p value was 0.05.



Supplementary Figure 15 The false positive rate (FPR) and false negative rate (FNR) of copy number calling, as functions of tumor purity.

Supplementary References

1 Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396-398, doi:10.1038/nmeth.2883 (2014).

2 Busch, C., Geisler, J., Lillehaug, J. R. & Lonning, P. E. MGMT expression levels predict disease stabilisation, progression-free and overall survival in patients with advanced melanomas treated with DTIC. *Eur J Cancer* **46**, 2127-2133, doi:10.1016/j.ejca.2010.04.023 (2010).

Poduval et al. Clinical Epigenetics (2020) 12:131
https://doi.org/10.1186/s13148-020-00920-7

Clinical Epigenetics

RESEARCH Open Access

Assessment of tumor suppressor promoter methylation in healthy individuals

Deepak B. Poduval[1,2], Elisabet Ognedal[1,2,3], Zuzana Sichmanova[1,2], Eivind Valen[4,5], Gjertrud T. Iversen[1,2], Laura Minsaas[1,2], Per E. Lønning[1,2] and Stian Knappskog[1,2*]

Abstract

Background: The number of tumor suppressor genes for which germline mutations have been linked to cancer risk is steadily increasing. However, while recent reports have linked constitutional normal tissue promoter methylation of *BRCA1* and *MLH1* to ovarian and colon cancer risk, the role of epigenetic alterations as cancer risk factors remains largely unknown, presenting an important area for future research. Currently, we lack fast and sensitive methods for assessment of promoter methylation status across known tumor suppressor genes.

Results: In this paper, we present a novel NGS-based approach assessing promoter methylation status across a large panel of defined tumor suppressor genes to base-pair resolution. The method omits the limitations related to commonly used array-approaches. Our panel includes 565 target regions covering the promoters of 283 defined tumor suppressors, selected by pre-specified criteria, and was applied for rapid targeted methylation-specific NGS. The feasibility of the method was assessed by analyzing normal tissue DNA (white blood cells, WBC) samples from 34 healthy postmenopausal women and by performing preliminary assessment of the methylation landscape of tumor suppressors in these individuals. The mean target coverage was 189.6x providing a sensitivity of 0.53%, sufficient for promoter methylation assessment of low-level methylated genes like *BRCA1*. Within this limited test-set, we detected 206 regions located in the promoters of 149 genes to be differentially methylated (*hyper*- or *hypo*-) at > 99% confidence level. Seven target regions in gene promoters (*CIITA*, *RASSF1*, *CHN1*, *PDCD1LG2*, *GSTP1*, *XPA*, and *ZNF668*) were found to be *hyper*-methylated in a minority of individuals, with a > 20 percent point difference in mean methylation across the region between individuals. In an exploratory hierarchical clustering analysis, we found that the individuals analyzed may be grouped into two main groups based on their WBC methylation profile across the 283 tumor suppressor gene promoters.

Conclusions: Methylation-specific NGS of our tumor suppressor panel, with detailed assessment of differential methylation in healthy individuals, presents a feasible method for identification of novel epigenetic risk factors for cancer.

Keywords: Methylation, Epimutations, Cancer risk, Promoter, Massive parallel sequencing

* Correspondence: stian.knappskog@uib.no
[1]K.G. Jebsen Center for Genome Directed Cancer Therapy, Department of Clinical Science, University of Bergen, Bergen, Norway
[2]Department of Oncology, Haukeland University Hospital, Bergen, Norway
Full list of author information is available at the end of the article

Introduction

The number of tumor suppressor genes for which germline mutations are linked to elevated cancer risk is steadily increasing [1–3]. Mutations across different genes present a continuum of penetrance, ranging from moderately to massively elevated risk of different cancer forms. Further, while mutations in some genes (so far) are restricted to increased risk of a single, or a few cancer forms, mutations in other genes may increase the risk of multiple different types of cancer [4, 5].

Some of the best described "classical" high penetrance genes include *BRCA1/2*, for which germline mutations are associated with an elevated risk of ovarian and breast cancer [6], *MLH1/MSH2* linked to colorectal cancer [7], *CDKN2A* and *RB1*, associated with melanoma and retinoblastoma, respectively [8–10], as well as *TP53*, associated with the Li-Fraumeni syndrome with an elevated risk for multiple cancer forms [11]. However, the list of genes for which germline mutations are ascertained to confer cancer risk is continuously increasing due to application of massive parallel sequencing [12, 13]. Still, for many families with multiple cases of a specific tumor form (like breast, ovary, or melanomas), no pathogenic germline gene variant has been identified.

Epigenetic gene inactivation may occur through different mechanisms [14, 15]. So far, promoter methylation is the best studied of all the epigenetic modifications, and such methylation is well established as a mechanism of inactivation of tumor suppressor genes. While many germline mutations affecting tumor suppressor genes are well studied as cancer risk factors, knowledge regarding constitutional epigenetic inactivation [16] as a potential cancer risk factor remains limited. Somatic promotor methylation in tumor suppressor genes is a common event in cancer [17], but the role of aberrant epigenetic events, or constitutional promoter methylation of tumor suppressor genes in normal cells as potential cancer risk factors, remains largely unexplored. While mosaic methylation of the *MLH1* gene in normal leukocytes has been observed in colorectal cancer patients [18, 19] and a haplotype leading to secondary constitutional methylation in the *MGM2* promoter [20] has been found in a cancer-prone family [21], in general, data on normal tissue methylation patterns and cancer risk are scarce [22].

Recently, in a large study, we reported low-grade mosaic (< 10% of alleles) normal tissue *BRCA1* promoter methylation to confer a significantly increased risk of high-grade serous ovarian cancer (HGSOC) [23]. In our study, we found > 4% of healthy adult females in a Caucasian population to harbor mosaic *BRCA1* promoter methylation in their normal white blood cells (WBC). Individuals carrying such methylation had a 2-3 fold increased risk of HGSOC. Importantly, WBC *BRCA1* promoter methylation was strongly associated with corresponding methylation in other normal tissues, and, in HGSOC patients, also associated with methylation in the tumor. Taken together, this indicated that methylated normal cells in the ovary may act as tumor precursors.

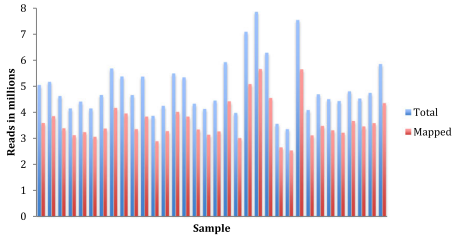
Based on these results and the findings of others [19, 24–29], we hypothesized that additional tumor suppressors could be hyper-methylated in normal cells, thereby causing an elevated risk for certain cancer forms within subgroups of healthy individuals in the general population [30].

To explore such a hypothesis, there is a need for improved methodologies. Although methylation status may be analyzed by conventional arrays, such assessments are limited to the selection of CpGs covered by the array probes. These selected CpGs may not necessarily represent all the CpGs crucial for gene silencing [23]. An alternative is methylation-specific whole genome sequencing, but this remains prohibitively costly. In the present study, we aimed to establish, and provide proof-of-concept for, a novel strategy assessing the full CpG spectrum across promoter areas of tumor suppressor genes. The assay applies methylation-specific massive parallel sequencing of the promoter areas of a panel of 283 tumor suppressor genes. We show the feasibility of the method by depicting promoter methylation variation across the promoter panel in a set of white blood cell (WBC) DNA obtained from 34 healthy individuals. Further, by performing an exploratory hierarchical clustering, our findings indicate that the profiles of normal cell promoter methylation of tumor suppressor genes fall into two main clusters defined by differences in genes regulating key biological pathways.

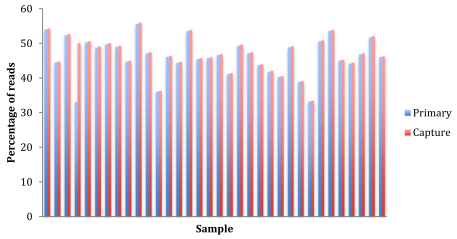
Results

Methylation specific sequencing

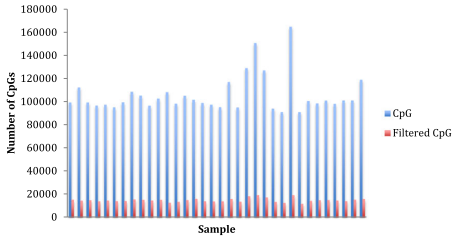
We analyzed WBC DNA from 34 healthy individuals. After bisulfite conversion of the DNA, we performed methylation-specific sequencing of 565 capture regions representing 356 target regions from 283 tumor suppressor gene promoters (the full list of genes and regions is presented as Supplementary Table S1). Sequencing was performed on an Illumina MiSeq, running 8 samples per run. Regarding average values per sample, we obtained 4.95×10^6 reads (range 3.36-7.85 $\times 10^6$) (Fig. 1a; for details per sample see Table 1). Subsequent to quality filtering, 88% of the reads, were retained. Thus, after filtering, 4.30×10^6 reads were attempted mapped to the genome, yielding 4.08×10^6 mapped single reads. Out of these, 3.6×10^6 reads mapped with properly paired reads for each sample (average values; Fig. 1a). These reads led to a mean primary target coverage of 189.6x (114.8x-269.5x) and a mean capture target coverage of 199.4x (120.7x-283.4x). Every sample had almost equal



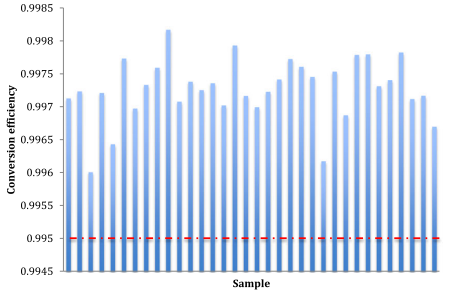
a: Output reads from methylation specific targeted sequencing. Bars indicate the output yield in terms of number of reads (millions) for each of the analysed individuals. Blue bars indicate the total number of reads, while red bars indicate number of reads mapping to the genome.



b: Percentage of mapped reads on primary (blue) and capture target regions (red) for each sample.



c: Number of CpGs called in the analysed samples. Blue bars show identified CpGs and red bars show CpGs with >10 reads in all samples and a non-zero methylation ratio.



d: Bisulfite Conversion Efficiency (> 99.5 % recommended; dotted line) for each sample.

Fig. 1 a Output reads from methylation specific targeted sequencing. Bars indicate the output yield in terms of number of reads (millions) for each of the analyzed individuals. Blue bars indicate the total number of reads, while red bars indicate number of reads mapping to the genome. **b** Percentage of mapped reads on primary (blue) and capture target regions (red) for each sample. **c** Number of CpGs called in the analyzed samples. Blue bars show identified CpGs and red bars show CpGs with > 10 reads in all samples and a non-zero methylation ratio. **d** Bisulfite conversion efficiency (> 99.5% recommended; dotted line) for each sample

percentage of reads mapped to capture targets and primary targets (Fig. 1b).

The overall number of informative CpGs identified for each sample were on average 1.0×10^5 (range 0.9×10^5 to 1.6×10^5). Restricting the CpGs to those with a methylation ratio > 0, and more than 10 reads in coverage, the number was reduced to 1.5×10^4 (range 1.1×10^4-1.9×10^4; Fig. 1c).

We defined the sensitivity of our strategy as 1/x, where x = sequencing depth at any given CpG. With the average primary target depth being 189.6x, the sensitivity was 0.53%. In theory, the fragility of this sensitivity estimate lies in that, for some samples, the results may depend on a single read, rendering them more sensitive to artifacts such as inadequate bisulfite conversion. However, assessing the bisulfite conversion rate (C to T) of the internal Lambda DNA control (see the "Methods" section), we found the conversion efficiency to be on average 99.7% (range 99.6-99.8%) across the analyzed samples (Fig. 1d). This indicates a rate of technical artifacts (falsely retained C's instead of T's) to be lower than 0.2-0.4%, thus approaching the error rate in the sequencing per se (Q30 threshold).

Reproducibility was assessed in a separate standard sample (pooled DNA from 5 healthy donors) that was run in 6 parallels per run over 2 independent runs. In a selection of 12 out of the 565 regions, we found the mean coefficient of variation to be 7.1% (median 4.4%; Supplementary Table S2). As such, the technical variability in this standard sample was considerably lower than the detected biological variation (see below) in our study set of 34. Variability was considerably lower when assessing all CpGs in a region than when limiting analyses to randomized selections of CpGs within the regions (e.g., for *PRDM2*, the coefficient of variation was 1.5% when considering all CpGs while it was on average 4.7% when assessing randomized selections of 5 CpGs within the region).

Methylation landscape of tumor suppressors

For each sample, we calculated the mean methylation for each of the 565 capture regions based on individual CpG methylation ratios within each actual region (see the "Methods" section for details). We observed large

Table 1 Summary of samples and analyses

Sample	Input reads	% reads post QC	% reads mapped	Reads (paired and clipped)	% reads on target (primary)	% reads on target (capture)	Coverage on target (primary)	Coverage on target (capture)
10046_S2	5043578	90.39	78.71	3588074	54.01	54.31	209.99	220.84
10071_S7	5167738	89.08	83.62	3849392	44.54	44.81	186.40	196.12
10077_S3	4622790	90.13	81.18	3382228	52.35	52.66	198.17	208.34
10078_S8	4148304	89.91	83.59	3117970	33.13	50.13	175.26	184.27
10081_S4	4408742	89.53	81.99	3236368	50.31	50.60	186.37	195.79
10082_S5	4146244	88.90	82.83	3053058	48.82	49.10	171.00	179.62
10086_S6	4665150	89.39	80.91	3373894	49.76	50.06	186.09	195.68
10088_S2	5683572	88.12	83.26	4169990	49.01	49.29	219.04	230.48
10097_S3	5372752	87.46	84.09	3951544	44.74	45.00	195.13	205.17
10107_S1	4659718	89.67	80.24	3352898	55.68	56.00	213.95	224.78
10110_S6	5369964	86.87	82.21	3835038	47.17	47.45	199.76	210.11
10113_S7	3862862	89.44	83.54	2886124	36.11	36.31	120.48	126.52
10117_S5	4243470	89.93	85.80	3274194	46.15	46.42	169.10	177.78
10126_S7	5490894	87.68	83.36	4013294	44.39	44.65	198.79	209.00
10131_S1	5338282	87.00	82.62	3836988	53.58	53.88	232.54	244.35
10146_S6	4326882	90.13	85.57	3337118	45.47	45.73	168.34	177.00
10149_S7	4129130	90.01	84.46	3139126	45.78	46.04	159.55	167.77
10155_S5	4450890	86.31	84.99	3264796	46.65	46.91	170.60	179.37
20011_S4	5923516	87.27	85.49	4419628	41.19	41.42	203.82	214.20
20019_S1	3972578	89.46	84.62	3007020	49.39	49.68	163.80	172.27
20022_S4	7091616	88.48	81.02	5083584	47.19	47.47	257.88	271.25
20023_S1	7854298	88.16	81.78	5662728	43.72	43.98	269.47	283.35
20024_S5	6284900	88.04	82.23	4549698	41.85	42.09	215.05	225.92
20062_S2	3554848	88.54	84.13	2648098	40.36	40.54	127.76	133.78
20068_S3	3355298	89.51	84.38	2534292	48.89	49.18	138.43	145.54
20078_S2	7541786	88.33	84.85	5651984	38.94	39.16	246.67	259.20
20088_S4	4083996	90.28	84.33	3109332	33.27	33.46	114.79	120.70
20092_S3	4686060	89.01	83.38	3477852	50.58	50.88	196.38	206.39
20098_S1	4501632	88.21	83.23	3305210	53.58	53.89	197.69	207.78
20106_S2	4430250	86.96	83.43	3214408	45.12	45.32	173.69	181.84
20117_S4	4802994	88.70	86.01	3664364	44.22	44.48	178.18	187.37
20119_S5	4525080	89.19	85.57	3453670	46.90	47.16	180.35	189.59
20122_S6	4741150	88.79	85.18	3585870	51.76	52.07	204.62	215.16
20160_S3	5848296	88.30	84.30	4353508	46.02	46.28	220.04	231.29

inter-region variation in the methylation levels of the regions within the 283 tumor suppressor gene promoters analyzed (Fig. 2). Some regions were completely methylated (e.g., regions within the promoters of *AIP*, *PRDM2*, *ATR*, *DICER1*, *SFPQ*), while others in general were non-methylated in most individuals (e.g., regions within *ARID2*, *TRIM33*, *SETD2*, *IKZF1*, and *ARID1B*; Supplementary Figure 1).

In some regions, there was a large variability between CpGs within the promoter region, indicating that some CpGs may be constitutively methylated, while others (perhaps more crucial for transcriptional regulation) had a lower methylation level and may be more dynamically methylated (Supplementary Figure 2).

Constitutional promoter hyper-methylation has been classified either as secondary due to a rare genetic/SNP variant [16], typically resulting in high methylation levels [31, 32] or primary, in which case, methylation may occur at a low mosaic level (VAF of < 10%) [23]. As for both cases, we may not expect identifying several

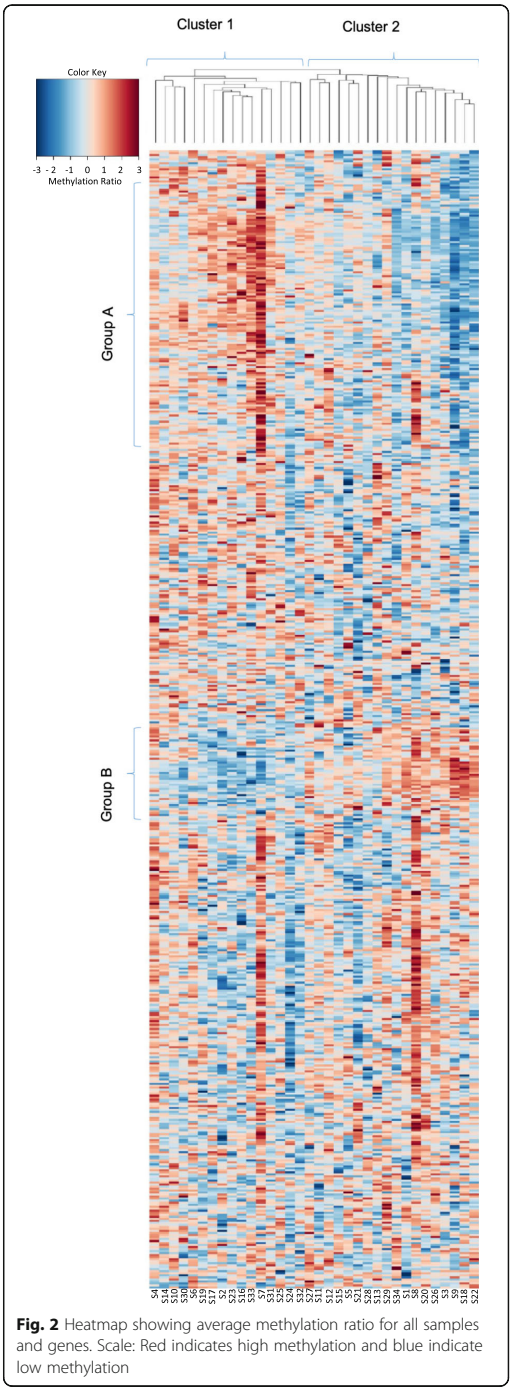


Fig. 2 Heatmap showing average methylation ratio for all samples and genes. Scale: Red indicates high methylation and blue indicate low methylation

affected individuals in a small dataset like the present; thus, lack of differential methylation here may not exclude a gene as a potential epigenetic pathogenic factor. Still, to validate the feasibility of our method, we aimed at exploring potential differential methylation between individuals across our data set. To do so, we took three approaches: first, we assessed differential methylation across the dataset in general. Second, we specifically assessed for individual hyper-methylation, assuming this to be the most relevant alteration regarding inactivating tumor suppressors. Third, we specifically assessed those tumor suppressors where previous data have linked promoter methylation to cancer risk.

Differential methylation

Subsequent to methylation calling, we identified promoter regions differentially methylated across our sample set. Although low levels of methylation (allele methylation frequency of < 5%) have been shown to affect cancer risk [23], in the present sample set we focused on identifying those genes presenting the largest inter individual methylation variation as a proof-of-concept for our methodological approach. We defined methylation variation in a region according to the difference in absolute but also relative methylation level. First, we assessed the difference in absolute methylation as the difference in percentage of alleles methylated between individual (i.e., difference presented as percent points). Second, we assessed the relative difference between individuals, i.e., the ratio between the highest and lowest methylated individual with respect to percentage of methylated alleles.

Based on a Z-score assessment of a methylation matrix consisting of averaged methylation ratios for each of the 565 capture regions across all 34 samples (see the "Methods" section for details), we identified 206 regions (within the promoters of 149 genes) where a minority (one-third or less) of the samples analyzed were significantly differentially methylated as compared to the majority of samples at a ≥ 99% confidence level (i.e., outside the 99% confidence interval; Supplementary Table S3). Assessing the difference between the samples with the highest and the lowest level of methylation within these 206 regions, about half of the regions (n = 101) displayed less than 5 percent point difference. However, several of the tumor suppressor regions displayed a large variation in methylation, with 72 regions displaying > 10 percent point difference and 22 regions displaying > 20 percent points difference between the highest and the lowest methylated samples (Table 2). The largest difference was observed for *GAS7*, where the difference between the highest and the lowest methylated sample was 66.6 percent points.

