



# LUND UNIVERSITY

## Ultrasonic Molecular Monitoring of Breast Cancer and Acute Myeloid Leukemia

Chen, Yilun

2021

*Document Version:*

Publisher's PDF, also known as Version of record

[Link to publication](#)

*Citation for published version (APA):*

Chen, Y. (2021). *Ultrasonic Molecular Monitoring of Breast Cancer and Acute Myeloid Leukemia*. [Doctoral Thesis (compilation), Department of Clinical Sciences, Lund]. Lund University, Faculty of Medicine.

*Total number of authors:*

1

### General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117  
221 00 Lund  
+46 46-222 00 00

# Ultrasensitive Molecular Monitoring of Breast Cancer and Acute Myeloid Leukemia

YILUN CHEN

DEPARTMENT OF CLINICAL SCIENCES, LUND | FACULTY OF MEDICINE | LUND UNIVERSITY





## FACULTY OF MEDICINE

Department of Clinical Sciences, Lund

Lund University, Faculty of Medicine

Doctoral Dissertation Series 2021:119

ISBN 978-91-8021-126-0

ISSN 1652-8220



# Ultrasensitive Molecular Monitoring of Breast Cancer and Acute Myeloid Leukemia

Yilun Chen



**LUND**  
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.

To be defended at Segerfalksalen, BMC, Lund.

Monday, November 22, 2021, at 13:00.

*Faculty opponent*

Professor Hans Petter Eikesdal, MD PhD

Department of Clinical Sciences

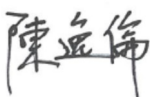
University of Bergen

Bergen, Norway



<b>Organization</b> LUND UNIVERSITY Faculty of Medicine Department of Clinical Sciences, Lund Division of Oncology  Author(s) Yilun Chen	<b>Document name</b> DOCTORAL DISSERTATION	
	<b>Date of issue</b> 2021-11-22	
	Sponsoring organization	
<b>Title and subtitle</b> Ultrasensitive Molecular Monitoring of Breast Cancer and Acute Myeloid Leukemia		
<b>Abstract</b>  <p>Cancer is the common name to a group of biologically diverse malignant neoplastic diseases. Approximately 18 million people are diagnosed with cancer annually and 8.8 million patients die from it. Tumorigenesis and progression of cancer are driven by alterations in the cancer cell genome. These alterations lead to gain of oncogenic functions, loss of tumor suppressor functions, or may be chromosomal rearrangements without obvious function, and these alterations themselves can serve as tumor-specific biomarkers that may have diagnostic and clinical utility.</p> <p>In this thesis, we investigated oncogenic and tumor suppressive genes in breast cancers and leukemias, with a focus on the PTEN/PIK3CA pathway as well as minimally-invasive monitoring of cancer patients using "liquid biopsies." We studied the underlying mechanism of PTEN protein loss in breast cancer, and showed how various types of tumor-specific mutations, including those in PIK3CA, can be used as biomarkers to monitor the dynamics of occult tumor burden, evaluate the degree of tumor content dissemination into the bloodstream with mammographic compression, and detect minimal residual disease in breast cancer and acute myeloid leukemia.</p> <p>In Paper I, we found that the frequent loss of PTEN protein in human breast cancer is not attributable to the overexpression of the E3 ubiquitin ligase NEDD4, and thus NEDD4 is unlikely to be a regulator of the oncogenic PI3K/PTEN signaling pathway. In Paper II, we showed that serial monitoring of tumor specific chromosomal rearrangements, identified with low coverage whole genome sequencing and then measured in blood samples by digital PCR (dPCR), is a highly sensitive and specific approach to detect occult breast cancer disease prior to the onset of symptoms and clinical detection. Detected plasma ctDNA level was a quantitative predictor of poor relapse-free and overall survival. In Paper III, we confirmed the general safety of mammography, using FDA approved CellSearch® and our ultrasensitive mutation detection dPCR technology IBSAFE, that mammographic compression of the breast with a breast tumor does not appear to lead to significant additional dissemination of CTCs and ctDNA into the bloodstream. In Paper IV, we showed that acute myeloid leukemia specific mutations can be serially monitored in follow-up bone marrow samples by IBSAFE, providing an insight in subclonal evolution of the leukemia and the status of minimal residual disease.</p> <p>These results and our mutation detection technology suggest they have high potential to be utilized in assessing treatment response, monitoring the disease course, detecting remnant tumor deposits with targetable mutations, and helping to speed the development of new drugs in the future.</p>		
<b>Key words</b> breast cancer, AML, PTEN, PI3K, liquid biopsy, sequencing, cfDNA, ctDNA, structural variant, mammography, CTC, mutation detection, IBSAFE, MRD		
Classification system and/or index terms (if any)		
Supplementary bibliographical information		<b>Language</b> English
<b>ISSN</b> and key title 1652-8220		<b>ISBN</b> 978-91-8021-126-0
Recipient's notes	<b>Number of pages</b> 125	
	Price	
Security classification		

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature 

Date 2021-10-15

# Ultrasensitive Molecular Monitoring of Breast Cancer and Acute Myeloid Leukemia

Yilun Chen



**LUND**  
UNIVERSITY

Cover photo by © J'z Jamason Chen 2021

Copyright © Yilun Chen 2021

Faculty of Medicine  
Department of Clinical Sciences, Lund

ISBN 978-91-8021-126-0  
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University  
Lund 2021



Media-Tryck is a Nordic Swan Ecolabel  
certified provider of printed material.  
Read more about our environmental  
work at [www.mediatryck.lu.se](http://www.mediatryck.lu.se)

**MADE IN SWEDEN** 

*To my mother*

*“All parabolas are similar.”  
Euclid, Elements of Conics*

*“The world is neither meaningful, nor absurd. It quite simply  
is, and that, in any case, is what is so remarkable about it.”  
Alain Robbe-Grillet*

# Table of Contents

<b>List of papers included in the thesis.....</b>	<b>1</b>
<b>Other published papers.....</b>	<b>3</b>
<b>Abstract .....</b>	<b>5</b>
<b>Popular summary .....</b>	<b>7</b>
<b>Abbreviations.....</b>	<b>9</b>
<b>Introduction .....</b>	<b>15</b>
Mutations in the cancer genome.....	17
Breast cancer .....	20
Risk factors.....	20
Diagnosis .....	21
Histopathology .....	21
Tissue biomarker status and breast cancer classification .....	22
Treatment.....	24
Minimally-invasive liquid biopsy.....	26
PI3K/PTEN signaling pathway .....	28
Acute myeloid leukemia (AML) .....	30
Diagnosis .....	32
Treatment.....	32
Minimal residual disease (MRD) .....	33
<b>Aims .....</b>	<b>37</b>
<b>Methods .....</b>	<b>39</b>
Patients, samples, and ethics .....	39
Immunohistochemistry (IHC) .....	40
DNA microarray.....	42
High-throughput sequencing (HTS).....	43
RNA sequencing (RNA-seq).....	46
DNA Sequencing.....	50
DNA purification from the follow-up samples .....	53
Multicolor Flow Cytometry (MFC) .....	55

Circulating Tumor Cell (CTC) analysis .....	56
Polymerase Chain Reaction (PCR) .....	57
Quantitative PCR (qPCR).....	60
Digital PCR (dPCR) .....	62
<b>Results and Discussion .....</b>	<b>75</b>
<b>Paper I</b> .....	<b>75</b>
<b>Paper II</b> .....	<b>77</b>
<b>Paper III</b> .....	<b>79</b>
<b>Paper IV</b> .....	<b>81</b>
<b>Conclusions .....</b>	<b>83</b>
<b>Future perspectives .....</b>	<b>85</b>
<b>Acknowledgements .....</b>	<b>87</b>
<b>References .....</b>	<b>89</b>



# List of papers included in the thesis

- I. **Chen Y**, van de Vijver MJ, Hibshoosh H, Parsons R, Saal LH.  
PTEN and NEDD4 in Human Breast Carcinoma.  
*Pathology and Oncology Research*. 2016 Jan;22(1):41-7.
- II. Olsson E\*, Winter C\*, George A, **Chen Y**, Howlin J, Tang MH, Dahlgren M, Schulz R, Grabau D, van Westen D, Fernö M, Ingvar C, Rose C, Bendahl PO, Rydén L, Borg Å, Gruvberger-Saal SK, Jernström H, Saal LH.  
Serial Monitoring of Circulating Tumor DNA in Patients with Primary Breast Cancer for Detection of Occult Metastatic Disease.  
*EMBO Molecular Medicine*. 2015 Aug;7(8):1034-47.
- III. Förnvik D, Aaltonen KE, **Chen Y**, George AM, Brueffer C, Rigo R, Loman N, Saal LH, Rydén L.  
Detection of Circulating Tumor Cells and Circulating Tumor DNA Before and After Mammographic Breast Compression in a Cohort of Breast Cancer Patients Scheduled for Neoadjuvant Treatment.  
*Breast Cancer Research and Treatment*. 2019 Sep;177(2):447-455.
- IV. Pettersson L\*, **Chen Y\***, George AM, Rigo R, Lazarevic V, Juliusson G, Saal LH, Ehinger M.  
Subclonal Patterns in Follow-up of Acute Myeloid Leukemia Combining Whole Exome Sequencing and Ultrasensitive IBSAFE Digital Droplet Analysis.  
*Leukemia and Lymphoma*. 2020 Sep;61(9):2168-2179.

\* = shared first-authorship





# Other published papers

- Dahlgren M, George A, Brueffer C, Gladchuk S, **Chen Y**, Vallon-Christersson J, Hegardt C, Häkkinen J, Rydén L, Malmberg M, Larsson C, Gruvberger-Saal S, Ehinger A, Loman N, Borg Å, Saal LH. Preexisting Somatic Mutations of Estrogen Receptor Alpha (ESR1) in Early-Stage Primary Breast Cancer. *JNCI Cancer Spectrum*, 2021.
- Pettersson L, Johansson Alm S, Almstedt A, **Chen Y**, Orrsjö G, Shah-Barkhordar G, Zhou L, Kotarsky H, Vidovic K, Asp J, Lazarevic V, Saal LH, Fogelstrand L, Ehinger M. Comparison of RNA- and DNA-based methods for measurable residual disease analysis in NPM1-mutated acute myeloid leukemia. *International Journal of Laboratory Hematology*. 2021.
- Brueffer C, Gladchuk S, Winter C, Vallon-Christersson J, Hegardt C, Häkkinen J, George AM, **Chen Y**, Ehinger A, Larsson C, Loman N, Malmberg M, Rydén L, Borg Å, Saal LH. The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome. *EMBO Molecular Medicine*. 2020.
- **Chen Y**, George AM, Olsson E, Saal LH. Identification and Use of Personalized Genomic Markers for Monitoring Circulating Tumor DNA. *Methods in Molecular Biology*. 2018.
- Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, Malina J, **Chen Y**, Bendahl PO, Manjer J, Malmberg M, Larsson C, Loman N, Rydén L, Borg Å, Saal LH. Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report from the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO Precision Oncology*. 2018.

- She QB, Gruvberger-Saal SK, Maurer M, **Chen Y**, Jumppanen M, Su T, Dendy M, Lau YK, Memeo L, Horlings HM, van de Vijver MJ, Isola J, Hibshoosh H, Rosen N, Parsons R, Saal LH.  
Integrated molecular pathway analysis informs a synergistic combination therapy targeting PTEN/PI3K and EGFR pathways for basal-like breast cancer.  
*BMC Cancer*. 2016.
- Winter C, Nilsson MP, Olsson E, George AM, **Chen Y**, Kvist A, Törngren T, Vallon-Christersson J, Hegardt C, Häkkinen J, Jönsson G, Grabau D, Malmberg M, Kristoffersson U, Rehn M, Gruvberger-Saal SK, Larsson C, Borg Å, Loman N, Saal LH.  
Targeted sequencing of BRCA1 and BRCA2 across a large unselected breast cancer cohort suggests that one-third of mutations are somatic.  
*Annals of Oncology*. 2016.
- Olsson E, Winter C, George A, **Chen Y**, Törngren T, Bendahl PO, Borg Å, Gruvberger-Saal SK, Saal LH.  
Mutation Screening of 1,237 Cancer Genes across Six Model Cell Lines of Basal-Like Breast Cancer.  
*PLoS One*. 2015.
- Tang MH\*, Dahlgren M\*, Brueffer C, Tjitrowirjo T, Winter C, **Chen Y**, Ohlsson E, Wank K, Törngren T, Sjöström M, Grabau D, Bendahl PO, Rydén L, Nimeus E, Saal LH, Borg Å, Gruvberger-Saal SK.  
Remarkable similarities of chromosomal rearrangements between primary human breast cancers and matched distant metastases as revealed by whole-genome sequencing.  
*Oncotarget*. 2015.
- Alkner S\*, Tang MH\*, Brueffer C, Dahlgren M, **Chen Y**, Ohlsson E, Winter C, Baker S, Ehinger A, Rydén L, Saal LH, Fernö M, Gruvberger-Saal SK.  
Contralateral breast cancer can represent a metastatic spread of the first primary tumor: determination of clonal relationship between contralateral breast cancers using next-generation whole genome sequencing.  
*Breast Cancer Research*. 2015.

# Abstract

Cancer is the common name to a group of biologically diverse malignant neoplastic diseases. Approximately 18 million people are diagnosed with cancer annually and 8.8 million patients die from it. Tumorigenesis and progression of cancer are driven by alterations in the cancer cell genome. These alterations lead to gain of oncogenic functions, loss of tumor suppressor functions, or may be chromosomal rearrangements without obvious function, and these alterations themselves can serve as tumor-specific biomarkers that may have diagnostic and clinical utility.

In this thesis, we investigated oncogenic and tumor suppressive genes in breast cancers and leukemias, with a focus on the PTEN/PIK3CA pathway as well as minimally-invasive monitoring of cancer patients using “liquid biopsies.” We studied the underlying mechanism of PTEN protein loss in breast cancer, and showed how various types of tumor-specific mutations, including those in PIK3CA, can be used as biomarkers to monitor the dynamics of occult tumor burden, evaluate the degree of tumor content dissemination into the bloodstream with mammographic compression, and detect minimal residual disease in breast cancer and acute myeloid leukemia.

In Paper I, we found that the frequent loss of PTEN protein in human breast cancer is not attributable to the overexpression of the E3 ubiquitin ligase NEDD4, and thus NEDD4 is unlikely to be a regulator of the oncogenic PI3K/PTEN signaling pathway. In Paper II, we showed that serial monitoring of tumor specific chromosomal rearrangements, identified with low coverage whole genome sequencing and then measured in blood samples by digital PCR (dPCR), is a highly sensitive and specific approach to detect occult breast cancer disease prior to the onset of symptoms and clinical detection. Detected plasma ctDNA level was a quantitative predictor of poor relapse-free and overall survival. In Paper III, we confirmed the general safety of mammography, using FDA approved CellSearch® and our ultrasensitive mutation detection dPCR technology IBSAFE, that mammographic compression of the breast with a breast tumor does not appear to lead to significant additional dissemination of CTCs and ctDNA into the bloodstream. In Paper IV, we showed that acute myeloid leukemia specific mutations can be serially monitored in follow-up bone marrow samples by IBSAFE, providing an insight in subclonal evolution of the leukemia and the status of minimal residual disease.

These results and our mutation detection technology suggest they have high potential to be utilized in assessing treatment response, monitoring the disease course, detecting remnant tumor deposits with targetable mutations, and helping to speed the development of new drugs in the future.

# Popular summary

Cancer is one of the most prevalent and deadly disease in the modern world. It rises when healthy cells become cancerous and starts to reproduce themselves. If the human body is considered a nation, cells are its citizens, whose behaviors, in normal cases, are under strict regulation. However, if cells have minds, they might be perceived as selfish, wanting to evade regulation and become immortal. Most of such cells are brought to justice – they are programmed to die via certain biological processes, but some may survive, becoming fugitives at the beginning, making copies of themselves subsequently to form their own rogue cell gang and establish a gangland, and potentially subverting the reigning regulation of the human body. These criminal cells are cancer cells, the colony they form are cancer tumors, and the subversion of regulation leads to the death of the nation, i.e. the patient. It is believed that the underlying reason for a healthy cell to become a cancer cell is that its genome is changed in such a way that it out competes other cells to form its own population, and eventually misbehaves to harm the human body as a whole.

One cell behavior regulation scheme is called the PI3K/PTEN signaling pathway. It resides inside all cells, consisting of numerous proteins for control of its activity. When activated, the pathway instructs the cell to proliferate, and proliferation is a key hallmark of cancer. PI3K and PTEN are the key positive (tumor promoting) and negative (tumor suppressing) regulators, respectively, of this pathway. It was reported that in some cancer types, an enzyme called NEDD4 is responsible for the loss of PTEN, contributing to the activation of the PI3K/PTEN pathway, and thus cell proliferation and cancer progression. We demonstrated, in **Paper I**, that NEDD4 does not seem to cause a reduction of PTEN in breast cancer, and thus the frequently observed loss of PTEN must be explained by other mechanism. A better understanding of the biology of cancer helps cancer researchers and clinicians make better strategies on how the disease should be treated, so that outcome of cancer patients can be gradually improved.

Cancer cells are also aggressive. They do not necessarily stay at where they originate, but can migrate to other body locations and settle down if the environment fits. Even when the original colony of the cancer cells, the primary tumor, is surgically removed, as long as some cancer cells remain and are not completely eliminated by other therapies, they are likely to occupy a niche again, and this is when a cancer relapse or metastasis happens. In **Paper II**, we proved the concept that the existence of these hidden cancer cells at large can be found by detecting a

class of genomic markers, known as structural variants, that are specific to the tumor in bloodstream in early-stage breast cancer cases. The tumor-specific genomic markers were identified by sequencing the genome of the primary tumor specimen that was collected at initial surgery. The detection of cancer using the blood-based tests were highly sensitive and specific, meaning that patients with an eventual relapse had positive detection of the cancer biomarkers whereas disease-free patients did not. These blood tests also rang a bell for future metastasis many months and in some cases several years before a cancer relapse was detected by imaging technologies, potentially giving additional time for treatment and hopefully a better outcome.

Tumor-specific genomic markers of another type, point mutation, was used as surrogates of presence of cancer and to measure tumor content in **Paper III** and **IV**. Detection of point mutation is more prone to false positive results, and for this reason, we developed the IBSAFE mutation detection technology with improved sensitivity and specificity to meet the requirement. In **Paper III**, we investigated the general safety of mammography, a screening test given to women at certain risk of developing breast cancer, from the perspective of whether the mechanical compression of the breast involved causes the dissemination of tumor content into the bloodstream. The US FDA approved method CellSearch was used for counting the number of circulating tumor cells and IBSAFE was used for detection of tumor-specific point mutations. We found that compression of the breast with a tumor during mammography does not cause a significantly elevated level of circulating tumor cells or circulating tumor DNA and therefore can be considered safe in this aspect. In **Paper IV**, we moved onto another type of cancer, acute myeloid leukemia (AML), and tried the IBSAFE method in detection of minimal residual disease (MRD) in AML patients. Lacking an identifiable solid tumor, surgery cannot be done on most AML patients, making eradication of all tumor cells extraordinarily difficult. MRD is the main readout of AML relapse, and existing methods of MRD detection all have their limitations. We demonstrated that IBSAFE can detect AML specific mutations in follow-up bone marrow samples with good sensitivity and specificity, and therefore shows promise in the detection of MRD in the clinical routine.

We hope our knowledge of cancer biology will ever grow to help better treat the disease, and our mutation detection technology as well as future methods serving other scientific and clinical purposes will be useful in improving the outcomes and survival of patients with all forms of cancer.

# Abbreviations

<b>ADC</b>	Antibody-drug conjugate
<b>AJCC</b>	American joint committee on cancer
<b>ALL</b>	Acute lymphocytic leukemia
<b>AML</b>	Acute myeloid leukemia
<b>APC</b>	Allophycocyanin
<b>ARMS</b>	Amplification refractory mutation system
<b>BCT</b>	Blood collection tube
<b>BY-SA</b>	Attribution-ShareAlike
<b>CA</b>	Cancer antigen
<b>CBC</b>	Complete blood count
<b>CC</b>	Creative commons
<b>CDK</b>	Cyclin-dependent kinase
<b>cDNA</b>	Complementary DNA
<b>CEA</b>	Carcinoembryonic antigen
<b>cfDNA or ccfDNA</b>	Circulating cell-free DNA
<b>CK</b>	Cytokeratin
<b>CLL</b>	Chronic lymphocytic leukemia
<b>CLSI</b>	Clinical and laboratory standards institute
<b>CML</b>	Chronic myeloid leukemia
<b>CNA</b>	Circulating nucleic acid
<b>CNV</b>	Copy number variation
<b>COSMIC</b>	Catalogue of somatic mutations in cancer



<b>CR</b>	Complete remission
<b>CT</b>	Computerized tomography
<b>CTC</b>	Circulating tumor cell
<b>ctDNA</b>	Circulating tumor DNA
<b>DAPI</b>	4',6-diamidino-2-phenylindole
<b>DCIS</b>	Ductal carcinoma <i>in situ</i>
<b>ddNTP</b>	Dideoxynucleoside triphosphate
<b>ddPCR</b>	Droplet digital PCR
<b>DF</b>	Disease-free
<b>DNA</b>	Deoxyribonucleic acid
<b>DNA-seq</b>	DNA sequencing
<b>dNTP</b>	Deoxynucleoside triphosphate, including <b>dATP</b> – Deoxyadenosine triphosphate <b>dCTP</b> – Deoxycytidine triphosphate <b>dGTP</b> – Deoxyguanosine triphosphate <b>dTTP</b> – Deoxythymidine triphosphate <b>dUTP</b> – Deoxyuridine triphosphate
<b>dPCR</b>	Digital PCR
<b>EDTA</b>	Ethylenediaminetetraacetic acid
<b>EGFR</b>	Epidermal growth factor receptor
<b>EM</b>	Eventual metastatic
<b>EpCAM</b>	Epithelial cell adhesion molecule
<b>ER</b>	Estrogen receptor
<b>EtBr</b>	Ethidium bromide
<b>FDA</b>	Food and drug administration
<b>FFPE</b>	Formalin-fixed paraffin-embedded
<b>FISH</b>	Fluorescence <i>in situ</i> hybridization
<b>FS</b>	Forward scatter
<b>GCO</b>	Global Cancer Observatory

<b>HER2</b>	Human epidermal growth factor receptor 2
<b>HLA</b>	Human leukocyte antigen
<b>HSC</b>	Hematopoietic stem cell
<b>HTS</b>	High-throughput sequencing, also known as <b>NGS</b> – Next-generation sequencing <b>SGS</b> – Second-generation sequencing <b>MPS</b> – Massive(ly) parallel sequencing
<b>IDC - NOS</b>	Invasive ductal carcinoma - not otherwise specified
<b>IGV</b>	Integrative genomics viewer
<b>IHC</b>	Immunohistochemistry
<b>ILC</b>	Invasive lobular carcinoma
<b>Indel</b>	Short insertion or deletion
<b>ITD</b>	Internal tandem duplication
<b>LAIP</b>	Leukemia associated immunophenotype
<b>LCA</b>	Leukocyte common antigen
<b>LCIS</b>	Lobular carcinoma <i>in situ</i>
<b>LNA</b>	Locked nucleic acid
<b>LoB</b>	Limit of blank
<b>LoD</b>	Limit of detection
<b>LOH</b>	Loss of heterozygosity
<b>LoQ</b>	Limit of quantitation
<b>MDS</b>	Myelodysplastic syndrome
<b>MFC</b>	Multicolor flow cytometry
<b>MGB</b>	Minor groove binder
<b>MNC</b>	Mononuclear cell
<b>MRD</b>	Minimal/measurable residual disease
<b>MRI</b>	Magnetic resonance imaging
<b>mRNA</b>	Messenger RNA
<b>MSD</b>	Matched sibling donor

<b>MUD</b>	Matched unrelated donor
<b>NEDD4</b>	Neural precursor cell expressed, developmentally down-regulated 4
<b>NFQ</b>	Nonfluorescent quencher
<b>NHG</b>	Nottingham histological grade
<b>NKI</b>	The Netherlands Cancer Institute
<b>PCR</b>	Polymerase chain reaction
<b>PD-1</b>	Programmed cell death protein 1
<b>PD-L1</b>	Programmed cell death ligand 1
<b>PE</b>	Phycoerythrin
<b>PET</b>	Positron emission tomography
<b>PI3K</b>	Phosphoinositide 3-kinase
<b>PIP2</b>	Phosphatidylinositol (4,5)-biphosphate
<b>PIP3</b>	Phosphatidylinositol (3,4,5)-triphosphate
<b>PMF</b>	Probability mass function
<b>PR or PgR</b>	Progesterone receptor
<b>PTEN</b>	Phosphatase and tensin homolog
<b>qPCR</b>	Quantitative PCR
<b>r<sub>cf</sub></b>	Relative centrifugal force
<b>Real time RT-PCR or rRT-PCR</b>	Real time reverse transcription PCR
<b>RNA</b>	Ribonucleic acid
<b>RNA-seq</b>	RNA sequencing
<b>ROC</b>	Receiver operating characteristic
<b>RTK</b>	Receptor tyrosine kinase
<b>SBS</b>	Sequencing by synthesis
<b>SCAN-B</b>	Sweden Cancerome Analysis Network - Breast
<b>SCT</b>	Stem cell transplantation
	<b>Allo-SCT</b> – Allogeneic stem cell transplantation
	<b>Auto-SCT</b> - Autologous stem cell transplantation

<b>SERM</b>	Selective estrogen receptor modulator
<b>SNV</b>	Single nucleotide variant
<b>SRA</b>	Sequence read archive
<b>SS</b>	Side scatter
<b>SV</b>	Structural variant
<b>TCGA</b>	The cancer genome atlas
<b>TGS</b>	Third generation sequencing
<b>TMA</b>	Tissue microarray
<b>TNBC</b>	Triple negative breast cancer
<b>TNM</b>	Tumor, node, metastasis staging system
<b>UDG</b>	Uracil-DNA glycosylase
<b>UICC</b>	Union for international cancer control
<b>UMI</b>	Unique molecular identifier
<b>VAF</b>	Variant allele frequency, also known as <b>MAF</b> – Mutant allele frequency, in most cases
<b>WBC</b>	White blood cell
<b>WES</b>	Whole exome sequencing
<b>WGS</b>	Whole genome sequencing

Names of genes are referred to by their gene symbols according to the HUGO Gene Nomenclature Committee (<https://www.genenames.org>).



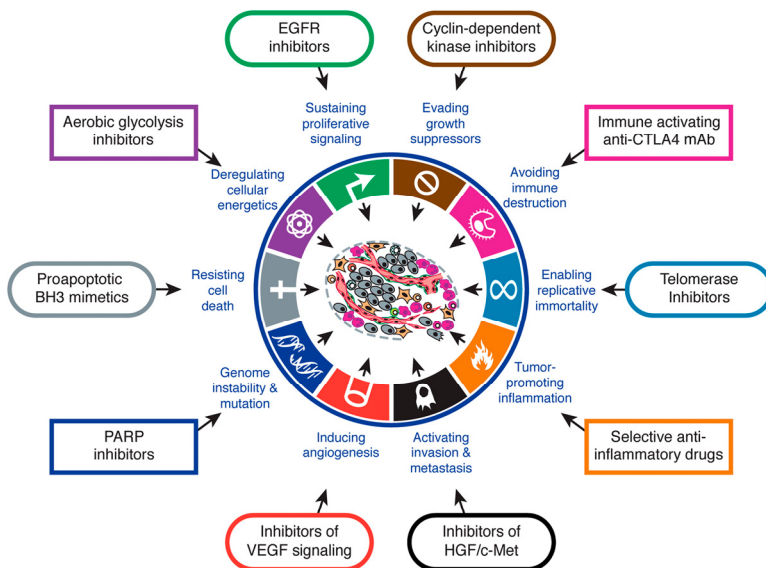
# Introduction

Cancer is the common name given to a variety of biologically diverse malignant neoplastic diseases. The etymology of the word *cancer* in English dates to ancient Greek, in which the original word *karkinos* means *crab* and later *tumor*, for Hippocrates, “father of medicine”, and his peers noted the resemblance between the swollen veins of tumors and the many legs of the arthropod. When the Greek word morphed into its Latin cognate, *cancer* was coined [1]. As for in the Chinese language, it has been considered the past few decades that the character for *cancer*, 癌 (ái), was a 19<sup>th</sup> century Japan-made kanji introduced back into Chinese, as a result of the language running out of existing words for new objects when Western texts of modern concepts were translated into the language en masse. Recently, however, classical Chinese literature scholars found that this word had actually been used in some 12<sup>th</sup> century traditional medicine notes, except it stood for another disease, probably abscess, and not cancer by today’s definition.

Historical records and paleopathological archeology have provided evidence that cancer has been plaguing all ethnicities since the dawn of human civilization [2-6]. In fact, cancer seems to be an exclusive disease for all vertebrate animals, as cancer cases (non-laboratory-induced) have been reported in almost all classes of the vertebrate subphylum except, perhaps, for amphibians, and no cancer disease entity has been observed in other living organisms [7].

As molecular genetics reveals, alterations in morphology, characteristics, and behavior of a cell, also known as phenotype, can often be directly attributable to alterations in its genetic material, also known as genotype, when the variable of environment remains the same. A large volume of studies have found strong correlations and/or causations between genetic mutations and cancer, such as *BRCA1/BRCA2* mutations in hereditary breast cancer [8, 9], *TP53* mutations in multiple cancer types [10], and *BCR-ABL* gene fusion, also known as the Philadelphia chromosome, in chronic myeloid leukemia (CML) [11], indicating cancer is a genetic disease [12]. Though functional studies are constantly being done to elucidate roles and regulation of genes in cellular information transduction cascades, called cell signaling pathways, and mechanisms underlying the interactions between different cell signaling pathways, called crosstalk, comprehensive analysis of individual cancer genomes has not been an easy task until early 2000’s, when the first drafts of human genome were assembled [13, 14]. With the advent of new technologies and accumulation of knowledge, Hanahan and

Weinberg summarized in 2000 [15] and 2011 [16] the commonalities of all cancers, which they called the hallmarks of cancer (Figure 1). These hallmarks enable the cell to circumvent growth control and programmed death, become nutritionally self-sufficient through sustained angiogenesis and thus achieve clonal immortality, evade the surveillance of the immune system, and eventually spread to nearby tissues through invasion, and to distant tissues through metastasis.



**Figure 1 Hallmarks of cancer and their example therapeutic agents**

From Hanahan, D. and Weinberg R.A., *Hallmarks of cancer: the next generation*. Cell, 2011. **144**(5): p. 646-74. Reprinted with permission from Elsevier Science & Technology Journals.

Based on the cell types of origin, most cancers fit into one of the four major types: 1) carcinoma, originating from epithelial cells lining the inner or outer surfaces of the body; 2) sarcoma, originating from non-hematopoietic mesenchymal cells that form the connective tissues; 3) leukemia, originating from hematopoietic cells that mature in blood; and 4) lymphoma, originating from hematopoietic cells that mature in the lymphatic system. Significant cancer types that do not fit into any of the four major types above also exist, such as melanoma of the skin, originating from melanocytes, and glioblastoma with unknown cell of origin.

# Mutations in the cancer genome

In a broad sense, all alterations in DNA sequence of the genome can be considered as a mutation. In this regard, mutations in cancer genome can be categorized in different ways as follows.

Depending on the type of the cells harboring the mutation, mutations can be divided into 1) germline mutations, which are in reproductive cells that can be inherited from parents and/or passed on to descendants, and 2) somatic mutations, which are in non-reproductive cells, and often occur sporadically from DNA replication error, DNA repair defects, and/or mutagen inductions.

Several germline mutations have been identified as strong hereditary cancer risk factors and drivers, for example, *BRCA1* and *BRCA2* mutations in hereditary breast cancer and ovarian cancer [8, 9, 17-20], *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM* mutations in hereditary nonpolyposis colorectal cancer (also known as Lynch syndrome) [21, 22], and *TP53* mutations in Li-Fraumeni syndrome which increases the risk of developing multiple primary cancers [23, 24].

Nonhereditary cancers, which comprises the majority of cancer cases, are predominantly caused by somatic mutations acquired postnatally [25] and are also termed ‘sporadic’ cancers. As cancer cells develop into tumors by gaining selective growth advantage over normal cells, the somatic mutations contributing to this advantage are called “driver” mutations, and those that are effectively neutral are called “passenger” mutations. It has been estimated that passenger mutations outnumber driver mutations by several orders of magnitude in the course of tumor development and progression due to their relatively small impact on natural selection of the cancer cell population [26]; however, accumulation of passenger mutations may still contribute to cancer progression [27, 28]. Despite their relative contribution to carcinogenesis and cancer progression, both driver and passenger mutations can be used as tumor specific biomarkers for monitoring of tumor burden, detecting occult metastatic disease, measuring response to treatments, and quantifying minimal residual disease.

From gene function point of view, there are 1) gain-of-function mutations, also known as activating mutations, and 2) loss-of-function mutations, also known as inactivating mutations.

In cancer cells, gain-of-function mutations often occur in proto-oncogenes such as *PIK3CA*, *AKT1*, *EGFR*, *KRAS*, and *BRAF*, which, in normal physiological conditions, are responsible for cell growth, proliferation, or survival. When mutations that upregulate these functions occur in a proto-oncogene and the cell evades programmed death (called apoptosis) the proto-oncogene turns into an oncogene, contributing to carcinogenesis [29].



On the other hand, loss-of function mutations often occur in tumor suppressor genes such as *TP53*, *PTEN*, *RBI*, *APC*, and *BCL2*, which regulate the cell cycle at checkpoints, repair damaged DNA, or induce apoptosis [30]. As somatic cells are diploid, sporadic loss-of-function mutations only inactivate one allele of the gene, with the other allele's function unaltered. Therefore, unlike gain-of-function mutations in oncogenes which lead to dominant traits, loss-of-function mutations in tumor suppressor genes are usually haploinsufficient, and often require deletion or mutation of the other allele for the phenotype to be altered. The deletion of the other allele and its surrounding region is called loss of heterozygosity (LOH), and LOH has been widely observed in breast cancer [31, 32], lung cancer [33, 34], leukemia [35], ovarian cancer [36], prostate cancer [37, 38], and gastric cancer [39]. LOH events also have been used to map the genomic locations of tumor suppressor genes [40-42].

From a genomic structure point of view, mutations occur at large and small genomic scales. Small-scaled mutations include 1) single nucleotide variants (SNVs), also known as point mutations, which substitutes one nucleotide with another, and 2) small insertions and deletions, also called indels, which insert or delete a short fragment of DNA, usually below 50 base pairs. Small-scaled mutations can happen at any genomic location. When occurring in regions translated into protein products, they are called coding mutations, otherwise they are non-coding mutations. For coding SNVs, the reading frame of the DNA always remains the same. Therefore, the SNVs that lead to an amino acid change in the protein product are called non-synonymous SNVs, and those that do not change the amino acid are called synonymous SNVs, also known as silent SNVs or silent point mutations. Depending on what the original amino acid is changed to, non-synonymous SNVs can be further identified as missense or nonsense SNVs, with the former having a codon for an alternative amino acid created by the mutation, and the latter having a stop codon created, which truncates the protein product. For coding indels, not only is the amino acid sequence of the protein product changed, but also a shift in reading frame may happen when the number of inserted or deleted nucleotide is not a multiple of three (number of nucleotides in a codon). Therefore, when the number of bases inserted or deleted is divisible by 3, the indel is an in-frame insertion or deletion, otherwise it is a frameshift insertion or deletion. In-frame small indels lead to insertion or deletion of a certain number of amino acids in the protein product of the gene, resulting in a relatively mild alteration of the protein function, whereas frameshift small indels change all the amino acids after them, and usually introduce a stop codon, causing a premature termination of the protein product and typically a devastating change in the protein's function. Several small-scaled mutation events of different types can happen at the same locus, resulting in complex mutations. Many somatic SNVs and small indels have been repeatedly found in different types of cancer, and online databases like Catalogue Of Somatic Mutations In Cancer (COSMIC) have been established as a resource for future cancer research [43]. Although DNA sequence alterations in coding regions of the genome directly

translate into altered protein products, transcription of DNA into mRNA and translation of mRNA into protein are regulated by a complex web of mechanisms. Non-coding mutations in, for example, cis- and trans- regulatory elements and microRNA loci may be critical drivers in various cancer types [44-46].

