Molecular classification of breast carcinomas

Cand. Med. Hege Elisabeth Gierckskey Russnes

The Department of Pathology
and
The Department of Genetics
Institute for Cancer research

Oslo University Hospital Radiumhospitalet

Oslo 2009, Thesis for the degree Doctor Philosophiae

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 979

ISBN 978-82-8072-585-1

Cover: Inger Sandved Anfinsen.
Printed in Norway: AiT e-dit AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Table of contents

Acknowledgements

As a third year student in Medical School I was introduced to surgical pathology, and after spending an entire summer as an assistant at the Department of Pathology at Radiumhospitalet, I was captured by the mysteries of pathogenesis. I am indebted to Professor Jahn M. Nesland for teaching me surgical pathology and for shearing his enthusiasm for the cellular and sub-cellualar microcosmos of complex tissues with me. I appreciate that you always had your door open for me and for making me reach a little higher than I dared to. I would also like to thank you for giving me the responsibility first for the Laboratory of Diagnostic Immunohistochemistry, later for the Laboratory of Molecular Pathology. Both times I felt the task frightening, but with the support from you and the excellent staff it turned out to be the most important periods in my career as a medical doctor. I am also grateful to everyone at Laboratory of Electron Microscopy who gave me my first laboratory training, teaching me Immunohistochemistry and for still helping me with tissue arrays and troublesome antibodies! I want to acknowledge all colleagues at the Department of Pathology for widening my perspective of pathological processes and for teaching me the fundamental principles in classification during my three years as a resident in surgical pathology. Of special importance was Dr. Wenche Reed who guided my education and helped keeping me on track with the residency program, and Prof. Aasmund Berner and Prof. Bjørn Risberg who encouraged me and inspired me to aim at combining residency with research. A special thank to Dr. Elin Borgen who is always ready for a good discussion and for being a supportive friend. I would also like to thank Prof. Jan Delabie for teaching me all I know about Molecular Pathology. It was an almost impossible task to take over the responsibility for the two "mol pat" laboratories after him, but his and Anne Tierens' continuous presence and participation in the diagnostic work have been crucial. Thank you also for freeing me from all diagnostic work in the last phase of my PhD project! This thesis has gained so much from my training in surgical and molecular pathology and I have tried to keep it as a backbone throughout this project.

I was introduced by Prof. Jahn M. Nesland to the Dept. of genetics and Prof. Anne-Lise Børresen-Dale and Prof. Ragnhild Lothe in my fourth year at Medical School. I would like to thank Prof. Ragnhild Lothe who taught me how to design and plan a study and guided me carefully through my first project. That project was the start of a more than ten years relationship between the Dept. of Pathology, the Dept. of Genetics and me. The major reason that this has been successful is the dynamic leadership of Prof. Anne-Lise Børresen-Dale. Trying to do a PhD at the same time as a residency, leading a lab and raising a family has been challenging and would have been undoable was it not for your enthusiasm and encouragement! Your clear mind, exceptional memory and broad knowledge combined with your generosity are irresistible and truly inspiring! I appreciate that you are open-minded and care to hear about what ideas I might have, and for introducing me to so many people that turned out to be crucial for this project. I have learned so much from you from numerous larger and smaller discussions ranging from molecular biology and cancer pathogenesis to ethical concerns and politics. I admire that you always share your thoughts and ideas and strive to make people cooperate to be able to develop better diagnostics and treatment for cancer patients! I feel lucky to be part of the Dept. of Genetics, a large department with many fellow PhD students, master students, post docs, researchers and technicians with different background, personalities and projects. I have always felt welcomed and integrated into this "family", and have learned so much from you all! I would especially like to thank Dr. Therese Sørlie for teaching me about gene expression analysis and for numerous discussions about the intrinsic classification in particular, and Eldri, Gry, Hilde, Laila and Anita who has always, since my first time in the lab as a student, had an answer to my questions.

I am grateful to all coauthors from USA, Sweden, Belgium, the Netherlands, United Kingdom, Russia and Norway. I would in particular thank Prof. James Hicks for enlightening discussions via numerous emails and phone calls and for hosting me during my visits at Cold Spring Harbor. I appreciate that you always wanted to hear my opinions and gave me constructive feedback! Likewise have I learnt so much from Prof. Anders Zetterberg. I don't know how many times you have answered one of my questions with "…yes, we looked at that several years ago and found that…". From our conversations I have learnt that not all findings by modern "high-tech" technologies are novel! These close collaborations would have been difficult without support from Radiumhospitalets Legater, who funded all my travels to Stockholm and to Cold Spring Harbor. I am also grateful for Prof. Ole Christian Lingjærdes pursuits to make the world of bioinformatics less scaring and more comprehensible for me. When one of the projects almost stalled, your persistence and all your intricate ideas was the clue to progression! Working closely with you and with Dr. Hans-Kristian Moen Vollan lifted this work several levels both academically and socially! Our frequent meetings have been very intense and constructive, but almost always intermingled with jokes and laughter.

Despite the molecular focus of this project I have tried to keep it clinically relevant. This has only been possible by Dr. Bjørn Naume and his dedication to breast cancer diagnostics, treatment and research. I am impressed by your wide knowledge and the wise decisions you make, and for your encouragement and the support you have shown me. Likewise has the "MicroMet" lab with all its dedicated ladies been of outmost importance.

This work would never been done without the endless support from family and friends. Thanks to my father for encouraging creativity of all kinds, to my mother for always being concerned about me and my family, to my mother-in-law for her care and to Tone & Esben, Erik & Gyda, Inger Marie & Chuck and Jan Helge & Rita and all my 11 nephews and nieces; me and my family always enjoy spending time with you! And to all friends and my "dame doktor klubb" in particular; thank you for always encouraging me to keep working and for good company, long talks and much fun!

But most important, thank you dear Kjell Magne and Jørgen, Anna, Ragne and Inger; for your love, affection and joy and for being patient with me in periods where I have been working too much and spending too many days away from you! Thank you Kjell Magne for your never ending love, for making me feel important, for cheering me up and for devoting yourself to the care of our children.

Last I would like to acknowledge all breast cancer patients who participate in projects. Even knowing that this kind of research do not have impact on their own fate, they answer questions and undergo additional procedures to donate tissue, blood and bone marrow. This project would have been impossible without such a commitment from both patients and health workers, and I hope the results will be a small contribution to the emerging knowledge finally leading to a cure for breast cancer.

Boston, May 25. 2010

Hege G. Russnes

Abbreviations

ABC: Avidin-Biotin Complex technique
ASCAT: Allele Specific Copy number Aberrations in Tumors
BAC: Bacterial Artificial Chromosomes
BRCA1: Breast Cancer gene 1
BRCA2: Breast Cancer gene 2
CAAI: Complex Arm Aberration Index
CK18: Cytokeratin 18
CK5/6: Cytokeratin 5 and 6
CNP: Copy Number Polymorphisms
CNV: Copy Number Variations
CpG: Cytocine and guanine base separated by phosphate (C-phosphate-G)
DAPI: 4'-6-diamino-2-phenylindole
DCIS: Ductal Carcinoma In Situ
EGFR: human Epidermal Growth Factor Receptor
ER: Estrogen Receptor
EST: Expressed Sequence Tags
FFPE: Formalin-Fixated, Paraffin-Embedded tissue
FISH: Fluorescent In Situ Hybridization
GA: Genetic Algorithm
GATA3: GATA binding protein 3
HSR: Homologous Staining Regions
IDC: Infiltrating Ductal Carcinoma
IHC: Immunohistochemistry
ILC: Infiltrating Lobular Carcinoma
LCIS: Lobular Carcinoma In Situ
LOH: Loss of heterozygosity
MOMA: Methylation Oligonucleotide Microarray Analysis
MSP: Methylation specific PCR
PCF: Piecewise Constant Fit
PCR: Polymerase Chain Reaction
PgR: Progesteron receptor
ROMA: Representational Oligo Microarray Analysis
RT-PCR: Reverse transcriptase PCR
SMA: Smooth Muscle Actin
SNP: Single Nucleotide Polymorphism
SSI: Stemline Scatter Index
TDLU: Terminal Ductal-Lobular Unit
WAAI: Whole Arm Aberration Index
aCGH: array Comparative Genomic Hybridization
cDNA: complementary DNA
erbB2/ERBB2/HER2: Human Epithelial growth factor Receptor 2

List of papers

Paper I:
Paired distribution of molecular subtypes in bilateral breast carcinomas
Hege G. Russnes, Ekatherina Sh. Kuligina, Evgeny N. Suspitsin, Ekaterina S. Jordanova, Cees J. Cornelisse, Anne-Lise Børresen-Dale, Evgeny N. Imyanitov
Under review, Molecular Oncology

Paper II:
Genomic architecture characterizes tumor progression paths and fate in breast cancer patients
Hege G. Russnes, Hans Kristian Moen Vollan, Ole Christian Lingjærde, Alexander Krasnitz, Pär Lundin, Bjørn Naume, Therese Sørlie, Elin Borgen, Inga H. Rye, Anita Langerød, Suet-Feung Chin, Andrew E. Teschendorff, Philip J. Stephens, Susanne Månér, Ellen Schlichting, Lars O. Baumbusch, Rolf Kåresen, Michael P. Stratton, Michael Wigler, Carlos Caldas, Anders Zetterberg, James Hicks, Anne-Lise Børresen-Dale
Submitted, Nature Medicine

Paper III:
Subtype dependent alterations of the DNA methylation landscape in breast cancer and implications for prognosis
Sitharthan Kamalakaran, Hege G. Russnes, Angel Janevski, Dan Levy, Jude Kendall, Vinay Varadan, Michael Riggs, Nilanjana Banerjee, Marit Synnestvedt, Ellen Schlichting, Rolf Kåresen, Robert Lucito, Michael Wigler, Nevenka Dimitrova, Bjørn Naume, Anne-Lise Børresen-Dale, James B. Hicks
Manuscript

Paper IV:

Novel tool reveals Allele Specific Copy number Aberrations in Tumors (ASCAT)

Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Charles M. Perou, Bjørn Naume, Anne-Lise Børresen-Dale, Vessela N. Kristensen

Submitted, Nature Biotechnology

Preface

A brief history of breast cancer treatment, diagnostics and research

Tumors in the breast was described as early as on papyrus from ancient Egypt (3000-2500 BC) but until the 19th century the only treatment offered women with breast carcinoma was high risk surgery. The 19th century reformed the diagnostics and treatment of cancer in general as both anesthetics and antiseptic surgery was introduced. In 1895 Wilhelm von Roentgen discovers the x-rays, which in 1899 is reported to be used to cure a cancer patient. Marie and Pierre Curies discovery of the radioactive element Radium in 1898 was later of major importance in cancer treatment. At both sides of the Atlantic, radical mastectomy was introduced and further developed by Charles Moore, William banks and William Halsted. There were debates concerning the type of surgery; some claimed that women's ribs should be removed while others tried to minimize the surgery and instead combine the treatment with radiation. Other important debates were whether tumor cells spread through lymph- or blood vessels. The treatment of breast cancer made a shift during the fifties with the introduction of chemotherapy, and in the following decades both the combination strategy and adjuvant chemotherapy were major breakthroughs in breast cancer treatment. At the same time, as the results from independent randomized trials lead by Veronesi and Fisher were published, breast conserving surgical techniques were favored. The development of lymph node mapping/sentinel node biopsy technique led to less extensive axillary surgery, reducing the negative side effects of surgery for women without lymph node involvement.

The pathologist Rudolph Virchow (1821-1902) was crucial in the development of microscopic examination of tissue and in defining cellular pathology as a medical discipline. He demonstrated that cancer rises from collections of diseased cells, and is known for his statement "omnis cellula e cellula" meaning that every cell has risen from another cell. Von Hansemann and Boveri were crucial for the discovery of chromosomes being the seats of cell hereditary and for describing the disruption of these highly organized structures in cancer cells. In 1925 Greenough proposed that breast cancer is more than one disease, and from survival data he deduce that there are three different classes of malignancy. In 1957 the Bloom and Richardson grading was published, a

modified form of this is the histological grading system used today. Steinthals division of tumors into stages (later developed by Denoix (the TNM classification) was a significant improvement in preoperative assessment, and a modified version is used today combining pathology and clinical information to guide treatment choices for the individual patient.

A major contribution to the improved outcome of the disease is the introduction of systemic adjuvant treatment and radiotherapy. The discovery of the effect of removing the ovaries on breast cancer growth was published in 1896 by George Beatson, but estrogen was first discovered in 1925 in urine from pregnant women, and estrogen receptor (ER) was frequently found in breast carcinomas. Tamoxifen (a drug proposed to have anti estrogen effect) was first used as a treatment for breast cancer in 1969, and the largest effect was seen in postmenopausale women. Brodie discovered in 1982 that a known aromataseinhibitor could stop tumor growth. In 1995 Gustafsson discovered a second estrogen receptor and the dual effect on hormone receptor therapy get more evident leading to the concept "SERMs", selective estrogen receptor modulators. In 1965 started Nissen Meyer the first multicentre trial with cyclophosfamide and showed an increased survival rate. This was followed by several studies showing a survival benefit for the combination regimen of cyclophosfamide, metotrexate and 5-fluorouracil (CMF). There have been performed several large scaled clinical trials addressing the effect of adjuvant systemic treatment on breast cancer. Furthermore, the results of these studies have also been registered in the European Breast Cancer Trialist Collaborative Group (EBCTCG). Analysis of these pooled data with a high number of individuals with long clinical follow up provide a strong basis for developing guidelines for evidence based clinical treatment of this complex and important patient population. Adjuvant treatment is now evolving rapidly with more drugs to choose from. Therapy targeted to a specific molecule is proposed to be the next revolution in cancer treatment; it makes it possible to tailor the choice of therapy for each woman aiming at getting maximum effect with a minimum of side effects. One example of this approach is Trastuzumab, the HER2 receptor binding drug that has been introduced to women whose tumors have increased number of the receptor. The research focusing on molecular alterations in breast carcinomas have been enormous. In 1979 the tumor protein 53 (TP53) was identified by Levine, Lane and Old and the gene was cloned in 1983. One year later the human

epidermal growth factor receptor EGFR was discovered and in the following year human epidermal growth factor receptor 2 (HER2/neu/erbB2) by Weinberg. The breast cancer gene 1 and 2 (*BRCA1* and *BRCA2*) was discovered by Skolnic in 1994 and by Stratton in 1995 respectively, pinpointing genomic alterations explaining a fraction of hereditary breast cancer.

Mammography used for early detection of breast cancer at an early phase was introduced a century ago, but was systemized first in 1963 by Shapiro and Strax. This was followed by several studies of mammography as screening of healthy individuals confirming the advantages in increased survival among patients detected by mammography. The official advice in Norway is now mammography screening of all females in the age group 50-69 years.

The focus on women's physical but also psychological condition after breast cancer diagnosis and treatment became more in focus during the 70's and 80's. It is fascinating to see the historical shift in the perception of this "common" disease. New knowledge and improved techniques have made it possible to move from the conception of breast cancer as "one disease-one treatment" to the more ominous view that both patient related factors such as age, tumor characteristics (such as molecular alterations) and clinical findings must all be taken into consideration to tailor the therapy. The last decade's research performed on large national and international trials testing new drugs, combination of drugs or drugs tailored to selected groups of patients show promising results As will be discussed later in this thesis, the introduction of high resolution methods such as microarrays and more recently deep-sequencing has increased the knowledge of molecular alterations in breast cancer enormously. More detailed diagnostics are already making attribution to the clinical decision making, and this will continue resulting in better disease control and less side effect of treatment for the individual woman.

(Sources; Brystkreft- diagnostikk og behandling; Novartisserien, faghefte nr. 12, 2007, The history of Breast Cancer; Breast Cancer Campaign, London 2009 and Weinberg RA, In retrospect: The chromosome trail, Nature 453, 725, 2008)

General Introduction

Classification aims at defining groups of distinct entities and to specify a relationship between them. Scientific taxonomy is applied to several disciplines including cancer biology. To date the classification of breast carcinomas are based on morphological criteria and molecular analyses applied in breast cancer diagnostics have been of prognostic or predictive value. This study has been focusing on identifying robust subgroups of breast cancer by analyzing multiple different features in breast tumors. The conclusions from the four separate studies presented in this thesis add knowledge about breast cancer subtypes and tumor progression and are presented and discussed together with a review of other studies in the field. The advantages and limitations of the materials and methods used are discussed separately after a summary of each paper.

Epidemiology

Incidence and risk factors

The incidence of female breast cancer varies worldwide and is markedly higher in high income countries such as North America and Western Europe[1]. Breast cancer is rarely diagnosed before age 30 but risk increases with age, and BC is the most frequent diagnosed cancer in women in Norway (2761 new cases in 2007) and has the highest cumulative risk with about 1:12 women diagnosed with breast cancer during lifetime[2]. The incidence has been and is still increasing, this is considered both as a result of demographic changes (population growth and ageing), increased ability to diagnose the disease and mass screening but also reflects a real increase in risk[2, 3].

Breast cancer is partly a hormone related disease, the most important risk factors being early menarche, low parity, late age at first pregnancy, late menopause and hormonal exposure[4]. More recently ageing is also considered a major risk factor[1]. Age specific incidence of breast cancer shows a plateau midlife termed Clemmesen's hook, often attributed to menopause[5]. Another interpretation of this phenomenon is that the

incidence curve reflects two major types of breast cancer; one ER negative, early onset type and one ER positive with late onset.

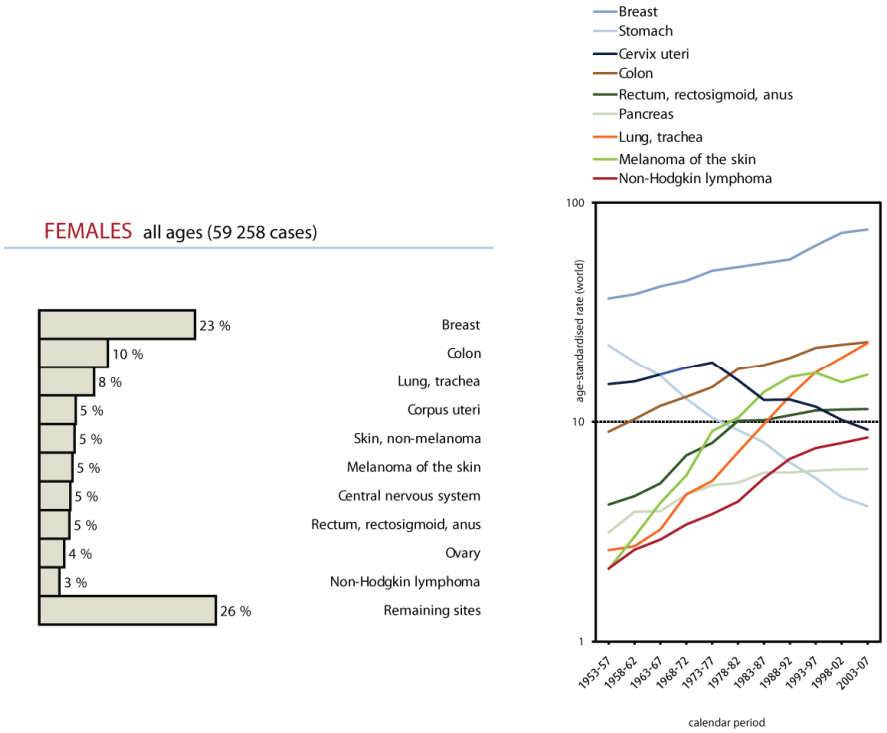


Figure 1: The barchart to the left illustrates that breast cancer has the highest incidence among Norwegian women (2003-2007). The graph to the right show the increase in incidence seen in the period 1953-2007(From The Norwegian Cancer Registery[2])

Bilateral breast cancer is rare and accounts for approximately 5% of breast cancer cases, and women with bilateral disease have a higher mortality than women with unilateral disease[6]. The incidence of bilateral disease diagnosed at the same time or within a short time span (synchronous disease) is increasing, while the incidence of bilateral tumors with a longer time span (metachronous disease) is decreasing[6]. This is probably reflecting the effect of increased use of adjuvant therapy; it having a preventive effect on developing contralateral disease. Daughters of mothers with bilateral disease have a higher risk of breast cancer[7] reflecting a hereditary component in bilateral disease.

Breast cancer in patients with either a strong family history of breast cancer or harboring a germline defect in high penetrance cancer susceptibility genes such as *BRCA1*, *BRCA2*, *TP53*, *PTEN* and *ATM* are defined as hereditary breast cancer and is estimated to be contributing up to 10% of all cases[4].

Mortality

Breast cancer is the major cause of death among adult women in high income countries but in Norway, the risk of dying of the disease seems to decline[3]. Both the incidence and survival was found to be increasing rapidly in Norway during the 1990's, partially because of the introduction of mass screening and increased use of adjuvant therapy [8]. The 15 year survival is slightly above 70%, but markedly less for the lower and higher age groups (<30 and >75 years). Survival increases to 90% given they survive 5 years, but the long term cumulative survival continues to decline many years after diagnosis[3].

Anatomy and histology of the breast gland

The breast gland

The female breast, serving the important function of producing and providing milk to our offspring, has a dynamic response to the changing hormonal phases during a woman's lifetime. Prepubertal breasts have rudimentary glandular structures, which during the extreme hormonal changes during puberty develops into 15-20 lobes that terminate into separate openings in the nipple (Fig. 2). Every lobe has a branching network of ducts draining smaller units called lobules, each composed of smaller secretory units called alveoli. This unit is called TDLU (terminal ductal-lobular unit) and is considered the functional unit of the breast. Both the amount of glandular structures and the surrounding fibro-adipose tissue are dependent on the hormonal status (menstrual cycle, pregnancy, lactating-, premenopausal- and postmenopausal state). The final differentiation stage is achieved during pregnancy and lactation by the formation of lobulo-acinar structures.

The breast epithelium is two layered surrounded by a basement membrane. The outer layer is composed of contractile myoepithelial cells and the inner layer of polarized, luminal cells where some have exocrine properties (Fig. 3B).

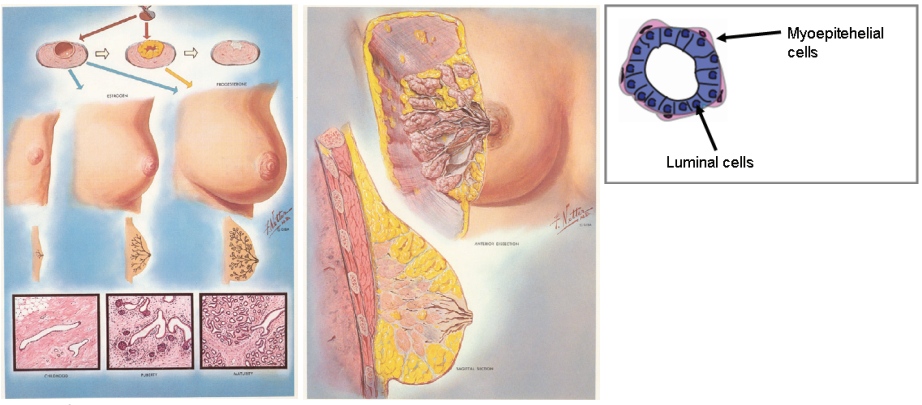


Figure 2:

Left: The changes of the female breast during puberty with development of lobes with ducts and lobules.

Middle:The branching network of ducts draining the lobules surrounded by tissue rich in fat. From Netter/Elsevier.

Right: An illustration of the organization of the two main celltypes in a duct.

The hierarchy of breast epithelial cells

In hematology, the knowledge about hierarchical relationship between stem cells and mature cells of different lineages have been acknowledge for some time[9], but for the cell types in the breast such relationship has just started to emerge[10]. The hierarchical relationship was suspected more than a decade ago as cells with specific combinations of cytokeratins was found by IHC in fetal and infant breasts[11]. The dynamic properties of breast epithelium demand compartments of stem cells and progenitor cells; i.e. cells with high proliferation potential and ability to differentiate. They reside in a protective and highly controlled region called the stem cell niche, and it seems evident that this is located in the TDLU regions[10, 12, 13]. The main cell-types, luminal and myoepithelial cells, likely represent mature cells from separate lineages but originating from the same stem cell and bipotent progenitor as is illustrated in Figur 3[14-16].

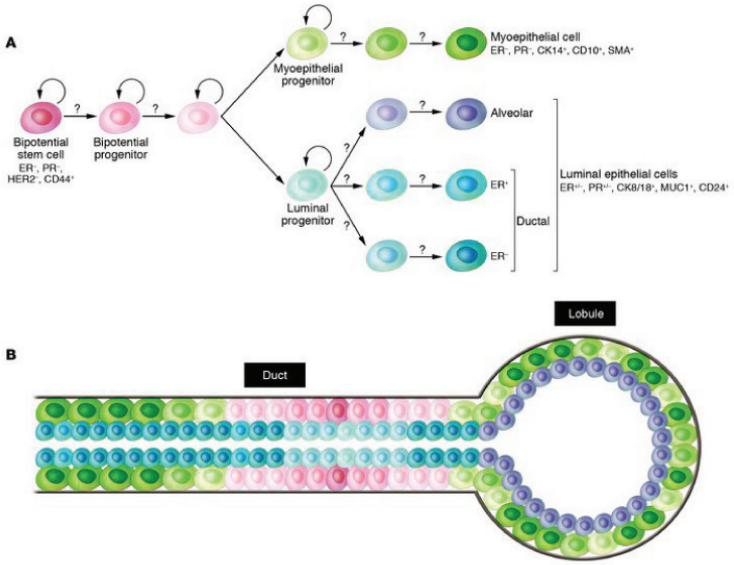


Figure 3:

A: An illustration of the assumed hierarchy of breast epithelial cells reflecting the relationship between the stem cell, the various progeny and the major mature cells.

B: The stem cell and the bipotent progenitors reside in the TDLU area while the more differentiated cells are residing either in the basal layer (myoepithelial cells) or the inner, duct-lining layer (luminal epithelial cells). From Polyak 2007[15].

A stem cell has the ability to self renew and to generate more specialized cells by differentiation. This is stepwise, where the first (and less differentiated) offspring are called progenitor cells. These cells have lost the capacity to self-renew, but are rapid proliferating cells capable to give rise to more differentiated cells needed as a response to external signals due to puberty, pregnancy or other demands. As indicated in Figure 3, several molecular markers seem to identify cells at different stages, but as the hierarchy probably is much more complex than the one exemplified, there are to be expected that this will change[15].

Morphological classification of breast cancer

Invasive carcinoma

Microscopic examination of BC reveals heterogeneity both at the architectonical and the cellular level[17]. The compositions of carcinomas can range from stroma rich tumors with glandular structures of tumor cells with minimal atypia to solid growth of large, highly atypic carcinoma cells. Breast carcinomas are commonly classified according to the World Health Organization's (WHO) recommendations[17]. The dominating growth pattern determines the type; this way a tumor with predominant tubular differentiation will be recognized as a distinct entity as will a tumor with either apocrine, lobular, cribriform, mucinous, medullary features etc. Such tumors are called 'special types', and WHO recognizes 18 different types (Fig. 4). Of the special types lobular carcinomas are most common (10-15%) while others are extremely rare (<1%). The most frequent histological type is ductal carcinomas ('invasive ductal carcinoma not otherwise specified (NOS)')[17]. Ductal carcinomas are a heterogeneous group of tumors that do not have sufficient characteristics of either of the special differentiation patterns to fall into any of those groups. Several of the rare subgroups have different clinical course and outcome[17, 18]. Mixed types are common and histological type has no major part in the Norwegian treatment guidelines to date.

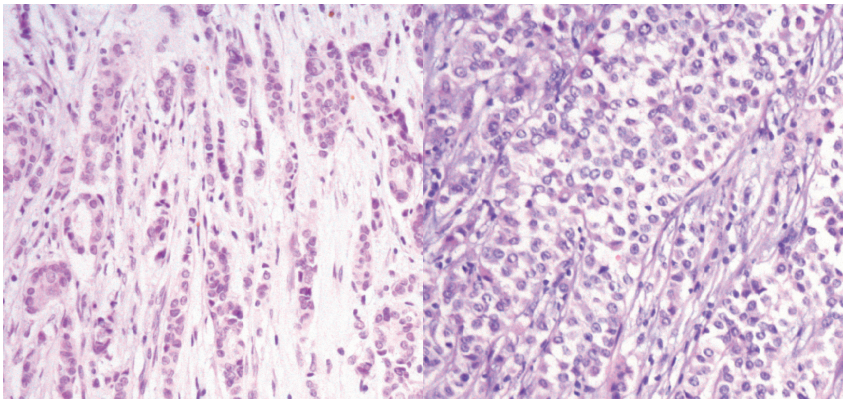


Figure 4: Left: Illustration of a of stroma rich ductal carcinoma with high differentiation, to the right a solid growing high grade invasive ductal carcinoma (HEx20)

Preinvasive neoplasia

Among intraductal proliferative lesions, WHO recognizes usual ductal hyperplasia, flat epithelial atypia and atypical ductal hyperplasia in addition to ductal and lobular carcinoma in situ (DCIS and LCIS). The relationship between such lesions and invasive carcinoma is much debated and will be further discussed later in this thesis. The DCIS and LCIS are heterogeneous entities. This is reflected in the grading system used for DCIS; low grade DCIS have cells with only subtle atypia and distinct architectural features in contrast to high grade DCIS having highly atypic cells without orientation often with a solid growth pattern and necroses[17].

Prognostic and predictive markers in breast carcinomas

A vast number of predictive and/or prognostic factors have been proposed for BC. Some factors are strictly prognostic (i.e. predicting the risk of recurrence and/or death from disease), predictive (predicts the likelihood of response to a given therapy) and others are both prognostic and predictive. The most established markers are histological grade, stage (size, lymph node involvement and metastases), steroid receptors, HER2, age at diagnosis and vascular invasion[17, 19].

Histological grade

Various systems for grading aggressiveness based on histopathological assessment of differentiation pattern (luminal/glandular) and nuclear features have been developed. Bloom et al. presented one system in 1950[20, 21], this has been the fundament for the grading system used today; "the Nottingham modification of the Bloom and Richardson method" which was introduced in 1991[22]. The degree of luminal differentiation, nuclear atypia/pleomorphism and mitotic index is combined in a single numerical score called histological grade. Each factor is assessed separately by examination of histological sections, given a numeric value (1-3) which is added into a score from 3-9. Tumors of grade 1 (score 3-5) have cells with tubular differentiation, few mitoses and lack of pleomorphia, this in contrast to grade 3 tumors (score 8-9) which are poorly

differentiated, have high mitotic index and are often highly pleoemorphic. Although histological grade is an independent prognostic index[22], the major difference in outcome is seen by comparing Grade 1 to Grade 3 tumors. This was the focus of the study by Sotiriou et al. defining genes able to subdivide grade 2 tumors into two groups with better and worse outcome[23]. That a binary grading of DCIS based on molecular observation improve the clinical evaluation is supported by others[24].