Table 2 Differentially methylated genes. Gene regions with > 20 percent points difference in methylation ratio, between least methylated sample to most methylated sample along with fold change differences are listed. Hyper-methylated target region of those genes are shown in bold

Gene name	Gene capture region	Min. methylation ratio	Max. methylation ratio	Difference in methylation ratio	Fold change
GAS7	chr17: 10199716 - 10200316	0.2670	0.9332	0.6662	3.4951
ELAC2	chr17: 13019069 - 13019845	0.4581	0.8332	0.3751	1.8188
GSTM1	chr1: 109686327 - 109687046	0.6438	1.0000	0.3562	1.5533
THBS1	chr15: 39579298 - 39579871	0.4671	0.7885	0.3214	1.6881
CIITA	**chr16: 10874982 - 10875928**	**0.2511**	**0.5577**	**0.3066**	**2.221**
RASSF1	**chr3: 50339388 - 50340021**	**0.1786**	**0.4720**	**0.2934**	**2.6428**
CHN1	**chr2: 174846842 - 174848034**	**0.2141**	**0.5074**	**0.2933**	**2.3699**
MSH2	chr2: 47401613 - 47402319	0.5897	0.8734	0.2838	1.4811
PALB2	chr16: 23642511 - 23643136	0.6333	0.9134	0.2801	1.4423
RUNX3	chr1: 24964233 - 24965550	0.3920	0.6479	0.2559	1.6528
TP63	chr3: 189789769 - 189790448	0.6612	0.9059	0.2446	1.3701
PDCD1LG2	**chr9: 5510022 - 5511326**	**0.3182**	**0.5511**	**0.2330**	**1.7319**
AIP	chr11: 67481632 - 67482276	0.6716	0.9002	0.2286	1.3404
GPC3	chrX: 133986729 - 133987434	0.5842	0.8036	0.2194	1.3756
AIP	chr11: 67482202 - 67482880	0.1035	0.3214	0.2180	3.1053
GSTP1	**chr11: 67581895 - 67582976**	**0.2673**	**0.4834**	**0.2162**	**1.8085**
AIP	chr11: 67481257 - 67481869	0.7857	1.0000	0.2143	1.2728
XPA	**chr9: 97698585 - 97699193**	**0.6991**	**0.9130**	**0.2139**	**1.306**
APC	chr5: 112736082 - 112736959	0.6361	0.8479	0.2118	1.333
CTCFL	chr20: 57524096 - 57527440	0.6554	0.8663	0.2109	1.3218
CASP8	chr2: 201259179 - 201260169	0.3698	0.5799	0.2102	1.5681
ZNF668	**chr16: 31064314 - 31065859**	**0.4513**	**0.6584**	**0.2070**	**1.4589**
--	--	--	--	--	--
RABEP1	**chr17: 5281240 - 5283045**	**0.0415**	**0.1033**	**0.0618**	**2.4902**
AIP	**chr11: 67482382 - 67483805**	**0.0517**	**0.1229**	**0.0712**	**2.3783**
RASSF1	**chr3: 50338258 - 50339618**	**0.0976**	**0.2178**	**0.1202**	**2.2322**
FOXO4	chrX: 71094692 - 71096928	0.1256	0.2592	0.1335	2.0629
ZRSR2	chrX: 15789350 - 15791219	0.0559	0.1021	0.0462	1.8252
RUNX1T1	**chr8: 92102449 - 92105016**	**0.0558**	**0.1008**	**0.0450**	**1.8068**
RHOH	**chr4: 40196452 - 40197679**	**0.1059**	**0.1914**	**0.0855**	**1.8067**

Assessing the relative difference (ratio between the highest and lowest methylated sample), again *GAS7* was the top-ranking promoter, showing a relative difference of 3.5 fold between the highest and the lowest methylated sample. As expected, in addition to *GAS7*, we found a substantial overlap between top-ranking regions based on absolute differences and the top-ranking regions based on relative differences (ratio) in methylation levels (Table 2). Especially the *AIP* gene also had a region that was highly differentially methylated both in terms of percentage difference (> 20%) and fold difference (> 3 fold). The only regions with less than 20 percent point difference but a high fold difference (> 2 fold), were regions in *RABEP1*, *RASSF1*, *AIP*, and *FOXO4* (Table 2, lower section).

Hyper-methylated tumor suppressors

Regarding tumor suppressor genes, we hypothesized that in case constitutional methylation is associated with a significantly elevated cancer risk, we may expect a minor sub-fraction of healthy individuals to have hyper-methylated promoters. We therefore performed additional sub-analyses restricting 206 genes identified above, to the genes/region with positive Z-scores with > 99% confidence level, i.e., genes/regions that were

significantly hyper-methylated in a minority of individuals as compared to the majority of individuals (see the "Methods" section). Among the 206 differentially methylated regions, 115 revealed positive Z-scores. Out of these 115, 25 displayed > 10 percent points difference from the highest to the lowest methylated sample. The corresponding number of regions revealing > 20 percent points difference was 7. These 7 regions were within the promoters of *CIITA*, *RASSF1*, *CHN1*, *PDCD1LG2*, *GSTP1*, *XPA*, and *ZNF668*, with the three former genes revealing a difference of more than 30 percent points (Table 2). Re-assessing these data based on fold difference instead of percent points, we identified three regions (in *AIP*, *RABEP1*, and *RASSF1*) with a lower than 20 percent point absolute difference but a relative ratio > 2. Since another region of *RASSF1* was already identified as having a difference > 20 percent points, this left us with 9 different genes with substantial differences in methylation levels.

Further, we reasoned that if methylation of any of these genes may act as a cancer risk factor, then somatic methylation of the same genes should be present in a fraction of human cancers. We therefore mined the COSMIC data base [33] for reported somatic methylation of the 9 genes. Six of these genes (*CHN1*, *PDCD1LG2*, *XPA*, *ZNF668*, *RABEP1*, *AIP*) were not reported to be aberrantly somatically methylated in tumors, while one gene (*CIITA*) was reported to be *hypo*-methylated in a very small fraction (0.19-1.53%) of various solid tumors. In contrast, somatic *hyper*-methylation of *RASSF1* was reported in > 4% of endometrial cancers and > 1% of breast cancers. Further, somatic *hyper*-methylation of *GSTP1* was reported in > 7% of prostate cancers and > 1% of breast cancers. Thus, this finding indicates that some genes found hyper-methylated in tumor tissue are also differentially methylated in normal tissue of healthy individuals. Although these data do not provide any conclusive evidence per se, the findings warrant further investigations exploring constitutional methylation as a potential cause of cancer risk.

Methylation in established cancer risk genes

Among some of the best-characterized cancer risk genes in terms of mutations (*BRCA1*, *TP53*, and *RB1*), we found the mean methylation level to be 0.7% in the known regulatory region of the *BRCA1* promoter, in line with our previous findings [23]. For *TP53*, the mean methylation level was 7.9%, while the corresponding number for *RB1* was 24.9%. For some additional genes where methylation has been found as a cancer risk factor, *MLH1* and *MGMT*, these revealed mean methylation levels of 6.4% and 18.6%, respectively. Among these established cancer risk genes (*BRCA1*, *TP53*, *RB1*, *MLH1*, and *MGMT*), we found no significant differences between the individuals in the present data set.

Co-methylated tumor suppressors

The cause of differential DNA methylation, and, in particular, tumor suppressor promoter methylation, remains poorly understood. Thus, in an exploratory analysis, we assessed potential covariation between promoter methylation on an individual basis. For this purpose, we performed hierarchical clustering of the samples by applying the Z-scores from average methylation ratio across the 565 capture regions. Doing so, all samples could be classified into two distinct major clusters, each harboring distinguishable sub-clusters (Fig. 2). Interestingly, the two major clusters (1 and 2) were characterized by different promoter methylation in two groups of genes (A and B), where cluster 1 had high methylation in genes in group A and low methylation in genes in group B, while the opposite methylation pattern was seen for samples in cluster 2 (Fig. 2).

We identified genes falling into these two groups (A and B), and analyzed their involvement in functional pathways by KEGG pathway analysis and GO enrichment analysis via Gather. Many of the genes involved in group A were important in development and regulation of cellular processes like Wnt signaling and TGF-beta signaling pathways. In contrast, genes from group B showed involvement in apoptotic pathways and leukocyte differentiation (Supplementary Table S4).

Notably, some individuals were characterized by having a majority of genes either *hyper*- or *hypo*-methylated as compared to the rest of individuals. Applying a 95% confidence interval across samples with respect to the overall methylation level of the regions analyzed, one sample (S24) fell below the lower limit of the CI, while three fell above the upper limit of the CI (Supplementary Figure 3). However, these individuals were distributed across the two main clusters with no preference for one group over the other. Assessing the available general clinical data for these individuals, no notable associations were observed between methylation and factors such as age or BMI (data not shown).

Validations in external data sets

Although our data are unique since they are generated by targeted massive parallel sequencing analyses, we sought to validate our biological findings by mining available data sets generated by application of methylation arrays.

A technical concern is that methylation could potentially vary between subfractions of leukocytes and differential methylation between individuals could then potentially be a result of individuals having different

compositions of leukocyte subfractions in their blood. Assessing the 7 most differentially methylated regions in our data set, in the leukocyte subfractions published in the Bioconductor Experiment Data Package FlowSorted.Blood.450K revealed no major difference in any of the 7 regions (Supplementary Table S5, with figures). In *GSTP1*, 6 out of 19 CpGs revealed lower methylation in CD14+ T cells and/or CD56+ NK cells than other subfractions, but the impact of this on the average levels in total WBC was negligible. Very similar observations were made in another data set of cord blood (R package FlowSorted. CordBlood Norway.450 K in Bioconductor [34]; Supplementary Table S6, with figures). This confirmed potentially varying composition of leukocyte subfractions not to be a likely cause of the observed methylation differences.

Further, we sought to validate the biological differences observed for the 7 most differentially methylated regions in our sample set, by assessing their methylation in a sample of blood DNA from 845 individuals (GSE51032). In this sample set, data was available for *CHN1*, *PDCD1LG2*, *GSTP1*, and *ZNF668*. In addition, we here included the two top-ranking genes with high differential methylation calculated as ratio, but where percent point difference was below 20 (see above; *RABEP1* and *AIP*; Table 2). In general, the methylation levels were called as slightly higher in the GSE51032 set than by our own sequencing. However, the differences between individuals were confirmed for all genes and the difference in percent points between the highest and lowest methylated individual was similar (Supplementary Table S7). The exception was *ZNF668*, where our maximum observation was 66% methylation, while in the GSE51032 set, some individuals were scored as 100% methylated. This difference probably relates to a substantially higher number of individuals analyzed in the validation set increasing the chance of observing outliers.

Discussion

While to this end constitutional epimutations of tumor suppressors have been linked to cancer risk for a few genes only [23, 27, 31, 35–37], one may postulate that constitutional epimutations affect other tumor suppressors as well. This may have implications to our understanding of cancer risk. A substantial number of cancer-prone families in which no underlying germline mutation have been identified, and it is tempting to postulate that some of these individuals may be at increased cancer risk due to constitutional epimutations in tumor suppressor genes [30]. In addition, germline mutations in several tumor suppressor genes have been associated with other conditions such as skin and limb development deficiencies, Cowden syndrome, and Fanconi anemia [38–40]. Thus, exploring constitutional promoter methylation across tumor suppressor genes may be of importance to other medical conditions as well.

To this end, the vast majority of epigenetic data reported in respect to different health conditions are based on global methylation-array analyses or single gene promoter analyses by methods like MSP or MLPA. While the array-based approaches do provide data for single CpGs, a large number of (potentially important) CpGs are lacking from the arrays, limiting the possibilities to identify methylation pattern across all regions of interest (e.g., as seen for *BRCA1* [23]). As for MSP and MLPA, such methods are fast and cheap but they are sensitive only to a general methylation presence in the CpGs covered by the primers and probes, precluding assessment at a single CpG resolution level.

Here, we established a massive parallel sequencing-based approach, enabling base-pair resolution analyses of methylation status in gene promoters. The method provides several advantages as compared to previous methods. First, as compared to conventional methods like MSP and MLPA, our method allows for detailed single-CpG resolution analyses of multiple promoter regions in concert. Second, our method limits both workload and costs compared to application whole-genome methylation sequencing for promoter methylation analysis. Third, the benefit of determining exact methylation levels, instead of binary assessments, has been confirmed in clinical studies [23], underlining the importance of high sensitivity required to detect low-grade mosaic methylation [30]. Fourth, as compared to available array-based approaches, our NGS-assay allows for methylation assessment of all CpGs in the region of interest, not only those covered by array probes. As mentioned above, this proved to be crucial in analyses of the cancer risk associated with mosaic *BRCA1* methylation [23].

In principle, the sequencing of the DNA-libraries we prepared could be run on any Illumina instrument. As such, the method is flexible and scalable. Here, we used the MiSeq instrument due to the rapid run time. In our set-up, we chose to run 8 samples in one run, yielding an average coverage of 189.6x, corresponding to a mean sensitivity limit of 0.53%. Although indicating a very sensitive method, this is an average value, and some regions reveal lower coverage. If needed, however, coverage could be increased in order to improve the sensitivity of the method [23]. Notably, the reproducibility of the assay may vary between the different covered regions. However, we show that the reproducibility is very good even in regions with low levels of methylation. Importantly, the observed technical variation was consistently negligible compared to the biological variations described. Further, we found that technical variations were lower when assessing all CpGs across

a given region than when assessing randomized selections of CpGs as "representative" for a region. This emphasizes the value of applying assays where all CpGs in a given region are covered, instead of relying on scattered, selected CpGs.

While constitutional methylation is considered an early life event affecting different germinal layers, methylation status is also prone to environmental influences and other factors and has been found to change during lifetime [41], causing differential methylation of many genes across different tissues [42]. One potentially important caveat when analyzing WBCs as surrogate markers for constitutional methylation is the fact that different leukocyte fractions may harbor different methylation patterns [43]. While such differences, so far, have been linked to global methylation patterns, it remains unclear whether this may represent a problem with respect to specific tumor suppressor methylation. Notably, differential methylation across WBC subfractions was found not be an issue regarding *BRCA1* promoter methylation [23], and in the present study, it was not found to be an issue in the most differentially methylated promoter regions either.

The methylation level of the genes found to confer cancer risk, so far, is highly variable. Regarding *MLH1*, normal cell methylation affecting ~ 50% of the alleles has been reported in a limited number of probands with familial colorectal cancer (for original references, see [30]). Recently, two families with a high breast and ovarian cancer incidence were found to harbor secondary constitutional *BRCA1* methylation, also with a methylation level of ~ 50% [31]. In contrast, about 4% of females in a Caucasian population was found to carry low-level mosaic constitutional *BRCA1* methylation (4-10% of alleles). Among these low-level methylated individuals, the incidence of high-grade serous ovarian cancer was significantly elevated with an odds ratio between 2 and 3 across two large cohorts [23]. As for the method presented here, this has the sensitivity required for exploring both scenarios.

While the limited number of samples analyzed precludes formal assessments of methylation frequency and/or potential correlations to health outcome, importantly, our findings confirm differential constitutional promoter methylation across a panel of tumor suppressor genes in healthy individuals. Interestingly, among those promoter regions found to be hyper-methylated in the normal tissue of some of the analyzed individuals, we found promoters in genes previously reported to be hyper-methylated in tumors (such as *RASSF1* and *GSTP1*). The presence of epigenetic deregulation of a distinct tumor suppressor at the somatic (tumor) level provides no evidence for constitutional methylation of the same gene. However, the examples related to *MLH1* and *BRCA1* suggest that potential relationships may occur for other genes as well. Thus, it is tempting to speculate that, at least some of the genes detected here (e.g., *RASSF1* and *GSTP1*) could be constitutionally methylated and, in such cases, methylated tumor cells may have originated from the constitutionally methylated normal cells [30]. Notably, although not directly comparable to our data, due to a restricted selection of CpGs covered, mining of a large external data set revealed similar interindividual differences largely confirming our findings.

Interestingly the methylation patterns revealed across our gene panel indicated that the individuals analyzed could be classified into two different methylation clusters. These findings should be interpreted with caution due to the limited number of individuals analyzed. However, the fact that the clusters were separated by differential methylation across important biological pathways involving Wnt- and TGF-beta signaling pathways as well as genes involved in apoptotic pathways and leukocyte differentiation indicate potential underlying biological differences to be explored in future studies.

Conclusions

We provide a relatively fast and affordable strategy for detailed assessments of differential methylation of tumor suppressors. This strategy is attractive in the warranted search for additional tumor suppressors that may be cancer risk factors when methylated in normal tissues.

Methods

Samples

The samples analyzed in the present study were from 34 individuals, selected from a set of 114 healthy postmenopausal women previously described [44]. Subsequent to providing informed consent, each individual donated anonymized blood samples in accordance with Norwegian regulations. All women were recruited during routine mammographic screening at Haukeland University Hospital, Bergen, Norway. Individuals with diabetes or other types of endocrine diseases as well as individuals using hormone replacement therapy were excluded. All samples were drawn > 2 years after the last menstrual period. Within the selection of 34 individuals analyzed in the present study, the mean age was 64 years (range 56-71 years) and the mean BMI was 24.8 (range 19.4-39.6) at the time of sample collection.

DNA isolation

Genomic DNA was extracted from EDTA-whole blood, using QIAamp DNA Mini kit (Qiagen). The procedure was performed according to the manufacturer's instructions with the exception that 400 μl of whole blood was used as input.

Selection of tumor suppressor promoter regions

Regions of interest were defined as 356 regions from the promoters of 283 tumor suppressor genes. The selection of genes was based on the cancer gene panel previously described as "CGPv2/3" [45, 46], Roche's "Comprehensive Cancer Design" as well as a manual literature review, in order to cover all well-established tumor suppressor genes, independent of cancer type. As such, the selection was independent of previous knowledge about methylation status. For each transcription start site (TSS), we designed probes covering a region spanning from −1500 to +500 relative to TSS. Positions of TSS were determined by NCBI and Ensembl-curated transcripts, literature search, and use of the FANTOM5 RNA expression resource (fantom.gsc.riken.jp/5/). Probes for hybridization to the included regions were manufactured by Roche and designed to bind the target DNA of all possible methylation configurations (fully methylated, partially methylated, and completely unmethylated). Importantly, both strands were targeted, in order to enable correction for potential overlap between CpGs and SNPs. By probe design, the 356 target regions were split into 565 capture regions. Full lists of included tumor suppressor genes and target regions are given in Supplementary Table S1.

Library preparation and methylation sequencing

Processing of the sample libraries was performed using the solution-based bead capture method for enrichment of bisulfite-converted DNA, SeqCap Epi Enrichment System (Roche) according to the user guide (version 1.2).

For each sample, 1 μg DNA isolated from blood was mixed with bisulfite-conversion control (Lambda DNA, negative for methylation). DNA was fragmented to the range of 180-220 bp using Covaris M220 followed by end repair, A-tailing, ligation of index/adapters, and dual size selection. Using the Zymo Research EZ DNA Methylation-Lightning kit, the DNA was bisulfite-converted according to manufactures protocol, and the resulting sample was amplified prior to nanodrop quantification. Based on these measurements, 1 μg bisulfite-converted DNA was put into the hybridization with custom-made probes for 68 h prior to capture by streptavidin-coated beads, extensive washing, and a final library amplification step.

The protocol was combined with the use of a custom-made probe design enabling analysis of only regions of interest (consisting of 356 promoter regions from 283 tumor suppressor genes, described above and in Supplementary Table S1). In addition, the probe set included probes targeting (Lambda DNA for conversion control). The targeted regions were enriched by a bead capturing method that captures both strands of DNA. Purified libraries were pooled, spiked with 10% PhiX, and sequenced on an Illumina MiSeq sequencer, using v2 chemistry and 2 × 100 (200 cycles) paired-end reads. RTA v1.18.54 and MCS v2.5.0.5 software was used to generate data. Eight samples were multiplexed per run, and resulting data were de-multiplexed based on sample-specific indexes attached to the sequencing adaptors. De-multiplexing was run automatically by the MiSeq Reporter software before further processing.

Methylation calling

Raw sequencing data was analyzed using an in-house workflow designed in collaboration with Roche, comprised of publicly available tools, implemented using shell script (Fig. 3; for a detailed description see Supplementary information). In brief, the first analytic steps involved quality checking of fastq files by FASTQC. Paired-end reads were filtered based on quality and clipped using Trimmomatic [47]. Trimmed sequences were aligned to the human genome (GRCh38) from NCBI as well as Enterobacteria phage lambda (NC_001416.1) complete genome, added for bisulfite conversion efficiency control using the bisulfite mapping algorithm BSMAP [48]. The aligned read statistics and format conversions were carried out using SAMtools [49]. After bisulfite conversion, the DNA strands are no longer complementary. To achieve methylation information from both strands, aligned reads were split into the top and bottom strand [50]. Subsequently, the sequences were sorted, and duplicates were removed and merged back using Picard tools. In the next step, the analysis was further restricted to those read pairs where both mates in the pair could be mapped in the correct orientation and at given distance consistent with the library insert size (properly paired reads) using BamTools [51]. To avoid bias, overlapping reads were clipped using BamUtils. Various statistics for reads, alignment, and coverage were calculated using SamTools.

For each sample, methylation analysis was carried out using methratio.py package in BSMAP by calculating methylation percentage. An additional step involves SNP calling for the targeted regions with BisSNP [52] from aligned reads.

DNA conversion rate was calculated based on all original Cs in the Lambda DNA sequence. For all Cs in the untreated sequence the following formula was used on sequencing data post bisulfite treatment:

$$\text{Conversion}\ (\%) = T/(C+T) \times 100$$

Assay reproducibility

To assess reproducibility of the assay, we performed 2 independent experiments with 6 parallels of a standard

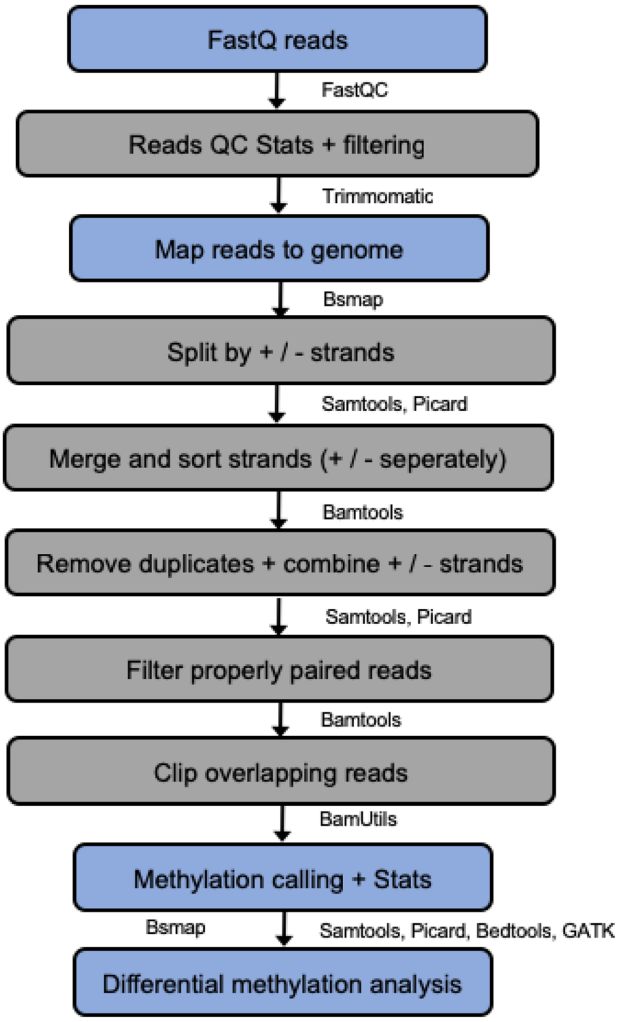


Fig. 3 Workflow of the methylation analysis. Flow chart of the steps taken within the informatics analysis pipeline from raw FastQ files to processed data used for biological interpretations. Main steps are indicated by blue background; smaller steps are indicated by gray background (figure adapted from original design by Roche)

sample in each experiment. The standard sample consisted of pooled DNA of equal amounts from WBC of 5 healthy donors. Reproducibility was assessed across 12 regions, selected based on three separate criteria: First, we selected 4 regions found to have high biological variance in our original sample set of 34 healthy women (*GAS7*, *ELAC2*, *AIP*, *ZRSR2*). Further, we selected 6 regions in genes known to be high penetrance genes when either mutated or hypermethylated (*BRCA1*, *TP53* (2 regions), *RB1*, *MLH1*, *MGMT*). Finally, we selected 2 regions at random (*PRDM2*, *TMEFF2*). Based on the 12 replicate analyses, we calculated mean methylation, standard deviation and coefficient of variation for all the regions (Supplementary Table S2a). Further, within the 2 randomly selected genes (*PRDM2*, *TMEFF2)*, we performed a randomized selection of 5 CpGs per region, using the mean methylation in these 5 as "representative" for the region. Then, we calculated mean methylation, standard deviation and coefficient of variation across the 12 replicate analyses of these 5 CpGs. This

randomization was repeated 5 times, yielding a general overview of the variability when applying limited numbers of CpGs as "representative" for a region (Supplementary Table S2b).