Large-scaled mutations can occur across distant genomic regions. Chromosomal rearrangements, a major type of structural variant (SV), are typical large-scale mutations. In sequencing data, by inspecting the junction sequences, chromosomal rearrangements can be stratified into inter-chromosomal rearrangements, wherein the two ends of the junction map to different chromosomes, and intra-chromosomal rearrangements, where both sides of the junction map to the same chromosome. Within the group of intra-chromosomal rearrangements, after aligned to the reference genome, if the split reads from a sequencing read pair point to each other, it is a deletion, and if both reads point the same direction, it is an inversion. When the reads point away from each other, an insertion or duplication could have happened, but without a whole picture of the consecutive up and downstream DNA sequence, the exact type of the rearrangement remains uncertain. Chromosomal rearrangements could cause copy number variation (CNV) and gene fusion. CNVs arise as a result of nonreciprocal chromosomal rearrangement, whereas when the two chromosomal segments involved in the rearrangement simply swap their positions, the event is reciprocal, and each rearranged locus is copy number neutral. Gene fusion forms when parts of two protein coding genes are spliced together and a chimeric protein is synthesized. Both copy number variation [47-50] and gene fusion [51, 52] can have deleterious effects in cancer. Other types of chromosomal aberrations that can be spotted with cytogenetics methods or sequencing include aneuploidy, where an entire chromosome is present at an abnormal number, isochromosome, where one part of a chromosome is lost while the rest part is copied as a mirror image, ring chromosome, where the two ends of a chromosome fuse to form a ring, etc. Together with SNVs and small indels, chromosomal rearrangements can often be passenger events, but can also be drivers that initiate the development of cancer and its progression [53-56].

In addition to mutations in the genome, abnormal phenotypes can also pass through cancer cell generations without changes in DNA sequence. These are studied by cancer epigenetics. Epigenetic mechanisms contributing to the change of phenotype in cancer cells include methylation in gene regulatory regions, histone modifications, effects of miRNA in sequestering or inhibiting transcripts, mRNA trans-splicing, post-transcriptional modifications of the mRNA, and post-translational modification of the protein [57-59].

By definition, mutations are classified as ‘somatic’ if they are only found in the cancer cell genome but not the normal cell genome of the patient, and thus can be used as tumor-specific biomarkers to monitor the progression of the tumor. Ultrasensitive molecular monitoring of tumor-specific mutations of different types was performed in **Papers II, III, and IV**.

# Breast cancer

Though men could also develop breast cancer, ~99% of breast cancer cases are diagnosed in women [60]. Despite the nearly exclusive presence in women, breast cancer is still the most diagnosed cancer globally. According to the Global Cancer Observatory (GCO), in 2020 an estimated 2.3 million new breast cancer cases were diagnosed in women, accounting for 11.7% of all cancer cases in both men and women. Not only does breast cancer lead in cancer incidence, but also it is the leading cause of cancer mortality in women, with 685,000 women estimated to have died from the disease in 2020 worldwide. The total mortality of breast cancer is only surpassed by that of lung cancer, liver cancer, stomach cancer, and colorectal cancer [61]. In Sweden, according to the Swedish National Board of Health and Welfare, Socialstyrelsen, the incidence of breast cancer in the country was over 150 per 100,000 women per year in 2019, with 8,288 new cases and 1,353 deaths registered [62]. This incidence means that about 1 in 8 women on average is expected to develop breast cancer in her lifetime in Sweden, which is similar to the rate in most Western countries.

## Risk factors

The development of breast cancer is influenced by a combination of intrinsic and extrinsic reasons known as risk factors. Intrinsic risk factors are inherent to individuals at a given time point, whereas extrinsic risk factors are from outside the body and generally are modifiable during one's lifetime.

Aside from being born as a woman, age is one of the most major intrinsic risk factors for breast cancer. As mutations rising from DNA replication error accumulate over time, the likelihood for some cells to gain selective advantage and become malignant is proportional to the person's age. This theory is corroborated by the observation that the probability of developing breast cancer in women < 49 years old was 2% in the United States between 2013 and 2015, as opposed to 6.7% in women ≥ 70 years old [63]. Another intrinsic risk factor, at least in the US, is related to ethnicity, although this may be confounded by socioeconomic factors. Non-Hispanic white women are most have the highest incidence (130.1 per 100,000) followed by black (126.5), while the mortality rate in black people (28.9 per 100,000) is much higher than all other ethnic groups [63]. Family history is also a very strong intrinsic risk factor. About 13-16% of breast cancer patients have a first-degree female relative with breast cancer. The risk for a woman to develop breast cancer, when she has 1, 2, and 3 or more first-degree female relatives diagnosed with the disease, is 1.80, 2.93, and 3.90 times as high as that for a woman who does not have such relatives [64]. About 20% of hereditary breast cancers are believed to be directly caused by high-risk germline mutations in *BRCA1* or *BRCA2*. Germline mutations in other genes such as *PTEN*, *TP53*, *STK11*, *CHEK2*, *ATM*, *BRIP1*, and

*PALB2* have also been found in a significant number of familial breast cancer cases. Genetic testing of *BRCA1/2* may be offered to higher-risk individuals according to different screening criteria per country [65]. Intrinsic reproductive factors such as early menarche, late menopause, late first pregnancy, and low parity can also increase the risk for breast cancer [66].

Extrinsic risk factors include 1) lifestyle-related risk factors, such as obesity, alcohol consumption, active and passive smoking, dietary consumption of meat, lack of physical exercise, lack of vitamin D intake; 2) hormonal factors, such as hormonal contraception and postmenopausal hormone replacement therapy; 3) environmental factors, such as air pollution and ionizing radiation; and 4) other factors, such as being diabetic and poor socioeconomic status [67].

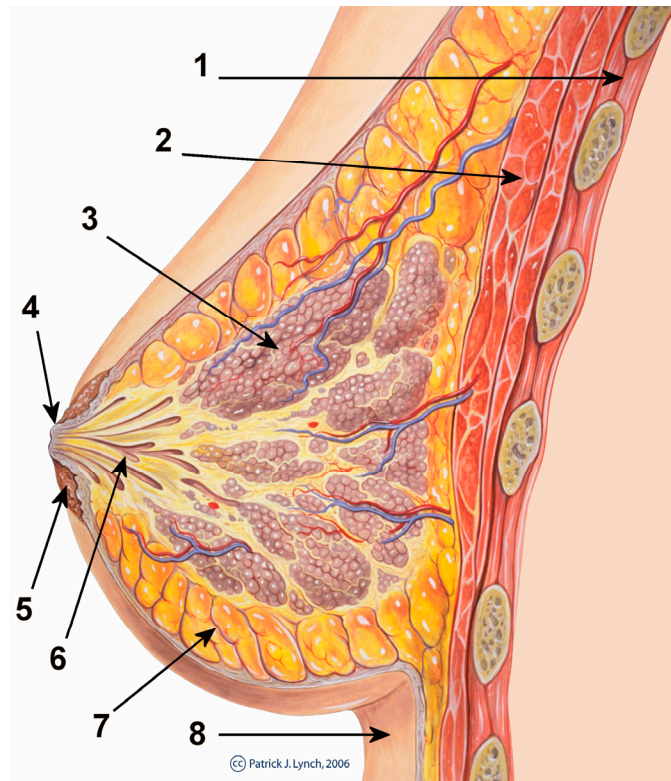
## Diagnosis

Screening for breast cancer is done with mammography, a low dose X-ray test to examine the breast tissues. In Sweden, women between 40 and 74 years old are offered mammography exams about once every other year [68]. The aim of the screening is to find early onset of breast cancer, so that further diagnostic tests and potential treatments could start as soon as possible to improve the outcome. However, it has been debated that many *in situ* carcinomas or other types of breast abnormalities may never develop into invasive cancers, and thus the benefit of early detection of breast cancer by mammography may be offset by the detrimental effect from overtreatment caused by false positive results and treatment of harmless lesions [69-71]. Other than screening, when a woman is aware of signs or symptoms of breast cancer, for example from self breast exam, she may turn to a doctor for diagnostic tests. Depending on the clinical situation, imaging studies of the breast are performed, such as ultrasound, magnetic resonance imaging (MRI), computerized tomography (CT) scans, and positron emission tomography (PET) scans. In addition, if a suspicious area is spotted during imaging, a small piece of the specimen may be removed from the site via a process called core needle biopsy. The specimen can be histopathologically examined and/or molecularly tested to give a definitive diagnosis of breast cancer.

## Histopathology

Illustrated in Figure 2, the human female breast is mainly made up of glandular tissue (the lobules and the duct) and supportive tissue (fatty tissue that supports the structure). The glandular tissue with an epithelial origin is where breast cancers predominantly develop, therefore, most breast cancers are carcinomas. Depending on whether the disease is localized or has invaded into nearby tissues, the carcinomas can be divided into carcinoma *in situ*, comprising 15-30% of all breast carcinomas, or invasive carcinoma, comprising the rest 70-85%. Carcinoma *in situ*

consists of ductal carcinoma *in situ* (DCIS, ~80%) and lobular carcinoma *in situ* (LCIS, ~20%) [72]. Since the malignancy is localized, *in situ* carcinomas are often regarded as benign, but studies have shown that they can increase the risk of future invasive carcinomas, and therefore follow-up care and/or localized treatment may be advisable for patients diagnosed with these lesions [73-75]. Within invasive carcinoma, ~80% are invasive ductal carcinoma (IDC) of not otherwise specified (NOS) type and ~10% are invasive lobular carcinoma (ILC). The remaining invasive carcinomas include more rare forms, such as tubular, cribriform, medullary, mucinous, papillary or micropapillary carcinomas [76].



**Figure 2 Anatomy of the human female breast.** 1. Chest wall, 2. Pectoralis muscles, 3. Lobules, 4. Nipple, 5. Areola, 6. Lactiferous duct, 7. Fatty tissue, 8. Skin

From [https://commons.wikimedia.org/wiki/File:Breast\\_anatomy\\_normal\\_scheme.png](https://commons.wikimedia.org/wiki/File:Breast_anatomy_normal_scheme.png) CC BY-SA 3.0

## Tissue biomarker status and breast cancer classification

Breast cancer is a heterogenous disease with diverse biology. Robust classification of breast cancer has prognostic and predicative value and guides patient decision making and the selection of treatment strategies.

Certain cell surface receptors and nuclear proteins dictate the behavior of tumor cells to a great extent. Routinely examined cell surface markers are estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2. Estrogen receptor (ER), encoded by the *ESR1* gene, and progesterone receptor (PgR or PR), encoded by the *PGR* gene, are hormone receptors and are major drivers of breast cancer cell survival and proliferation. Human epidermal growth factor receptor 2 (HER2), encoded by *ERBB2* gene, is a receptor tyrosine kinase, which dimerizes with other epidermal growth factor receptors upon binding of a ligand, switching on the downstream proliferation and cell growth signaling transduction. In recent years, the nuclear protein Ki-67, encoded by the *MKI67* gene, is routinely analyzed as a surrogate marker of cell proliferation [77, 78]. The statuses of these markers are used to dictate treatment, for example to endocrine regimens such as tamoxifen for ER-positive disease, and targeted treatments such as trastuzumab for HER2-positive disease. The laboratory method predominantly used to determine the status of these markers is immunohistochemistry (IHC) staining, detailed in the Methods section.

Based on the status of these markers, breast cancers can be classified into four major subtypes: 1) the luminal A subtype (ER+ and/or PR+, HER2- and Ki-67 low), 2) the luminal B subtype (ER+ and/or PR+, HER2+ and/or Ki-67 high), 3) the HER2-positive subtype (ER-, PR-, and HER2+), and 4) the triple negative breast cancer (TNBC) subtype (ER-, PR-, and HER2-) [79], each corresponding to different treatment options and prognosis.

In addition to subtyping, breast cancers are graded and staged.

Grading describes how morphologically similar the tumor cells are to normal cells. The most commonly used breast cancer grading system is the Nottingham Histological Grade (NHG), modified from the Scarff-Bloom-Richardson system [80]. Scores from 1 to 3 are given by pathologists to three categories: tubule formation, nuclear pleomorphism, and mitotic activity. Grade I are assigned to tumors with a total score between 3 and 5, Grade II to 6-7 scored tumors and Grade III to 8-9 scored tumors. The Grade I, II, and III tumors are also known as well differentiated, moderately differentiated, and poorly differentiated breast tumors, and increasing grade is strongly associated to worse prognosis.

Staging describes how advanced the disease is. Breast cancer stages are evaluated using a so-called TNM scale, where the T stands for tumor size, N stands for lymph nodes, and M stands for distant metastasis. The TNM staging system is proposed by the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC). Scores for each category and their conversion into stages I to IV are shown in Table 1. Similar to grade, increasing stage is strongly associated to worse prognosis.

Gene expression measured by DNA microarray and sequencing in the past two decades has enabled a modern molecular classification of breast cancers. For

example, by unsupervised hierarchical clustering on global gene expression profile, intrinsic subtypes of luminal A, luminal B, ERBB2+, basal-like, and normal-like subtypes were identified and associated with different outcomes [81, 82]. The histologic status of ER, PR, HER2, Ki-67 could also be accurately reproduced by measuring the expression levels of their coding genes *ESR1*, *PGR*, *ERBB2*, and *MKI67* from RNA-seq results [83]. Since many hospitals do not have access to gene expression profiling, the IHC surrogates for the molecular subtypes were developed (as described above).

**Table 1. TNM staging of breast cancer**

Stage	TMN score	Interpretation
0	Tis B0 M0	Pre-invasive stage
I	T1 N0 M0	Low stage
	T0 N1mi M0	
	T1 N1mi M0	
II	T0 N1 M0	Intermediate stage
	T1 N1 M0	
	T2 N0 M0	
	T2 N1 M0	
III	T3 N0 M0	High stage
	T0 N2 M0	
	T1 N2 M0	
	T2 N2 M0	
	T3 N1 M0	
	T3 N2 M0	
IV	T4 N0 M0	Metastatic stage
	T4 N2 M0	
	Any T N3 M0	
IV	Any T Any N M1	Metastatic stage

## Treatment

Local and systemic treatment options are available for breast cancer. The primary local treatment is surgery, including 1) removal of the primary tumor and its surrounding normal tissue while keeping the rest of the breast intact, called lumpectomy, and 2) removal of the entire breast, called mastectomy. It is worth noting that prophylactic mastectomy could also be electively performed on women with a family history and confirmed to be a carrier of a *BRCA1/BRCA2* mutation, in order to reduce the risk of developing breast cancer in the future. Depending on characteristics of the tumor, drugs can be administered before the surgery to shrink the tumor in size, or after the surgery to cleanse any remnant of the tumor. These therapeutic options are called neoadjuvant therapy and adjuvant therapy, respectively. Another local treatment option is radiation therapy, which utilizes ionizing radiation on the primary cancer site to kill any remnant tumor cells. Radiation therapy is often given to women after lumpectomy but not mastectomy.

Conventional systemic treatments include 1) chemotherapy, 2) hormone therapy, also known as endocrine therapy, and 3) targeted therapy. All these therapies involve anti-cancer pharmaceuticals delivered, in most cases, intravenously, but they function with different mechanisms. Chemotherapy, nowadays, specifically means the use of non-specific intracellular cytotoxic agents that produce severe DNA damage, impede DNA/RNA synthesis, or thwart formation of microtubules, thereby killing cells which are highly proliferative. Doxorubicin, epirubicin, paclitaxel, docetaxel, 5-fluorouracil, cyclophosphamide, and carboplatin, are common chemotherapy drugs used to treat breast cancer. Hormone therapy, on the other hand, is used for tumors that express ER and consists of full or partial hormone receptor antagonists to block the response and activity of the hormone receptors. Selective estrogen receptor modulators (SERMs) such as tamoxifen, raloxifene, and toremifene are hormone therapy agents commonly administered in breast cancer patients. As for targeted therapy, rather than simply interacting with rapidly proliferating cells, the agents target molecules that are specific to the tumor cells, paralyzing their growth and proliferation while having relatively low effect against normal cells. A frequently used targeted therapy agent is trastuzumab, a monoclonal antibody that binds to HER2 and induces internalization and recycling of the receptor, reducing cell growth and proliferation triggered by HER2's heterodimerization. Cyclin-dependent kinases (CDKs) are catalytic subunits of protein kinase complexes responsible for regulation of cell cycle and are often overactivated in certain subtypes of breast cancer, resulting in cell proliferation and tumor progression [84]. CDK4/6 inhibitors such as palbociclib, ribociclib (US FDA approved for treatment of ER+/HER2- advanced breast cancer), and abemaciclib have gone through stages of clinical trials and been used, to different extents, as targeted therapy agents in recent years [85-87]. As TNBC does not express the hormone receptors nor HER2, nonspecific chemotherapy was conventionally one of the few pharmacotherapy options left for this subtype of breast cancer.

It is worth mentioning that cancer cell phenotypes are subject to change, so that tumors once susceptible to certain pharmaceuticals may become insensitive after a period of treatment. This phenomenon is called drug resistance. Recently, immunotherapy has become another promising option for treatment of breast cancer, especially TNBC. The programmed cell death protein 1 (PD-1, encoded by *PDCDI*) and its ligand (PD-L1, encoded by *CD274*) function as suppressors of the adaptive immune system in physiological conditions, and are upregulated in certain cancers so that the tumor cells can evade anti-tumor immunity [88, 89]. Blockade of PD-1 by pembrolizumab, a humanized monoclonal antibody, has been shown to be an effective immunotherapy option for advanced melanoma [90, 91], metastatic lung cancer [92-94], and advanced urothelial cancer [95]. Clinical trials evaluating the efficacy of pembrolizumab in TNBC [96, 97] as well as other anti-PD-1 antibodies have been committed [98-100]. Not only could pharmacotherapies be administered independently, but also multiple pharmaceuticals may work synergistically. Combined administration of drugs with synergy can significantly



reduce the effective dose to minimize side effects, avoid or postpone drug resistance, and maximize treatment response.

Due to the screening options and generally early diagnosis, comprehensive classification systems, and advanced treatment options, the overall 5-year survival rate of breast cancer, with all stages combined, is outstanding at 90% in the United States between 2008 and 2014 [63]. However, despite this high 5-year survival, it is less appreciated that breast cancer can have late relapses that occur as late as 15-20 years after diagnosis; and metastatic disease is essentially incurable [101-104]. On the other hand, overdiagnosis and overtreatment are also significant in breast cancer, bringing treatment side-effects and unnecessary anxiety to patients [105, 106]. Robust surveillance methods to monitor the occult tumor burden, or the lack of it, therefore, is highly desirable for improvement of clinical management and outcome of breast cancer.

### Minimally-invasive liquid biopsy

One way to detect the occult tumor burden, which may be too small to be detected by imaging technologies, is to look for cancer biomarkers in the blood circulation using a minimally-invasive “liquid biopsy”. Samples of 10-20 mL of whole blood may be taken during follow-up visits for testing.

A class of protein biomarkers known as cancer antigens (CAs) can be examined in the blood samples. Particularly, CA 125 is recognized as an ovarian cancer specific biomarker, whose serum concentration is elevated with an increased ovarian cancer burden [107]. Alongside other serum antigens, the diagnostic value of CA 125 has also been evaluated in metastatic breast cancer, and correlations were observed between elevated concentrations of one or several of the cancer antigens in serum and breast cancer metastasized to different body locations [108]. Other cancer antigens such as CA 15-3 and carcinoembryonic antigen (CEA) have been proposed as prognostic markers for primary, and monitoring markers for metastatic breast cancers [109-111]. However, the sensitivity and specificity of CA testing as a surrogate for occult tumor burden is generally poor, in that not only different individuals have different baseline CA levels, but also lots of other physiological and pathological processes can cause fluctuation and alteration of CA levels [112].

Circulation tumor cell (CTC) is another liquid biopsy cancer biomarker. A widely recognized model of distant metastasis proposes that tumor cells from the primary tumor invade and disrupt the basement membrane, and can extravasate and travel to and colonize a distant body tissue via blood and/or lymphatic vessels, adapt to the local microenvironment, and eventually form a metastatic tumor [113, 114]. CTC concentration in peripheral blood may have prognostic value in primary [115-117] and metastatic breast cancer [118], and clusters of CTCs, rather than single CTCs, were found to be a major predictor of a potential metastatic disease [119]. One

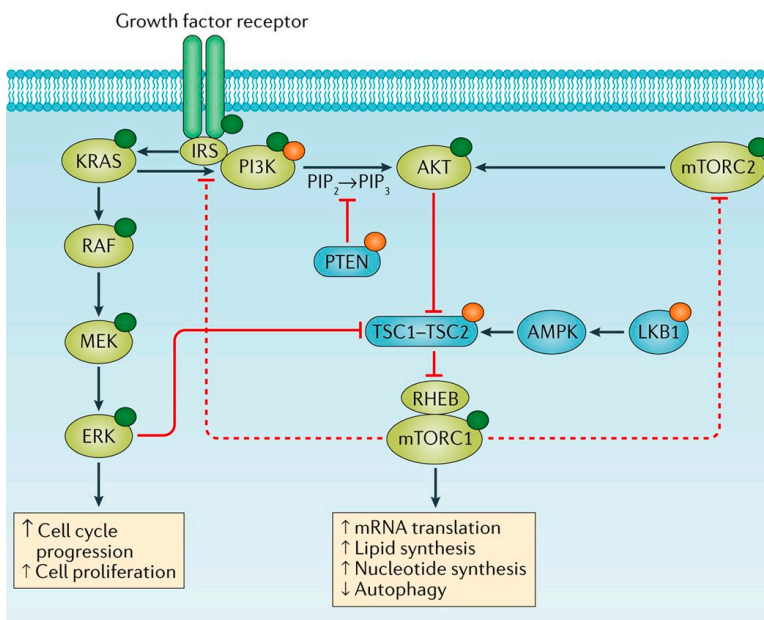
method to enumerate CTCs have been established and validated for clinical use, notably the CellSearch<sup>®</sup> system approved by the US Food and Drug Administration (FDA) [120], detailed in the Methods section. To date, though, detection of CTCs in early-stage breast cancer remains a challenge, in that either no tumor cell is disseminated to the circulatory system, or that the number of CTC can be so low per volume of blood that they are simply not sampled due to sampling error [121]. The inability to sample rare CTCs may be alleviated to some extent by drawing the blood not from a vein in the arm, but from the superior vena cava via, for example, a central venous portal [122], but this practice increases inconveniency in many cases, and the degree of improvement in CTC counts can be marginal. Moreover, an increased CTC count may not necessarily indicate a progression or cellular relapse of the cancer but may attribute to mechanical pressure from clinical interventions such as surgery [123-127]. The robustness of CTC being a routinely monitored liquid biopsy cancer biomarker has not improved in recent years and the use of CTC enumeration in the clinic is relatively low [128, 129].

Circulating tumor DNA (ctDNA) is circulating cell-free DNA (cfDNA or ccfDNA) that originates from a tumor cell. cfDNA arises from a number of sources including cell apoptosis and spontaneous active DNA release [130], and in a cancer patient, ctDNA usually comprises a small fraction of total cfDNA [131]. In addition to its low relative abundance, ctDNA seems to be more fragmented than normal cfDNA, with average fragment sizes at ~140 bp vs 167 bp, respectively [132]. Although total cfDNA level might be correlated to tumor burden of the patient [133], it may only be considered as a risk factor, but not a diagnostic or prognostic marker in that it lacks specificity. In contrast, ctDNA is highly specific to the tumor, suitable to be used as a personalized tumor biomarker, but robust detection of ctDNA was not attainable for long time until high throughput sequencing technology and polymerase chain reaction (PCR) assays with good analytical performance were developed. As a tumor cell genome retains much of the normal cell genome, most cfDNA originating from tumor cells cannot be easily differentiated from that from normal cells. One identifiable manifestation of ctDNA is somatic mutations specific to the tumor genome. For example, the somatic mutation profile of a tumor can be established by sequencing the primary tumor, and assays targeting these mutations can be applied in cfDNA samples to detect the presence and quantity of ctDNA. Both small-scaled mutations such as SNVs and small indels and large-scaled mutations such as chromosomal rearrangements have been successfully detected in follow-up liquid biopsy samples in different types of cancers using deep sequencing or PCR-based methods, and the dynamics of ctDNA was shown to be correlated to outcome and/or relapse in these studies [134-139]. Measurements of ctDNA using chromosomal rearrangements in Paper II, and as SNVs in Paper III were performed in this thesis, in order to evaluate the feasibility and clinical value of serial monitoring of ctDNA in early stage breast cancer, and the safety of mammography from the perspective of tumor cell dissemination into the bloodstream.

## PI3K/PTEN signaling pathway

Tumor cells have abnormal phenotype and behavior. In normal circumstances, cells are programmed to respond to extracellular molecules, ligands, via transmembrane receptor proteins. When ligands bind to their corresponding receptors, a cascade of biochemical reactions happen in the cytoplasm and the nucleus, changing the expression of a certain collection of genes and subsequently the cell behavior. These signal transduction cascades are called cell signaling pathways. Deregulation of cell signaling pathways are frequently observed in cancer cells and impart selective advantages and promote their progression into malignant cells.

The PI3K/PTEN signaling pathway, illustrated in Figure 3, is an evolutionarily conserved pathway regulating a variety of processes in normal cells, including metabolism, survival, proliferation, apoptosis, growth, and migration [140], and one of the most frequently hyperactivated pathways in breast cancer and other cancer types [141].



Nature Reviews | Clinical Oncology

**Figure 3 The PI3K/PTEN signaling pathway**

From Janku, F., Yap T.A. and Meric-Bernstam F., *Targeting the PI3K pathway in cancer: are we making headway?* Nat Rev Clin Oncol. 2018. 15: p. 273-91. Reprinted with permission from Springer Nature.

When extracellular growth factors or insulin are presented to receptor tyrosine kinases (RTKs) on the cell surface, the receptors are dimerized and cross phosphorylate each other leading to activation. The activated RTKs recruit phosphoinositide 3-kinase, (also known as phosphatidylinositol 3-kinase, PI3 kinase, or PI3K) to the cell membrane, phosphorylates and activates it. The activated PI3K then catalyzes the reaction to turn phosphatidylinositol (4,5)-biphosphate (PIP2) into phosphatidylinositol (3,4,5)-triphosphate (PIP3). Increased levels of PIP3 then recruits the serine/threonine kinase AKT to the plasma membrane, where it is activated by a variety of other kinases such as the mammalian target of rapamycin (mTOR). The activated AKT, in turn, phosphorylates a spectrum of substrates, leading to oncogenic behaviors of the cell. Acting in opposition to PI3K, the phosphatase and tensin homolog (PTEN) tumor suppressor gene catalyzes the precise opposite reaction to dephosphorylate PIP3 back into PIP2, thus acting as a negative regulator of the signaling pathway. As the catalytic subunit of PI3K, called p110alpha, is encoded by the oncogene *PIK3CA*, and the PTEN phosphatase by the tumor suppressor gene *PTEN*, it is conceivable that aberrations in these two genes are likely to have direct impact on carcinogenesis and tumor progression. In fact, mutations in *PIK3CA* are observed in nearly 30% of breast cancer cases (using standard methods with standard sensitivity), and a large fraction of them happen at glutamic acid residue 542 and glutamic acid 545 within the helical domain, or at histidine 1047 within the kinase domain of the gene [142]. As for *PTEN*, the mutation rate in breast cancer is much lower, at about 5%, however the expression of the gene and protein product is also down regulated in about 25% of cases [143-145]. Interestingly, PTEN loss is essentially mutual exclusive with *PIK3CA* mutations in breast cancer, suggesting the reciprocal driving role of *PIK3CA* mutations or PTEN protein loss in promoting tumorigenesis [144]. Mutant *PIK3CA* is associated to resistance to HER2 targeted treatments [146], and *PIK3CA* inhibitors, have been assessed and in the case of alpelisib, approved, for the treatment of *PIK3CA*-mutated, hormone receptor-positive advanced breast cancer patients [147, 148]. Mutations in *PTEN* or PTEN protein loss are usually indicators of poor prognosis in breast cancers [145], and they are associated with resistance to trastuzumab treatment [149].

The mechanism of PTEN protein loss remains understudied. One mechanism is PTEN gene disruption, particularly in some basal-like breast cancers associated with BRCA1 mutation [150]. Another common mechanism is cellular protein degradation, proteolysis, mediated by the addition of a chain of small proteins, known as ubiquitin, to the target protein. This process is called polyubiquitination and is catalyzed by E3 ubiquitin ligases specific to the target protein. It was proposed that the neural precursor cell expressed, developmentally down-regulated 4 (NEDD4) protein is the E3 ubiquitin ligase of the PTEN protein, responsible for polyubiquitination and downregulation of PTEN protein in mouse prostate and human bladder cancer models [151]. This theory of PTEN protein degradation by NEDD4-mediated polyubiquitination was corroborated by observations in axon

branching [152, 153], T-cell activation [154], keloid formation [155], and insulin-mediated glucose metabolism [156], and *NEDD4* and *PTEN* expression levels were inversely correlated human non-small cell lung cancer [157] and colon cancer [158] cohorts. However, several reports present data which do not support this mechanism for PTEN protein degradation by NEDD4 [159-163]. Understanding the biology behind PTEN protein loss in breast cancer may establish new breast cancer biomarkers and help clinicians make better decisions in management of the disease.

It is worth noting that tumor cell behavior is not regulated just by a few critical genes or signaling pathways, but rather by a complex network of them. Genes may play different roles in different pathways, and different pathways may act independently or cooperatively, the latter called signaling pathway crosstalk. For example, in addition to its cytoplasmic phosphatase activity, the PTEN protein was also found in the cell nucleus, contributing to stabilization of the genome [159, 164], and an elongated version of PTEN protein, known as PTEN-long, was reported to be secreted from cells in exosomes, and may inhibit the PI3K/PTEN signaling pathway in a paracrine-like fashion [165]. Overexpression of the RTK epidermal growth factor receptor (EGFR) was found to strongly correlate to PTEN protein loss in basal-like breast cancers, suggesting the coadministration of EGFR and PI3K inhibitors in this subtype may be a viable treatment strategy in the future [166]. Moreover, though both being well-established tumor suppressors, high *PTEN* expression was reported to be a negative prognostic marker in advanced local breast cancers with wild-type *TP53*, and the biological reason was not fully understood [167].

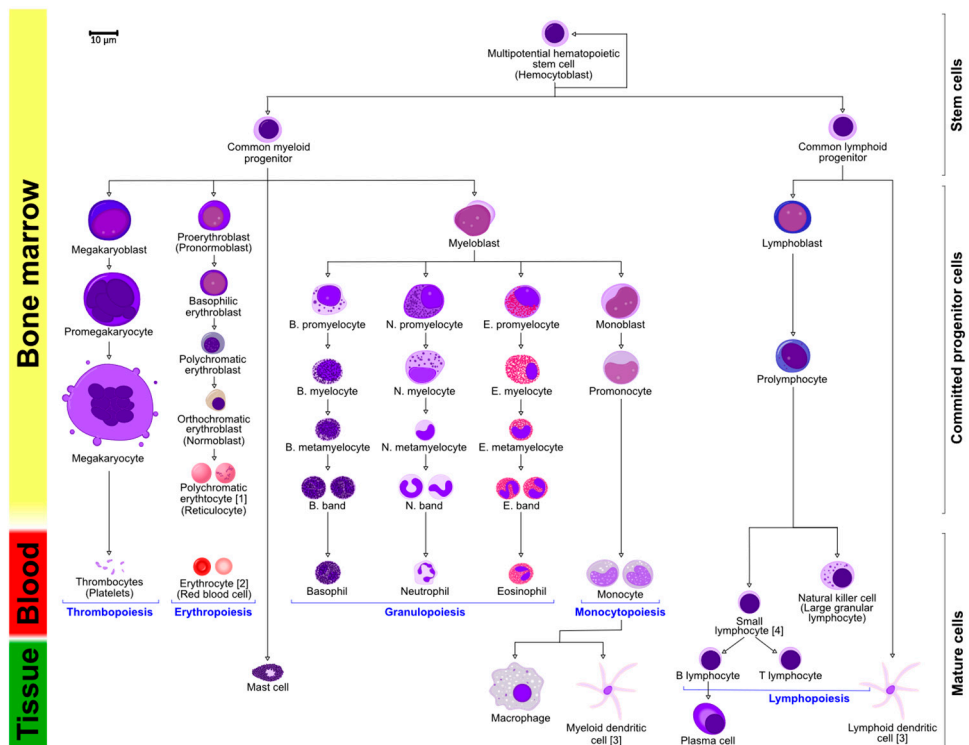
## Acute myeloid leukemia (AML)

Leukemia refers to a group of blood cancer that affects white blood cells (WBCs), known as the leukocytes, and their progenitors, called blasts, in the bone marrow. It was estimated that in 2020, nearly half a million people were diagnosed with leukemia, and more than 300,000 patients died from it around the world [61]. Leukemia can be divided into four major types, given in Table 2, based on the type of blasts of origin and rapidness of disease development. These four major types of leukemia account for ~92% of all leukemias in the United States in 2019 [63].

**Table 2. Major types of leukemia**

Hematopoietic lineage	Acute	Chronic
Myeloid	Acute myeloid leukemia (AML)	Chronic myeloid leukemia (CML)
Lymphocytic	Acute lymphocytic leukemia (ALL)	Chronic lymphocytic leukemia (CLL)

As depicted in Figure 4, the process of blood cell formation, called hematopoiesis, produces functional blood cells from hematopoietic stems cells and precursor blasts in the bone marrow of an adult, and releases them into peripheral blood and the lymphatic system. Specifically, myeloblasts differentiate into WBCs in peripheral blood, including basophils, neutrophils, eosinophils, and monocytes. Acute myeloid leukemia (AML) is the malignancy of bone marrow descendants of the common myeloid progenitor cells, which releases immature blasts into peripheral blood instead of functional leukocytes, sabotaging the normal blood cell function. AML is the most frequently diagnosed type of leukemia among the four, with an estimate of 21,450 cases diagnosed in the US in 2019 (~38% of the four major types combined, or 35% of all types of leukemia). AML is primarily diagnosed in late adulthood with a median age of over 65 years at first diagnosis [62, 63, 168, 169]. The prognosis of AML, especially in older patients, is poor, with more than half of the older patients succumbing within 2 months after diagnosis, and the median survival is only about 6 months [170].



**Figure 4 Schematic of hematopoiesis**

From [https://commons.wikimedia.org/wiki/File:Hematopoiesis\\_\(human\)\\_diagram\\_en.svg](https://commons.wikimedia.org/wiki/File:Hematopoiesis_(human)_diagram_en.svg) CC BY-SA 3.0.

The cause of AML is not well understood. However, a range of risk factors contributing to the disease have been identified, including 1) genetic disorders of Down syndrome, Fanconi anemia, and Li-Fraumeni syndrome, 2) exposure to physical and chemical carcinogens of benzene, pesticides, cigarette smoking, and herbicides, 3) exposure to therapeutic and non-therapeutic radiation, and 4) previous chemotherapy [171]. As an individual is aging, somatic mutations in driver genes of leukemia, such as *DNMT3A*, *TET2*, *AXL1*, and others, accumulate in the hematopoietic stem cells (HSCs), causing HSCs harboring such mutations to have a selective advantage over other HSCs, and thus most of the mature leukocytes are differentiated from these preleukemic HSCs. This is called clonal hematopoiesis and is a preceding clinicopathological event of the genesis of leukemia [172, 173].

## Diagnosis

There is no screening test for AML. When individuals show symptoms such as bruising, bleeding, fever, fatigue, pain in bone, joint, and/or muscle, they may be referred for further workup and diagnostic testing for AML. Peripheral blood, potentially followed by a bone marrow specimen, is sampled and tested for diagnosis of AML. A complete blood count (CBC) and/or a peripheral blood smear is performed, with the former showing reduced number of red blood cells and the latter showing poorly differentiated leukemic blasts in the peripheral blood. If the blood test shows a sign of leukemia, a bone marrow specimen will be obtained through aspiration or core biopsy from, in most cases, the pelvis of the patient for definitive diagnosis. Myeloblasts within the bone marrow cell population are to be identified, morphologically evaluated (whether the blasts are from the myeloid or the lymphoid lineage), and counted by hematopathologists, and a  $\geq 20\%$  presence of blasts is sufficient for diagnosis. Due to their distinct prognoses, AML needs to be further distinguished from acute lymphocytic leukemia (ALL) and myelodysplastic syndrome (MDS) by hematopathological evaluation [174]. In addition to microscopic check of the bone marrow smear, chromosomal aberration can be karyotyped by fluorescence *in situ* hybridization (FISH) and characteristic cell surface markers can be checked with multicolor flow cytometry (MFC) to diagnose with leukemia even if the blasts are  $< 20\%$ . Recently, molecular methods such as sequencing, quantitative polymerase chain reaction (qPCR), and digital PCR (dPCR) have been added to the toolkit for diagnosis of AML.