Staging of the disease

Both the size of the tumor and nodal involvement (i.e. metastases in regional lymphnodes) has independent prognostic value[25]. These two factors are positively correlated, but tumors size is found to be more important in lymph node positive patients than in negative[26]. Both tumor size and lymphnode involvement are, in addition to metastases, used for staging a womans disease. Staging of breast cancer follow the guidelines from The European and the American cancer unions (UICC (Union Contre le Cancer) and AJCC (American Joint Committee on Cancer))[27] and is based on the TNM classification[28]. The combined information of tumor size, nodal involvement and distant metastases will define the disease stage of each individual from, Stage I-Stage IV, each with different prognostic profiles (Fig. 5).

A widely used system integrating size, lymph node metastases and grade is the Nottingham Prognostic Index (NPI), a numerical categorization stratifying patients into three prognostic groups[29]. The NPI is not in clinical use in Norway today.

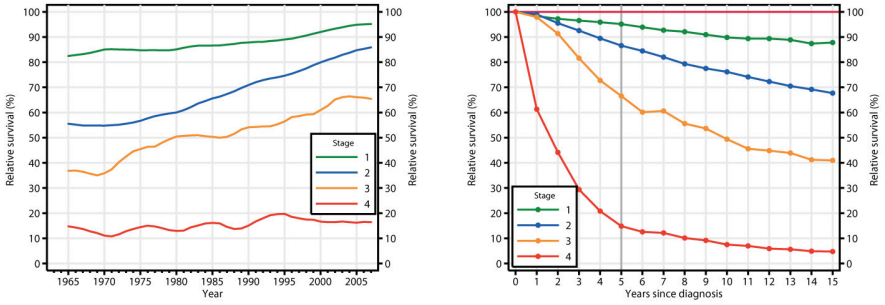


Figure 5: Breast cancer by stage. Left: trends in 5-year relative survival by stage show an increase in survival for patients with stage I, II and also II during the last two decades. Right: The long term relative survival by stage show a huge variation from stage I to IV. From Småstuen et al.[3].

Other prognostic or predictive parameters

The steroid receptors, ER and PgR, have predictive and a medium to weak prognostic value[30-32]. Stimulation of ER increase mitogen activity and induce expression of PgR[33]. The most widely used technique to measure ER and PgR protein expression is by immunohistochemistry (IHC). The advantage is that visual evaluation confirms that normal glandular epithelium in the biopsy do not cause false positive results, and the number and intensity of stained cells can be quantified in a semi-quantitative way. The major disadvantages are the use of different antibodies, detection systems and protocols causing inter-laboratory differences, so participation in quality assessment programs are of major importance. HER2/erbB2/neu is a protein with thyrosine kinase activity involved in regulation of cellular growth and is regarded as a prognostic and predictive factor in breast cancer (for review; [34]).

Gene expression signatures

In the last decade several gene expression studies have defined groups of genes that subdivide breast carcinomas into different prognostic groups, regardless of histopathological classification, and several are commercialized (for review;[35, 36]). Among the first microarray based studies were the identification of 'the intrinsic subtypes'[37], 'the 70-gene metastasis predictive signature'[38, 39] and the 'wound healing signature'[40]. Others have used PCR based techniques to identify responders and non responders to Tamoxifen [41]. Two of the gene lists are forming the basis for large prospective studies (MINDACT and TAILORx). Such studies are useful to identify responders and non-responders to existing therapeutic regimen, but few have per se an approach aiming at classification of breast carcinomas.

Diagnosis and Treatment

In Norway, NBCG defines and updates guidelines for diagnosis and treatment of breast cancer (Norsk Bryst Cancer Gruppe, NBCG; http://www.nbcg.no/nbcg.blaaboka.html). Tumors recognized as cancer will undergo histopathological examination including classification into histological type, histological grade and estimation of the size of the tumor. Lymphnodes will be carefully investigated to detect micro- or macro metastases. Only ER, PgR and HER2 status are molecular markers with predictive or prognostic value included in the national guidelines today.

Breast cancer is today with a multi-disciplinary approach (NBCG guidelines). The cornerstone of all curative breast cancer management is surgical removal of the primary tumor with either breast conserving surgical technique or surgical removal of the whole breast and removal of lymph nodes, either by sentinel node biopsy or axillary lymph node dissection. Locally advanced-primarily inoperable tumors will often be offered neo-adjuvant chemotherapy. Post operative radiation to the breast is offered all women with breast conserving surgery and no lymph node involvement and to women where histopathology showed positive or marginal distance to resection margin. Post operative radiation involving regional lymph node areas is offered individuals with positive lymph nodes depending on age and number of positive lymph nodes. Adjuvant systemic treatment is based on the use of both prognostic and predictive markers to all women with node positive disease and women with node negative disease depending on age, size, grade and HER2 and ER/PgR status. Women with hormone receptor positive disease will be offered 5 years of adjuvant endocrine treatment. The basis of adjuvant chemotherapy regimen is anthracyclins, and in Norway the standard regimen now is the FEC (Fluorouracil, epirubicin, cyclofosfamide) regimen. The benefit of taxanes has been studied the later years and the best effect is observed in lymph node positive disease and estrogen receptor negative disease. It is today standard treatment combined with FEC in these patients groups aged below 70. HER2 positivity is usually associated with more aggressive clinical behavior. The monoclonal antibody Trastuzumab blocks the activity in the receptors tyrosine kinase and is now a part of the standard adjuvant treatment in individuals with HER2 positive tumors. For women with distant metastasis at the time of

diagnosis or distant disease relapse after primary treatment, the treatment will be palliative. Endocrine therapy, chemotherapy, Trastuzumab and local radiotherapy are all possible options to consider.

Why Classifiy? Review and discussion

Grouping of tumors into classes or entities is of importance for several reasons. In clinical management, categorization of tumors is a tool to decide or standardize treatment and patients care. In a classification distinct entities should be recognizable in an objective way. The traditional way of constructing taxonomy in biology is by using a tree based approach where major classes can have smaller subgroups. A robust and objective classification is of importance when performing large clinical studies where clinical behavior and response to therapy are evaluated in order to standardize or tailor therapy. In haematopoietic and lymphoid neoplasia the classification has shifted from being descriptive to an integrative approach also including molecular alterations with features from the hierarchical relationship between mature haematopoietic cells, their progenity cells and stem cells. The knowledge about different lineages and molecular mechanisms determining the direction of differentiation have been the backbone for the modern classification of leukemias and lymphomas[9, 42]. As the hierarchical relationship between the epithelial cell-types of the breast have become more recognized, it is tempting to speculate that the same approach can be used to modernize breast cancer classification. In a Darwinian way of thinking, tree based taxonomy is not a static hierarchy. Offspring will show alterations in a progressive way leading to diversity. The time course of such progression has for mammals been millions of years, but a tumor with rapid growth will produce several levels of offspring during months or even weeks. If the daughter cells have acquired new characteristics compared to the parent cell, this can be defined as progression. Breast tumors in humans are recognized clinically at different stages of progression. One challenge in building a classification based on molecular alterations is that little is know about which lineages exist and at which stage or along which linage breast tumors develop. Whether tumors follow one path of progression or several, or which alterations characterize the different levels of progression still remains to be defined. To be able to relate findings of molecular subtypes to this, a review over tumor initiation and progression will be given.

Initiation and progression by successive genetic changes

Cancer being caused by alterations in hereditary material was suspect before the discovery of DNA[43], and genomic instability was shown decades ago to be a hallmark of cancer[44]. At that time it was acknowledge that transformation of cells into neoplasia required only a limited number of genomic changes[45]. This was also the main focus of the review by Hanahan and Weinberg[46] defining different characteristics being essential for cancer development. Reflecting the enormous increase in knowledge in this field just in the last decade, a recent publication defines even more 'hallmarks of cancers'[47]. The underlying defects of these hallmarks can prove to be important targets for treatment, but represent a complexity not captured by the standard classifications of today. As reviewed by Stratton et al., cancer can be considered an evolutionary process analogous to Darwinian evolution[48]. Two main processes are required; continuous acquisition of heritable genetic variation in individual cells and natural selection of cells with higher capability to proliferate and survive. If a single cell get sufficient advantageous alterations and reside in an environment providing 'matching' conditions, the result can be a tumor progressing into cancer. This is reflecting the heterotypic view on tumor formation and progression in contrast to the reductionist view[46]. The first focus on the fact that tumors are composed of other cell types such as endothelial cells, fibroblast, lymphatic cells etc. as well, but in the reductionist view the alterations in the tumor cells are the only ones considered. Normal development of the breast are dependent on stimuli from the environment and that tumor cells collaborate with or dictate other cells to provide an advantageous micro-environment is continuously more recognized[49-51].

Studies of rodent breast tumor development and progression as reviewed by Foulds in 1954 revealed some interesting features[52]. Spontaneous mammary tumors in rabbits begin either as adenomas in otherwise normal breast or in breast with cystic disease. The progression follows successive stages through non-invasive to invasive tumor and eventually to metastatic disease. Foulds concluded that cancer is the final step in a developmental process where the early neoplasia is not an invasive disease (i.e. cancer) either in structure or behavior. In studies of mice strain developing multiple tumors at the same time, the effect of host related factors on tumor progression could be studied. The breast tumors seemed to be of two types; 'unresponsive' tumors where

growth did not depend on hormonal related factors and 'responsive' tumors where the tumor growth was related to the hormonal state of the host. The studies showed that progression of one tumor was independent of other tumors and probably reflected a regulation by 'intrinsic' properties. Fould made six statements concerning tumor progression:

1. Tumors progress independently
2. Characters such as growth rate, responsiveness, invasiveness and the ability to disseminate are independent of each other.
3. Progression is independent of growth rate
4. Progression is continuous or discontinuous by gradual change or by abrupt steps
5. Progression follows one of alternative paths of development, but can change course into a different path
6. Progression does not always reach an end-point within the life-span of the host

These statements were based on observations from rodent experiments performed in the same decade the structure and composition of DNA were revealed, and therefore without any of the knowledge we have today about genomic related alterations in tumors. Much of the knowledge we have about molecular subclasses in breast carcinomas are based on clinical samples, and knowing that such samples are analyzed at individual progression levels, Foulds hypotheses can serve as a backbone for discussing the molecular types of breast carcinomas.

Genomic data indicate at least two types of breast cancer

Several studies analyzing genomewide DNA alterations have tried to identify groups of tumor with distinct features. Four different patterns of alterations were identified by Hicks et al. with high resolution aCGH analyses of two breast tumor cohorts[53]. The 'Simplex' pattern had broad segments of duplications and deletions. Deletion of 16q, 8p and 22 as well as gain of 1q, 8q and 16p was dominating. 'Complex I' had either a "sawtooth" appearance with narrow segments of deletions and duplications affecting more or less all chromosomes. 'Complex II' resembled the 'simplex' but had at least one localized region of clustered peaks of amplifications called 'firestorm'. The fourth pattern was called "flat" defining profiles with no clear gains or losses except from copy number

polymorphism. Interestingly, all four patterns were found both in diploid and aneuploid tumors. The same groups have been identified in other datasets[54]. A study by Chin et al. using aCGH identified three subtypes of breast carcinomas that varied with respect to level of genomic instability[55]. The groups had overlapping characteristics with the classes in Hicks' work. One group of tumors had few alterations and was dominated by 1q amplification and 16q deletion (the 1q/16q group), another group had more complex alterations (complex group), and the third displayed frequently high level amplifications (mixed amplifier group). Tumors with *BRCA1* mutation had similar changes as the complex group. In this cohort it was also observed that shorter telomeres were associated with greater number of amplifications[56, 57]. Several studies have had quite divergent definitions on which genomic alterations characterize distinct subgroups of breast carcinomas, but that 1q and 16q alterations dominate in one type and multiple alterations on several arms dominate another are found by most[58-64].

Genomic changes in early stages of breast carcinogenesis

The *in situ* breast carcinoma, DCIS, considered as a true precursor to invasive ductal carcinoma, is a heterogeneous group probably reflecting multiple types of breast tumors[65-67]. The loss of 16q is frequently found in DCIS, but also in proliferative and premalignant lesions such as usual ductal hyperplasia, columnar cell lesions, atypical ductal hyperplasia and in a substantial proportion of invasive carcinomas (ILC and also IDC), often in combination with 1q gain[68-76]. Low grade DCIS frequently display loss of 16q and gain of 1q, while high grade DCIS have more alterations including high level amplifications of 6q22, 8q22, 11q13, 17q12 and 17q22-24 [54, 65, 77, 78]. The few CGH data that exists from LCIS are showing overall less gains than invasive carcinoma, and that the alterations partly overlap with grade I invasive carcinomas [66, 79, 80]. In invasive tumors, deletion of 16q is more frequent a physical loss of the whole arm in grade I tumors, while alterations of 16q in grade II and grade III are more complex[78, 81-83]. Grade I tumors have fewer genomic alterations compared to grade III carcinomas that often have numerous genomic changes with chromosome arms 8q, 17q and/or 20q frequently altered[84].

Molecular studies of near-diploid invasive tumors probably give insight into early genomic changes in tumor progression. The most frequent rearrangements seen in such

cases by karyotyping are unbalanced translocations where a majority resulted in loss of one of the derivative chromosomes[85, 86]. Dutrillaux et al. reported that near diploid cases with less than four rearrangements almost always involved alterations of 1q and/or 16q while losses of chromosome segments were more prominent than gains in cases with more than four rearrangements[85]. This is in line with the findings from aCGH analyses of diploid tumors; some tumors were of the simplex type, other of the complex 1 or complex 2 type[87]. A translocation resulting in a der(1;16)(10p;10p) is identified by karyotypic studies and considered an early event in mammary carcinogenesis[88, 89]. Another early event seems to be formation of isochromosome 1q, this gain is also seen in numerous studies using array comparative hybridization (aCGH), making 1q gain one of the most frequent alterations in breast carcinomas.

Subgrouping breast cancer by ploidy measurements

The prognostic value of measurements of DNA content in breast carcinomas have been debated for decades but it seem evident that breast tumors can be grouped by different levels of DNA content[90, 91]. Breast carcinomas display a wide range of modal values from less than 30 to more than 200 chromosomes per cell[64]. Kronenwett at al. subdivided a tumor set into diploid (modal value 1.8c-2.2c), tetraploid (3.8-4.2c) or aneuploid groups (one peak or more outside the diploid or tetraploid range)[92]. By adding a stemline scatter index (SSI), each of the three groups was subdivided into being stable or unstable. Their study showed that is was of minor importance where the stemline was situated, but the scatter indicating an unstable genome reflected a significantly worse prognosis. Aneuploid tumors had frequently a hypotetraploid modal value, but a minor group of aneuploid tumors were hypodiploid, hyperdiploid, triploid or hypertetraploid. Structural chromosomal aberrations and losses of entire chromosomes have been suggested to occur first during genetic evolution of breast tumors, and would lead to a transient hypodiploid cell clone[85]. A succeeding doubling of DNA by endoreduplication would result in a DNA content ranging from triploid to hypotetraploid tumor depending on the amount of initial losses. Alternatively the endoreduplication can occur early and additional rearrangements will result in a hypo or hypertetraploid tumor. Hypodiploid tumors have been considered

a distinct entity with both clinical and genomic characteristics dominated by losses on multiple chromosomes and is associated with a worse outcome[93, 94].

Subclasses defined by gene expression patterns

The intrinsic subtypes

The gene expression based classification defining five subtypes was the result of the works of Perou and Sorlie a decade ago[95, 96] in neoadjuvant treated breast carcinomas. The expression of approximately 12000 genes was measured by cDNA arrays[95]. Thereafter, genes that had low variation in expression in samples taken before and after treatment for each patient and at the same time varied most between all patients were extracted. A total of more than 550 genes were thus identified and named the "intrinsic gene list" as they were thought to be reflecting the individual tumors phenotype. By hierarchical clustering, a pattern of two main clusters with a total of five subclusters emerged in several independent cohorts[96-100]. The largest cluster has frequently two groups dominated by ER positive and Luminal cell related genes, one having more proliferation related genes upregulated than the other (Luminal A and Luminal B respectively). The other main cluster had three groups. One related to myoepitel/basal epithelial cell gene expression (such as basal cytokeratins and thus called Basal-like), another were dominated by high expression of *erbB2* related genes (called erbB2+ group) and the third had gene expression not very dissimilar from patterns found in normal breast tissue samples (called Normal-like).

The robustness of the subtypes

By calculated centroids for each of the five main subtypes (Luminal A, Luminal B, erbB2+, Basal-like and Normal-like), class prediction can be made for individual samples. When making class predictions for the cohort analyzed in paper II, III and IV, several of the samples correlated to more than one centroid[100]. A heat map generated by a cluster algorithm illustrates the heterogeneity of the centroid correlation in the sample set (Fig. 6).

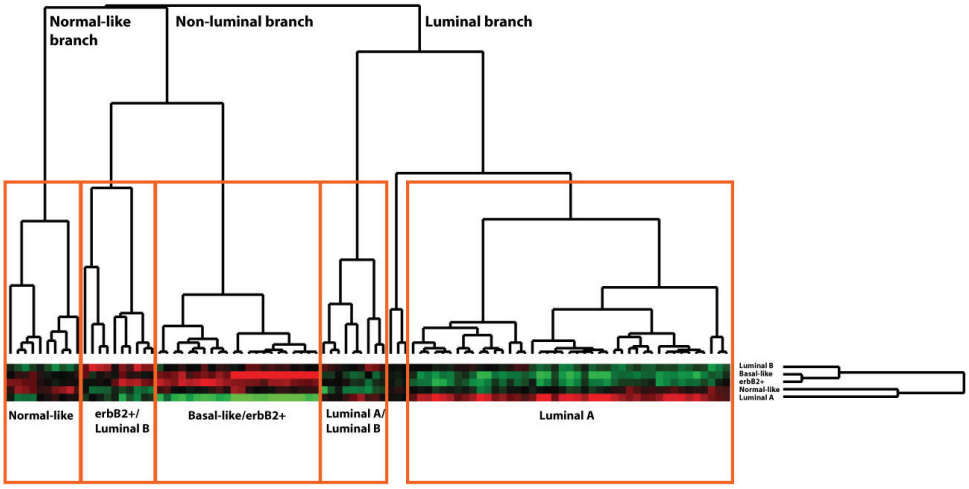


Figure 6: Hierarcical clustering of 123 MicMa samples based on the calculated correlation to the centroid for each of the five subgroups. Red indicates positive correlation, green indicates negative (anti-) correlation. Dark color indicates correlation close to zero. The rows of the heat map indicate the centroid correlation values to Luminal B (1. row), followed by the Basal-like, the erbB2+,the Normal-like and Luminal A at the bottom. The clusters reflect the relationship between the different subtypes. .

By using this approach two conclusions can be drawn:

1: There are two main branches, one dominated by samples correlated to the Luminal A centroid, the other correlated to the ERBB2+ and/or Basal-like centroid. Samples do not have a strong correlation to both the Luminal A centroid and the Basal-like and/or erbB2+ centroid. The Basal-like samples have almost always a positive correlation to ERBB2+.

2: Samples highly correlated to the Luminal B centroid are found in both main branches, some have additional correlation to the Luminal A centroid, others to the Basal-like or erbB2+ centroid. Samples highly correlated to the Normal-like centroid are also in both main branches, some have additional correlation to the luminal A centroid, others to the Basal-like or erbB2+ centroid.

An interesting notion is that samples with a high correlation to Normal-like are always anti correlated to Luminal B.

From this we can hypothesize that Luminal A and Basal-like are phenotypically diverse with regard to intrinsic characteristics.

Surrogate markers for the subtypes

Immunohistochemical (IHC) staining of tumor sections has revealed that the Luminal A tumors are often ER and/or PgR positive while the Basal-like are not. The former have several proteins in common with the luminal cell type of the breast (such as ER, PgR, CK18, GATA3) while the latter resemble to some extent the myoepithelial cell type, such as CK5, 6, 14, 17 and SMA[97, 101-103] (for review: [104]). Basal-like tumors are often said yto be 'triple negative' (i.e. negative IHC for ER and PgR and negative IHC/FISH for HER2), but is known to be heterogeneous[105]. Another major difference between Luminal A and Basal-like tumors are the frequent finding of single base mutations in genes such as TP53 and BRCA1 in Basal-like tumors. Those genes are only rarely mutated in Luminal A tumors. Histological patterns of differentiation are linked to the subtypes. Carcinomas with lobular and tubular differentiation are almost always of Luminal A type while tumors with medullary, adenoid cystic or metaplastic differentiation are of Basal-like type[106, 107].

Accepting that the phenotype of the tumor is influenced by extra-tumoral factors such as tumor microenvironment (stroma, inflammation, endothelium, fat) and endogenous and exogenous components such as hormones and other substances, the search for genomic alterations for each of the subtypes was important. Several groups have found genomic alterations by aCGH that seem to be more frequent in one or more of the intrinsic classes[56, 57, 108]. Bergamaschi showed, in an advance stage cohort, that the intrinsic subclasses harbored different genomic alterations[108]. The Basal-like had higher numbers of gains and losses than Luminal A and the Luminal B and erbB2+ had more frequent high-level amplifications. Chin and Fridlyand compared their aCGH groups to the expression subtypes, and found that Luminal A tumors were dominating the 1q/16q group, Luminal A and erbB2+ the mixed amplifier group and Basal-like and Luminal B

the complex group[56, 57]. Another study identified a group of tumors with low genomic instability, and found these tumors to be enriched by the Basal-like subtype[109]. Normal-like samples are often too few to be studied, and Luminal B can be hard to identify in some datasets[99]. The erbB2+ group was dissolved when the *erbB2* amplicon was removed from the data in one CGH based study[110], but are more distinct as a subgroup in others[111].

Expression subtypes and epidemiology

It seems evident that of the molecular expression subclasses, the Luminal A and the Basal-like group are regarded as distinct diseases with different genomic changes, expression patterns and clinical and histopathological profiles. By using IHC markers several epidemiological studies have been perform to identify differences in etiological factors[101, 112-114]. The distribution varies among different ethnical populations with Basal-like tumors more frequent in African-American than in non African-American women[101]. It is also shown that increasing body mass index reduces the risk of Luminal tumors in premenopausal women, and that late menarche reduces the risk of Basal-like carcinomas[113]. Acknowledged risk factors for breast cancer in general seem to only be valid for Luminal A tumors; women with fewer children and high age at first full term pregnancy had a higher risk of Luminal A carcinomas than Basal-like[114]. The increased risk of Basal-like carcinomas observed in women with young age at first full time pregnancy and in women with high parity and short duration of breast-feeding indicate the complementary nature of these two diseases[114]. Basal-like tumors are also known to have an earlier age distribution compared to the Luminal type[112].

Breast cancer progression from a molecular point of view

Several observations of Foulds can now be viewed with the knowledge of molecular alterations as seen by multiple different methods investigating different characteristics of breast tumors.

Tumors progress independently

The notion of this came from studies in mice, by having five to six pairs of breast glands the probability of having several tumors at the same time is much larger than in humans. An interesting aspect is that tumors in the same host can have different paths of progression. In a study we performed on bilateral human tumors we saw that the distribution of molecular subtypes followed some patterns (paper I). Women with a luminal tumor in one breast had almost always a luminal tumor in the other breast. Luminal tumors were defined as having either ER or PgR expression, and represent the tumor type dependent on the host for instance by hormonal influence ('responsive tumors'). Interestingly, the triple negative tumors in this study had a more heterogeneous distribution and are probably of a more 'unresponsive' type.

Progression follows alternative paths in luminal and basal related carcinomas

The findings reviewed above about molecular types of breast carcinomas indicate that separate breast cancer tumor types exist and Luminal-A and Basal-like are the most acknowledged.

One type of carcinomas evolves from hyperplasia through low grade pre-invasive tumors into invasive carcinoma (IDC/ILC) predominantly of low grade. It also seems evident that several tumors do not follow this path but have genome wide rearrangements already at the pre-invasive stage. They probably evolve from high grade DCIS into high grade invasive carcinomas[115]. The high grade tumors are frequently ER negative in contrast to the low grade tumors dominated by loss of 16q and gain of 1q[78, 116, 117]. In paper II we studied the genomic alterations in 595 tumors aiming at combining the knowledge supporting the existence of two main classes of tumors; 1) Luminal A/simplex type and 2) the Basal-like/erbB2+/complex type. As seen by others, the alterations 1q gain and/or 16q loss recognized a majority of Luminal A tumors (called *A* tumors) and tumors with genome wide alterations were dominated by Basal-like tumors (called *B* tumors).

The frequent concordance of 1q gain and 16q losses is shown by karyotyping to represent centromere close translocations. As shown in Figure 7 multigene interphase

FISH identified this translocations in several of the *A* tumors included in paper II (unpublished data).

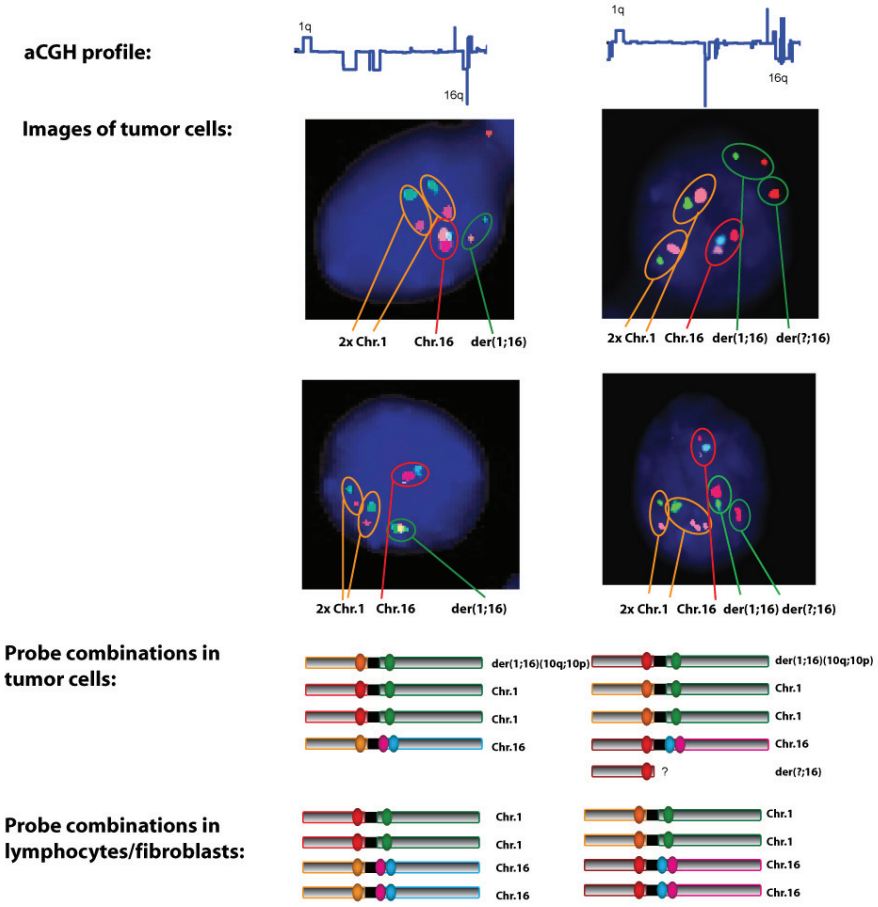


Figure 7:
Example of aCGH and FISH analyses from two Luminal A tumors. At the top is the aCGH profile with 1q gain and 16q loss in addition to some other alterations. The pictures show two cells from each tumor hybridized with five different FISH probes. The illustrations below illustrate the observed combinations of the FISH probes compared to the expected combination as it is seen in normal cells indicating a der(1;16)(10q;10p) in the tumors.

The abundance of heterochromatin and segmental duplications close to the centromere on chromosome 1 might make this a vulnerable area for mitotic over-crossing and

subsequent translocation[118]. Interestingly, chromosome 16 has duplication rich centromeric regions with homologous sequences to several chromosomes including chr. 1, this might also make chromosome 16 vulnerable for such changes [119].

The data analyzed in paper II suggest that a progression occur in *A* tumors when the tumor genome are able to undergo complex rearrangements. As illustrated in paper II the tumors with complex rearrangements (*A2* tumors) have overall more alterations than those without (*A1* tumors) and the clinico-pathological data are in favor of *A2* tumors representing more advanced progression levels of *A* tumors. This is in line with Foulds hypothesis; tumors can progress by a shift of path. Complex alterations of the firestorm type in aCGH profiles are showing high-level gains of regions with intermittent losses. Both karyotyping and advanced sequencing of such tumors has revealed that several different chromosomes can be involved in complex combinations[120, 121]. In contrast to karyotyping and sequencing, aCGH can only give indications of which arms are involved in such complex rearrangements. One mechanism explaining this type of rearrangements is the breakage-fusing-bonding principle (BFB cycles), where double strand DNA breaks in cells with repair defects can lead to either sister chromatin or non homologous end joining followed by a new break during the next mitosis creating amplifications and deletions[122, 123]. The most frequent arms with complex rearrangements in *A* tumors were 8p and 11q. Bautista et al. showed by FISH that alterations on these two chromosome arms can be rearranged together in a derivative chromosome, probably due to BFB cycles[124], although other groups have shown that these events can occur unconnected as well[125]. In MCF7, a well characterized ER positive cell line with complex rearrangements on several chromosomes including 17q and 20q, the same phenomenon is seen, resulting in functional fusion genes from the two chromosomes[121, 126, 127]. The results from paired-end sequencing from one of the *A2* tumors reveal the same complex pattern of several chromosome arms being intermingled and causing fusion genes (Stephens at al. under review, Nature). Recurrent fusion genes rare in breast cancer[128], but can be explained in wide range of breakpoints from tumor to tumor. High-level amplifications of selected regions like 8p11, 11q13, 12p13, 17q12 and/or 20q13 are strong predictors of reduced survival [110, 129].

Intra-tumor heterogeneity has been acknowledged in breast carcinomas[130]. One study by Navin et al (in press, Genome Research) different parts of tumors were sorted into cell fractions with regard to ploidy. This study showed two main types of progression; one monogenomic, stable type and one polygenomic more genomic unstable type. The latter type had one clone dominated by hypodiploid cells, but also additional clones with aneuploid DNA index (triploid area) indicating that a doubling of DNA content from a hypodiploid phase has occurred. This is in line with the findings of Dutrillaux at al.[85]. In paper IV the ploidy measurements of Basal-like tumors by ASCAT correspond to the distribution seen in the polygenomic group and the measurements for Luminal A the distribution of the monogenomic type. Coinciding with the aneuploidization of the polygenomic tumors, complex rearrangements occur, in line with our findings of *B1* tumors being dominated by large regions of losses while the related group, *B2* tumors, had more gains in addition to complex rearrangements (paper II). This switch can explain the close relationship between erbB2+/Luminal B and Basal-like tumors; complex rearrangements have frequently amplifications of growth promoting genes found, and this can shift the phenotypic pattern more towards the expression subtypes such as Luminal B and erbB2+. As also seen by Chin et al.; if genes whose expression was correlating with amplification were removed, the erbB2+ cases did not cluster together. This can indicate erbB2+ tumors do not represent a separate path of progression but reflects a 'side-path' for the main types[110]. Data from paired-end sequencing revealed a very dissimilar rearrangement pattern compared to Luminal A tumors. Basal-like tumors had multiple segmental duplications genome wide (paper II). The mechanism behind is not known, but in the MicMa cohort we identified two tumors of the *AB2* and *C2* type with this pattern in addition to more complex rearrangements. One was Basal-like by expression, the other were erbB2+, again strengthening the suspicion of a close relationship between these groups. In addition, this latter case was by SNP analyses (paper IV) found to have allelic imbalance of the same type as seen for the Basal-like tumors.