Differential methylation assessment

Among all CpGs in the 565 capture regions as well as 250 flanking bps at each end, the analysis was restricted only to include CpGs with minimum of 10 reads in coverage in all of the 34 samples. For each sample, we then calculated the mean methylation per region, based on individual CpG methylation ratios within the region. Based on these data, we generated a methylation matrix for all the common regions across all the samples (n = 34 in the present study), and calculated Z-scores for that matrix. Then we assessed the Z-scores and identified all the regions where a minority of individuals were differentially methylated as compared to the majority. Differential methylation was here defined as Z-scores that were outside of the 99% confidence interval. We used an arbitrary definition of minority, set to one-third, or less, of the total number of samples, i.e., minimum 1 individual and maximum 12 individuals (this definition may need adjustment according to the size of subsequent studies). Regions that had confidence level more than 99% were then categorized into negatively and positively methylated regions based on the Z-score value and whether the minority of individuals had higher or lower methylation levels than the majority.

To find the differentially methylated regions, we calculated the mean methylation for these regions across CpGs within individual samples and measured the difference in methylation between individuals with the lowest and highest methylation mean. Although relatively small differences in methylation levels have been shown to modulate cancer risk [23], we here sought to identify the regions with larger differences, applying arbitrary thresholds of 5, 10, and 20 percent point difference in methylation. Further, we performed additional analyses assessing ratios (fold difference) between individuals, taking into account that biological important differences may have high ratios, not necessarily reaching a certain threshold set by percent point difference (e.g., a difference between 1% and 10% may be important, even if the percent point difference is only 9).

Hierarchical clustering

We created a matrix of methylation ratios for all genes across patients. We then calculated a variance for each gene across patients to identify differential methylation. Heatmap was produced with heatplot function from made 4 package [53], with mean linkage cluster analysis and a correlation metric distance. For the purpose of clustering, missing values for regions in individual patients were filled in using the impute R package [54, 55]. (Impute-knn function from impute R package, finds k-nearest neighbors using a Euclidean metric and uses their mean to substitute the missing value). Missing values affected one region of *GSTM1* in 16 samples, another region of *GSTM1* in 7 samples, and a region of *AIP* in 3 samples.

Pathway analysis

We identified groups of genes from cluster analysis and explored their functional roles by pathway analyses with GATHER. GATHER is an online platform that predicts functional molecular patterns and biological context by incorporation of several biological databases [56]. In GATHER, we analyzed KEGG pathways and gene ontology enrichment analyses [57].

External data sets

We performed data mining and extracted detailed methylation status for all available CpGs for a given region (defined by our NGS-panel) from the Bioconductor Experiment Data Package FlowSorted.Blood.450K (https://bioconductor.org/packages/release/data/experiment). This data set was generated by methylation array analyses across 6 independent samples from adult individuals and contains information on 10 different categories of leukocytes. The categories include the major groups of granulocytes and lymphocytes.

We obtained similar data for umbilical cord blood from newborns [34]. These data were available as the R package FlowSorted.CordBloodNorway.450K in Bioconductor. This data set was also based on methylation array and holds information about 7 categories of leukocytes, including the major groups of granulocytes and lymphocytes, across 11 independent cord blood samples from newborns.

For validation of methylation differences in blood DNA from healthy individuals, we mined data from GSE51032, available through Gene Expression Omnibus (GEO). This data set was generated by methylation array and consists of 845 samples from the EPIC-Italy cohort (out of which 188 were males and 657 were females).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13148-020-00920-7.

Additional file 1: Supplementary Figure 1. Fraction of methylated alleles in promoter region of selected tumour suppressor genes. (A) Regions with high methylation levels across samples from all 34 healthy individuals. (B) Regions with low methylation levels across the same samples. Note the different scale on the Y-axis for panel A and B. Data for AIP were lacking for samples 32, 33, 34 due to low coverage (see details in Materials and methods). **Supplementary Figure 2.** Plot examplifying consistent high and low methylated CpGs in the same promoter, across patients. Fraction of methylated alleles across CpGs in the promoter region of *RB1* in the two samples S7 and S24 are displayed. These two

samples were selected because they were the one with highest and lowest overall methylation across the 283 investigated tumour suppressor genes, respectively (ref. Supplementary figure 3), and as such should represent the extremes. Still within the *RB1* promoter, they reveal a very similar pattern of some CpGs being highly methylated, while others are hardly methylated at all. **Supplementary Figure 3.** Distribution of overall average methylation across 283 tumour suppressor gene promoters in 34 healthy individuals. (A) Bars indicate the average fraction of methylated alleles for all CpGs covered per patient. Dotted red lines indicate the upper and lower border of the 95% confidence interval for the average values per patient (CI for individual observations). Sample S24 falls below the lower border of the CI, indicating general *hypo*-methylation. Samples S4, S8 and S7 fall above the upper border of the CI, indicating general *hyper*-methylation. (B) Q-Q plot based on the same data as displayed in (A). S24 is encircled in green, while S4, S8 and S7 are encircled in red.

Additional file 2: Supplementary information – workflow

Additional file 3: Supplementary Table S1. Pan-cancer panel of 283 tumor suppressor genes for which promoters are included in methylation analyses. The panel was generated based on CGPv2/3-panels [1], Roche's Comprehensive Cancer Design along with manual literature search.

Additional file 4: Supplementary Table S2a. Reproducibility test. **Supplementary Table S2b.** Reproducibility test restricted to randomised CpGs.

Additional file 5: Supplementary Table S3. Genes with >99 confidence level difference in methylation ratio between a minority (one third or less) of samples versus the majority.

Additional file 6: Supplementary Table S4. groupAB_GE

Additional file 7: Supplementary Table S5. WBC fractions

Additional file 8: Supplementary Table S6. Coord blood

Additional file 9: Supplementary Table S7. EPIC

Abbreviations

BMI: Body mass index; CpG: Cytosine-phosphate-guanine; HGSOC: High grade serous ovarian cancer; MSP: Methylation-specific polymerase chain reaction; MLPA: Multiplex ligation-dependent probe amplification; NGS: Next-generation sequencing; SNP: Single nucleotide polymorphism; WBC: White blood cells

Acknowledgements

We thank Beryl Leirvaag and Christine Eriksen for technical assistance.

Authors' contributions

Study design: DBP, EO, PEL, SK. Generation of data: DBP, EO, ZS, EV, GTI, LM. Interpretation of data: DBP, EO, ZS, PEL, SK. Funding/grants: PEL, SK. Writing of manuscript: DBP, PEL, SK. Approval of final manuscript: All authors

Funding

This work was performed in the Mohn Cancer Research Laboratory and was funded by grants from the Bergen Research Foundation, the Norwegian Cancer Society, the Norwegian Research Council, and the Norwegian Health Region West.

Availability of data and materials

All data generated or analyzed during this study are included in this published article (and its supplementary information files). Unprocessed raw files are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Subsequent to providing informed consent, each individual included in the present work donated anonymized blood samples for research purposes. This was done in accordance with Norwegian legislation at the time of sample collection (late 1990s).

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Author details

[1]K.G. Jebsen Center for Genome Directed Cancer Therapy, Department of Clinical Science, University of Bergen, Bergen, Norway. [2]Department of Oncology, Haukeland University Hospital, Bergen, Norway. [3]Present address: Department of Medical Genetics, Haukeland University Hospital, Bergen, Norway. [4]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway. [5]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway.

Received: 15 April 2020 Accepted: 17 August 2020
Published online: 28 August 2020

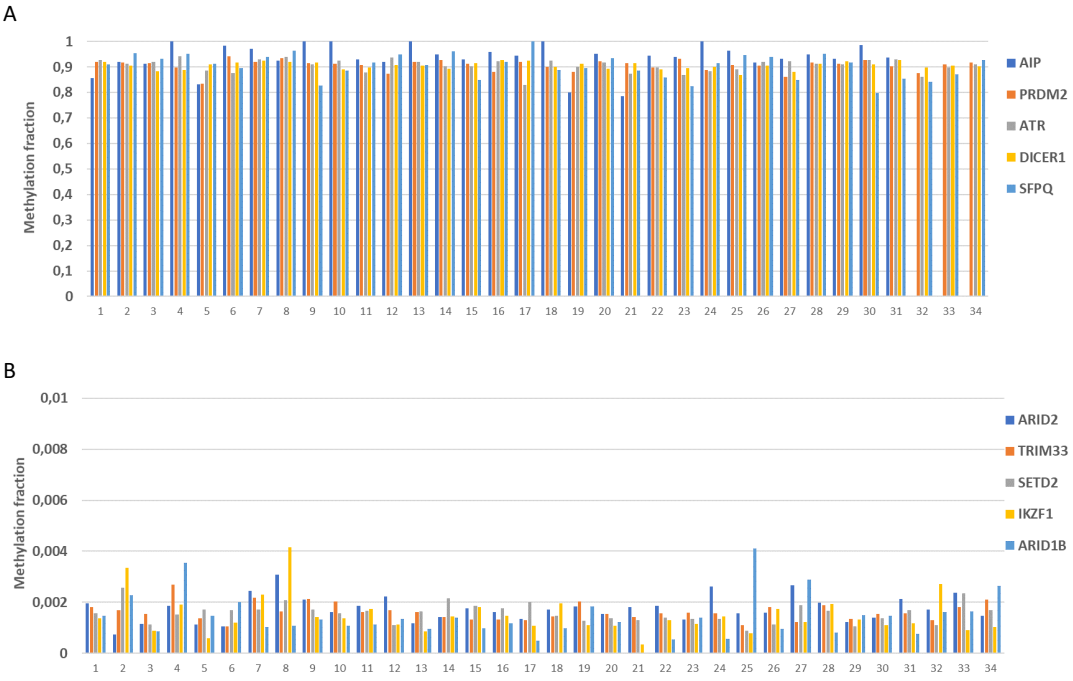
References

1. Clark DF, Maxwell KN, Powers J, Lieberman DB, Ebrahimzadeh J, Long JM, et al. Identification and confirmation of potentially actionable germline mutations in tumor-only genomic sequencing. JCO Precis Oncol. 2019;3 Epub 2019/09/13. doi: 10.1200/PO.19.00076. PubMed PMID: 31511844; PubMed Central PMCID: PMCPMC6738953.
2. Hata C, Nakaoka H, Xiang Y, Wang D, Yang A, Liu D, et al. Germline mutations of multiple breast cancer-related genes are differentially associated with triple-negative breast cancers and prognostic factors. J Hum Genet. 2020. https://doi.org/10.1038/s10038-020-0729-7 Epub 2020/02/08PubMed PMID: 32029870.
3. Jansen AML, Ghosh P, Dakal TC, Slavin TP, Boland CR, Goel A. Novel candidates in early-onset familial colorectal cancer. Familial Cancer. 2020; 19(1):1–10. https://doi.org/10.1007/s10689-019-00145-5 Epub 2019/09/27. PubMed PMID: 31555933.
4. Li FP, Fraumeni JF Jr. Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome? Ann Intern Med. 1969;71(4):747–52 PubMed PMID: 5360287.
5. Nichols KE, Malkin D, Garber JE, Fraumeni JF Jr, Li FP. Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. Cancer Epidemiol Biomark Prev. 2001;10(2):83–7 PubMed PMID: 11219776.
6. Donaldson A, Murray A, Antoniou AC, Brewer C, Houghton C, Evans DG, et al. Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. J Natl Cancer Inst. 2013;105(11):812–22. https://doi.org/10.1093/jnci/djt095.
7. Bonadona V, Bonaiti B, Olschwang S, Grandjouan S, Huiart L, Longy M, et al. Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. JAMA. 2011;305(22):2304–10. https://doi.org/10.1001/jama.2011.743 Epub 2011/06/07. PubMed PMID: 21642682.
8. Hussussian CJ, Struewing JP, Goldstein AM, Higgins PA, Ally DS, Sheahan MD, et al. Germline p16 mutations in familial melanoma. Nat Genet. 1994; 8(1):15–21. https://doi.org/10.1038/ng0994-15 Epub 1994/09/01. PubMed PMID: 7987387.
9. Borg A, Sandberg T, Nilsson K, Johannsson O, Klinker M, Masback A, et al. High frequency of multiple melanomas and breast and pancreas carcinomas in CDKN2A mutation-positive melanoma families. J Natl Cancer Inst. 2000;92(15):1260–6 Epub 2000/08/03. PubMed PMID: 10922411.
10. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971;68(4):820–3 Epub 1971/04/01. PubMed PMID: 5279523; PubMed Central PMCID: PMC389051.
11. Li FP, Fraumeni JF Jr. Rhabdomyosarcoma in children: epidemiologic study and identification of a familial cancer syndrome. J Natl Cancer Inst. 1969; 43(6):1365–73 PubMed PMID: 5396222.
12. Tung N, Domchek SM, Stadler Z, Nathanson KL, Couch F, Garber JE, et al. Counselling framework for moderate-penetrance cancer-susceptibility mutations. Nat Rev Clin Oncol. 2016;13(9):581–8 Epub 2016/06/15. doi: 10.1038/nrclinonc.2016.90. PubMed PMID: 27296296; PubMed Central PMCID: PMCPMC5513673.
13. Plichta JK, Griffin M, Thakuria J, Hughes KS. What's new in genetic testing for cancer susceptibility? Oncology (Williston Park, NY). 2016;30(9):787–99 Epub 2016/09/17. PubMed PMID: 27633409.
14. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nat Rev Genet. 2019;20(2):109–27. https://doi.org/10.1038/s41576-018-0074-2 PubMed PMID: 30479381. Epub 2018/11/28.
15. Llinas-Arias P, Esteller M. Epigenetic inactivation of tumour suppressor coding and non-coding genes in human cancer: an update. Open Biol. 2017;7(9) Epub 2017/09/22. doi: 10.1098/rsob.170152. PubMed PMID: 28931650; PubMed Central PMCID: PMCPMC5627056.

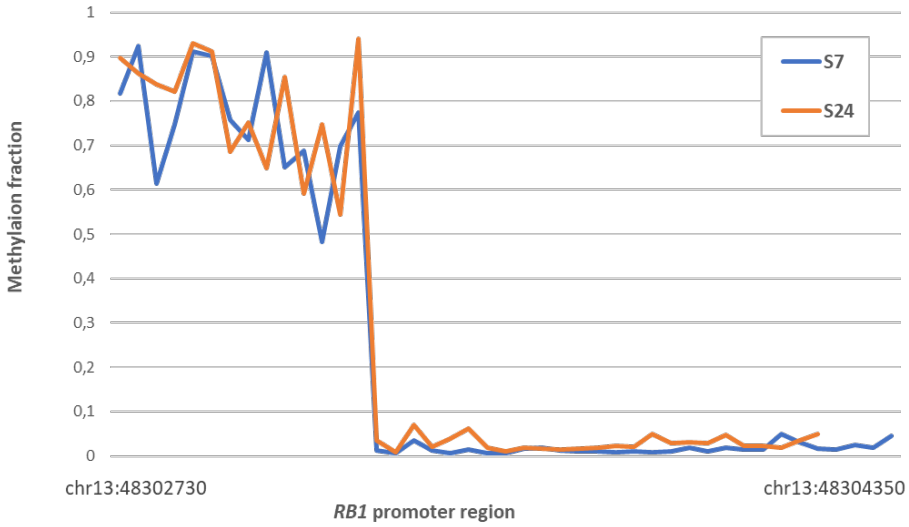
16. Sloane MA, Ward RL, Hesson LB. Defining the criteria for identifying constitutional epimutations. Clin Epigenetics. 2016;8:39 Epub 2016/04/21. doi: 10.1186/s13148-016-0207-4. PubMed PMID: 27096027; PubMed Central PMCID: PMCPMC4835913.
17. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene. 2002;21(35):5427–40 PubMed PMID: 12154405.
18. Damaso E, Canet-Hermida J, Vargas-Parra G, Velasco A, Marin F, Darder E, et al. Highly sensitive MLH1 methylation analysis in blood identifies a cancer patient with low-level mosaic MLH1 epimutation. Clin Epigenetics. 2019; 11(1):171 Epub 2019/11/30. doi: 10.1186/s13148-019-0762-6. PubMed PMID: 31779681; PubMed Central PMCID: PMCPMC6883525.
19. Gazzoli I, Loda M, Garber J, Syngal S, Kolodner RD. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. Cancer Res. 2002;62(14):3925–8 PubMed PMID: WOS:000176871500006.
20. Welin S, Sorbye H, Sebjornsen S, Knappskog S, Busch C, Oberg K. Clinical effect of temozolomide-based chemotherapy in poorly differentiated endocrine carcinoma after progression on first-line chemotherapy. Cancer. 2011;117(20):4617–22. https://doi.org/10.1002/cncr.26124 PubMed PMID: 21456005.
21. Hitchins MP. Constitutional epimutation as a mechanism for cancer causality and heritability? Nat Rev Cancer. 2015;15(10):625–34. https://doi.org/10.1038/nrc4001 Epub 2015/09/19. PubMed PMID: 26383139.
22. Verma M, Rogers S, Divi RL, Schully SD, Nelson S, Su LJ, et al. Epigenetic research in cancer epidemiology: trends, opportunities, and challenges. Cancer Epidemiol Biomark Prev. 2014;23(2):223–33. https://doi.org/10.1158/1055-9965.epi-13-0573 PubMed PMID: 24326628; PubMed Central PMCID: PMCPMC3925982.
23. Lonning PE, Berge EO, Bjornslett M, Minsaas L, Chrisanthar R, Hoberg-Vetti H, et al. White blood cell BRCA1 promoter methylation status and ovarian cancer risk. Ann Intern Med. 2018;168(5):326–34. https://doi.org/10.7326/M17-0101 Epub 2018/01/18. PubMed PMID: 29335712.
24. Chan TL, Yuen ST, Kong CK, Chan YW, Chan ASY, Ng WF, et al. Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. Nat Genet. 2006;38(10):1178–83. https://doi.org/10.1038/ng1866.
25. Hesson LB, Hitchins MP, Ward RL. Epimutations and cancer predisposition: importance and mechanisms. Curr Opin Genet Dev. 2010;20(3):290–8. https://doi.org/10.1016/j.gde.2010.02.005 Epub 2010/04/03. PubMed PMID: 20359882.
26. Hitchins MP. The role of epigenetics in Lynch syndrome. Familial Cancer. 2013;12(2):189–205. https://doi.org/10.1007/s10689-013-9613-3 Epub 2013/03/07. PubMed PMID: 23462881.
27. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, et al. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat Genet. 2009;41(1):112–7. https://doi.org/10.1038/ng.283 Epub 2008/12/23. PubMed PMID: 19098912.
28. Miyakura Y, Sugano K, Akasu T, Yoshida T, Maekawa M, Saitoh S, et al. Extensive but hemiallelic methylation of the hMLH1 promoter region in early-onset sporadic colon cancers with microsatellite instability. Clin Gastroenterol Hepatol. 2004;2(2):147–56 Epub 2004/03/16. PubMed PMID: 15017620.
29. Shen L, Kondo Y, Rosner GL, Xiao L, Hernandez NS, Vilaythong J, et al. MGMT promoter methylation and field defect in sporadic colorectal cancer. J Natl Cancer Inst. 2005;97(18):1330–8. https://doi.org/10.1093/jnci/dji275.
30. Lonning PE, Eikesdal HP, Loes IM, Knappskog S. Constitutional mosaic epimutations - a hidden cause of cancer? Cell Stress. 2019;3(4):118–35. https://doi.org/10.15698/cst2019.04.183 Epub 2019/06/22. PubMed PMID: 31225507; PubMed Central PMCID: PMCPMC6551830.
31. Evans DGR, van Veen EM, Byers HJ, Wallace AJ, Ellingford JM, Beaman G, et al. A dominantly inherited 5′ UTR variant causing methylation-associated silencing of BRCA1 as a cause of breast and ovarian cancer. Am J Hum Genet. 2018;103(2):213–20. https://doi.org/10.1016/j.ajhg.2018.07.002 Epub 2018/08/04. PubMed PMID: 30075112; PubMed Central PMCID: PMCPMC6080768.
32. Morak M, Schackert HK, Rahner N, Betz B, Ebert M, Walldorf C, et al. Further evidence for heritability of an epimutation in one of 12 cases with MLH1 promoter methylation in blood cells clinically displaying HNPCC. Eur J Hum Genet. 2008;16(7):804–11. https://doi.org/10.1038/ejhg.2008.25 Epub 2008/02/28. PubMed PMID: 18301449.
33. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1): D777–D83. https://doi.org/10.1093/nar/gkw1121.
34. Gervin K, Page CM, Aass HC, Jansen MA, Fjeldstad HE, Andreassen BK, et al. Cell type specific DNA methylation in cord blood: a 450 K-reference data set and cell count-based validation of estimated cell type composition. Epigenetics. 2016;11(9):690–8. https://doi.org/10.1080/15592294.2016.1214782 Epub 2016/08/06. PubMed PMID: 27494297; PubMed Central PMCID: PMCPMC5048717.
35. Hansmann T, Pliushch G, Leubner M, Kroll P, Endt D, Gehrig A, et al. Constitutive promoter methylation of BRCA1 and RAD51C in patients with familial ovarian cancer and early-onset sporadic breast cancer. Hum Mol Genet. 2012;21(21):4669–79. https://doi.org/10.1093/hmg/dds308 Epub 2012/07/31. PubMed PMID: 22843497; PubMed Central PMCID: PMCPMC3471399.
36. Hitchins MP, Wong JJ, Suthers G, Suter CM, Martin DI, Hawkins NJ, et al. Inheritance of a cancer-associated MLH1 germ-line epimutation. N Engl J Med. 2007;356(7):697–705. https://doi.org/10.1056/NEJMoa064522 Epub 2007/02/16. PubMed PMID: 17301300.
37. Prajzendanc K, Domagala P, Hybiak J, Rys J, Huzarski T, Szwiec M, et al. BRCA1 promoter methylation in peripheral blood is associated with the risk of triple-negative breast cancer. Int J Cancer. 2020;146(5):1293–8. https://doi.org/10.1002/ijc.32655 Epub 2019/08/31. PubMed PMID: 31469414.
38. Blumenthal GM, Dennis PA. PTEN hamartoma tumor syndromes. Eur J Hum Genet. 2008;16(11):1289–300. https://doi.org/10.1038/ejhg.2008.162 Epub 2008/09/11. PubMed PMID: 18781191; PubMed Central PMCID: PMCPMC6939673.
39. Celli J, Duijf P, Hamel BC, Bamshad M, Kramer B, Smits AP, et al. Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome. Cell. 1999;99(2):143–53. https://doi.org/10.1016/s0092-8674(00)81646-3 Epub 1999/10/27. PubMed PMID: 10535733.
40. Fiesco-Roa MO, Giri N, McReynolds LJ, Best AF, Alter BP. Genotype-phenotype associations in Fanconi anemia: a literature review. Blood Rev. 2019;37:100589. https://doi.org/10.1016/j.blre.2019.100589 Epub 2019/07/29. PubMed PMID: 31351673; PubMed Central PMCID: PMCPMC6730648.
41. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604–9. https://doi.org/10.1073/pnas.0500398102 Epub 2005/07/13. PubMed PMID: 16009939; PubMed Central PMCID: PMCPMC1174919.
42. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. Nature. 2019;571(7766):489–99. https://doi.org/10.1038/s41586-019-1411-0 Epub 2019/07/26. PubMed PMID: 31341302.
43. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30(10):1431–9. https://doi.org/10.1093/bioinformatics/btu029 Epub 2014/01/24. PubMed PMID: 24451622; PubMed Central PMCID: PMCPMC4016702.
44. Helle SI, Ekse D, Holly JMP, Lønning PE. The IGF-system in healthy pre- and postmenopausal women: relations to demographic variables and sex-steroids. J Steroid Biochem Mol Biol. 2002;81(1):95-102. doi: http://dx.doi.org/https://doi.org/10.1016/S0960-0760(02)00052-3.
45. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21(7):751–9. https://doi.org/10.1038/nm.3886 PubMed PMID: 26099045; PubMed Central PMCID: PMC4500826.
46. Yates LR, Knappskog S, Wedge D, Farmery JHR, Gonzalez S, Martincorena I, et al. Genomic evolution of breast cancer metastasis and relapse. Cancer Cell. 2017;32(2):169-184.e7. doi: https://doi.org/10.1016/j.ccell.2017.07.005. PubMed PMID: 28810143; PubMed Central PMCID: PMCPMC5559645.
47. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PubMed PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.
48. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10:232. https://doi.org/10.1186/1471-2105-10-232 PubMed PMID: 19635165; PubMed Central PMCID: PMC2724425.
49. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93. https://doi.org/10.1093/bioinformatics/btr509 Epub 2011/09/10. PubMed PMID: 21903627; PubMed Central PMCID: PMCPMC3198575.

50. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11(3):204–20. https://doi.org/10.1038/nrg2719 PubMed PMID: 20142834; PubMed Central PMCID: PMCPMC3034103.
51. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–2. https://doi.org/10.1093/bioinformatics/btr174 PubMed PMID: 21493652; PubMed Central PMCID: PMCPMC3106182.
52. Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 2012;13(7):1–14. https://doi.org/10.1186/gb-2012-13-7-r61.
53. Culhane AC, Thioulouse J, Perriere G, Higgins DG. MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics. 2005;21(11):2789-2790. doi: https://doi.org/10.1093/bioinformatics/bti394. Epub 2005/03/31. PubMed PMID: 15797915.
54. Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstein D. Imputing missing data for gene expression arrays. Stanford University Statistics Department Technical report; 1999.
55. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5 Epub 2001/06/08. PubMed PMID: 11395428.
56. Chang JT, Nevins JR. GATHER: a systems approach to interpreting genomic signatures. Bioinformatics. 2006;22(23):2926–33. https://doi.org/10.1093/bioinformatics/btl483 Epub 2006/09/27. PubMed PMID: 17000751.
57. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30 Epub 1999/12/11. PubMed PMID: 10592173; PubMed Central PMCID: PMCPMC102409.

Publisher's Note

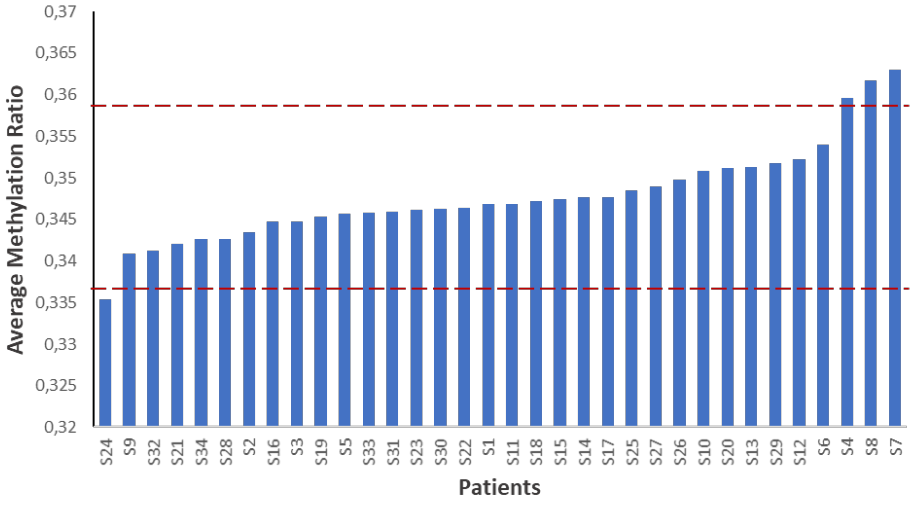


Supplementary figure 1. Fraction of methylated alleles in promoter region of selected tumour suppressor genes. (A) Regions with high methylation levels across samples from all 34 healthy individuals. (B) Regions with low methylation levels across the same samples. Note the different scale on the Y-axis for panel A and B. Data for AIP were lacking for samples 32, 33, 34 due to low coverage (see details in Materials and methods).

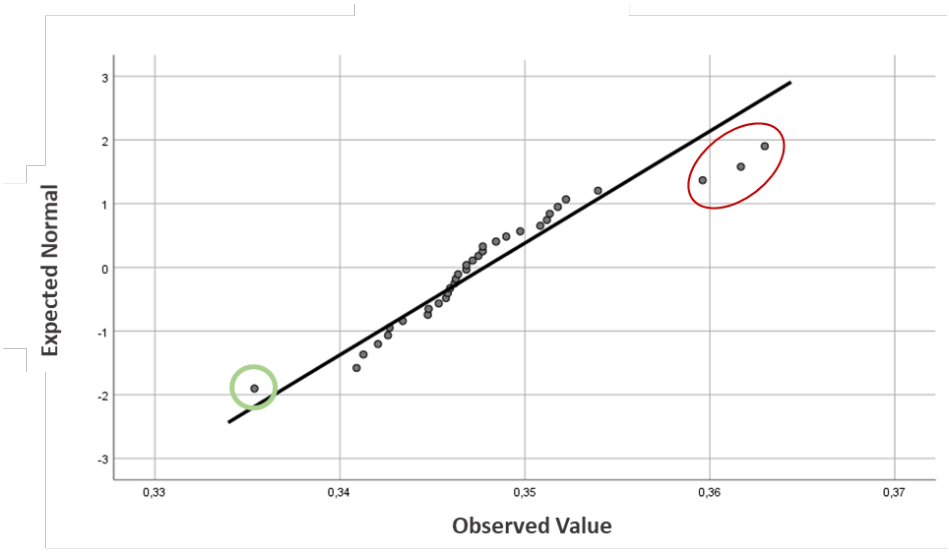


Supplementary figure 2. Plot examplifying consistent high and low methylated CpGs in the same promoter, across patients. Fraction of methylated alleles across CpGs in the promoter region of *RB1* in the two samples S7 and S24 are displayed. These two samples were selected because they were the one with highest and lowest overall methylation across the 283 investigated tumour suppressor genes, respectively (ref. Supplementary figure 3), and as such should represent the extremes. Still within the *RB1* promoter, they reveal a very similar pattern of some CpGs being highly methylated, while others are hardly methylated at all.

A



B



Supplementary figure 3. Distribution of overall average methylation across 283 tumour suppressor gene promoters in 34 healthy individuals. (A) Bars indicate the average fraction of methylated alleles for all CpGs covered per patient. Dotted red lines indicate the upper and lower border of the 95% confidence interval for the average values per patient (CI for individual observations). Sample S24 falls below the lower border of the CI, indicating general hypomethylation. Samples S4, S8 and S7 fall

above the upper border of the CI, indicating general hypermethylation. (B) Q-Q plot based on the same data as displayed in (A). S24 is encircled in green, while S4, S8 and S7 are encircled in red.

PLOS ONE

RESEARCH ARTICLE

The novel microRNAs hsa-miR-nov7 and hsa-miR-nov3 are over-expressed in locally advanced breast cancer

Deepak Poduval[1], Zuzana Sichmanova[1], Anne Hege Straume[1¤], Per Eystein Lønning[1,2], Stian Knappskog[1,2]*

1 Section of Oncology, Department of Clinical Science, University of Bergen, Bergen, Norway, **2** Department of Oncology, Haukeland University Hospital, Bergen, Norway

¤ Current address: Norwegian Institute of Marine Research, Bergen, Norway.
* stian.knappskog@uib.no

Citation: Poduval D, Sichmanova Z, Straume AH, Lønning PE, Knappskog S (2020) The novel microRNAs hsa-miR-nov7 and hsa-miR-nov3 are over-expressed in locally advanced breast cancer. PLoS ONE 15(4): e0225357. https://doi.org/10.1371/journal.pone.0225357

Editor: Bernard Mari, Institut de Pharmacologie Moleculaire et Cellulaire, FRANCE

Received: November 2, 2019

Accepted: March 16, 2020

Published: April 16, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0225357

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files and the raw data are available

Abstract

miRNAs are an important class of small non-coding RNAs, which play a versatile role in gene regulation at the post-transcriptional level. Expression of miRNAs is often deregulated in human cancers. We analyzed small RNA massive parallel sequencing data from 50 locally advanced breast cancers aiming to identify novel breast cancer related miRNAs. We successfully predicted 10 novel miRNAs, out of which 2 (hsa-miR-nov3 and hsa-miR-nov7) were recurrent. Applying high sensitivity qPCR, we detected these two microRNAs in 206 and 214 out of 223 patients in the study from which the initial cohort of 50 samples were drawn. We found hsa-miR-nov3 and hsa-miR-nov7 both to be overexpressed in tumor versus normal breast tissue in a separate set of 13 patients ($p = 0.009$ and $p = 0.016$, respectively) from whom both tumor tissue and normal tissue were available. We observed *hsa-miR-nov3* to be expressed at higher levels in ER-positive compared to ER-negative tumors ($p = 0.037$). Further stratifications revealed particularly low levels in the her2-like and basal-like cancers compared to other subtypes ($p = 0.009$ and 0.040, respectively). We predicted target genes for the 2 microRNAs and identified inversely correlated genes in mRNA expression array data available from 203 out of the 223 patients. Applying the KEGG and GO annotations to target genes revealed pathways essential to cell development, communication and homeostasis. Although a weak association between high expression levels of *hsa-miR-nov7* and poor survival was observed, this did not reach statistical significance. *hsa-miR-nov3* expression levels had no impact on patient survival.

Introduction

miRNAs are an important class of small non-coding RNAs, playing a versatile role in the gene regulation at the post–transcriptional level [1–5]. These molecules have proven to be involved in vital cellular functions, such as development, differentiation and metabolism [6–8]. In recent years there has been increased focus on the role of miRNAs in cancer [9], and the

through the Gene Expression Omnibus (accession: GSE145151).

Funding: This work was performed in the Mohn Cancer Research Laboratory. The project was funded by grants from the Trond Mohn Research Foundation, The Norwegian Cancer Society, The Norwegian Research Council and the Norwegian Health Region West.

Competing interests: The authors have declared that no competing interests exist

implementation of next generation sequencing (NGS) has led to the identification of multiple novel miRNAs as well as linked individual miRNA expression and combined signatures to tumor characteristics [10]. Currently there are 2656 distinct human miRNAs identified in the miRbase v22 [11], including more than 700 found to be deregulated in cancers [12].

Breast cancer is the most common malignancy in women. While outcome has improved significantly over the last three decades, resistance to therapy still presents a major challenge causing breast cancer related deaths [13]. As for chemoresistance in general, the underlying biological mechanisms remain poorly understood [14].

Merging evidence has indicated miRNA deregulation to play a role in breast cancer biology and outcome. Dysregulation of miRNAs may affect signal transduction pathways by targeting oncogenes and tumor suppressor genes [15], important to cancer development, progression, metastasis and potentially therapy response [16, 17]. Thus, while miR-10b, miR-125b, and miR-145 are generally downregulated, other miRNAs, like miR-21 and miR-155, are generally upregulated in breast cancer as compared to normal breast tissue [18]. Further, several miRNAs have revealed strong associations to clinical parameters [19, 20]: For example, differential expression of miR-210, miR-21, miR-106b*, miR-197, miR-let-7i, and miR-210, have been identified as a signature with prognostic value and also linked to invasiveness [21]. Moreover, miR-21 has been found linked to breast cancer metastasis and poor survival [22], while mir-29a overexpression has been shown to reduce the growth rate of breast cancer cells [23]. Given that many of the observed miRNA alterations are strongly cancer specific, this has inspired investigations into the potential use of miRNA as diagnostic biomarkers. Since miRNA are relatively stable molecules, they may be particularly attractive biomarkers to screen for in liquid biopsies (for original references, see [24])

miRNAs are also known to be differentially regulated across different subclasses of breast cancer. E.g. while members of the mir-181 family are up regulated in breast cancer in general, miR-181c in particular is activated by the expression of HER2 gene [25]. Also, miR-140 has been found suppressed by estrogen stimulation in ERα-positive breast cancer cells, most likely due to ER response elements in the flanking element of the miR-140 promoter [26].

In the present study, we analyzed global miRNA expression in 50 locally advanced breast cancers using NGS, aiming to identify novel, potentially breast cancer specific miRNAs. We identified and validated two novel miRNAs (one not previously described and one not previously reported in breast cancer), and subsequently evaluated their expression in an extended patient series (n = 223), by high sensitivity qPCR. Both were found over-expressed in breast cancer as compared to normal breast tissue. Considering different breast cancer subtypes, *hsa-miR-nov3* was expressed at particular high levels in ER-positive tumors contrasting lower levels in basal-like and Her2-like tumors. No similar patterns were observed for *hsa-miR-nov7*.

Materials and methods

Patients

In the present work we have analyzed biopsy material from two breast cancer studies.

1) In the first study, incisional biopsies were collected before chemotherapy from 223 patients with locally advanced breast cancer in a prospective study designed to identify the response to epirubicin (n = 109) and paclitaxel (n = 114) monotherapy. Primary response to therapy as well as long-term follow up (>10 years or death) was recorded for all patients. This cohort has been described in detail previously [27].

2) In the second study, tumor breast tissue and normal breast tissue from tumor bearing and non-tumor bearing quadrants were collected from 46 anonymous breast cancer patients

undergoing mastectomy, with the purpose of determining tissue estrogens. This cohort is described in detail in [28].

Using NGS, we analyzed miRNA expression in 50 patients from study 1). Next, candidate miRNAs were quantified using qPCR in all 223 patients from study 1), as well as 13 randomly selected patients from study 2), where RNA was available from tumor tissue and matching normal breast tissue (7 ER-positive and 6 ER-negative tumors). In addition, mRNA expression array data was available for 203 out of the 223 patients in study 1).

All patients provided written informed consent, and the studies conducted in accordance to national laws, regulation and ethical permissions (Norwegian health region West; REK Vest).

Tissue sampling and RNA extraction

Tissue samples were snap-frozen in liquid nitrogen in the operating theatre and stored in liquid nitrogen until further processing. Total RNA was extracted from the biopsies using miRvanaTM kit (ThermoFisher), according to the manufacturer's instructions. RNA integrity and concentration were determined using Bioanalyzer 2000 and Nanodrop ND2000 spectrophotometer, respectively.

miRNA-sequencing

Sample preparation and single-end sequencing were performed at the core facility of the Norwegian Genomics Consortium in Oslo, on Illumina HiSeq 2500, 1x50bp. De-multiplexing was performed using the Illumina CASAVA software. FastQC was run on all samples with the main purpose to assess sequence quality. The raw data are available through the Gene Expression Omnibus (accession: GSE145151).

Novel miRNA prediction

The raw sequencing files (fastq) were processed using the novel miRNA prediction algorithm mirdeep v2.0.0.5 [10]. Potential novel miRNAs were identified using the human reference genome (hg19) and already identified miRNAs from humans and other hominids from miRbase 20 [29]. In the mirdeep2 algorithm, filtering parameters randfold P-value less than 0.05 and scores greater than or equal to 10 were applied. Precursor structures obtained after filtering were manually identified based on the presence of 1–2 mismatches in the stem region, a loop sequence of 4–8 nt, and the presence of mature sequence in the stem region (See S1 File.) [30].

Validation of predicted novel miRNAs

Validation of the predicted novel miRNAs was performed by qPCR-based amplification of the miRNAs, with subsequent cloning and capillary sequencing of the products, to pinpoint the exact size and sequence of the miRNAs (see sections below for details).

cDNA synthesis and qPCR

cDNA from miRNAs was prepared using Exiqon's Universal cDNA synthesis kit II, with 20 ng of total RNA as input. qPCR was performed using Exiqon's miRCURY LNA™ Universal RT microRNA PCR system, with custom Pick-&-Mix ready to use PCR plates with an inter-plate calibrator, on a LightCycler 480 instrument (Roche). Relative expression levels for each sample were calculated by dividing the expression of the gene of interest on the average expression of two reference miRNAs: miR-16-5p and miR-30b-5p.

miRNA cloning and capillary sequencing

End products from custom miRNA specific qPCR were cloned into pCR 2.1 TOPO-TA vector (Life Technologies) by TOPO-TA cloning according to the manufacturer's instructions. The generated plasmids were amplified by transformation and cultivation of E. coli TOP10 cells (Life Technologies). The plasmids were then isolated using the Qiagen miniprep kit according to the manufacturer's instructions.

Sequencing was performed using the BigDye v.1.3 system (Applied Biosystems) and the primers following thermocycling conditions as previously described [31]. Capillary electrophoresis and data collection were performed on an automated capillary sequencer (ABI3700).

Target prediction and pathway analysis

Target prediction was performed using the offline algorithm miRanda [32, 33] and the online algorithms miRDB [34] and TargetScanHuman Custom (Release 5.2) [35].

miRanda predicts gene targets based on position specific sequence complementarity between miRNA and mRNA using weighted dynamic programming, an extension of the Smith-Waterman algorithm [36]. Also, the miRanda algorithm uses the free energy estimation between duplex of miRNA: mRNA (Vienna algorithm [37]) as an additional filter.

The miRDB is an online database of animal miRNA targets, which uses SVM (Support Vector Machine) machine-learning algorithm trained with miRNA-target binding data from already known and validated miRNA-mRNA interactions [34, 38].

TargetScanHuman Custom predicts biological miRNA targets by searching for match for the seed region of the given miRNA that is present in the conserved 8-mer and 7-mer sites [35]. It also identifies sites with conserved 3' pairing from the mismatches in the seed region [39, 40].

An in-house pan-cancer panel of 283 tumor suppressor genes was used to filter target genes of interest. The panel was generated based on the tumor suppressors within the CGPv2/3-panels [41], Roche's Comprehensive Cancer Design as well as a manual literature search (S1 Table).

Further, we used GATHER, a functional gene enrichment tool, which integrates various available biological databases to find functional molecular patterns, in order to find biological context from the target gene list [42]. With the help of GATHER, we did KEGG pathway [43], and GO (gene ontology) enrichment analyses for the common genes predicted by all three prediction algorithms. Further, validations were performed using DAVID [44] and topGO [45].

mRNA expression

In the interest of validating miRNA targets, we analyzed inverse correlations between miRNA expression and mRNA levels. mRNA expression levels were extracted from microarray analyses performed on a Human HT-12-v4 BeadChip (Illumina) after labeling (Ambion; Aros Applied Biotechnology). Illumina BeadArray Reader (Illumina) and the Bead Scan Software (Illumina) were used to scan BeadChips. Expression signals from the beads were normalized and further processed as previously described [46]. We re-annotated the data set using illuminaHumanv4.db from AnnotationDbi package, built under Bioconductor 3.3 in R [47], to select only probes with "Perfect" annotation[48]. The probes represented 21043 identified and unique genes (13340 represented by single probe and 7703 represented by multiple probes). In the cases of multiple probes targeting the same gene, we calculated fold difference for these probes. This was done to avoid losing potentially relevant biological information if expression of one probe was significantly higher that expression of another. However, for no genes did we find a fold difference >2 fold. Therefore, the mean expression for each such gene, was

calculated based on the values form each probe, weighted according to the number of beads per probe.

Statistics

Expression levels of miRNAs in tumor versus normal tissue were compared by Wilcoxon rank tests for paired samples. Inverse correlations between miRNA expression and mRNA expression were assessed by Spearman tests. The potential impact of the novel miRNAs on long-term outcome (relapse-free survival and disease-specific survival) in breast cancer patients was calculated by Log-rank tests and illustrated by Kaplan-Meier curves, using the SPSS software v.19. All p-values are reported as two-sided.

Results

Novel miRNA prediction

In order to identify novel miRNAs, 50 patients with locally advanced breast cancer (from study 1, see materials and methods) were subject to global miRNA-sequencing using massive parallel sequencing. On average, the dataset resulted in 3 million reads per sample. Using the miRNA identifier module in miRDeep2, we detected 10 novel miRNAs (Table 1). Eight out of these 10 miRNAs were detected in a single sample only, while two were expressed in two or more patients and therefore regarded as the most reliable predictions. These two miRNAs, here temporarily named *hsa-miR-nov3* and *hsa-miR-nov7*, were found in tumor samples from 2 and from 6 patients, respectively. For both of these novel miRNAs, we identified precursor structures with not more than one or two mismatches in the stem region, as well as the presence of mature miRNA sequences (Fig 1; S1 Fig). Therefore, we selected these two miRNAs for further analyses. Cross-checking the miRCarta database [49], no hits were found for either of the two, but notably, while this work was conducted, *hsa-miR-nov7* was identified by another team in lymphomas, and reported as miR-10393-3p [50].