## Treatment

Since AML is a “liquid cancer” in which a solid mass is usually absent (except for myeloid sarcoma, a rare extramedullary solid manifestation of myeloid leukemia formed from myeloblasts [175]), surgery is not a treatment option. Treatment of AML is divided into two phases: 1) the induction phase, aiming to kill as many

leukemic blasts as possible, and 2) the consolidation phase, also known as the post-remission phase, aiming to eliminate or keep the leukemic blasts population as small as possible and prevent relapse.

Chemotherapy is the primary treatment of the induction phase, and a cocktail of multiple drugs are often co-administered. Commonly used chemotherapy drugs include anthracyclines such as doxorubicin, daunorubicin, and idarubicin, small molecules such as midostaurin, venetoclax, glasdegib, and ivosidenib, and antibody-drug conjugate (ADC) such as gemtuzumab ozogamicin. In fact, daunorubicin and cytarabine mixed with a molar ratio 1:5 have been commercialized as Vyxeos, and approved by US FDA for treatment of high-risk AML [175].

Most AMLs are going to relapse after induction, if no further clinical intervention is given even when complete remission (CR) is achieved [174, 176]; therefore, consolidation treatments are highly necessary. In the consolidation phase, intensive chemotherapy, similar to that used in the induction phase, and/or hematopoietic stem cell transplantation (SCT) are administered. SCT intends to restore the hematopoietic system within the bone marrow, after the existing system, containing both leukemic and healthy blasts, is eradicated by intensive chemotherapy. Two types of SCT are available: 1) allogeneic SCT (allo-SCT), using hematopoietic stem cells from a healthy donor, and 2) autologous SCT (auto-SCT), using hematopoietic stem cells from the patient him- or herself. Allo-SCT is by far the more common SCT approach for AML. Human leukocyte antigen (HLA) class I and II antigens are genotyped for the patient and the potential donor, and those that have at least 9 or 10 alleles matched are considered immunologically compatible donors [177]. In the best-case scenario, the donor is related to the patient, and is most often a matched sibling donor (MSD). If such donors are not available, a matched unrelated donor (MUD) may be available. Allo-SCT has been shown to significantly improve the risk-free survival of high- and intermediate-risk AML patients in first CR [178], however, finding matched donors and subsequent immunosuppressive therapy are limitations of this approach. Alternatively, auto-SCT can be performed when no matched donors are available. Non-leukemic bone marrow stem cells are taken from the patients and stored. After intensive myeloablative therapy, the stored cells are purged of residual leukemic cells and infused back into the patient to reconstruct the hematopoietic system. The drawback of auto-SCT is that the stored stem cells could still be contaminated by leukemic cells even after purging, omening a future relapse of the disease.

### Minimal residual disease (MRD)

Although CR is achieved in a large fraction of AML patients, the relapse rate remains high, leading to a poor prognosis of the malignancy. Relapse of AML is mainly due to so-called minimal residual disease (MRD), defined as leukemic cells remain and persist but may not (yet) form a clinically overt leukemia. Monitoring



and/or detection of MRD in AML, is a requisite in the standard clinical management of AML.

Classic MRD detection methods are bone marrow blast morphology and percentage assessment using light microscopy and leukemia associated immunophenotype (LAIP) detection using multicolor flow cytometry. Both these methods are not without drawbacks [179]: light microscopy is impaired by the limited sensitivity of the method and variability between different assessing hematopathologists, and flow cytometry relies on an established LAIP from the diagnostic sample, which is subject to change during treatment. Treatment acts as a strong selection pressure in AML, and the leukemic cell population at relapse can be significantly different from that at diagnosis. A subclone within the original malignancy can be selected over other subclones, and new mutations can be acquired in the course of treatment, both contributing to a change of the LAIP of the AML [180, 181], a process called clonal evolution. In cases where notable chromosomal abnormality is present in the diagnostic sample, FISH can also be done on follow-up samples for MRD detection [182].

With the advent of modern molecular biology technologies, new methods have been introduced for detection of MRD in AML including qPCR and sequencing. Mutational landscape studies launched by The Cancer Genome Atlas (TCGA) found that AML has fewer mutations per genome than most other cancers, and an average of 5 mutations per case was found in one of the 23 significantly mutated genes, while more than 200 other genes had mutations found in at least two samples. This focused pattern of mutation landscape makes targeted sequencing a highly feasible approach for molecular MRD detection in AML [183]. Among the small number of recurrently mutated genes, *FLT3*, *NPM1*, and *DNMT3A* are most frequently mutated [184, 185]. Mutations within *FLT3* and *NPM1* are mostly internal tandem duplications (ITDs) and insertions respectively, with the latter occur predominantly in a conservative site in exon 12 of the gene [186]. Highlighted in Figure 5, after local realignment, more than a dozen recurrently observed small insertions exist in *NPM1*, and MRDs represented by these *NPM1* insertions are detectable with high sensitivity by qPCR [187-189].

As sequencing and qPCR both involve DNA polymerase, random nucleotide base misincorporation caused by the inherent polymerase error is inevitable, giving rise to false positive mutation calls, hampering the usability of these methods for accurate and sensitive MRD detection. We aimed to prove the concept that our innovative dPCR-based mutation detection technology IBSAFE can essentially bypass this type of error, yielding a much-improved analytical performance of MRD detection. Further details are provided in the Methods section of this thesis.





# Aims

This thesis began with a project investigating the underlying mechanism of PTEN protein loss in breast cancer. Following on this pathway and an interest in liquid biopsies, we sought to investigate the concept that various types of tumor-specific mutations, including those in *PIK3CA*, can be used as biomarkers to monitor the dynamics of occult tumor burden, evaluate the degree of tumor content dissemination into the bloodstream during mammographic compression, and detect minimal residual disease in multiple types of cancer.

Specifically

- Paper I** aimed to investigate whether the ubiquitin ligase NEDD4 is responsible for the PTEN protein loss phenotype in breast cancer to better understand the PI3K/PTEN signaling pathway in breast cancer and develop new actionable biomarkers.
- Paper II** aimed to prove the concept that tumor specific chromosomal rearrangements found by low coverage sequencing can be detected in cell-free DNA from plasma, and that this minimally invasive liquid biopsy approach can detect occult disease with high sensitivity and specificity, thus predicting future relapses and sparing disease-free patients from over treatment.
- Paper III** aimed to investigate whether mechanical compression of the breast in mammography has a risk of releasing circulating tumor cells and/or circulating tumor DNA into the bloodstream. IBSAFE was tested as an ultrasensitive method of detecting ctDNA.
- Paper IV** aimed to explore the value of IBSAFE as an ultrasensitive ddPCR-based mutation detection method in MRD detection of AML. Leukemia-specific genomic variants were identified by whole exome sequencing at diagnosis, and IBSAFE was used to monitor the variants in bone marrow samples collected at follow-up visits and relapse(s).



# Methods

## Patients, samples, and ethics

Given the nature of the divergent scientific questions addressed in this thesis, different patient and sample cohorts were used in each study, and respective ethical review were evaluated and approved by relevant bodies.

The main tumor sample cohort in **Paper I**, the Swedish Cohort in the published paper, were obtained from the South Sweden Breast Cancer Group collected at the Lund University Hospital, Lund, between 1986 and 1994. The cohort are sporadic stage II breast tumors that had received adjuvant tamoxifen treatment for 2 years and are from a series of larger patient cohort [190]. The tumors were selected such that approximately one third were PTEN protein negative, and the rest are PTEN protein positive, wherein node and hormone receptors status distributions are roughly matched. Protein levels of NEDD4 and PTEN were obtained from immunohistochemistry (IHC) staining results of the tissue microarray (TMA) of the tumors, and mRNA levels of *NEDD4* and *PTEN* were available from gene expression microarray analyses. Thus, a subset of 186 tumors with IHC and/or gene expression microarray results were included in this study. The collection of the samples was approved by the Regional Ethics Committee at Lund University. In addition to this main cohort, samples with mRNA and/or protein level results from two independent breast cancer cohorts, the Netherlands Cancer Institute (NKI, N = 295) and The Cancer Genome Atlas (TCGA, N = 970), were analyzed to corroborate the findings from the Swedish Cohort.

Patients enrolled in the Breast Cancer and Blood Study (BC Blood, Sweden) [191] were included in **Paper II**. A selection was applied on this starting cohort of 725 patients with inclusion criteria of 1) non-metastatic breast cancer at diagnosis with no neoadjuvant therapy administered, and 2) availability of fresh frozen primary tumor specimen and at least two follow-up plasma samples, resulting in 71 total patients passing the eligibility requirements. The patients were divided into the eventual metastatic (EM) group (N = 24) if a distant metastasis was clinically identified 1-6 years after diagnosis, and the long-term disease-free (DF) group (N = 47) if disease-free survival was observed upon the last follow-up > 7 years after diagnosis. From these, 14 EM and 6 DF patients were randomly selected to be investigated in this study. DNA extracted from fresh frozen tumor tissues, and from plasma samples taken prior to the surgery and normally at 3-, 8-, 12-, 24-, and 36-

month follow-up timepoints after the surgery were analyzed in this study. The study was approved by the Regional Ethics Committee at Lund University. All patients were provided with written and oral information about the study by trained health professionals, and written consent were signed.

The cohort in **Paper III** comprise 31 patients from ongoing perspective Sweden Cancerome Analysis Network – Breast (SCAN-B) trial, an initiative aiming to improve diagnosis, treatment, survival, and quality of life for breast cancer patients by joining together expertise from biomedicine researchers, physicians, nurses, and other health-care specialists. These 31 patients were diagnosed with breast cancer between 2015 and 2016 and volunteered to have an extra mammography scan after diagnosis. For each patient, blood samples were collected before and after the mammography, from superior vena cava via a central venous access planted prior to this study to allow for neoadjuvant chemotherapy administration, and from a peripheral vein. These central and peripheral blood samples, collected before and after the mammography, were used to do the circulating tumor cell (CTC) and the circulating tumor DNA (ctDNA) analyses in this study. All patients signed written consent and the study was approved by the Regional Ethical Review Board at Lund University.

**Paper IV** involves 14 patients diagnosed with acute myeloid leukemia (AML). The patients were divided into relapsing (N = 10) and non-relapsing (N = 4) groups by clinical standard at the end of follow-up in September 2019. DNA samples extracted from bone marrow aspirates taken at diagnosis, clinical relapse, and follow-up visits were analyzed in this study. All patients signed written informed consent and the study was approved by the Regional Ethical Review Board at Lund University.

## Immunohistochemistry (IHC)

Immunohistochemistry (IHC) is a technique that uses antibodies to detect the presence of cellular antigens of interest in the tissue being tested. Especially in breast cancer, IHC has been established as a major clinical laboratory tool to evaluate steroid hormone receptors, tumor proliferation markers, and angiogenesis and apoptosis markers in tumor tissue [192]. According to the St. Gallen breast cancer guidelines, IHC is routinely done in breast cancer tissues to check the statuses of estrogen receptor (ER), progesterone receptor (PgR or PR), both being hormone receptors, human epidermal growth factor receptor 2 (HER2), an oncogenic growth factor receptor, and Ki-67, a proliferation marker [79]. The staining results of these biomarkers are useful for guiding management of breast cancer, fitting the tumor into a histological subtype, making prognosis, etc. There are many ways to evaluate the status of a biomarker. For example, for the hormone receptors, the status can be determined by H-score. For ER, H-score is the percentage of weakly stained cells

plus two times the percentage of moderately stained cells plus three times the percentage of strongly stained cells, whereas for PR, H-score is the percentage of cells with > 10% staining intensity. Cells with H-scores > 1 are considered positive for the hormone receptor [193, 194]. For HER2, a commonly scoring system gives points to a tumor on a 0-3+ scale, in which 0 stands for no stain, 1+ stands for weak stain, 2+ stands for equivocal or borderline stain, and 3+ stands for strong or positive stain. Tumors with 3+ are classified as HER2-positive, while those with 0 and 1+ are classified as HER2-negative. For tumors with equivocal HER2 scores, a FISH test can be conducted to verify the HER2 protein amplification by checking status of *ERBB2*, the gene responsible for coding the HER2 protein [195, 196]. As for Ki-67, the percentage of cells with positive staining is considered, and it has been proposed that  $\leq 15\%$ , 16-30%, and  $> 30\%$  are markers for low, intermediate, and high proliferation in breast cancer [197]. Ki-67 evaluation is varying, with no established consensus criteria and many regional/national recommendations.

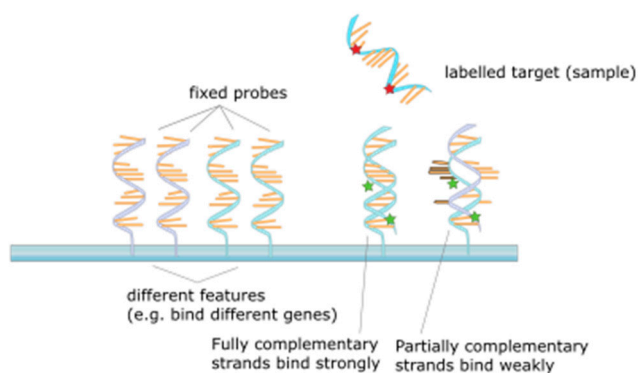
Conventionally, tissues for IHC assessment are acquired from core biopsy, a clinical practice where a core needle penetrates and takes a small piece of sample from the suspected malignancy site, or from specimens obtained at surgery. The morphological structure and antigenicity of the tissues are retained in a process called fixation, usually involving formalin. The fixed tissues are then sliced into  $\mu\text{m}$  thick pieces and mounted onto glass slides for IHC staining and microscopic inspection. The technology of tissue microarray (TMA) was invented to improve the threshold of histological analysis, including IHC. TMAs are typically prepared by taking biopsies using 0.6 mm cylinders from donor tissue blocks and transferring the specimens to a recipient block. The recipient block containing up to 1,000 tissues are then cut into 4-8  $\mu\text{m}$  pieces for downstream analysis. At least 200 such slices can be prepared from such recipient blocks [198]. For the IHC done in **Paper I**, breast cancer tissues in triplicates were assessed for NEDD4 protein status on TMA.

Since the NEDD4 protein is not a commonly investigated epitope in breast cancer, a good antibody, a proper dilution factor for the antibody, and a scoring system needed to be established. For the antibody and its dilution factor, after reading the literatures and doing pilot tests, a rabbit polyclonal antibody targeting the WW2 domain of the NEDD4 protein (#07-049 Millipore) was chosen and diluted 1:500 for use in IHC [151]. The NEDD4 staining was evaluated with a scale similar to the HER2 scoring system, with 0 given to specimens with no NEDD4 staining, 1+ to weak staining, 2+ to intermediate staining, and 3+ to strong staining. Specimens with discordant scores from the triplicates were given the majority score, if two of the three scores were the same to each other, or excluded from the analysis, if all three scores were different. Specimens with IHC scores of 0 and 1+ were categorized as NEDD4 protein negative and 2+ and 3+ were categorized as NEDD4 protein positive for the statistical analyses in this study.



## DNA microarray

Gene expression data from DNA microarrays were used in **Paper I**. Upon its invention in mid 1990's, DNA microarray has, in a way, served as a quantitative high-throughput version of classical nucleic acid research methods such as Southern blot and Northern blot [199, 200]. Contrary to Southern and Northern blot, a DNA microarray has single stranded DNA pieces with known sequences, called probes, fixed to a solid surface. A typical DNA microarray can contain up to tens of thousands of such DNA piece clusters, called spots, or features, or reporters. After the sample containing fluorescently or radioactively labelled nucleic acids is applied to the microarray, nucleic acids with complementary sequences bond to the fixed probes non-covalently, whereas those with nonspecific sequences are washed away. By measuring the signal intensity of the spots, the amounts of nucleic acids the spots hybridize with can be measured. The mechanism is depicted in Figure 6.



**Figure 6 Schematic of DNA microarray mechanism**

From [https://commons.wikimedia.org/wiki/File:NA\\_hybrid.svg](https://commons.wikimedia.org/wiki/File:NA_hybrid.svg) public domain.

In **Study I**, gene expression data was retrieved from DNA microarray analyses done in previous studies [145, 201, 202].

DNA microarray has a few limitations. First, as the quantity of a nucleic acid is measured as the signal intensity of its corresponding spot, which in turn is proportional to the efficiency of hybridization, the variation in hybridization conditions contribute a lot to the variation of the results, hurting the reproducibility of microarray results. The same mechanism also leads to microarray's inability to reliably detect genes expressed at extremely low or high abundances, limiting its dynamic range to a couple orders of magnitude [203]. Moreover, DNA microarrays are only able to relatively quantify genes present on the microarray, and new

transcripts with previously unknown sequences cannot be interrogated. As a comparison, RNA-seq offers a way to study the transcriptome that overcomes essentially all shortcomings of DNA microarrays – it has a much wider dynamic range of expression levels; it does absolute quantification; it is independent from environment change and has great reproducibility; and it can detect transcripts with previously unknown sequences [204-209]. The TCGA gene expression data used in **Paper I** was generated from RNA-seq results [210]. As the price for sequencing is getting lower and lower, the technology of DNA microarray may fade as a legacy approach, with RNA-seq taking a central role in transcriptome research today.

## High-throughput sequencing (HTS)

Molecular biology research in the past few decades has demonstrated that in most eukaryotic organisms, the phenotype of a cell, essentially manifested as its proteins, are translated from RNA, which in turn is transcribed from DNA, the manifestation of the cell's genotype [211]. This genetic information flow has been known as the central dogma of molecular biology [212]. Conceivably, changes in nucleotide sequence of DNA or RNA can cause changes in amino acid sequence of the protein that use it as a blueprint, altering its function and potentially leading to deleterious consequences, such as cancer. Therefore, knowing the exact sequence of a particular DNA or RNA region, or even that of the entire genome or transcriptome is highly desirable in the realm of molecular biology, particularly in cancer research and clinical management. The methods for determination of nucleotide sequences are collectively known as sequencing. Figure 7 is a schematic illustrating the evolution of sequencing from its genesis.

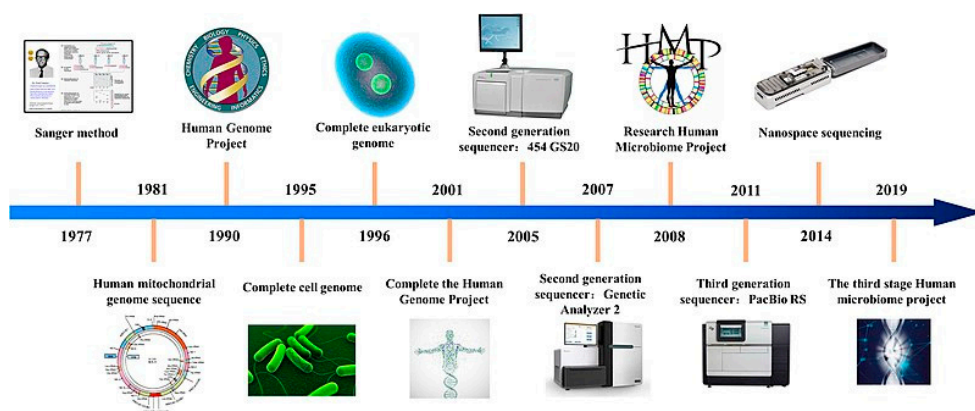


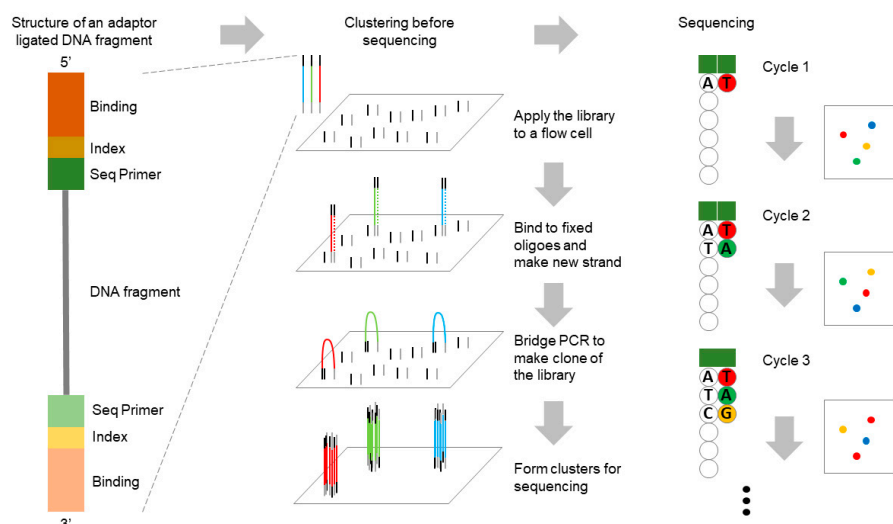
Figure 7 History of sequencing technologies

From [https://commons.wikimedia.org/wiki/File:History\\_of\\_sequencing\\_technology.jpg](https://commons.wikimedia.org/wiki/File:History_of_sequencing_technology.jpg) CC BY-SA 4.0.

The first-generation sequencing technology was invented by Sanger et al. in 1977, and is thus known as Sanger sequencing [213]. Sanger DNA sequencing sequences the fragmented PCR-amplified single-stranded DNA with a mixture DNA polymerase, sequencing primers, dNTP, and ddNTP tagged with four fluorophores with different colors. The dNTP and ddNTP are competitively added to the newly synthesized DNA strand, with the former allowing the synthesis to keep going, while the latter terminating the synthesis and emitting a fluorescence signal corresponding to the nucleobase of the ddNTP. With capillary electrophoresis, the sizes of the newly synthesized double-stranded DNA molecules can be separated with a single-base-pair resolution, and by recording the colors of the differently sized molecules, the sequence of the original DNA fragment can be reconstructed. Given its mechanisms, Sanger sequencing has the limitations such as 1) it can only sequence one type of DNA fragment at a time, 2) the first few dozen bases are often of low quality, and 3) large-size double-stranded DNA molecules are difficult to be separated with electrophoresis, limiting the maximal total length of Sanger sequencing to ~1,000 bp.

In the mid-noughties of the new century, invention of new sequencing technologies drastically changed the landscape of genomic and transcriptomic research [214-222]. The commonly known names of the new technologies include next-generation sequencing (NGS), second-generation sequencing (SGS), massive(ly) parallel sequencing (MPS), and high-throughput sequencing (HTS). Contrary to Sanger sequencing, HTS can sequence an enormous number of different DNA molecules, called sequencing library, simultaneously and in parallel in the same instrument run. Preparation of the sequencing library starts with DNA fragmentation, ligation to a pair of engineered DNA pieces, called adapters, that the sequencing machine, also known as a sequencer, can work with, and usually PCR amplification of the adapter-ligated DNA fragments. This library is ready for sequencing or can be selected for regions of interest and then sequenced. Since the Illumina sequencing systems of HiSeq 2000, HiSeq 2500 and NextSeq 500 were the ones used throughout this thesis, and the Illumina approach is the most common, the HTS discussions in the next subsections are mainly about these platforms. The workflow of Illumina's sequencing by synthesis (SBS) approach is illustrated in Figure 8.

The DNA library is denatured (converted to single strand with physical or chemical methods) and loaded into a sequencing flow cell, a chamber where the library is sequenced. The surface of a flow cell lane is coated with DNA oligoes that bind to the binding region of the sequencing adapters. After binding, the different DNA molecules in the library are cloned through bridge PCR amplification to form clusters from which fluorescence signals can be detected during sequencing. This step is called cluster generation, or simply clustering. Each cluster represents a clone from one library DNA fragment.



**Figure 8 Workflow of Illumina sequencing by synthesis (SBS)**

The clusters are then sequenced by synthesis – sequencing primers bind to the sequencing primer binding region of the adaptors, and dNTPs with different fluorescence dyes or dye combinations are incorporated to synthesize the new strand within each physical cluster in stepwise sequencing cycles, with one base added per cycle. Images are taken after each cycle to capture the fluorescence signals representing the single nucleotide added to each cluster in this cycle. The Illumina HiSeq 2000 platform has 4 exposures per cycle for the 4 different dyes conjugated to the different bases, whereas NextSeq 500 only exposes two times, because the single nucleotides used in this platform are tagged with a red dye for C's, a green dye for T's, both dyes for A's, and no dye for G's, so that two wavelength channels are sufficient for all 4 bases. The clustered DNA fragments can be sequenced from one side, called single-end sequencing, or from both sides, called paired-end sequencing. In summary, in HTS, the number of clusters represents the number of different DNA fragments in the library that can be sequenced in parallel, the number of sequencing cycles represents the number of bases of each DNA fragment that can be sequenced, and the product of the number of clusters and the number of cycles is called the output of the sequencing experiment.

Illumina provides a variety of sequencing platforms matched with different choices of sequencing reagents and consumables to meet different scientific requirements. For example, an Illumina NextSeq 500 sequencer with a High Output Kit, 300 cycles can generate up to 400 million clusters, and sequence up to 300 bases of the DNA fragment within each cluster.

Libraries prepared from different original samples can also be pooled and sequenced together through the use of sample barcodes in the adapters. After sequencing, the index region of the sequencing adapter, which is composed of a 6- to 8-base sample specific DNA sequence can be used to identify which original sample within the library pool the DNA fragment in a cluster was from. This step is done bioinformatically and is called demultiplexing. In addition to the sample-specific index sequences, another type of index sequence can be added to the adapters as well, called the unique molecular identifier (UMI), which are small strings of random oligonucleotides [223]. As PCR amplification may be performed in library preparation, it is possible that multiple PCR product molecules synthesized from the same original DNA fragment are sequenced in different clusters, causing sequencing data duplication. Moreover, wrong DNA bases may be incorporated into some PCR product molecules due to polymerase error, resulting in false positive mutation calling from the sequencing results. The addition of UMIs effectively helps to identify PCR-duplicated sequencing reads and collapse them into one original DNA fragment, improving the accuracy of quantification of the sequencing reads [224-226]. It also enables the exclusion of a large fraction of potential false positive mutation calls rising from PCR polymerase error [227, 228].

The HTS experiments often generate large amounts of data, which is to be processed bioinformatically. Depending on the type of sample the sequencing library is built from and the scientific questions to be answered with downstream analysis, different bioinformatics pipelines are employed to process the raw data. The details are provided in the next subsections.

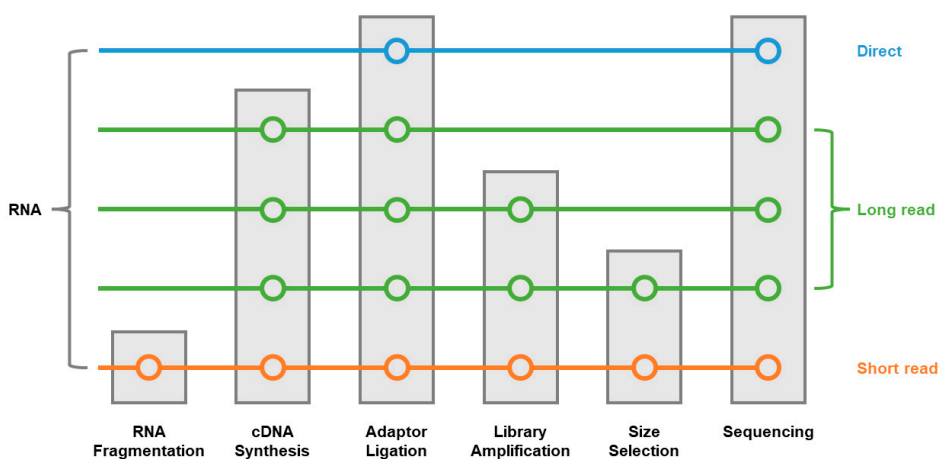
The 2010's saw the rising of the third-generation sequencing (TGS), also known as long-read sequencing [229]. Although methods developed by Pacific Biosciences [230-232] and Oxford Nanopore [233-237] have caught the attention of the research community and made major accomplishments, long-read sequencing seems to be more error prone and have a lower throughput than the second-generation HTS technologies, and also a considerably higher price point per megabase of sequence [232, 238].

## RNA sequencing (RNA-seq)

The technology of RNA sequencing came into wide use in molecular biology research in year 2007 and 2008 [239-243], and since then has developed into a powerful tool for sample transcriptome profiling [205]. As RNA-seq and DNA microarray both take a sample's transcriptome as object of the experiments, both have been established as methods for analyzing gene expression levels. Over the last decade, comparisons of the two technologies have had most researchers convinced that RNA-seq is a better technology at differential gene expression analysis in that 1) RNA-seq does not rely on a well-designed set of hybridization DNA probes and thus can not only measure the expression levels of known

transcripts, but also detect previously unknown gene fusions, single nucleotide variants, small insertions and deletions, and alternative splice sites; 2) RNA-seq has a wider dynamic range and better reproducibility than microarray; 3) RNA-seq has a better sensitivity than microarray at detecting weakly expressed genes [209, 244-247]. Due to its technical advancement and cheaper cost, RNA-seq has essentially made DNA microarray an obsolete technology in recent years.

Despite termed RNA-seq, the technology, in most cases, does not sequence RNA directly, but rather takes complementary DNA (cDNA), synthesized via reverse transcription from the sample's mRNA, as the input to the instrument and does the sequencing. Although long-read sequencing technologies are emerging, allowing for identification of alternative transcript isoforms and even RNA modifications [234, 248-250], Illumina's SBS short-read sequencing approach is by far the most popular approach. According to National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) repository, over 95% of all RNA-seq reads in the database were acquired from Illumina's short-read sequencing platforms, illustrating Illumina's near monopolistic dominance of the market [251-253]. The difference in workflows of the RNA-seq approaches is illustrated in Figure 9. Since only short-read cDNA sequencing was used in this thesis, description of RNA-seq hereafter refers to this method without further annotation.

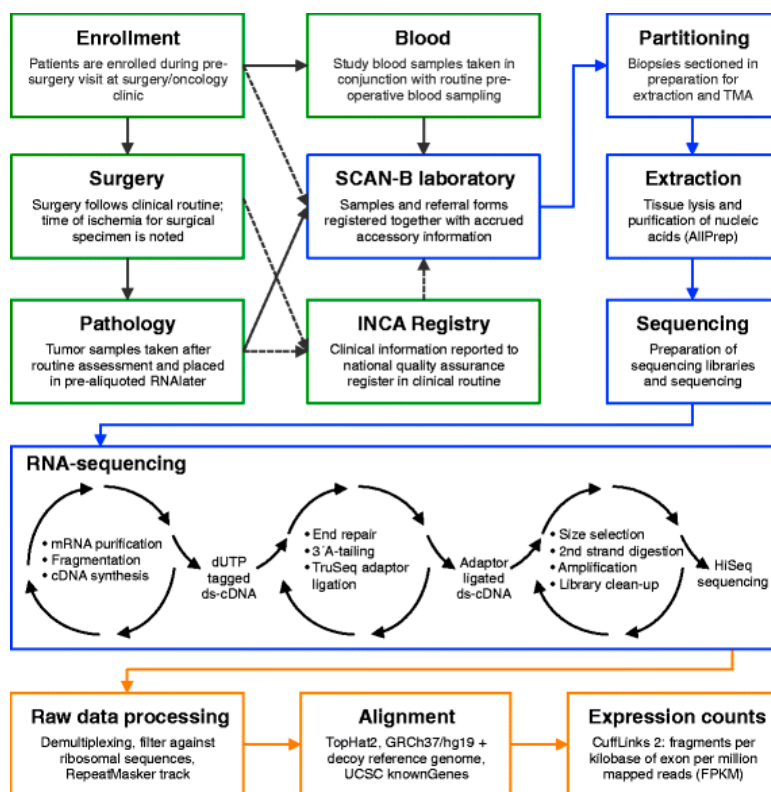


**Figure 9** Short-read, long read, and direct RNA-seq workflows

Inspired by Stark R., Grzelak M., and Hadfield J. *RNA sequencing: the teenage years*. Nat Rev Genet, 2019.

The 31 patients in **Paper III** enrolled in the SCAN-B project, and RNA-seq was performed on their primary tumor tissues. A protocol of mRNA purification and

fragmentation, cDNA synthesis, library preparation, and sequencing are described in detail by Saal et al. [254], and is illustrated in Figure 10.



**Figure 10 Detailed SCAN-B RNA-seq workflow**

From Saal L.H. et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med*, 2015. 7(1): p. 20. Reprinted with permission of the authors.

In brief, ~30 mg of primary tumor tissue fully immersed in 1 mL of RNAlater (Ambion) for at least 16 hours at 4 °C was dissected per sample for tissue lysis. At least 800 µL of lysis buffer per sample was prepared by adding 8 µL of 2-Mercapatoethanol to 790 µL of RLT Plus buffer (Qiagen). Tissue lysis was done with two 5mm stainless steel beads (Qiagen), 400 µL of lysis buffer, and 2 µL of Reagent DX Antifoaming reagent (Qiagen) on a prechilled TissueLyser (Qiagen) at 50 Hz for two rounds of 4 minutes. After the lysis, another 400 µL of lysis buffer was added to the lysed sample, and sample was centrifuged at 16,000 relative centrifugal forces (rcf) for 5 minutes at room temperature in a QIAshredder column (Qiagen). The flowthrough, which contains total DNA, RNA, and protein of the

sample, was homogenized at 80 °C for at least 30 minutes. 350 µL of the flowthrough was used as input of the AllPrep RNA/DNA/flowthrough isolation protocol with minor modifications [254] semi-automatedly done on the QIAcube instrument (Qiagen), and the purified total RNA fraction was used in the downstream experiments.

For each sample, 1 µg of purified total RNA diluted in 50 µL of water was subjected to two rounds of Dynabeads mRNA purification (Thermo Fisher Scientific) followed by a 1.5-minute incubation at 70 °C with zinc RNA fragmentation reagents (Ambion), yielding, on average, ~10 ng (1%) of mRNA fragmented to ~240 bases. Strandedness is important information in RNA-seq that should not be lost, for the same double-stranded DNA segment can be transcribed in both strands, and thus be a part of different RNA transcripts. To preserve the strandedness information, the fragmented mRNA was reversely transcribed to cDNA in two steps, with step one synthesizing the first strand of cDNA with reverse transcriptase, random hexamer as primers, buffers necessary for the reactions, and dNTP. After cleaning up of the product, the second-strand synthesis was done in essentially the same way, except dUTP was used instead of dTTP to make up the dNTP component of the reaction. The double-stranded cDNA molecules then underwent standard Illumina TruSeq end repair, A-tailing, adapter ligation, PCR amplification, and size selection steps to construct indexed sequencing libraries. Before pooling and loading the libraries onto a sequencer, the cDNA libraries were treated with uracil-DNA glycosylase (UDG, New England Bio Labs) to digest the second strand in which uridine was incorporated instead of thymidine. This step of UDG treatment was adapted from a method published by Parkhomchuk et al. [255], and could effectively retain strandedness information of the original mRNA sample.

Each indexed library pool was sequenced with the Illumina HiSeq 2000 Sequencing System (Illumina) in dual flow cell mode across two flow cells. Paired-end reads were generated per sequencing experiment, and approximately 30 million read pairs were generated per sample.

The sequencing results make no sense until they are bioinformatically processed. In **Paper III**, among many other attainable usages of the data, the primary goal was to call somatic single nucleotide variants (SNVs) and small insertions and deletions (indels) against which IBSAFE mutation detection assays could be designed, and the tumor specific mutations could be used as a circulating tumor biomarker in central and peripheral blood samples drawn before and after the mammographic compression. The pipeline used for calling, annotating, and filtering genomic variants from RNA-seq data was described in detail in another paper [256]. Basically, after the sequencing experiments were done, base-calling was performed using Illumina's on-instrument software and stored in Illumina's BCL format. With the IlluminaBasecallsToFastq tool from the Picard suite, the reads were then converted to the more commonly used FASTQ format, and demultiplexed into per-sample FASTQ files according to the sample specific index sequence ligated to the



reads [257]. Trimmomatic was used to remove adapter sequences and poor-quality bases at the end of the reads [258]. To assemble loose reads into a transcriptomic profile specific to the sample, a step called alignment or mapping was done. HISAT2 was used to map the reads to the human reference genome version GRCh38 (hg38), from which aligned BAM files are made for variant calling. Before the variant calling was performed, duplicate reads, arising mainly from PCR amplification of the libraries, were identified and filtered away by the MarkDuplicates tool of the Picard suite [257]. This tool flags sequencing reads to be duplicates if 1) the reads have the same starting coordinates, and 2) the overlapping region of the reads have identical sequences. Of all the duplicates, only the read with the highest total base calling quality scores gets to be kept. Depending on what the downstream analysis is, it has been debated whether duplicate removal is advisable in RNA-seq data analysis. Removal of the apparent duplicates, especially if the downstream analysis is differential gene expression analysis, would risk underestimation of expression of certain transcripts. However, in the case of SNVs and small indels calling in **Paper III**, removal of duplicates would significantly reduce the number of identical reads with wild-type genomic sequences, avoiding underestimation of the VAF should a variant is detected. Variants were called upon the aligned BAM file using VarDict-Java with a limit of detection at about 1% VAF [259]. Raw variant calling files without further filtering steps often harbor a considerable number of false positive variant calls and germline variants. To only keep true somatic mutations for IBSAFE assay design, several filters were applied on the raw variants called for the samples. A called variant was kept only if 1) it was mapped to an exonic region of a gene with a quality score larger than 30, 2) the VAF was no lower than 5%, 3) the mean position of the variant in a read was greater than 10, 4) it was a known somatic mutation in COSMIC [260, 261], 5) it was not in predefined regions of low complexity [262, 263], involving RNA editing [264, 265], in SweGen database of genetic variability of the Swedish population [266], or in NCBI's database of single nucleotide polymorphism dbSNP [267]. SNVs and small indels from the filtered and annotated list were selected for IBSAFE assay design.