Progression does not always reach an end-point

This reflect a phenomenon widely known to be true for some types of prostate carcinomas, and when mammography was introduced, it was debated whether the increased incidence in the same time reflected tumors that never would have progressed to become a clinical detectable tumor during a woman lifetime. Breast tumors are estimated to have very different growth rate[131]. Highly differentiated tumors such as tubular carcinomas with only one or two genomic changes (such as 16q loss) might represent such tumors[83, 131]. As mentioned above, no data up to now have been able to identify which tumors have the propensity to have secondary changes and develop via another path into more aggressive disease.

Epigenetic alterations in breast cancer

Epigenetic modifications both at the chromatin and DNA level affect the structure and the expression of genes and is essential both for normal development but also for regulation of tissue specific processes. Several mechanisms are of importance, such as histone modification, DNA methylation, non-coding RNA's and nucleosome position (for review; [132, 133]. Probably the most widely studied epigenetic modification is the cytosine methylation in the context of the dinucleotide CpG. In embryonic stem cells such modifications is of major importance in regulating genes important for cell differentiation and function[134]. Altered regulation of CpG methylation is implicated in many diseases. Specifically, in cancer, methylation of CpG islands proximal to tumor suppressor genes such as *p16*, *Rassf1a*, and *BRCA1*, is a frequent event, and methylation of several gene are found to be linked to breast cancer[135-138]. Knowledge about different methylation states characterizing cells at different levels in the breast cell hierarchy is emerging[139], and in paper III we found a correlation between subgroups of tumors and methylation patterns more common in the luminal lineage compared to myoepithelial lineage strengthening the relationship between the Luminal A and Basal-like carcinomas with the different levels in the hierarchy of normal breast epithelium.

Tumor stem cell models

A key event in carcinogenesis are the acquisition of self renewal capacity[46]. Self renewal capacity is a hallmark of stem cells, and the discovery of subpopulations of cells with phenotypic resemblance with stem cells opened for a debate concerning the existence of 'cancer stem cells'[140]. The cancer stem cell theories can be viewed as two different models; the cancer stem cell model and the clonal selection and evolution model (Fig. 8)[15, 141].

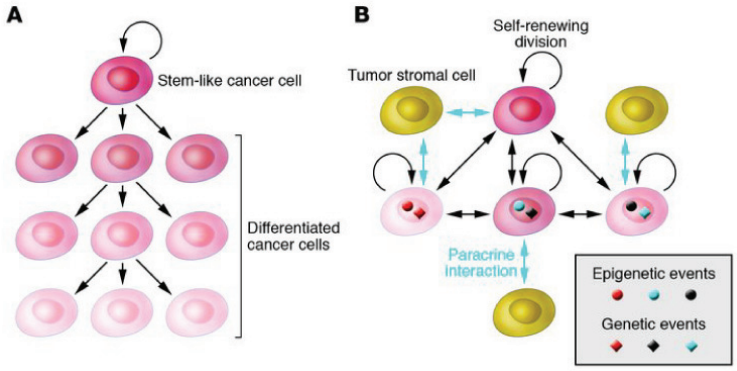


Figure 8: The two main models of cancer stem cells;
A: The cancer stem cell model and B: the clonal selection and evolution model. The dark red cells illustrate tumor cells with stem cell capacity, while the brighter cells represent more differentiated progeny causing tumor heterogeneity. From Polyak 2007[15].

Tumor heterogeneity is explained differently in the two models, with programmed aberrant differentiation in the first or as a mixture of subclones with difference in the latter. The former hypothesis defines the cancer stem cells to be the driver of the progression, while the latter defines the clone with the most advantageous aberrations as the driver. Both models can explain treatment resistance; either is the stem cell resistant or a development of a resistant subclone will explain the progression of the disease. Although the models are different, they are not mutually exclusive. It is possible that some breast tumor types have a cell of origin with stem cell properties and can develop heterogeneous subclones if the ability to differentiate is intact. Others might origin from more mature and linage restricted progenitors and subclones with additional alterations can explain progression and resistance to treatment. In others and our data the Luminal

related tumors fit into clonal selection and evolution model, while Basal-related tumor progression can be explained by a stem cell model[142, 143]. It is of major importance to reveal more about the properties and relationship between mammary epithelial cells and their predecessors[144].

Aims of the study

The primary aim of this study was to explore breast carcinomas at the genomic, transcriptomic and epigenetic level to identify distinct molecular subgroups of tumors, and explore their different progression paths and the clinical impact.

The secondary objectives were:

- to define the relationship between host-related influence and the molecular expression subtypes by classifying bilateral synchronous and metachronous breast tumors using IHC surrogate markers.

- to elucidate the relationship between genomic alterations, molecular expression subtypes, structural rearrangements, ploidy, pathology and clinical data by exploring genomic architectural alterations in high-resolution aCGH data from different breast cancer cohorts

- To explore genome wide methylation patterns to identify subgroups and their relationship to molecular expression subtypes and clinical data.

- To develop bioinformatical tools enabling objective measurements of genomic events.

- To develop bioinformatical tools to elucidate the heterogeneity and ploidy in tumors in order to adjust genomic copy number values.

Material and methods

Patient material

This study has been analyzing several clinical breast cancer cohorts, but the main focus has been on the "MicMa" samples. Four other cohorts were used; one with bilateral tumors ("Russian") and three cohorts with primary tumors (Sweden; "WZ", Oslo; "Ull" and England; "ChinUC"). The details of these cohorts and analyzes performed for each study is given in Table 1. Demographic data for all cohorts are given in Paper II (Supplementary Table 1).

As a part of the "micrometastasis" research group at the Norwegian Radiumhospital, a study concerning the implication of micrometastasis for breast carcinoma patients were launched in 1993 (The DNK study, supported by The Norwegian Cancer Association). A total of 921 breast carcinoma patients from five different hospitals were enrolled between 1995 and 1998 into the study. Blood, bone marrow, tumor tissue and lymph nodes were collected if possible, as well as clinical data including 10 years follow up[145]. Fresh-frozen tumor tissue was available from 130 patients, and this sub cohort of the DNK study is referred to as the MicMa cohort. The cohort consists mainly of primary operable tumors of stage I-III where almost 40% received no adjuvant therapy[100].

The ChinUC cohort is selected from a clinical tissue bank to represent low stage tumors[109]. All tumors were primary operable invasive carcinomas collected from 1990-1996.

The WZ cohort was highly selected as it was drawn from a tissue bank to study diploid tumors with different outcome[87]. In addition to 100 diploid tumors, 41 aneuploid tumors were included.

The Ull cohort was sequentially collected at a single Norwegian hospital between 1990-1994 and was dominated by primary operable breast carcinomas of low to intermediate stage[99].

The Russian cohort was collected retrospectively to include equal numbers of metachronous and synchronous breast carcinomas.

Table 1:

Paper:	Cohort	Cohort Nationality	Inclusion periode	No. of samples in the study	Tumor types	Clinical follow up	main type of analyzes	Data available from other analyzes	Statistics and bioinformatics
Paper I	BiBC	Russia	Retrospective	50 patients (100 tumors)	Bilateral breast carcinomas	No	IHC and Immunofluorescent	BRCA1 mut. Status HER2 IHC Clinical and pathology data	SPSS 15.0
Paper II	MicMa	Norway	1995-1998	125	Primary operable breast carcinoma	10 yrs.	aCGH	Molecular expression subtyping DNA Ploidy measurements HER2 IHC and FISH TP53 mut. status Clinical and pathology data	Novel algorithms in Java, Matlab, R. SPSS 15.0
	Ull	Norway	1990-1994	167	Primary operable breast carcinoma	Up to 20 yrs.	aCGH	Molecular expression subtyping TP53 mut. status Clinical and pathology data	
	WZ	Sweden	Retrospective	141	primary operable breast carcinomas; diploid and aneuploid survivors and non-survivors	20 yrs.	aCGH	DNA Ploidy measurements Clinical and pathology data	
	ChinUC	England	1990-1996 (retrospective)	162	Primary operable breast carcinoma	Up to 20 yrs.	aCGH	Molecular expression subtyping Clinical and pathology data	
Paper III	MicMa	Norway	1995-1998	114	Primary operable breast carcinoma	10 yrs.	MOMA	Molecular expression subtyping Clinical and pathology data	R
	Controls	USA		11	'Normal' breast tissue		MOMA		
Paper IV	MicMa	Norway	1995-1998	114	Primary operable breast carcinoma	10 yrs.	snp array DNA Ploidy measurements FISH analyzes	Molecular expression subtyping Clinical and pathology data	Novel algorithms in Matlab, R.

Methods

As noted in Table 1, several different methods were used in this study. An overview of all methods is given in Table 2 at the end of this section. By combining data from different types of analyzes, we have been able to characterize breast tumor both at the phenotypic (Protein, RNA), epigenetic (methylation) and genomic (DNA) level.

Immunohistochemistry

Immunohistochemistry for protein detection was chosen as it is convenient on FFPE tissue and because it allows visually interpretation of which cell type express the chosen protein. The method is based on antibodies binding to the chosen antigen (protein) and thereafter visualized by different detection systems. The main detection system used in this study was Envision+ (DAKO) which has less background and is easier to interpret than the previously more common techniques such as the ABC (Avidin-Biotin-Peroxidase) method [146, 147]. The bound antibodies are recognized by a secondary antibody coupled to a dextran polymer with enzymes, and after biotin treatment it gives a strongly enhanced visual signal as illustrated in Figure 9. Antibody based assays with detection by Fluorochromes can also be used as demonstrated in Paper I.

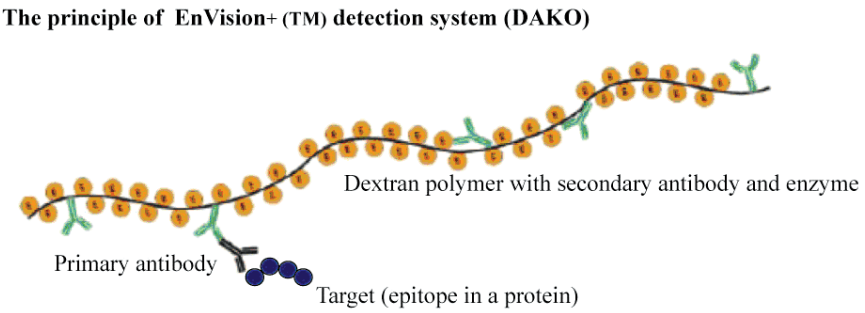


Figure 9: The principle behind protein detection by polymer based IHC. Modified from Wiedorn et al. 2001 [148].

Gene expression microarray analysis

Measurements of RNA levels of different genes were not a specific part of this study, but the classification by expression data are fundamental for all the four papers, so the methods will be reviewed briefly. The RNA levels of expressed genes can be made individually by quantitative RT-PCR (reverse transcriptase polymerase chain reaction), but microarray technology opened for expression analyses of thousands genes at the same time. The array type used for expression based classification in paper II and III were cDNA arrays consisting of 42 000 cDNA clones selected from expressed a sequence tags (EST) library and spotted on glass slides[100, 149]. Both sample and reference RNA were converted into cDNA, labeled by different fluorescent dyes mixed and hybridized to the array. An optical reader measured fluorescent at both wavelengths to be able to calculate an intensity ratio (Fig. 10). The ratio reflects genes that are over-, under- or equally expressed compared to the reference cDNA.

The molecular classification of breast carcinomas is based on previous studies of Sorlie and Perou[37, 95, 96]. Although some samples have almost equally high correlation to more than one, the most common way of classifying is to designate each sample to the centroid with highest correlation as illustrated in Figure 11.

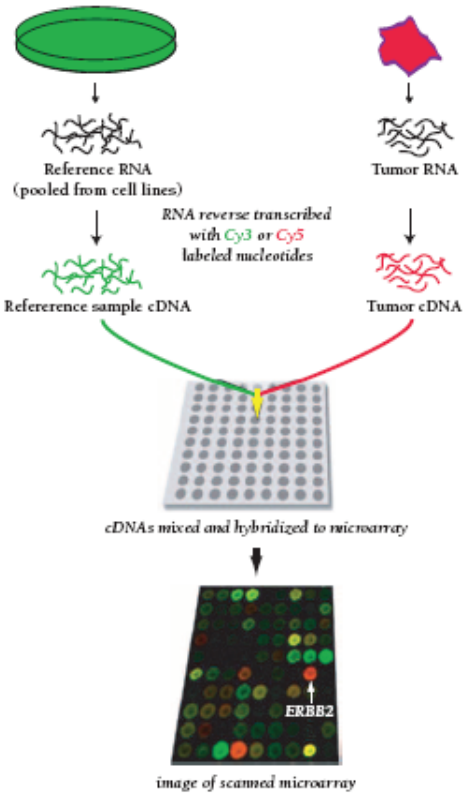


Figure 10: A schematic illustration of cDNA expression array hybridization. From Jeffrey et al.[149].

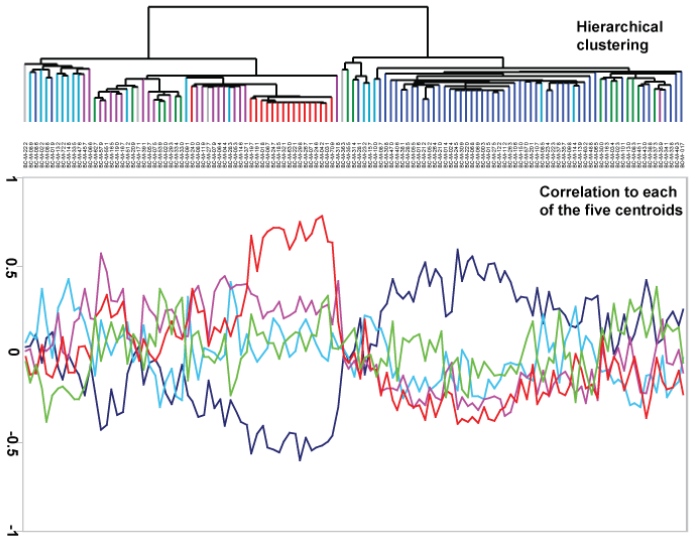


Figure 11: The gene expression data from the intrinsic gene list is used for hierarchical clustering of the MicMa samples (top). The corresponding chart (bottom) illustrates the correlation value to all five centroids for each sample. The color of the bars in the cluster diagram show the centroid the given sample had the highest correlation to. From Naume et al.[100].

Measurement of DNA content

Measurement of DNA content of the MicMa samples was performed on imprints, made by lightly pressing frozen tumor tissue onto glass slides followed by fixation in formalin. The staining of DNA was performed by Feulgen reaction; hydrolysis of DNA followed by a color reaction (Schiff) as previous described[150]. The cells were identified visually as tumor cells or non-tumor cells (such as lymphocytes and fibroblasts). By image cytometry the DNA content was measured in approximately 200 tumor cells and in representative non-tumor cells. The optical density of each cell was compared to the density of the non-tumor cells and the result from each tumor was viewed in a histogram. The histograms in this study was interpreted visually where the mode value of each peak were selected as the ploidy value of the tumor. Tumors with mode values between 1.8 and 2.2 were called diploid while tumors with mode values higher that 2.2 was called

aneuploid. Some tumors were purely diploid while others were aneuploid often displaying a broad specter of DNA content (Fig. 12).

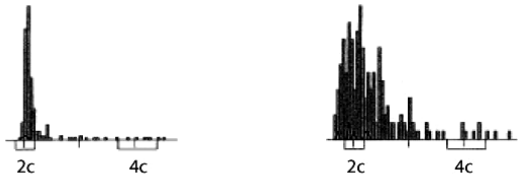


Figure 12: To the left a histogram from a diploid tumor (top) where almost all measured cells have DNA content equally to non-tumor cells (2c). To the right is an aneuploid tumor displaying a broad specter of cells with DNA content ranging from below 2c to more than 3c.

Fluorescence In Situ Hybridization; FISH

To visualize alterations of the DNA structure in more detail, FISH is a technique that can both show copy number alterations and structural rearrangements. The HER2 copynumber was measured by FISH in the MicMa cohort on TMA (tissue micro arrays) by commercial probes hybridizing to the gene (Vysis)[151]. A fluorescent labeled DNA probe is designed to be complementary to the target DNA and after hybridization the signal will be detected by a fluorescent microscope. DAPI (4',6-diamino-2-phenylindole) are frequently used to visualize the nuclei. Probes can be made in-house both by using BAC (bacterial artificial chromosomes) clones and by PCR based techniques. Absence of signals can be interpreted as genomic loss, while extra signals indicate gains (Fig. 13). By selecting probes close to each other, translocations can be detected either as split signals or fused signals. In paper II we designed BAC probes with different fluorescence to DNA loci on each side of the centromeres on chromosome 1 and 16 to visualize a translocation between the two chromosomes. In paper IV we used probes tailored to frequently amplified regions on chr. 8 for validating the copy number estimates made by ASCAT (Fig. 14)

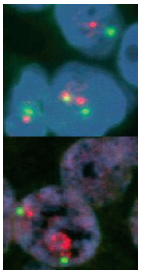


Figure 13: FISH analysis revealing copy numbers of the HER2 gene (red) compared to centromere 17 (green) in tumor cells (the nuclei are blue by DAPI). On top is nuclei from a tumor with no increase in HER2 copy number, below is a tumor with two copies of the centromere 17 and >20 copies of HER2.

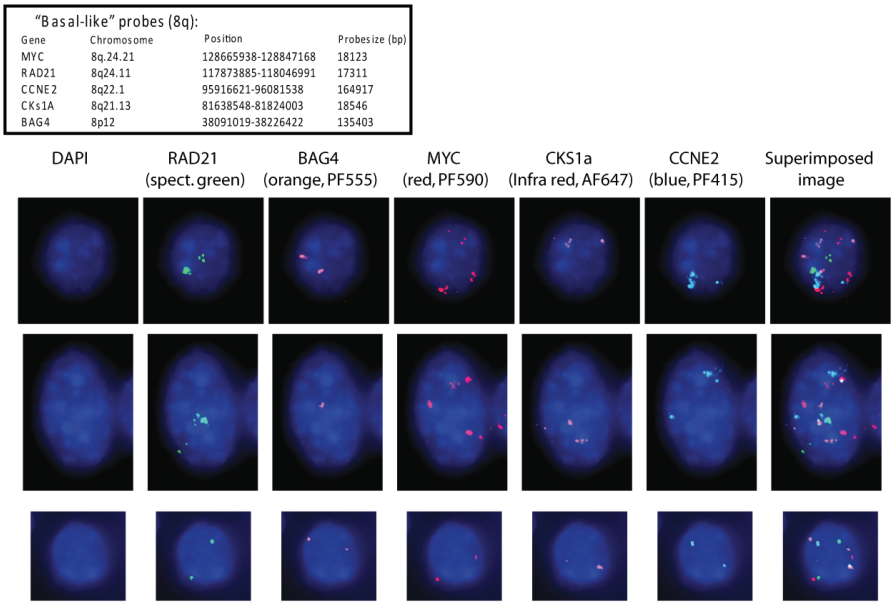


Figure14: Multigene FISH analyses with five probes targeting different genes on chromosome 8. The columns represents photographs taken from each specter, the last column are the superimposed image with all signals. The two first rows represent tumor cells, the third row is a lymphocyte serving as an internal control.

Copy number microarray analysis

Measurement of genomic variations was traditionally performed by karyotyping. Later, comparative hybridization with reference DNA on metaphases improved the detection[152]. Almost a decade ago, the first maps of the sequences in the human genome was published[153]. This, together with the technical improvement of array analyses and bioinformatical methods, opened for high resolution DNA analyses such as aCGH (array comparative genomic hybridization). The first published work with aCGH used a 2400 BAC array[154]. The most common type of aCGH is constructed by spotting DNA sequences (BAC, PCR fragments or synthetic oligonucleotides based) on glass slides (for Review;[155]). Sample DNA are compared to "standard" DNA (such as DNA from pooled blood cells from healthy individuals) marked with different fluorochromes and

hybridized to the array. The arrays are scanned to measure the intensities of the two fluorochromes per spot, and the ratio indicate if it is more (gain), less (loss) or no difference (no alteration) between the sample DNA and the reference. The amount of information from such experiments is enormous and different types of bioinformatical algorithms are used for quality control, adjustment of variation and visualization of the results.

As summarized in Table 2, data from three different types of aCGH platforms was analyzed in Paper II, and copy number variation deduced from SNP array was used in Paper IV. The Roma array (Representational Oligo Microarray Analysis) was developed at Cold Spring Harbor Laboratories (CSHL) to identify copy number polymorphisms and variations (CNP and CNV)[156]. In this method DNA was digested by BglII to reduce the complexity of the genome but still keep the analysis at a high resolution. The ROMA array is spotted with >83 000 DNA fragments distributed throughout the genome as illustrated in Figure 15.

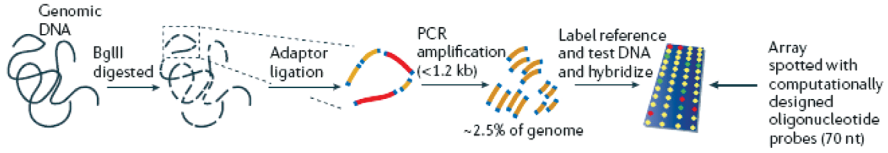


Figure 15: The ROMA platform. DNA is cut by BglII prior to adaptor ligation, DNA amplification, labeling, mixing with reference DNA and hybridization to the array. From Feuk et al.[157].

The Ull cohort was analyzed by an array designed total genomic DNA analyses without PCR amplification [158](Agilent). The array is spotted with 244 000 probes with a genome wide distribution. The data from the ChinUC cohort was from a custom made oligonucleotide based array with approximately 30 000 probes[159]. All three platforms are arrayed with oligonucleotides, but the Agilent and the custom made (ChinUC) are biased towards intragenic probes in contrast to the ROMA array. The Illumina SNP array used in Paper IV is based on a bead principle[160] and measure both signal intensity and changes in allelic composition identifying both copy number change and copy number neutral events (LOH; loss of heterozygosity)[161]. Comparison of data obtained from the ROMA, Agilent (44K) and Illumina platform has shown only minor discrepancies[162].

Methylation status analysis

For study of genome wide methylation we used MOMA (Methylation Detection Oligonucleotide Microarray Analysis), also developed at CSHL[163]. MOMA allows for high throughput analysis of classical CpG islands of size 200-2000bp. As for ROMA, MOMA is based on representations of DNA. After cutting and ligation with adapters each sample is divided into two. One part is digested with McrBC (cleaving DNA at methylated cytosine residues), the other part is mock digested to serve as a reference for comparative hybridization on the array (Fig. 16).

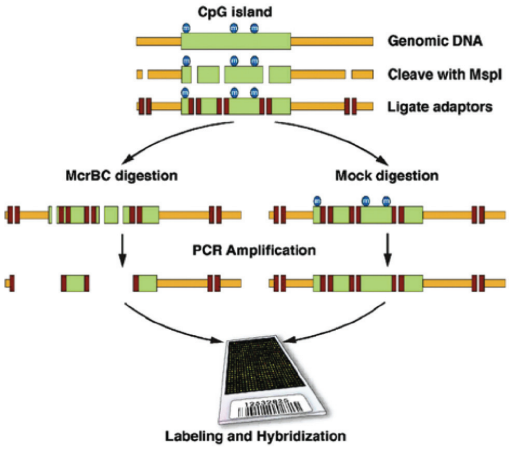


Figure 16: Schematic illustration of the principle behind MOMA. Genomic DNA is cleaved in CG rich areas, ligated to adaptors and split into two. One part is digested at methylated cytosine residues, the other not. After a balanced PCR reaction the two parts are mixed and hybridized to the array. From Kamalakaran et al.[163].

Paired-end sequencing

In a separate study a minor subset of samples from the MicMa and Ull cohort were analyzed by Paired end sequencing (Stephens et al, resubmitted, Nature). In Paper II the results from five of the tumors are used to illustrate the differences between the subgroups defined by different genomic architecture. In this method, DNA is fragmented into 400bp fragments where each end of every fragment are ligated to adapters and then sequenced[164]. The first 37 bases are sequenced from each end of the strands, and then mapped to the genome. Ends that do not map as expected indicate a structural rearrangement, such as a translocation, duplication, inversion or amplification (Figure

17). By mapping several overlapping fragments both breakpoints and type of rearrangement can be identified.

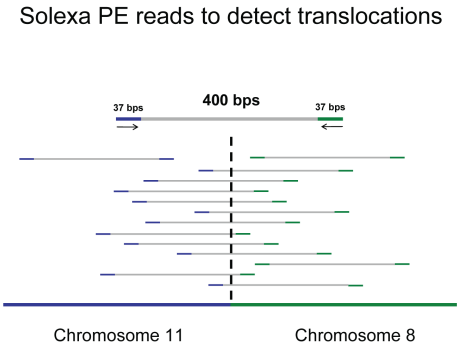


Figure 17: DNA fragments are sequenced only from each end and mapped to the genome. This illustrates the identification of a translocation between Chromosome 11 and 8, mapped by several overlapping fragments. Published with permission from P. Stephens.

Bioinformatical and statistical methods

Most of the bioinformatical based tools developed in paper II, III and IV are in Java, R or MATLAB codes. Statistical analyses were performed by SPSS 15.0 in paper I and II and by R in paper III and IV. In paper I and II, associations between categorical or continuous values were assessed by Pearson chi square, Fishers exact or Kruskal-Wallis tests[165]. Hierarchical cluster analyses and t-tests used to identify subgroups and altered loci of methylation in paper III were performed in R as were the Genetic Algorithm (GA) and the survival analyses. The survival analyses in paper II and III were based on the Cox' proportional hazard method and log rank tests[165]. More detailed description of both the bioinformatical codes and statistical methods are given in the respective papers.

Table 2:

Method	Name of system/platform	Type of tissue used	Structure of interest (target)	Appr. no. of individual targets per analysis	Advantages	Limitations/challenges
IHC	DAKO Envision	FFPE: TMA, whole sections, imprints	Protein	1	Identifies the localization in different types of cells in the tissue	One target analyzed at the same time
					Identifies the localization in subcellular compartments	The specificity and sensitivity of the antibodies are crutical
					TMA: Many patients analyzed under the same conditions	Subjective, visual scoring
Expression analysis (cDNA arrays)	Stanford arrays	Fresh frozen tissue	RNA (cDNA)	42 000	Extreme number of targets analysed at the same time with same conditions per sample	Dependent on high quality and not degraded RNA
					Quantification of transcripts	Probe annotations make comparison different
					Custom made to identify only targets of interest	Demanding Bioinformatics, and huge amounts of data needs to handled
						Information is from the mean value, cannot reveal heterogeneity easily
					Small amounts of tissue/RNA needed	Studies uses different platforms making comparison and validation challenging
DNA ploidy measurements	Feulgen staining	Imprints (frozen cells on glass slide)	DNA	DNA content per cellnucleus	Measure ever nucleus individually and can identify subclones with aberrant DNA content	Need nuclei from normal cells for comparison
						Can not sort cells into different fractions
						Rough estimates; minor genomic aberrations are not seen
FISH	QM-FISH	Imprints (frozen cells on glass slide)	DNA	1 to 5	Identifies the localization in different types of cells in the tissue	Few targets analyzed at the same time
					Can identify subclones	The specificity and the size of the probes are crutical
	Vysis	TMA			Identifies the physical relationship between targets (genomic rearrangements)	Subjective, visual scoring
					Quantification	
aCGH	ROMA	Fresh frozen tissue	DNA (representations)	83 000	Extreme number of targets analysed at the same time with same conditions per sample	Information is from the mean value, cannot reveal heterogeneity easily
	Agilent	Fresh frozen tissue	DNA	244 000	Quantification of copynumber	Dependent on selection of probes/targets
					Custom made to identify only targets of interest	Hybridizing conditions are crucial
	Illumina	Fresh frozen tissue	DNA (snp)	100 000	Can identify both germline variations and somatic alterations	Demanding Bioinformatics, and huge amounts of data needs to handled
					Can reveal allel specific copynumber variation (LOH)	Studies uses different platforms making comparison and validation challenging
	Custom made (ChinUC)	Fresh frozen tissue	DNA	32 000	Small amounts of tissue/DNA needed	
Methylation array	MOMA	Fresh frozen tissue	DNA (CpG islands)	390 000	Extreme number of targets analysed at the same time with same conditions per sample	Information is from the mean value, cannot reveal heterogeneity easily
					Measure the level of methylation of targets	Dependent on selection of probes/targets (MsPI)
					Custom made to identify only targets of interest	Ligation, digestions and hybridizing, conditions are crucial
						Demanding Bioinformatics, and huge amounts of data needs to handled
Paired-end sequencing	Pres1/Illumina	Fresh frozen tissue	DNA	Individual per sample	Extreme number of targets analysed at the same time with same conditions per sample	Need several DNA fragments (reads) to define a rearrangements makes sensitivity low
					No preselection of targets	No reads from repetetive segments or heterochromatin
					Identify both inter and intra chromosomal targets	Selection of reads due to lenght
					Identify minor rearrangments	Demanding Bioinformatics, and huge amounts of data needs to handled

Summary of results

Paper I: "Paired distribution of molecular subtypes in bilateral breast carcinomas"

Hege G. Russnes, Ekatherina Sh. Kuligina, Evgeny N. Suspitsin, Ekaterina S. Jordanova, Cees J. Cornelisse, Anne-Lise Børresen-Dale, Evgeny N. Imyanitov

Under review, Molecular Oncology

Tumors arising in both breasts in a female are rare but represent a unique setting to explore the relationship between host-related factors and tumor phenotype. In this study, we analyzed 100 tumors from fifty women with bilateral disease. Of these, 23 had synchronous disease (tumor in the contralateral breast diagnosed within a year from the first) and 27 had metachronous disease (tumor in the contralateral breast diagnosed more than a year after the first). As the tumors had been preserved as FFPE tissue, we chose to classify them into molecular subtypes by IHC. Six antibodies were selected as surrogate IHC markers to identify tumors as Luminal ('Luminal'), triple-negative Basal-like ('TN-Basal'), triple-negative unclassified ('TN-UNC') or heterogeneously ('Heterogenous') stained tumors. Clinico-pathological data as well as *BRCA1* mutation status were available. We found that in bilateral disease, synchronous tumors showed a slightly higher rate of concordant pairs than metachronous tumors, and Luminal tumors were highly concordant regardless of being synchronous or metachronous. Metachronous cases had a higher degree of discordance if the time interval was more than 10 years, and this was especially pronounced when the first tumor was of the TN-Basal type. The TN-Basal tumors with a short time interval were all concordant, while those with a long time interval were highly discordant. These findings points to host related factors being important for the development of Luminal-like tumors. The TN-Basal tumors of synchronous and metachronous type with short time span were also highly concordant, pointing to host related factor in this type of carcinomas as well. In addition, the data reflect the acknowledged heterogeneity of Basal-like carcinomas. Metachronous TN-Basal and TN-UNC tumors with longer time span than five years were highly discordant

and suggest that the second tumor arising in these women have different causes dominated by stronger environmental influences than genetic factors.