In-vitro validation of novel micro RNAs

Next, we aimed to validate our *in-silico* predictions and confirm that the sequences from which we identified *hsa-miR-nov3* and *hsa-miR-nov7* represented bona-fide novel miRNAs expressed in the patients. Utilizing total RNA from the patients found to express the two predicted novel miRNAs, we performed global poly-adenylation and cDNA synthesis followed by miRNA-specific qPCR amplification. For both miRNAs we observed positive qPCR reactions. Further, end products of the qPCRs were then ligated into carrier-plasmids and sequenced.

Table 1. Novel miRNA sequences as predicted by mirdeep v2.0.0.5 from massive parallel sequencing of total miRNA in 50 locally advanced breast cancers.

miRNA	Co-ordinate	Mature sequence	Strand	Number of samples
hsa-miR-nov2	chr2:36662749..36662809	AAAAACTGCGATTACTTTTGCA	-	1
hsa-miR-nov3	chr3:186505088..186505149	AAAGCAGGATTCAGACTACAATAT	+	2
hsa-miR-nov3_2	chr3:132393169..132393224	CAAAAACTGCAATTACTTTTGC	+	1
hsa-miR-nov4	chr4:155140075..155140134	AAAAGTAATCGCTGTTTTTG	+	1
hsa-miR-nov7	chr7:138728845..138728903	AATTACAGATTGTCTCAGAGA	-	6
hsa-miR-nov8	chr8:116546693..116546762	TTAGAGCTTCAACCTCCAGTGTGA	-	1
hsa-miR-nov10	chr10:31840034..31840078	CGCGGGTGCTTACTGACCCT	+	1
hsa-miR-nov10_2	chr10:72163928..72163994	GCGGCGGCGGCGGCGGCG	+	1
hsa-miR-nov17	chr17:36760852..36760906	CCCAGCCCCACGCGTCCCCATG	-	1
hsa-miR-nov20	chr20:26189318..26189366	TGGCCGAGCGCGGCTCGTCGCC	-	1

https://doi.org/10.1371/journal.pone.0225357.t001

We confirmed the resulting plasmids to contain the predicted miRNA sequences. Further, in both cases, the sequences were flanked by a poly A tail, confirming that the original molecules used as input in the poly-adenylation were present as short 22nt RNAs (Fig 2). Thus, we confirmed the presence of miRNAs with the exact sequence as predicted from the NGS-based data.

Overexpression of hsa-miR-nov7 and hsa-miR-nov3 in breast cancer

Given that the sensitivity for the novel miRNAs was better in the qPCR than in the miRNA massive parallel sequencing (MPS) analysis, we aimed to assess whether the miRNAs were expressed in a limited number of breast cancer patients only (as indicated by their detection in 2 and 6 out of 50 patients in the MPS analysis), or if they were detectable in a higher fraction of patients, when applying a more sensitive detection method. We therefore performed qPCR to quantify the expression levels of *hsa-miR-nov7 and hsa-miR-nov3* in tumor tissue across the entire cohort of patients from study 1 (n = 223). With this method, we detected *hsa-miR-nov7* and *hsa-miR-nov3* in 206 and 214 samples out of total 223 samples respectively, albeit at variable levels (Fig 3).

Interestingly, while no difference in the expression levels of *hsa-miR-nov7* was observed between breast cancer subgroups, we found a significant difference in the expression levels of *has-miR-nov3* related to estrogen receptor status. Thus, the expression levels of *has-miR-nov3* were higher in ER-positive as compared to ER-negative tumors (p = 0.037; Fig 4A). Further,

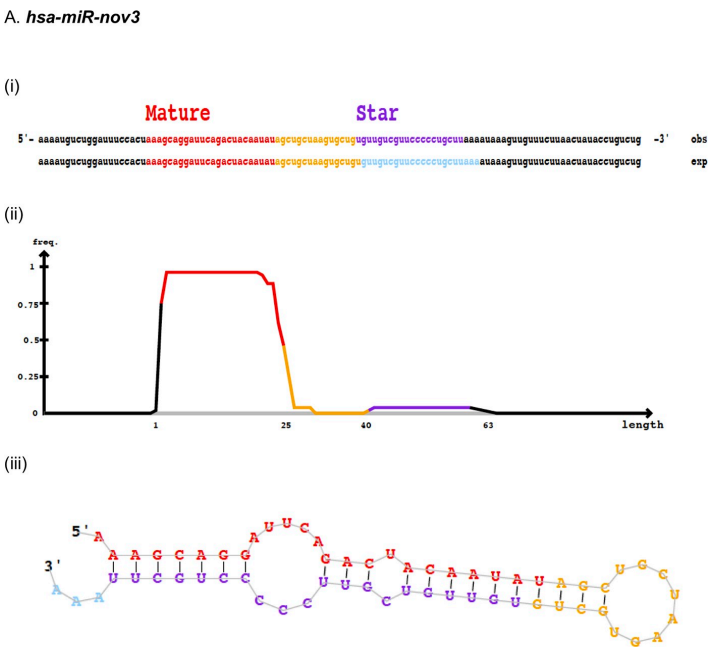


Fig 1. Predicted novel miRNAs. Depiction of novel miRNAs (A) *hsa-miR-nov3* and (B) *hsa-miR-nov7*, identified by miRDeep2, showing (i) predicted mature and star sequences, **exp**, probabilistic model expected from Drosha/Dicer processing and **obs**, observed sequences from sequencing data (ii) density plot for read counts for mature and star sequences as well as (iii) miRNA secondary structure.

https://doi.org/10.1371/journal.pone.0225357.g001

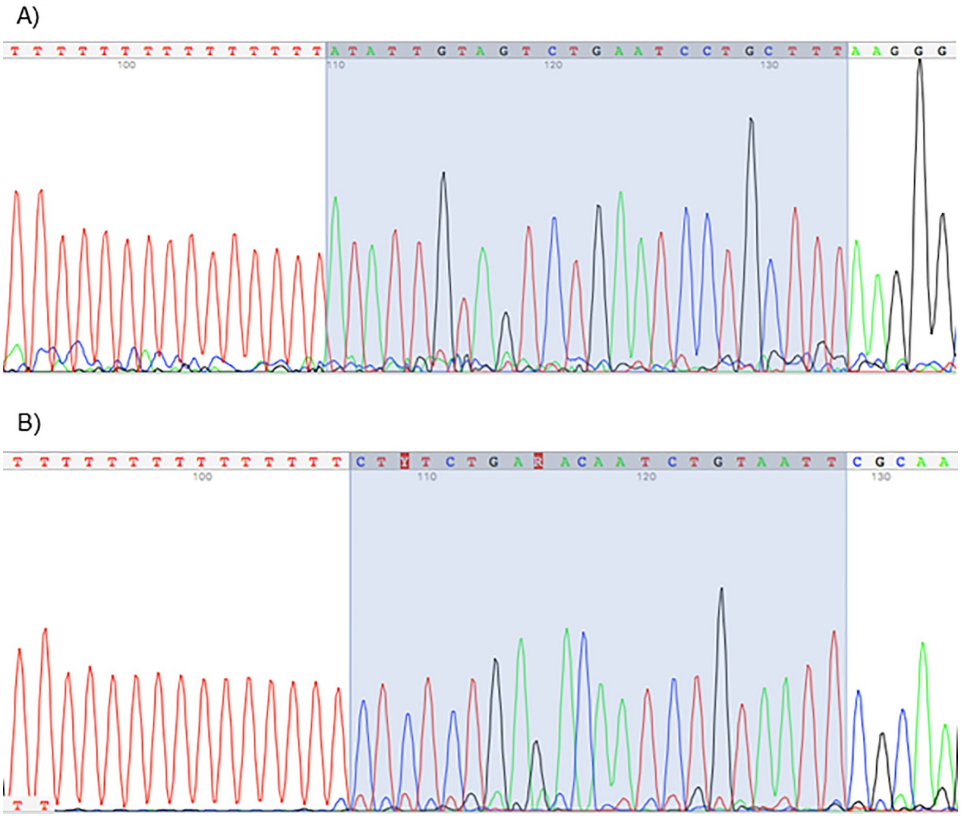


Fig 2. miRNA sequences. Chromatogram of capillary-sequenced qPCR products after *hsa-miR-nov3* (A) and *hsa-miR-nov7* (B) amplification. Highlighted background indicates the 22nt miRNA-sequence region (reverse complementary), followed by the Adenine homopolymer indicating *in vitro* adenylation at the expected site, confirming the exact size and sequence of the predicted miRNAs.

https://doi.org/10.1371/journal.pone.0225357.g002

assessing the expression levels of the two miRNAs in mRNA-based subclasses of breast cancer according to the Perou classification [51], comparing all five classes, we observed a significant difference between the subtypes with respect to *miR-nov3* expression ($p = 0.041$; Kruskal-Wallis test; Fig 4C). We found *hsa-miR-nov3* levels to be lower in HER2 like ($p = 0.009$; Mann-Whitney test) and basal-like ($p = 0.04$; Mann-Whitney) tumors as compared to tumors of the other classes.

Following the finding that the two miRNAs were detectable in more than 90 percent of patients, in order to assess whether the expression of these miRNAs were tumor specific we compared the levels of *hsa-miR-nov7* and *hsa-miR-nov3* expression in breast cancer tissue versus normal breast tissue. For this purpose, we randomly selected 13 patients from a study where samples of breast tumor tissue and matching normal tissue from a non-tumor bearing quadrant of the same breast were available (study 2, see materials and methods) [28]. We detected expression of the novel miRNAs in both tumor- and normal tissue samples for all 13 patients. Notably, we found *hsa-miR-nov3* expression to be elevated in tumor compared to normal tissue in 10 out of the 13 patients ($p = 0.009$; Wilcoxon test; Fig 5A). Similar findings

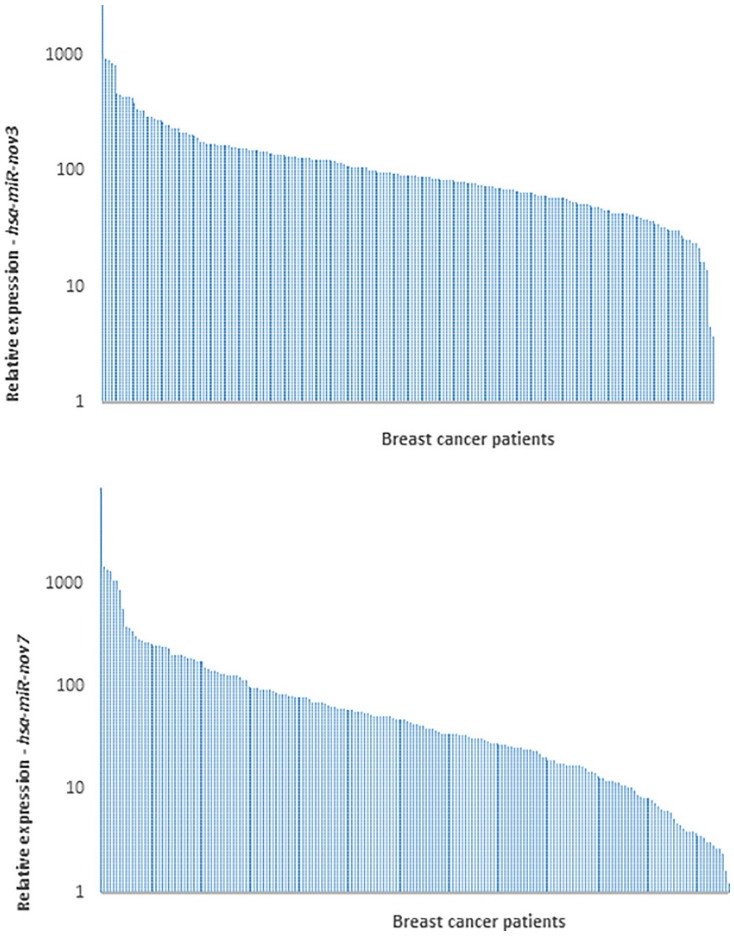


Fig 3. Expression of novel miRNAs in breast cancer tissue. Bars indicate the relative expression of *hsa-miR-nov3* (A) and *hsa-miR-nov7* (B) in 223 breast cancer patients.

https://doi.org/10.1371/journal.pone.0225357.g003

were observed for *hsa-miR-nov7* with elevated expression in 10 out of 13 tumors (Wilcoxon: $p = 0.016$; Fig 5B). The level of overexpression (i.e. the ratio of expression levels in tumor versus normal tissue) for the two miRNAs did not correlate to each other ($p>0.2$; Spearman).

Notably, overexpression of *hsa-miR-nov7* in tumor versus normal tissue was observed predominantly in ER-positive tumors (overexpression in 7 out of 7 ER-positive tumors, contrasting 3 out of 6 ER-negative tumors; $p = 0.070$; Fischer exact test).

hsa-miR-nov7 and hsa-miR-nov3 target prediction

Based on our finding of both novel miRNAs to be overexpressed in breast cancer, we next aimed to elucidate the functional roles for *hsa-miR-nov7 and hsa-miR-nov3* by identifying

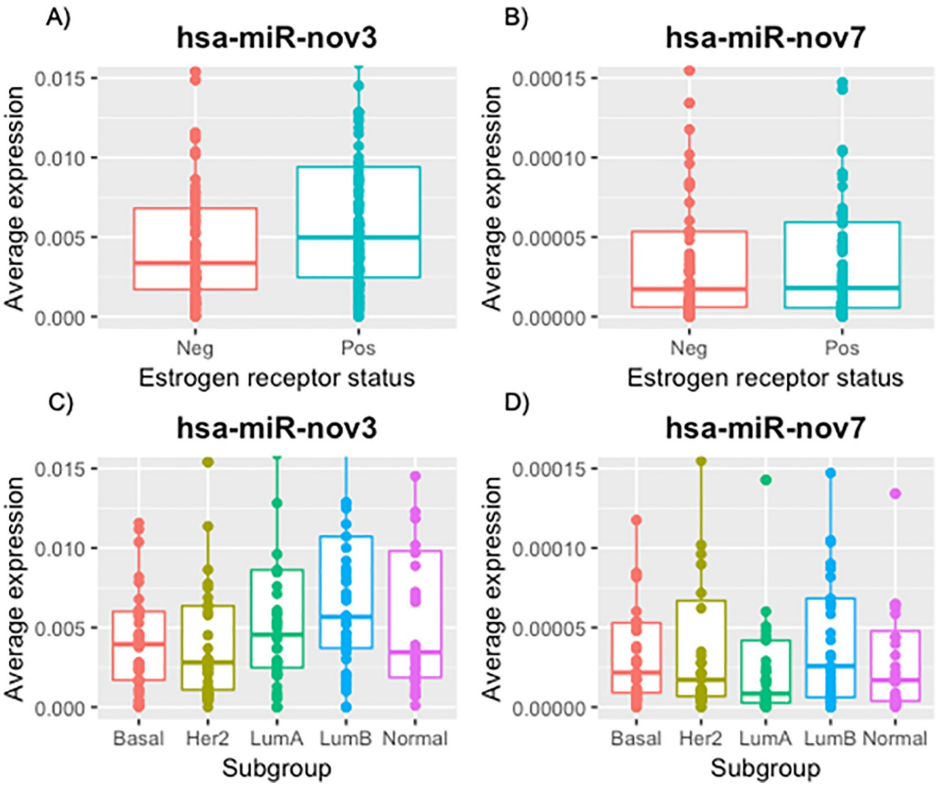


Fig 4. Expression of novel miRNAs in breast cancer tissue. Expression levels stratified by ER-status (A, B) and by expression subtypes (C, D).

https://doi.org/10.1371/journal.pone.0225357.g004

potential targets. We performed *in silico* target predictions using three different algorithms–miRanda, miRDB and TargetScan Human Custom. miRanda, which predicts possible targets from human transcripts in general, predicted 9200 and 12315 target genes for *hsa-miR-nov7 and hsa-miR-nov3*, respectively. miRDB, which contains curated and possible miRNA targets, predicted 570 and 530 target genes each for *hsa-miR-nov7 and hsa-miR-nov3*, respectively, while TargetScanHuman custom predicted 633 target genes for *hsa-miR-nov7*, and 282 target genes for *hsa-miR-nov3*. For increased stringency in our predictions, we restricted the potential targets to the ones called by all three algorithms (Fig 6). This left a total of 97 and 180 potential targets for *hsa-miR-nov3* and *hsa-mir-nov7*, respectively.

The two lists of 97 and 180 predicted gene targets were then used for KEGG pathway analysis and GO enrichment analysis using GATHER. The top 10 KEGG pathways and GO terms for each microRNAs are listed in Table 2. The KEGG and GO annotations for *hsa-miR-nov3* showed pathways that are important in cell development, communication and cytoskeletal organization. Similar analysis for *hsa-miR-nov7* unveiled pathways playing a vital role in cell functions such as communication and homeostasis. These findings were largely validated by performing the same analyses applying alternative tools (DAVID and topGO; S2 Table).

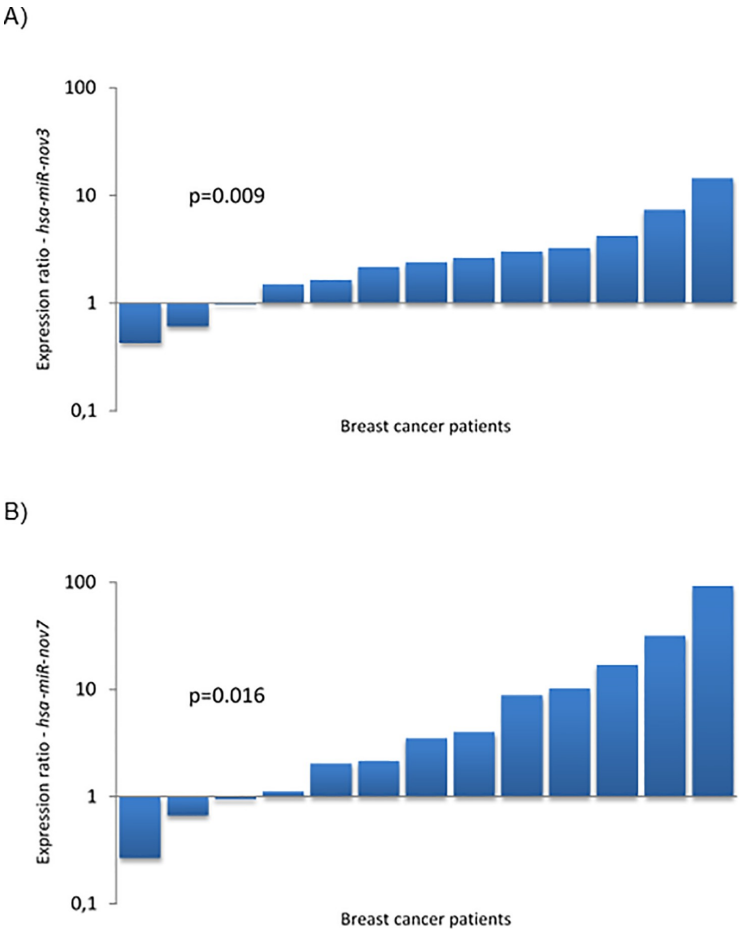


Fig 5. Expression of novel miRNAs in breast cancer tissue. Bars indicate the ratio of expression in tumour tissue vs. matched normal breast tissue in 13 breast cancer patients, for *hsa-miR-nov3* (A) and *hsa-miR-nov7* (B).

https://doi.org/10.1371/journal.pone.0225357.g005

Thus, both these miRNAs implied cell functions that are vital to cancer development and progression.

In order to further substantiate these *in-silico* predictions, we performed a complete Spearman correlation analysis between the expression levels of *hsa-miR-nov7 and hsa-miR-nov3* and mRNA expression array data available for 203 out of the 233 patients in study 1. Assuming the miRNAs, in general, to execute their function by suppressing gene expression (mRNA degradation), we restricted the analysis to genes which were negatively correlated to expression of the miRNAs. The top ranking negatively correlated genes are listed in Table 3. Notably, the only genes with Rho-values < -0.2 were *RMND5A* for *hsa-miR-nov3* and *GLUD1* and *SASH1* for *hsa-miR-nov7*. Given that the two novel miRNAs were overexpressed in breast cancer tissue, we went on to restrict the correlation analysis to an in-house list of 283-tumor suppressor

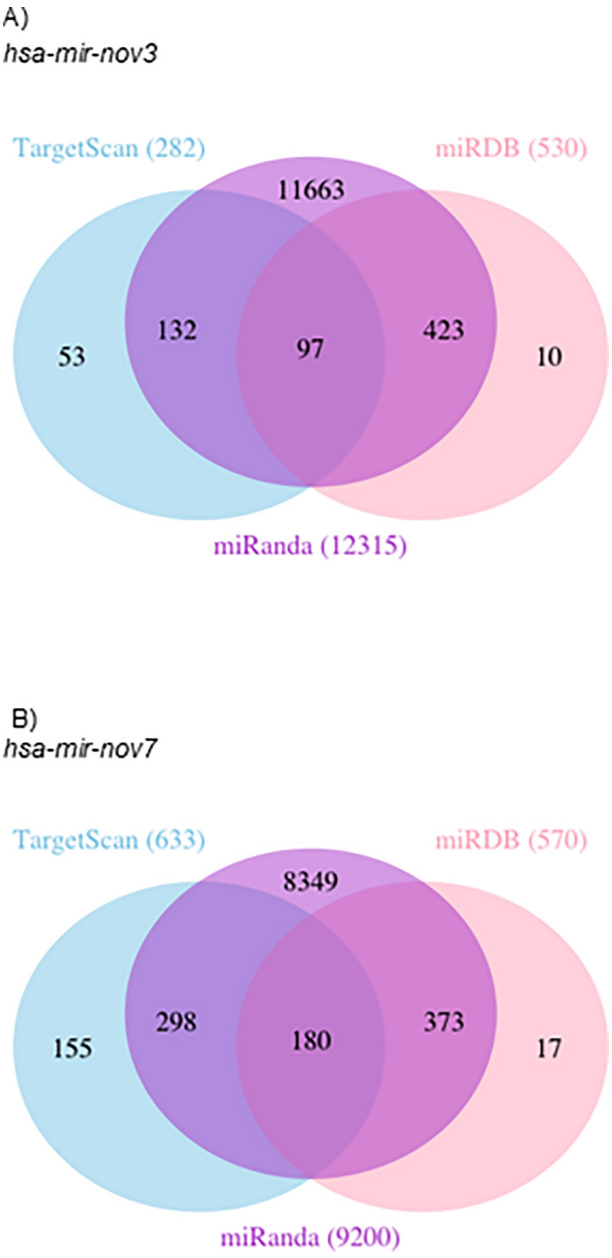


Fig 6. Target genes predicted. Venn-diagrams illustrating the number of target genes predicted by TargetScan, mirDB and Miranda for the two novel miRNAs *hsa-mir-nov3* (A) and *hsa-mir-nov7* (B).

https://doi.org/10.1371/journal.pone.0225357.g006

Table 2. Top 10 (arbitrary cut-off) GO and KEGG annotation.