## DNA Sequencing

DNA from the cell nucleus is the embodiment of genomic information that leads to all functionalities of the cell, therefore sequencing of the whole genome, or certain regions of the genome, is of particular interest in molecular biology and cancer research. The principles of DNA sequencing with HTS technologies are largely the same as those of RNA sequencing described in the previous subsection, except the synthesis of cDNA from RNA is not applicable. Whole genome sequencing was performed in **Study II**, while whole exome sequencing was performed in **Paper IV**.

### Whole Genome Sequencing (WGS)

Different numbers of chromosomal rearrangements are often featured in breast cancer cell genomes [268-271] and can be used as a highly personalized cancer biomarker [272]. Without prior knowledge of where the hotspots of rearrangements are for the type of cancer or the individual tumors, whole genome sequencing (WGS) needs to be done to find the rearrangements. In **Paper II**, the fresh frozen tumor tissue genomic DNA libraries were sequenced on a HiSeq 2000 or a HiSeq 2500 platform, with 2×50 bp, 2×100 bp, or 2×150 bp paired-end sequencing settings. The methods are described in detail in Supplementary Methods of **Paper II** and a standalone *Methods in Molecular Biology* book series chapter [273]. It might be a concern that sequencing the entire human genome needs a significant sequencing capacity, and thus can be quite costly and rather unfeasible. However, since the objective of the WGS employed in **Paper II** was to find some, but not necessarily all, chromosomal rearrangements that can be monitored as circulating biomarkers in the plasma, the sequencing did not need to be very deep. In fact, several whole genome libraries could be sequenced in one experiment. The number of libraries  $N_{library}$  that is suitable to be pooled in one sequencing flow cell is given by Equation (1)

$$N_{library} = \frac{N_{clusters} \times N_{cycles}}{S_{region} \times C / (1 - R_{duplication}) / R_{on.target}} \quad (1)$$

where  $N_{clusters}$  is number of clusters,  $N_{cycles}$  is number of sequencing cycles (or the total number of base pairs sequenced per cluster),  $S_{region}$  is the size of the region of interest, in base pairs ( $S_{region} = 3.09 \times 10^9$  bp [274]),  $C$  is the aimed sequence coverage,  $R_{duplication}$  is duplication rate, and  $R_{on.target}$  is on target rate. In the case of **Paper II**,  $N_{clusters}$  was up to  $1 \times 10^9$  for the HiSeq 2000 sequencer,  $R_{on.target}$  can be considered 100% as the entire genome is the target and therefore almost no read in theory should be left unmapped. Let the  $N_{cycles}$  be 200 (2×100 bp), the  $R_{duplication}$  be 2% (the default value given by Illumina's sequencing coverage calculator [275]), and  $C$  be 5. It can be calculated that up to 12 whole genome libraries may be sequenced together, proving the viability of this setup.

Mapping of the sequencing reads to the reference genome was done in a similar way to that of the RNA-seq pipeline, except since Study II was done earlier, the reference genome version was GRCh37 (hg19), and the aligner was Novoalign. BreakDancer was used to make a preliminary list of candidate chromosomal rearrangements from the aligned BAM files [276]. The preliminary list was then annotated and filtered to deplete potential non-specific rearrangements and false-positive calls. The filters include 1) the rearrangements must have at least two supporting discordant read pairs, 2) no satellite region is within 1kb on either side of the breakpoint, 3) no sequencing gap is within 1kb on either side of the breakpoint, 4) no matching rearrangement in other sequenced tumor or normal samples from the same cohort, 5) the distance between two breakpoints for intra-chromosomal rearrangements

mush be larger than 1kb, 6) both parts of the breakpoints must be mapped to chromosomes 1-22 or X, involving no non-standard sequence contigs. See Supplementary Methods of **Paper II** and the book chapter for further technical details. The exact breakpoint fusion sequences were then reconstructed using our self-developed pipeline SplitSeq [273, 277, 278]. Four classes of chromosomal rearrangements were identified by BreakDancer: CTXs (inter-chromosomal translocations), ITXs – (intra-chromosomal translocations), INVs (inversions), and DELs (deletions).

Although low coverage WGS was proved powerful enough to report somatic chromosomal rearrangements in fresh frozen breast cancer tissue, the question remains whether it is robust enough to find true positive rearrangements from formalin-fixed paraffin-embedded (FFPE) DNA. Formaldehyde, a major compound in FFPE is known to fragment [279] and cross link [280] DNA molecules, thus artificially creating a lot of chimeric DNA molecules that can appear as false positive calls [281]. The applicability of this sequencing setup in FFPE samples needs further investigation and technical troubleshooting.

### *Whole Exome Sequencing (WES)*

Several mutational landscape studies have revealed that AML, compared with other common types of cancer, has fewer mutations per million base pairs of the genome, and mutations tend to be focused on a few dozens of genes, for example *NPM1*, *FLT3*, *NRAS*, *KRAS*, *DNMT3A*, *IDH1*, and *RUNX1*, among others. [184-186, 282-289]. For this reason, targeted sequencing, which enriches the genomic region of interest before sequencing, rather than WGS, was more suitable for determination of the mutational profile for each AML. According to Equation (1), when the other variables are unchanged, i.e., if the sequencing experiments are done with the same technological settings, the smaller the target region size  $S_{region}$ , the higher the sequencing coverage  $C$ , and thus it is more likely to find mutations at a low allele frequency when the target region is small. Although several myeloid leukemia sequencing panels were designed targeting dozens to several hundreds of curated genes or exons to massively reduce  $S_{region}$  [290-294], each panel was built based on prior knowledge and/or data mining, and therefore, was not immune from biased choices of AML-associated regions. In contrast, whole genome sequencing (WES), as a special case of targeted sequencing, unbiasedly sequences all the exonic regions of the genome, increasing the chance to catch personalized AML-specific mutations. Although much bigger than other AML panels, the exome still only makes up ~ 1% of the human genome [295, 296], or ~30 million base pairs, making WES a highly viable approach for **Paper IV**.

There are two ways to enrich the genomic region of interest for targeted sequencing – selective capture and amplification. The selective capture method uses a set of capture probes to hybridize with targeted genomic region. Comparing with the amplification method, it is less prone to false positive variant callings, suitable for large target

regions, and can identify not only SNVs and small indels, but also some structural variants. However, the capture method is usually more expensive and laborious, and thus if the region of interest is small, and the aim is only to call simple types of variants, the amplification method can be a proper choice. In **Paper IV**, whole exome libraries of the bone marrow specimens were prepared using the Nextera Rapid Capture Exome Kit (Illumina), which is a capture-based whole exome enrichment method. Paired-end sequencing with 2×150 bases were performed on a NextSeq 500 sequencer, resulting in a median sequencing coverage of 152X. The reads were mapped to the reference genome version GRCh37 (hg19) using BWA 0.7.9a [297] and PCR duplicates were removed by SAMBLASTER [298]. Somatic variants were called using Strelka [299] with cultured fibroblasts as germline control samples. Variants with allele frequencies above 3% were kept as preliminary candidates for the downstream IBSAFE assay design. The WES experiments and variant calling were done by our collaborator's research group [300].

## DNA purification from the follow-up samples

No good mutation detection result in the follow-up samples can be produced without effective and efficient DNA purification from the samples.

The type of follow-up samples in **Paper II** and **III** was plasma. The plasma samples were separated from venous whole blood collected at different follow-up visits. The blood collection tubes used in **Paper II** were the BD Vacutainer® K2EDTA tubes (Becton Dickinson). EDTA (ethylenediaminetetraacetic acid) is an in vitro anticoagulant widely used for clinical purposes [301, 302], which is dried coated to the walls of the BD K2EDTA tubes. Use of EDTA as an anticoagulating preservative in blood collection is recommended by Clinical Laboratory Standards Institute (CLSI) and International Council for Standardization in Hematology (ICSH). After blood collection, the tubes were either processed within 2 hours, or stored at 4 °C and processed within 24 hours. The BD K2EDTA tubes, though, were designed for generic blood collection, not specifically optimized for cfDNA and CTC preservation and analysis. The fact that blood collected in this type of tubes must be kept chilled at all times and processed rapidly adds uncertainties to the quality of the purified cfDNA. Moreover, the residue of the coating chemical EDTA may act as an inhibitor in the downstream PCR analysis [303]. Due to these reasons, after reading the literatures and testing other types of blood collection tubes, the Streck Cell-Free DNA BCT were chosen as the blood collection tube for cfDNA analysis in **Paper III** [304]. Comparison studies have shown that the Streck tubes can preserve cfDNA at room temperature (6 °C -37 °C) for up to two weeks, and that their secret recipe of preservatives can suppress white blood cells lysis, limiting genomic DNA contamination of cfDNA [305-307]. For CTC analysis in **Paper III**, CellSave Preservative Tubes (Menarini Silicon Biosystems) were used to collect the

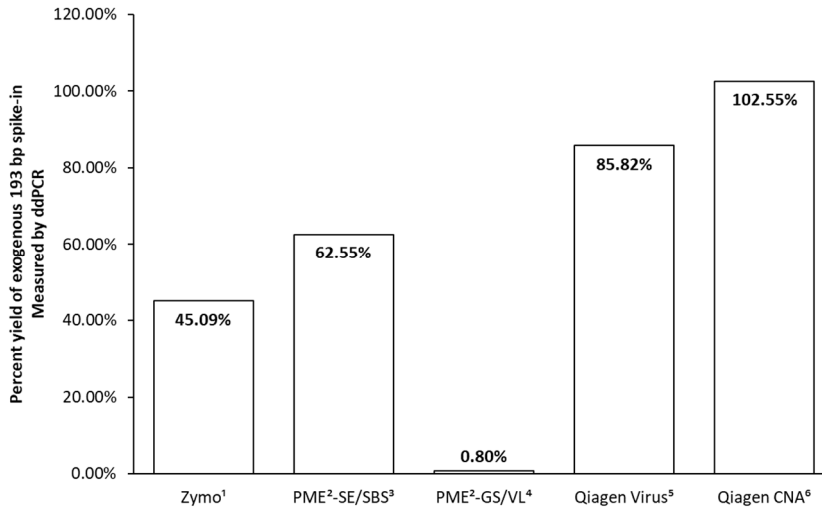
blood due to their exclusive compatibility with the downstream CTC analysis procedures [308]. The CellSave tubes are coated with cellular stabilizing chemicals that prevent cells from degrading for up to 96 hours [308-310].

Separation of plasma from whole blood, termed blood fractionation and plasma collection, for cfDNA purification, was basically done the same way in **Papers II** and **III**. The blood collection tubes were centrifuged at 2,000 relative centrifugal forces (rcf) for 10 or 15 minutes with a slow acceleration and deceleration at the beginning and the end of the centrifuge program to avoid disturbance of the fractions. The top fraction of plasma was transferred to 1.5 mL mini spin tubes and further centrifuged at 10,000 rcf to deplete any leftover cells. The middle layer of white blood cells was saved as a source of personalized normal genomic DNA for exclusion of germline mutations in the downstream experiments, and the bottom fraction of red blood cells was discarded. All fractionation and collection steps were done at 4 °C in **Paper II**, and at room temperature in **Paper III**.

cfDNA extractions in **Paper III** were done using the QIAamp UltraSens Virus Kit from Qiagen following the standard protocol with minor modifications. 1 µg of ploy(A) carrier RNA was used per extraction, regardless of input plasma volume, to increase the yield of cfDNA [311, 312]. For **Paper IV**, a proteinase K treatment step at an elevated temperature for extended duration was determined to improve quality and yields. Proteinase K is a broad-spectrum serine protease widely used in purification of nucleic acids [313, 314]. Without this step, downstream PCR will be inhibited by the residual chemical in the purified cfDNA sample, hurting the sensitivity of mutation detection [310, 315]. The Virus kit reagents do not endure warmer temperatures, and therefore relevant studies were referred to and several extraction kits were compared in-house to find a replacement solution. In the end, Qiagen's QIAamp Circulating Nucleic Acid (CNA) Kit was selected as a substitute for the Virus kit for its outstanding performance [316]. The CNA kit protocol features a step of proteinase K incubation at 60 °C for at least 30 minutes, and has all other steps done at room temperature, fully compatible with the Streck tubes. In addition, the CNA kit not only utilizes a silica membrane to capture the cfDNA – carrier RNA complex with high efficiency like the Virus kit, but also takes a maximum 5 mL plasma as input per extraction, as opposed to only 2 mL for the Virus kit (Figure 11), allowing for a condensation of cfDNA in the purified sample, potentially leading to a higher sensitivity of ddPCR mutation detection.

As for CTC in **Paper III**, a brief isolation was done to collect CTCs from whole blood. As breast carcinoma was the type of cancer investigated in this study, CTCs all express the transmembrane glycoprotein epithelial cell adhesion molecule (EpCAM) [317, 318], distinguishing them from the overwhelmingly outnumbering population of leukocytes in the specimens. For this reason, a ferrofluid-conjugated antibody recognizing EpCAM was used to selectively bind to the CTCs, followed by a capture using magnetism. These briefly purified CTC samples were subjected to analysis described in the Circulating Tumor Cell (CTC) analysis section.

In **Paper IV**, bone marrow specimens were obtained via aspiration from the pelvis at follow-up visits according to routine clinical protocols at Department of Pathology, Lund University. The specimens were collected in heparinized tubes, from which mononuclear cells (MNCs) were isolated using the Lymphoprep Density Gradient Medium (Stemcell Technologies). DNA was purified from the MNCs with the QIAamp DNA Blood Mini Kit (Qiagen) following the standard protocol [188].



**Figure 11 Comparison of cDNA extraction kits.** A synthetic double-stranded exogenous DNA fragment at 193 bp was spiked into the plasma before extraction. The concentrations of the spike-in molecule in the source and in the purified cfDNA sample were measured by ddPCR, and the percent yield of each kit was calculated by dividing the latter by the former. Percent yields of the Qiagen kits were much better than other candidates, with the CNA kit surpassing 100%. This observation could be caused by errors from pipetting and measurements. 1. ZR Viral DNA Kit™, D3015, Zymo Research; 2. PME free-circulating DNA Extraction Kit, 845-IR-0003010, Analytik Jena; 3. Lysis system SE/binding system SBS; 4. Lysis system GS/binding system VL; 5. QIAamp UltraSens Virus Kit, 53704, Qiagen; 6. QIAamp Circulating Nucleic Acid Kit, 55114, Qiagen

## Multicolor Flow Cytometry (MFC)

Flow cytometry is a laboratory technology to sequentially determine optical and fluorescence characteristics of cell-sized particles in a fluid stream [319]. In a flow cytometer, a laboratory instrument that does the analysis, cells in a sample are resuspended in a saline solution and pass through a checkpoint one by one, where certain characteristics of the cells are measured. Modern flow cytometers are usually able to measure multiple variables simultaneously [320]. For example, when a cell passes through the laser beam at the checkpoint, optical information for the cell is generated in that the cell causes the laser beam to scatter in multiple directions. The

amount of forward scatter (FS) is proportional to the size of the cell and the amount of side scatter (SS) is correlated to the complexity of the cell's shape. Likewise, fluorescence signals emitted from fluorophore-conjugated antibodies that bind to certain antigens on the cell could also be detected by a flow cytometer. Using the optical and fluorescence information, profiles of the cells can be established, and different populations of the cells can thus be identified.

Flow cytometry has been widely applied in clinical practices in hematology [321, 322]. In **Paper IV**, multicolor flow cytometry (MFC) was used to establish the phenotypic profile for the leukemic myeloblasts found in the diagnostic bone marrow aspirate. This profile, called the leukemia-associated immunophenotype (LAIP), was used to identify the population of leukemic myeloblasts in the follow-up bone marrow samples. The detailed steps were described in Pettersson et al [188]. Briefly, myeloblasts were defined as CD45 positive cells [323]. FS and SS information was also used to filter away cell debris while retain the intact hematopoietic cells. A range of monoclonal antibodies targeting blast markers, myeloid antigens, and aberrantly expressed markers, tagged with fluorophores in different colors, were applied in both the leukemic bone marrow and the normal cells to find the aberrant immunophenotype of the leukemic myeloblasts [324]. The antibodies used to establish the patient specific LAIPs include those against CD56, CD13, CD34, CD117, CD33, CD11b, HLA-DR, CD36, CD64, CD14, CD15, NG2, CD19, CD7, CD2, CD7, CD96, CD123, CD38, CD99, CD135, CD133, and CD4. All antibodies were ordered from Becton Dickinson and Beckman Coulter. The flow cytometry was done on a Beckman Coulter Navios flow cytometer, and data was analyzed with the Kaluza software (Beckman Coulter).

As reviewed by Jaso et al, the achievable LoD of MFC-based MRD detection is only in the range between 0.1% and 1% leukemic blasts, limited by the level of background noise, total number of cells analyzed, and the distinction of the LAIP comparing with the phenotype of the normal myeloblasts [325]. Moreover, the LAIP of a leukemia may change between the diagnostic and the relapse time-points, adding an extra layer of complexity in detection of MRD using the MFC [180, 181].

## Circulating Tumor Cell (CTC) analysis

The immunomagnetically selected CTCs were analyzed with the US FDA approved CellSearch® system in **Paper III** [120]. The samples were stained with DAPI (4',6-diamidino-2-phenylindole), and hybridized with an allophycocyanin (APC) conjugated antibody recognizing CD45 and phycoerythrin (PE) conjugated antibody recognizing cytokeratin (CK) 8, 18, and 19 to further identify CTCs from leftover leukocytes for enumeration. DAPI is a fluorescent stain that binds to the minor groove of AT-rich regions in DNA and emits a blue fluorescence (461 nm)

upon excitation. It is widely used to reveal the cell nucleus [326-328]. CD45, also known as the leukocyte common antigen (LCA), as suggested by its name, is a protein tyrosine phosphatase expressed across the plasma membrane of essentially all leukocytes [329-331]. The APC conjugator emits red fluorescence (660 nm). CKs are keratins found in cytoskeleton of epithelial cells and has been used as markers for cells with breast carcinoma origin [332-335]. The PE conjugator emits orange-yellow fluorescence (575 nm). CTCs were defined as nuclear cells (DAPI+) that lack CD45 (CD45-) and express CKs [336]. The CTCs were automatically sorted and counted by the CellTracks Analyzer II instrument [337], and were manually confirmed by two independent technicians. Samples with CTC counts  $\geq 1$  per 7.5 mL of original blood specimen were regarded positive for CTC [338].

## Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR), since its invention in mid 1980s by biochemist Kary Mullis and his research team, has developed into an indispensable tool in various aspects of molecular biology research, genetic engineering, a variety of clinical practices, and forensic approaches [339-342]. Outside of the scientific community, PCR may not have been a particularly familiar term to the general public until the COVID-19 pandemic, in which real time reverse transcription PCR (real time RT-PCR, or rRT-PCR), a modern variation of the classic PCR, was used as a gold standard for the detection of active infection caused by this single-stranded RNA virus [343-345].

A shared characteristic of essentially all types of PCR is that the reaction takes a small amount of nucleic acid sample (DNA or RNA), called template, as input, and makes copies of it exponentially. Suggested by its name, the cornerstones of PCR are 1) a DNA polymerase that has acceptable fidelity and can sustain the reaction conditions throughout its duration and 2) a laboratory instrument capable of carrying out the conditions in which the chain reaction could happen.

The first requirement had been met even before the invention of PCR, when the thermostable DNA polymerase *Taq* was purified from *Thermus aquaticus* [346]. Following the discovery of *Taq* polymerase, other DNA polymerases with improved fidelity, such as *Pfu*, were discovered and engineered for PCR [347, 348]. Other essential components of PCR include

- a. A pair of short single-stranded DNA flanking the template sequence, called primers, as DNA polymerases require them to bind to the template and start the DNA base incorporation.
- b. Deoxynucleoside triphosphate (dNTP) at an excessive amount. This is a collective term given to the four molecules of such type – dATP, dGTP,



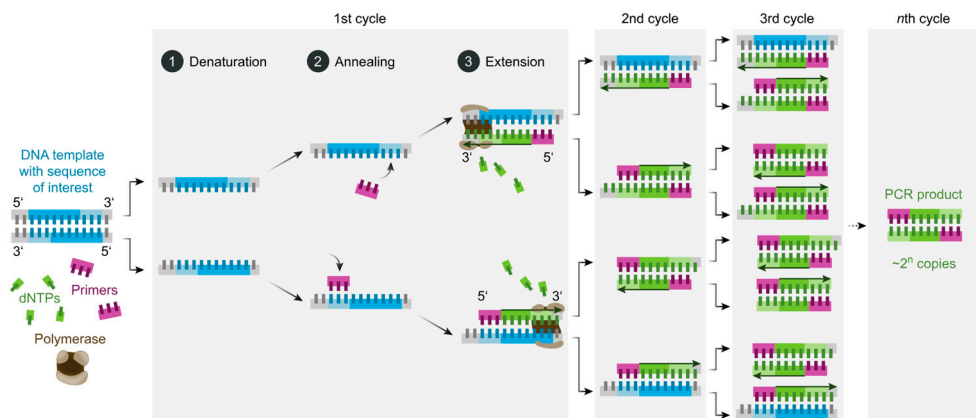
dCTP, and dTTP. The molecules are all composed of a ribose backbone connected to three consecutive phosphate groups, and a nucleobase of either adenine (A), guanine (G), cytosine (C), or thymine (T). dNTP is the material from which DNA is synthesized in PCR.

- c. A buffer solution in order to provide a stable chemical environment for the polymerase to work. Nowadays, the reaction buffer often contains bivalent and monovalent cations, such as  $Mg^{2+}$  and  $K^{+}$ , at certain concentrations to improve the stability and fidelity of the polymerase.

Since the PCR components of polymerase, dNTP, and buffer are almost universally applicable, they are usually premixed and sold under the name of *Super Mix* or *Master Mix* by many modern vendors.

To meet the second requirement, so that the original template can be amplified in an exponential fashion, the PCR is carried out in repeated reaction cycles, within each of which the newly synthesized DNA strand in the previous reaction cycle is added to the repository of template molecules, and thus doubling the speed of DNA synthesis per reaction cycle. A schematic depicting basic principles of PCR is shown in Figure 12. Each reaction cycle typically features

- a. A *Denaturation* step in which double-stranded DNA is transformed into single strands so that DNA primers can bind
- b. An *Annealing* step that allows for binding of the DNA primers and the polymerase to the template DNA
- c. An *Extension* (also known as *Elongation*) step in which the polymerase becomes enzymatically active and incorporates dUTP, according to the template DNA sequence, to synthesize the new DNA strand



**Figure 12 Schematic of PCR mechanism**

From [https://commons.wikimedia.org/wiki/File:Polymerase\\_chain\\_reaction-en.svg](https://commons.wikimedia.org/wiki/File:Polymerase_chain_reaction-en.svg) CC BY-SA 4.0.

Shifting between the steps requires rapid and accurate manipulation of reaction temperatures, which is achieved by the lab instrument called thermocycler or thermal cycler. Modern thermocyclers are typically able to change the reaction temperatures at a rate of up to 5 °C/second, and reach the target temperatures with less than  $\pm 0.3$  °C [349, 350].

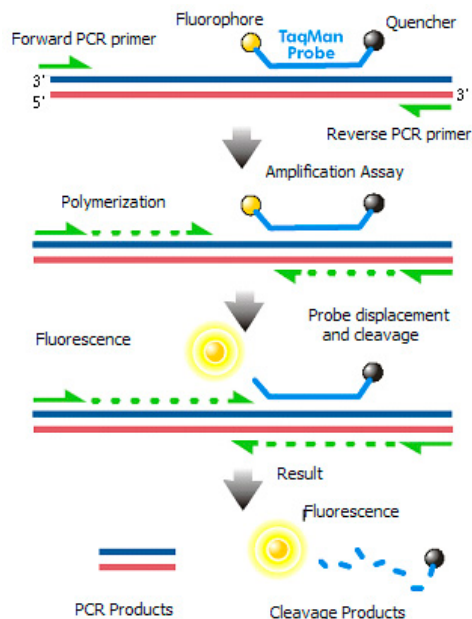
Given the characteristics of PCR, a common usage of this technology is naturally to enrich the nucleic acid template originally at a low concentration or total amount so that the amplified template DNA molecule can be retrieved and used as input for the downstream laboratory processes. This type of PCR is necessary for building sequencing libraries when the original amount of nucleic acid to be sequenced is too low to be loaded on a sequencer. Sequencing library enrichment by PCR was performed in **Papers II, III, and IV**.

PCR is also used to detect the presence of nucleic acid template molecule of interest in the given sample, qualitatively and/or quantitatively. To achieve this purpose, fluorescence signal reporters that bind to nonspecific double-stranded DNA molecules (PCR products) are often added to the reaction. Frequently used fluorescence dyes include ethidium bromide (EtBr) [351, 352], SYBR Green I [353, 354], and EvaGreen [355, 356]. The PCR products are often analyzed with agarose gel electrophoresis subsequently, which separates double-stranded DNA molecules into different bands by their lengths. The fluorescence signal intensity within each band is proportional to the amount of DNA with the dye intercalated, hence serving as a semi-quantitative surrogate of DNA amount.

An alternative approach to detect PCR products, instead of having a fluorescence dye that unselectively binds to all double-stranded DNA molecules, is to include a piece of single-stranded DNA tagged with a fluorophore and a nonfluorescent quencher (NFQ), sometimes called a TaqMan probe [357]. When the fluorophore and the quencher are tethered by the intact probe, they are in close spatial proximity, and the fluorescence is inhibited. The TaqMan probe has a sequence complementary to that of the DNA template and is flanked by the DNA primers. Mechanism of TaqMan probe in PCR is depicted in Figure 13. In the annealing step, the probe binds to the template. In the extension step, the polymerase not only incorporates dNTP according to the reverse complement strand's sequence, but also hydrolyzes the probe using its 5'-3' exonuclease activity, spatially separating the fluorophore and the quencher, and thus create a detectable fluorescence signal when excited by an energy source such as a laser. This approach, comparing with probe-free PCR, significantly improves the specificity of the technology, which is critical for reliable detection of SNVs, small indels, and genomic translocations at low abundance, where the intended DNA template shares high degree of sequence similarity with the background DNA at an overwhelming amount. In recent years, base modifications, such as minor groove binder (MGB) [358, 359], and nucleic acid analogs, such as locked nucleic acid (LNA) [360, 361] in primers and probes have been shown to further increase the

specificity of PCR. A mixture of the primer pairs and one or several probes with a well titrated ratio at a desired concentration is called an assay.

The variations of PCR used in this thesis were quantitative PCR (qPCR) in **Paper IV** and digital PCR (dPCR) in **Papers II, III, and IV**. Details about qPCR and dPCR are described in the following sections.



**Figure 13 Schematic of mechanism of TaqMan probe**

From <https://commons.wikimedia.org/wiki/File:Taqman.png> public domain.

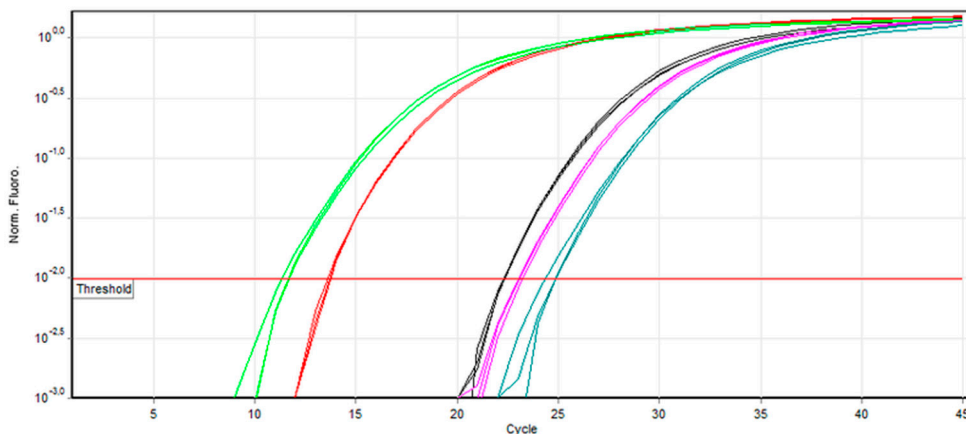
## Quantitative PCR (qPCR)

To accurately quantify the target DNA molecule, in the case of **Paper IV**, the *NPM1* type A insertion [362], the real-time quantitative PCR (qPCR) technology was employed. Based on regular endpoint PCR, qPCR adds a step at the end of each PCR cycle to measure the fluorescence signal intensity of the reactions [363, 364]. Empirically, the fluorescence signal intensity increases first in a geometric manner, followed by a linear phase, until it hits the plateau. The intensities, on a logarithmic scale, are often plotted against the number of cycles, and a threshold intensity is defined within the geometric phase cycles (Figure 14). The exact cycle number when the threshold intensity is reached is defined as the cycle of quantification ( $C_q$ ),

also known as the threshold cycle ( $C_t$ ). The quantity of the target DNA molecule in the original sample has a relationship to the  $C_q$  value given by Equation (1)

$$Quantity \sim (c \cdot 2)^{C_q} \quad (1)$$

where  $c$  stands for efficiency of the PCR (0-100%).



**Figure 14** An example of logarithmically transferred intensity values versus cycle numbers. Threshold is set

From <https://commons.wikimedia.org/wiki/File:Qpcr-cycling.png> CC BY-SA 3.0.

Typically, a standard curve of  $C_q$  values versus known concentrations of the DNA molecule in a set of serially diluted control samples needs to be constructed for any qPCR experiment. For a tester sample, the concentration of the DNA molecule can be queried on this standard curve when the  $C_q$  value of the sample is determined. In **Paper IV**, DNA extracted from a NPM1 type A insertion positive cell line, OCI-AML3 (DSMZ # ACC 582, Leibniz Institutes) was diluted in DNA extracted from a NPM1 wild-type cell line, NB4 (DSMZ # ACC 207, Leibniz Institutes) at a constant total DNA concentration of 100 ng/ $\mu$ L, to plot the standard curve.

The assay design and PCR conditions were detailed by Pettersson et al. in a previous study [188]. In short, 250-500 ng of bone marrow DNA was tested per follow-up sample for the insertion. The concentration of albumin, a housekeeping gene at a constant concentration, was also measured using methods published before, to normalize the measured NPM1 type A insertion concentrations [187, 365]. As heparin in the coated collection tubes is a known PCR inhibitor [366], bovine serum albumin (BSA) was added to the PCR reactions at a final concentration of 0.32  $\mu$ g/mL to alleviate the inhibition according to a previously published protocol [367].

The percentage of the residual leukemic DNA content can be calculated by Equation (2)

$$MRD\% = \frac{C^{NPM1 \text{ type A insertion}}}{C^{albumin}} \times 100\% \quad (2)$$

where  $C^{NPM1 \text{ type A insertion}}$  is the concentration of *NPM1* type A insertion and  $C^{albumin}$  is the concentration of albumin, measured by qPCR. Given the high amount of total DNA input, and the fact that the type A insertion is not as similar to the wild-type allele as a SNV would be, the *NPM1* type A insertion qPCR assay was able to detect MRD down to 0.001%. However, if the mutation is a SNV, or the input amount is not as much, the LoD will be affected. Detection of SNVs and small indels with IBSAFE ddPCR is discussed in the following sections.

### Digital PCR (dPCR)

Digital PCR (dPCR) is a powerful variation of traditional endpoint PCR. The original concept of dPCR was to dilute the nucleic acid analyte and evenly aliquot the diluted sample into many replicates, such that there is either zero, one, or a few target molecules in each component reaction (compartment). After amplification, the results of the component reactions are to be recorded in a dichotomous manner – either negative for those that contain zero target molecule, or positive for those that contain any non-zero number of target molecules [368, 369]. The technology also got its name “digital” because of the discrete and binary nature of its results, where each compartment is scored as positive or negative, or 1 or 0. As each target nucleic acid molecule ends up in one of the component reactions independently and randomly, the numbers of the target molecules in the component reactions, according to probability theory, are distributed following the Poisson distribution. Derived from the probability mass function (PMF) of the Poisson distribution, the probability  $P(X = k)$  for a component reaction to contain  $k$  ( $k = 0, 1, 2, \dots$ ) target molecules, when the average number of target molecules per component reaction is  $\lambda$ , is given by Equation (3)

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \quad (3)$$

where  $e$  is the base of the natural logarithm ( $e = 2.71828$ ) and  $!$  is the factorial function sign. It is self-evident that when the values of  $P(X = k)$  and  $k$  are set, the value of  $\lambda$  can be uniquely determined. Let  $k = 0$ ,  $P(X = k)$  becomes the probability for component reaction to contain 0 target molecule, represented by the occurrence rate of negative component reactions. The probability  $P(X = k)$  can be estimated using the calculation given by Equation (4)

$$P(X = 0) = \frac{N}{N+P} = 1 - \frac{P}{T} = e^{-\lambda} \quad (4)$$

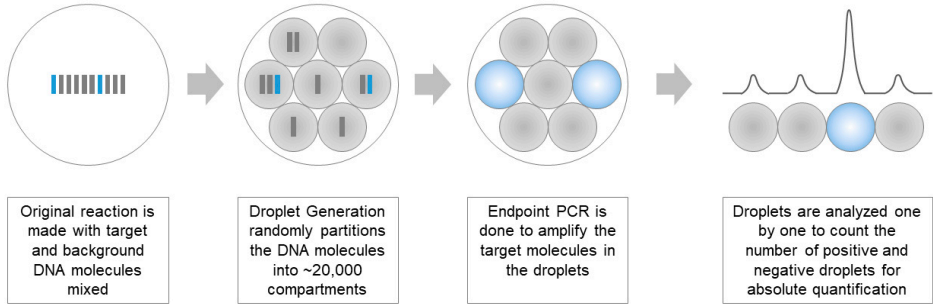
where  $N$  is the observed number of negative component reactions,  $P$  is the observed number of positive component reactions, and  $T$  is the total number of component reactions ( $T = N + P$ ). The average number of target molecules per component reaction,  $\lambda$ , can thus be calculated using Equation (5)

$$\lambda = -\ln\left(1 - \frac{P}{T}\right) \quad (5)$$

At this point, the advantage of dPCR has emerged, in that it can directly measure the absolute copy number concentration of the target molecule, whereas other methods like qPCR makes relative quantification based on a standard curve built from measurements in external control samples with known target molecule concentrations. However, to achieve an acceptable statistical power so that the objective occurrence rate  $P(X = 0)$  can be accurately calculated from the empirically observed numbers,  $N$  and  $P$  must be large. In addition, to make results of dPCR robust and reliable, the volume of the component reactions must be highly consistent, and the limiting dilution steps must be free from sample contamination. For these reasons, despite being a conceptually tempting new technology, the wide application of dPCR in research had long been hindered by its laborious, costly, and technique-demanding workflow. The situation changed in early 2010s, when technological advances led to the launching of droplet digital PCR (ddPCR) platforms, such as the QuantaLife QX100 system [370] and the RainDance system [371] (both were later acquired by Bio-Rad), which create component reactions in the form of nanoliter- to picoliter-sized compartments called droplets. As the ddPCR experiments in this thesis were all done with the QX system (QX100 and QX200), the discussions about ddPCR hereafter are all about this system unless specified otherwise.