This study provides additional evidence for the role of host factors determining the molecular subtypes of breast cancer disease, indicating that both germline variations and hormonal status are of importance. Such knowledge can provide important information about selection of treatment for the first cancer that would also provide as prevention for contralateral breast cancer.

Paper II: "Genomic architecture characterizes tumor progression paths and fate in breast cancer patients"

Hege G. Russnes, Hans Kristian Moen Vollan, Ole Christian Lingjærde, Alexander Krasnitz, Pär Lundin, Bjørn Naume, Therese Sørlie, Elin Borgen, Inga H. Rye, Anita Langerød, Suet-Feung Chin, Andrew E. Teschendorff, Philip J. Stephens, Susanne Månér, Ellen Schlichting, Lars O. Baumbusch, Rolf Kåresen, Michael P. Stratton, Michael Wigler, Carlos Caldas, Anders Zetterberg, James Hicks, Anne-Lise Børresen-Dale

Submitted Nature Medicine

The era of genome-wide high resolution analyses have increased the amount of detailed knowledge about molecular alterations in breast cancer, but the physical distortion of the genome is seldom attributed. In breast carcinomas a variety of structural distortion patterns have been identified by karyotyping and this is now supported by detailed sequencing analyses. Karyotyping require viable tumor cells and is only appropriate for smaller studies, it is time consuming and does not reveal detailed information about rearrangements. Sequencing analyzes are costly and time consuming in contrast to aCGH analyses. The aim of this study was therefore to construct objective estimates of genomic architectural alterations in high-resolution aCGH data and to apply this to several breast cancer cohorts to increase sample size in order to be able to explore the relationship between genomic alterations, molecular expression subtypes, structural rearrangements, pathology and clinical data. By making platform independent scores to 1) identify either

gain or loss of whole chromosome arms (WAAI) and 2) identifying complex rearrangements of chromosome arms (CAAI), we were able to merge four different breast cancer cohorts analyzed on three different aCGH platforms and thus relate genomic architectural distortion to various types of data from a total of 595 breast cancer patients.

By using WAAI, we sub-stratified the merged cohort into Luminal (*A* tumors) and non-luminal tumors (*B* tumors) based on selected genomic surrogate markers known to distinguish the Luminal A and Basal-like subtype. By doing this we also found a group with combination of Luminal A and Basal-like markers (*AB* tumors), and a group with none of the markers (*C* tumors). The selected markers for *A* tumors were either gain of 1q (whole arm) and/or loss of 16q (whole arm), while regional loss on 5q and/or gain of 10p were selected as markers for *B* tumors. The four groups showed that the A group was enriched in Luminal A tumors and the B group in Basal-like tumors. Interestingly, Luminal B, erbB2 and Normal-like tumors were found in all groups, but the latter two subtypes were more frequent in the *C* tumors. Complex rearrangements as defined by CAAI occurred in all subgroups, and were used to subdivide each of them making a total of eight different WAAI/CAAI defined groups (*A1*, *B1*, *AB1*, *C1* with no/low CAAI and *A2*, *B2*, *AB2*, *C2* with high CAAI). The groups displayed very different types of genomic distortion. The *A* tumors were dominated by gain or loss of whole chromosomes and chromosome arms and *B* tumors by genomic losses and more regional aberrations. This difference were also evident by the few samples selected for paired-end sequencing; the *A* tumor had only one alteration, compared to the *B* tumor having genome wide duplications and several translocations. The complex rearrangements measured by CAAI had distinct patterns, with chromosome arms 8p and 11q most frequently affected in *A* tumors in contrast to *B* tumors having 17q and 20q as frequent affected arms. The pattern of genomic distortion and the ploidy status of *A* and *B* tumors indicated that a progression from *A1* to *A2* probably occurs along a linear path. Such a progression was less clear for the *B* tumors. A resemblance between *B*, *AB* and groups of *C* tumors probably reflect a relationship between the non-Luminal tumors. The WAAI defined groups had significant differences in outcome (breast cancer specific death) and CAAI had a strong prognostic impact, reflecting that patients with tumors with complex rearrangements, even of only one chromosome arm, had a worse outcome independently of other factors. An

established prognostic index such as histological grad had a strong prognostic impact in *A* tumors but not in *B* and *AB* tumors, reflecting the importance of acknowledging the different properties of molecular subgroups. This study show how genomic architecture can be used to more robustly define molecular subtypes of breast carcinomas and that genomic distortion such as complex rearrangements constitute a new prognostic tool in breast cancer.

Paper III: "Subtype dependent alterations of the DNA methylation landscape in breast cancer and implications for prognosis"

Sitharthan Kamalakaran, Hege E. Gierckskey Russnes, Vinay Varadan, Dan Levy, Jude Kendall, Angel Janevski, Michael Riggs, Nilanjana Banerjee, Marit Synnestvedt, Ellen Schlichting, Rolf Kåresen, Robert Lucito, Michael Wigler, Nevenka Dimitrova, Bjørn Naume, Anne-Lise Børresen-Dale, James B. Hicks

Manuscript

This study was designed to measure the levels of DNA methylation of breast carcinomas by performing high-throughput genome-wide scans of CpG methylation by the MOMA technology. By analyzing breast carcinomas (n=114) and normal breast tissue (n=11) we aimed at 1) identifying tumor specific methylation patterns, 2) subgroup tumors based on methylation patterns and 3) identifying loci with prognostic value. Unsupervised hierarchical clustering using the 500 most differentially methylated loci across all tumors and the 100 most significant altered loci between tumors and normal tissues clustered the tumors into 3 major clusters. As the cohort previously had been classified into the five gene expression subtypes, a comparison between the three groups and the molecular subtypes was performed. Cluster I, was enriched in luminal subtypes (Luminal A and Luminal B) in contrast to cluster II which were dominated by the Basal-like and erbB2+ subtypes. Cluster III did not show any expression subtype specific enrichment, the majority of the samples belonged to a group of tumors having inconclusive or only weak correlations to multiple expression subtypes. The three groups showed a high correlation to the DNA based WAAI/CAAI groups as well; cluster I was dominated by *A* tumors,

cluster II by *B* tumors and cluster III by *C* and *A* tumors. Interestingly, the latter cluster had only few samples with complex rearrangements. Methylation loci that contributed to this clustering were only infrequently localized to CpG islands upstream of genes, suggesting that there are subtype dependant genome-wide alterations in the methylation landscape in breast cancers. Of the loci mapped to known genes, more than half of them showed significant correlation to gene expression, implying possible functional effects of the methylation on gene expression. Additionally, distinct expression subtype specific patterns of methylation could be detected in known cancer associated genes. CpG islands in the HOXA gene cluster and many other homeobox genes were significantly more methylated in Luminal A tumors. Several of the loci discriminating between Basal-like and Luminal A are known to be differentially methylated in myoepithelial and luminal progenitor cells in the normal breast. The methylation patterns of genes characterizing Luminal A tumors resemble those identified in CD24+ luminal epithelial cells and the loci in Basal-like tumors resemble CD44+ breast progenitor cells indicating that Basal-like and Luminal A tumors might originate from cells at different levels in the breast epithelial cell hierarchy. Furthermore, analysis of these tumors by using follow-up survival data allowed an identification of genes whose methylation state was associated to poor outcome.

Paper IV: "Novel tool reveals copy number aberrations in tumors (ASCAT)"
Peter Van Loo, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M. Perou, Anne-Lise Børresen-Dale, Vessela N. Kristensen
Submitted Nature Biotechnology

In this study, SNP array data from the 102 breast carcinomas were used to deduce tumor ploidy, contaminating tissue involvement, intra-tumor heterogeneity and allele specific aberrations by a novel bioinformatics approach, ASCAT. SNP arrays measure both signal intensity and changes in allelic composition and, in contrast to aCGH, it is possible to identify both copy number change and copy number neutral events. ASCAT's consistency and sensitivity to a lowering percentage of aberrant tumor cells was validated

by applying the algorithm to a dilution series of a tumor sample mixed with different proportions of its germline DNA. In addition, FISH analyses of selected, frequently amplified genomic regions were performed on 11 tumors. The ploidy estimations by ASCAT were validated by image DNA cytometry of 79 tumors. The copy number counts from FISH analyses were highly concordant with the copy number estimates by ASCAT in the selected loci, as were the ploidy estimates compared to the results from image DNA cytometry. Together, these validation experiments confirm that ASCAT accurately predicts allele-specific copy number profiles of tumors over a broad range of tumor ploidy and fraction of aberrant tumor cells.

Furthermore, ASCAT revealed differences in non-aberrant cell infiltration, ploidy, gains, losses, LOH and copy number neutral events between the five molecular breast cancer subtypes. Finally, ASCAT allowed a detection of allelic skewness and by this we identified several novel markers of breast cancer.

Methodological considerations

All four studies included in this thesis were based upon analyzes of clinical tumor samples, either as frozen tumor biopsies or FFPE, in addition were matching blood samples used for SNP analyses. An advantage in usingng clinical samples from different patient cohorts is that they a spectrum of the disease including different subtype and progression levels can be represented. This is in contrast to functional studies based on cell-lines and xenografts, where the diversity of a cohort is lost. The limitations are to be acknowledged. It would be unethical to study progression of individual tumors as not removing a tumor by surgery would be unethical. Clinical cohorts can have a selection bias related to many factors, a bias towards heavily treated patients with large tumors are not uncommon. In series such as the Ull cohort sequentially collected by surgeons, smaller tumors are often not included as the doctors will not dare to ruin the histopathological examination. The MicMa series is part of a larger cohort of patients collected at several different hospitals and not only on university hospitals (which can have an overrepresentation of large and rarer tumor types). In this cohort 130 tumors had fresh frozen tumor tissue available, and these seem to have a skewed distribution towards more advanced tumors than the rest of the cohort. Both the WZ and the ChinUC cohort was drawn from tissue banks, an advantage is that the tumor samples often are collected by a pathologist which can secure also pieces from minor tumors. In the merged cohort analyzed in paper II, the WZ tumors (selected for diploids) and the ChinUC tumors were important contributors to the descriptive analyses as the set got enrich in tumors probably at an early stage of progression. The WZ set was omitted from analyses regarding outcome; its selection criteria was to have equal distribution of survivors and non survivors. In addition the clinical information was not collected and secured by a clinician in contrast to the three other cohorts.

Frozen sections were analyzed by microscopy to secure tumor representativity of the biopsies, but the variations even in small tumor pieces can be huge. Some tumors have huge DCIS components and only minor areas with invasion. An example of tumor heterogeneity influencing the analyses was seen in a tumor classified as Basal-like by gene expression analyses. The part of the tumor investigated by image DNA cytometry

showed a clear diploid profile, while the DNA extracted from another part of the tumor showed by ASCAT an aneuploid profile. This is not misinterpretation of the analyses but reflect the heterogeneity in some carcinomas.

All methods in these four papers have advantages and disadvantages as summarized in table 2. The design and properties of each method vary enormously from analyzing one target (FISH/IHC) to multiple predefined targets (microarrays) or unknown targets (paired-end sequencing). As separate clinical cohorts often have tissue preserved differently and of limited amounts, applying the same method on several cohorts is in often impossible. In paper I, the state of the tissue made IHC analyses of TMA the method of choice to classify the tumors into the molecular subtypes. Selection of markers was, as reviewed in the paper, based on previous literature. Due to limitations in tissue availability and the work being performed as a collaboration between two laboratories, two different detection systems were used. The major problem in classifying this cohort was the sample size, and recognizing the major groups was therefore the focus. The HER2 marker was of that reason not used as a surrogate marker. It is also debatable if it was wise to split the triple negative cases into a 'TN-Basal' and a 'TN-unclassified', but this decision was based on the known heterogeneity of non-Luminal tumors.

FISH analyses is to date still difficult to score in an objectively way. A major advantage is the in situ visualization of signals in each nucleus. The size, shape and location of the signals are important to be able to avoid false interpretations. All counting of signals in paper IV was performed visually. To be representative, three to four different areas on the imprints were used, and the mean value from 20 cells were chosen to represent the copy number for a given gene/probe from each tumor. To identify translocations, the same combination of probes needed to be in close approximation to each other in several cells to be regarded as a translocation. This is easy to interpret in diploid tumors, but much more challenging in aneuploid tumors where the signals were more numerous. In such comprehensive FISH analyses using multiple probes it is important to keep in mind the few number of cells analyzed, and in heterogeneous tumors the findings can probably not be generalized.

Ploidy measurements performed by image DNA cytometry and are in this work scored visually by choosing the mode value as the DNA index. The histograms from

some tumors show broad distributions of DNA content, and a more dynamic type of measurements could have been advantageous. In paper II ploidy was used solely as a measure for progression and a rough estimate and categorization into diploid and aneuploid tumors were therefore used. In paper IV the mode values were compared to the ASCAT estimates. Interestingly, most of the tumors that ASCAT could not be applied to were highly aneuploid with a broad distribution probably reflecting multiple subclones.

Microarray analyses are designed to give information about numerous targets; in this thesis the patterns of aberrations have been the main focus and not single genes/loci or groups of genes. The use of expression array analyses to deduce the molecular subtypes have been discussed previously, but it is important to keep in mind that this classification is based on few genes extracted from analyzes on a small tumor set with advanced tumors. It is shown that by adding genes to the list, additional expression subtypes seem to emerge[166]. Microarrays measuring copy number variation have various types of design, but this thesis point to one major feature; the SNP arrays ability to deduce allele specific alterations, measure the influence of non-aberrant cells and deduce ploidy state compared to the CGH arrays. The differences between various types of CGH arrays can be overcome as illustrated by paper II. Construction of bioinformatical codes that easily can be tailored to each type of aCGH data (i.e. centering and PCF segmentation) gave data that could be the input to the WAAI and CAAI algorithm making these them platform independent. It has to be acknowledged that the probe selections on the three types of arrays are fundamentally different, the 32K customized array and Agilent being gene centered while ROMA were not. The WAAI and CAAI scores were validated primarily in the ROMA data as HER2 FISH and paired-end sequencing was available for several of the tumors. CAAI and WAAI were carefully tailored to recognize complex rearrangements and whole arm alterations, and this was confirmed by visual inspection of all aCGH profiles. Visual inspection is a subjective estimate not good enough as validation, but it was important to do as a quality control of the estimates. Samples with whole arm alterations but with either a high standard deviation or low amplitude would not get an elevated WAAI score. Such samples are difficult to interpret and WAAI was designed to take this into account not to get too many false positive scores. This resulted in false negative samples (samples where visual

inspection indicated whole arm gain or loss of 1q and 16q but not classified as *A* tumors by WAAI classification). CAAI was designed to recognize the complex alterations defined as regions with high-level amplicons separated by short deletions (firestorms). Although the BFB mechanism can in theory explain such alterations, more detailed analyses by paired end sequencing indicate that several mechanisms are involved. It is to be mentioned that CAAI is not reflecting the complex type of rearrangements called 'saw-tooth', and comparing the WAAI and CAAI distribution in B tumors with the frequency plots (Figure 3 and Supplementary Figure 5 in paper II) it is obvious that defining a third parameter to capture such rearrangements as well would be an advantage.

The nature of genomic rearrangements is until now defined primarily by cytogenetics, and the transfer of concepts and definitions from karyotyping to detailed studies such as paired-end sequencing is difficult. The details about intra- and inter-chromosomal rearrangements are starting to emerge, and alterations discovered can not be fully covered by the existing 'nomenclature' of cytogenetics. It is to be expected that the new level of resolution in genomic analyses will demand and define such a nomenclature, and will bring new insight into the mechanisms behind the different architectural distortion patterns we observe in breast carcinomas.

The studies in this thesis combine information from different analyses in a pragmatic way based on established statistical methods. We are in an era where integrative approaches are being the main focus in a recently established science called 'systems biology', and major achievements in statistics and bioinformatics are to be expected leading to new understanding in the complex field of cancer biology.

Main conclusions and future aspects

The studies included in this thesis support the existence of at least two major types of breast carcinomas, one with features related to Luminal cells, the other to cells with stemcell/Basal-like properties. The tumors progressing along a luminal path are often ER and/or PgR positive and can have luminal differentiation as seen by histopathology. This phenotype is also reflected by gene expression. At the genomic level such tumors are often diploid but have characteristic gains and losses of whole chromosomes or chromosome arms, the latter can be explained by whole arm translocation frequently involving chromosome 1 or 16. The tumors have a good prognosis, but if luminal tumors get more complex rearrangements, the outcome is worse. Such tumors probably reflect a more advance stage in progression as they are frequently aneuploid and have high proliferation and are less differentiated. An established prognostic factor such as histological grade is important to identify patients with a worse prognosis, but this is only to be of benefit in luminal related tumors. At the genomic level luminal tumors rarely have mutations in TP53, and have few structural genomic rearrangements.

The tumors progressing along the Basal-like/stem cell path are typical ER and/or PgR negative; expression analyses and methylation patterns link this subtype to basal- and stem cells. They have a distinct ploidy pattern being diploid/hypodiploid in a early phase and aneuploid close to the triploid region in a later phase. Tumors in the first phase are dominated by genomic losses, while tumors in the aneuploid phase are showing genome wide complexity including complex rearrangements with high level amplicons. Basal-like tumors have minor genomic duplications scattered in the genome, and the more advanced tumors seem to have complex rearrangements in addition. This is also reflected by the expression pattern as a shift can be observed towards the erbB2+ and Luminal B centroid. The mechanism behind this unstable genome is unknown, but mutations in TP53 are a frequent alteration.

In paper II we found that several tumors could be grouped both as an *A* and a *B* tumor and were thus designated as AB tumors. By visual inspection they rarely belonged to the typical ‘simplex’ (luminal) pattern but had often rearrangements on almost all chromosomes, including loss and gain interpreted by WAAI, with whole arm alteration of

1q or 16q. Most of the additional analyzes on such samples support that they are related to the B/Basal-like type of tumors, but only functional studies can tell if a tumor can switch from a luminal path to a Basal-like or vice versa.

In addition to these entities minor groups have also emerged; in paper II we identify a group of tumors without any of the selected markers which frequently had complex rearrangements on 17q. This group was dominated by erbB2+ and normal-like samples, and two of the samples were DCIS. As shown in Fig.6 the clustering of the five centroid values revealed a group dominated by expression towards the same two centroids also indicating an independent type of tumors.

To get closer to defining distinct entities and their relationship, the next step will be functional studies. As a part of the OSLO2 study, fresh tumor samples are collected and disaggregated into single cells preserved in a viable state. Fluorescence Activated Cell Sorting (FACS) will be used to sort tumor cells into different fractions by applying different antibodies targeting various cell surface markers. The markers will identify cells representing different stages in differentiation (such as breast stem cells and more mature myoepithelial or luminal related cells) in addition to other cells in breast tissue such as fibroblasts, lymphocytes, adipocytes and endothelial cells. Sorted subpopulations will further be analyzed both at the genomic level (sequencing/SNP/copy number variation/methylation) and at the expression level (RNA/miRNA/protein)

If some of the collected tumor samples have viable cells that grow in culture, the level of environmental stimuli can be mimicked and varied. The level of differentiation can be measured both visually and by gene expression analyzes aiming at identifying subgroups of tumor cells that show more or less plasticity with regard to direction of differentiation and to investigate whether such changes imply genomic aberrations as well.

Molecular alterations characterizing a subclone of importance will be selected to be analyzed by technologies such as IHC and FISH using tissue sections and TMAs to be able to go back to the cohorts used in this thesis where so much additional information is available. Such in situ studies will serve as an important validation of findings; it also makes it possible to visually identify which cells have the alterations, and where they reside in a tissue architectonical context. Larger sample sets of breast carcinoma can thus

be analyzed, the clinical impact can be evaluated and the search for a robust molecular classification based on more knowledge from the hierarchical relationship can continue. If individualized therapy is to become a reality in the near future, a robust molecular based classification of breast cancer will be of major importance.

Reference list

1. Kamangar,F., Dores,G.M., & Anderson,W.F. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J. Clin. Oncol.* **24**, 2137-2150 (2006).

2. Cancer Registry of Norway. Cancer in Norway 2007 - Cancer incidence, mortality, survival and prevalence in Norway. 2009. 2008.
Ref Type: Report

3. Småstuen M, Aagnes B, ,J.T., ,M.B., & ,B.F. Long-term cancer survival: patterns and trends in Norway 1965-2007. ed. Oslo: Cancer Registry of Norway, 2. 2008.
Ref Type: Report

4. Dumitrescu,R.G. & Cotarla,I. Understanding breast cancer risk -- where do we stand in 2005? *J. Cell Mol. Med.* **9**, 208-221 (2005).

5. CLEMMESEN,J. Carcinoma of the breast; results from statistical research. *Br. J. Radiol.* **21**, 583-590 (1948).

6. Hartman,M. *et al.* Incidence and prognosis of synchronous and metachronous bilateral breast cancer. *J. Clin. Oncol.* **25**, 4210-4216 (2007).

7. Hemminki,K. & Vaittinen,P. Familial risks in second primary breast cancer based on a family cancer database. *Eur. J. Cancer* **35**, 455-458 (1999).

8. Kalager,M. *et al.* Improved breast cancer survival following introduction of an organized mammography screening program among both screened and unscreened women: a population-based cohort study. *Breast Cancer Res.* **11**, R44 (2009).

9. Swerdlow,S.H. *et al. WHO classification of Tumors of Haematopoietic and Lymphoid Tissue*(WHO Press,2008).

10. Villadsen,R. *et al.* Evidence for a stem cell hierarchy in the adult human breast. *J. Cell Biol.* **177**, 87-101 (2007).

11. Anbazhagan,R. *et al.* The development of epithelial phenotypes in the human fetal and infant breast. *J. Pathol.* **184**, 197-206 (1998).

12. Dontu,G., El-Ashry,D., & Wicha,M.S. Breast cancer, stem/progenitor cells and the estrogen receptor. *Trends Endocrinol. Metab* **15**, 193-197 (2004).

13. Smalley,M. & Ashworth,A. Stem cells and breast cancer: A field in transit. *Nat. Rev. Cancer* **3**, 832-844 (2003).

14. Clarke,R.B., Anderson,E., Howell,A., & Potten,C.S. Regulation of human breast epithelial stem cells. *Cell Prolif.* **36 Suppl 1**, 45-58 (2003).

15. Polyak,K. Breast cancer: origins and evolution. *J. Clin. Invest* **117**, 3155-3163 (2007).

16. Stingl,J., Raouf,A., Emerman,J.T., & Eaves,C.J. Epithelial progenitors in the normal human mammary gland. *J. Mammary. Gland. Biol. Neoplasia.* **10**, 49-59 (2005).

17. Tavassoli,F.A. & Devilee,P. *World Health Organization Classification of Tumors. Pathology and Genetics of Tumors of the Breast and Female Genital Organs.*(IARC press,2003).

18. Reis-Filho,J.S. & Lakhani,S.R. Breast cancer special types: why bother? *J. Pathol.* **216**, 394-398 (2008).

19. Rakha,E.A. & Ellis,I.O. An overview of assessment of prognostic and predictive factors in breast cancer needle core biopsy specimens. *J. Clin. Pathol.* **60**, 1300-1306 (2007).

20. BLOOM,H.J. Further studies on prognosis of breast carcinoma. *Br. J. Cancer* **4**, 347-367 (1950).

21. BLOOM,H.J. Prognosis in carcinoma of the breast. *Br. J. Cancer* **4**, 259-288 (1950).

22. Elston,C.W. & Ellis,I.O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**, 403-410 (1991).

23. Sotiriou,C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262-272 (2006).

24. Balleine,R.L. *et al.* Molecular grading of ductal carcinoma in situ of the breast. *Clin. Cancer Res.* **14**, 8244-8252 (2008).

25. Carter,C.L., Allen,C., & Henson,D.E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63**, 181-187 (1989).

26. Rosen,P.P., Groshen,S., & Kinne,D.W. Survival and prognostic factors in node-negative breast cancer: results of long-term follow-up studies. *J. Natl. Cancer Inst. Monogr*159-162 (1992).

27. Singletary,S.E. & Greene,F.L. Revision of breast cancer staging: the 6th edition of the TNM Classification. *Semin. Surg. Oncol.* **21**, 53-59 (2003).

28. Sobin LH & Wittekind C. The TNM classification of Mailgnant tumors. 2002.
Ref Type: Generic

29. Haybittle,J.L. *et al.* A prognostic index in primary breast cancer. *Br. J. Cancer* **45**, 361-366 (1982).

30. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365**, 1687-1717 (2005).

31. Fisher,B., Redmond,C., Fisher,E.R., & Caplan,R. Relative worth of estrogen or progesterone receptor and pathologic characteristics of differentiation as indicators of prognosis in node negative breast cancer patients: findings from National Surgical Adjuvant Breast and Bowel Project Protocol B-06. *J. Clin. Oncol.* **6**, 1076-1087 (1988).

32. Cleator,S.J., Ahamed,E., Coombes,R.C., & Palmieri,C. A 2009 update on the treatment of patients with hormone receptor-positive breast cancer. *Clin. Breast Cancer* **9 Suppl 1**, S6-S17 (2009).

33. McGuire,W.L., Chamness,G.C., Costlow,M.E., & Richert,N.J. Steroids and human breast cancer. *J. Steroid Biochem.* **6**, 723-727 (1975).

34. Dhesy-Thind,B. *et al.* HER2/neu in systemic therapy for women with breast cancer: a systematic review. *Breast Cancer Res. Treat.* **109**, 209-229 (2008).

35. Meltzer,P.S. Gene expression profiling in breast cancer research. *Breast Dis.* **19**, 23-27 (2004).

36. Ross,J.S., Hatzis,C., Symmans,W.F., Pusztai,L., & Hortobagyi,G.N. Commercialized multigene predictors of clinical outcome for breast cancer. *Oncologist.* **13**, 477-493 (2008).

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

38. 't Veer,L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).

39. Van De Vijver,M.J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999-2009 (2002).

40. Chang,H.Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. U. S. A* **102**, 3738-3743 (2005).

41. Paik,S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817-2826 (2004).

42. Cammenga,J. Gatekeeper pathways and cellular background in the pathogenesis and therapy of AML. *Leukemia* **19**, 1719-1728 (2005).

43. WATSON,J.D. & CRICK,F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

44. NOWELL,P.C. & HUNGERFORD,D.A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **25**, 85-109 (1960).

45. Hahn,W.C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400**, 464-468 (1999).

46. Hanahan,D. & Weinberg,R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

47. Luo,J., Solimini,N.L., & Elledge,S.J. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* **136**, 823-837 (2009).

48. Stratton,M.R., Campbell,P.J., & Futreal,P.A. The cancer genome. *Nature* **458**, 719-724 (2009).

49. Howlett,A.R. & Bissell,M.J. The influence of tissue microenvironment (stroma and extracellular matrix) on the development and function of mammary epithelium. *Epithelial Cell Biol.* **2**, 79-89 (1993).

50. Bissell,M.J. *et al.* Tissue structure, nuclear organization, and gene expression in normal and malignant breast. *Cancer Res.* **59**, 1757-1763s (1999).

51. Hu,M. & Polyak,K. Molecular characterisation of the tumour microenvironment in breast cancer. *Eur. J. Cancer* **44**, 2760-2765 (2008).

52. FOULDS,L. The experimental study of tumor progression: a review. *Cancer Res.* **14**, 327-339 (1954).

53. Hicks,J. *et al.* High-resolution ROMA CGH and FISH analysis of aneuploid and diploid breast tumors. *Cold Spring Harb. Symp. Quant. Biol.* **70**, 51-63 (2005).

54. Natrajan,R. *et al.* Tiling path genomic profiling of grade 3 invasive ductal breast cancers. *Clin. Cancer Res.* **15**, 2711-2722 (2009).

55. Chin,K. *et al.* In situ analyses of genome instability in breast cancer. *Nat. Genet.* **36**, 984-988 (2004).

56. Chin,K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529-541 (2006).

57. Fridlyand,J. *et al.* Breast tumor copy number aberration phenotypes and genomic instability. *BMC. Cancer* **6**, 96 (2006).

58. Andre,F. *et al.* Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin. Cancer Res.* **15**, 441-451 (2009).

59. Korsching,E. *et al.* Deciphering a subgroup of breast carcinomas with putative progression of grade during carcinogenesis revealed by comparative genomic hybridisation (CGH) and immunohistochemistry. *Br. J. Cancer* **90**, 1422-1428 (2004).

60. Tirkkonen,M. *et al.* Molecular cytogenetics of primary breast cancer by CGH. *Genes Chromosomes. Cancer* **21**, 177-184 (1998).

61. Baudis,M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC. Cancer* **7**, 226 (2007).

62. Climent,J., Garcia,J.L., Mao,J.H., Arsuaga,J., & Perez-Losada,J. Characterization of breast cancer by array comparative genomic hybridization. *Biochem. Cell Biol.* **85**, 497-508 (2007).

63. Rennstam,K. *et al.* Patterns of chromosomal imbalances defines subgroups of breast cancer with distinct clinical features and prognosis. A study of 305 tumors by comparative genomic hybridization. *Cancer Res.* **63**, 8861-8868 (2003).

64. Teixeira,M.R., Pandis,N., & Heim,S. Cytogenetic clues to breast carcinogenesis. *Genes Chromosomes. Cancer* **33**, 1-16 (2002).

65. Buerger,H. *et al.* Comparative genomic hybridization of ductal carcinoma in situ of the breast-evidence of multiple genetic pathways. *J. Pathol.* **187**, 396-402 (1999).

66. Buerger,H. *et al.* Genetic relation of lobular carcinoma in situ, ductal carcinoma in situ, and associated invasive carcinoma of the breast. *Mol. Pathol.* **53**, 118-121 (2000).

67. Vos,C.B. *et al.* Genetic alterations on chromosome 16 and 17 are important features of ductal carcinoma in situ of the breast and are associated with histologic type. *Br. J. Cancer* **81**, 1410-1418 (1999).

68. Lakhani,S.R., Collins,N., Stratton,M.R., & Sloane,J.P. Atypical ductal hyperplasia of the breast: clonal proliferation with loss of heterozygosity on chromosomes 16q and 17p. *J. Clin. Pathol.* **48**, 611-615 (1995).