A) GO annotation—*hsa-miR-nov3*

#	Annotation	ln(Bayes factor)[a]	neg ln(p value)[b]	FE: neg ln (p value)[c]	FE: neg ln(FDR)[d]
1	GO:0009653 [3]: morphogenesis	94.88	7.98	100.5	92.91
2	GO:0007275 [2]: development	87.32	7.58	92.89	85.99
3	GO:0007154 [3]: cell communication	85.41	7.46	90.99	84.5
4	GO:0009887 [4]: organogenesis	74.65	6.99	80.24	74.26
5	GO:0048513 [3]: organ development	74.65	6.99	80.24	74.26
6	GO:0007165 [4]: signal transduction	74.2	6.97	79.77	73.97
7	GO:0007242 [5]: intracellular signaling cascade	66.52	6.59	72.18	66.53
8	GO:0007010 [6]: cytoskeleton organization and biogenesis	55.54	6.04	61.15	55.63
9	GO:0009790 [3]: embryonic development	48.63	5.7	54.28	48.88
10	GO:0006928 [4]: cell motility	47.82	5.65	53.49	48.2

B) KEGG annotation—*hsa-miR-nov3*

#	Annotation	Total Genes With Ann	ln(Bayes factor)[a]	neg ln(p value)[b]	FE: neg ln (p value)[c]	FE: neg ln(FDR)[d]
1	path:hsa04810: Regulation of actin cytoskeleton	35	9.03	4.07	13.96	9.57
2	path:hsa04010: MAPK signaling pathway	36	6.93	3.75	11.8	8.1
3	path:hsa04510: Focal adhesion	32	4.15	3.26	8.94	5.94
4	path:hsa04110: Cell cycle	18	4.1	3.25	9.08	5.95
5	path:hsa04060: Cytokine-cytokine receptor interaction	33	3.23	3.07	7.97	5.24
6	path:hsa04620: Toll-like receptor signaling pathway	17	2.97	3.01	7.92	5.24
7	path:hsa04210: Apoptosis	16	2.13	2.82	7.06	4.55
8	path:hsa04512: ECM-receptor interaction	14	1.17	2.55	6.09	3.72
9	path:hsa04630: Jak-STAT signaling pathway	21	1.01	2.51	5.77	3.52
10	path:hsa05050: Dentatorubropallidoluysian atrophy (DRPLA)	5	0.7	2.41	5.9	3.59

C) GO annotation—*hsa-miR-nov7*

#	Annotation	ln(Bayes factor)[a]	neg ln(p value)[b]	FE: neg ln(p value)[c]	FE: neg ln(FDR)[d]
1	GO:0007154 [3]: cell communication	60.17	6.3	65.79	58.14
2	GO:0007275 [2]: development	54.83	6	60.38	53.43
3	GO:0007165 [4]: signal transduction	50.84	5.81	56.44	49.89
4	GO:0009653 [3]: morphogenesis	48.96	5.72	54.56	48.3

(*Continued*)

Table 2. (Continued)

5	GO:0050794 [3]: regulation of cellular process	41.31	5.3	46.94	40.9
6	GO:0009987 [2]: cellular process	40.56	5.26	46.33	40.48
7	GO:0009887 [4]: organogenesis	40.37	5.25	45.94	40.24
8	GO:0048513 [3]: organ development	39.98	5.23	45.54	40.09
9	GO:0007242 [5]: intracellular signaling cascade	39.87	5.22	45.58	40.09
10	GO:0050789 [2]: regulation of biological process	39.18	5.18	44.62	39.27

D) KEGG annotation—*hsa-miR-nov7*

#	Annotation	Total Genes With Ann	ln(Bayes factor)[a]	neg ln(p value)[b]	FE: neg ln(p value)[c]	FE: neg ln(FDR)[d]
1	path:hsa04630: Jak-STAT signaling pathway	27	5.53	3.5	10.48	6.6
2	path:hsa04350: TGF-beta signaling pathway	18	5.22	3.45	10.3	6.6
3	path:hsa04010: MAPK signaling pathway	33	3.15	3.04	7.94	4.57
4	path:hsa04210: Apoptosis	17	2.51	2.91	7.49	4.27
5	path:hsa04620: Toll-like receptor signaling pathway	17	2.28	2.85	7.25	4.24
6	path:hsa04020: Calcium signaling pathway	4	2.23	2.84	0	0
7	path:hsa00471: D-Glutamine and D-glutamate metabolism	3	1.12	2.54	6.48	3.7
8	path:hsa04510: Focal adhesion	29	0.96	2.49	5.64	3.23
9	path:hsa05030: Amyotrophic lateral sclerosis (ALS)	5	-0.17	0	5.04	2.78
10	path:hsa04512: ECM-receptor interaction	13	-0.28	0	4.61	2.39

[a] Measure of the strength of annotation

[b] p-value for the Bayes factor estimate

[c] p-value for Fishcer's exact test

[d] FDR for Fishcer's exact test

https://doi.org/10.1371/journal.pone.0225357.t002

Table 3. Spearman correlation table for *hsa-miR-nov3* and *hsa-miR-nov7* and their top 25 target genes (arbitrary cut-off for inclusion in the table; ranked by inverse correlation).

A) *hsa-miR-nov3*

Gene Symbol	Estimate	P.value	Expression (mean)
RMND5A	-0.2018	0.0038	14.0750
YES1	-0.1649	0.0184	17.0218
PALM2-AKAP2	-0.1455	0.0378	13.0997
SLC7A1	-0.1224	0.0811	16.8650
RAPGEF5	-0.1208	0.0853	14.9652
CTDSPL2	-0.1196	0.0885	15.4945
SLC4A5	-0.1077	0.1251	15.0101
HIPK1	-0.1046	0.1366	13.3737
ABHD12	-0.0998	0.1555	16.2313
FMNL2	-0.0982	0.1624	16.0939
POU4F1	-0.0933	0.1844	13.4684
RPS6KA3	-0.0905	0.1981	14.6430
LARP1	-0.0890	0.2054	15.0210
WIPI2	-0.0702	0.3184	14.7316
MTCH1	-0.0575	0.4139	18.6604
DIAPH1	-0.0528	0.4530	16.7109
MARCKS	-0.0481	0.4946	18.6286
LUZP1	-0.0453	0.5200	17.1097
DNAJC8	-0.0449	0.5238	18.2152
CLOCK	-0.0436	0.5354	15.7894
SLAMF6	-0.0415	0.5557	15.4277
CDAN1	-0.0405	0.5655	16.6394
PCDH11X	-0.0359	0.6104	13.4661
RYBP	-0.0346	0.6234	16.9184
FGF1	-0.0344	0.6249	13.9423

B) *hsa-miR-nov7*

Gene Symbol	Estimate	P.value	Expression (Mean)
GLUD1	-0.2274	0.0011	18.0399
SASH1	-0.2095	0.0026	16.9164
MARK1	-0.1883	0.0070	15.0356
ARID5B	-0.1877	0.0072	17.7569
ELOVL5	-0.1854	0.0079	17.5656
PUM1	-0.1707	0.0147	17.8295
PNRC2	-0.1599	0.0224	15.4674
UNC13B	-0.1583	0.0238	15.5633
FLRT2	-0.1581	0.0239	15.7323
ZFHX4	-0.1482	0.0344	14.7383
CHIC1	-0.1479	0.0348	13.5807
MAN1A1	-0.1457	0.0375	15.4956
CPEB2	-0.1387	0.0478	14.6995
PDE4D	-0.1377	0.0495	13.9823
TMED7	-0.1366	0.0514	17.1083
NDFIP1	-0.1280	0.0680	16.1458
CSMD1	-0.1269	0.0704	13.8158
MITF	-0.1187	0.0908	14.0482

(*Continued*)

Table 3. (Continued)

ITSN1	-0.1185	0.0915	14.8011
CTDSPL2	-0.1178	0.0932	15.4945
ATAD2B	-0.1178	0.0932	14.9892
SFRP2	-0.1129	0.1080	18.4511
DPP10	-0.1119	0.1110	13.4306
BMPR2	-0.1107	0.1149	17.1664
EIF5A2	-0.1100	0.1174	14.5450

https://doi.org/10.1371/journal.pone.0225357.t003

genes previously described. Among these tumor suppressors, we found 115 to be negatively correlated to *hsa-miR-nov7* and 119 to *hsa-miR-nov3* (S3 Table). Assessing the intersection between these negatively correlated tumor suppressor genes and the predicted targets, we obtained a list of one gene for *hsa-miR-nov3* (*ATRX*) and three genes for *hsa-miR-nov7* (*APC*, *SFRP2* and *CDH11*), but the correlations were non-significant in all 4 cases (Table 4, Fig 7).

In order to get a broader overview of potential biological function, we selected the 100 gene transcripts with the strongest positive and the top 100 gene transcripts with the strongest negative correlation to the two miRNAs (independent of previous target-predictions) and performed gene ontology analyses. We detected no cancer related pathways or cellular functions to be significantly associated with *hsa-miR-nov7* (S4 Table). However, for *hsa-miR-nov3*, KEGG analysis of the negatively correlated genes revealed associations to Hepatorcellular carcinoma as well as several pathways related to drug metabolism (S5 Table). Notably, when seeking to validate these findings by application of alternative tools (DAVID and topGO) the latter was not validated. (S6 and S7 Tables).

Expression of hsa-miR-nov7 and hsa-miR-nov3 and clinical outcome in breast cancer

Since both *hsa-miR-nov7* and *hsa-miR-nov3* were overexpressed in the tumor tissue of breast cancer patients, we assessed whether any of the two novel miRNAs were associated to clinical outcomes in study 1 (223 breast cancer patients). Given that these patients were enrolled in a prospective study specifically designed to assess response to primary chemotherapy administered as epirubicin or paclitaxel monotherapy in a neoadjuvant setting [27, 52], we assessed the association of *hsa-miR-nov7 and hsa-miR-nov3* levels with primary therapy response and with long term survival (10-years).

We found no association between any of the two novel miRNAs and primary response to either epirubicin or paclitaxel (S8 Table). Regarding survival, we observed a weak association between high levels of *hsa-miR-nov7* and poor survival in the paclitaxel treated arm of the study, with the strongest associations observed for relapse free survival, however, none of these associations reached statistical significance (Fig 8). No effect was observed in the epirubicin treated arm. Further, for *hsa-miR-nov3*, no significant correlation to outcome was recorded.

Table 4. List of intersection between correlated tumour suppressor genes and the predicted targets of *hsa-miR-nov3* and *hsa-miR-nov7*.

hsa-miR-nov3	***hsa-miR-nov7***
ATRX	APC
	CDH11
	SFRP2

https://doi.org/10.1371/journal.pone.0225357.t004

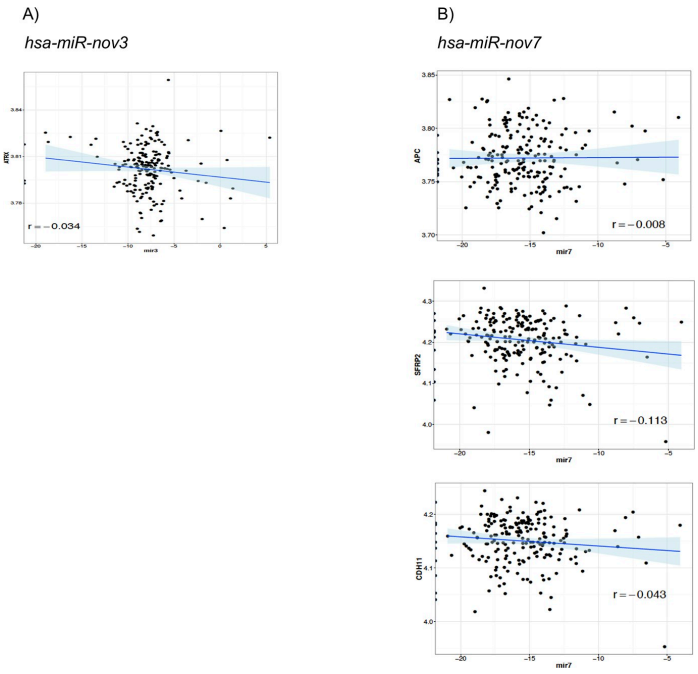


Fig 7. Correlations to tumor suppressor genes. Scatter plots showing correlation of target tumor suppressors with A) *hsa-miR-nov3* and B) *hsa-miR-nov7*.

https://doi.org/10.1371/journal.pone.0225357.g007

Given the skewed expression levels between breast cancer subtypes for *hsa-miR-nov3*, we performed survival analyses stratified for ER-status and subtypes. These analyses revealed no significant associations to survival (Log rank test p-values ranging from 0.09 to 0.98).

Discussion

We investigated whether we could detect novel, previously undescribed miRNAs and, if so, address their potential association to other defined biological parameters and to outcome in a cohort of locally advanced breast cancer. We successfully predicted 10 new miRNAs, out of which 2 were deemed reliable because of their detected presence in more than one patient. Although these two novel miRNAs (preliminary termed *hsa-miR-nov7* and *hsa-miR-nov3*) were only predicted from 8 samples among the 50 initially sequenced biopsies, we found them to be expressed in all patients by highly sensitive qPCR at varying levels. In addition to our *in vitro* validations, the qPCR detection validated the initial NGS based analysis, detecting these two miRNAs.

Since expression of the two miRNAs was confirmed in breast tumor tissue from the majority of patients analyzed, we went on to assess the relative expression levels in tumor versus matched normal breast tissue, collected from a non-tumor bearing quadrant. Our finding that both novel miRNAs had higher expression levels in tumor than in normal tissue indicates a potential functional role in breast cancer. However, although being overexpressed, the biological role of these two miRNAs in cancer should be interpreted with caution. The expression

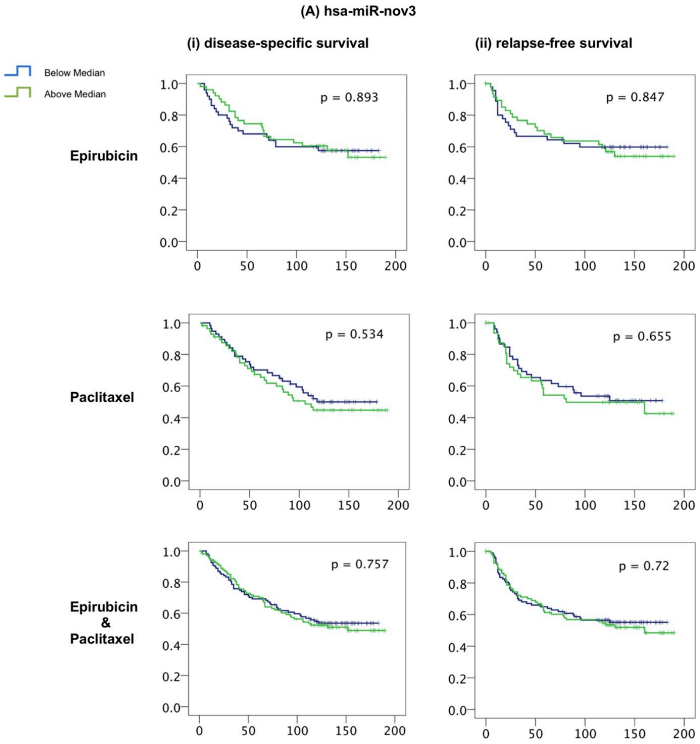


Fig 8. miRNAs and breast cancer survival. Kaplan-Meier curves showing (i) disease-specific and (ii) relapse-free survival of locally advanced breast cancer patients treated with epirubicin or paclitaxel monotherapy in the neoadjuvant setting (study 1), with respect to expression levels of (A) hsa-miR-nov3 and (B) hsa-miR-nov7 on all samples.

https://doi.org/10.1371/journal.pone.0225357.g008

levels are very low, and it is therefore uncertain whether they will have a major impact on cellular functions. Notably, given our approach and identification of the two miRNAs with low expression level, this indicates that there may currently be a limited potential for new discoveries of miRNAs high expression levels and strong functional roles in breast cancer. However, when assessing the potential functional roles of these microRNAs by *in silico* prediction of targets followed by validation using correlation to mRNA-array data, the KEGG and GO annotations for these targets revealed cellular functions of potential importance in development and progression of cancer. As such, our present findings may warrant further investigations into the functions of the two miRNAs. Notably, regarding *hsa-miR-nov3*, it was of particular interest that this miRNA was significantly higher expressed in ER-positive as compared to ER-negative breast cancers. Accordingly, we found relatively high expression levels of *hsa-miR-nov3* in tumors of the luminal and normal-like subtypes, contrasting low expression levels in basal-like and her2-like tumors [53, 54]. This finding may indicate a potential role for *hsa-miR-nov3* restricted to ER-positive tumors.

Regarding potential specific targets, we narrowed these down by first assessing the intersect of three different target prediction algorithms, and then the intersect of this result with a

predefined list of tumor suppressors. Although none of the remaining genes after this filtering had a statistically significant inverse correlation with the miRNAs, we identified some potentially interesting connections: For *hsa-miR-nov3*, we propose *ATRX* as a target. This is a gene in the SWI/SNF family, involved in chromatin remodelling, and it has previously been found subject to loss of heterozygosity (LOH) in breast cancer [55]. Importantly, we recently reported mutations in the SWI/SNF family genes to be enriched in relapsed breast cancer as compared to primary cancers [56]. Thus, this supports the hypothesis of a breast cancer promoting function for *hsa-miR-nov3*. For *hsa-miR-nov7*, we propose *APC*, *SFRP2*, and *CDH11* as potential targets. Interestingly, the two former are involved in regulation of the Wnt-signalling pathway [57–59] and both have previously been reported as targets for several miRNAs in breast cancer [60–62]. Taken together, this may imply a role for *hsa-miR-nov7* in Wnt signaling. Notably, during our work with the present project, *hsa-miR-nov7*, was identified by Lim and colleagues and coined miR-10393-3p [50]. They found this miRNA to target genes involved in chromatin modifications associated with pathogenesis of Diffuse large B-cell lymphoma (DLBCL). While this differs from our present finding, it may likely be explained by tissue specific effects of the miRNA.

Regarding any predictive or prognostic role for the two investigated miRNAs, we found no significant impact on survival. While we recorded a non-significant trend towards an association between miRnov7 expression and overall survival in the paclitaxel arm, further studies on larger patient cohorts are warranted to clarify this issue. Alternatively, the miRNAs could play a role in tumorigenesis but not later tumor progression. As such, the observed overexpression in tumor tissue compared to normal breast tissue may be a remaining signal from tumorigenesis.

Whether cancer related overexpression of the two miRNAs described here is merely consequences of other molecular mechanisms in cancer cells or whether the two miRNAs may be involved in tumorigenesis, but not subsequent cancer progression, remains unknown.

Supporting information

S1 Fig. Predicted novel miRNAs. Table on the upper left shows miRDeep2 scores and read counts. RNA secondary structure for miRNA on the top right. Color code for depiction as follows mature sequence in red, loop sequence in yellow and purple for star sequences. Density plot in the middle shows distribution of reads in precursor reads predicted. Dotted lines illustrate alignment and mm, number of mismatches. Exp, is potential precursor model predicted by algorithm with taking accounts of stability based on free energy, position and read frequencies according to Dicer/Drosha processing of miRNA. Obs, is position and reads found from deep sequencing data. (A) *hsa-miR-nov3* and (B) *hsa-miR-nov7*.
(DOCX)

S1 Table. In-house pan-cancer panel of 283 tumor suppressor genes. Panel generated based on CGPv2/3-panels [41], Roche's Comprehensive Cancer Design along with manual literature search, to filter target genes of interest.
(DOCX)

S2 Table. Predicted miRNA-targets by DAVID and topGO.
(XLSX)

S3 Table. Correlation miRNAs and tumour suppressor genes. Spearman correlation table for *hsa-miR-nov3* (A) and *hsa-miR-nov7* (B) inversely correlated tumor suppressor genes.
(DOCX)

S4 Table. Correlations mir7 and gene ontology.
(XLS)

S5 Table. Correlations mir3 and gene ontology.
(XLS)

S6 Table. Negatively correlated genes (validation analyses).
(XLSX)

S7 Table. Positively correlated genes (validation analyses).
(XLSX)

S8 Table. Statistics mir3 and mir7 versus response to treatment.
(XLSX)

S1 File. Supporting information mirDeep.
(DOCX)

Acknowledgments

We thank Beryl Leirvaag and Gjertrud T. Iversen for technical assistance.

Author Contributions

Conceptualization: Deepak Poduval, Stian Knappskog.

Data curation: Deepak Poduval, Zuzana Sichmanova, Stian Knappskog.

Formal analysis: Deepak Poduval, Zuzana Sichmanova, Anne Hege Straume, Stian Knappskog.

Funding acquisition: Per Eystein Lønning, Stian Knappskog.

Methodology: Anne Hege Straume, Stian Knappskog.

Supervision: Per Eystein Lønning, Stian Knappskog.

Writing – original draft: Deepak Poduval, Stian Knappskog.

Writing – review & editing: Deepak Poduval, Per Eystein Lønning, Stian Knappskog.

References

1. Beezhold KJ, Castranova V, Chen F. Microprocessor of microRNAs: regulation and potential for therapeutic intervention. Mol Cancer. 2010; 9:134. Epub 2010/06/03. https://doi.org/10.1186/1476-4598-9-134 PMID: 20515486; PubMed Central PMCID: PMC2887798.
2. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993; 75(5):843–54. Epub 1993/12/03. https://doi.org/10.1016/0092-8674(93)90529-y PMID: 8252621.
3. Lee Y, Jeon K, Lee JT, Kim S, Kim VN. MicroRNA maturation: stepwise processing and subcellular localization. The EMBO journal. 2002; 21(17):4663–70. Epub 2002/08/29. https://doi.org/10.1093/emboj/cdf476 PMID: 12198168; PubMed Central PMCID: PMC126204.
4. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000; 403(6772):901–6. Epub 2000/03/08. https://doi.org/10.1038/35002607 PMID: 10706289.
5. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. Science (New York, NY). 2001; 294(5543):853–8. Epub 2001/10/27. https://doi.org/10.1126/science.1064921 PMID: 11679670.

6. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, et al. A uniform system for microRNA annotation. RNA (New York, NY). 2003; 9(3):277–9. Epub 2003/02/20. https://doi.org/10.1261/rna.2183803 PMID: 12592000; PubMed Central PMCID: PMC1370393.

7. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D. MicroRNAs and other tiny endogenous RNAs in C. elegans. Current biology: CB. 2003; 13(10):807–18. Epub 2003/05/16. https://doi.org/10.1016/s0960-9822(03)00287-2 PMID: 12747828.

8. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, et al. The small RNA profile during Drosophila melanogaster development. Developmental cell. 2003; 5(2):337–50. Epub 2003/08/16. https://doi.org/10.1016/s1534-5807(03)00228-4 PMID: 12919683.