The schematic of ddPCR workflow is shown in Figure 15. Generation of droplets, sometimes also called partitioning, is reproducibly done using the QX Droplet Generator, resulting in  $\sim 20,000$  droplets from a  $20\ \mu\text{L}$  reaction, with a high-degree uniformity in droplet volume. This machine-driven droplet generation step effectively eliminates the possibility of sample contamination, reduces variance in component reaction volume, and makes the dPCR practical and affordable. After thermocycling, the fluorescence signal results of the droplets are analyzed, at a speed of  $\sim 10,000$  droplets/minute, by the system's Droplet Reader [372]. With this streamlined workflow, ddPCR has become a readily viable laboratory tool for absolute quantification [373-376]. It also has been reported that ddPCR outperforms qPCR in gene expression analysis [377], copy number variation analysis [378], mutation detection [379], noninvasive prenatal testing [380], and even SARS-CoV-2 detection [381]. Moreover, ddPCR has a better potential to be multiplexed [382] and is more tolerant to PCR inhibitors than traditional endpoint PCR [383, 384]. The QX systems have channels to simultaneously detect blue (FAM) and green (VIC/HEX) fluorescence signals. Using this feature, assays targeting the variant allele (usually reported in the FAM channel) and the wild-type allele (usually

reported in the VIC/HEX channel) of the same genomic locus can be applied in the same reaction to measure the absolute concentration of both alleles at the same time, and thus the variant allele frequency (VAF) can be calculated.



**Figure 15 Schematic of ddPCR workflow**

With the QX system's setup, absolute copy number concentration of the target molecule (the variant allele molecule or the wild-type allele molecule) in the original sample  $C_{V_i}$  can be calculated by Equation (6)

$$C_{V_i} = \frac{\lambda}{V_d} \times \frac{V_r}{V_i} = \frac{-\ln(1-\frac{P}{T})}{V_d} \times \frac{V_r}{V_i} \quad (6)$$

where  $V_d$  is the volume of a droplet ( $0.91 \times 10^{-3} \mu\text{L}$  at the time of **Paper II**, later changed to  $0.85 \times 10^{-3} \mu\text{L}$  at the time of **Papers III** and **IV**, after the system's firmware was upgraded);  $V_r$  is the total volume of a ddPCR reaction ( $20 \mu\text{L}$ );  $V_i$  is the input volume of the sample. When the concentrations of the variant allele  $C_{V_i}^{\text{variant}}$  and the concentration of the wild-type allele  $C_{V_i}^{\text{wild-type}}$  are both measured,  $VAF$  can be calculated by Equation (7)

$$VAF = \frac{C_{V_i}^{\text{variant}}}{C_{V_i}^{\text{variant}} + C_{V_i}^{\text{wild-type}}} \times 100\% \quad (7)$$

Whether the absolute concentration of the variant allele  $C_{V_i}^{\text{variant}}$  or the variant allele frequency  $VAF$  is more clinically meaningful is debatable. On one hand, recovery rates of the upstream DNA extraction may vary between specimens even if the same protocol is used. Despite this, cfDNA molecules of both alleles are presumably affected equally. By using  $VAF$ , this variance in extraction efficiency is normalized, and therefore the variant allele burdens between different specimens can be compared. On the other hand,  $VAF$  in an original specimen may have already changed prior to DNA extraction. For example, when DNA is extracted from plasma, which in turn is fractionated from whole blood, in vitro hemolysis could

happen as the result of certain mishandlings of the whole blood sample. This could not only cause a deviation in measurements of certain circulating biomarkers [385], but also lyse the leukocytes and contaminate cfDNA in the plasma fraction with genomic DNA released from the lysed cells, resulting in an underestimated *VAF* [386]. In this situation,  $C_{V_i}^{variant}$  may be a preferred metric. In this thesis, *VAF* was chosen as the indicator for variant allele burdens in the plasma samples.

Although the mechanism of the ddPCR technology, by design, enables absolute quantification of target nucleic acid molecules, this promise cannot be fulfilled in real world without savvily designed assays, especially in situations where the target molecule is a DNA variant at low abundance that bear some degree of resemblance to the abundant background wild-type allele molecules. As plasma was the material for mutation detection in **Papers II, III, and IV**, which usually only contains thousands to tens of thousands of cfDNA genome-equivalents per mL of plasma [131, 387-392], the assays must be able to reliably detect single-digit numbers of variant allele molecules, while give low numbers of false positive signal, if not none, in true negative samples. Due to the difference in characteristics of the variant types investigated in the studies, assays were design with different strategies.

#### *Detection of chromosomal rearrangements*

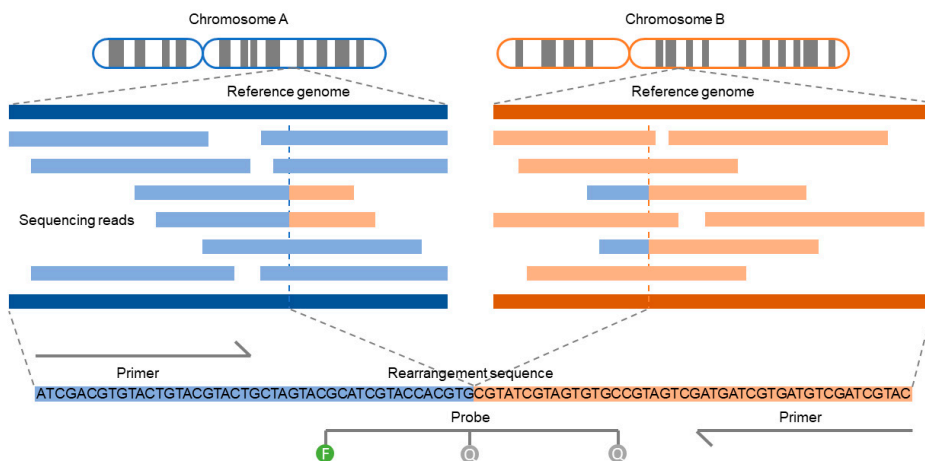
Assays were designed for detection of chromosomal rearrangements in **Paper II**. Details were published as a standalone chapter in *Digital PCR – Methods and Protocols*, a part of the *Methods in Molecular Biology* book series [273]. Figure 16 is a schematic to show the workflow from chromosomal rearrangements identification by low coverage WGS in the tumor tissue genome to ddPCR assay design.

Briefly, about 10 rearrangements per sample were initially selected from the list for ddPCR assay design. The rearrangements with a high number of reads that span across the junction point and/or a high number of read pairs that are mapped to two far-apart genomic locations were prioritized for assay development, for they are important indicators of the fidelity of the rearrangement. In addition, if there were still a lot of rearrangements to choose from, those that were mapped to different chromosomes were prioritized so that the chance that they represent different tumor subclones was maximized. The selected rearrangements were then visually inspected using the Integrative Genomics Viewer (IGV, Broad Institute) to exclude artifacts from bioinformatics [393].

For ddPCR assay design, the Primer Express 3.0.1 software (Applied Biosystems) was used following typical guidelines for TaqMan probe-based PCR assay design. As there was no guarantee that the exact sequence of the junction region can be reconstructed, a universal strategy was to place one primer and the probe on one side of the junction, and the other primer on the other side. For those rearrangements whose exact break point sequences were determined, the probe could also be placed



across the break point, instead of being solely on either side. The primers and the probe did not overlap each other, but were placed as close to each other as possible, so that the size of the PCR product in base pairs, called amplicon, could remain as small as possible. A small amplicon size is a prerequisite for detection of cfDNA with high sensitivity, as most cfDNA molecules are fragmented to ~160 bp [394, 395].



**Figure 16** Schematic of workflow from chromosomal rearrangements identification to ddPCR assay design

The biophysics of the primers and probes are as follows. Melting temperatures ( $T_m$ ) of the primers were between 58 and 60 °C, with the difference in  $T_m$  as small as possible. A G or C was placed at the 3' end of the primers to form a GC clamp, so that the specificity of the primer binding to its template was increased. As for the probe, the  $T_m$  was between 68 and 70 °C, ~10 °C higher than those of the primers, so that they can bind to the template before the binding of the primers and synthesis of the new DNA strand. The terminal 5' end of the probe was not a G for all probes, and specifically for probes tagged with the FAM fluorophore, the second base from 5' end was not a G either. The percentage of guanine (G) and cytosine (C) bases, called GC content, in both the primers and the probe were between 30 and 80%. Polynucleotide repeats, especially GGGG (4 G's), CCCC (4 C's), and AAAAAA (6 A's), were avoided in both types of DNA oligoes. In addition to all these, the DNA oligoes must remain single-stranded under reaction conditions for them to bind to the template and keep the PCR going. For this reason, the secondary structure of the oligoes was also checked with the software, and those with strong secondary structures of hairpin, self-dimers, and cross dimers were excluded from

the analysis. Due to the demanding biophysical requirements, not all selected rearrangements could have an assay designed.

For the designed assays, first the primer pairs, at 250 nM molar concentration each in reaction, were tested in 10 ng of tumor tissue DNA as positive control and the same amount of matched normal genomic DNA extracted from buffy coat leukocytes from the same patient as negative control. A “step-down” PCR program was used for all the primers with slight differences in their biophysics characteristics. Briefly, the annealing temperature of the PCR program was 70 °C in the first cycle, only to decrease by 1 °C per cycle, until it reached 60 °C, and last 29 extra cycles with the annealing temperature at 60 °C. The purpose of this setup was to let the most biophysically stringent primers to bind and amplify first, while still let as many types of primers as possible to start working in later cycles, so that the PCR program is specific and yet highly generic. The results were analyzed with the Caliper LabChip XT System (Perkin Elmer), a microcapillary agarose gel electrophoresis system, to select for primer pairs that can make PCR products in the tumor tissue DNA, and against those that also amplify in the matched normal DNA. The remaining rearrangements were somatic ones for which an assay could likely be validated. The probe was added to the reaction in the next round of assay validation for these assays, also at 250 nM in reaction. An assay was considered validated and ready to be used in follow-up DNA samples when the intensity of the signal in the positive control was high and the measured copy number concentration was concordant with WGS results, and the negative control had no false positive signal. Following the same assay design rules, an assay targeting a copy-number stable region in breast cancer located in chromosome 2, 2p14, was designed to measure the total number of genome equivalents analyzed per reaction. On average, ~5 somatic rearrangements per patient were monitored by the successfully designed assays, with a mean amplicon size of 101 bp.

The results could not be compared between samples and different time-points until certain normalizations. To normalize the results, ddPCR thresholds were set for the reactions in an unbiased, automated, reproducible way. The droplet amplitudes files were exported from the QuantaSoft software (Bio-Rad) for this normalization. Basically, the intensity of the highest droplet in the negative control was defined as negMax. For each reaction, droplets  $\leq 2 \times \text{negMax}$  were defined as lower droplets, and those  $> 2 \times \text{negMax}$  were upper droplets. Find the median of the lower droplets of each reaction and subtract it from all the droplet intensities to bring the lower droplets median intensity to 0. Then find the median of the upper droplets of the positive control reaction and divide all the droplet intensities of all reactions by it to set the upper droplet median intensity of the positive control to 1. Thus far, the droplet intensities of different assays in different samples were normalized. For thresholding, a receiver operating characteristic (ROC) analysis was done [396] and it was determined that a threshold at 0.5 after normalization gave the best

performance, and was therefore selected. An example of thresholding after normalization is shown in Figure 17.

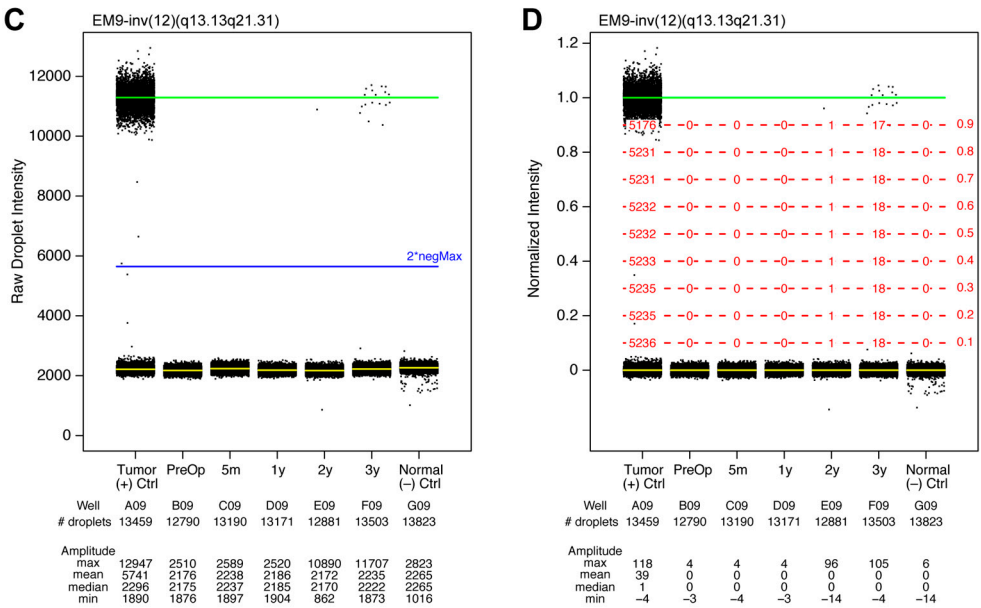


Figure 17 Example of unbiased ddPCR thresholding

### Detection of SNVs and small indels

Assays were designed for SNVs in **Paper III**, and for SNVs and small indels in **Paper IV**. SNVs, by definition, only involve a change in a single DNA base, and the local genomic sequence context is identical to the wild-type reference genome. Moreover, somatic SNVs are often present at a low allele frequency. The combination of high similarity and low abundance not only makes SNVs relatively difficult to be called from HTS data [397-399], but also evokes complexities in PCR-based mutation detection assay design.

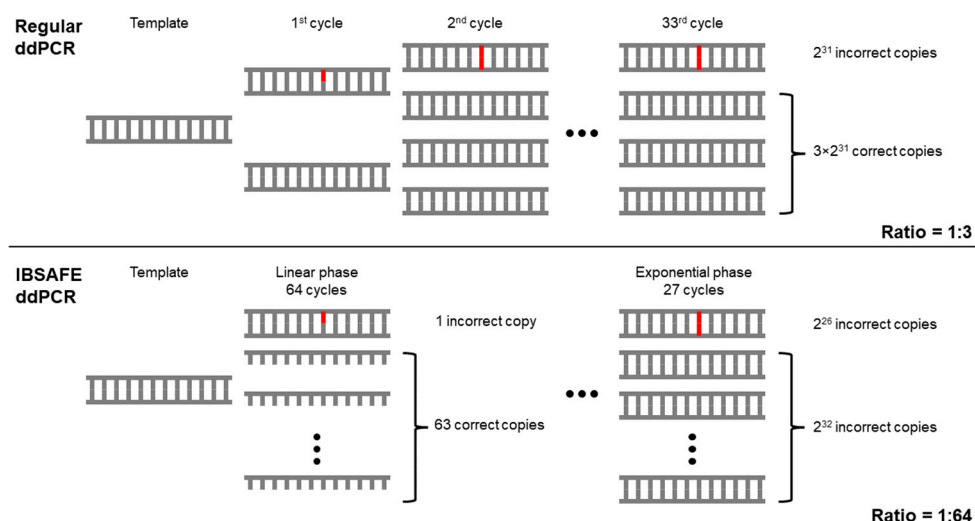
A typical PCR assay targeting a SNV has a variant-allele-specific probe covering the SNV. The probe has a perfect reverse-complementary sequence to the variant allele template sequence, the melting temperature of which, therefore, is termed the perfect match  $T_m$ . Besides the perfect binding, the variant allele probe is also to inevitably bind to the wild-type allele template with certain degrees of affinity, especially as the wild-type allele is often at a higher concentration in a given sample than the variant allele. The melting temperature of the probe against the wild-type allele, as there is a single-base mismatch, is called the mismatch  $T_m$ . It is obvious that the larger the difference between the perfect match  $T_m$  and the mismatch  $T_m$ , the higher specificity the assay has for the variant allele against the wild-type allele.

In practice, most assays cannot prevent the variant-allele-specific probe from binding to the wild-type allele with a reduced efficiency, and therefore a background fluorescence signal for the variant allele is to be created from the nonspecific binding in essentially all types of samples, including pure wild-type control samples, hurting the specificity of the mutation detection. To overcome this issue, researchers have suggested alternative assay design strategies, of which the amplification refractory mutation system (ARMS) is perhaps the most embraced one [400, 401]. Generally, the ARMS design strategy instructs researchers to put the SNV base at the 3' end of a primer, while have the second last base from the 3' end a mismatch to both the variant and the wild-type alleles. For the variant allele template, the perfect match of the 3' end base overrides the mismatch effect of the penultimate base, keeping the PCR rolling, whereas for the wild-type allele template, the combo effect of two consecutive mismatches at the 3' end of the primer effectively terminates the extension. This assay design strategy, however, is not a once-and-for-all solution to SNV detection using PCR in that the imperfect match between the variant-allele-specific primer to the variant allele template may reduce the sensitivity of the assay, especially in situations when the SNV is an adenosine or thymidine, precluding the possibility of forming an extension facilitating GC clamp structure at the 3' end of the primer.

To make things more complicated, all DNA polymerases are not free from base misincorporation errors [402, 403], and thus there is a chance that in the process of PCR a wrong DNA base is incorporated to the newly synthesized wild-type DNA strand at the position of the SNV. When the mistakenly incorporated base is that of the SNV, and the misincorporation event happens in early cycles of the PCR, the SNV is "created" in vitro and copied exponentially in the subsequent cycles, and eventually a false positive result will be reported. Although many modern thermostable DNA polymerases utilized in PCR have a low error rate in the range of 1 base misincorporation per 100,000 to 1,000,000 base pairs [347, 404], this advancement is counteracted by the fact that variant-allele molecules, if any, are usually at very low absolute concentrations and allele frequencies in cfDNA. Therefore, an ultrasensitive mutation detection method that is essentially free from false positives, despite the limited fidelity of all polymerases, is necessary for circulating tumor DNA detection in liquid biopsy to become a useful diagnostic and prognostic tool in clinical practices.

The IBSAFE technology was invented to achieve a better-than-ever sensitivity for digital PCR-based mutation detection. The technology has been used to reliably detect SNVs at single digit copy number concentrations in lung cancer [405], ovarian cancer [406], breast cancer [407] (and **Paper III**), and AML (**Paper IV**), and a manuscript on the ontology of the method is also in preparation [George AM, Chen Y, Saal LH, et al]. IBSAFE does not achieve its enhanced analytical performance via eliminating the innate polymerase base misincorporation error per se, but instead utilizes an alternative chemistry alongside a modified thermocycling

program to reduce the negative consequence of polymerase error. Defying canonical PCR assay design paradigms, the primers of an IBSAFE assay are deliberately designed with asymmetrical biophysical attributes, such that in each ddPCR droplet, correctly copied PCR product is enriched during a “linear” phase of 64 cycles when only one of the primers binds, followed by an “exponential” phase of 27 cycles when both primers and the probe bind. The primers and the probes were designed using IDT’s OligoAnalyzer Tool [408], which predicts the biophysical characteristics of the DNA oligoes based on previously published studies [409-415].

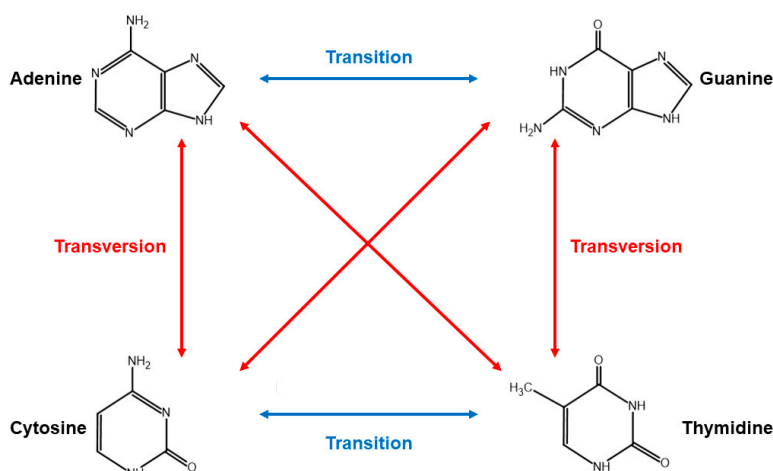


**Figure 18 Theoretical comparison between regular ddPCR and IBSAFE ddPCR performances.** This comparison features the worst-case scenario of polymerase error, that a base misincorporation event happens at the variant site in the first cycle in a droplet. The result shows that IBSAFE can effectively eliminate a false positive call for the droplet, whereas regular ddPCR cannot.

The amount of correct PCR product produced in the 64 linear cycles is equal to the amount that can be synthesized in 6 exponential cycles (as  $64 = 2^6$ ). As the DNA molecules are already partitioned into droplets prior to thermocycling, this linear step of enrichment de facto gives the true positive signal approximately 6 leading exponential PCR cycles over any potential false positive signal, without changing the number of positive and negative droplets. This way, the false positive signal, if any, can be effectively suppressed, and the quantity of the true positive variant allele molecule can be accurately measured. Figure 18 shows a theoretical performance comparison between a regular ddPCR and an IBSAFE droplet, when the template is wild-type DNA, and a base misincorporation event happens exactly at the variant position in the first cycle. In the regular ddPCR droplet, a quarter of the final fluorescence signal would support the existence of a false positive variant, whereas

in the IBSAFE droplet, it would only be 1/128. Note that the worst-case scenario for the base misincorporation error to happen is when it happens in an early cycle of the PCR. Such errors work as a founding synthetic mutation, causing amplified false positive signal in the final result. Therefore, such situations are where IBSAFE has the most benefit over regular ddPCR experiments.

Different types of SNVs need different IBSAFE assay design strategies. SNVs are divided into two classes: 1) *transition*, which involves base changes within the purines (adenine and guanine) or the pyrimidines (cytosine and thymine), and 2) *transversion*, which changes the type of the base from purine to pyrimidine or the other way round (Figure 19). Although there are twice as many types of transversion as transition, transitions are observed at a much higher frequency between species in the course of evolution [416, 417], and usually occur at an elevated rate in various types of tumor tissue [184, 418] and cancer cell lines [419]. Evolutionarily, when transitions occur at the third base of a codon, known as the “wobble” position, the amino acid could remain the same in many cases, reducing the natural selection pressure against these mutations. In biochemistry, changes between the same type of bases (“one ring” for the purines and “two rings” for the pyrimidines) are more likely to happen.



**Figure 19 Transition and transversion of DNA base changes**

The same underlying mechanisms that lead to the prevalence of transitions in vivo, perhaps, are also the factor to make transition the type of mutation vulnerable to false positives in ddPCR mutation detection in vitro. Conventional ddPCR mutation detection assays, when the target is a transition, usually has a background false-positive level of at least 0.01% VAF, whereas IBSAFE can reliably detect transitional SNVs down to 0.005% VAF and lower. See Figure 20 for an

experimental performance comparison between a Bio-Rad off-the-shelf assay versus an IBSAFE assay, both targeting the *EGFR* resistance mutation p.T790M.

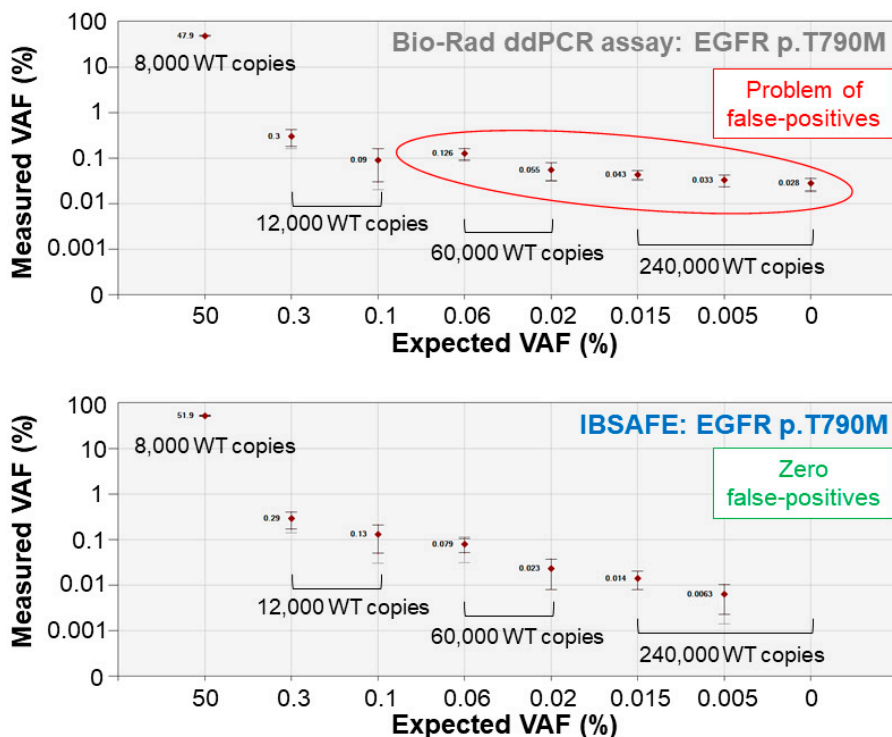


Figure 20 Comparison between Bio-Rad and IBSAFE assays targeting EGFR p.T790M

Although our data has suggested that in principle, IBSAFE has an LoD approaching 0.001% VAF, the performances of different assays are still slightly different. Therefore, determination of the analytical performance of each individual ddPCR assay provides a closer insight on to what degree the analysis results using the assay are to be trusted. The Clinical and Laboratory Standards Institute (CLSI) has published a series of clinical method evaluation standards to instruct establishment of analytical methods. According to the document EP17, Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures, at least 60 technical replicates with permutations in multiple variables such as control sample lot, date of experiment, person to run the experiment, etc. are needed to establish limit of blank (LoB), limit of detection (LoD), limit of quantitation (LoQ) etc. for the method. Despite its extreme rigorousness, this guideline is not easy to carry out, as the workflow is quite costly and laborious, especially for academic projects in which

a specific assay may only be used one or a few times. Armbruster and Pry published a paper in 2008 to simplify the workflow [420], but it still requires relatively large amounts of experimental data. For **Papers III** and **IV** and other ongoing research projects involving the IBSAFE technology, an approximation method was used to calculate LoB and achievable LoD for the assays using droplet level false-positive rate measured in wild-type control samples.

According to definition, LoB of an IBSAFE assay is the highest apparent VAF that could be detected in a wild-type control sample, or in other words, the upper 95% confidence interval of the droplet-level false positive rate. The droplet-level false positive rate of an assay can be determined by running the assay in a large amount of wild-type control sample. If zero false positive droplet is observed in the wild-type control sample, the number of false positives is arbitrarily adjusted to 1, and the number of total genome equivalents is multiplied by  $e$ , the base of the natural logarithm ( $e = 2.71828$ ). Let the adjusted number of false positives be  $N_{FP}$  and the adjusted number of total genome equivalents be  $N_{Genome}$ . The LoB is the lowest true positive VAF that has 5% chance to have no more than  $N_{FP}$  true positive molecules sampled when  $N_{Genome}$  total copies of genome equivalents are tested. In terms of LoD, it is defined as the lowest objective VAF that could be reliably distinguished from the LoB. As most of the validated IBSAFE assays had LoBs lower than 0.005%, the limiting factor of ultrasensitive mutation detection became the amounts of available DNA. Thus, the achievable LoD was calculated as the lowest true positive VAF that allows at least one copy of the variant allele molecule to be sampled into the reaction, given the amount of available total copies of genome equivalents. The achievable LoD must also be higher than the upper 95% confidence interval of the LoB, otherwise the latter was taken as the achievable LoD.

In **Paper III**, the validated IBSAFE assays were applied in ~20% of the total cfDNA extracted from ~5 mL of plasma on average, or ~1 mL plasma equivalent. On average, 3,628 total copies of genome equivalents (range 415-35,965 copies) were analyzed per sample, in line with plasma cfDNA concentrations reported in literature [131, 387-392]. In **Paper IV**, the assays were used in two replicate reactions containing 60 ng of bone marrow DNA each (or ~36,000 copies of genome equivalents). The achievable LoDs for these cohorts of samples were determined to be at ~0.028% and ~0.003%, respectively.





# Results and Discussion

## Paper I

### **PTEN and NEDD4 in human breast carcinoma**

*PTEN* (Phosphatase and tensin homolog) is an important tumor suppressor gene with multiple functions. In the cytoplasm, with its protein tyrosine phosphatase activity, the PTEN protein catalyzes dephosphorylation of phosphatidylinositol (3,4,5)-triphosphate (PIP3) to become phosphatidylinositol (4,5)-biphosphate (PIP2), the inverse reaction that the phosphoinositide 3-kinase (PI3 kinase, PI3K, also known as phosphatidylinositol 3-kinase) catalyzes to activate AKT and the downstream cell survival signaling pathway, effectively antagonizing the oncogenic PI3K/PTEN signaling pathway. When localized in the nucleus, the PTEN protein is involved in maintenance of genomic integrity. Germline mutations in *PTEN* cause PTEN hamartoma tumor syndrome (PHTS) and are associated with an increased risk of development tumor in breast, thyroid, and endometrium, among other potential sites. Moreover, somatic loss of heterozygosity (LOH) of *PTEN* have been observed in essentially all types of cancers, with an overall estimated prevalence of 30%. The key positive regulator of the PI3K/PTEN pathway, PI3 kinase, is the antagonist of PTEN in regulating the activity of this pathway. PI3 kinase has its catalytic subunit encoded by the oncogene *PIK3CA*, which has activating mutations in ~30% of all breast tumors. The mutation rate of *PTEN*, in contrast, is about 5%. Despite the low mutation rate, PTEN protein is lost in at least 25% of breast tumors and this loss-of-PTEN-protein phenotype and *PIK3CA* mutations are almost mutually exclusive [144].

Various mechanisms of loss of PTEN have been hypothesized and investigated, including gene mutations, copy number loss, chromosomal rearrangements, epigenetic silencing, as well as post-translational downregulation, but it was, and, to my best knowledge, still is, largely unknown in breast cancer.

One mechanism was theorized in mouse prostate and human bladder cancer models that NEDD4 (neural precursor cell expressed, developmentally down-regulated 4, E3 ubiquitin protein ligase) catalyzes poly-ubiquitination of the PTEN protein in the cytosol, and thus lead to proteolysis of the PTEN protein [151]. This theory of PTEN protein degradation by NEDD4-mediated poly-ubiquitination has been observed in other cases such as axon branching [152, 153], T-cell activation [154], keloid

formation [155], and insulin-mediated glucose metabolism [156], and an inverse correlation between *NEDD4* and *PTEN* expression levels were also found in human non-small cell lung cancer [157] and colon cancer [158] cohorts. However, numerous reports also found evidence to contradict this mechanism [159-163].

We attempted to investigate whether NEDD4 is a negative regulator of *PTEN* expression in breast cancer in this study. Patients with gene expression microarray data available or tissue microarray (TMA) of FFPE tumor specimens were selected from a Swedish cohort into this study (N = 186). Immunohistochemistry (IHC) staining targeting the NEDD4 protein was done on the 132 TMA samples, of which 123 samples had had their PTEN protein levels scored in previous studies [144, 145]. After being semi-quantitatively scored, the samples were divided into groups of NEDD4 negative (N = 60, 45%) and NEDD4 positive (N = 72, 55%). NEDD4 protein has no correlation to progesterone receptor (PR,  $P = 0.12$ ), human epidermal growth factor receptor 2 (HER2,  $P = 0.12$ ), Nottingham Histologic Grade ( $P = 0.57$ ), and Ki-67 ( $P = 0.40$ ), but was correlated with estrogen receptor (ER,  $P = 0.0017$ ). When compared with the PTEN protein levels, a positively trended correlation, although not significant ( $P = 0.12$ ), was observed, with 77% of the NEDD4-positive samples being PTEN-positive samples, compared to only 64% of the NEDD4-negative samples being PTEN-positive. Gene expression data was retrieved for 105 samples, of which 42 had matched IHC results for NEDD4. A significant correlation was observed between the gene expression and protein levels of PTEN (N = 105,  $P < 0.001$ ) and NEDD4 (N = 42,  $P = 0.04$ ), indicating gene expression levels of both genes can be a good surrogate of their respective protein levels. *PTEN* mRNA levels were significantly correlated with *NEDD4* mRNA levels (N = 105,  $P = 0.03$ ) and NEDD4 protein levels (N = 42,  $P = 0.02$ ), in contrast to the theory that NEDD4 is a negative regulator of the PTEN protein.

These findings were confirmed in two independent breast cancer cohorts from NKI (N = 295) and TCGA (N = 970) where positively trended correlations between gene expression and/or protein levels of PTEN and NEDD4 were observed, with or without significance.

This study ruled out NEDD4 as a negative regulator of the PTEN protein in breast cancer.

## Paper II

### **Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease**

Twenty patients from a larger cohort of the Breast Cancer and Blood Study (BC Blood, Sweden) [191] enrolled in this project, of which 6 had long-term disease-free survival after a median of 9.2 years of follow-up (termed DF patients), and 14 had eventual metastasis 1.2-5.1 years after surgery (termed EM patients). For each patient, a low-coverage whole genome sequencing was done on DNA extracted from the primary tumor, in order to identify tumor-specific chromosomal rearrangements to be used as liquid biopsy biomarkers in the follow-up plasma samples. All patients had one tumor sequenced, except for patient EM6, who had bilateral primary tumors and therefore two tumors were sampled for this patient. For these 21 sequenced tumors, an average of 93 million DNA fragments were sequenced from both ends (range 54-160 million), resulting in a mean genome coverage of 5.3-fold (range 1.8-12.9) and a mean physical coverage of 15.6-fold (range 9.2-28.2). The raw sequencing data was analyzed by our bioinformatics pipeline of SplitSeq to identify intra- and inter-chromosomal rearrangements (see the appended method paper for details). In short, SplitSeq scans through the sequenced read pairs and identifies those pairs that A) the two component reads are both perfectly mapped to discordant positions in the genome, or B) one of the components has the starting part perfectly mapped to a genomic position different than the other half of itself as well as the other component of the read pair. Chromosomal rearrangements supported by two or more such sequencing read pairs were enumerated for each tumor (92 chromosomal rearrangements per tumor, range 21-305), and ddPCR assays were designed for rearrangement detection in follow-up plasma samples.

Considering the possible intra-tumoral heterogeneity, chromosomal rearrangements with different copy numbers, as supported by the number of sequencing reads, and mapped to as many different chromosomes as possible, were selected for ddPCR assay design to represent the potential subclones of the tumor. For the 21 tumors, a total of 237 selected candidate rearrangements were selected for preliminary assay design attempts. Limited by the complexity of the local sequences, 197 (83%) rearrangements were able to have their specific assays designed, and, after tested in matched normal DNA with conventional PCR, 167 rearrangements were confirmed to be somatic. Due to the limited available volume of the follow-up plasma samples, 4-6 rearrangements per tumor (122 assays in total by this standard, of which, 113 assays were successfully designed for final use) were selected to be followed-up with ddPCR. In addition to these assays targeting the rearrangements, and assay specific to a non-rearranged copy-number neutral region located at 2p14 was designed to measure the total loading amounts of normal genomic DNA in the ddPCR wells. Experiments were done to confirm that the ddPCR analyses are highly

linear over at least 3 orders of magnitudes of variant allele frequency (VAF) and can reliably detect tumor-specific rearrangements down to 0.01% VAF.

Circulating cell-free DNA was extracted from 93 plasma samples for patients in both groups for chromosomal rearrangement detection, of which, 29 samples had positive results detected down to 0.45% VAF, corresponding to 13 of the 14 EM patients, and none from the DF patients. The frequencies of the chromosomal rearrangement allele detected in these 29 samples ranged from 1.4 to 72.4%, (mean 19.3%), and the copy number concentrations ranged from 38 to 2,617 copies/mL of plasma (mean 552 copies/mL plasma). There was no significant difference between the 2p14 fragment concentrations of the EM and the DF samples, with an average of 1,908 copies/mL plasma measured (range 280-8,960 copies/mL plasma).

Thresholds for each ddPCR test was set at 50% between the normalized positive and negative droplet clusters. A receiver operating characteristic (ROC) curve analysis was done to evaluate the performance of the ddPCR analysis, and the area under curve (AUC) was observed at 0.98 ( $P = 0.001$ ) with a sensitivity of 93% and a specificity of 100%, indicating an excellent performance of this method. Of the 13 EM patients with positive ddPCR results, 12 had the occult disease detected prior to clinical relapse, with an average lead time window of 11 months (0-37 months) from the first positive blood test time point to clinical relapse. The two EM patients with no positive blood test results before clinical relapse had their last blood samples taken 4.5 and 12 months before the clinical metastasis, which might have been too big of an interval for monitoring ctDNA. We also found that ctDNA levels is a significant predictor of poor disease-free and overall survival.