69. Lakhani,S.R. *et al.* Detection of allelic imbalance indicates that a proportion of mammary hyperplasia of usual type are clonal, neoplastic proliferations. *Lab Invest* **74**, 129-135 (1996).

70. Lakhani,S.R., Collins,N., Sloane,J.P., & Stratton,M.R. Loss of heterozygosity in lobular carcinoma in situ of the breast. *Clin. Mol. Pathol.* **48**, M74-M78 (1995).

71. Vos,C.B. *et al.* E-cadherin inactivation in lobular carcinoma in situ of the breast: an early event in tumorigenesis. *Br. J. Cancer* **76**, 1131-1133 (1997).

72. Waldman,F.M. *et al.* Genomic alterations in tubular breast carcinomas. *Hum. Pathol.* **32**, 222-226 (2001).

73. Pandis,N. *et al.* Chromosome analysis of 97 primary breast carcinomas: identification of eight karyotypic subgroups. *Genes Chromosomes. Cancer* **12**, 173-185 (1995).

74. Nishizaki,T. *et al.* Genetic alterations in lobular breast cancer by comparative genomic hybridization. *Int. J. Cancer* **74**, 513-517 (1997).

75. Simpson,P.T. *et al.* Columnar cell lesions of the breast: the missing link in breast cancer progression? A morphological and molecular analysis. *Am. J. Surg. Pathol.* **29**, 734-746 (2005).

76. Friedrich,K. *et al.* Comparative genomic hybridization-based oncogenetic tree model for genetic classification of breast cancer. *Anal. Quant. Cytol. Histol.* **31**, 101-108 (2009).

77. Roylance,R. *et al.* Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. *Cancer Res.* **59**, 1433-1436 (1999).

78. Natrajan,R. *et al.* Loss of 16q in high grade breast cancer is associated with estrogen receptor status: Evidence for progression in tumors with a luminal phenotype? *Genes Chromosomes. Cancer* **48**, 351-365 (2009).

79. Etzell,J.E. *et al.* Loss of chromosome 16q in lobular carcinoma in situ. *Hum. Pathol.* **32**, 292-296 (2001).

80. Hwang,E.S. *et al.* Patterns of chromosomal alterations in breast ductal carcinoma in situ. *Clin. Cancer Res.* **10**, 5160-5167 (2004).

81. Cleton-Jansen,A.M. *et al.* Different mechanisms of chromosome 16 loss of heterozygosity in well- versus poorly differentiated ductal breast cancer. *Genes Chromosomes. Cancer* **41**, 109-116 (2004).

82. Roylance,R. *et al.* A comprehensive study of chromosome 16q in invasive ductal and lobular breast carcinoma using array CGH. *Oncogene* **25**, 6544-6553 (2006).

83. Waldman,F.M. *et al.* Genomic alterations in tubular breast carcinomas. *Hum. Pathol.* **32**, 222-226 (2001).

84. Buerger,H. *et al.* Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. *J. Pathol.* **194**, 165-170 (2001).

85. Dutrillaux,B., Gerbault-Seureau,M., & Zafrani,B. Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes. *Cancer Genet. Cytogenet.* **49**, 203-217 (1990).

86. Gerbault-Seureau,M., Vielh,P., Zafrani,B., Salmon,R., & Dutrillaux,B. Cytogenetic study of twelve human near-diploid breast cancers with chromosomal changes. *Ann. Genet.* **30**, 138-145 (1987).

87. Hicks,J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16**, 1465-1479 (2006).

88. Flagiello,D. *et al.* Highly recurrent der(1;16)(q10;p10) and other 16q arm alterations in lobular breast cancer. *Genes Chromosomes. Cancer* **23**, 300-306 (1998).

89. Tsarouha,H. *et al.* Karyotypic evolution in breast carcinomas with i(1)(q10) and der(1;16)(q10;p10) as the primary chromosome abnormality. *Cancer Genet. Cytogenet.* **113**, 156-161 (1999).

90. Wenger,C.R. *et al.* DNA ploidy, S-phase, and steroid receptors in more than 127,000 breast cancer patients. *Breast Cancer Res. Treat.* **28**, 9-20 (1993).

91. Millot,C. & Dufer,J. Clinical applications of image cytometry to human tumour analysis. *Histol. Histopathol.* **15**, 1185-1200 (2000).

92. Kronenwett,U. *et al.* Improved grading of breast adenocarcinomas based on genomic instability. *Cancer Res.* **64**, 904-909 (2004).

93. Chavez-Uribe,E. *et al.* Hypoploidy defines patients with poor prognosis in breast cancer. *Oncol. Rep.* **17**, 1109-1114 (2007).

94. Tanner,M.M. *et al.* Genetic aberrations in hypodiploid breast cancer: frequent loss of chromosome 4 and amplification of cyclin D1 oncogene. *Am. J. Pathol.* **153**, 191-199 (1998).

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

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

97. Calza,S. *et al.* Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res.* **8**, R34 (2006).

98. Hu,Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC. Genomics* **7**, 96 (2006).

99. Langerod,A. *et al.* TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res.* **9**, R30 (2007).

100. Naume,B. *et al.* Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Mol. Oncol.* **1**, 160-171 (2007).

101. Carey,L.A. *et al.* Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* **295**, 2492-2502 (2006).

102. Laakso,M. *et al.* Basoluminal carcinoma: a new biologically and prognostically distinct entity between basal and luminal breast cancer. *Clin. Cancer Res.* **12**, 4185-4191 (2006).

103. Nielsen,T.O. *et al.* Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res.* **10**, 5367-5374 (2004).

104. Fadare,O. & Tavassoli,F.A. Clinical and pathologic aspects of basal-like breast cancers. *Nat. Clin. Pract. Oncol.* **5**, 149-159 (2008).

105. Rakha,E.A. *et al.* Triple-negative breast cancer: distinguishing between basal and nonbasal subtypes. *Clin. Cancer Res.* **15**, 2302-2310 (2009).

106. Vincent-Salomon,A. *et al.* Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. *Breast Cancer Res.* **9**, R24 (2007).

107. Weigelt,B. *et al.* Refinement of breast cancer classification by molecular characterization of histological special types. *J. Pathol.* **216**, 141-150 (2008).

108. Bergamaschi,A. *et al.* Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes. Cancer* **45**, 1033-1040 (2006).

109. Chin,S.F. *et al.* High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.* **8**, R215 (2007).

110. Chin,S.F. *et al.* Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*(2006).

111. Bergamaschi,A. *et al.* Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes. Cancer* **45**, 1033-1040 (2006).

112. Anderson,W.F., Pfeiffer,R.M., Dores,G.M., & Sherman,M.E. Comparison of age distribution patterns for different histopathologic types of breast carcinoma. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1899-1905 (2006).

113. Yang,X.R. *et al.* Differences in risk factors for breast cancer molecular subtypes in a population-based study. *Cancer Epidemiol. Biomarkers Prev.* **16**, 439-443 (2007).

114. Millikan,R.C. *et al.* Epidemiology of basal-like breast cancer. *Breast Cancer Res. Treat.* **109**, 123-139 (2008).

115. Boecker,W. *et al.* Ductal epithelial proliferations of the breast: a biological continuum? Comparative genomic hybridization and high-molecular-weight cytokeratin expression patterns. *J. Pathol.* **195**, 415-421 (2001).

116. Farabegoli,F. *et al.* Simultaneous chromosome 1q gain and 16q loss is associated with steroid receptor presence and low proliferation in breast carcinoma. *Mod. Pathol.* **17**, 449-455 (2004).

117. Karayiannakis,A.J. *et al.* Immunohistochemical detection of oestrogen receptors in ductal carcinoma in situ of the breast. *Eur. J. Surg. Oncol.* **22**, 578-582 (1996).

118. Gregory,S.G. *et al.* The DNA sequence and biological annotation of human chromosome 1. *Nature* **441**, 315-321 (2006).

119. Martin,J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**, 988-994 (2004).

120. Greenman,C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).

121. Hahn,Y. *et al.* Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc. Natl. Acad. Sci. U. S. A* **101**, 13257-13261 (2004).

122. McClintock,B. The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. *Proc. Natl. Acad. Sci. U. S. A* **25**, 405-416 (1939).

123. McClintock,B. The Stability of Broken Ends of Chromosomes in Zea Mays. *Genetics* **26**, 234-282 (1941).

124. Bautista,S. & Theillet,C. CCND1 and FGFR1 coamplification results in the colocalization of 11q13 and 8p12 sequences in breast tumor nuclei. *Genes Chromosomes. Cancer* **22**, 268-277 (1998).

125. Paterson,A.L. *et al.* Co-amplification of 8p12 and 11q13 in breast cancers is not the result of a single genomic event. *Genes Chromosomes. Cancer* **46**, 427-439 (2007).

126. Ruan,Y. *et al.* Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* **17**, 828-838 (2007).

127. Hampton,O.A. *et al.* A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*(2008).

128. Mitelman,F., Johansson,B., & Mertens,F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233-245 (2007).

129. Letessier,A. *et al.* Frequency, prognostic impact, and subtype association of 8p12, 8q24, 11q13, 12p13, 17q12, and 20q13 amplifications in breast cancers. *BMC. Cancer* **6**, 245 (2006).

130. Heim,S., Teixeira,M.R., Dietrich,C.U., & Pandis,N. Cytogenetic polyclonality in tumors of the breast. *Cancer Genet. Cytogenet.* **95**, 16-19 (1997).

131. Weedon-Fekjaer,H., Lindqvist,B.H., Vatten,L.J., Aalen,O.O., & Tretli,S. Breast cancer tumor growth estimated through mammography screening data. *Breast Cancer Res.* **10**, R41 (2008).

132. Gronbaek,K., Hother,C., & Jones,P.A. Epigenetic changes in cancer. *APMIS* **115**, 1039-1059 (2007).

133. Sharma,S., Kelly,T.K., & Jones,P.A. Epigenetics in Cancer. *Carcinogenesis*(2009).

134. Jackson,M. *et al.* Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Mol. Cell Biol.* **24**, 8862-8871 (2004).

135. Hoque,M.O. *et al.* Changes in CpG islands promoter methylation patterns during ductal breast carcinoma progression. *Cancer Epidemiol. Biomarkers Prev.* **18**, 2694-2700 (2009).

136. Dammann,R. *et al.* Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat. Genet.* **25**, 315-319 (2000).

137. Merlo,A. *et al.* 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.* **1**, 686-692 (1995).

138. Rice,J.C., Massey-Brown,K.S., & Futscher,B.W. Aberrant methylation of the BRCA1 CpG island promoter is associated with decreased BRCA1 mRNA in sporadic breast cancer cells. *Oncogene* **17**, 1807-1812 (1998).

139. Bloushtain-Qimron,N. *et al.* Cell type-specific DNA methylation patterns in the human breast. *Proc. Natl. Acad. Sci. U. S. A* **105**, 14076-14081 (2008).

140. Al Hajj,M., Wicha,M.S., Benito-Hernandez,A., Morrison,S.J., & Clarke,M.F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A* **100**, 3983-3988 (2003).

141. Shackleton,M., Quintana,E., Fearon,E.R., & Morrison,S.J. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell* **138**, 822-829 (2009).

142. Lim,E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907-913 (2009).

143. Prat,A. & Perou,C.M. Mammary development meets cancer genomics. *Nat. Med.* **15**, 842-844 (2009).

144. Stingl,J. & Caldas,C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat. Rev. Cancer* **7**, 791-799 (2007).

145. Wiedswang,G. *et al.* Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer. *J. Clin. Oncol.* **21**, 3469-3478 (2003).

146. Sabattini,E. *et al.* The EnVision++ system: a new immunohistochemical method for diagnostics and research. Critical comparison with the APAAP, ChemMate, CSA, LABC, and SABC techniques. *J. Clin. Pathol.* **51**, 506-511 (1998).

147. Vosse,B.A., Seelentag,W., Bachmann,A., Bosman,F.T., & Yan,P. Background staining of visualization systems in immunohistochemistry: comparison of the Avidin-Biotin Complex system and the EnVision+ system. *Appl. Immunohistochem. Mol. Morphol.* **15**, 103-107 (2007).

148. Wiedorn,K.H., Goldmann,T., Henne,C., Kuhl,H., & Vollmer,E. EnVision+, a new dextran polymer-based signal enhancement technique for in situ hybridization (ISH). *J. Histochem. Cytochem.* **49**, 1067-1071 (2001).

149. Jeffrey,S.S., Fero,M.J., Borresen-Dale,A.L., & Botstein,D. Expression array technology in the diagnosis and treatment of breast cancer. *Mol. Interv.* **2**, 101-109 (2002).

150. Auer,G., Caspersson TO, & Wallgren AS DNA content and survival in mammary carcinoma. *Anal. Quant. Cytol. Histol.* **2**, 161-165 (1980).

151. Pauletti,G. *et al.* Assessment of methods for tissue-based detection of the HER-2/neu alteration in human breast cancer: a direct comparison of fluorescence in situ hybridization and immunohistochemistry. *J. Clin. Oncol.* **18**, 3651-3664 (2000).

152. Kallioniemi,A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-821 (1992).

153. Lander,E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

154. Snijders,A.M. *et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* **29**, 263-264 (2001).

155. Gresham,D., Dunham,M.J., & Botstein,D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* **9**, 291-302 (2008).

156. Lucito,R. *et al.* Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* **13**, 2291-2305 (2003).

157. Feuk,L., Carson,A.R., & Scherer,S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85-97 (2006).

158. Barrett,M.T. *et al.* Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci. U. S. A* **101**, 17765-17770 (2004).

159. van den Ijssen *et al.* Human and mouse oligonucleotide-based array CGH. *Nucleic Acids Res.* **33**, e192 (2005).

160. Gunderson,K.L. *et al.* Decoding randomly ordered DNA arrays. *Genome Res.* **14**, 870-877 (2004).

161. Peiffer,D.A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136-1148 (2006).

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

163. Kamalakaran,S. *et al.* Methylation detection oligonucleotide microarray analysis: a high-resolution method for detection of CpG island methylation. *Nucleic Acids Res.* **37**, e89 (2009).

164. Campbell,P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722-729 (2008).

165. Kirkwood B & Sterne J *Essential Medical Statistics*(WileyBlackwell,2003).

166. Hennessy,B.T. *et al.* Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res.* **69**, 4116-4124 (2009).

I

This article is removed.

II

Genomic architecture characterizes tumor progression paths and fate in breast cancer patients

Hege G. Russnes[1,2,4], Hans Kristian Moen Vollan[1,4,5], Ole Christian Lingjærde[6,] Alexander Krasnitz[7], Pär Lundin[8], Bjørn Naume[3], Therese Sørlie[1], Elin Borgen[2], Inga H. Rye[2], Anita Langerød[1], Suet-Feung Chin[9], Andrew E. Teschendorff[9,10], Philip J. Stephens[12], Susanne Månér[8], Ellen Schlichting[5], Lars O. Baumbusch[1,4], Rolf Kåresen[5], Michael P. Stratton[12,13], Michael Wigler[7], Carlos Caldas[9,11], Anders Zetterberg[8], James Hicks[7], Anne-Lise Børresen-Dale[1,4].

[JH and A-L B-D are shared last]

[1]Department of Genetics, Institute for Cancer Research, [2]Division of Pathology and [3]Department of Oncology, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway.
[4]Faculty Division The Norwegian Radium Hospital, Faculty of Medicine, University of Oslo
[5]Department of Breast and Endocrine Surgery, Ullevål Hospital, Oslo University Hospital, 0450 Oslo, Norway.
[6]Biomedical Research Group, Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, 0316 Oslo, Norway.
[7]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.
[8]Department of Oncology–Pathology, Karolinska Institutet, Cancer Center Karolinska, SE-171 76 Stockholm, Sweden.
[9]Breast Cancer Functional Genomics, Cancer Research UK Cambridge Research Institute and Department of Oncology, University of Cambridge, Li Ka-Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.
[10]UCL Cancer Institute, University College London, WC1E 6BT, UK.
[11]Cambridge Breast Unit, Addenbrookes Hospital and Cambridge NIHR Biomedical Research Centre, Cambridge University Hospitals NHS Foundation Trust, Hills Road, Cambridge, UK.
[12]Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
[13]Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.

Correspondence should be addressed to: Anne-Lise Børresen-Dale (a.l.borresen-dale@medisin.uio.no)

Abstract

Distinct molecular subtypes of breast carcinomas have been identified, but translation into clinical use has been limited. We have developed two platform independent algorithms to explore genomic architectural distortion using aCGH data to measure 1) whole arm gains and losses (WAAI) and 2) complex rearrangements (CAAI). By applying CAAI and WAAI to data from 595 breast cancer patients we were able to separate the cases into eight subgroups with different distribution of genomic distortion. Within each subgroup data from expression analyses, sequencing and ploidy indicated that progression occurs along separate paths into more complex genotypes. Histological grade had prognostic impact only in the Luminal related groups while the complexity identified by CAAI had an overall independent prognostic power. This study emphasizes the relationship between structural genomic alterations, molecular subtype and clinical behavior, and provides a score of genomic complexity as a new tool for prognostication in breast cancer.

Breast cancer is a heterogeneous disease as reflected by histopathology, molecular alterations and clinical behavior. In order to relate cellular and sub-cellular features to clinical parameters and outcome, substantial effort has been exerted towards identifying tumor groups with distinct molecular features. Estrogen receptor (ER) status was early shown to be a major discriminating factor and is still of clinical importance[1]. The more recent gene expression based classification proposed by Perou et al. in 2000[2] identified five different subgroups where one was Luminal-cell related (Luminal A) and another were myoepithelial-cell related group (Basal-like). Three additional groups were identified, but these are less characterized (erbB2+, Luminal B and Normal-like). Basal-like and Luminal A carcinomas have different etiologies and for most purposes may be considered as distinct diseases[3-6]. This is also reflected in the genomic portraits defined by aCGH (array Comparative Genomic Hybridization), and it seems evident that the history of molecular subgroups is written in the DNA alterations[7-9].

Despite the power of RNA and DNA based profiling, translating complex molecular classifications into clinical practice has proven challenging. Clinical cohorts are often selected to have tumors of a certain category, and might not include all subtypes or outcome groups. The size of sample sets available for microarray studies has so far been limited, and combining sets to increase size has been challenging since various types of array platforms have been used.

Array CGH does not reveal the chromosomal pattern associated with copy number alterations; however much can be inferred from cytogenetic studies. The genomic architectural changes in breast tumors revealed by karyotyping follow some main traits. One type of events seen early in tumor progression is loss or gain of whole chromosome arms[10]. Another type is more complex rearrangements, often involving several different chromosomes with inversions, deletions and amplifications[10]. Previously we found that invasive breast tumors had different patterns of aCGH aberrations[11]. Tumors of the simplex type had few alterations with loss or gain of whole arms dominating, while tumors of the complex type had either many chromosomes altered with multiple regions with low level loss and gain (sawtooth pattern) or had a few selected regions with high copy number gains with intermittent losses (firestorms). We hypothesized that distinct molecular mechanisms underlie such patterns of aberrations.

In this paper, we have developed objective estimates of genome-wide architectural distortion. For each chromosome arm, two platform independent scores were defined: one measures the deviation from normal copy number (Whole Arm Aberration Index; WAAI) and the other the degree of local distortion (Complex Arm Aberration Index; CAAI). The clinical impact of WAAI and CAAI was studied using aCGH data from 595 breast carcinomas belonging to four clinical cohorts profiled by three different aCGH platforms (30K-244K resolution). This revealed patterns of genomic architectural distortion recognizing Luminal and Basal related tumors with distinct subgroups and outcome. The study illustrates the importance of dividing breast cancer into molecularly defined subgroups as they have independent progression paths and clinical outcome.

Results

Genomic architecture characterized by CAAI and WAAI

Two novel algorithms were constructed; one to identify complex architectural distortions characterized by physically tight clusters of break points with large changes of amplitude, and another to recognize gains and loss of whole chromosome arms (CAAI: Complex Arm Aberration Index and WAAI: Whole Arm Aberration Index, respectively). Segmented data from one tumor with corresponding CAAI values are illustrated for selected chromosome arms in Figure 1a. The circos plot from Paired End Sequencing of the same sample (Fig. 1b) shows that CAAI recognizes regions with structural complexity (Stephens et al., resubmitted). Areas of complex rearrangements were found by selecting chromosome arms with CAAI ≥ 0.5. Comparison in one cohort of HER2 copy number gains estimated by FISH and the CAAI score showed that all but one sample with high CAAI had more than four copies of HER2 (Supplementary Fig. 1).

For most chromosome arms, the distribution of WAAI is approximately symmetric around zero (Supplemental Figure 2). For some arms however, WAAI is skewed towards positive values (1q, 8q and 16p) and for others towards negative values (16q and 17q), reflecting a bias towards gain or loss. This pattern was seen in all cohorts, independent of platform. Arms with WAAI ≥ 0.8 were defined as whole arm gains and arms with WAAI ≤ -0.8 as whole arm losses. An example of a tumor with whole arm gain of 1q and whole arm loss of 16q is shown in Supplementary Figure 3a. FISH analyses of this case identified a combination of probes indicating a centromere-close translocation t(1q;16p) (Supplementary Fig. 3b).

Demographic data for the four cohorts are presented in Supplementary Table 1 and overall aberration frequencies are found in Supplementary Fig. 4. The four cohorts were merged for the analysis of association to clinico-pathological information, and the frequency plot in Figure 2 shows an aberration pattern typical for breast cancer. Several of the most frequent events such as gain of 1q and loss of 16q/17q are whole arm events, while the majority of gains on 17q and losses on 11q have CAAI ≥ 0.5 and are likely caused by complex rearrangements (Fig. 2b). A few alterations such as gain on 8q and 20q displayed both whole arm gain and high CAAI.

Defining subgroups based on genomic architecture

Several studies have shown that the number of genomic alterations and the regions preferentially altered differ between the molecular expression subtypes[7, 8, 12, 13]. Luminal A/ER positive tumors often have few alterations with gain of 1q and loss of 16q dominating[7, 8, 12-14] while Basal-like have many alterations affecting most of the chromosomes. Loss on 5q and gain on 10p have been proposed as specific Basal-like alterations[7, 8, 12, 15], similar to findings in breast carcinomas from BRCA1 carriers[16, 17]. Based on this, we distinguish between four "WAAI groups" of tumors: those with whole arm gain of 1q and/or loss of 16q (group), those with regional loss on 5q and/or gain on 10p (group), those with both (group), and those with neither (group) (see M&M). To further characterize these groups we split each into two "CAAI subgroups" depending on the level of complex rearrangement: those with CAAI < 0.5 for all arms (, , A ,) and those with CAAI ≥ 0.5 for at least one arm (, , ,). The group distribution was similar for all four cohorts, except for the WZ which had more samples of type and less samples with elevated CAAI, most likely due to selection

of diploid tumors (Supplementary Table 2)[11]. The sample size of the eight groups and the arm-wise distribution of WAAI and CAAI for all 595 samples are shown in Figure 3a.

Patterns of genomic architecture in the WAAI/CAAI groups

WAAI and CAAI revealed different chromosomal event distributions in the eight subgroups (Fig. 3a). This is also reflected in the frequency plots of individual subgroups (Supplementary Fig. 5). The subgroups displayed pronounced differences with respect to the number of whole chromosome arm loss or gain events (Fig. 3a and Supplementary Fig. 6). For each of the four WAAI groups, the tumors with complex rearrangements (i.e. , , and) had more whole arms affected, mostly by gains (WAAI ≥ 0.8), than the corresponding group without complex rearrangements.

Tumors of type were frequently ER positive, of low or intermediate grade, diploid and included a majority of the invasive lobular carcinomas (Supplementary Table 3). Group was the only group with frequent alterations of whole chromosomes; particularly prominent were gain of 5, 7, 8 and 20 and loss of 18 (Fig. 3a), in line with previous cytogenetic findings[18, 19]. Supplementary Fig. 7 illustrates that and tumors had the same distributions of altered arms, and the increased number of gains seen in tumors were mainly affecting 8q, 16p, 20p and 20q. In tumors of type , complex rearrangements were most frequent on 11q and 8p, followed by 17q and 8q (Supplementary Fig. 8). The high level amplifications on 8p and 11q includes genes of interest such as and loci known to be frequently amplified in ER positive breast carcinomas[20-22].

Tumors of type were more frequently of high grade, aneuploid and mutated than tumors of type (Supplementary Table 3). Tumors of type were dominated by whole arm losses, most frequently of 17p, 4p, 4q and 5q, while tumors of type had complex alterations often affecting many arms, most frequently 17q, followed by 8p and 20q (Fig. 3a and Supplementary Figs. 6, 7 and 8). The overall frequencies of aberrations were quite similar in B1 and B2 (Supplementary Fig. 5).

tumors had elements of both and tumors, were dominated by aneuploid tumors of intermediate or high grade, and had the highest frequency of whole arm alterations (both gains and losses) (Supplementary Table 3, Supplementary Fig. 6 and 7). The tumors with complex rearrangements had a heterogeneous distribution pattern of arms with high CAAI (Fig. 3a, Supplementary Fig. 8).

tumors had the fewest numbers of whole arm alterations with 8q and 16p gain and 17p and 22 loss as the most frequent (Fig. 3a and Supplementary Fig. 6 and 7). This was seen both in and carcinomas, with 17p being more frequently lost in than in . High CAAI was frequent on 17q but rare on 11q (Fig. 3a and Supplementary Fig. 8). The clinico-pathological parameters had similarities with the A group, but with fewer ER positive and more mutated tumors (Supplementary Table 3). Interestingly almost half of all tumors with histological grade 1 and most carcinomas of a special histological type such as lobular, tubulolobular and mucinous were grouped as

Paired-end sequencing was performed on a few selected samples (Stephens et al., resubmitted and Fig. 3b). The analyzed tumor showed a single rearrangement, in contrast to the tumor which had a larger number of complex inter- and intra-chromosomal rearrangements, in

line with the high CAAI score. The 1q/16q translocation in the tumor is missed as the paired-end sequencing method does not detect alterations involving centromere-close heterochromatin. The tumor showed numerous smaller structural rearrangements ("mutator phenotype") in contrast to the pattern seen in the and tumors. The tumor showed a mutator phenotype pattern, but with more inter-chromosomal rearrangements than the tumor. The tumor had some segmental duplications/inversions in addition to complex rearrangements involving chromosome arm 17q.

For the 298 tumors with available gene expression data, the correlation to the five intrinsic subtype centroids was calculated[5]. Both and tumors showed strong correlation to the Luminal A subtype (Fig. 3c). Luminal B tumors were more frequent in the group, indicating that tumors represent more advanced tumors with high proliferation and increased growth factor signaling than [23] (Supplementary Table 4). This was also supported by ploidy data as the group had a higher fraction of aneuploid tumors (Supplementary Fig. 9). The tumors were dominated by the Basal-like subtype. The subtype correlation patterns of and / were quite similar, dominated by negative correlation to the Luminal A subtype, and overall had a closer resemblance to than to . A majority of erbB2+ and Normal-like tumors were classified as tumors. Normal-like tumors are rare and often omitted from breast cancer expression classification studies, but Normal-like cell lines have shown an enrichment in stem-cell related features[24]. Almost 30% of all Basal-like tumors were classified as tumors, in line with a previous study identifying a subgroup of Basal-like having low genomic instability[13].

WAAI and CAAI groups as prognostic markers

DCIS patients and the WZ cohort were omitted from survival and risk analyses to avoid bias as they were highly selected, leaving 451 cases. Both WAAI and CAAI classification identified subgroups with significant difference in breast cancer related death ($p=0.009$ and $p<0.001$ respectively; see Fig. 4a and b). Bivariate Cox regression analysis showed that CAAI classification had predictive power independently of age, lymph node status, tumor size, histological grade, ER status, mutation status, vascular invasion, intrinsic subtype and adjuvant treatment (Supplementary File 1). Furthermore, the increased risk of breast cancer specific death in patients with high CAAI was independent of known risk factors (multivariate Cox analysis; HR:1.92, 95% CI [1.33-2.78], $p<0.001$) (Table 1a).

For the WAAI classification, patients with tumors had an almost twofold risk of death from breast cancer compared to patients with tumors (Supplementary File 1). Bivariate Cox analysis showed that this was independent of age, tumor size, lymph node status, vascular invasion and adjuvant treatment (Supplementary File 1). High histological grade, large tumor size and positive node status indicated increased risk for breast cancer specific death for patients with tumors, as opposed to tumor patients (Supplementary File 1). Interestingly, histological grade was non-informative for patients with and tumors but of high importance for patients belonging to the and groups ($p=0.58$, $p=0.68$, $p=0.02$ and $p=0.03$; Figure 4e-h). mutation status and high CAAI were the only factors having prognostic value (though borderline) in patients with tumors, while histological grade, tumor size, lymph node status all were of importance in the group patients (Supplementary File 1).

The classification obtained by combining WAAI and CAAI also revealed distinct patterns of clinical behavior; the worst clinical outcome was seen in the / groups with a 2.6 fold increase in breast cancer death risk compared to the groups without high CAAI ($p<0.001$) (Fig. 4c and Table 1b). The same trend in survival was seen for patients with lymph node negative disease (Fig. 4d).

Discussion

Genome-wide, high resolution analyses of both DNA and RNA have brought novel insights into breast carcinoma classification[8, 13, 25], but conclusions have been limited by small samples sizes. By developing platform independent algorithms, we could merge aCGH data from several clinical cohorts and perform DNA based grouping of breast carcinomas, utilizing previous DNA and RNA classifications. Defining surrogate markers for Luminal and Basal-like breast cancer, we observed several distinct patterns of aberrant genomic architecture. Tumors of type are dominated by ER positive, Luminal A tumors with large WAAI magnitude (both gains and losses), and by concomitant 1q gain and 16q loss caused by unbalanced centromere-close translocations between the two chromosomes[26]. The same mechanism affecting other arms might explain the frequent losses and gains of whole chromosome arms in group Several studies have indicated that Luminal tumors have a distinct progression path[27-30]. This is reflected in our study by tumors having more arms with high WAAI magnitude, being more frequently aneuploid, of high grade and with worse outcome than tumors (Fig. 3a). Amplification is found to precede aneuploidization in breast cancer cell lines[31], and our study indicates that the same switch also occurs in vivo. Progression from to seems to induce a shift in gene expression pattern with increased correlation to the Luminal B centroid and worse outcome (Figs. 3c and 4c).

The tumors had a completely different and more heterogeneous genomic pattern. Group tumors were dominated by losses, and the single case investigated by paired-end sequencing had in addition the typical mutator phenotype pattern reflecting multiple segmental duplications. In two separate studies we have found that a subgroup of Basal-like tumors are characterized by losses and progress from hypodiploid to aneuploid, often with complex rearrangements (Navin N. et al., in press Genome Research, van Loo P. et al., submitted), in line with the group being dominated by losses. Both and some tumors had an expression pattern pointing towards a Basal-like relationship (Fig. 3c), In addition, both and some tumors had the highest genomic distortion, were often aneuploid and had short survival, and we hypothesize that , and some cases reflect more advanced Basal related tumors. Interestingly, the ER status cannot be used as a surrogate marker for these groups as a large number is ER positive.