9. Mishra S, Yadav T, Rani V. Exploring miRNA based approaches in cancer diagnostics and therapeutics. Critical reviews in oncology/hematology. 2016; 98:12–23. Epub 2015/10/21. https://doi.org/10.1016/j.critrevonc.2015.10.003 PMID: 26481951.

10. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008; 26(4):407–15. Epub 2008/04/09. https://doi.org/10.1038/nbt1394 PMID: 18392026.

11. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42(Database issue):D68–73. Epub 2013/11/28. https://doi.org/10.1093/nar/gkt1181 PMID: 24275495; PubMed Central PMCID: PMC3965103.

12. Wang Y, Yu Y, Tsuyada A, Ren X, Wu X, Stubblefield K, et al. Transforming growth factor-beta regulates the sphere-initiating stem cell-like feature in breast cancer through miRNA-181 and ATM. Oncogene. 2011; 30(12):1470–80. Epub 2010/11/26. https://doi.org/10.1038/onc.2010.531 PMID: 21102523; PubMed Central PMCID: PMC3063856.

13. Lonning PE, Knappskog S, Staalesen V, Chrisanthar R, Lillehaug JR. Breast cancer prognostication and prediction in the postgenomic era. Ann Oncol. 2007; 18(8):1293–306. Epub 2007/02/24. mdm013 [pii] https://doi.org/10.1093/annonc/mdm013 PMID: 17317675.

14. Lonning PE, Knappskog S. Mapping genetic alterations causing chemoresistance in cancer: identifying the roads by tracking the drivers. Oncogene. 2013; 32:5315–30. https://doi.org/10.1038/onc.2013.48 PMID: 23474753.

15. Gotte M, Mohr C, Koo CY, Stock C, Vaske AK, Viola M, et al. miR-145-dependent targeting of junctional adhesion molecule A and modulation of fascin expression are associated with reduced breast cancer cell motility and invasiveness. Oncogene. 2010; 29(50):6569–80. Epub 2010/09/08. https://doi.org/10.1038/onc.2010.386 PMID: 20818426.

16. Rask L, Balslev E, Jorgensen S, Eriksen J, Flyger H, Moller S, et al. High expression of miR-21 in tumor stroma correlates with increased cancer cell proliferation in human breast cancer. APMIS: acta pathologica, microbiologica, et immunologica Scandinavica. 2011; 119(10):663–73. Epub 2011/09/16. https://doi.org/10.1111/j.1600-0463.2011.02782.x PMID: 21917003.

17. Harquail J, Benzina S, Robichaud GA. MicroRNAs and breast cancer malignancy: an overview of miRNA-regulated cancer processes leading to metastasis. Cancer biomarkers: section A of Disease markers. 2012; 11(6):269–80. Epub 2012/12/19. https://doi.org/10.3233/cbm-120291 PMID: 23248185.

18. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005; 65(16):7065–70. Epub 2005/08/17. https://doi.org/10.1158/0008-5472.CAN-05-1783 PMID: 16103053.

19. Nassar FJ, Nasr R, Talhouk R. MicroRNAs as biomarkers for early breast cancer diagnosis, prognosis and therapy prediction. Pharmacology & Therapeutics. 2017; 172:34–49. https://doi.org/10.1016/j.pharmthera.2016.11.012.

20. Goh JN, Loo SY, Datta A, Siveen KS, Yap WN, Cai W, et al. microRNAs in breast cancer: regulatory roles governing the hallmarks of cancer. Biological Reviews. 2016; 91(2):409–28. https://doi.org/10.1111/brv.12176 PMID: 25631495

21. Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, et al. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci U S A. 2012; 109(8):3024–9. https://doi.org/10.1073/pnas.1200010109 PMID: 22315424; PubMed Central PMCID: PMC3286983.

22. Yan LX, Huang XF, Shao Q, Huang MAY, Deng L, Wu QL, et al. MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. RNA (New York, NY). 2008; 14(11):2348–60. https://doi.org/10.1261/rna.1034808 PMID: 18812439; PubMed Central PMCID: PMC2578865.

23. Wu JZ, Yang TJ, Lu P, Ma W. Analysis of signaling pathways in recurrent breast cancer. Genetics and Molecular Research. 2014; 13(4):10097–104. https://doi.org/10.4238/2014.December.4.4 WOS:000350229200042. PMID: 25501221

24. Kashyap D, Kaur H. Cell-free miRNAs as non-invasive biomarkers in breast cancer: Significance in early diagnosis and metastasis prediction. Life Sci. 2020:117417. Epub 2020/02/12. https://doi.org/10.1016/j.lfs.2020.117417 PMID: 32044304.

25. Tashkandi H, Shah N, Patel Y, Chen H. Identification of new miRNA biomarkers associated with HER2-positive breast cancers. Oncoscience. 2015; 2(11):924–9. Epub 2015/12/24. https://doi.org/10.18632/oncoscience.275 PMID: 26697527; PubMed Central PMCID: PMC4675790.

26. Güllü G. Clinical significance of miR-140-5p and miR-193b expression in patients. 2015; 38(1):21–9. https://doi.org/10.1590/s1415-475738120140167 PMID: 25983620; PubMed Central PMCID: PMC4415571.

27. Chrisanthar R, Knappskog S, Lokkevik E, Anker G, Ostenstad B, Lundgren S, et al. CHEK2 mutations affecting kinase activity together with mutations in TP53 indicate a functional pathway associated with resistance to epirubicin in primary breast cancer. PLoS ONE. 2008; 3(8):e3062. Epub 2008/08/30. https://doi.org/10.1371/journal.pone.0003062 PMID: 18725978.

28. Lonning PE, Helle H, Duong NK, Ekse D, Aas T, Geisler J. Tissue estradiol is selectively elevated in receptor positive breast cancers while tumour estrone is reduced independent of receptor status. J Steroid Biochem Mol Biol. 2009; 117(1–3):31–41. Epub 2009/07/14. https://doi.org/10.1016/j.jsbmb.2009.06.005 PMID: 19591931.

29. Sasidharan V, Lu YC, Bansal D, Dasari P, Poduval D, Seshasayee A, et al. Identification of neoblast- and regeneration-specific miRNAs in the planarian Schmidtea mediterranea. RNA (New York, NY). 2013; 19(10):1394–404. Epub 2013/08/27. https://doi.org/10.1261/rna.038653.113 PMID: 23974438; PubMed Central PMCID: PMC3854530.

30. Krishna S, Nair A, Cheedipudi S, Poduval D, Dhawan J, Palakodeti D, et al. Deep sequencing reveals unique small RNA repertoire that is regulated during head regeneration in Hydra magnipapillata. Nucleic Acids Res. 2013; 41(1):599–616. Epub 2012/11/21. https://doi.org/10.1093/nar/gks1020 PMID: 23166307; PubMed Central PMCID: PMC3592418.

31. Knappskog S, Chrisanthar R, Lokkevik E, Anker G, Ostenstad B, Lundgren S, et al. Low expression levels of ATM may substitute for CHEK2 /TP53 mutations predicting resistance towards anthracycline and mitomycin chemotherapy in breast cancer. Breast Cancer Res. 2012; 14(2):R47. https://doi.org/10.1186/bcr3147 PMID: 22420423; PubMed Central PMCID: PMC3446381.

32. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome biology. 2003; 5(1):R1. Epub 2004/01/08. https://doi.org/10.1186/gb-2003-5-1-r1 PMID: 14709173; PubMed Central PMCID: PMC395733.

33. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS biology. 2004; 2(11):e363. Epub 2004/10/27. https://doi.org/10.1371/journal.pbio.0020363 PMID: 15502875; PubMed Central PMCID: PMC521178.

34. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic Acids Res. 2015; 43(Database issue):D146–52. Epub 2014/11/08. https://doi.org/10.1093/nar/gku1104 PMID: 25378301; PubMed Central PMCID: PMC4383922.

35. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120(1):15–20. Epub 2005/01/18. https://doi.org/10.1016/j.cell.2004.12.035 PMID: 15652477.

36. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981; 147 (1):195–7. Epub 1981/03/25. https://doi.org/10.1016/0022-2836(81)90087-5 PMID: 7265238.

37. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999; 49(2):145–65. Epub 1999/03/10. https://doi.org/10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G PMID: 10070264.

38. Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics. 2008; 24(3):325–32. Epub 2007/12/01. https://doi.org/10.1093/bioinformatics/btm595 PMID: 18048393.

39. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009; 19(1):92–105. Epub 2008/10/29. https://doi.org/10.1101/gr.082701.108 PMID: 18955434; PubMed Central PMCID: PMC2612969.

40. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular cell. 2007; 27(1):91–105. Epub 2007/07/07. https://doi.org/10.1016/j.molcel.2007.06.017 PMID: 17612493; PubMed Central PMCID: PMC3800283.

41. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015; 21(7):751–9. https://doi.org/10.1038/nm.3886 PMID: 26099045; PubMed Central PMCID: PMC4500826.

42. Chang JT, Nevins JR. GATHER: a systems approach to interpreting genomic signatures. Bioinformatics. 2006; 22(23):2926–33. Epub 2006/09/27. https://doi.org/10.1093/bioinformatics/btl483 PMID: 17000751.

43. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28 (1):27–30. Epub 1999/12/11. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173; PubMed Central PMCID: PMC102409.

44. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007; 8(9):R183. Epub 2007/09/06. https://doi.org/10.1186/gb-2007-8-9-r183 PMID: 17784955; PubMed Central PMCID: PMC2375021.

45. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. 2.38.1 ed2019. p. R package

46. Curtis C. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. 486(7403):346–52. https://doi.org/10.1038/nature10983 PMID: 22522925; PubMed Central PMCID: PMC3440846.

47. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43(7):e47. Epub 2015/01/22. https://doi.org/10.1093/nar/gkv007 PMID: 25605792; PubMed Central PMCID: PMC4402510.

48. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res. 2010; 38(3):e17. Epub 2009/11/20. https://doi.org/10.1093/nar/gkp942 PMID: 19923232; PubMed Central PMCID: PMC2817484.

49. Backes C, Fehlmann T, Kern F, Kehl T, Lenhof HP, Meese E, et al. miRCarta: a central repository for collecting miRNA candidates. Nucleic Acids Res. 2018; 46(D1):D160–D7. Epub 2017/10/17. https://doi.org/10.1093/nar/gkx851 PMID: 29036653; PubMed Central PMCID: PMC5753177.

50. Lim EL, Trinh DL, Scott DW, Chu A, Krzywinski M, Zhao Y, et al. Comprehensive miRNA sequence analysis reveals survival differences in diffuse large B-cell lymphoma patients. Genome biology. 2015; 16:18. Epub 2015/02/28. https://doi.org/10.1186/s13059-014-0568-y PMID: 25723320; PubMed Central PMCID: PMC4308918.

51. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747–52. http://www.nature.com/nature/journal/v406/n6797/suppinfo/406747a0_S1.html. https://doi.org/10.1038/35021093 PMID: 10963602

52. Chrisanthar R, Knappskog S, lokkevik E, Anker G, Ostenstad B, Lundgren S, et al. Predictive and prognostic impact of TP53 mutations and MDM2 promoter genotype in primary breast cancer patients treated with epirubicin or paclitaxel. PLoS ONE. 2011; 6(4):e19249. https://doi.org/10.1371/journal.pone.0019249 PMID: 21556366

53. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747–52. https://doi.org/10.1038/35021093 PMID: 10963602.

54. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001; 98 (19):10869–74. https://doi.org/10.1073/pnas.191367098 PMID: 11553815; PubMed Central PMCID: PMC58566.

55. Roy D, Guida P, Zhou G, Echiburu-Chau C, Calaf GM. Gene expression profiling of breast cells induced by X-rays and heavy ions. International journal of molecular medicine. 2008; 21(5):627–36. Epub 2008/04/22. https://doi.org/10.3892/ijmm.21.5.627 PMID: 18425356.

56. Yates LR, Knappskog S, Wedge D, Farmery JHR, Gonzalez S, Martincorena I, et al. Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell. 2017; 32(2):169–84.e7. https://doi.org/10.1016/j.ccell.2017.07.005 PMID: 28810143; PubMed Central PMCID: PMC5559645.

57. von Marschall Z, Fisher LW. Secreted Frizzled-Related Protein-2 (sFRP2) Augments Canonical Wnt3a-induced Signaling. Biochem Biophys Res Commun. 2010; 400(3):299–304. https://doi.org/10.1016/j.bbrc.2010.08.043 PMID: 20723538; PubMed Central PMCID: PMC2952323.

58. Rattner A, Hsieh JC, Smallwood PM, Gilbert DJ, Copeland NG, Jenkins NA, et al. A family of secreted proteins contains homology to the cysteine-rich ligand-binding domain of frizzled receptors. Proc Natl Acad Sci U S A. 1997; 94(7):2859–63. https://doi.org/10.1073/pnas.94.7.2859 PMID: 9096311; PubMed Central PMCID: PMC20287.

59. Hankey W, Frankel WL, Groden J. Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: Implications for therapeutic targeting. Cancer metastasis reviews. 2018; 37(1):159–72. https://doi.org/10.1007/s10555-017-9725-6 PMC5803335. PMID: 29318445

60. Isobe T, Hisamori S, Hogan DJ, Zabala M, Hendrickson DG, Dalerba P, et al. miR-142 regulates the tumorigenicity of human breast cancer stem cells through the canonical WNT signaling pathway. eLife.

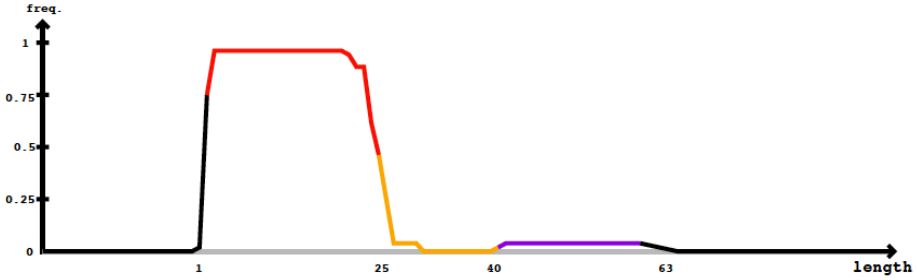
2014;3. Epub 2014/11/19. https://doi.org/10.7554/eLife.01977 PMID: 25406066; PubMed Central PMCID: PMC4235011.

61. Tan Z, Zheng H, Liu X, Zhang W, Zhu J, Wu G, et al. MicroRNA-1229 overexpression promotes cell proliferation and tumorigenicity and activates Wnt/beta-catenin signaling in breast cancer. Oncotarget. 2016; 7(17):24076–87. Epub 2016/03/19. https://doi.org/10.18632/oncotarget.8119 PMID: 26992223; PubMed Central PMCID: PMC5029685.

62. Liu S, Wang Z, Liu Z, Shi S, Zhang Z, Zhang J, et al. miR-221/222 activate the Wnt/β-catenin signaling to promote triple negative breast cancer. Journal of Molecular Cell Biology. 2018:mjy041–mjy. https://doi.org/10.1093/jmcb/mjy041 PMID: 30053090

A)

Provisional ID	:	chr3_6970
Score total	:	25.6
Score for star read(s)	:	3.9
Score for read counts	:	20.5
Score for mfe	:	0.2
Score for randfold	:	1.6
Score for cons. seed	:	-0.6
Total read count	:	52
Mature read count	:	50
Loop read count	:	0
Star read count	:	2

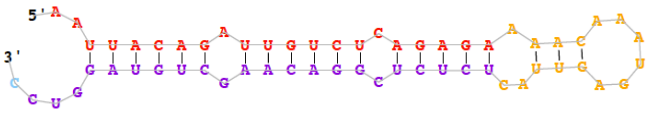


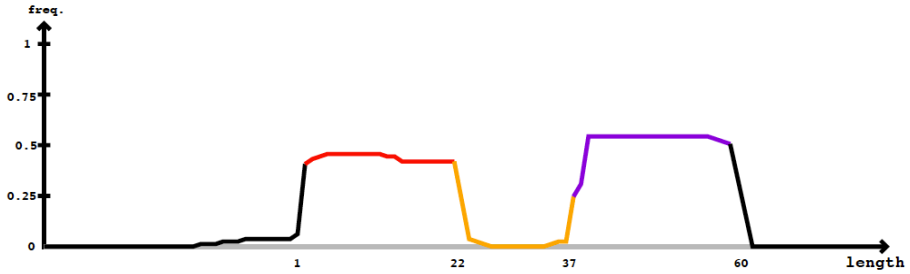
Mature Star

```
5'- aaaaugucuggauuuccacuaaagcaggauucagacuacaauauagcugcuaagugcugugguugucguuccccccugcuuaaaauaaaguuguuucuuaacuauaccugucug  -3'   obs
    aaaaugucuggauuuccacuaaagcaggauucagacuacaauauagcugcuaagugcugugguugucguuccccccugcuuaaaauaaaguuguuucuuaacuauaccugucug        exp
    ........(((....)))...(((((((.....(((.((((((((((........)))))))))).)))...)))))))........(((((.....)))))..........  reads  mm
    ..................uaaagcaggauucagacuacaau.....................................................................  1      0
    ..................aaagcaggauucagacuac.....................................................................  1      0
    ..................aaagcaggauucagacuaca.....................................................................  2      0
    ..................aaagcaggauucagacuaAa.....................................................................  1      1
    ..................aaagcaggauucagacuacaau...................................................................  10     0
    ..................aaagcaggauucagacuacaaua..................................................................  6      0
    ..................aaagcCggauucagacuacaauau.................................................................  2      1
    ..................aaagcaggauucagacuacaauau.................................................................  14     0
    ..................aaagcaggauucagacuacaauaua................................................................  1      0
    ..................aaagcaggauucagacuacaauauagcug............................................................  1      0
    ..................aagcaggauucagacuacaau....................................................................  3      0
    ..................aagcaggauucagacuacaaua...................................................................  2      0
    ..................aagcaggauucagacuacaauau..................................................................  5      0
    ..................aagcaggauucagacuacaauauagcug.............................................................  1      0
    ..........................................................uguugucguuccccccugcuu.....................................  1      0
    ..........................................................guugucguuccccccugcuu......................................  1      0
```

B)

Provisional ID	:	chr7_2556
Score total	:	38.3
Score for star read(s)	:	3.9
Score for read counts	:	32.8
Score for mfe	:	0.7
Score for randfold	:	1.6
Score for cons. seed	:	-0.6
Total read count	:	76
Mature read count	:	34
Loop read count	:	0
Star read count	:	42





Mature Star

```
5'- aauuaccuuacuuaaggauaauaggacaagacaaauuacagauugucucagagaaaacaaaugaguuacucucucggacaagcuguaggucuaccuaaauguccagcagg -3'   obs
    aauuaccuuacuuaaggauaauaggacaagacaaauuacagauugucucagagaaaacaaaugaguuacucucucggacaagcuguaggucccuaccuaaauguccagcagg   exp
    ..............(((((.((((....(((....(((((.(((((((.(((((.(((......))))..))))))..)))))).)))))).)))...))))..))))).....  reads  mm
    ..................aauaggacaagacaaauuacagauugu.....................................................................  1      0
    .....................aggUcaagacaaauuacagauu.....................................................................  1      1
    ........................acaagCcaaauuacagauugu.....................................................................  1      1
    ..............................aaauuacagauugucucagaga.....................................................................  2      0
    ..............................aauAacagauugucucagaga.....................................................................  1      1
    ..............................aauuacagauugucucagaga.....................................................................  17     0
    ..............................aauuacagauugucucagagaa....................................................................  10     0
    ..............................auuacagauugucucagaga.....................................................................  1      0
    ..............................auuacagauugucucagagaaa....................................................................  1      0
    ..............................uuacagauugucucagagaaaa...................................................................  1      0
    ..............................uacagauugucucagagaaaac...................................................................  1      0
    ..........................................................uacucucucggacaagcuguag.....................................  1      0
    ..........................................................acucucucggacaagcuguagA.....................................  1      1
    ..........................................................ucucucggacaagcuguaggu.....................................  1      0
    ..........................................................ucucucggacaagcuguagguU.....................................  16     1
    ..........................................................ucucucggacaagcuguagguA.....................................  1      1
    ..........................................................cucucggacaagcuguagguU.....................................  5      1
    ...........................................................ucucggacaagcuguagguU.....................................  4      1
    ...........................................................ucucggacaagcuguagguc.....................................  4      0
    ...........................................................ucucggacaagcuguaggucU....................................  7      1
    ...........................................................ucucggacaagcuguaggucc....................................  4      0
```

Figure 1: Predicted novel miRNAs. Table on the upper left shows miRDeep2 scores and read counts. RNA secondary structure for miRNA on the top right. Color code for depiction as follows mature sequence in red, loop sequence in yellow and purple for star sequences.

Density plot in the middle shows distribution of reads in precursor reads predicted. Dotted lines illustrate alignment and mm, number of mismatches. Exp, is potential precursor model predicted by algorithm with taking accounts of stability based on free energy, position and read frequencies according to Dicer/Drosha processing of miRNA. Obs, is postion and reads found from deep sequencing data. (A) *hsa-miR-nov3* and (B) *hsa-miR-nov7*.

Supporting Table S1. In-house pan-cancer panel of 283 tumor suppressor genes, generated based on CGPv2/3-panels[1], Roche's Comprehensive Cancer Design along with manual literature search, to filter target genes of interest.