This study proved the concept that serial monitoring of circulating cell-free DNA originating from tumor is a sensitive and specific tool with good feasibility to detect occult disease and predict outcome in breast cancer.

## Paper III

### Detection of circulating tumor cells and circulating tumor DNA before and after mammographic breast cancer compression in a cohort of breast cancer patients scheduled for neoadjuvant treatment

Thirty-one patients diagnosed with breast cancer from a large, population-based cohort of SCAN-B enrolled in this study. As a routine of the SCAN-B analysis, mRNA of the primary tumors was sequenced, and RNA-seq data was used for bioinformatic identification of tumor-specific single nucleotide variants (SNV, also known as point mutations) and small insertions and deletions (indels). Twenty-nine of the 31 tumors had at least one mutation identified for detection in the blood samples taken at mammography. One IBSAFE mutation detection assay was designed per tumor, including hotspot mutations such as *PIK3CA* p.H1047R, *PIK3CA* p.E545K, *PIK3CA* p.E542K, *TP53* p.R248W, *TP53* p.R110P, *TP53* p.V272E, *TP53* p.Y220C, *ARID1A* p.R1989\*, *PTEN* p.R130Q, etc. For samples where no known hotspot mutations were identified, an assay for another mutation merely used as a tumor-specific biomarker was designed. In total, 20 assays were designed for the 29 tumors, with *PIK3CA* p.H1047R being the most common mutation, found in 8/29 (27.6%) of the cases. The IBSAFE assays were tested in tumor tissue DNA as the positive control, and matched normal DNA extracted from the blood leukocytes as the negative control. Such validation experiments guaranteed that the selected mutations are somatic, and that the validated IBSAFE assays for the mutations have acceptable sensitivity and specificity given the settings of this study. IBSAFE assays for 20 tumor samples passed the validation. When applied in the pairs of central/peripheral blood samples taken before and after mammography, the validated IBSAFE assays for somatic mutations were able to detect ctDNA in 9/20 central and 12/20 peripheral blood sample pairs, with an increased level of ctDNA measured in post-mammographic samples in both central ( $P = 0.0756$ ) and peripheral ( $P = 0.0108$ ) cases. The average increases of VAF percentages were 0.77% in central and 0.35% in peripheral sample pairs (medians are 0.35% and 0.22%), respectively. These small VAF percentage increments indicated that the levels of ctDNA changes, insignificant in central and significant in peripheral blood, may have little biological impact. Notably, all 4 triple-negative breast cancer (TNBC) and 4 T4 staged cancers were ctDNA positive.

The CTC analysis, on the other hand, seemed to have better applicability than the ctDNA analysis, in that it does not require tumor-specific mutation assay design, but relies on CTC-specific antigen-based cell sorting. As a result, all patients with available pre- and post-mammographic central ( $N = 30$ ) and peripheral ( $N = 29$ ) blood samples were able to be analyzed for CTC levels. However, CTC analysis may have a poorer sensitivity than ctDNA analysis, given that CTCs were only detected in 8/30 and 2/29 central and peripheral blood sample pairs, respectively. An average increase of 3.2 cells ( $P = 0.188$ ) and decrease of 17 cells ( $P = 0.371$ )

were observed in the central and peripheral sample pairs, and no significant agreement was achieved between the ctDNA and the CTC analyses ( $\kappa = 0.02$ ,  $P = 0.92$ ). When the numbers of CTCs measured in the central and the peripheral samples were pairwise compared, 8/10 cases had more CTCs detected in central than in peripheral ( $P = 0.04$ ) samples, but such a trend was not observed in the ctDNA comparisons (8/20 had higher % VAF measured in central than in peripheral,  $P = 0.50$ ).

The results of this study showed that ctDNA, represented by tumor specific SNVs and indels, is a useful type of circulating biomarker for non-/minimally-invasive detection of tumor derived content in the body, and can be detected with the IBSAFE mutation detection assays with high sensitivity and specificity. CTCs are more abundant in central venous blood from superior vena cava, whereas ctDNA is at similar concentration levels in both central and peripheral blood. The study also showed, from the viewpoints of CTC and ctDNA, that mammography is a safe breast cancer diagnosis and screening tool that the risk of disseminating tumor content into the bloodstream to cause metastasis is low.

## Paper IV

### Subclonal patterns in follow-up of acute myeloid leukemia combining whole exome sequencing and ultrasensitive IBSAFE digital droplet analysis

Fourteen patients were included in this retrospective study aiming to prove the concept that IBSAFE is a useful tool to detect MRD in AML. The cohort consisted of 10 relapsing and 4 non-relapsing patients, of which the somatic mutational profiles were determined by WES done on tumor samples retrieved at diagnosis and clinical relapses and cultured matched normal skin fibroblast samples, yielding 10-30 (mean = 18) somatic mutations found at diagnosis per patient. Somatic SNVs and small indels, with those in recurrently mutated genes in AML (such as *NPM1*, *DNMT3A*, *RUNX1*, *FLT3*, *IDH1*, etc.) prioritized, were selected for IBSAFE assay design. A total of 86 assays, corresponding to 5-9 assays per patient, were validated in DNA extracted from the diagnostic samples as the positive controls, and at least 180 ng of normal human genomic DNA as the negative control. The validated IBSAFE assays were used to monitor molecular MRD in follow-up bone marrow aspirates with 120 ng of extracted DNA as input per sample. In parallel, qPCR analysis, if the patient was positive for one of the recurrent *NPM1* insertions at diagnosis, and MFC analysis were done in the follow-up samples.

For the 10 relapsing patients, one or several mutations, representing different tumor subclones, were tested positive for one or several time-points prior to the clinical relapses, indicating persisting or emerging MRD in the patients. Of the total 66 follow-up samples (relapsing samples included) tested in the relapsing patients, 35 samples from 9 patients (2-7 samples per patient, or 50-100% of all samples tested for the patient) had at least one mutation detected by IBSAFE at VAF < 0.1%, which could have been missed should another MRD detection method, for example MFC, was used instead of IBSAFE. In fact, of the 43 follow-up samples, not including the relapsing samples, in which both MFC and IBSAFE results were positive, MFC was only able to detect MRD in 7 (16.3%) samples from 5 (50%) patients, whereas IBSAFE reported positive results in 42 (97.7%) samples from all 10 (100%) patients, only missing one sample that was negative by both IBSAFE and MFC. Moreover, all 10 relapsing patients had at least one follow-up sample in which at least one mutation was tested positive by IBSAFE at a VAF > 0.1%.

Despite the small sample size, three patterns of tumor subclonal evolution seemed to manifest. In the first pattern featuring four patients, serial monitoring of the mutations revealed one or several subclones that responded differently during treatment, only to all come back as positive at relapse. Pattern two featuring another four patients, in comparison, had only some, but not all, of the subclones come back at relapse. The rest two patients showed a third pattern, in which one subclone represented by at least two mutations stayed at a high level throughout the treatment, even though they achieved clinical complete remission.



For the four non-relapsing patients, SCT was operated on two patients, and only 1 of the 10 monitored mutations was tested positive at 0.01% VAF by IBSAFE in the last follow-up samples. For the two other patients, both had one subclone, represented by 2 or more tracked mutations, persisting in all the follow-up samples, but the measured VAFs were below 0.08% in all follow-up samples.

As a potential new MRD detection method under assessment, IBSAFE was compared with the established MRD detection methods of 1) WES on VAFs of mutations followed by both IBSAFE and WES in the diagnostic and relapse samples, and 2) qPCR on VAFs of *NPM1* insertions in samples that both IBSAFE and qPCR were done. The results, illustrated in Figure 1 and Supplementary Figure 2 of Paper IV, show a balanced high-degree agreement between VAF measurements by IBSAFE and WES (N = 144), and by IBSAFE and qPCR (N = 34), across a wide range of VAFs, indicating the reliability of IBSAFE in detecting mutant molecules in follow-up bone marrow samples. Of note, for the comparison between IBSAFE and WES, 15 of the 144 pairs of mutation detections turned out negative by WES, of which IBSAFE reported positive results in 5. None of the mutation detections negative by IBSAFE was positive by WES, indicating a gained sensitivity of IBSAFE compared with WES. In contrast, for the IBSAFE vs qPCR comparison, 20 of the 34 pairs were negative by IBSAFE, of which 6 were positive by qPCR. None of the qPCR-negative sample was positive by IBSAFE. The explanation to this apparent poorer sensitivity of IBSAFE is that the total input DNA amount for IBSAFE was 120 ng compared to up to 500 ng for qPCR, resulting in the chance of mutant molecules sampled in the reaction simply higher in qPCR than in IBSAFE. This explanation is corroborated by the fact that the qPCR measured VAFs in 5 of the 6 samples were lower than the achievable limit of detection with the input amount of 120 ng of DNA in IBSAFE analyses.

The results of this study showed that IBSAFE has the potential to become a useful MRD detection tool in clinical practices of AML. IBSAFE can detect tumor specific mutations in follow-up bone marrow samples of AML, thus monitor the evolution of tumor subclones, measure the patients' response to treatments, and predict clinical relapses. Prospective clinical studies of larger scales in other types of follow-up samples and hematopoietic malignancies should be done to have a complete assessment of the clinical value of IBSAFE, and at the same time help us better understand the biology of the hematopoietic malignancies.

# Conclusions

In **Paper I**, we found that the frequent loss of PTEN protein in human breast cancer is not attributable to the overexpression of the E3 ubiquitin ligase NEDD4. In **Paper II**, we showed that serial monitoring, using ddPCR, of tumor specific chromosomal rearrangements identified with low coverage whole genome sequencing is a highly sensitive and specific approach to detect occult breast cancer disease prior to the onset of symptoms and clinical detection. Detected plasma ctDNA level also predicts poor relapse-free and overall survival. In **Paper III**, we confirmed the general safety of mammography that it does not appear to lead to additional dissemination of CTCs and ctDNA into the bloodstream. In **Paper IV**, we showed that acute myeloid leukemia specific mutations can be reliably detected with IBSAFE, and thus IBSAFE can be a clinically useful tool for MRD detection in AML.



# Future perspectives

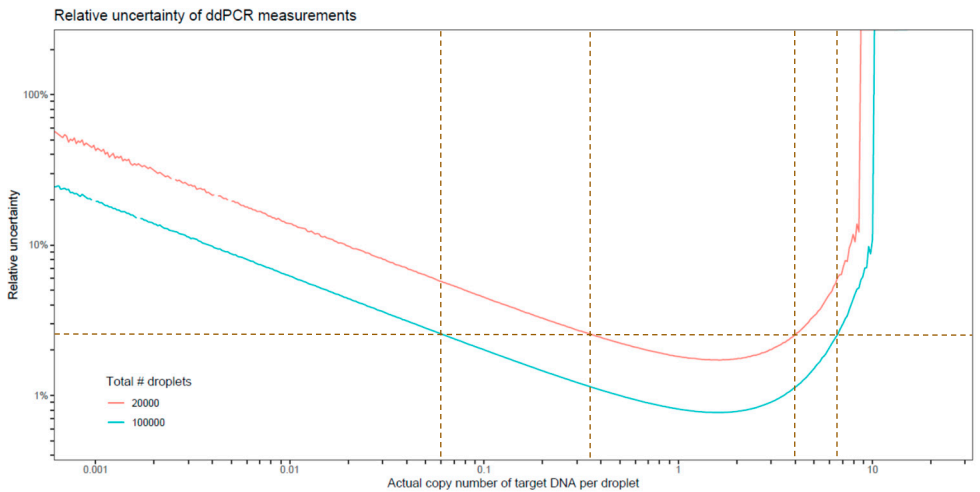
As cancer is a genetic disease, better understanding of its biology can lead to a better clinical management of the disease and eventually a better outcome for patients. Specifically, PTEN protein in breast cancer is apparently regulated with diverse mechanisms, and the interaction of these mechanisms may lead to different alterations of the PTEN protein levels. Functional studies on these regulation mechanisms and their interactions are worth studying in the future.

Knowledge of cancer biology can also help build more efficient pan-cancer or cancer type specific sequencing panels, so that the sequencing capacity can be allocated to relevant genomic regions, improving the sensitivity and feasibility of mutation detection by sequencing in the cancer genome. From both the laboratory and the bioinformatics points of view, the sequencing pipelines have room for improvement. For example, identification of chromosomal translocations from FFPE tumor tissue samples is challenging, in that the DNA is often degraded and cross linked. Direct sequencing of cfDNA requires a great depth, and thus is still expensive despite the cost of sequencing is constantly reducing. Although promising library preparing and target capture methods have been proposed, such as the ATOM-seq method using linear amplification to enrich the correctly copied DNA template [421], their performance and feasibility are yet to be assessed with real world samples. In addition, third generation sequencing methods, such as Oxford Nanopore, may change the landscape of genomic and transcriptomic research in the near future.

ddPCR also has the potential to serve a variety of research purposes, such as to monitor the methylation status of certain regions, assess microsatellite instability (MSI), and measure copy number variation of the genome. Recently, new ddPCR systems are constantly being marketed by different manufacturers, aiming to 1) increase the number of detectors for different fluorescence wavelengths, 2) increase the number of droplets or partitions to widen the dynamic range of absolute DNA copy number concentration measurement, and 3) decrease the reaction volume left out of analysis, known as the dead volume, to increase the sensitivity of rare event mutation detection.

Increased number of detection channels would facilitate the development of multiplexed mutation detection assays and allow for discriminatory detection of the component mutation which is otherwise hard to achieve if the number of channels

is fewer. Increased number of partitions gives more statistical power to the Poisson statistics underlying the idea of ddPCR, making the range of accurate copy number concentration detection wider, illustrated in Figure 21. Decreased dead volume makes it less likely that a true positive target DNA molecule fails to be sampled because of sampling error, which will render a false negative result no matter how sensitive the target DNA detection method is.



**Figure 21 Dynamic range of accurate measurement of target DNA concentration with 20,000 or 100,000 droplets.** The measurand in this plot is the average copy number of actual target DNA molecule per droplet. Relative uncertainty is defined as 95% confidence interval (CI) divided by the mean of the measurement. The dynamic range is defined as 2.5% relative uncertainty and is wider when 100,000 droplets are used for the analysis than when only 20,000 droplets are used.

Table 3 features the comparison of major multicolor digital PCR platforms already or soon to be released to the market.

Table 3. Major multicolor dPCR platforms			
Company	Bio-Rad	Stilla	Qiagen
System	QX600	Naica	QIAquity
Number of droplets/partitions	100,000	20,000 (Opal chips) 30,000 (Sapphire chips)	8,500 (96-well plates) 26,000 (24-well chips)
Number of colors	6	6	5
Droplet volume	Unrevealed	0.59 nL	0.34 nL (96-well plates) 0.91 nL (24-well chips)
Dead volume	Unrevealed (35-50% for QX200)	30-40%	76% (96-well plates) 42% (24-well chips)

# Acknowledgements

In China where I am from, a question is frequently given to young children as what they want to do when they grow up, and one of the most common answers is that they want to become a scientist. This was my childhood answer as well. Along the way, this answer has not changed much, for the life of a scientist to me is about perpetually exploring the vast unknown, which addresses the curiosity that lies deeply in our nature, and about improving the well-being of people, which is an obligation of all humans and especially for those that are capable. This lifestyle is desirable and moral and happens to be the lifestyle I dreamed to have.

The completion of my PhD thesis work may be a mark of this dream coming true, and for this I am wholeheartedly grateful. The work could not have been done without the help and support from a tremendous number of people, among whom I would especially thank:

*Lao Saal*, my main supervisor, for all the education, training, mentorship, supervision, discussion, inspiration, and everything, from when I took a course in which he gave a lecture, to my master's thesis work, lab internship, and doctorate program.

*Åke Borg*, *Göran B Jönsson*, and *Johan Staaf* for being my co-supervisors.

Current and former members of the Translational Oncogenomics Unit research group and our closest collaborators: *Anthony George*, *Malin Dahlgren*, *Robert Rigo*, *Sergii Gladchuk*, *Heena Saini*, *Xu Chi*, *Yoshiko Nanki*, *Christof Winter*, *Man-Hung Eric Tang*, *Eleonor Olsson*, *Tamara Tjitrowirjo*, *Barbara Lettiero*, *Stefano Misino*, *Willow Hight-Warburton*, *Michelle Lee*, *Madeline Dixon*, *Marina Villamor*, *Sofia Gruvberger-Saal*, and *Jill Howlin*, for their help for making me a better researcher.

*Louise Pettersson* and *Mats Ehinger* for our very productive collaboration and for teaching me all about leukemia from A to Z.

*Miguel Alcaide* and *Sofia Birkeälv* for valuable discussions.

To the *SCAN-B infrastructure* and all personnel of the *SCAN-B central laboratory* for continuously processing the samples and data at high standard.

*Susanne André* and *Björn Frostner* for administrative support.

*Co-authors* and *contributors* of all the research projects for their expertise, input, and feedback.

And all the *clinicians, nurses, and patients* for participating in the project, donating their time, effort, and tissues, so that current and future patients may benefit.

I would also like to thank my mother, *Yanhua Shen* 沈燕华, father *Jianzhong Jamason Chen* 陈建中, aunt *Chenxi Chen* 陈晨曦, and grandma *Minlan Niu* 钮敏兰 for being role models for me, and their love and care.

Yilun Chen

October 2021

# References

1. Merriam-Webster Inc., The Merriam-Webster dictionary. 2019, Springfield, Massachusetts: Merriam-Webster, Incorporated. (18, 701) pages.
2. Zink, A., Rohrbach, H., Szeimies, U., Hagedorn, H.G., Haas, C.J., Weyss, C., et al., Malignant tumors in an ancient Egyptian population. *Anticancer Res*, 1999. **19**(5B): p. 4273-7.
3. Nerlich, A.G., Rohrbach, H., Bachmeier, B., and Zink, A., Malignant tumors in two ancient populations: An approach to historical tumor epidemiology. *Oncol Rep*, 2006. **16**(1): p. 197-202.
4. Zweifel, L., Buni, T., and Ruhli, F.J., Evidence-based palaeopathology: meta-analysis of PubMed-listed scientific studies on ancient Egyptian mummies. *Homo*, 2009. **60**(5): p. 405-27.
5. David, A.R. and Zimmerman, M.R., Cancer: an old disease, a new disease or something in between? *Nat Rev Cancer*, 2010. **10**(10): p. 728-33.
6. Dageforde, K.L., Vennemann, M., and Ruhli, F.J., Evidence based palaeopathology: meta-analysis of Pubmed-listed scientific studies on pre-Columbian, South American mummies. *Homo*, 2014. **65**(3): p. 214-31.
7. Capasso, L.L., Antiquity of cancer. *Int J Cancer*, 2005. **113**(1): p. 2-13.
8. Ford, D., Easton, D.F., Bishop, D.T., Narod, S.A., and Goldgar, D.E., Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Lancet*, 1994. **343**(8899): p. 692-5.
9. Breast Cancer Linkage, C., Cancer risks in BRCA2 mutation carriers. *J Natl Cancer Inst*, 1999. **91**(15): p. 1310-6.
10. Caron de Fromentel, C. and Soussi, T., TP53 tumor suppressor gene: a model for investigating human mutagenesis. *Genes Chromosomes Cancer*, 1992. **4**(1): p. 1-15.
11. Druker, B.J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G.M., Fanning, S., et al., Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med*, 1996. **2**(5): p. 561-6.
12. Stratton, M.R., Campbell, P.J., and Futreal, P.A., The cancer genome. *Nature*, 2009. **458**(7239): p. 719-24.
13. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. **409**(6822): p. 860-921.
14. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., et al., The sequence of the human genome. *Science*, 2001. **291**(5507): p. 1304-51.
15. Hanahan, D. and Weinberg, R.A., The hallmarks of cancer. *Cell*, 2000. **100**(1): p. 57-70.



16. Hanahan, D. and Weinberg, R.A., Hallmarks of cancer: the next generation. *Cell*, 2011. **144**(5): p. 646-74.
17. Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., et al., A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 1994. **266**(5182): p. 66-71.
18. Easton, D.F., Ford, D., and Bishop, D.T., Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am J Hum Genet*, 1995. **56**(1): p. 265-71.
19. Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., et al., Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 1995. **378**(6559): p. 789-92.
20. Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., et al., Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet*, 1998. **62**(3): p. 676-89.
21. Vasen, H.F., Wijnen, J.T., Menko, F.H., Kleibeuker, J.H., Taal, B.G., Griffioen, G., et al., Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis. *Gastroenterology*, 1996. **110**(4): p. 1020-7.
22. Plazzer, J.P., Sijmons, R.H., Woods, M.O., Peltomaki, P., Thompson, B., Den Dunnen, J.T., et al., The InSiGHT database: utilizing 100 years of insights into Lynch syndrome. *Fam Cancer*, 2013. **12**(2): p. 175-80.
23. Hisada, M., Garber, J.E., Fung, C.Y., Fraumeni, J.F., Jr., and Li, F.P., Multiple primary cancers in families with Li-Fraumeni syndrome. *J Natl Cancer Inst*, 1998. **90**(8): p. 606-11.
24. Varley, J.M., Germline TP53 mutations and Li-Fraumeni syndrome. *Hum Mutat*, 2003. **21**(3): p. 313-20.
25. Moolgavkar, S.H. and Knudson, A.G., Jr., Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst*, 1981. **66**(6): p. 1037-52.
26. Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., et al., Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A*, 2010. **107**(43): p. 18545-50.
27. McFarland, C.D., Korolev, K.S., Kryukov, G.V., Sunyaev, S.R., and Mirny, L.A., Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci U S A*, 2013. **110**(8): p. 2910-5.
28. McFarland, C.D., Yaglom, J.A., Wojtkowiak, J.W., Scott, J.G., Morse, D.L., Sherman, M.Y., et al., The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Res*, 2017. **77**(18): p. 4763-4772.
29. Cantley, L.C., Auger, K.R., Carpenter, C., Duckworth, B., Graziani, A., Kapeller, R., et al., Oncogenes and signal transduction. *Cell*, 1991. **64**(2): p. 281-302.
30. Marshall, C.J., Tumor suppressor genes. *Cell*, 1991. **64**(2): p. 313-26.
31. Carter, S.L., Negrini, M., Baffa, R., Gillum, D.R., Rosenberg, A.L., Schwartz, G.F., et al., Loss of heterozygosity at 11q22-q23 in breast cancer. *Cancer Res*, 1994. **54**(23): p. 6270-4.

32. Yee, C.J., Roodi, N., Verrier, C.S., and Parl, F.F., Microsatellite instability and loss of heterozygosity in breast cancer. *Cancer Res*, 1994. **54**(7): p. 1641-4.
33. Naylor, S.L., Johnson, B.E., Minna, J.D., and Sakaguchi, A.Y., Loss of heterozygosity of chromosome 3p markers in small-cell lung cancer. *Nature*, 1987. **329**(6138): p. 451-4.
34. Tseng, R.C., Chang, J.W., Hsien, F.J., Chang, Y.H., Hsiao, C.F., Chen, J.T., et al., Genomewide loss of heterozygosity and its clinical associations in non small cell lung cancer. *Int J Cancer*, 2005. **117**(2): p. 241-7.
35. Christiansen, D.H., Andersen, M.K., and Pedersen-Bjergaard, J., Mutations with loss of heterozygosity of p53 are common in therapy-related myelodysplasia and acute myeloid leukemia after exposure to alkylating agents and significantly associated with deletion or loss of 5q, a complex karyotype, and a poor prognosis. *J Clin Oncol*, 2001. **19**(5): p. 1405-13.
36. Abkevich, V., Timms, K.M., Hennessy, B.T., Potter, J., Carey, M.S., Meyer, L.A., et al., Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br J Cancer*, 2012. **107**(10): p. 1776-82.
37. Dong, J.T., Boyd, J.C., and Frierson, H.F., Jr., Loss of heterozygosity at 13q14 and 13q21 in high grade, high stage prostate cancer. *Prostate*, 2001. **49**(3): p. 166-71.
38. Gao, X., Zacharek, A., Salkowski, A., Grignon, D.J., Sakr, W., Porter, A.T., et al., Loss of heterozygosity of the BRCA1 and other loci on chromosome 17q in human prostate cancer. *Cancer Res*, 1995. **55**(5): p. 1002-5.
39. Uchino, S., Tsuda, H., Noguchi, M., Yokota, J., Terada, M., Saito, T., et al., Frequent loss of heterozygosity at the DCC locus in gastric cancer. *Cancer Res*, 1992. **52**(11): p. 3099-102.
40. Koufos, A., Hansen, M.F., Copeland, N.G., Jenkins, N.A., Lampkin, B.C., and Cavenee, W.K., Loss of heterozygosity in three embryonal tumours suggests a common pathogenetic mechanism. *Nature*, 1985. **316**(6026): p. 330-4.
41. Deng, G., Lu, Y., Zlotnikov, G., Thor, A.D., and Smith, H.S., Loss of heterozygosity in normal tissue adjacent to breast carcinomas. *Science*, 1996. **274**(5295): p. 2057-9.
42. Thiagalingam, S., Foy, R.L., Cheng, K.H., Lee, H.J., Thiagalingam, A., and Ponte, J.F., Loss of heterozygosity as a predictor to map tumor suppressor genes in cancer: molecular basis of its occurrence. *Curr Opin Oncol*, 2002. **14**(1): p. 65-72.
43. Wellcome Sanger Institute. Catalogue Of Somatic Mutations In Cancer (COSMIC). Available from: <https://cancer.sanger.ac.uk/cosmic> [accessed 2021-09-08].
44. Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M., Role of non-coding sequence variants in cancer. *Nat Rev Genet*, 2016. **17**(2): p. 93-108.
45. Rheinbay, E., Nielsen, M.M., Abascal, F., Wala, J.A., Shapira, O., Tiao, G., et al., Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 2020. **578**(7793): p. 102-111.
46. Elliott, K. and Larsson, E., Non-coding driver mutations in human cancer. *Nat Rev Cancer*, 2021. **21**(8): p. 500-509.
47. Taylor, B.S., Barretina, J., Socci, N.D., Decarolis, P., Ladanyi, M., Meyerson, M., et al., Functional copy-number alterations in cancer. *PLoS One*, 2008. **3**(9): p. e3179.

48. Shlien, A. and Malkin, D., Copy number variations and cancer. *Genome Med*, 2009. **1**(6): p. 62.
49. Kuiper, R.P., Ligtenberg, M.J., Hoogerbrugge, N., and Geurts van Kessel, A., Germline copy number variation and cancer risk. *Curr Opin Genet Dev*, 2010. **20**(3): p. 282-9.
50. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., et al., Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, 2013. **45**(10): p. 1134-40.
51. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., et al., Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 2005. **310**(5748): p. 644-8.
52. Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., et al., Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, 2007. **448**(7153): p. 561-6.
53. Feuk, L., Carson, A.R., and Scherer, S.W., Structural variation in the human genome. *Nat Rev Genet*, 2006. **7**(2): p. 85-97.
54. Stankiewicz, P. and Lupski, J.R., Structural variation in the human genome and its role in disease. *Annu Rev Med*, 2010. **61**: p. 437-55.
55. Alkan, C., Coe, B.P., and Eichler, E.E., Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011. **12**(5): p. 363-76.
56. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., et al., An integrated map of structural variation in 2,504 human genomes. *Nature*, 2015. **526**(7571): p. 75-81.
57. Jones, P.A. and Baylin, S.B., The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 2002. **3**(6): p. 415-28.
58. Esteller, M., Epigenetics in cancer. *N Engl J Med*, 2008. **358**(11): p. 1148-59.
59. Sharma, S., Kelly, T.K., and Jones, P.A., Epigenetics in cancer. *Carcinogenesis*, 2010. **31**(1): p. 27-36.
60. Weiss, J.R., Moysich, K.B., and Swede, H., Epidemiology of male breast cancer. *Cancer Epidemiol Biomarkers Prev*, 2005. **14**(1): p. 20-6.
61. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., et al., Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, 2021. **71**(3): p. 209-249.
62. Socialstyrelsen. Statistics on Cancer Incidence 2019. Available from: <https://www.socialstyrelsen.se/en/statistics-and-data/statistics/> [accessed 2021-09-10].
63. Siegel, R.L., Miller, K.D., and Jemal, A., Cancer statistics, 2019. *CA Cancer J Clin*, 2019. **69**(1): p. 7-34.
64. Collaborative Group on Hormonal Factors in Breast, C., Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet*, 2001. **358**(9291): p. 1389-99.

65. Balmana, J., Diez, O., Rubio, I.T., Cardoso, F., and Group, E.G.W., BRCA in breast cancer: ESMO Clinical Practice Guidelines. *Ann Oncol*, 2011. **22 Suppl 6**: p. vi31-4.
66. Anderson, K.N., Schwab, R.B., and Martinez, M.E., Reproductive risk factors and breast cancer subtypes: a review of the literature. *Breast Cancer Res Treat*, 2014. **144**(1): p. 1-10.
67. Momenimovahed, Z. and Salehiniya, H., Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer (Dove Med Press)*, 2019. **11**: p. 151-164.
68. 1177 Vårdguiden. Mammogram – breast screening. Available from: <https://www.1177.se/en/other-languages/other-languages/undersokningarprover---andra-sprak/mammografi---andra-sprak/> [accessed 2021-09-10].
69. Ernster, V.L., Ballard-Barbash, R., Barlow, W.E., Zheng, Y., Weaver, D.L., Cutter, G., et al., Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst*, 2002. **94**(20): p. 1546-54.
70. Gotzsche, P.C. and Jorgensen, K.J., Screening for breast cancer with mammography. *Cochrane Database Syst Rev*, 2013(6): p. CD001877.
71. Marmot, M.G., Altman, D.G., Cameron, D.A., Dewar, J.A., Thompson, S.G., and Wilcox, M., The benefits and harms of breast cancer screening: an independent review. *Br J Cancer*, 2013. **108**(11): p. 2205-40.
72. Kumar, V., Abbas, A.K., and Aster, J.C., Robbins and Cotran pathologic basis of disease. Ninth edition. ed. 2015, Philadelphia, PA: Elsevier/Saunders. xvi, 1391 pages.
73. Li, C.I., Malone, K.E., Saltzman, B.S., and Daling, J.R., Risk of invasive breast carcinoma among women diagnosed with ductal carcinoma in situ and lobular carcinoma in situ, 1988-2001. *Cancer*, 2006. **106**(10): p. 2104-12.
74. Kerlikowske, K., Epidemiology of ductal carcinoma in situ. *J Natl Cancer Inst Monogr*, 2010. **2010**(41): p. 139-41.
75. Xie, Z.M., Sun, J., Hu, Z.Y., Wu, Y.P., Liu, P., Tang, J., et al., Survival outcomes of patients with lobular carcinoma in situ who underwent bilateral mastectomy or partial mastectomy. *Eur J Cancer*, 2017. **82**: p. 6-15.
76. Tavassoli, F.A. and Devilee, P., Pathology and genetics of tumours of the breast and female genital organs. 2003, Lyon: International Agency for Research on Cancer ; Oxford : Oxford University Press [distributor].
77. Urruticoechea, A., Smith, I.E., and Dowsett, M., Proliferation marker Ki-67 in early breast cancer. *J Clin Oncol*, 2005. **23**(28): p. 7212-20.
78. Li, L.T., Jiang, G., Chen, Q., and Zheng, J.N., Ki67 is a promising molecular target in the diagnosis of cancer (review). *Mol Med Rep*, 2015. **11**(3): p. 1566-72.
79. Goldhirsch, A., Wood, W.C., Coates, A.S., Gelber, R.D., Thurlimann, B., Senn, H.J., et al., Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol*, 2011. **22**(8): p. 1736-47.

80. Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., et al., Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res*, 2010. **12**(4): p. 207.
81. Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., et al., Molecular portraits of human breast tumours. *Nature*, 2000. **406**(6797): p. 747-52.
82. Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., et al., Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 2003. **100**(14): p. 8418-23.
83. Brueffer, C., Vallon-Christersson, J., Grabau, D., Ehinger, A., Hakkinen, J., Hegardt, C., et al., Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO Precis Oncol*, 2018. **2**.
84. Reddy, H.K., Mettus, R.V., Rane, S.G., Grana, X., Litvin, J., and Reddy, E.P., Cyclin-dependent kinase 4 expression is essential for neu-induced breast tumorigenesis. *Cancer Res*, 2005. **65**(22): p. 10174-8.
85. Dean, J.L., Thangavel, C., McClendon, A.K., Reed, C.A., and Knudsen, E.S., Therapeutic CDK4/6 inhibition in breast cancer: key mechanisms of response and failure. *Oncogene*, 2010. **29**(28): p. 4018-32.
86. O'Leary, B., Finn, R.S., and Turner, N.C., Treating cancer with selective CDK4/6 inhibitors. *Nat Rev Clin Oncol*, 2016. **13**(7): p. 417-30.
87. Kwapisz, D., Cyclin-dependent kinase 4/6 inhibitors in breast cancer: palbociclib, ribociclib, and abemaciclib. *Breast Cancer Res Treat*, 2017. **166**(1): p. 41-54.
88. Okazaki, T. and Honjo, T., The PD-1-PD-L pathway in immunological tolerance. *Trends Immunol*, 2006. **27**(4): p. 195-201.
89. Okazaki, T. and Honjo, T., PD-1 and PD-1 ligands: from discovery to clinical application. *Int Immunol*, 2007. **19**(7): p. 813-24.
90. Robert, C., Schachter, J., Long, G.V., Arance, A., Grob, J.J., Mortier, L., et al., Pembrolizumab versus Ipilimumab in Advanced Melanoma. *N Engl J Med*, 2015. **372**(26): p. 2521-32.
91. Eggermont, A.M.M., Blank, C.U., Mandala, M., Long, G.V., Atkinson, V., Dalle, S., et al., Adjuvant Pembrolizumab versus Placebo in Resected Stage III Melanoma. *N Engl J Med*, 2018. **378**(19): p. 1789-1801.
92. Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., Eder, J.P., et al., Pembrolizumab for the treatment of non-small-cell lung cancer. *N Engl J Med*, 2015. **372**(21): p. 2018-28.
93. Reck, M., Rodriguez-Abreu, D., Robinson, A.G., Hui, R., Csoszi, T., Fulop, A., et al., Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med*, 2016. **375**(19): p. 1823-1833.
94. Gandhi, L., Rodriguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., et al., Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *N Engl J Med*, 2018. **378**(22): p. 2078-2092.

95. Bellmunt, J. and Bajorin, D.F., Pembrolizumab for Advanced Urothelial Carcinoma. *N Engl J Med*, 2017. **376**(23): p. 2304.
96. Nanda, R., Chow, L.Q., Dees, E.C., Berger, R., Gupta, S., Geva, R., et al., Pembrolizumab in Patients With Advanced Triple-Negative Breast Cancer: Phase Ib KEYNOTE-012 Study. *J Clin Oncol*, 2016. **34**(21): p. 2460-7.
97. Schmid, P., Cortes, J., Puzstai, L., McArthur, H., Kummel, S., Bergh, J., et al., Pembrolizumab for Early Triple-Negative Breast Cancer. *N Engl J Med*, 2020. **382**(9): p. 810-821.
98. Topalian, S.L., Hodi, F.S., Brahmer, J.R., Gettinger, S.N., Smith, D.C., McDermott, D.F., et al., Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med*, 2012. **366**(26): p. 2443-54.
99. Hamid, O., Robert, C., Daud, A., Hodi, F.S., Hwu, W.J., Kefford, R., et al., Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. *N Engl J Med*, 2013. **369**(2): p. 134-44.
100. Chen, L. and Han, X., Anti-PD-1/PD-L1 therapy of human cancer: past, present, and future. *J Clin Invest*, 2015. **125**(9): p. 3384-91.
101. Tabar, L., Yen, M.F., Vitak, B., Chen, H.H., Smith, R.A., and Duffy, S.W., Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. *Lancet*, 2003. **361**(9367): p. 1405-10.
102. Early Breast Cancer Trialists' Collaborative, G., Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*, 2005. **365**(9472): p. 1687-717.
103. Early Breast Cancer Trialists' Collaborative, G., Darby, S., McGale, P., Correa, C., Taylor, C., Arriagada, R., et al., Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. *Lancet*, 2011. **378**(9804): p. 1707-16.
104. Pan, H., Gray, R., Braybrooke, J., Davies, C., Taylor, C., McGale, P., et al., 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *N Engl J Med*, 2017. **377**(19): p. 1836-1846.
105. Alvarado, M., Ozanne, E., and Esserman, L., Overdiagnosis and overtreatment of breast cancer. *Am Soc Clin Oncol Educ Book*, 2012: p. e40-5.
106. Esserman, L.J., Thompson, I.M., Jr., and Reid, B., Overdiagnosis and overtreatment in cancer: an opportunity for improvement. *JAMA*, 2013. **310**(8): p. 797-8.
107. Canney, P.A., Moore, M., Wilkinson, P.M., and James, R.D., Ovarian cancer antigen CA125: a prospective clinical assessment of its role as a tumour marker. *Br J Cancer*, 1984. **50**(6): p. 765-9.
108. Wang, W., Xu, X., Tian, B., Wang, Y., Du, L., Sun, T., et al., The diagnostic value of serum tumor markers CEA, CA19-9, CA125, CA15-3, and TPS in metastatic breast cancer. *Clin Chim Acta*, 2017. **470**: p. 51-55.
109. Geraghty, J.G., Coveney, E.C., Sherry, F., O'Higgins, N.J., and Duffy, M.J., CA 15-3 in patients with locoregional and metastatic breast carcinoma. *Cancer*, 1992. **70**(12): p. 2831-4.