We find that and tumors are different both at the genomic, transcriptomic and clinical level. It has been shown that amplifications on 8p/11q and 8q/17q occurs preferentially in two phenotypically diverse groups of breast cancer[32], consistent with the different CAAI distribution in and tumors. In a study using high resolution methylation arrays on one of the cohorts, we found patterns of methylation in tumors pointing towards CD24+/luminal cell relationship and likewise a connection between tumors and CD44+/progenitor cell methylation patterns (Kamalakaran et al., manuscript). There are several indicators that molecular subgroups of breast cancer reflect transformation of different breast epithelial cell progenitors[33-35]. Our study indicates that molecular subgroups can be recognized by differences

in genomic architecture. This is probably reflecting underlying subgroup-specific defects linked to different cell of origin. As illustrated in Figure 5, we hypothesize that the genomic architectural pattern reflects tumor subgroups related to different cell of origin. Tumors of type originate from Luminal-committed progenitors and are prone to whole arm translocations. They have a linear progression path with complex rearrangements with more arms affected. Tumors of type , and have a much more complex progression path, possibly originating from less differentiated progenitors. Basal-like carcinomas are composed of several subtypes[36-38], and recent work indicates that a Luminal progenitor on a background of deficiency may be the cell of origin of such Basal-like tumors[39]. We suggest that the heterogeneity seen in groups , and with respect to the distribution of WAAI and CAAI, indicates that tumors of these types descend from different but related early progenitors, and that alternative combination of repair defects defines several progression paths as illustrated in Figure 5.

Complex rearrangements as defined by CAAI occurred in all subgroups, and CAAI had a strong prognostic impact independent of other factors, even if it only occurred on one chromosome arm. The mechanisms behind complex rearrangements are not completely understood, but one type is breakage-fusion-bridge cycles due to double strand repair defects[40, 41] resulting in high level amplicons with intermittent deletions. As high level amplicons are seen even in DCIS[42] and in diploid tumors[11], this opens the possibility for a distinct subtype of carcinomas having complex alterations at an early stage of progression ("de novo complexity"). As illustrated in Figure 5, we speculate that the group might have a subset of tumors with a non- , non- relationship.

The present study indicates that the type of architectural distortion is of major importance in determining the tumor phenotype and can be used to group tumors into Luminal and Basal-related tumors. This is of major importance, since the value of established prognostic markers is subgroup dependent. We also find that even in biological distinct subtypes of breast cancer, the addition of complex rearrangements seem to be of major importance for patient outcome. A strong hierarchical relationship between subtypes of breast carcinomas is yet to be defined, but our findings provide a background for further functional studies aiming to elucidate the relationship between genomic architecture, phenotypic traits and the cell of origin in breast cancer. Our study demonstrates that the patterns of genomic architecture described here constitute a new prognostic tool in breast cancer.

Acknowledgements

This work was supported by grants to; ALBD: The Norwegian Research council, grant no. 155218/V40, 175240/S10 and The Norwegian Cancer Society, grant D99061; CC: Cancer Research UK; MW: National Institute of Health and Dept. of the Army; AZ: the Swedish Cancer Society and the Swedish Research Council. HGR has received grants from Radiumhospitalets Legater for travel and lab assistance. The authors would like to thank Eldri U. Due for technical assistance.

Authors contribution

HGR, JH, ALBD conceived and designed the study with valuable input from HKMV and OCL. HGR, HKMV, OCL and ALBD developed CAAI, WAAI and the WAAI/CAAI classification. OCL developed Java software for PCF and centering, with valuable input from HGR and HKMV. HGR performed statistical analyses with valuable input from HKMV, OCL, ALBD, CC and BN. HKMV, JH, AK, MW, LOB, AET, SFC, CC planned and performed experiments and/or contributed aCGH data. BN, RK, ES, AL, EB and ALBD collected samples and clinical data. EB and HGR contributed with pathology data. AZ, HGR, IHR, PL and SM planned and performed FISH experiments with input from ALBD and JH. HGR and AZ planned and performed ploidy experiments with input from ALBD and JH. MPS and PJH contributed with paired end sequencing data. AL, TS, SFC, ALBD, AET and CC contributed with gene expression data. AL and TS contributed with TP53 sequencing data. HGR, HKMV, OCL, ALBD, JH, AZ, CC, BN and LOB provided valuable discussion. HGR, HKMV, OCL and ALBD wrote the paper.

Competing interests
The authors declare that they have no competing interests.

Figure legends

Figure 1: CAAI pattern compared to structural rearrangements identified by paired-end sequencing
a: Raw (dots) and segmented (line) data for chromosome arms 7p and 8p and chr.15 from sample 595. Red segments correspond to the 20 Mb window with highest CAAI; the corresponding CAAI was 7.04, 1.04 and 4.74 respectively. Chromosome arms 7p had an additional region with elevated CAAI, but as this score was lower than 7.04 it was neglected.
b: Structural sequence alterations identified by genome wide paired-end sequencing for the same sample. Outer circle show the cytobands for each chromosome, followed by a plot indicating the copy number variation. The green bars in the centre refer to smaller intra-chromosomal changes such as duplications and inversions while pink lines indicate inter-chromosomal translocations. In this sample 13 chromosome arms had CAAI>0, six of these had CAAI≥0.5, these are in bold and marked with *. The two regions with most rearrangements showed the highest CAAI (chromosome arm 7p and chr.15). Areas with few rearrangements had low or zero CAAI.

Figure 2: Genome wide distribution of genomic loss and gain compared to frequencies of WAAI and CAAI in 595 breast carcinomas
a: Frequency plot illustrating the percentage of samples with gain and loss genome wide (red: gain, green: loss).
b: The frequency of samples scored with whole arm changes identified by WAAI and complex rearrangements scored by CAAI are shown in the heatmap. The color indicates the percentage arms with WAAI over and under the chosen threshold and the percentage of arms with CAAI

higher than the threshold for each chromosome arm with: WAAI≥0.8 (red, top row), WAAI≤-0.8 (green, middle row) and CAAI≥0.5 (blue, bottom row).

Figure 3: Genome wide distribution of WAAI and CAAI for all samples sorted into WAAI and CAAI groups, examples of identified structural aberrations and corresponding gene expression patterns.
a: The heat map illustrate the WAAI and CAAI score for all 595 samples sorted into , , and tumors and thereafter into groups of tumors with and without high CAAI on one chromosome arm or more. Each row in the heatmap corresponds to one sample, and each column to a chromosome arm (from 1p to 22). The left panel indicate WAAI alterations for each chromosome arm (red: WAAI≥0.8, green; WAAI≤-0.8, black: 0.8>WAAI<-0.8). The right panel indicate the corresponding CAAI score for each chromosome arm for the same samples (no rearrangements=white. The CAAI scale is indicated below the figure).
b: Structural sequence alterations identified by genome wide paired-end sequencing for selected samples from the various WAAI groups. Outer circle show the cytobands for each chromosome, followed by the copy number variation. The green bars in the center indicate smaller intra chromosomal changes while pink lines indicate inter chromosomal translocations. The lines indicate the position of the selected samples in the WAAI/CAAI groups.
c: Correlation to each of the five intrinsic subtypes for a total of 185 cases sorted into WAAI/CAAI groups.

Figure 4: WAAI and CAAI groups and breast cancer specific survival in the merged clinical dataset (n=454 cases)
The Kaplan Meier plots illustrate that breast cancer patients with tumors with high or low CAAI (a) had significant difference in survival ($p<0.001$). A difference was also found between patients with , , and tumors (the WAAI groups) (b) and between patients subdivided into the combined WAAI/CAAI groups (c). The Kaplan Meier curves showed that and had the worst survival; in a multivariate Cox regression model these patients had an increased hazard of 2.6 of dying from breast cancer compared to the , , and patients with a 95% CI: [1.66-4.16] and <0.001 (Table 1b).
Patients with lymph node negative disease (n=231) showed the same trend in survival for the different WAAI/CAAI subclasses, with and having a worse prognosis and the and having better compared to the whole cohort ($p=0.057$).
In e-f, the different impact of histological grade is illustrated. Patients with an or tumor were stratified into good, intermediate and bad prognosis by histological grade ($p=0.02$ and $p=0.03$) in contrast to patients with and tumors where we could not show any difference in breast cancer specific survival according to histological grade.

Figure 5: A hypothetical relationship between observed patterns of genomic architecture, expression subtype and cell of origin in breast carcinomas
We hypothesize here that a luminal developmental pathway originates from a dedicated luminal progenitor cell. Tumors of the type have 'simplex' aCGH profiles with whole chromosome or chromosome arm rearrangements dominating. In an early phase a Normal-like or Luminal A expression pattern dominates. This in contrast to tumors that are more advanced with increased numbers of chromosome arms affected in addition to complex rearrangements in preferential regions such as 8p and 11q. These tumors have frequent

expression correlation to Luminal B in addition to Luminal A but rarely to the erbB2+ centroid. The simplicity of tumors are illustrated by the circos plot with only one structural rearrangement and the histology illustrating the frequent finding of high luminal differentiation in this group. In the group the circos plot show more inter- and intra chromosomal rearrangements, and the histology show the more frequent low differentiation pattern. Likewise we observe that tumors are different from / tumors and hypothesize that they originate from a less dedicated or a myoepithelial progenitor. They are dominated by 'sawtooth pattern' and genomic losses in an early phase and with a high correlation to the Basal-like expression subtype. Related groups are , and representing tumors with numerous aberrations genome wide including complex rearrangements such as firestorms. This is supported by the circos plots from a , and tumor, all having segmental duplications genome wide. The histology rarely showed any luminal differentiation and had a solid growth pattern with and without lymphoid infiltration. These tumors have correlation towards the erbB2+ and Luminal B expression centroids in addition to the Basal-like. The , , and tumors might represent different stages in a non-linear progression. We also speculate that a subgroup of represent tumors with complexity present already in an early phase ('de novo complexity').

Table 1: Multivariate Cox regression analysis, breast cancer specific death

a)

Variable	Multivariate Cox regression p value	HR	95% CI Lower	Upper
n=398				
CAAI (high vs. low)	0.001	1.92	1.31	2.81
Lymph node status (pos. vs. neg.)	0.002	1.81	1.24	2.63
Tumor size				
pT2 (vs. pT1)	0.055	1.47	0.99	2.17
pT3 and pT4 (vs. pT1)	<0.001	3.08	1.70	5.60
Histological grade				
Grade 2 (vs. Grade 1)	0.100	1.95	0.88	4.34
Grade 3 (vs. Grade 1)	0.007	2.98	1.34	6.63

ER status and WAAI classes were also in the model but did not reach statistical significance.

b)

Variable	Multivariate Cox regression p value	HR	95% CI Lower	Upper
n=398				
aCGH/CAAI grouped into three:				
A2, C2 (vs. A1, B1, AB1, C1)	0.033	1.59	1.04	2.44
B2, AB2 (vs. A1, B1, AB1, C1)	<0.001	2.63	1.66	4.16
Lymph node status (pos. vs. neg.)	0.003	1.79	1.23	2.61
Tumor size				
pT2 (vs. pT1)	0.122	1.37	0.92	2.04
pT3 and pT4 (vs. pT1)	<0.001	3.02	1.66	5.48
Histological grade				
Grade 2 (vs. Grade 1)	0.105	1.94	0.87	4.31
Grade 3 (vs. Grade 1)	0.010	2.88	1.29	6.43

ER status was also in the model but did not reach statistical significance.

Methods

Patient samples and gene expression data

Two cohorts from Norway (MicMa and Ull), one from Sweden (WZ) and one from England (ChinUC) were included in this study and the clinical and pathological descriptions are available in Supplemental Table 1. Gene expression data, ploidy, sequencing and clinical data are previously published [13, 43, 44](ploidy: van Loo P. et al., submitted, sequencing: Stephens et al., resubmitted).

The ethical boards of all institutions involved for the different cohorts have approved the study.

aCGH platforms and preprocessing of raw copy number data

DNA from the MicMa cohort were hybridized to the ROMA (Representational Oligonucleotide Microarray Analysis) 85k microarray, developed at Cold Spring Harbor Laboratory[45]. The method is based on oligonucleotide probes designed after the restriction fragments from digestion with . The platform is manufactured by NimbleGen, and the experiments followed the ROMA/NimbleGen protocol as previously described[11]. Probe intensities were read with the GenePix Pro 4.0 software and used for ratio calculation. The data from both the MicMa and WZ cohort were normalized using an intensity-based lowess curve fitting algorithm. The aCGH data from WZ is also published[11] and accessible from http://roma.cshl.edu.

DNA from the Ull samples was analyzed using 244k CGH microarrays (Hu-244A, Agilent technologies, Santa Clara, California, USA). This platform contains over 236.000 mapped in-situ synthesized oligonucleotide probes representing coding and non-coding sequences of the genome[46]. The standard Agilent protocol was used, without pre-labeling amplification of input genomic DNA. Scanned microarray images were read and analyzed with Feature Extraction v9.5 (Agilent Technologies), using protocols (CGH-v4_95_Feb07 and CGH-v4 91 2) for aCGH-preprocessing which included linear normalization.

DNA from the Caldas cohort were as previously described[13] analyzed with a customized oligonucleotide microarray containing 30k 60-mer oligonucleotide probes representing 27800 mapped sequences of the human genome[47]. Signal intensities and fluorescent ratios were obtained with BlueFuse version 3.2 (Bluegnome). Raw data were preprocessed using the R[48] with the bioconductor package limma[49].

The raw data and preprocessed data can be accessed from NCBI's GEO (http://www/ncbi.nlm.nih.gov/geo/) with accession number GSE8757 (ChinUC), GSE…. (Ull), GSE…. (MicMa) and GSE…. (WZ).

Statistical methods and analytical tools

We fit for each sample a piecewise constant regression function to the log-transformed aCGH data, using the PCF algorithm (9,43). For each probe a fitted value ("PCF-value") is thus obtained. The user controls the sensitivity of the method (via a "penalty parameter" gamma) and the least allowed number of probes in a segment (kmin). In our case, segmentation was to be performed on data from three different platforms with relative probe densities (average number of probes per unit distance) 0.12 (ChinUC), 0.34 (MicMa/WZ) and 1.00 (244k Ull). As we aimed to pool all the segmented aCGH profiles, we scaled the parameters gamma and kmin to obtain roughly equal segmentation resolutions in the three platforms (thus essentially favoring variance reduction over bias reduction in the estimated copy number profiles for

increasing probe densities). The chosen values for (gamma,kmin) were (100, 20) for Ull, (34, 7) for MicMa/WZ and (12, 3) for ChinUC and are consistent with this.

To center the segmented data, we find the density of the PCF-values using a kernel smoother with an Epanechnikov kernel and a window size of 0.03. Consider the three tallest peaks P_1 , P_2 , P_3 in the density, in decreasing order of height (if there are less than three peaks, we replicate the highest one to obtain three peaks). For each, we find the location and relative height (i.e. the absolute height of the peak divided by the sum of the heights of the three highest peaks). Select among P_1, P_2 the peak P with location closest to the median of the PCF-values. If the relative height of P is at least 0.2, then the PCF-values are centered by subtracting the location of P; otherwise, the PCF-values are centered by subtracting the location of the tallest of all the three peaks.

WAAI is found separately for each arm and sample. Define normalized PCF (NPCF) values as centered PCF-values divided by the residual standard deviation. Average NPCF over all probes on the arm to obtain s. If s>0, WAAI is the 5% quantile of NPCF; if s≤0, WAAI is the 95% quantile of NPCF (in practice constrained to a predefined grid). Arms with WAAI≥0.8 are called as whole-arm gains, and arms with WAAI≤-0.8 are called as whole arm losses. See Supplemental Figure 3 for an example.

CAAI is found separately for each arm and sample. For each break point found by PCF, we calculate three scores P, Q and W reflecting the proximity to neighboring break points, the magnitude of change and a weight of importance:

$$P = \tanh\left(\frac{\alpha}{L_1 + L_2}\right), \qquad Q = \tanh(|H_2 - H_1|), \qquad W = \frac{1}{2}\cdot\left[1 + \frac{\tanh\left(10\left(P - \frac{1}{2}\right)\right)}{\tanh(5)}\right]$$

where α is a constant, L_1, L_2 are the number of probes and H_1, H_2 the PCF-values for the segments joined at the break point. For any genomic subregion R we may define

$$S_R = \sum W \cdot \min(P, Q),$$

summing over all break points in R. Define CAAI as the maximal value of S_R across all subregions R of a predefined size (in this paper: 20 Mb).

The software used in this paper is partially written in Java and partially in Matlab, and is available at http://www.ifi.uio.no/bioinf/Projects/GenomeArchitecture. For statistical analysis SPSS 15.0 was used. The clinical data and WAAI and CAAI estimates are available in Supplementary File 2.

Reference List

1. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. **365**, 1687-1717 (2005).

2. Perou,C.M. Molecular portraits of human breast tumours. **406**, 747-752 (2000).

3. Carey,L.A. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. **295**, 2492-2502 (2006).

4. Millikan,R.C. Epidemiology of basal-like breast cancer. **109**, 123-139 (2008).

5. Sorlie,T. Repeated observation of breast tumor subtypes in independent gene expression data sets. **100**, 8418-8423 (2003).

6. Dalgin,G.S. Portraits of breast cancer progression. **8**, 291 (2007).

7. Bergamaschi,A. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. **45**, 1033-1040 (2006).

8. Chin,K. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. **10**, 529-541 (2006).

9. Chin,S.F. Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. (2006).

10. Dutrillaux,B., Gerbault-Seureau,M., & Zafrani,B. Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes. **49**, 203-217 (1990).

11. Hicks,J. Novel patterns of genome rearrangement and their association with survival in breast cancer. **16**, 1465-1479 (2006).

12. Adelaide,J. Integrated profiling of basal and luminal breast cancers. **67**, 11565-11575 (2007).

13. Chin,S.F. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. **8**, R215 (2007).

14. Farabegoli,F. Simultaneous chromosome 1q gain and 16q loss is associated with steroid receptor presence and low proliferation in breast carcinoma. **17**, 449-455 (2004).

15. Vincent-Salomon,A. Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. **9**, R24 (2007).

16. Johannsdottir,H.K. Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors. **119**, 1052-1060 (2006).

17. Tirkkonen,M. Distinct somatic genetic changes associated with tumor progression in carriers of BRCA1 and BRCA2 germ-line mutations. **57**, 1222-1227 (1997).

18. Molist,R., Remvikos,Y., Dutrillaux,B., & Muleris,M. Characterization of a new cytogenetic subtype of ductal breast carcinomas. **23**, 5986-5993 (2004).

19. Teixeira,M.R., Pandis,N., & Heim,S. Cytogenetic clues to breast carcinogenesis. **33**, 1-16 (2002).

20. Letessier,A. Frequency, prognostic impact, and subtype association of 8p12, 8q24, 11q13, 12p13, 17q12, and 20q13 amplifications in breast cancers. **6**, 245 (2006).

21. Paterson,A.L. Co-amplification of 8p12 and 11q13 in breast cancers is not the result of a single genomic event. **46**, 427-439 (2007).

22. Reis-Filho,J.S. Cyclin D1 protein overexpression and CCND1 amplification in breast carcinomas: an immunohistochemical and chromogenic in situ hybridisation analysis. **19**, 999-1009 (2006).

23. Loi,S. Gene expression profiling identifies activated growth factor signaling in poor prognosis (Luminal-B) estrogen receptor positive breast cancer. **2**, 37 (2009).

24. Sieuwerts,A.M. Anti-epithelial cell adhesion molecule antibodies and the detection of circulating normal-like breast tumor cells. **101**, 61-66 (2009).

25. Sorlie,T. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. **98**, 10869-10874 (2001).

26. Tsarouha,H. Karyotypic evolution in breast carcinomas with i(1)(q10) and der(1;16)(q10;p10) as the primary chromosome abnormality. **113**, 156-161 (1999).

27. Buerger,H. Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. **194**, 165-170 (2001).

28. Korsching,E. Deciphering a subgroup of breast carcinomas with putative progression of grade during carcinogenesis revealed by comparative genomic hybridisation (CGH) and immunohistochemistry. **90**, 1422-1428 (2004).

29. Abdel-Fatah,T.M. Morphologic and molecular evolutionary pathways of low nuclear grade invasive breast cancers and their putative precursor lesions: further evidence to support the concept of low nuclear grade breast neoplasia family. **32**, 513-523 (2008).

30. Natrajan,R. Loss of 16q in high grade breast cancer is associated with estrogen receptor status: Evidence for progression in tumors with a luminal phenotype? **48**, 351-365 (2009).

31. Rennstam,K., Baldetorp,B., Kytola,S., Tanner,M., & Isola,J. Chromosomal rearrangements and oncogene amplification precede aneuploidization in the genetic evolution of breast cancer. **61**, 1214-1219 (2001).

32. Courjal,F. Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups. **57**, 4360-4367 (1997).

33. Dontu,G., El-Ashry,D., & Wicha,M.S. Breast cancer, stem/progenitor cells and the estrogen receptor. **15**, 193-197 (2004).

34. Polyak,K. Breast cancer: origins and evolution. **117**, 3155-3163 (2007).

35. Sims,A.H., Howell,A., Howell,S.J., & Clarke,R.B. Origins of breast cancer subtypes and therapeutic implications. **4**, 516-525 (2007).

36. Kao,J. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. **4**, e6146 (2009).

37. Neve,R.M. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. **10**, 515-527 (2006).

38. Teschendorff,A.E., Miremadi,A., Pinder,S.E., Ellis,I.O., & Caldas,C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. **8**, R157 (2007).

39. Lim,E. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. **15**, 907-913 (2009).

40. McClintock,B. The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. **25**, 405-416 (1939).

41. McClintock,B. The Stability of Broken Ends of Chromosomes in Zea Mays. **26**, 234-282 (1941).

42. Iakovlev,V.V. Genomic differences between pure ductal carcinoma in situ of the breast and that associated with invasive disease: a calibrated aCGH study. **14**, 4446-4454 (2008).

43. Langerod,A. TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. **9**, R30 (2007).

44. Naume,B. Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. **1**, 160-171 (2007).

45. Lucito,R. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. **13**, 2291-2305 (2003).

46. Barrett,M.T. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. **101**, 17765-17770 (2004).

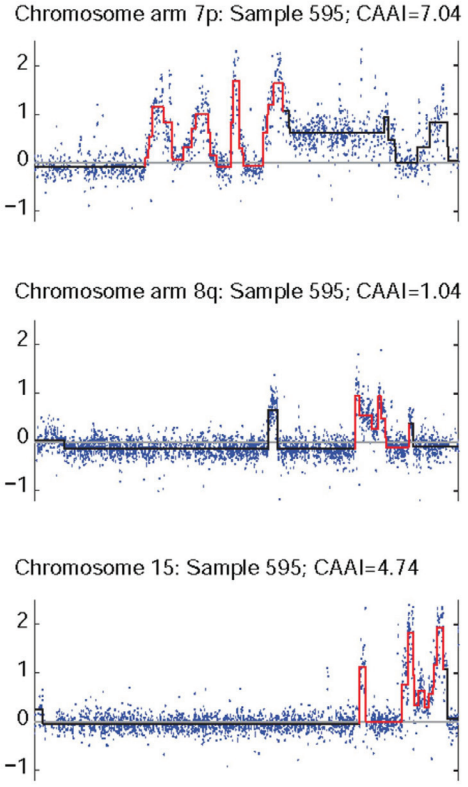
47. van den Ijssen Human and mouse oligonucleotide-based array CGH. **33**, e192 (2005).

48. R Development Core Team (2009). (R Foundation for Statistical Computing, Vienna, Austria.,2009).

49. Gentleman,R.C. Bioconductor: open software development for computational biology and bioinformatics. **5**, R80 (2004).

Figure 1:

1a:



1b:

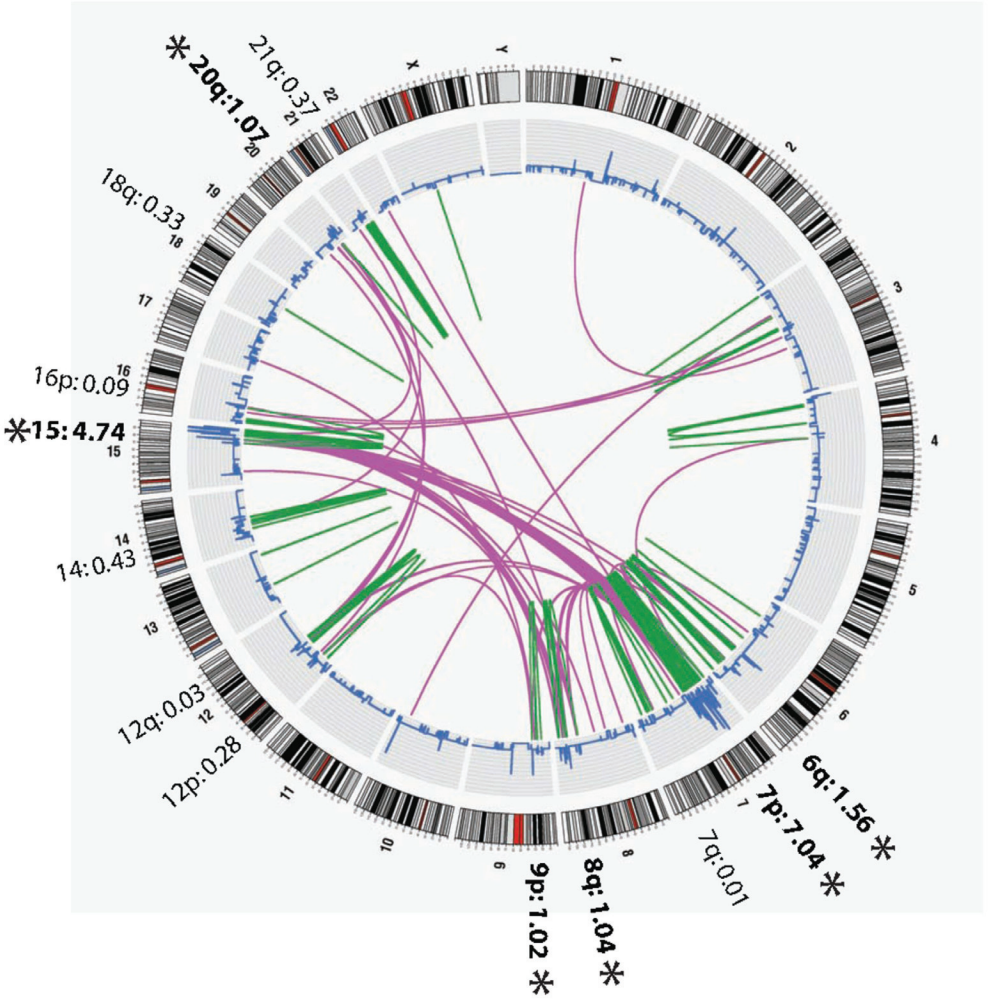


Figure 2:

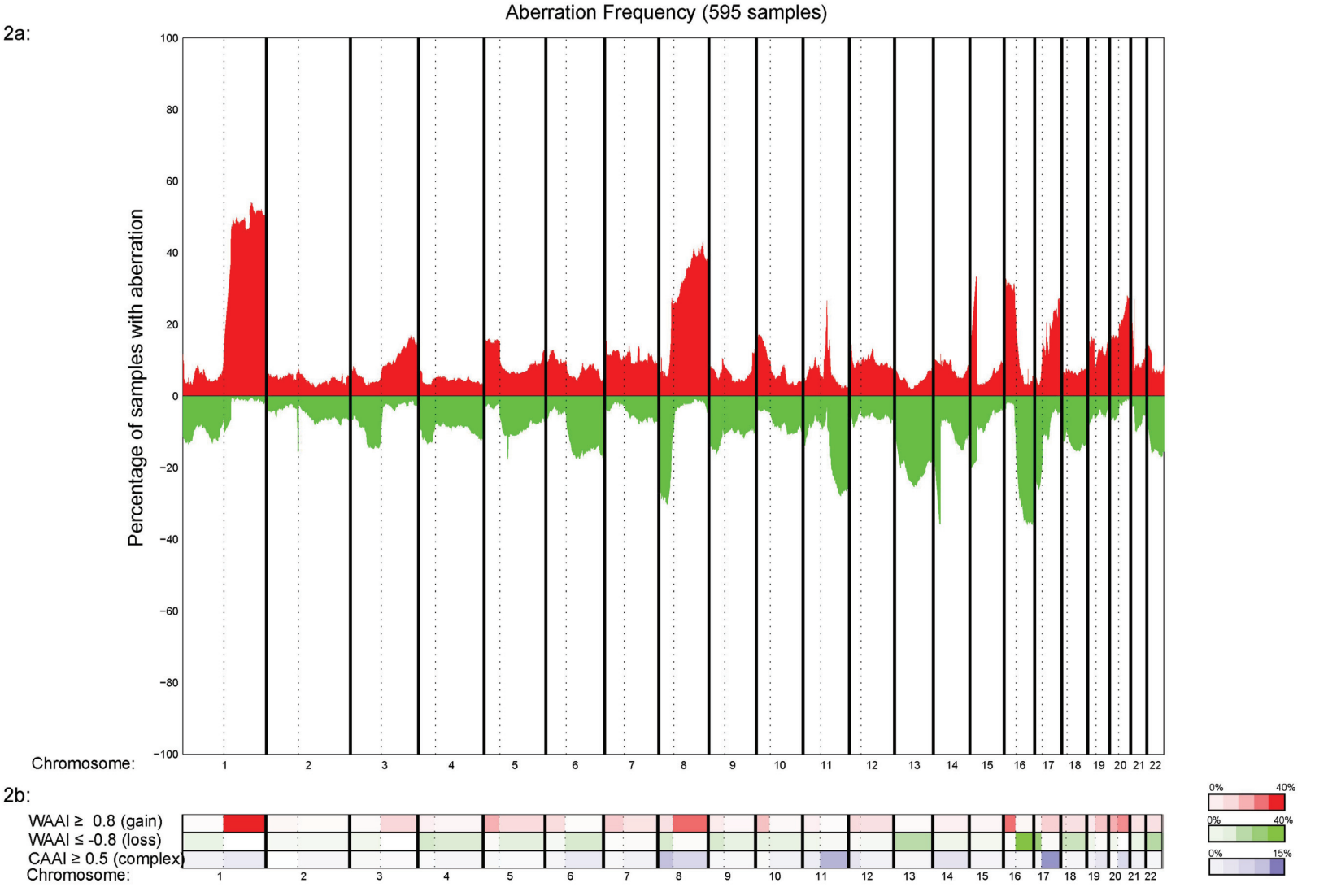

2a:
Aberration Frequency (595 samples)
Percentage of samples with aberration
100
80
60
40
20
0
−20
−40
−60
−80
−100
Chromosome:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
2b:
WAAI ≥ 0.8 (gain)
WAAI ≤ -0.8 (loss)
CAAI ≥ 0.5 (complex)
Chromosome:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
0% 40%
0% 40%
0% 15%