Gene Name	Chromosome
AIP	chr11
ALDH2	chr12
AMER1	chrX
APC	chr5
AR	chrX
ARHGAP26	chr5
ARHGEF12	chr11
ARID1A	chr1
ARID1B	chr6
ARID2	chr12
ARID4A	chr14
ASXL1	chr20
ATM	chr11
ATR	chr3
ATRX	chrX
AXIN2	chr17
BAP1	chr3
BARD1	chr2
BCL7A	chr12
BLID	chr11
BLM	chr15
BMP2	chr20
BMP3	chr4
BMP4	chr14
BMP7	chr20
BMPR1A	chr10
BRCA1	chr17
BRCA2	chr13
BRIP1	chr17
BTG1	chr12
BUB1B	chr15
CARS	chr11
CASC5	chr15
CASP8	chr2
CCDC6	chr10
CCNB1IP1	chr14

CD2	chr1
CDC73	chr1
CDH1	chr16
CDH11	chr16
CDH13	chr16
CDK12	chr17
CDK2AP2	chr11
CDKN1A	chr6
CDKN1B	chr12
CDKN1C	chr11
CDKN2A	chr9
CDKN2B	chr9
CDKN2C	chr1
CDKN2D	chr19
CDX2	chr13
CEBPA	chr19
CHD5	chr1
CHD6	chr20
CHEK1	chr11
CHEK2	chr22
CHFR	chr12
CHN1	chr2
CIC	chr19
CIITA	chr16
CLDN3	chr7
CLDN4	chr7
CLTCL1	chr22
CNBP	chr3
COX6C	chr8
CREB3L1	chr11
CREBBP	chr16
CTCFL	chr20
CTNNB1	chr3
CYLD	chr16
DAPK1	chr9
DDB2	chr11
DDIT3	chr12
DDX53	chrX
DICER1	chr14
DKK1	chr10
DNMT3A	chr2
EBF1	chr5
EIF4A2	chr3

ELAC2	chr17
EMP3	chr19
EP300	chr22
EPHA5	chr4
EPHA6	chr3
EPHB6	chr7
ERCC1	chr19
ERCC2	chr19
ERCC3	chr2
ERCC4	chr16
ERCC5	chr13
ERG	chr21
ESR1	chr6
EXT1	chr8
EXT2	chr11
FAM46C	chr1
FANCA	chr16
FANCB	chrX
FANCC	chr9
FANCD2	chr3
FANCE	chr6
FANCF	chr11
FANCG	chr9
FANCI	chr15
FANCL	chr2
FANCM	chr14
FAS	chr10
FAT1	chr4
FBXO11	chr2
FBXW7	chr4
FH	chr1
FHIT	chr3
FHL1	chrX
FLCN	chr17
FOXL2	chr3
FOXO1	chr13
FOXO3	chr6
FOXO4	chrX
FUS	chr16
GAS7	chr17
GATA1	chrX
GATA2	chr3
GATA3	chr10

GATA4	chr8
GATA5	chr20
GMPS	chr3
GPC3	chrX
GSTM1	chr1
GSTP1	chr11
HAND2	chr4
HECW1	chr7
HERPUD1	chr16
HIC1	chr17
HNF1A	chr12
HOXA10	chr7
HOXA11	chr7
HOXA9	chr7
ID4	chr6
IGFBP3	chr7
IKZF1	chr7
IL21R	chr16
KDM5C	chrX
KDM6A	chrX
KDSR	chr18
KEAP1	chr19
KL	chr13
KLF6	chr10
KMT2C	chr7
KMT2D	chr12
LMNA	chr1
LRP5	chr11
LTBP2	chr14
MAL	chr2
MC1R	chr16
MEN1	chr11
MGMT	chr10
MIR124-1	chr8
MIR127	chr14
MIR155	chr21
MLF1	chr3
MLH1	chr3
MLLT11	chr1
MNX1	chr7
MRE11A	chr11
MSH2	chr2
MSH6	chr2

MTUS2	chr13
MUTYH	chr1
NBN	chr8
NCKIPSD	chr3
NDRG1	chr8
NF1	chr17
NF2	chr22
NFKB2	chr10
NTRK3	chr15
NUMA1	chr11
OPTN	chr10
PALB2	chr16
PAX5	chr9
PBRM1	chr3
PDCD1LG2	chr9
PER1	chr17
PGR	chr11
PHF6	chrX
PLAG1	chr8
PML	chr15
PMS1	chr2
PMS2	chr7
PRDM1	chr6
PRDM16	chr1
PRDM2	chr1
PREX2	chr8
PRKAR1A	chr17
PRKDC	chr8
PRLR	chr5
PTCH1	chr9
PTEN	chr10
PTGS2	chr1
PTPN6	chr12
PTPRD	chr9
PYCARD	chr16
RAB40AL	chrX
RABEP1	chr17
RAD51B	chr14
RAD51C	chr17
RAD51D	chr17
RANBP17	chr5
RAP1GDS1	chr4
RASSF1	chr3

RASSF5	chr1
RB1	chr13
RBBP8	chr18
RBM15	chr1
RBP1	chr3
RHOH	chr4
RMI2	chr16
RNASEL	chr1
RPTOR	chr17
RRM1	chr11
RUNX1	chr21
RUNX1T1	chr8
RUNX3	chr1
SARDH	chr9
SBDS	chr7
SDHAF2	chr11
SDHB	chr1
SDHC	chr1
SDHD	chr11
SETD2	chr3
SFPQ	chr1
SFRP1	chr8
SFRP2	chr4
SFRP5	chr10
SLC5A8	chr12
SLX4	chr16
SMAD2	chr18
SMAD3	chr15
SMAD4	chr18
SMARCA4	chr19
SMARCB1	chr22
SNCG	chr10
SOCS1	chr16
SOCS3	chr17
SPECC1	chr17
SPEN	chr1
SRGAP3	chr3
STK11	chr19
SUFU	chr10
SYK	chr9
TCEA1	chr8
TET1	chr10
TFAP2A	chr6

TFG	chr3
TGFBR2	chr3
THBS1	chr15
THRAP3	chr1
TIMP3	chr22
TLX3	chr5
TMEFF2	chr2
TMEM127	chr2
TNFAIP3	chr6
TOP2A	chr17
TP53	chr17
TP63	chr3
TP73	chr1
TRIM33	chr1
TSC1	chr9
TSC2	chr16
TSHR	chr14
TTL	chr2
TUBB3	chr16
VDR	chr12
VHL	chr3
WIF1	chr12
WRN	chr8
XPA	chr9
XPC	chr3
YWHAE	chr17
ZBTB16	chr11
ZMYM2	chr13
ZNF331	chr19
ZNF668	chr16
ZRSR2	chrX
sep.09	chr17

1. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21(7):751-9. Epub 2015/06/23. doi: 10.1038/nm.3886. PubMed PMID: 26099045; PubMed Central PMCID: PMCPMC4500826.

Supporting Table S2. Spearman correlation table for *hsa-miR-nov3* (A) and *hsa-miR-nov7* (B) inversely correlated tumor suppressor genes.

A)

Gene	Estimate	P value	Mean Expression
ARHGAP26	-0.0065	0.9262	14.4132
ARID2	-0.0573	0.4160	15.5138
ARID4A	-0.0134	0.8493	14.6255
ATRX	-0.0483	0.4923	13.9469
BCL7A	-0.0926	0.1876	14.1301
BLID	-0.0529	0.4521	13.9765
BMP3	-0.0124	0.8600	13.3724
BMP7	-0.0004	0.9954	14.3881
BRIP1	-0.0486	0.4899	13.8642
CARS	-0.0213	0.7627	15.8107
CDH11	-0.0518	0.4618	17.7525
CDKN1A	-0.0217	0.7581	15.6141
CDKN1C	-0.0144	0.8379	16.0217
CDKN2A	-0.0842	0.2309	14.3086
CDKN2B	-0.0310	0.6595	15.0449
CDX2	-0.0527	0.4537	13.5056
CHD5	-0.0314	0.6555	13.7651
CHD6	-0.0826	0.2401	15.8235
CHFR	-0.0805	0.2522	17.0289
CHN1	-0.0013	0.9850	14.8979
CIITA	-0.0062	0.9298	14.5010
CLTCL1	-0.1132	0.1071	14.1098
CNBP	-0.0703	0.3177	16.3166
CTNNB1	-0.1208	0.0853	15.9912
DAPK1	-0.1016	0.1480	15.6762
DDB2	-0.0454	0.5191	15.2302
DDIT3	-0.0077	0.9129	16.0630
DDX53	-0.0291	0.6794	13.5698
DKK1	-0.0004	0.9954	14.3922
ELAC2	-0.1488	0.0336	16.1477

EMP3	-0.0232	0.7418	17.4996
EP300	-0.0645	0.3596	16.1435
EPHA6	-0.0616	0.3815	13.3497
ERCC1	-0.0019	0.9788	15.8715
ERCC3	-0.1056	0.1326	15.5411
ERCC4	-0.0049	0.9449	13.9151
EXT2	-0.0937	0.1824	15.4237
FANCF	-0.0183	0.7952	13.9110
FANCI	-0.0072	0.9187	15.5721
FANCL	-0.0878	0.2119	14.9136
FAT1	-0.0280	0.6909	17.3311
FBXO11	-0.0567	0.4205	16.6127
FHIT	-0.0265	0.7071	14.5957
FHL1	-0.0015	0.9831	16.5500
FLCN	-0.0163	0.8168	14.3452
FOXL2	-0.0424	0.5466	13.3440
GATA2	-0.0145	0.8364	14.8464
GATA3	-0.0301	0.6686	16.8666
GATA4	-0.0356	0.6128	13.4347
GATA5	-0.1094	0.1193	13.5412
GSTM1	-0.0315	0.6543	15.4934
HAND2	-0.0792	0.2600	13.8374
HIC1	-0.0022	0.9746	15.0303
HOXA10	-0.0662	0.3468	14.3840
HOXA11	-0.0274	0.6970	13.1974
HOXA9	-0.0077	0.9126	13.3157
IGFBP3	-0.0380	0.5898	17.3472
KDM6A	-0.0797	0.2569	16.1162
KDSR	-0.0348	0.6209	16.9388
KMT2C	-0.0108	0.8783	14.4106
LMNA	-0.0361	0.6079	17.5236
MLH1	-0.0409	0.5617	16.5589
MLLT11	-0.0638	0.3646	16.2646
MSH2	-0.0441	0.5311	14.7562
MSH6	-0.0715	0.3095	17.5512
MUTYH	-0.0348	0.6212	15.0585

NF1	-0.0844	0.2298	14.0124
NTRK3	-0.0140	0.8428	13.8522
NUMA1	-0.0489	0.4877	16.7333
PAX5	-0.0796	0.2576	13.2127
PGR	-0.0248	0.7245	14.3941
PHF6	-0.0058	0.9346	13.4835
PMS1	-0.0219	0.7559	15.2495
PRDM1	-0.0004	0.9959	14.3366
PRDM2	-0.0488	0.4879	14.1393
PREX2	-0.0313	0.6568	13.2312
PRKAR1A	-0.0308	0.6624	16.5292
PRKDC	-0.0809	0.2502	15.4145
PTPRD	-0.0449	0.5235	14.4119
RAB40AL	-0.0991	0.1583	13.2453
RABEP1	-0.0687	0.3290	17.3346
RANBP17	-0.0771	0.2728	13.2919
RAP1GDS1	-0.0263	0.7083	15.7314
RASSF1	-0.0445	0.5278	15.2786
RBBP8	-0.0025	0.9714	14.6984
RBP1	-0.1739	0.0128	16.9774
RNASEL	-0.0623	0.3763	16.1561
RPTOR	-0.0351	0.6183	14.5302
RUNX1	-0.0227	0.7477	15.3497
RUNX1T1	-0.0367	0.6023	14.0611
SDHAF2	-0.0895	0.2033	17.3818
SDHB	-0.1287	0.0665	17.9386
SETD2	-0.0426	0.5454	17.2496
SFRP2	-0.0279	0.6922	18.4511
SMAD2	-0.0062	0.9302	14.1585
SMAD3	-0.0963	0.1708	16.4175
SMAD4	-0.0969	0.1679	17.3092
SMARCA4	-0.0297	0.6736	17.6729
SOCS1	-0.0933	0.1843	15.7624
SOCS3	-0.0454	0.5192	14.8089
SPEN	-0.0853	0.2251	17.2409
SRGAP3	-0.0210	0.7651	14.4660

STK11	-0.0854	0.2244	15.3924
TCEA1	-0.0642	0.3619	15.5099
TET1	-0.0380	0.5899	14.1420
TFAP2A	-0.0262	0.7096	15.2176
THBS1	-0.0850	0.2270	17.5950
THRAP3	-0.1031	0.1421	15.7921
TIMP3	-0.0010	0.9886	17.6612
TLX3	-0.0153	0.8277	13.4299
TMEFF2	-0.0201	0.7758	13.5501
TMEM127	-0.0346	0.6228	15.6207
TP73	-0.0772	0.2723	13.1015
TRIM33	-0.0358	0.6112	15.7066
TSC1	-0.0164	0.8158	15.6086
TSC2	-0.0982	0.1622	15.2237
YWHAE	-0.0193	0.7836	15.5636
ZNF331	-0.0028	0.9683	14.5922
ZNF668	-0.0229	0.7448	16.0503

B)

Gene	Estimate	P value	Mean Expression
ATRX	-0.0013	0.9850	13.9469
BLID	-0.1188	0.0906	13.9765
BLM	-0.0764	0.2777	14.6813
BMP7	-0.0179	0.7990	14.3881
BRCA1	-0.0176	0.8027	14.9580
BRCA2	-0.0017	0.9810	13.4017
BRIP1	-0.1504	0.0318	13.8642
BTG1	-0.0686	0.3295	18.6391
BUB1B	-0.0014	0.9841	14.8483
CASC5	-0.0248	0.7244	13.9560
CCNB1IP1	-0.0924	0.1886	14.7104
CD2	-0.0996	0.1563	16.4594
CDH1	-0.0434	0.5375	18.1575
CDH11	-0.0247	0.7257	17.7525
CDK12	-0.1043	0.1375	16.3371

CDKN1A	-0.0502	0.4759	15.6141
CDKN1C	-0.0564	0.4230	16.0217
CDKN2A	-0.1455	0.0379	14.3086
CHD5	-0.0255	0.7168	13.7651
CHD6	-0.0342	0.6275	15.8235
CHEK1	-0.0315	0.6542	15.1712
CHEK2	-0.1116	0.1120	14.5868
CHFR	-0.0462	0.5116	17.0289
CHN1	-0.1318	0.0602	14.8979
CIITA	-0.1004	0.1530	14.5010
CLTCL1	-0.0321	0.6484	14.1098
CNBP	-0.1700	0.0151	16.3166
COX6C	-0.0124	0.8605	19.2678
CTNNB1	-0.0667	0.3435	15.9912
DAPK1	-0.1735	0.0131	15.6762
DDIT3	-0.0390	0.5795	16.0630
DKK1	-0.0305	0.6653	14.3922
ELAC2	-0.0832	0.2367	16.1477
EMP3	-0.0929	0.1864	17.4996
EPHA5	-0.0245	0.7284	13.2599
EPHA6	-0.0975	0.1653	13.3497
ERCC2	-0.0053	0.9396	15.5522
ERCC3	-0.0620	0.3782	15.5411
EXT2	-0.0684	0.3307	15.4237
FAM46C	-0.0795	0.2583	16.9997
FANCA	-0.0477	0.4977	13.7107
FANCC	-0.0026	0.9707	14.0886
FANCE	-0.0240	0.7338	15.8101
FANCI	-0.0836	0.2344	15.5721
FANCL	-0.1365	0.0515	14.9136
FAS	-0.0390	0.5793	14.8132
FAT1	-0.0256	0.7162	17.3311
FBXW7	-0.0699	0.3205	14.7353
FHIT	-0.1178	0.0932	14.5957
FOXL2	-0.1411	0.0441	13.3440
FOXO1	-0.0258	0.7146	16.4032

GATA5	-0.1373	0.0502	13.5412
GMPS	-0.0775	0.2704	17.0642
GSTP1	-0.0014	0.9842	18.4426
HERPUD1	-0.0438	0.5337	17.6315
IGFBP3	-0.0180	0.7983	17.3472
IKZF1	-0.0348	0.6209	14.8119
IL21R	-0.1125	0.1092	14.0628
LMNA	-0.0583	0.4075	17.5236
MAL	-0.0373	0.5964	14.6452
MLF1	-0.0316	0.6541	14.8246
MLH1	-0.0214	0.7617	16.5589
MLLT11	-0.0240	0.7332	16.2646
MSH2	-0.1006	0.1521	14.7562
MSH6	-0.0823	0.2420	17.5512
MTUS2	-0.0431	0.5402	13.2167
MUTYH	-0.0187	0.7906	15.0585
NF2	-0.0952	0.1756	13.5838
PALB2	-0.0779	0.2681	15.3981
PAX5	-0.0889	0.2063	13.2127
PHF6	-0.0703	0.3176	13.4835
PLAG1	-0.0161	0.8191	13.6951
PML	-0.0076	0.9144	13.8870
PMS1	-0.0374	0.5955	15.2495
PRDM1	-0.0945	0.1790	14.3366
PREX2	-0.0556	0.4293	13.2312
PRKAR1A	-0.0942	0.1804	16.5292
PRKDC	-0.0798	0.2563	15.4145
PRLR	-0.0225	0.7497	14.7949
PTPRD	-0.0128	0.8563	14.4119
RAB40AL	-0.1972	0.0047	13.2453
RAD51B	-0.0780	0.2675	13.9763
RAD51C	-0.1491	0.0333	15.4363
RAD51D	-0.0693	0.3245	14.4466
RANBP17	-0.0678	0.3356	13.2919
RASSF5	-0.0508	0.4708	14.5639
RBBP8	-0.0517	0.4629	14.6984

RBP1	-0.1176	0.0940	16.9774
RMI2	-0.0087	0.9022	16.1039
RPTOR	-0.0373	0.5968	14.5302
RUNX3	-0.0945	0.1786	15.1754
SDHAF2	-0.2143	0.0021	17.3818
SDHB	-0.1346	0.0550	17.9386
SLX4	-0.0092	0.8961	15.8230
SMAD2	-0.0204	0.7720	14.1585
SMARCA4	-0.0266	0.7057	17.6729
SOCS1	-0.0468	0.5064	15.7624
SPECC1	-0.0448	0.5249	14.1781
SRGAP3	-0.0357	0.6122	14.4660
STK11	-0.0150	0.8319	15.3924
SYK	-0.0274	0.6969	16.3501
TCEA1	-0.0029	0.9671	15.5099
TET1	-0.0135	0.8485	14.1420
TFAP2A	-0.0213	0.7627	15.2176
THRAP3	-0.0139	0.8436	15.7921
TMEFF2	-0.0639	0.3639	13.5501
TMEM127	-0.0150	0.8318	15.6207
TNFAIP3	-0.0886	0.2078	16.0023
TOP2A	-0.0328	0.6418	16.9282
TP73	-0.0708	0.3145	13.1015
TUBB3	-0.0276	0.6949	16.2461
VDR	-0.0587	0.4045	14.2406
XPA	-0.0031	0.9653	15.5822
ZNF668	-0.0059	0.9328	16.0503
ZRSR2	-0.0079	0.9110	16.0608

miRDeep algorithm

This algorithm exploits Dicer's miRNA precursor processing along with integrating massively parallel sequencing data into a simple probabilistic model. First it searches for potential precursor secondary structure and reads corresponding to them. It makes sure that precursor sequence has reads aligning to mature, star and loop of precursor, by-products of Dicer processing. Algorithm searches for both structural and miRNA signatures and scores precursor sequences based on of energetically stable it is, conservation on phylogenetic distance, number of reads mapping to it as wells number of reads aligning to all three Dicer products. The algorithm rejects reads that align to multiple positions (>5) in the genome. It also does not consider reads that map in to already annotated non - coding RNA regions in the genome. Rest of the reads are assessed both structural stability and closeness to miRNA signatures. The precursor sequence is assigned mature position based on where majority of the reads align followed potential star sequence with a base pairing with an overhang that is phylogenetically conserved typically of lengths 2 – 3 nts. Stem loop sequence with at least 14 nts between mature and star sequence that can form an unbifurcated hairpin structure is defined. Finally miRDeep2 give probability score for each predicted sequence, ruling out the possibility of background hairpin formation.
[1, 2].

1. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008;26(4):407-15. Epub 2008/04/09. doi: 10.1038/nbt1394. PubMed PMID: 18392026.
2. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40(1):37-52. doi: 10.1093/nar/gkr688. PubMed PMID: 21911355; PubMed Central PMCID: PMCPMC3245920.

Corrigendum

In Paper III, during printing at the publisher figures 1B and 8B were left out. These are now in the process of being published separately from the publisher as a corrigendum and will be available online in the very near future.

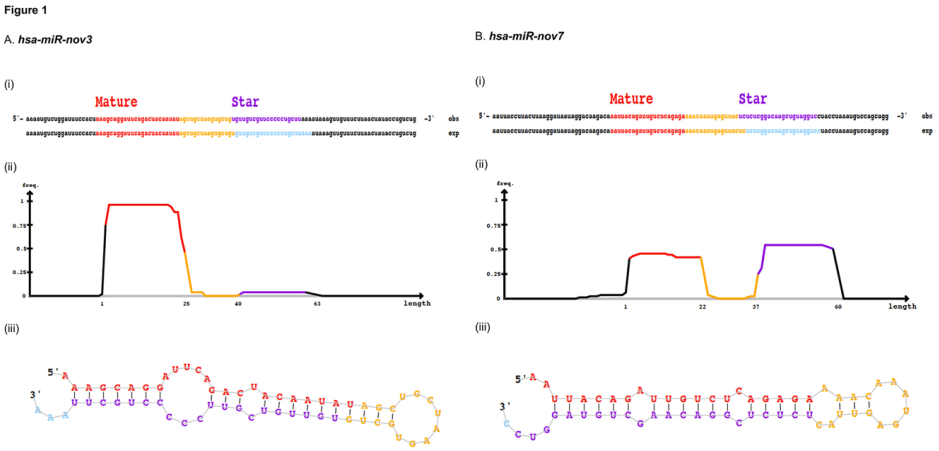
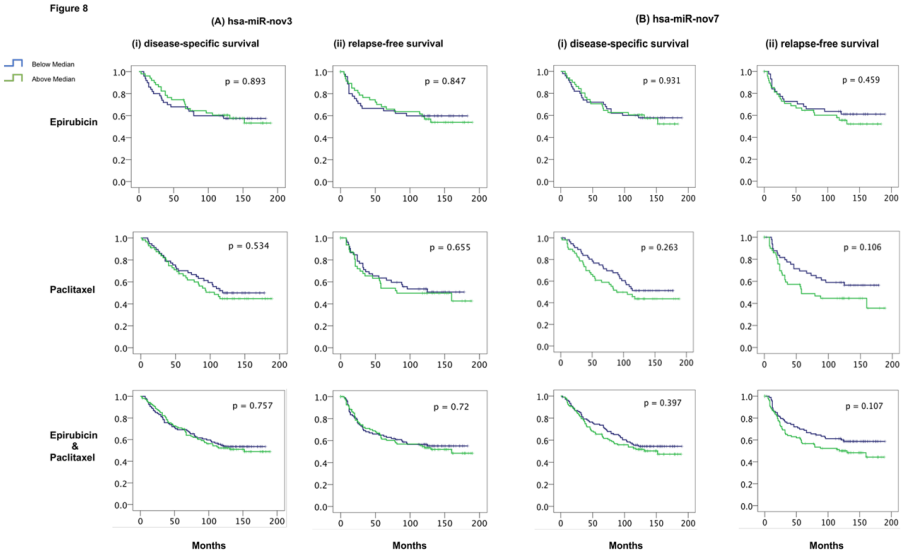


Figure 8


(A) hsa-miR-nov3
(B) hsa-miR-nov7
(i) disease-specific survival
(ii) relapse-free survival
Below Median
Above Median
Epirubicin
Paclitaxel
Epirubicin & Paclitaxel
p = 0.893
p = 0.847
p = 0.931
p = 0.459
p = 0.534
p = 0.655
p = 0.263
p = 0.106
p = 0.757
p = 0.72
p = 0.397
p = 0.107
Months

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230845424 (print)
9788230842942 (PDF)