Figure 3:

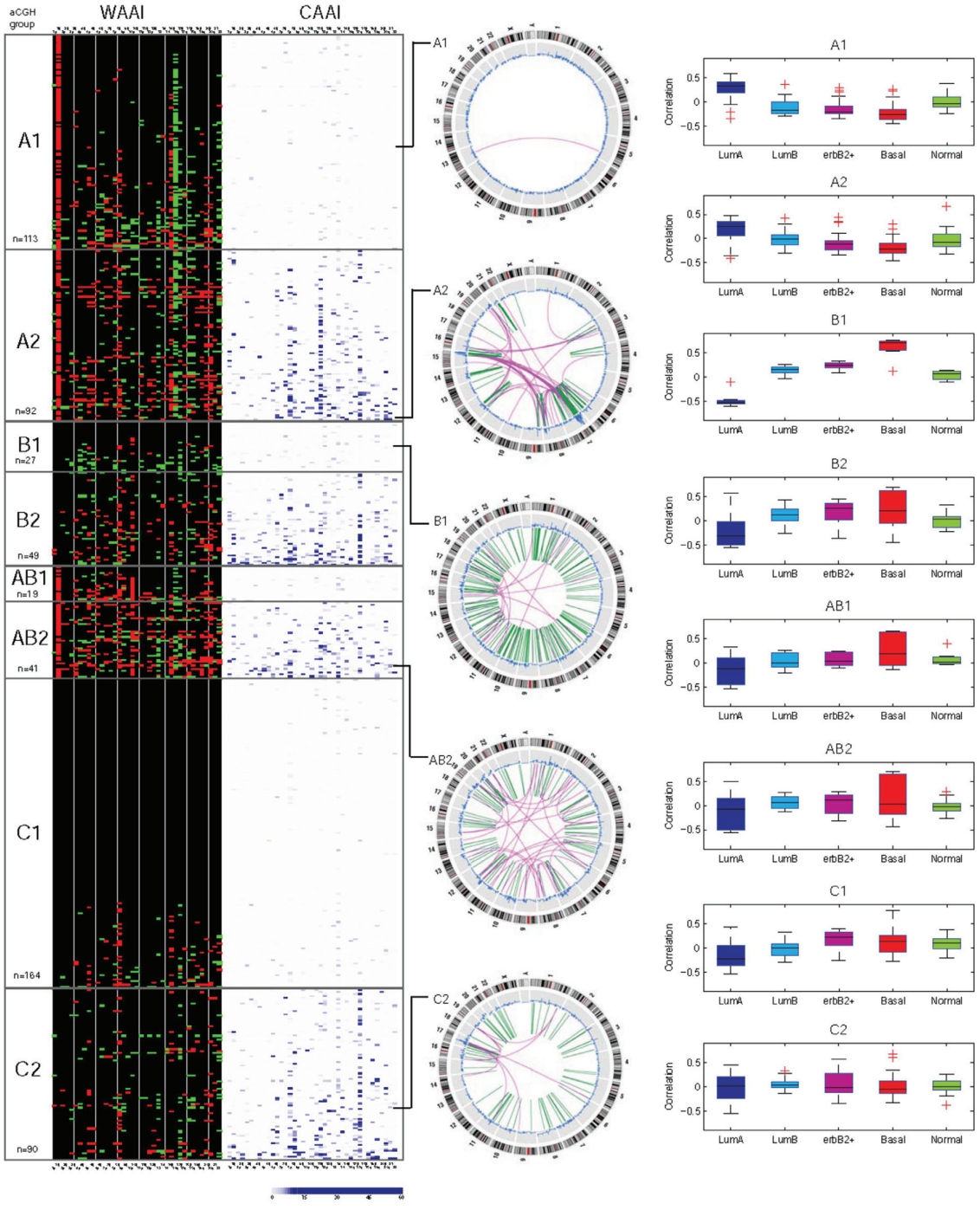

3a:
3b:
3c:
aCGH group
WAAI
CAAI
A1
n=113
A2
n=92
B1
n=27
B2
n=49
AB1
n=19
AB2
n=41
C1
n=164
C2
n=90
Correlation
LumA
LumB
erbB2+
Basal
Normal

Figure 4:

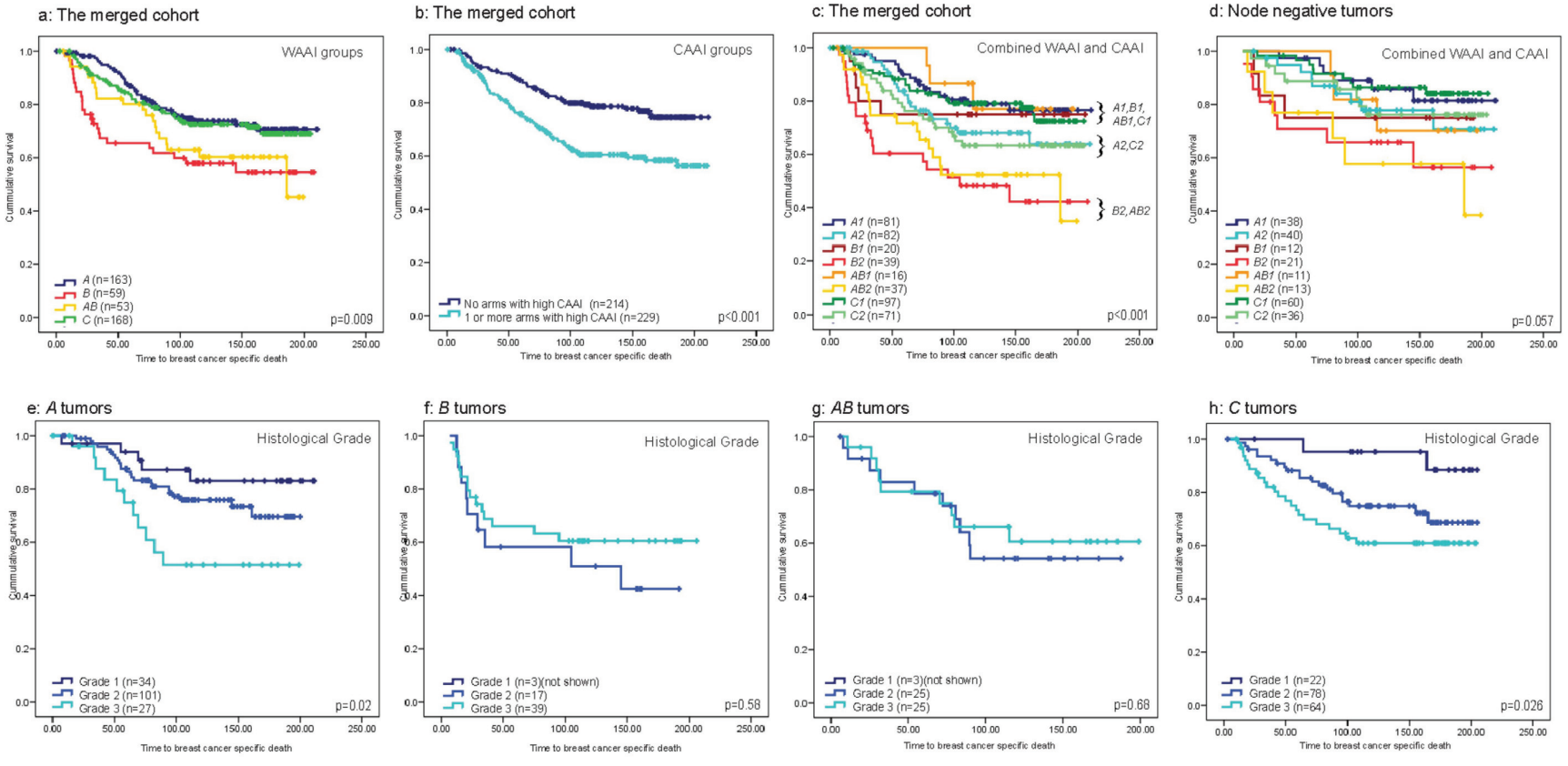
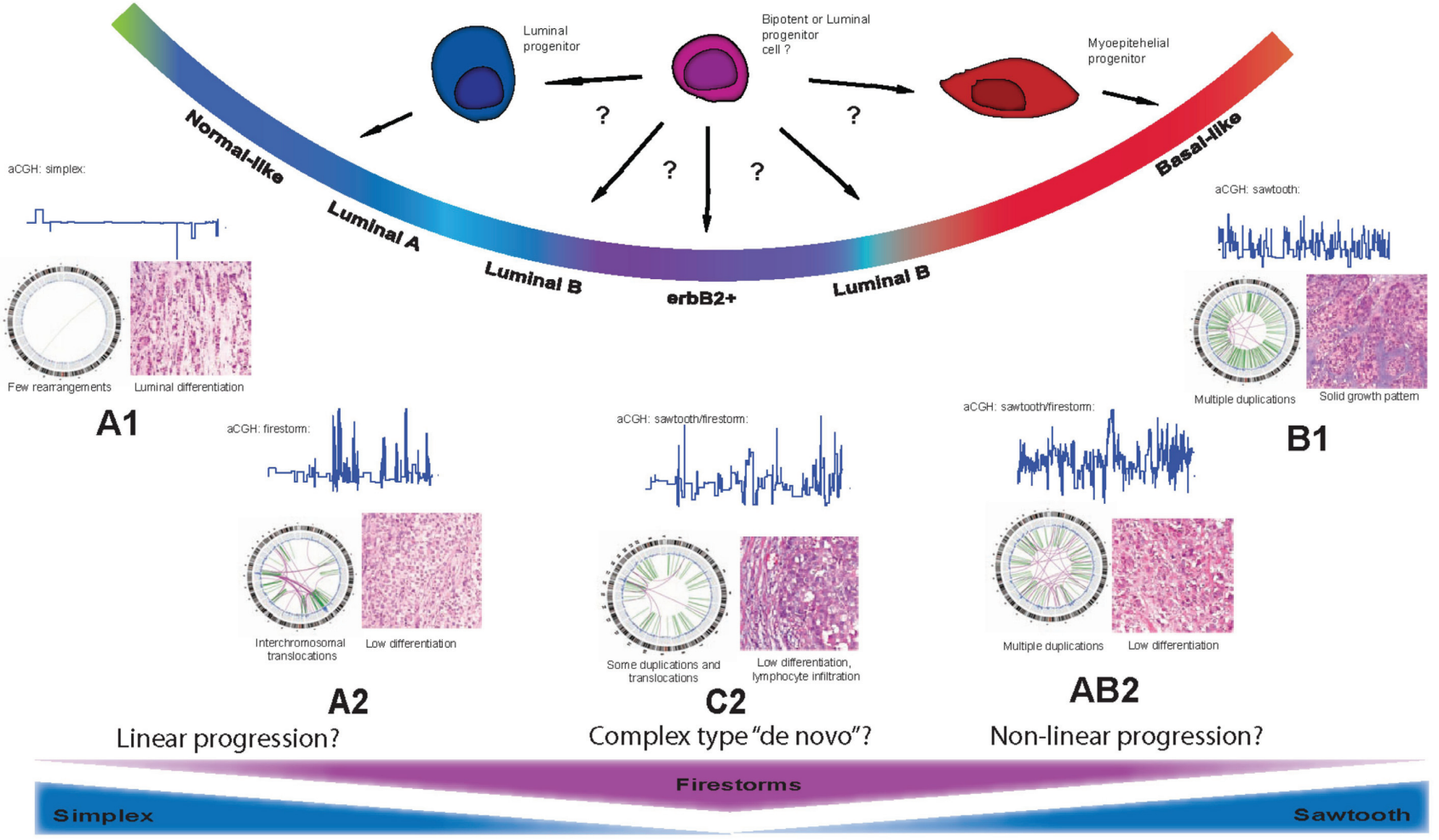

a: The merged cohort
WAAI groups
A (n=163)
B (n=59)
AB (n=53)
C (n=168)
p=0.009
b: The merged cohort
CAAI groups
No arms with high CAAI (n=214)
1 or more arms with high CAAI (n=229)
p<0.001
c: The merged cohort
Combined WAAI and CAAI
A1 (n=81)
A2 (n=82)
B1 (n=20)
B2 (n=39)
AB1 (n=16)
AB2 (n=37)
C1 (n=97)
C2 (n=71)
A1,B1, AB1,C1
A2,C2
B2,AB2
p<0.001
d: Node negative tumors
Combined WAAI and CAAI
A1 (n=38)
A2 (n=40)
B1 (n=12)
B2 (n=21)
AB1 (n=11)
AB2 (n=13)
C1 (n=60)
C2 (n=36)
p=0.057
e: A tumors
Histological Grade
Grade 1 (n=34)
Grade 2 (n=101)
Grade 3 (n=27)
p=0.02
f: B tumors
Histological Grade
Grade 1 (n=3)(not shown)
Grade 2 (n=17)
Grade 3 (n=39)
p=0.58
g: AB tumors
Histological Grade
Grade 1 (n=3)(not shown)
Grade 2 (n=25)
Grade 3 (n=25)
p=0.68
h: C tumors
Histological Grade
Grade 1 (n=22)
Grade 2 (n=78)
Grade 3 (n=64)
p=0.026
Cummulative survival
Time to breast cancer specific death

Figure 5:


Luminal progenitor
Bipotent or Luminal progenitor cell ?
Myoepitehelial progenitor
?
?
?
?
Normal-like
Luminal A
Luminal B
erbB2+
Luminal B
Basal-like
aCGH: simplex:
Few rearrangements
Luminal differentiation
A1
aCGH: sawtooth:
Multiple duplications
Solid growth pattern
B1
aCGH: firestorm:
Interchromosomal translocations
Low differentiation
A2
Linear progression?
aCGH: sawtooth/firestorm:
Some duplications and translocations
Low differentiation, lymphocyte infiltration
C2
Complex type "de novo"?
aCGH: sawtooth/firestorm:
Multiple duplications
Low differentiation
AB2
Non-linear progression?
Firestorms
Simplex
Sawtooth

Supplementary Table 1: Demographic data for the four cohorts

		MicMa no=125	WZ no=141	Ull no=167	Caldas n=162
		cases (% of available cases)	cases (% of available cases)	cases (% of available cases)	cases (% of available cases)
Age (mean, min-max)		61 (33-93)	53 (31-82)	63 (28-90)	57 (32-71)
Histologic type					
	IDC	98 (78%)	124 (88%)	110 (67%)	
	ILC	24 (19%)	11 (8%)	40 (25%)	
	Others	2 (2%)	4 (3%)	11 (7%)	
	DCIS	1 (1%)	2 (1%)	2 (1%)	
	Not available	*0*	*0*	*4/167 (2%)*	*162/162 (100%)*
Histologic grade					
	Grade I	14 (12%)	11 (11%)	11 (7%)	37 (23%)
	Grade II	60 (50%)	23 (22%)	110 (67%)	55 (34%)
	Grade III	47 (39%)	70 (67%)	43 (26%)	68 (43%)
	Not available	*4/125 (3%)*	*37/141 (26%)*	*3/167 (2%)*	*2/162 (1%)*
ER status					
	Positive	75 (60%)	92 (79%)	86 (57%)	107 (66%)
	Negative	49 (40%)	25 (21%)	65 (43%)	54 (34%)
	Not available	*1/125 (1%)*	*24/141 (17%)*	*16/167 (10%)*	*1/162 (1%)*
PgR status					
	Positive	58 (48%)		98 (59%)	
	Negative	64 (52%)		66 (40%)	
	Not available	*1/125 (1%)*	*141/141 (100%)*	*3/167 (2%)*	*162/162 (100%)*
HER2 FISH status					
	HER2/cent17 ≤2	84 (82%)			
	HER2/cent17 >2	19 (18%)			
	Not available	*22/125 (18%)*	*141/141 (100%)*	*167/167(100%)*	*162/162 (100%)*
TP53 status					
	TP53 wt	83 (66%)		124 (74%)	
	TP53 mut	42 (34%)		43 (26%)	
	Not available	*0/125 (0%)*	*141/141 (100%)*	*0/167 (0%)*	*162/162 (100%)*
Tumor size					
	T1	52 (43%)	71 (51%)	57 (35%)	113 (71%)
	T2	57 (47%)	65 (47%)	86 (53%)	47 (29%)
	T3	8 (7%)	2 (2%)	12 (8%)	0

	T4	4 (3%)	0	6 (4%)	0
	Not available	*4/125 (3%)*	*3/141 (2%)*	*6/167 (3%)*	*2/162 (1%)*
Node status					
	Node negative	51 (44%)	69 (49%)	73 (51%)	109 (69%)
	Node positive	64 (56%)	71 (51%)	70 (49%)	49 (31%)
	Not available	*10/125 (8%)*	*1/141 (1%)*	*24/167 (14%)*	*4/162 (3%)*
Ploidy					
	Diploid	41 (41%)	100 (71%)		
	Aneuploid	60 (59%)	41 (29%)		
	Not available	*24/125 (19%)*	*0/141 (0%)*	*167/167 (100%)*	*162/162 (100%)*
Expression class					
	Luminal A	49 (43%)		34 (47%)	54 (48%)
	Luminal B	14 (12%)		6 (8%)	13 (12%)
	erbB2+	19 (17%)		12 (16%)	14 (13%)
	Basal-like	14 (12%)		13 (18%)	19 (17%)
	Normal-like	14 (12%)		8 (11%)	12 (11%)
	Unclassified	3 (3%)		0	0
	Not available	*12/125 (10%)*	*141/141 (100%)*	*94/167 (56%)*	*50/162 (31%)*
Treatment, chemotherapy					
	No Chemotherapy	48 (40%)		139 (84%)	154 (96%)
	Chemotherapy	71 (60%)		26 (16%)	6 (4%)
	Not available	*6/125 (5%)*	*141/141 (100%)*	*2/167 (1%)*	*2/162 (1%)*
Treatment, Tamoxifen					
	No Tamoxifen	63 (53%)		125 (75%)	82 (51%)
	Tamoxifen	55 (47%)		41 (25%)	78 (49%)
	Not available	*7/125 (6%)*	*141/141 (100%)*	*2/167 (1%)*	*2/162 (1%)*
Adjuvant, general					
	No adjuvant	48 (40%)		108 (65%)	75 (47%)
	Adjuvant	71 (60%)		58 (35%)	86 (53%)
	Not available	*6 (5%)*	*141/141 (100%)*	*1/167 (1%)*	*1/162 (1%)*

Supplementary Table 2: Distribution between the WAAI groups and CAAI groups in the four cohorts.

	All four cohorts n=595	MicMa n=125	WZ n=141	Ull n=167	Caldas n=162
WAAI groups¥					
A	204/595 (34%)	49/125 (39%)	38/141 (27%)	66/167 (39%)	51/162 (31%)
B	76/595 (13%)	16/125 (13%)	13/141 (9%)	25/167 (15%)	22/162 (14%)
AB	60/595 (10%)	16/125 (13%)	6/141 (4%)	26/167 (16%)	12/162 (7%)
C	255/595 (43%)	44/125 (35%)	84/141 (60%)	50/167 (30%)	77/162 (48%)
CAAI*					
No CAAI	323/595 (54%)	68/125 (54%)	103/141 (73%)	64/167 (38%)	88/162 (54%)
High CAAI	272/595 (46%)	57/125 (46%)	38/141 (27%)	103/167 (62%)	74/162 (46%)

¥WAAI groups: A: WAAI≥0.8 on 1q and/or WAAI≤ -0.8 on 16q
B: Regional loss of 5q and/or gain of 10p
AB: Samples scored by the criteria for both A and B
C: Samples scored by neither of the criteria for A and B

*High CAAI is defined as CAAI≥0.5.

Supplementary Table 3: Clinico-pathological characteristics of the four WAAI groups

(total cases with available data)		A no=204 cases (% of all available A)	B no=76 cases (% of all available B)	AB no=60 cases (% of all available AB)	C n=255 cases (% of all available C)	Chi square:
Age (mean, min-max) (n=594)		62 (28-90)	57 (33-88)	56 (28-90)	57 (28-93)	P<0.001*
Histologic type						
n=429	IDC	108/153 (71%)	47/54 (87%)	44/48 (92%)	133/174 (76%)	
	ILC	42/153 (27%)	5/54 (9%)	3/48 (6%)	25/174 (14%)	
	Others	2/153 (1%)	1/54 (2%)	1/48 (2%)	13/174 (8%)	p=0.001
	DCIS	1/153 (1%)	1/54 (2%)	0/48 (0%)	3/174 (2%)	
Histologic grade						
n=549	Grade I	37/195 (19%)	3/71 (4%)	3/55 (6%)	30/228 (13%)	
	Grade II	115/195 (59%)	20/71 (28%)	25/55 (45%)	88/228 (39%)	p<0.001
	Grade III	43/195 (22%)	48/71 (68%)	27/55 (49%)	110/228 (48%)	
ER status						
n=553	Positive	153/187 (82%)	28/74 (38%)	34/59 (58%)	145/233 (62%)	
	Negative	34/187 (18%)	46/74 (62%)	25/59 (42%)	88/233 (38%)	p<0.001
PgR status						
n=286	Positive	70/112 (63%)	15/40(37%)	22/42 (52%)	49/92 (53%)	
	Negative	42/112 (37%)	25/40 (63%)	20/42 (48%)	43/92 (47%)	p=0.053
HER2 status						
n=103	Positive	4/43 (9%)	4/12 (33%)	3/11 (27%)	8/37 (22%)	
	Negative	39/43 (91%)	8/12 (67%)	8/11 (72%)	29/37 (78%)	p=0.174
TP53 status						
n=292	Positive	13/115 (11%)	28/41 (68%)	19/42 (45%)	25/94 (27%)	
	Negative	102/115 (89%)	13/41 (32%)	23/42 (55%)	69/94 (73%)	p<0.001
Tumor size						
n=580	pT1	94/194 (48%)	38/75 (51%)	22/59 (37%)	139/252 (55%)	
	pT2	89/194 (46%)	36/75 (48%)	31/59 (53%)	99/252 (39%)	
	pT3	8/194 (4%)	1/75 (1%)	4/59 (7%)	9/252 (4%)	p=0.275
	pT4	3/194 (2%)	0/75 (0%)	2/59 (3%)	5/252 (2%)	
Node status						
n=556	Node neg.	96/186 (52%)	41/72 (57%)	24/55 (44%)	141/243 (58%)	
	Node pos.	90/186 (48%)	31/72 (43%)	31/55 (56%)	102/243 (42%)	n=0.202

Expression class						
n=298	Luminal A	86/115 (75%)	2/38 (5%)	10/33 (30%)	39/112 (35%)	
	Luminal B	9/115 (8%)	10/38 (26%)	5/33 (15%)	9/112 (8%)	
	erbB2+	6/115 (5%)	7/38 (19%)	2/33 (6%)	30/112 (27%)	p<0.001
	Basal-like	1/115 (1%)	18/38 (47%)	14/33 (43%)	13/112 (13%)	
	Normal-like	12/115 (10%)	1/38 (3%)	2/33 (6%)	19/112 (17%)	
	Unclassified	1/115 (1%)	0/38 (0%)	0	2/112 (2%)	
Ploidy						
n=242	Diploid	54/80 (68%)	7/25 (28%)	6/18 (33%)	74/119 (62%)	
	Aneuploid	26/80 (32%)	18/25 (72%)	12/18 (67%)	45/119 (38%)	p=0.001
Treatment, Tamoxifen						
n=444	No Tam.	94/160 (59%)	42/62 (68%)	31/52 (60%)	103/170 (61%)	
	Tam.	66/160 (41%)	20/62 (32%)	21/52 (40%)	67/170 (39%)	p=0.666
Treatment, CMF						
n=444	No CMF	125/160 (78%)	48/63 (76%)	33/51 (65%)	135/170 (79%)	
	CMF	35/160 (22%)	15/63 (24%)	18/51 (35%)	35/170 (21%)	p=0.171
Adjuvant, general						
n=446	No adjuvant	86/161 (53%)	32/63 (51%)	22/52 (42%)	91/170 (54%)	P=0.517
	Adjuvant	65/161 (47%)	31/63 (49%)	30/52 (58%)	79/170 (46%)	

* Kruskal Wallis test

Supplementary Table 4: Correlation between molecular expression subgroups and WAAI groups:

	A1	A2	B1	B2	AB1	AB2	C1	C2	Total:
Luminal A	52/65 (80%)	34/50 (68%)	0/14 (0%)	2/24 (8%)	4/12 (33%)	6/21 (29%)	20/61 (33%)	19/51 (37%)	**137**
Luminal B	2/65 (3%)	7/50 (14%)	3/14 (21%)	7/24 (29%)	1/12 (8%)	4/21 (19%)	4/61 (7%)	5/51 (10%)	**33**
erbB2+	2/65 (3%)	4/50 (8%)	1/14 (7%)	6/24 (25%)	0/12 (0%)	2/21 (10%)	15/61 (25%)	15/51 (29%)	**45**
Basal-like	1/65 (2%)	0/50 (0%)	10/14 (71%)	8/24 (33%)	5/12 (42%)	9/21 (42%)	7/61 (11%)	6/51 (12%)	**46**
Normal-like	8/65 (12%)	4/50 (8%)	0/14 (0%)	1/24 (4%)	2/12 (17%)	0/21 (0%)	14/61 (23%)	5/51 (10%)	**34**
NC*	0/65 (0%)	1/50 (2%)	0/14 (0%)	0/24 (0%)	0/12 (0%)	0/21 (0%)	1/61 (1%)	1/51 (2%)	**3**
Total:	**65**	**50**	**14**	**24**	**12**	**21**	**61**	**51**	**298**

*NC: samples with low correlation to all five centroids.

Supplementary Figure legends:

Supplementary Figure 1: Validation of CAAI
Scatter plot of CAAIs for 17q compared to mean HER2 copy number measured by FISH (MicMa cohort). The broken lines indicate the selected threshold (CAAI=0.5). Samples with CAAI>=0.5 all except one had 4 or more copies of HER2. A few samples had increased HER2 copy number but CAAI<0.5. Inspection of the corresponding aCGH profile in such cases revealed that increased copy number was due to narrow amplicons and not to complex rearrangements of the firestorm type.

Supplementary Figure 2: Arm wise distribution of WAAI
The box plots showing the arm-wise distribution of WAAI are illustrating the non-random distribution of positive and negative WAAI scores in the ChinUC, MicMa/WZ and Ull cohort. The chromosomes arms are on the x-axis, and the WAAI sores on the y-axis. The distribution of WAAI is approximately symmetric around zero for most arms, but for others, such as 1q, 8q and 16p, WAAI is skewed towards positive values. For others, such as 16q and 17q, WAAI is skewed towards negative values.

Supplementary Figure 3: WAAI and centromere close translocation
a: Plotted aCGH values for chromosome arm 1q and 16q from case WZ061; unsegmented data as blue points and PCF values as black line showed whole chromosome arm gain of 1q and loss of 16q. This was reflected in the estimated WAAI; WAAI= 1.221 for 1q and WAAI= -1.465 for 16q.
b: Multi gene FISH analyses with five selected probes derived from centromere close BAC clones on chr.1 and chr.16 were hybridized to tumor cells (imprint) from WZ061. The image at the top show a tumor cell with all fluorescent probes superimposed revealing two green signals together, one orange and red and one green and orange (note that the probes will never be fused due to the large stretches of heterochromatin around the centromere). The illustration at bottom left show the combination of fluorochromes observed in nuclei from lymphocytes with non-translocated chr.1 and chr.16. To the right the observed combination in the tumor cells demonstrating a translocation and a derivative chromosome; der(1;16)(10q;10p) is illustrated.

Supplementary Figure 4: Frequencies of gains and loss in the four cohorts.
Frequency plots illustrating the percentage of samples with gain and loss within each cohort (red; gain, green; loss). The WZ cohort is enriched in diploid tumors by selection and has fewer events in total than the others, but the dominating alterations such as gains on 1q, 8q, 16p and 20q and loss on 6q, 8p, 11q,13, 16q, 17p and 22 is seen in all four cohorts.

Supplementary Figure 5: Frequencies of gain and loss of the eight WAAI/CAAI defined groups.
Frequency plots illustrating the percentage of samples with gain and loss within each WAAI/CAAI group (red; gain, green; loss). *A1* tumors are dominated by gain on 1q and 16p and loss on 16q. These alterations are frequent in *A2* tumors, in addition to gain on 8q, 17q and 20q and loss on 6q, 8p, 11q, 13 and 17p. *B1*, *B2*, *AB1* and *AB2* tumors have

almost similar patterns of gain and loss where almost all chromosomes are affected, a pattern very dissimilar from aberrations in *A1* and *A2* tumors. *C1* tumors have few alterations, with gain of 8q dominating. This is the most frequent aberration in *C2* tumors as well, followed by gain on 1q, 17q and 20q.

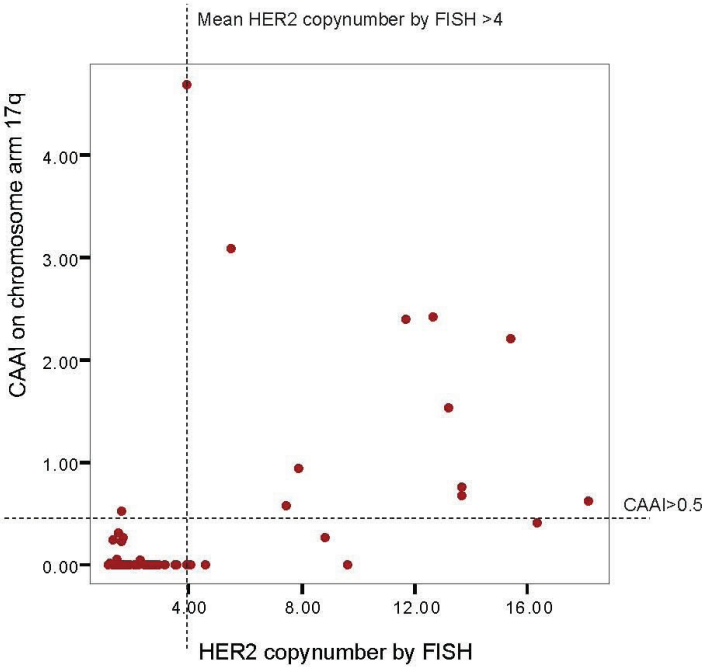
Supplementary Figure 6: Frequencies of WAAI in the WAAI/CAAI groups.
Top: Bar plots illustrating the mean number of altered arms, either gain or loss (WAAI≥0.8 or WAAI≤ -0.8) for each WAAI/CAAI group. *C* tumors had fewest alterations, and this persisted even if we omitted all cases without any alterations ('flat' aCGH profiles). *A* and *B* tumors had intermediate number of arms altered, with slightly more in the latter group. *AB* tumors had the highest mean value of altered whole arms. In all four groups more arms were altered in the tumors with high CAAI on one arm or more.
Middle: Bar plots illustrating the mean number of gained arms (WAAI≥0.8) for each WAAI/CAAI group. The same tendency reflected by the total number of alterations was seen for gains alone.
Bottom: Bar plots illustrating the mean number of lost arms (WAAI≤ -0.8) for each WAAI/CAAI group. In contrast to gains, the WAAI/CAAI groups seemed to have almost equal mean number of arms altered. This illustrate that the *B1* group was dominated by tumors with losses, and that the total increase in altered arms seen in samples with high CAAI compared to those with low CAAI mainly was due to gains and not losses.

Supplementary Figure 7: Chromosome wise frequencies of WAAI in the WAAI/CAAI groups.
The four plots show the arm wise frequency of samples with whole arm gain or loss as measured by WAAI (whole arm gain; WAAI≥0.8, whole arm loss; WAAI≤ -0.8). The plot at the top show *A1* and *A2* samples (dark and light blue bars), followed by *B1* and *B2* samples (dark and bright red), *AB1* and *AB2* samples (orange and yellow) and at the bottom the *C1* and *C2* samples (dark and light green).The *A1* and *A2* tumors had the same distributions of altered arms, but *A2* tumors had more frequent gain of 8q, 16p, 20p and 20q. *B1* tumors were dominated by whole arm losses (such as 17p, 4p, 4q and 5q), while *B2* tumors had more frequent gain of 8q, 10p16p and 20q. *AB* tumors had whole arm alterations resembling both the loss and gain pattern of both *A* and *B* tumors, with only little difference between *AB1* and *AB2* tumors. *C* tumors had the fewest numbers of whole arm alterations with gain of 8q and 16p and loss of 17p and 22 as the most frequent.

Supplementary Figure 8: Chromosome wise frequencies of CAAI in the WAAI/CAAI groups.
The four plots show the arm wise frequency of samples with complex rearrangements as measured by CAAI (CAAI≥0.5). The plot at the top show *A2* samples (light blue bars), followed by *B2* samples (bright red), *AB2* samples (yellow) and at the bottom the *C2* samples (light green). *A2* tumors had high CAAI most frequent on 11q, 8p, 17q and 8q, B2 tumors had high CAAI on more arms (such 17q, 20q and 8p) while AB2 had a more heterogeneous distribution pattern of arms with high CAAI. In C tumors, high CAAI was most frequent on 17q.

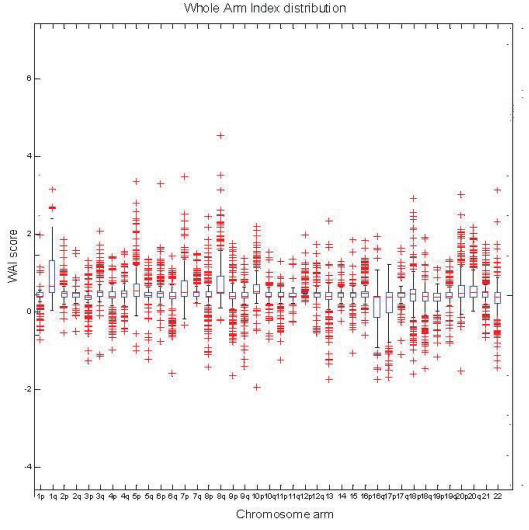
Supplementary Figure 9: Ploidy measurements and histological grade in the WAAI/CAAI groups.
Top: Bar plot illustrating the distribution of aneuploid and diploid samples in each of the eight WAAI/CAAI groups. All groups had both diploid and aneuploid tumors, it was a higher percentage of diploids in *A* and *C* tumors compared to *B* and *AB*, and aneuploid tumors were more frequent in all groups with high CAAI compared to the respective groups with low CAAI.
Bottom: Bar plot illustrating the distribution of histological grade in the eight WAAI/CAAI groups. Grade 1 tumors were most frequent in *A1* and *AB1* tumors, and rarely found in the other groups. In *A* tumors, there were a reduced proportion of grade 1 and grade 2 tumors in *A2* compared to *A1*, the same was seen for the *C* tumors. The highest percentage of grade 3 tumors was found in the *B* group.

Supplementary Figure 1:

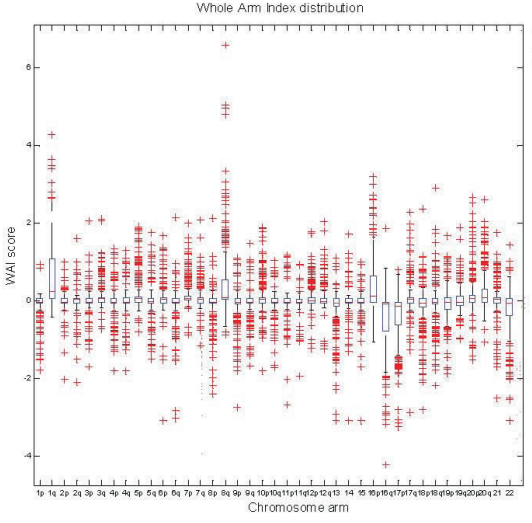


Supplementary Figure 2:

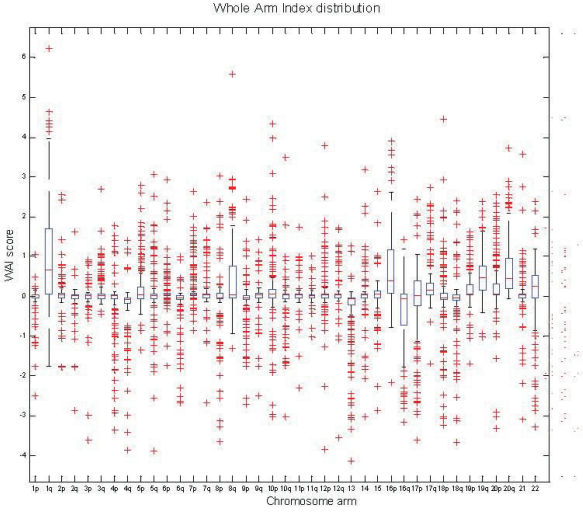
ChinUC (32k):



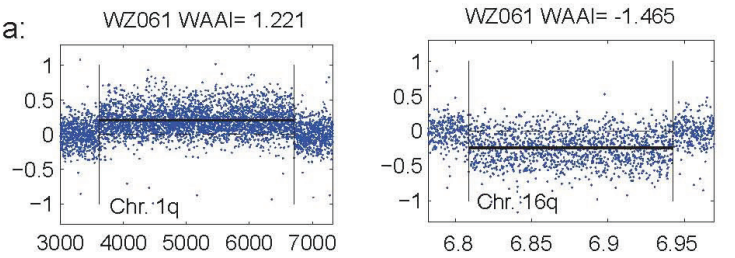
MicMa/WZ (82k):

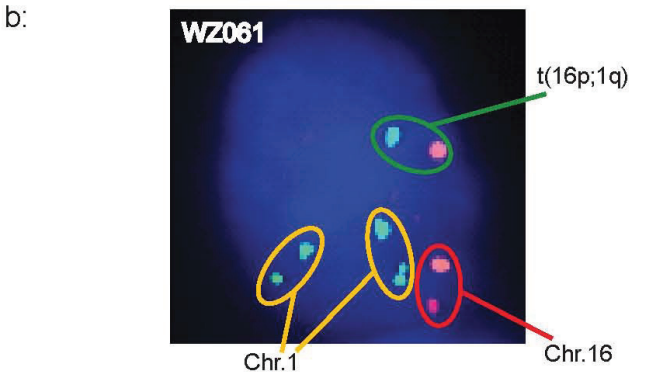


UII (244k):

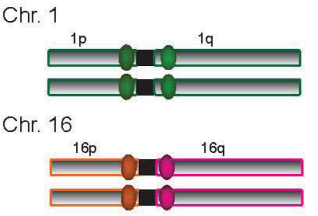


Supplementary Figure 3:

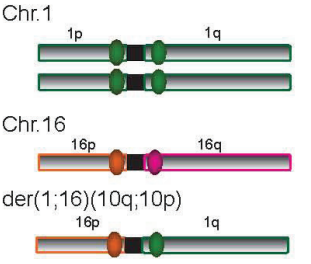




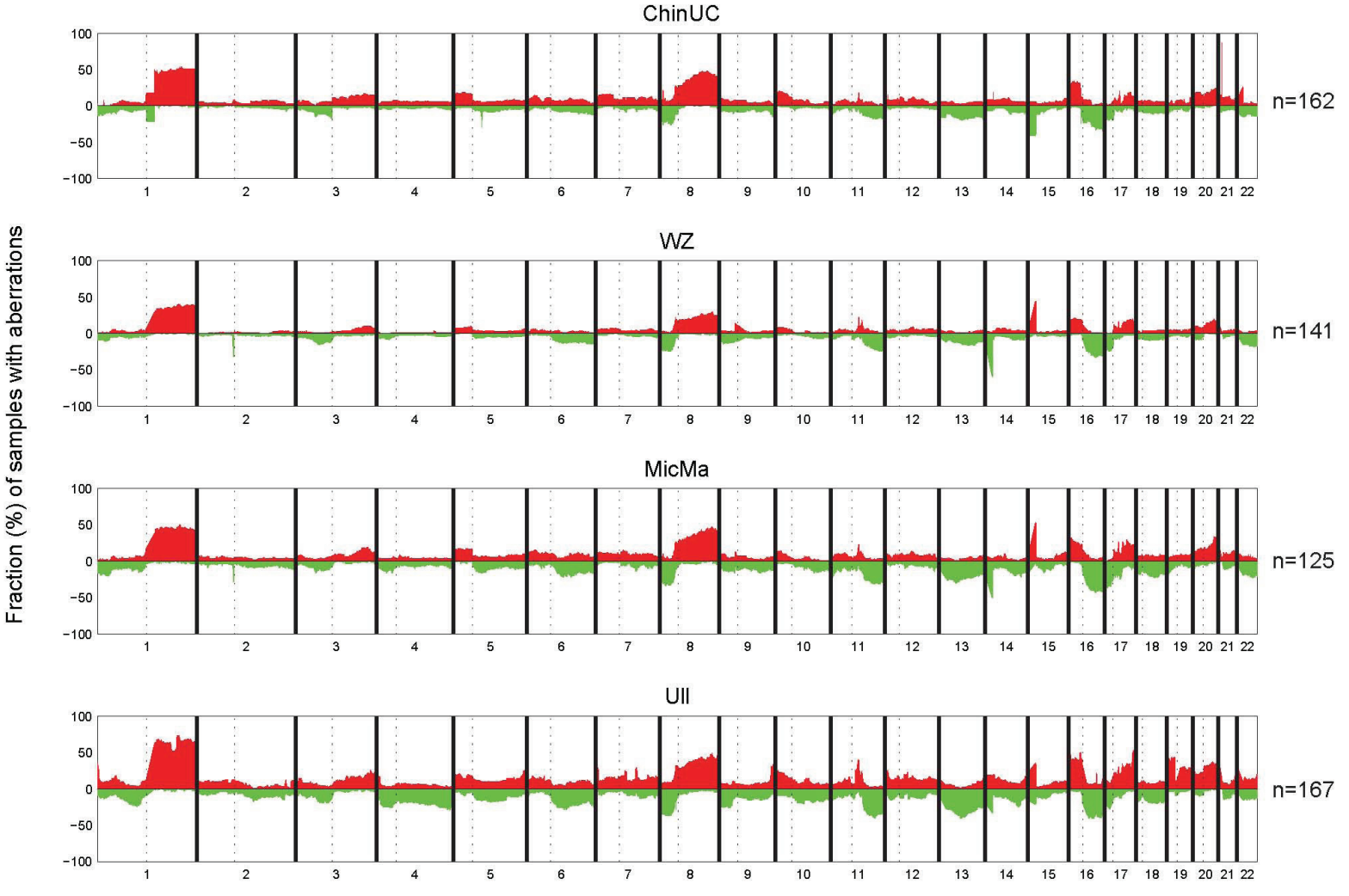
Probe positions and colours ("normal" genome):



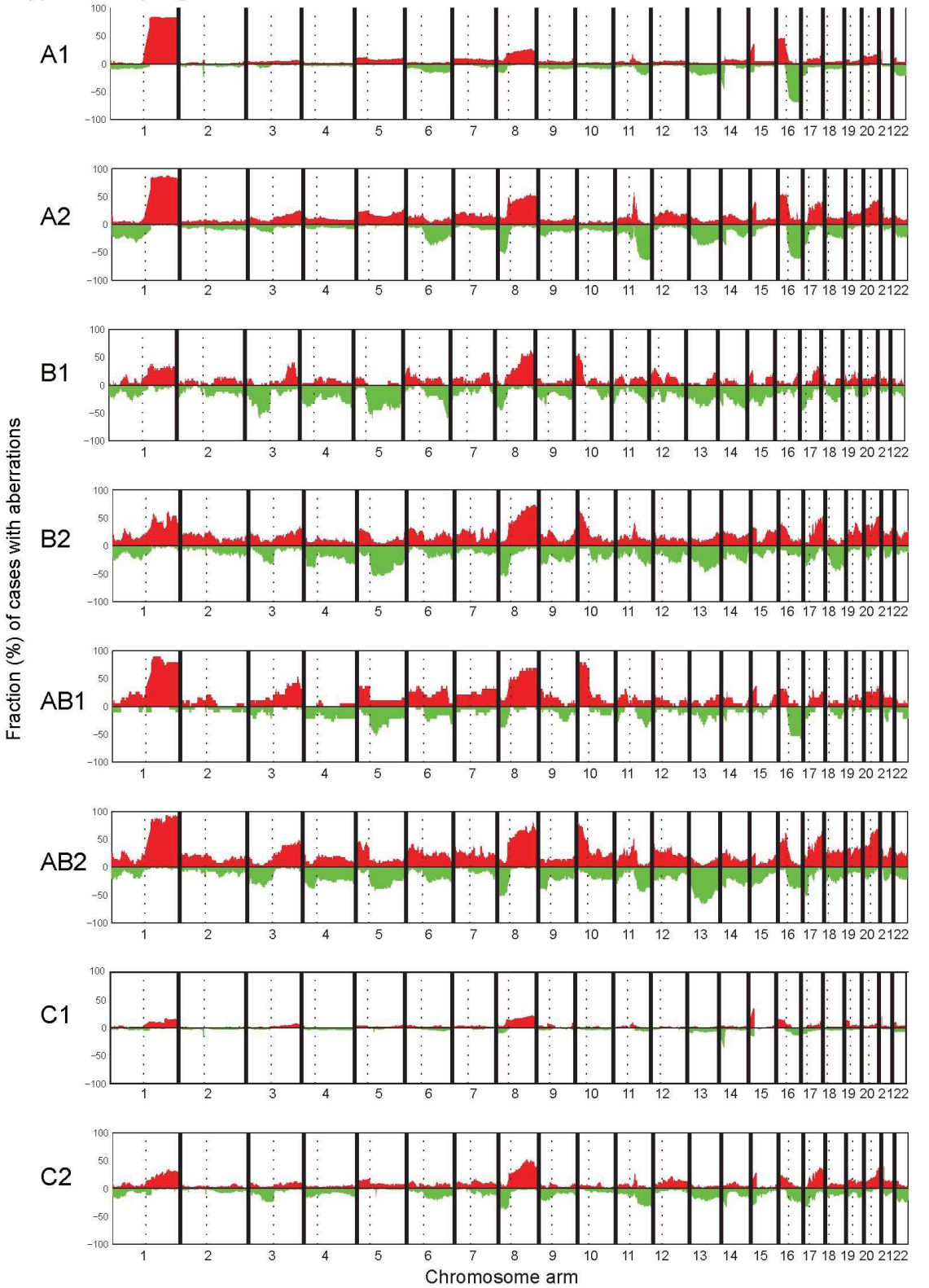
Observed probe combination:



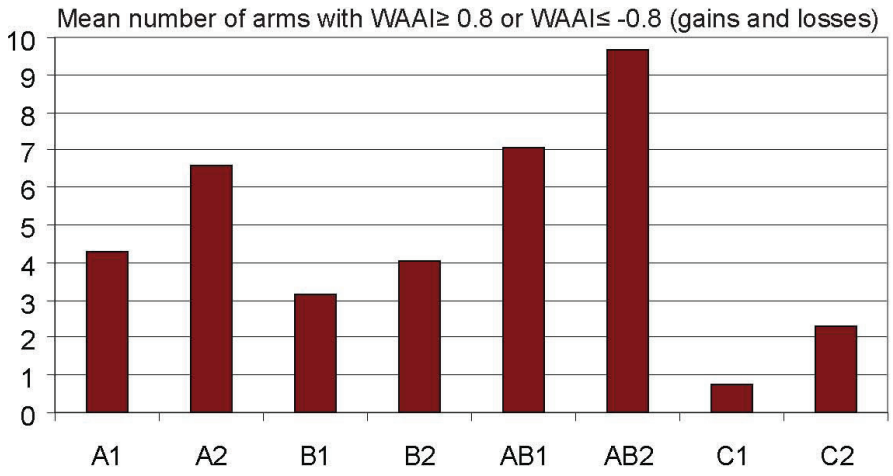
Supplementary Figure 4:

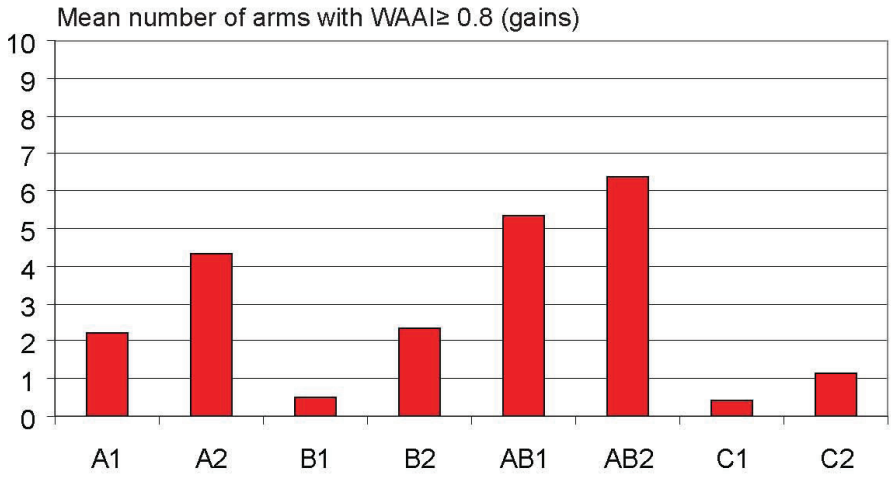


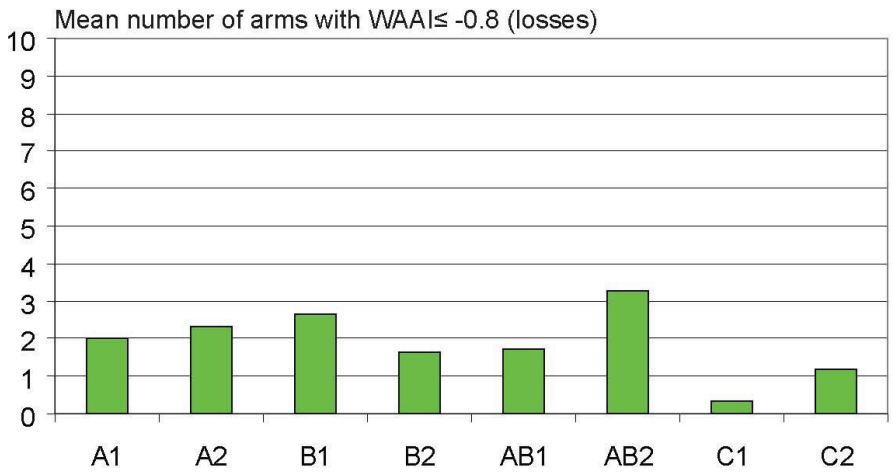
Supplementary Figure 5



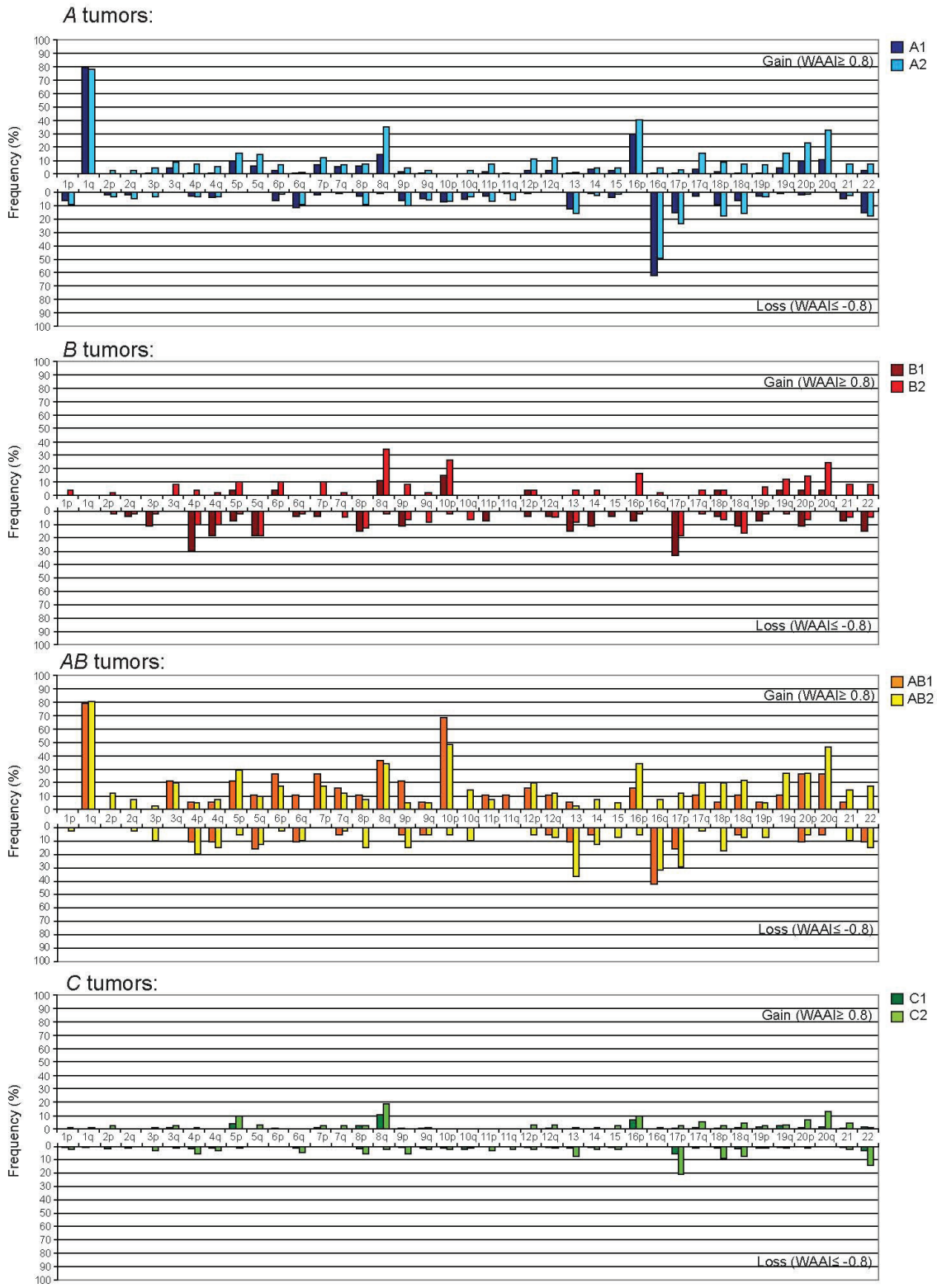
Supplementary Figure 6:



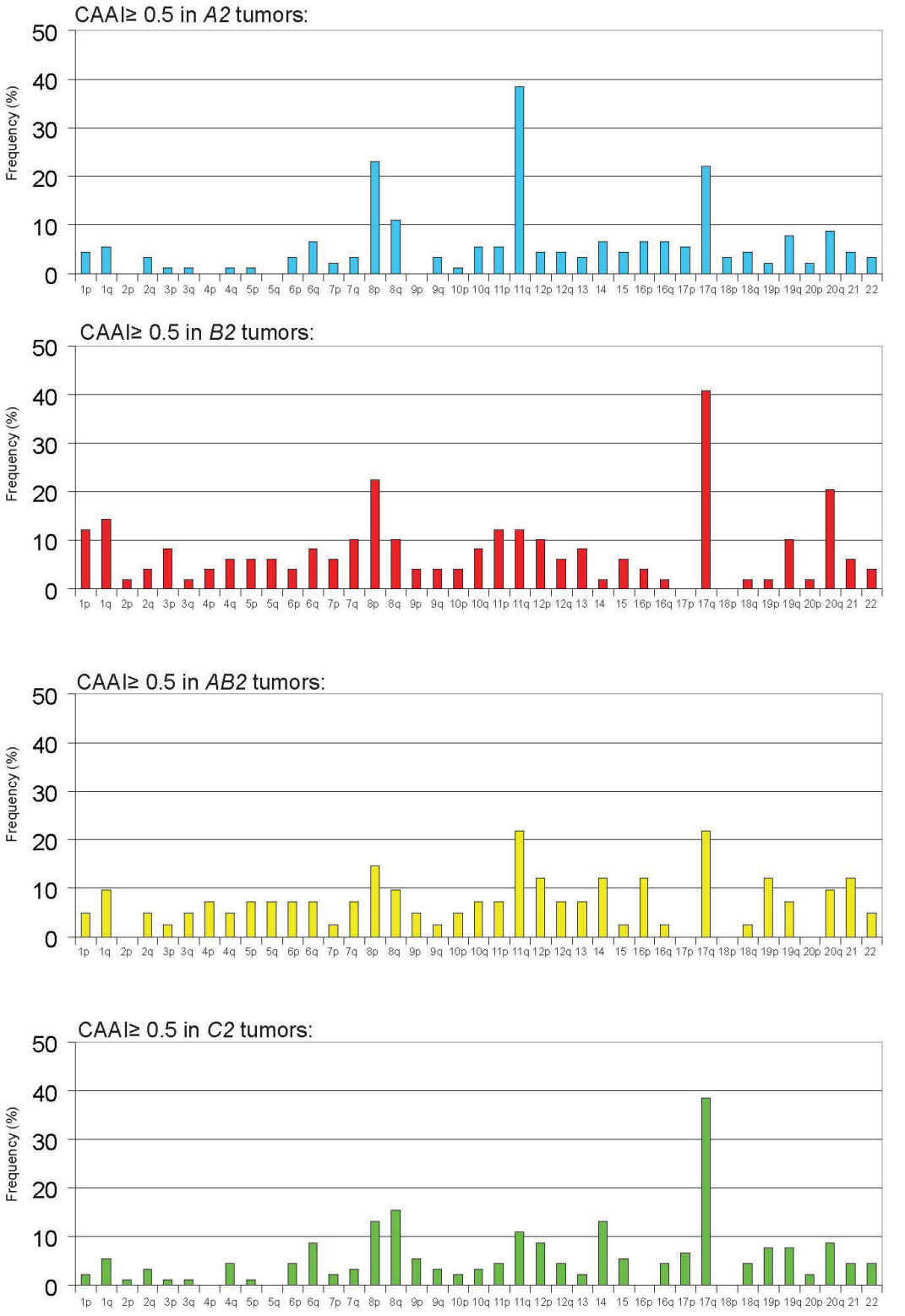




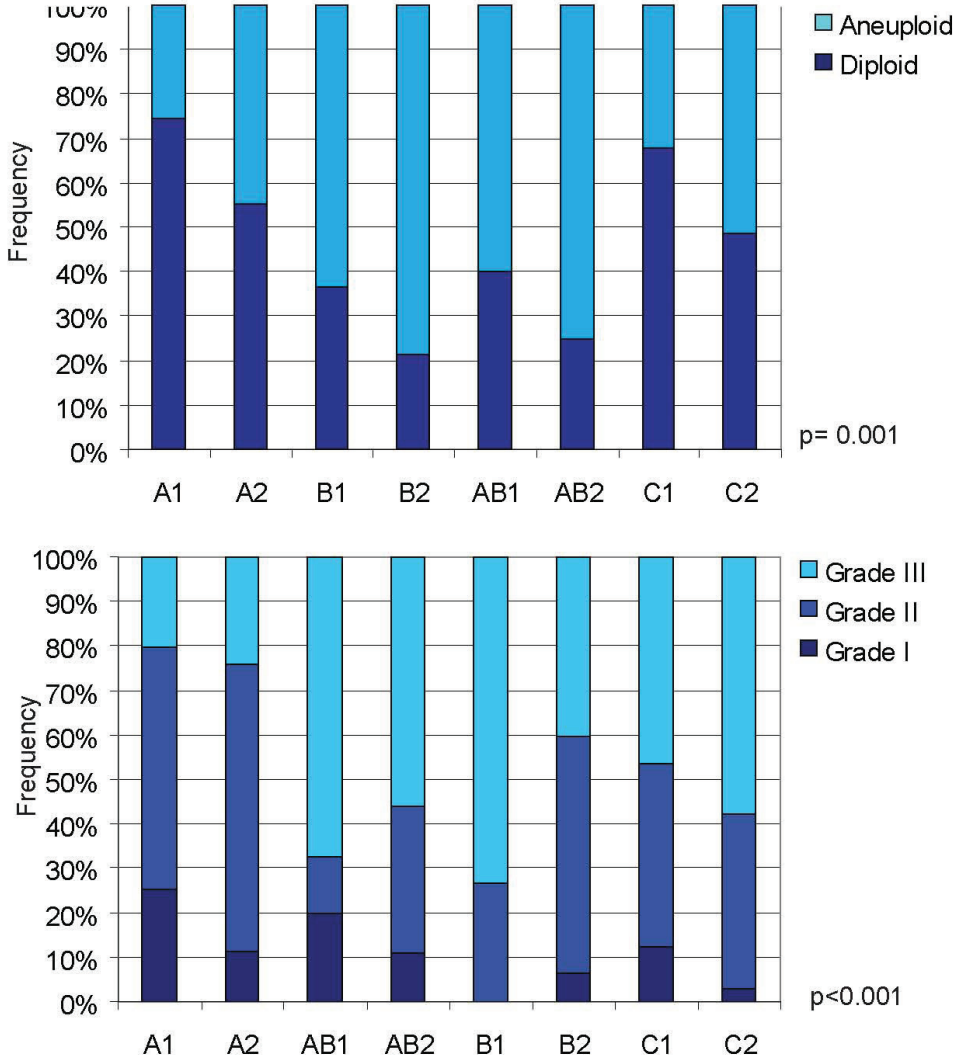
Supplementary Figure 7:



Supplementary Figure 8:



Supplementary Figure 9:



III

This article is removed.

IV

This article is removed.