110. Park, B.W., Oh, J.W., Kim, J.H., Park, S.H., Kim, K.S., Kim, J.H., et al., Preoperative CA 15-3 and CEA serum levels as predictor for breast cancer outcomes. *Ann Oncol*, 2008. **19**(4): p. 675-81.
111. Lee, J.S., Park, S., Park, J.M., Cho, J.H., Kim, S.I., and Park, B.W., Elevated levels of serum tumor markers CA 15-3 and CEA are prognostic factors for diagnosis of metastatic breast cancers. *Breast Cancer Res Treat*, 2013. **141**(3): p. 477-84.
112. Miralles, C., Orea, M., Espana, P., Provencio, M., Sanchez, A., Cantos, B., et al., Cancer antigen 125 associated with multiple benign and malignant pathologies. *Ann Surg Oncol*, 2003. **10**(2): p. 150-4.
113. Chaffer, C.L. and Weinberg, R.A., A perspective on cancer cell metastasis. *Science*, 2011. **331**(6024): p. 1559-64.
114. Plaks, V., Koopman, C.D., and Werb, Z., Cancer. Circulating tumor cells. *Science*, 2013. **341**(6151): p. 1186-8.
115. Bidard, F.C., Mathiot, C., Delaloge, S., Brain, E., Giachetti, S., de Cremoux, P., et al., Single circulating tumor cell detection and overall survival in nonmetastatic breast cancer. *Ann Oncol*, 2010. **21**(4): p. 729-733.
116. Lucci, A., Hall, C.S., Lodhi, A.K., Bhattacharyya, A., Anderson, A.E., Xiao, L., et al., Circulating tumour cells in non-metastatic breast cancer: a prospective study. *Lancet Oncol*, 2012. **13**(7): p. 688-95.
117. Rack, B., Schindlbeck, C., Juckstock, J., Andergassen, U., Hepp, P., Zwingers, T., et al., Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *J Natl Cancer Inst*, 2014. **106**(5).
118. Nole, F., Munzone, E., Zorzino, L., Minchella, I., Salvatici, M., Botteri, E., et al., Variation of circulating tumor cell levels during treatment of metastatic breast cancer: prognostic and therapeutic implications. *Ann Oncol*, 2008. **19**(5): p. 891-7.
119. Aceto, N., Bardia, A., Miyamoto, D.T., Donaldson, M.C., Wittner, B.S., Spencer, J.A., et al., Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell*, 2014. **158**(5): p. 1110-1122.
120. Riethdorf, S., Fritsche, H., Muller, V., Rau, T., Schindlbeck, C., Rack, B., et al., Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. *Clin Cancer Res*, 2007. **13**(3): p. 920-8.
121. Kowalik, A., Kowalewska, M., and Gozdz, S., Current approaches for avoiding the limitations of circulating tumor cells detection methods-implications for diagnosis and treatment of patients with solid tumors. *Transl Res*, 2017. **185**: p. 58-84 e15.
122. Peeters, D.J., Van den Eynden, G.G., van Dam, P.J., Prove, A., Benoy, I.H., van Dam, P.A., et al., Circulating tumour cells in the central and the peripheral venous compartment in patients with metastatic breast cancer. *Br J Cancer*, 2011. **104**(9): p. 1472-7.
123. Koch, M., Kienle, P., Hinz, U., Antolovic, D., Schmidt, J., Herfarth, C., et al., Detection of hematogenous tumor cell dissemination predicts tumor relapse in patients undergoing surgical resection of colorectal liver metastases. *Ann Surg*, 2005. **241**(2): p. 199-205.

124. Sandri, M.T., Zorzino, L., Cassatella, M.C., Bassi, F., Luini, A., Casadio, C., et al., Changes in circulating tumor cell detection in patients with localized breast cancer before and after surgery. *Ann Surg Oncol*, 2010. **17**(6): p. 1539-45.
125. Papavasiliou, P., Fisher, T., Kuhn, J., Nemunaitis, J., and Lamont, J., Circulating tumor cells in patients undergoing surgery for hepatic metastases from colorectal cancer. *Proc (Bayl Univ Med Cent)*, 2010. **23**(1): p. 11-4.
126. Hashimoto, M., Tanaka, F., Yoneda, K., Takuwa, T., Matsumoto, S., Okumura, Y., et al., Significant increase in circulating tumour cells in pulmonary venous blood during surgical manipulation in patients with primary lung cancer. *Interact Cardiovasc Thorac Surg*, 2014. **18**(6): p. 775-83.
127. van Dalum, G., van der Stam, G.J., Tibbe, A.G., Franken, B., Mastboom, W.J., Vermes, I., et al., Circulating tumor cells before and during follow-up after breast cancer surgery. *Int J Oncol*, 2015. **46**(1): p. 407-13.
128. Khatcheressian, J.L., Hurley, P., Bantug, E., Esserman, L.J., Grunfeld, E., Halberg, F., et al., Breast cancer follow-up and management after primary treatment: American Society of Clinical Oncology clinical practice guideline update. *J Clin Oncol*, 2013. **31**(7): p. 961-5.
129. Theriault, R.L., Carlson, R.W., Allred, C., Anderson, B.O., Burstein, H.J., Edge, S.B., et al., Breast cancer, version 3.2013: featured updates to the NCCN guidelines. *J Natl Compr Canc Netw*, 2013. **11**(7): p. 753-60; quiz 761.
130. Stroun, M., Lyautey, J., Lederrey, C., Olson-Sand, A., and Anker, P., About the possible origin and mechanism of circulating DNA apoptosis and active DNA release. *Clin Chim Acta*, 2001. **313**(1-2): p. 139-42.
131. Jung, K., Fleischhacker, M., and Rabien, A., Cell-free DNA in the blood as a solid tumor biomarker--a critical appraisal of the literature. *Clin Chim Acta*, 2010. **411**(21-22): p. 1611-24.
132. Underhill, H.R., Kitzman, J.O., Hellwig, S., Welker, N.C., Daza, R., Baker, D.N., et al., Fragment Length of Circulating Tumor DNA. *PLoS Genet*, 2016. **12**(7): p. e1006162.
133. Stroun, M., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., and Beljanski, M., Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology*, 1989. **46**(5): p. 318-22.
134. Dawson, S.J., Tsui, D.W., Murtaza, M., Biggs, H., Rueda, O.M., Chin, S.F., et al., Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med*, 2013. **368**(13): p. 1199-209.
135. Newman, A.M., Bratman, S.V., To, J., Wynne, J.F., Eclov, N.C., Modlin, L.A., et al., An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med*, 2014. **20**(5): p. 548-54.
136. Murtaza, M., Dawson, S.J., Tsui, D.W., Gale, D., Forshew, T., Piskorz, A.M., et al., Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature*, 2013. **497**(7447): p. 108-12.
137. Diaz, L.A., Jr., Williams, R.T., Wu, J., Kinde, I., Hecht, J.R., Berlin, J., et al., The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 2012. **486**(7404): p. 537-40.



138. McBride, D.J., Orpana, A.K., Sotiriou, C., Joensuu, H., Stephens, P.J., Mudie, L.J., et al., Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer*, 2010. **49**(11): p. 1062-9.
139. Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S., Li, M., et al., Circulating mutant DNA to assess tumor dynamics. *Nat Med*, 2008. **14**(9): p. 985-90.
140. Engelman, J.A., Luo, J., and Cantley, L.C., The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat Rev Genet*, 2006. **7**(8): p. 606-19.
141. Chalhoub, N. and Baker, S.J., PTEN and the PI3-kinase pathway in cancer. *Annu Rev Pathol*, 2009. **4**: p. 127-50.
142. Catalogue Of Somatic Mutations In Cancer (COSMIC). PIK3CA mutations. Available from: <https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=PIK3CA> [accessed 2021-09-14].
143. Catalogue Of Somatic Mutations In Cancer (COSMIC). PTEN mutations. Available from: <https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=PTEN> [accessed 2021-09-14].
144. Saal, L.H., Holm, K., Maurer, M., Memeo, L., Su, T., Wang, X., et al., PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Res*, 2005. **65**(7): p. 2554-9.
145. Saal, L.H., Johansson, P., Holm, K., Gruvberger-Saal, S.K., She, Q.B., Maurer, M., et al., Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc Natl Acad Sci U S A*, 2007. **104**(18): p. 7564-9.
146. Majewski, I.J., Nuciforo, P., Mittempergher, L., Bosma, A.J., Eidtmann, H., Holmes, E., et al., PIK3CA mutations are associated with decreased benefit to neoadjuvant human epidermal growth factor receptor 2-targeted therapies in breast cancer. *J Clin Oncol*, 2015. **33**(12): p. 1334-9.
147. Andre, F., Ciruelos, E., Rubovszky, G., Campone, M., Loibl, S., Rugo, H.S., et al., Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer. *N Engl J Med*, 2019. **380**(20): p. 1929-1940.
148. Andre, F., Ciruelos, E.M., Juric, D., Loibl, S., Campone, M., Mayer, I.A., et al., Alpelisib plus fulvestrant for PIK3CA-mutated, hormone receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: final overall survival results from SOLAR-1. *Ann Oncol*, 2021. **32**(2): p. 208-217.
149. Nagata, Y., Lan, K.H., Zhou, X., Tan, M., Esteva, F.J., Sahin, A.A., et al., PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer Cell*, 2004. **6**(2): p. 117-27.
150. Saal, L.H., Gruvberger-Saal, S.K., Persson, C., Lovgren, K., Jumppanen, M., Staaf, J., et al., Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair. *Nat Genet*, 2008. **40**(1): p. 102-7.
151. Wang, X., Trotman, L.C., Koppie, T., Alimonti, A., Chen, Z., Gao, Z., et al., NEDD4-1 is a proto-oncogenic ubiquitin ligase for PTEN. *Cell*, 2007. **128**(1): p. 129-39.

152. Drinjakovic, J., Jung, H., Campbell, D.S., Strohlic, L., Dwivedy, A., and Holt, C.E., E3 ligase Nedd4 promotes axon branching by downregulating PTEN. *Neuron*, 2010. **65**(3): p. 341-57.
153. Goh, C.P., Low, L.H., Putz, U., Gunnersen, J., Hammond, V., Howitt, J., et al., Ndfip1 expression in developing neurons indicates a role for protein ubiquitination by Nedd4 E3 ligases during cortical development. *Neurosci Lett*, 2013. **555**: p. 225-30.
154. Guo, H., Qiao, G., Ying, H., Li, Z., Zhao, Y., Liang, Y., et al., E3 ubiquitin ligase Cbl-b regulates Pten via Nedd4 in T cells independently of its ubiquitin ligase activity. *Cell Rep*, 2012. **1**(5): p. 472-82.
155. Chung, S., Nakashima, M., Zembutsu, H., and Nakamura, Y., Possible involvement of NEDD4 in keloid formation; its critical role in fibroblast proliferation and collagen production. *Proc Jpn Acad Ser B Phys Biol Sci*, 2011. **87**(8): p. 563-73.
156. Shi, Y., Wang, J., Chandarlapaty, S., Cross, J., Thompson, C., Rosen, N., et al., PTEN is a protein tyrosine phosphatase for IRS1. *Nat Struct Mol Biol*, 2014. **21**(6): p. 522-7.
157. Amodio, N., Scrima, M., Palaia, L., Salman, A.N., Quintiero, A., Franco, R., et al., Oncogenic role of the E3 ubiquitin ligase NEDD4-1, a PTEN negative regulator, in non-small-cell lung carcinomas. *Am J Pathol*, 2010. **177**(5): p. 2622-34.
158. Hong, S.W., Moon, J.H., Kim, J.S., Shin, J.S., Jung, K.A., Lee, W.K., et al., p34 is a novel regulator of the oncogenic behavior of NEDD4-1 and PTEN. *Cell Death Differ*, 2014. **21**(1): p. 146-60.
159. Trotman, L.C., Wang, X., Alimonti, A., Chen, Z., Teruya-Feldstein, J., Yang, H., et al., Ubiquitination regulates PTEN nuclear import and tumor suppression. *Cell*, 2007. **128**(1): p. 141-56.
160. Fouladkou, F., Landry, T., Kawabe, H., Neeb, A., Lu, C., Brose, N., et al., The ubiquitin ligase Nedd4-1 is dispensable for the regulation of PTEN stability and localization. *Proc Natl Acad Sci U S A*, 2008. **105**(25): p. 8585-90.
161. Maddika, S., Kavela, S., Rani, N., Palicharla, V.R., Pokorny, J.L., Sarkaria, J.N., et al., WWP2 is an E3 ubiquitin ligase for PTEN. *Nat Cell Biol*, 2011. **13**(6): p. 728-33.
162. Yang, Z., Yuan, X.G., Chen, J., and Lu, N.H., Is NEDD4-1 a negative regulator of phosphatase and tensin homolog in gastric carcinogenesis? *World J Gastroenterol*, 2012. **18**(43): p. 6345-8.
163. Eide, P.W., Cekaite, L., Danielsen, S.A., Eilertsen, I.A., Kjenseth, A., Fykerud, T.A., et al., NEDD4 is overexpressed in colorectal cancer and promotes colonic cell growth independently of the PI3K/PTEN/AKT pathway. *Cell Signal*, 2013. **25**(1): p. 12-8.
164. Liu, F., Wagner, S., Campbell, R.B., Nickerson, J.A., Schiffer, C.A., and Ross, A.H., PTEN enters the nucleus by diffusion. *J Cell Biochem*, 2005. **96**(2): p. 221-34.
165. Hopkins, B.D., Fine, B., Steinbach, N., Dendy, M., Rapp, Z., Shaw, J., et al., A secreted PTEN phosphatase that enters cells to alter signaling and survival. *Science*, 2013. **341**(6144): p. 399-402.

166. She, Q.B., Gruvberger-Saal, S.K., Maurer, M., Chen, Y., Jumppanen, M., Su, T., et al., Integrated molecular pathway analysis informs a synergistic combination therapy targeting PTEN/PI3K and EGFR pathways for basal-like breast cancer. *BMC Cancer*, 2016. **16**: p. 587.
167. Yndestad, S., Austreid, E., Knappskog, S., Chrisanthar, R., Lilleng, P.K., Lonning, P.E., et al., High PTEN gene expression is a negative prognostic marker in human primary breast cancers with preserved p53 function. *Breast Cancer Res Treat*, 2017. **163**(1): p. 177-190.
168. Forman, D., Stockton, D., Moller, H., Quinn, M., Babb, P., De Angelis, R., et al., Cancer prevalence in the UK: results from the EUROPREVAL study. *Ann Oncol*, 2003. **14**(4): p. 648-54.
169. Juliusson, G., Abrahamsson, J., Lazarevic, V., Antunovic, P., Derolf, A., Garelius, H., et al., Prevalence and characteristics of survivors from acute myeloid leukemia in Sweden. *Leukemia*, 2017. **31**(3): p. 728-731.
170. Oran, B. and Weisdorf, D.J., Survival for older patients with acute myeloid leukemia: a population-based study. *Haematologica*, 2012. **97**(12): p. 1916-24.
171. Deschler, B. and Lubbert, M., Acute myeloid leukemia: epidemiology and etiology. *Cancer*, 2006. **107**(9): p. 2099-107.
172. Genovese, G., Kahler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., et al., Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med*, 2014. **371**(26): p. 2477-87.
173. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., et al., Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med*, 2014. **371**(26): p. 2488-98.
174. Lowenberg, B., Downing, J.R., and Burnett, A., Acute myeloid leukemia. *N Engl J Med*, 1999. **341**(14): p. 1051-62.
175. Campidelli, C., Agostinelli, C., Stitson, R., and Pileri, S.A., Myeloid sarcoma: extramedullary manifestation of myeloid disorders. *Am J Clin Pathol*, 2009. **132**(3): p. 426-37.
176. Burnett, A.K., Acute myeloid leukemia: treatment of adults under 60 years. *Rev Clin Exp Hematol*, 2002. **6**(1): p. 26-45; discussion 86-7.
177. Niewerth, D., Creutzig, U., Bierings, M.B., and Kaspers, G.J., A review on allogeneic stem cell transplantation for newly diagnosed pediatric acute myeloid leukemia. *Blood*, 2010. **116**(13): p. 2205-14.
178. Koreth, J., Schlenk, R., Kopecky, K.J., Honda, S., Sierra, J., Djulbegovic, B.J., et al., Allogeneic stem cell transplantation for acute myeloid leukemia in first complete remission: systematic review and meta-analysis of prospective clinical trials. *JAMA*, 2009. **301**(22): p. 2349-61.
179. Grimwade, D. and Freeman, S.D., Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for "prime time"? *Blood*, 2014. **124**(23): p. 3345-55.

180. Voskova, D., Schoch, C., Schnittger, S., Hiddemann, W., Haferlach, T., and Kern, W., Stability of leukemia-associated aberrant immunophenotypes in patients with acute myeloid leukemia between diagnosis and relapse: comparison with cytomorphicologic, cytogenetic, and molecular genetic findings. *Cytometry B Clin Cytom*, 2004. **62**(1): p. 25-38.
181. Chen, W., Karandikar, N.J., McKenna, R.W., and Kroft, S.H., Stability of leukemia-associated immunophenotypes in precursor B-lymphoblastic leukemia/lymphoma: a single institution experience. *Am J Clin Pathol*, 2007. **127**(1): p. 39-46.
182. El-Rifai, W., Ruutu, T., Elonen, E., Volin, L., and Knuutila, S., Prognostic value of metaphase-fluorescence in situ hybridization in follow-up of patients with acute myeloid leukemia in remission. *Blood*, 1997. **89**(9): p. 3330-4.
183. Jongen-Lavrencic, M., Grob, T., Hanekamp, D., Kavelaars, F.G., Al Hinai, A., Zeilemaker, A., et al., Molecular Minimal Residual Disease in Acute Myeloid Leukemia. *N Engl J Med*, 2018. **378**(13): p. 1189-1199.
184. Kandath, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., et al., Mutational landscape and significance across 12 major cancer types. *Nature*, 2013. **502**(7471): p. 333-339.
185. Cancer Genome Atlas Research, N., Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., et al., Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 2013. **368**(22): p. 2059-74.
186. Thiede, C., Koch, S., Creutzig, E., Steudel, C., Illmer, T., Schaich, M., et al., Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood*, 2006. **107**(10): p. 4011-20.
187. Gorello, P., Cazzaniga, G., Alberti, F., Dell'Oro, M.G., Gottardi, E., Specchia, G., et al., Quantitative assessment of minimal residual disease in acute myeloid leukemia carrying nucleophosmin (NPM1) gene mutations. *Leukemia*, 2006. **20**(6): p. 1103-8.
188. Pettersson, L., Leveen, P., Axler, O., Dvorakova, D., Juliusson, G., and Ehinger, M., Improved minimal residual disease detection by targeted quantitative polymerase chain reaction in Nucleophosmin 1 type a mutated acute myeloid leukemia. *Genes Chromosomes Cancer*, 2016. **55**(10): p. 750-66.
189. Pettersson, L., Johansson Alm, S., Almstedt, A., Chen, Y., Orrsjo, G., Shah-Barkhordar, G., et al., Comparison of RNA- and DNA-based methods for measurable residual disease analysis in NPM1-mutated acute myeloid leukemia. *Int J Lab Hematol*, 2021. **43**(4): p. 664-674.
190. Ferno, M., Stal, O., Baldetorp, B., Hatschek, T., Kallstrom, A.C., Malmstrom, P., et al., Results of two or five years of adjuvant tamoxifen correlated to steroid receptor and S-phase levels. South Sweden Breast Cancer Group, and South-East Sweden Breast Cancer Group. *Breast Cancer Res Treat*, 2000. **59**(1): p. 69-76.
191. Borgquist, S., Hjertberg, M., Henningson, M., Ingvar, C., Rose, C., and Jernstrom, H., Given breast cancer, is fat better than thin? Impact of the estrogen receptor beta gene polymorphisms. *Breast Cancer Res Treat*, 2013. **137**(3): p. 849-62.
192. Zaha, D.C., Significance of immunohistochemistry in breast cancer. *World J Clin Oncol*, 2014. **5**(3): p. 382-92.

193. Dowsett, M., Allred, C., Knox, J., Quinn, E., Salter, J., Wale, C., et al., Relationship between quantitative estrogen and progesterone receptor expression and human epidermal growth factor receptor 2 (HER-2) status with recurrence in the Arimidex, Tamoxifen, Alone or in Combination trial. *J Clin Oncol*, 2008. **26**(7): p. 1059-65.
194. Cuzick, J., Dowsett, M., Pineda, S., Wale, C., Salter, J., Quinn, E., et al., Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *J Clin Oncol*, 2011. **29**(32): p. 4273-8.
195. Dowsett, M., Cooke, T., Ellis, I., Gullick, W.J., Gusterson, B., Mallon, E., et al., Assessment of HER2 status in breast cancer: why, when and how? *Eur J Cancer*, 2000. **36**(2): p. 170-6.
196. Rakha, E.A., Pinder, S.E., Bartlett, J.M., Ibrahim, M., Starczynski, J., Carder, P.J., et al., Updated UK Recommendations for HER2 assessment in breast cancer. *J Clin Pathol*, 2015. **68**(2): p. 93-9.
197. Goldhirsch, A., Ingle, J.N., Gelber, R.D., Coates, A.S., Thurlimann, B., Senn, H.J., et al., Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. *Ann Oncol*, 2009. **20**(8): p. 1319-29.
198. Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., et al., Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med*, 1998. **4**(7): p. 844-7.
199. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 1995. **270**(5235): p. 467-70.
200. Shalon, D., Smith, S.J., and Brown, P.O., A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, 1996. **6**(7): p. 639-45.
201. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., et al., A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 2002. **347**(25): p. 1999-2009.
202. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., et al., Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 2002. **415**(6871): p. 530-6.
203. Draghici, S., Khatri, P., Eklund, A.C., and Szallasi, Z., Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*, 2006. **22**(2): p. 101-9.
204. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y., RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 2008. **18**(9): p. 1509-17.
205. Wang, Z., Gerstein, M., and Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 2009. **10**(1): p. 57-63.
206. Marguerat, S. and Bahler, J., RNA-seq: from technology to biology. *Cell Mol Life Sci*, 2010. **67**(4): p. 569-79.

207. Xu, X., Zhang, Y., Williams, J., Antoniou, E., McCombie, W.R., Wu, S., et al., Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC Bioinformatics*, 2013. **14 Suppl 9**: p. S1.
208. Mantione, K.J., Kream, R.M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J.M., et al., Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med Sci Monit Basic Res*, 2014. **20**: p. 138-42.
209. Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K., and Liu, X., Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, 2014. **9**(1): p. e78644.
210. Li, B. and Dewey, C.N., RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 2011. **12**: p. 323.
211. Crick, F.H., On protein synthesis. *Symp Soc Exp Biol*, 1958. **12**: p. 138-63.
212. Crick, F., Central dogma of molecular biology. *Nature*, 1970. **227**(5258): p. 561-3.
213. Sanger, F., Nicklen, S., and Coulson, A.R., DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 1977. **74**(12): p. 5463-7.
214. Nyren, P., Pettersson, B., and Uhlen, M., Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem*, 1993. **208**(1): p. 171-5.
215. Ronaghi, M., Pyrosequencing sheds light on DNA sequencing. *Genome Res*, 2001. **11**(1): p. 3-11.
216. Rothberg, J.M. and Leamon, J.H., The development and impact of 454 sequencing. *Nat Biotechnol*, 2008. **26**(10): p. 1117-24.
217. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., et al., A large genome center's improvements to the Illumina sequencing system. *Nat Methods*, 2008. **5**(12): p. 1005-10.
218. Shendure, J. and Ji, H., Next-generation DNA sequencing. *Nat Biotechnol*, 2008. **26**(10): p. 1135-45.
219. Mardis, E.R., Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 2008. **9**: p. 387-402.
220. Metzker, M.L., Sequencing technologies - the next generation. *Nat Rev Genet*, 2010. **11**(1): p. 31-46.
221. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al., Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012. **2012**: p. 251364.
222. Merriman, B., Ion Torrent, R., Team, D., and Rothberg, J.M., Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, 2012. **33**(23): p. 3397-417.
223. Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., et al., Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*, 2011. **9**(1): p. 72-4.
224. Smith, T., Heger, A., and Sudbery, I., UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*, 2017. **27**(3): p. 491-499.

225. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al., Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*, 2014. **11**(2): p. 163-6.
226. Fu, Y., Wu, P.H., Beane, T., Zamore, P.D., and Weng, Z., Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics*, 2018. **19**(1): p. 531.
227. Kou, R., Lam, H., Duan, H., Ye, L., Jongkam, N., Chen, W., et al., Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLoS One*, 2016. **11**(1): p. e0146638.
228. Filges, S., Yamada, E., Stahlberg, A., and Godfrey, T.E., Impact of Polymerase Fidelity on Background Error Rates in Next-Generation Sequencing with Unique Molecular Identifiers/Barcodes. *Sci Rep*, 2019. **9**(1): p. 3503.
229. Schadt, E.E., Turner, S., and Kasarskis, A., A window into third-generation sequencing. *Hum Mol Genet*, 2010. **19**(R2): p. R227-40.
230. McCarthy, A., Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol*, 2010. **17**(7): p. 675-6.
231. Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., and DePristo, M.A., Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 2012. **13**: p. 375.
232. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., et al., Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, 2012. **7**(11): p. e47768.
233. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., et al., The potential and challenges of nanopore sequencing. *Nat Biotechnol*, 2008. **26**(10): p. 1146-53.
234. Depledge, D.P., Srinivas, K.P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D.G., et al., Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun*, 2019. **10**(1): p. 754.
235. Derrington, I.M., Butler, T.Z., Collins, M.D., Manrao, E., Pavlenok, M., Niederweis, M., et al., Nanopore DNA sequencing with MspA. *Proc Natl Acad Sci U S A*, 2010. **107**(37): p. 16060-5.
236. Jain, M., Olsen, H.E., Paten, B., and Akeson, M., The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*, 2016. **17**(1): p. 239.
237. Venkatesan, B.M. and Bashir, R., Nanopore sensors for nucleic acid analysis. *Nat Nanotechnol*, 2011. **6**(10): p. 615-24.
238. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q., Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, 2020. **21**(1): p. 30.
239. Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L., and Schnable, P.S., SNP discovery via 454 transcriptome sequencing. *Plant J*, 2007. **51**(5): p. 910-8.

240. Emrich, S.J., Barbazuk, W.B., Li, L., and Schnable, P.S., Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res*, 2007. **17**(1): p. 69-73.
241. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., et al., Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 2008. **133**(3): p. 523-36.
242. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J., Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 2008. **40**(12): p. 1413-5.
243. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., et al., A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 2008. **321**(5891): p. 956-60.
244. Wilhelm, B.T. and Landry, J.R., RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 2009. **48**(3): p. 249-57.
245. Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., et al., The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol*, 2014. **32**(9): p. 926-32.
246. Li, J., Hou, R., Niu, X., Liu, R., Wang, Q., Wang, C., et al., Comparison of microarray and RNA-Seq analysis of mRNA expression in dermal mesenchymal stem cells. *Biotechnol Lett*, 2016. **38**(1): p. 33-41.
247. Liu, Y., Morley, M., Brandimarto, J., Hannehalli, S., Hu, Y., Ashley, E.A., et al., RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics*, 2015. **105**(2): p. 83-9.
248. Lorenz, D.A., Sathe, S., Einstein, J.M., and Yeo, G.W., Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. *RNA*, 2020. **26**(1): p. 19-28.
249. Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D., et al., Nanopore direct RNA sequencing maps the complexity of *Arabidopsis* mRNA processing and m(6)A modification. *Elife*, 2020. **9**.
250. Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Holzer, M., et al., Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res*, 2019. **29**(9): p. 1545-1554.
251. Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database, C., The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, 2012. **40**(Database issue): p. D54-6.
252. Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C., The sequence read archive. *Nucleic Acids Res*, 2011. **39**(Database issue): p. D19-21.
253. Stark, R., Grzelak, M., and Hadfield, J., RNA sequencing: the teenage years. *Nat Rev Genet*, 2019. **20**(11): p. 631-656.



254. Saal, L.H., Vallon-Christersson, J., Hakkinen, J., Hegardt, C., Grabau, D., Winter, C., et al., The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med*, 2015. **7**(1): p. 20.
255. Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., et al., Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*, 2009. **37**(18): p. e123.
256. Brueffer, C., Gladchuk, S., Winter, C., Vallon-Christersson, J., Hegardt, C., Hakkinen, J., et al., The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome. *EMBO Mol Med*, 2020. **12**(10): p. e12118.
257. Broad Institute. Picard Tools - By Broad Institute. 2009; Available from: <https://broadinstitute.github.io/picard/>. [accessed 2021-09-03].
258. Bolger, A.M., Lohse, M., and Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014. **30**(15): p. 2114-20.
259. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., et al., VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*, 2016. **44**(11): p. e108.
260. Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., et al., The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet*, 2008. **Chapter 10**: p. Unit 10 11.
261. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al., COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, 2017. **45**(D1): p. D777-D783.
262. Chen, N., Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*, 2004. **Chapter 4**: p. Unit 4 10.
263. Tarailo-Graovac, M. and Chen, N., Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*, 2009. **Chapter 4**: p. Unit 4 10.
264. Ramaswami, G. and Li, J.B., RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res*, 2014. **42**(Database issue): p. D109-13.
265. Kiran, A. and Baranov, P.V., DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics*, 2010. **26**(14): p. 1772-6.
266. Ameur, A., Dahlberg, J., Olason, P., Vezzi, F., Karlsson, R., Martin, M., et al., SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet*, 2017. **25**(11): p. 1253-1260.
267. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., et al., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 2001. **29**(1): p. 308-11.
268. Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., et al., Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 2012. **486**(7403): p. 405-9.

269. Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., et al., Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 2012. **486**(7403): p. 353-60.
270. Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I., Pleasance, E.D., Simpson, J.T., et al., Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 2009. **462**(7276): p. 1005-10.
271. Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., et al., Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 2009. **461**(7265): p. 809-13.
272. Leary, R.J., Sausen, M., Kinde, I., Papadopoulos, N., Carpten, J.D., Craig, D., et al., Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med*, 2012. **4**(162): p. 162ra154.
273. Chen, Y., George, A.M., Olsson, E., and Saal, L.H., Identification and Use of Personalized Genomic Markers for Monitoring Circulating Tumor DNA. *Methods Mol Biol*, 2018. **1768**: p. 303-322.
274. Genome Reference Consortium. Human Genome Assembly GRCh38.p13. 2019; Available from: <https://www.ncbi.nlm.nih.gov/grc/human/data>.
275. Illumina. Sequencing Coverage Calculator. Available from: [https://emea.support.illumina.com/downloads/sequencing\\_coverage\\_calculator.html](https://emea.support.illumina.com/downloads/sequencing_coverage_calculator.html) [accessed 2021-09-03].
276. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., et al., BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 2009. **6**(9): p. 677-81.
277. Alkner, S., Tang, M.H., Brueffer, C., Dahlgren, M., Chen, Y., Olsson, E., et al., Contralateral breast cancer can represent a metastatic spread of the first primary tumor: determination of clonal relationship between contralateral breast cancers using next-generation whole genome sequencing. *Breast Cancer Res*, 2015. **17**: p. 102.
278. Tang, M.H., Dahlgren, M., Brueffer, C., Tjitrowirjo, T., Winter, C., Chen, Y., et al., Remarkable similarities of chromosomal rearrangements between primary human breast cancers and matched distant metastases as revealed by whole-genome sequencing. *Oncotarget*, 2015. **6**(35): p. 37169-84.
279. Wong, S.Q., Li, J., Salemi, R., Sheppard, K.E., Do, H., Tothill, R.W., et al., Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. *Sci Rep*, 2013. **3**: p. 3494.
280. Feldman, M.Y., Reactions of nucleic acids and nucleoproteins with formaldehyde. *Prog Nucleic Acid Res Mol Biol*, 1973. **13**: p. 1-49.
281. Do, H. and Dobrovic, A., Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin Chem*, 2015. **61**(1): p. 64-71.
282. Kiyoi, H., Naoe, T., Nakano, Y., Yokota, S., Minami, S., Miyawaki, S., et al., Prognostic implication of FLT3 and N-RAS gene mutations in acute myeloid leukemia. *Blood*, 1999. **93**(9): p. 3074-80.
283. Stirewalt, D.L., Kopecky, K.J., Meshinchi, S., Appelbaum, F.R., Slovak, M.L., Willman, C.L., et al., FLT3, RAS, and TP53 mutations in elderly patients with acute myeloid leukemia. *Blood*, 2001. **97**(11): p. 3589-95.

284. Suzuki, T., Kiyoi, H., Ozeki, K., Tomita, A., Yamaji, S., Suzuki, R., et al., Clinical characteristics and prognostic implications of NPM1 mutations in acute myeloid leukemia. *Blood*, 2005. **106**(8): p. 2854-61.
285. Tang, J.L., Hou, H.A., Chen, C.Y., Liu, C.Y., Chou, W.C., Tseng, M.H., et al., AML1/RUNX1 mutations in 470 adult patients with de novo acute myeloid leukemia: prognostic implication and interaction with other gene alterations. *Blood*, 2009. **114**(26): p. 5352-61.
286. Dohner, H., Estey, E.H., Amadori, S., Appelbaum, F.R., Buchner, T., Burnett, A.K., et al., Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*, 2010. **115**(3): p. 453-74.
287. Paschka, P., Schlenk, R.F., Gaidzik, V.I., Habdank, M., Kronke, J., Bullinger, L., et al., IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication. *J Clin Oncol*, 2010. **28**(22): p. 3636-43.
288. Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., et al., DNMT3A mutations in acute myeloid leukemia. *N Engl J Med*, 2010. **363**(25): p. 2424-33.
289. Gaidzik, V.I., Bullinger, L., Schlenk, R.F., Zimmermann, A.S., Rock, J., Paschka, P., et al., RUNX1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the AML study group. *J Clin Oncol*, 2011. **29**(10): p. 1364-72.
290. Li, Z., Stolzel, F., Onel, K., Sukhanova, M., Mirza, M.K., Yap, K.L., et al., Next-generation sequencing reveals clinically actionable molecular markers in myeloid sarcoma. *Leukemia*, 2015. **29**(10): p. 2113-6.
291. Alonso, C.M., Llop, M., Sargas, C., Pedrola, L., Panadero, J., Hervas, D., et al., Clinical Utility of a Next-Generation Sequencing Panel for Acute Myeloid Leukemia Diagnostics. *J Mol Diagn*, 2019. **21**(2): p. 228-240.
292. Integrated DNA Technologies. xGen Acute Myeloid Leukemia Cancer Panel. Available from: <https://eu.idtdna.com/pages/products/next-generation-sequencing/targeted-sequencing/hybridization-capture/predesigned-panels/xgen-aml-panel> [accessed 2021-09-06].
293. Thermo Fisher Scientific. Oncomine Myeloid Research Assay. Available from: <https://www.thermofisher.com/se/en/home/clinical/preclinical-companion-diagnostic-development/oncomine-oncology/oncomine-myeloid-research-assay.html> [accessed 2021-09-06].
294. Illumina. TruSight Myeloid Sequencing Panel. Available from: <https://emea.illumina.com/products/by-type/clinical-research-products/trusight-myeloid.html> [accessed 2021-09-06].
295. Sakharkar, M.K., Chow, V.T., and Kanguane, P., Distributions of exons and introns in the human genome. *In Silico Biol*, 2004. **4**(4): p. 387-93.

296. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., et al., Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 2009. **461**(7261): p. 272-6.
297. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009. **25**(14): p. 1754-60.
298. Faust, G.G. and Hall, I.M., SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 2014. **30**(17): p. 2503-5.
299. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Kallberg, M., et al., Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*, 2018. **15**(8): p. 591-594.
300. Lazarevic, V., Orsmark-Pietras, C., Lilljebjorn, H., Pettersson, L., Rissler, M., Lubking, A., et al., Isolated myeloid sarcoma is characterized by recurrent NFE2 mutations and concurrent preleukemic clones in the bone marrow. *Blood*, 2018. **131**(5): p. 577-581.
301. Banfi, G., Salvagno, G.L., and Lippi, G., The role of ethylenediamine tetraacetic acid (EDTA) as in vitro anticoagulant for diagnostic purposes. *Clin Chem Lab Med*, 2007. **45**(5): p. 565-76.
302. Lam, N.Y., Rainer, T.H., Chiu, R.W., and Lo, Y.M., EDTA is a better anticoagulant than heparin or citrate for delayed blood processing for plasma DNA analysis. *Clin Chem*, 2004. **50**(1): p. 256-7.
303. Rossen, L., Norskov, P., Holmstrom, K., and Rasmussen, O.F., Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions. *Int J Food Microbiol*, 1992. **17**(1): p. 37-45.
304. Streck. Streck Cell-Free DNA BCT®. Available from: <https://www.streck.com/products/stabilization/cell-free-dna-bct/> [accessed 2021-08-26].
305. Toro, P.V., Erlanger, B., Beaver, J.A., Cochran, R.L., VanDenBerg, D.A., Yakim, E., et al., Comparison of cell stabilizing blood collection tubes for circulating plasma tumor DNA. *Clin Biochem*, 2015. **48**(15): p. 993-8.
306. Medina Diaz, I., Nocon, A., Mehnert, D.H., Fredebohm, J., Diehl, F., and Holtrup, F., Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing. *PLoS One*, 2016. **11**(11): p. e0166354.
307. Risberg, B., Tsui, D.W.Y., Biggs, H., Ruiz-Valdepenas Martin de Almagro, A., Dawson, S.J., Hodgkin, C., et al., Effects of Collection and Processing Procedures on Plasma Circulating Cell-Free DNA from Cancer Patients. *J Mol Diagn*, 2018. **20**(6): p. 883-892.
308. CellSearch. CellSave Preservative Tubes. Available from: <https://www.cellsearchctc.com/product-systems-overview/cellsave-preservative-tubes> [accessed 2021-08-27].
309. Allard, W.J., Matera, J., Miller, M.C., Repollet, M., Connelly, M.C., Rao, C., et al., Tumor cells circulate in the peripheral blood of all major carcinomas but not in healthy subjects or patients with nonmalignant diseases. *Clin Cancer Res*, 2004. **10**(20): p. 6897-904.

310. Kang, Q., Henry, N.L., Paoletti, C., Jiang, H., Vats, P., Chinnaiyan, A.M., et al., Comparative analysis of circulating tumor DNA stability In K3EDTA, Streck, and CellSave blood collection tubes. *Clin Biochem*, 2016. **49**(18): p. 1354-1360.
311. Shaw, K.J., Thain, L., Docker, P.T., Dyer, C.E., Greenman, J., Greenway, G.M., et al., The use of carrier RNA to enhance DNA extraction from microfluidic-based silica monoliths. *Anal Chim Acta*, 2009. **652**(1-2): p. 231-3.
312. Kishore, R., Reef Hardy, W., Anderson, V.J., Sanchez, N.A., and Buoncristiani, M.R., Optimization of DNA extraction from low-yield and degraded samples using the BioRobot EZ1 and BioRobot M48. *J Forensic Sci*, 2006. **51**(5): p. 1055-61.
313. Ebeling, W., Hennrich, N., Klockow, M., Metz, H., Orth, H.D., and Lang, H., Proteinase K from *Tritirachium album* Limber. *Eur J Biochem*, 1974. **47**(1): p. 91-7.
314. Goldenberger, D., Perschil, I., Ritzler, M., and Altwegg, M., A simple "universal" DNA extraction procedure using SDS and proteinase K is compatible with direct PCR amplification. *PCR Methods Appl*, 1995. **4**(6): p. 368-70.
315. Warton, K., Yuwono, N.L., Cowley, M.J., McCabe, M.J., So, A., and Ford, C.E., Evaluation of Streck BCT and PAXgene Stabilised Blood Collection Tubes for Cell-Free Circulating DNA Studies in Plasma. *Mol Diagn Ther*, 2017. **21**(5): p. 563-570.
316. Sorber, L., Zwaenepoel, K., Deschoolmeester, V., Roeyen, G., Lardon, F., Rolfo, C., et al., A Comparison of Cell-Free DNA Isolation Kits: Isolation and Quantification of Cell-Free DNA in Plasma. *J Mol Diagn*, 2017. **19**(1): p. 162-168.
317. Baeuerle, P.A. and Gires, O., EpCAM (CD326) finding its role in cancer. *Br J Cancer*, 2007. **96**(3): p. 417-23.
318. Went, P.T., Lugli, A., Meier, S., Bundi, M., Mirlacher, M., Sauter, G., et al., Frequent EpCam protein expression in human carcinomas. *Hum Pathol*, 2004. **35**(1): p. 122-8.
319. Adan, A., Alizada, G., Kiraz, Y., Baran, Y., and Nalbant, A., Flow cytometry: basic principles and applications. *Crit Rev Biotechnol*, 2017. **37**(2): p. 163-176.
320. Wilkerson, M.J., Principles and applications of flow cytometry and cell sorting in companion animal medicine. *Vet Clin North Am Small Anim Pract*, 2012. **42**(1): p. 53-71.
321. Barlogie, B., Raber, M.N., Schumann, J., Johnson, T.S., Drewinko, B., Swartzendruber, D.E., et al., Flow cytometry in clinical cancer research. *Cancer Res*, 1983. **43**(9): p. 3982-97.
322. Brown, M. and Wittwer, C., Flow cytometry: principles and clinical applications in hematology. *Clin Chem*, 2000. **46**(8 Pt 2): p. 1221-9.
323. Lacombe, F., Durrieu, F., Briais, A., Dumain, P., Belloc, F., Bascans, E., et al., Flow cytometry CD45 gating for immunophenotyping of acute myeloid leukemia. *Leukemia*, 1997. **11**(11): p. 1878-86.
324. Ossenkoppele, G. and Schuurhuis, G.J., MRD in AML: does it already guide therapy decision-making? *Hematology Am Soc Hematol Educ Program*, 2016. **2016**(1): p. 356-365.

325. Jaso, J.M., Wang, S.A., Jorgensen, J.L., and Lin, P., Multi-color flow cytometric immunophenotyping for detection of minimal residual disease in AML: past, present and future. *Bone Marrow Transplant*, 2014. **49**(9): p. 1129-38.
326. Kapuscinski, J., DAPI: a DNA-specific fluorescent probe. *Biotech Histochem*, 1995. **70**(5): p. 220-33.
327. Otto, F., DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. *Methods Cell Biol*, 1990. **33**: p. 105-10.
328. Schweizer, D., Reverse fluorescent chromosome banding with chromomycin and DAPI. *Chromosoma*, 1976. **58**(4): p. 307-24.
329. Charbonneau, H., Tonks, N.K., Walsh, K.A., and Fischer, E.H., The leukocyte common antigen (CD45): a putative receptor-linked protein tyrosine phosphatase. *Proc Natl Acad Sci U S A*, 1988. **85**(19): p. 7182-6.
330. Holmes, N., CD45: all is not yet crystal clear. *Immunology*, 2006. **117**(2): p. 145-55.
331. Trowbridge, I.S., Ostergaard, H.L., and Johnson, P., CD45: a leukocyte-specific member of the protein tyrosine phosphatase family. *Biochim Biophys Acta*, 1991. **1095**(1): p. 46-56.
332. Braun, S., Pantel, K., Muller, P., Janni, W., Hepp, F., Kentenich, C.R., et al., Cytokeratin-positive cells in the bone marrow and survival of patients with stage I, II, or III breast cancer. *N Engl J Med*, 2000. **342**(8): p. 525-33.
333. Gusterson, B.A., Ross, D.T., Heath, V.J., and Stein, T., Basal cytokeratins and their relationship to the cellular origin and functional classification of breast cancer. *Breast Cancer Res*, 2005. **7**(4): p. 143-8.
334. Korsching, E., Packeisen, J., Agelopoulos, K., Eisenacher, M., Voss, R., Isola, J., et al., Cytogenetic alterations and cytokeratin expression patterns in breast cancer: integrating a new model of breast differentiation into cytogenetic pathways of breast carcinogenesis. *Lab Invest*, 2002. **82**(11): p. 1525-33.
335. Laakso, M., Loman, N., Borg, A., and Isola, J., Cytokeratin 5/14-positive breast cancer: true basal phenotype confined to BRCA1 tumors. *Mod Pathol*, 2005. **18**(10): p. 1321-8.
336. Cristofanilli, M., Budd, G.T., Ellis, M.J., Stopeck, A., Matera, J., Miller, M.C., et al., Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med*, 2004. **351**(8): p. 781-91.
337. CellSearch. CellTracks Analyzer II. Available from: <https://www.cellsearchctc.com/product-systems-overview/celltracks-analyzer> [accessed 2021-08-27].
338. Janni, W.J., Rack, B., Terstappen, L.W., Pierga, J.Y., Taran, F.A., Fehm, T., et al., Pooled Analysis of the Prognostic Relevance of Circulating Tumor Cells in Primary Breast Cancer. *Clin Cancer Res*, 2016. **22**(10): p. 2583-93.
339. Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B., and Erlich, H.A., Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature*, 1986. **324**(6093): p. 163-6.

340. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., et al., Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 1988. **239**(4839): p. 487-91.
341. Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., et al., Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 1985. **230**(4732): p. 1350-4.
342. Kasai, K., Nakamura, Y., and White, R., Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. *J Forensic Sci*, 1990. **35**(5): p. 1196-200.
343. Shen, M., Zhou, Y., Ye, J., Abdullah Al-Maskri, A.A., Kang, Y., Zeng, S., et al., Recent advances and perspectives of nucleic acid detection for coronavirus. *J Pharm Anal*, 2020. **10**(2): p. 97-101.
344. Tahamtan, A. and Ardebili, A., Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert Rev Mol Diagn*, 2020. **20**(5): p. 453-454.
345. Machhi, J., Herskovitz, J., Senan, A.M., Dutta, D., Nath, B., Oleynikov, M.D., et al., The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections. *J Neuroimmune Pharmacol*, 2020. **15**(3): p. 359-386.
346. Chien, A., Edgar, D.B., and Trela, J.M., Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. *J Bacteriol*, 1976. **127**(3): p. 1550-7.
347. Cline, J., Braman, J.C., and Hogrefe, H.H., PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res*, 1996. **24**(18): p. 3546-51.
348. Lundberg, K.S., Shoemaker, D.D., Adams, M.W., Short, J.M., Sorge, J.A., and Mathur, E.J., High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene*, 1991. **108**(1): p. 1-6.
349. Bio-Rad. C1000 Touch Thermal Cycler. Available from: [https://www.bio-rad.com/sites/default/files/2021-08/Bulletin\\_6095.pdf](https://www.bio-rad.com/sites/default/files/2021-08/Bulletin_6095.pdf) [accessed 2021-08-23].
350. Thermo Fisher Scientific. Thermo Fisher thermal cyclers. Available from: <https://assets.thermofisher.com/TFS-Assets/BID/brochures/qpcr-pcr-solutions-brochure.pdf> [accessed 2021-09-03].
351. LePecq, J.B. and Paoletti, C., A fluorescent complex between ethidium bromide and nucleic acids. Physical-chemical characterization. *J Mol Biol*, 1967. **27**(1): p. 87-106.
352. Waring, M.J., Complex formation between ethidium bromide and nucleic acids. *J Mol Biol*, 1965. **13**(1): p. 269-82.
353. Dragan, A.I., Pavlovic, R., McGivney, J.B., Casas-Finet, J.R., Bishop, E.S., Strouse, R.J., et al., SYBR Green I: fluorescence properties and interaction with DNA. *J Fluoresc*, 2012. **22**(4): p. 1189-99.
354. Zipper, H., Brunner, H., Bernhagen, J., and Vitzthum, F., Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic Acids Res*, 2004. **32**(12): p. e103.
355. Eischeid, A.C., SYTO dyes and EvaGreen outperform SYBR Green in real-time PCR. *BMC Res Notes*, 2011. **4**: p. 263.

356. Mao, F., Leung, W.Y., and Xin, X., Characterization of EvaGreen and the implication of its physicochemical properties for qPCR applications. *BMC Biotechnol*, 2007. **7**: p. 76.
357. Holland, P.M., Abramson, R.D., Watson, R., and Gelfand, D.H., Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci U S A*, 1991. **88**(16): p. 7276-80.
358. de Kok, J.B., Wiegerinck, E.T., Giesendorf, B.A., and Swinkels, D.W., Rapid genotyping of single nucleotide polymorphisms using novel minor groove binding DNA oligonucleotides (MGB probes). *Hum Mutat*, 2002. **19**(5): p. 554-9.
359. Kutuyavin, I.V., Afonina, I.A., Mills, A., Gorn, V.V., Lukhtanov, E.A., Belousov, E.S., et al., 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res*, 2000. **28**(2): p. 655-61.
360. Braasch, D.A. and Corey, D.R., Locked nucleic acid (LNA): fine-tuning the recognition of DNA and RNA. *Chem Biol*, 2001. **8**(1): p. 1-7.
361. Vester, B. and Wengel, J., LNA (locked nucleic acid): high-affinity targeting of complementary RNA and DNA. *Biochemistry*, 2004. **43**(42): p. 13233-41.
362. Falini, B., Mecucci, C., Tiacci, E., Alcalay, M., Rosati, R., Pasqualucci, L., et al., Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med*, 2005. **352**(3): p. 254-66.
363. Heid, C.A., Stevens, J., Livak, K.J., and Williams, P.M., Real time quantitative PCR. *Genome Res*, 1996. **6**(10): p. 986-94.
364. Arya, M., Shergill, I.S., Williamson, M., Gommersall, L., Arya, N., and Patel, H.R., Basic principles of real-time quantitative PCR. *Expert Rev Mol Diagn*, 2005. **5**(2): p. 209-19.
365. Chou, W.C., Tang, J.L., Wu, S.J., Tsay, W., Yao, M., Huang, S.Y., et al., Clinical implications of minimal residual disease monitoring by quantitative polymerase chain reaction in acute myeloid leukemia patients bearing nucleophosmin (NPM1) mutations. *Leukemia*, 2007. **21**(5): p. 998-1004.
366. Moppett, J., van der Velden, V.H., Wijkhuijs, A.J., Hancock, J., van Dongen, J.J., and Goulden, N., Inhibition affecting RQ-PCR-based assessment of minimal residual disease in acute lymphoblastic leukemia: reversal by addition of bovine serum albumin. *Leukemia*, 2003. **17**(1): p. 268-70.
367. van der Velden, V.H., Hochhaus, A., Cazzaniga, G., Szczepanski, T., Gabert, J., and van Dongen, J.J., Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. *Leukemia*, 2003. **17**(6): p. 1013-34.
368. Sykes, P.J., Neoh, S.H., Brisco, M.J., Hughes, E., Condon, J., and Morley, A.A., Quantitation of targets for PCR by use of limiting dilution. *Biotechniques*, 1992. **13**(3): p. 444-9.
369. Vogelstein, B. and Kinzler, K.W., Digital PCR. *Proc Natl Acad Sci U S A*, 1999. **96**(16): p. 9236-41.



370. Henrich, T.J., Gallien, S., Li, J.Z., Pereyra, F., and Kuritzkes, D.R., Low-level detection and quantitation of cellular HIV-1 DNA and 2-LTR circles using droplet digital PCR. *J Virol Methods*, 2012. **186**(1-2): p. 68-72.
371. Kiss, M.M., Ortoleva-Donnelly, L., Beer, N.R., Warner, J., Bailey, C.G., Colston, B.W., et al., High-throughput quantitative polymerase chain reaction in picoliter droplets. *Anal Chem*, 2008. **80**(23): p. 8975-81.
372. Bio-Rad. QX200 Droplet Digital PCR System. Available from: <https://www.bio-rad.com/en-se/product/qx200-droplet-digital-pcr-system?ID=MPOQQE4VY> [accessed 2021-08-25].
373. Hindson, C.M., Chevillet, J.R., Briggs, H.A., Gallichotte, E.N., Ruf, I.K., Hindson, B.J., et al., Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods*, 2013. **10**(10): p. 1003-5.
374. Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., et al., High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem*, 2011. **83**(22): p. 8604-10.
375. Sanders, R., Huggett, J.F., Bushell, C.A., Cowen, S., Scott, D.J., and Foy, C.A., Evaluation of digital PCR for absolute DNA quantification. *Anal Chem*, 2011. **83**(17): p. 6474-84.
376. Strain, M.C., Lada, S.M., Luong, T., Rought, S.E., Gianella, S., Terry, V.H., et al., Highly precise measurement of HIV DNA by droplet digital PCR. *PLoS One*, 2013. **8**(4): p. e55943.
377. Taylor, S.C., Laperriere, G., and Germain, H., Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data. *Sci Rep*, 2017. **7**(1): p. 2409.
378. Whale, A.S., Huggett, J.F., Cowen, S., Speirs, V., Shaw, J., Ellison, S., et al., Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. *Nucleic Acids Res*, 2012. **40**(11): p. e82.
379. Zhang, B.O., Xu, C.W., Shao, Y., Wang, H.T., Wu, Y.F., Song, Y.Y., et al., Comparison of droplet digital PCR and conventional quantitative PCR for measuring EGFR gene mutation. *Exp Ther Med*, 2015. **9**(4): p. 1383-1388.
380. Zimmermann, B.G., Grill, S., Holzgreve, W., Zhong, X.Y., Jackson, L.G., and Hahn, S., Digital PCR: a powerful new tool for noninvasive prenatal diagnosis? *Prenat Diagn*, 2008. **28**(12): p. 1087-93.
381. Suo, T., Liu, X., Feng, J., Guo, M., Hu, W., Guo, D., et al., ddPCR: a more accurate tool for SARS-CoV-2 detection in low viral load specimens. *Emerg Microbes Infect*, 2020. **9**(1): p. 1259-1268.
382. Zhong, Q., Bhattacharya, S., Kotsopoulos, S., Olson, J., Taly, V., Griffiths, A.D., et al., Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR. *Lab Chip*, 2011. **11**(13): p. 2167-74.
383. Racki, N., Dreo, T., Gutierrez-Aguirre, I., Blejec, A., and Ravnikar, M., Reverse transcriptase droplet digital PCR shows high resilience to PCR inhibitors from plant, soil and water samples. *Plant Methods*, 2014. **10**(1): p. 42.

384. Dingle, T.C., Sedlak, R.H., Cook, L., and Jerome, K.R., Tolerance of droplet-digital PCR vs real-time quantitative PCR to inhibitory substances. *Clin Chem*, 2013. **59**(11): p. 1670-2.
385. Yucel, D. and Dalva, K., Effect of in vitro hemolysis on 25 common biochemical tests. *Clin Chem*, 1992. **38**(4): p. 575-7.
386. Nishimura, F., Uno, N., Chiang, P.C., Kaku, N., Morinaga, Y., Hasegawa, H., et al., The Effect of In Vitro Hemolysis on Measurement of Cell-Free DNA. *J Appl Lab Med*, 2019. **4**(2): p. 235-240.
387. Gormally, E., Hainaut, P., Caboux, E., Airoidi, L., Autrup, H., Malaveille, C., et al., Amount of DNA in plasma and cancer risk: a prospective study. *Int J Cancer*, 2004. **111**(5): p. 746-9.
388. Meddeb, R., Dache, Z.A.A., Thezenas, S., Otandault, A., Tanos, R., Pastor, B., et al., Quantifying circulating cell-free DNA in humans. *Sci Rep*, 2019. **9**(1): p. 5220.
389. Mussolin, L., Burnelli, R., Pillon, M., Carraro, E., Farruggia, P., Todesco, A., et al., Plasma cell-free DNA in paediatric lymphomas. *J Cancer*, 2013. **4**(4): p. 323-9.
390. Jen, J., Wu, L., and Sidransky, D., An overview on the isolation and analysis of circulating tumor DNA in plasma and serum. *Ann N Y Acad Sci*, 2000. **906**: p. 8-12.
391. Schwarzenbach, H., Hoon, D.S., and Pantel, K., Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer*, 2011. **11**(6): p. 426-37.
392. Fleischhacker, M. and Schmidt, B., Circulating nucleic acids (CNAs) and cancer--a survey. *Biochim Biophys Acta*, 2007. **1775**(1): p. 181-232.
393. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., et al., Integrative genomics viewer. *Nat Biotechnol*, 2011. **29**(1): p. 24-6.
394. Breitbach, S., Tug, S., and Simon, P., Circulating cell-free DNA: an up-coming molecular marker in exercise physiology. *Sports Med*, 2012. **42**(7): p. 565-86.
395. Jiang, P. and Lo, Y.M.D., The Long and Short of Circulating Cell-Free DNA and the Ins and Outs of Molecular Diagnostics. *Trends Genet*, 2016. **32**(6): p. 360-371.
396. Metz, C.E., Basic principles of ROC analysis. *Semin Nucl Med*, 1978. **8**(4): p. 283-98.
397. Goya, R., Sun, M.G., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., et al., SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 2010. **26**(6): p. 730-6.
398. Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., et al., Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods*, 2015. **12**(7): p. 623-30.
399. Xu, C., A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J*, 2018. **16**: p. 15-24.
400. Newton, C.R., Graham, A., Heptinstall, L.E., Powell, S.J., Summers, C., Kalsheker, N., et al., Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Res*, 1989. **17**(7): p. 2503-16.
401. Little, S., Amplification-refractory mutation system (ARMS) analysis of point mutations. *Curr Protoc Hum Genet*, 2001. **Chapter 9**: p. Unit 9 8.
402. Eckert, K.A. and Kunkel, T.A., DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl*, 1991. **1**(1): p. 17-24.

403. Joyce, C.M. and Benkovic, S.J., DNA polymerase fidelity: kinetics, structure, and checkpoints. *Biochemistry*, 2004. **43**(45): p. 14317-24.
404. McNerney, P., Adams, P., and Hadi, M.Z., Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol Biol Int*, 2014. **2014**: p. 287430.
405. Isaksson, S., George, A.M., Jonsson, M., Cirenajwis, H., Jonsson, P., Bendahl, P.O., et al., Pre-operative plasma cell-free circulating tumor DNA and serum protein tumor markers as predictors of lung adenocarcinoma recurrence. *Acta Oncol*, 2019. **58**(8): p. 1079-1086.
406. Arildsen, N.S., Martin de la Fuente, L., Masback, A., Malander, S., Forslund, O., Kannisto, P., et al., Detecting TP53 mutations in diagnostic and archival liquid-based Pap samples from ovarian cancer patients using an ultra-sensitive ddPCR method. *Sci Rep*, 2019. **9**(1): p. 15506.
407. Dahlgren, M., George, A.M., Brueffer, C., Gladchuk, S., Chen, Y., Vallon-Christersson, J., et al., Preexisting Somatic Mutations of Estrogen Receptor Alpha (ESR1) in Early-Stage Primary Breast Cancer. *JNCI Cancer Spectr*, 2021. **5**(2): p. pkab028.
408. Integrated DNA Technologies, OligoAnalyzer Tool. (2021-09-06).
409. Allawi, H.T. and SantaLucia, J., Jr., Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry*, 1997. **36**(34): p. 10581-94.
410. Bommarito, S., Peyret, N., and SantaLucia, J., Jr., Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res*, 2000. **28**(9): p. 1929-34.
411. McTigue, P.M., Peterson, R.J., and Kahn, J.D., Sequence-dependent thermodynamic parameters for locked nucleic acid (LNA)-DNA duplex formation. *Biochemistry*, 2004. **43**(18): p. 5388-405.
412. Owczarzy, R., You, Y., Moreira, B.G., Manthey, J.A., Huang, L., Behlke, M.A., et al., Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry*, 2004. **43**(12): p. 3537-54.
413. SantaLucia, J., Jr. and Hicks, D., The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct*, 2004. **33**: p. 415-40.
414. Owczarzy, R., Moreira, B.G., You, Y., Behlke, M.A., and Walder, J.A., Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry*, 2008. **47**(19): p. 5336-53.
415. Owczarzy, R., You, Y., Groth, C.L., and Tataurov, A.V., Stability and mismatch discrimination of locked nucleic acid-DNA duplexes. *Biochemistry*, 2011. **50**(43): p. 9352-67.
416. Collins, D.W. and Jukes, T.H., Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics*, 1994. **20**(3): p. 386-96.
417. Yang, Z. and Yoder, A.D., Estimation of the transition/transversion rate bias and species sampling. *J Mol Evol*, 1999. **48**(3): p. 274-83.
418. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., et al., Patterns of somatic mutation in human cancer genomes. *Nature*, 2007. **446**(7132): p. 153-8.

419. Olsson, E., Winter, C., George, A., Chen, Y., Torngren, T., Bendahl, P.O., et al., Mutation Screening of 1,237 Cancer Genes across Six Model Cell Lines of Basal-Like Breast Cancer. *PLoS One*, 2015. **10**(12): p. e0144528.
420. Armbruster, D.A. and Pry, T., Limit of blank, limit of detection and limit of quantitation. *Clin Biochem Rev*, 2008. **29 Suppl 1**: p. S49-52.
421. Dunwell, T.L., Dailey, S.C., Ottestad, A.L., Yu, J., Becker, P.W., Scaife, S., et al., Adaptor Template Oligo-Mediated Sequencing (ATOM-Seq) is a new ultra-sensitive UMI-based NGS library preparation technology for use with cfDNA and cfRNA. *Sci Rep*, 2021. **11**(1): p. 3138.




# Paper I





# PTEN and NEDD4 in Human Breast Carcinoma

Yilun Chen<sup>1,2</sup> · Marc J. van de Vijver<sup>3</sup> · Hanina Hibshoosh<sup>4</sup> · Ramon Parsons<sup>5</sup> ·  
Lao H. Saal<sup>1,2,6</sup> 

Received: 14 March 2015 / Accepted: 4 August 2015 / Published online: 15 August 2015  
© The Author(s) 2015. This article is published with open access at Springerlink.com

**Abstract** PTEN is an important tumor suppressor gene that antagonizes the oncogenic PI3K/AKT signaling pathway and has functions in the nucleus for maintaining genome integrity. Although PTEN inactivation by mutation is infrequent in breast cancer, transcript and protein levels are deficient in >25 % of cases. The E3 ubiquitin ligase NEDD4 (also known as NEDD4-1) has been reported to negatively regulate PTEN protein levels through poly-ubiquitination and proteolysis in carcinomas of the prostate, lung, and bladder, but its effect on PTEN in the breast has not been studied extensively. To investigate whether NEDD4 contributes to low PTEN levels in human breast cancer, we analyzed the expression of these proteins by immunohistochemistry across a large Swedish cohort of breast tumor specimens, and their transcript expression levels by microarrays. For both NEDD4 and PTEN, their transcript expression was significantly correlated to their protein expression. However, comparing NEDD4 expression to PTEN expression, either no association or a positive correlation was observed at the protein and transcript levels. This unexpected observation

was further corroborated in two independent breast cancer cohorts from The Netherlands Cancer Institute and The Cancer Genome Atlas. Our results suggest that NEDD4 is not responsible for the frequent down-regulation of the PTEN protein in human breast carcinoma.

**Keywords** PTEN · NEDD4 · Breast carcinoma · IHC

## Introduction

PTEN is a phosphatase that plays an important role in tumor suppression by negatively regulating the oncogenic phosphatidylinositol 3-kinase (PI3K) pathway, as well as through functions in the nucleus that contribute to maintenance of genomic integrity [1]. Germline mutations of PTEN are found in patients with PTEN hamartoma tumor syndrome and are associated with an increased risk for breast, thyroid, and endometrial cancer [2–4]. Moreover, somatic loss-of-function mutations of *PTEN* are estimated to be present in 30 % of cancer and are found across the entire spectrum of tumor types [5–7]. The PTEN/PI3K pathway is one of the key pathways deregulated in breast cancer. *PIK3CA*, which encodes the p110- $\alpha$  catalytic subunit of PI3K, has activating mutations in one-third of breast tumors, and although mutation rate of *PTEN* is less than 5 % [8], PTEN expression is found to be greatly diminished in at least 25 % of breast tumors and in near mutual exclusivity to *PIK3CA* mutation [9, 10]. The mechanisms by which PTEN is down-regulated is poorly delineated in breast cancer, but mutations, copy number loss, rearrangements, epigenetic silencing, as well as post-translational regulation may contribute [9–13]. Of note, PTEN loss is frequent within the poor-prognosis basal-like molecular subtype of breast cancer [13].

✉ Lao H. Saal  
lao.saal@med.lu.se

<sup>1</sup> Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden

<sup>2</sup> Lund University Cancer Center, Lund, Sweden

<sup>3</sup> Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands

<sup>4</sup> Department of Pathology, Columbia University Medical Center, NY, USA

<sup>5</sup> Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, NY, USA

<sup>6</sup> CREATE Health Strategic Centre for Translational Cancer Research, Lund University, Lund, Sweden



Recently, Wang et al. reported that NEDD4 (neural precursor cell expressed, developmentally down-regulated 4, E3 ubiquitin protein ligase; also known as NEDD4-1) is an E3 ubiquitin ligase of PTEN and catalyzes poly-ubiquitination of PTEN in cells leading to proteolysis of the PTEN protein, thereby negatively regulating PTEN abundance [14]. Furthermore, in their analysis of mouse prostate and human bladder cancer samples, high expression of NEDD4 was inversely correlated to PTEN protein levels but not *PTEN* mRNA levels, suggesting that NEDD4 plays a proto-oncogenic role in tumorigenesis and cancer development via post-translational suppression of PTEN [14]. Negative regulation of PTEN by NEDD4-mediated poly-ubiquitination has since been reported to be involved in several biological and pathological processes, such as axon branching [15, 16], T-cell activation [17], keloid formation [18], and insulin-mediated glucose metabolism [19]. Inverse relationships between the expression of NEDD4 and PTEN have also been observed in human non-small cell lung carcinomas [20] and colon cancer [21].

However, the regulation of PTEN by NEDD4 may be microenvironment and/or cell-type specific. For example, Trotman et al. found that in addition to catalyzing poly-ubiquitination of PTEN, NEDD4 is also responsible for PTEN mono-ubiquitination that leads to PTEN nuclear import and protection from proteasomal degradation, making the role of NEDD4 in regulation of PTEN stability subtle and complex [22]. Moreover, some studies have called into question the interaction between NEDD4 and PTEN. For example, no discernible effect on Pten stability, subcellular localization, or downstream targets was observed in two separate *Nedd4* knock-out mouse models [23]. Furthermore, Maddika et al. failed to reproduce the functional interaction between NEDD4 and PTEN, and instead found that WWP2, another E3 ubiquitin ligase within the NEDD4-like protein family, mediated poly-ubiquitination of PTEN [24]. A third group has also failed to demonstrate that PTEN is a substrate of *Nedd4*, and rather found that PTEN regulated *Nedd4* by modulating mTORC1 activity [19]. Lastly, in gastric carcinoma, no relationship was observed between NEDD4 and PTEN expression [25], and in colorectal cancer cell lines and biopsies, NEDD4 modulation and expression level were not associated to the levels of PTEN [26].

NEDD4 and its potential role in PTEN regulation in breast cancer have not been studied. To reveal the pattern of expression of NEDD4 in human breast cancer, and to investigate whether NEDD4-mediated PTEN degradation is a factor that contributes to the frequent loss of PTEN protein, we analyzed NEDD4 and PTEN expression at the protein and mRNA levels in a large cohort of Swedish breast tumors, and verified our findings in two independent breast cancer cohorts from The Netherlands Cancer Institute (NKI) and The Cancer Genome Atlas (TCGA) (Table 1).

## Materials and Methods

### Breast Cancer Cohorts

Clinical and demographic information is provided for all cohorts in Table 1. For the Swedish cohort, 132 formalin-fixed paraffin-embedded (FFPE) tissue microarray (TMA) tumor specimens, arrayed in triplicates, were studied for NEDD4 protein expression by IHC, of which 123 had matched PTEN IHC scores previously evaluated [9, 27]. These 123 samples were analyzed for correlation between PTEN and NEDD4 protein levels. Correlation between the PTEN protein and *NEDD4* mRNA levels, and correlation between *PTEN* mRNA and *NEDD4* mRNA levels were analyzed in a subset of 105 samples with both PTEN IHC status and microarray gene expression data [27] (NCBI Gene Expression Omnibus accession GSE5325). Correlation between NEDD4 protein and *NEDD4* mRNA levels was performed in a subset of 42 samples with NEDD4 IHC and microarray data. For the NKI cohort, gene expression microarray data from 295 tumor samples was analyzed for correlation between gene expression levels of *PTEN* and *NEDD4* [28, 29]. Tissue microarrays containing these 295 NKI cases were stained for PTEN protein, of which 267 samples could be evaluated, and thereafter were analyzed for correlations between PTEN IHC scores and *PTEN* or *NEDD4* mRNA expression levels. For TCGA cohort, level 3 IlluminaHiSeq\_RNASeqV2 gene expression data for 970 primary breast tumor samples was used, as well as PTEN protein expression status for 407 cases derived from a reverse phase protein arrays platform. All TCGA data were downloaded from the TCGA data portal (<https://tcga-data.nci.nih.gov/tcga/>, downloaded on January 20, 2014). The study was approved by the Lund University Hospital ethics committee (LU240-01 and 2009/658), waiving the requirement for informed consent for the study, and all experimental protocols were performed in accordance with approved guidelines.

### Immunohistochemistry

The rabbit polyclonal anti-NEDD4 WW2 domain antibody #07–049 (EMD Millipore, Darmstadt, Germany), previously validated to be specific for NEDD4 [14], was used for IHC. The staining was done using an Autostainer Plus instrument and EnVision Plus system (Dako Denmark A/S, Glostrup, Denmark) following manufacturer's recommended protocol. Antigen retrieval was performed using Dako Targeted Retrieval Buffer pH 6.0 at 98 °C for 20 min, and the primary antibody was used at 1:500 dilution with 30 min incubation time at room temperature. The stained specimens were scanned using a MIRAX MIDI slide scanner (Carl Zeiss AG, Oberkochen, Germany) and viewed with Panoramic Viewer v1.15.3 (3DHISTECH, Budapest, Hungary). Semi-

**Table 1** Clinical demographics of the breast cancer patients

	Swedish Cohort						NKI cohort		TCGA cohort	
	With protein data				With mRNA data					
	<i>n</i> = 186 (%)		<i>n</i> = 123 (%)		<i>n</i> = 105 (%)		<i>n</i> = 295 (%)		<i>n</i> = 970 (%)	
Median age at diagnosis (y/o)	62	(range, 26–80)	64	(range, 31–80)	61	(range, 26–77)	44	(range, 26–53)	59	(range, 26–90)
Median tumor size (mm)	25	(range, 2–55)	25	(range, 10–55)	27	(range, 2–50)	20	(range, 2–50)	NA	(NA)
Estrogen receptor										
Positive	121	(65)	85	(69)	55	(52)	214	(73)	716	(74)
Negative	59	(32)	35	(28)	47	(45)	72	(24)	210	(22)
Unknown	6	(3)	3	(2)	3	(3)	9	(3)	44	(5)
Progesterone receptor										
Positive	78	(42)	55	(45)	35	(33)	185	(63)	622	(64)
Negative	98	(53)	64	(52)	62	(59)	101	(34)	301	(31)
Unknown	10	(5)	4	(3)	8	(8)	9	(3)	47	(5)
HER2										
Positive	27	(15)	16	(13)	18	(17)	56	(19)	148	(15)
Negative	113	(61)	84	(68)	55	(52)	217	(74)	496	(51)
Equivocal	NA	(NA)	NA	(NA)	NA	(NA)	NA	(NA)	156	(16)
Unknown	46	(25)	23	(19)	32	(30)	22	(7)	170	(18)
Nottingham histological grade										
1	3	(2)	1	(1)	3	(3)	60	(20)	NA	(NA)
2	47	(25)	15	(12)	37	(35)	99	(34)	NA	(NA)
3	37	(20)	14	(11)	28	(27)	136	(46)	NA	(NA)
Unknown	99	(53)	93	(75)	37	(35)	0	(0)	NA	(NA)
Lymph node										
Positive	118	(63)	79	(64)	65	(62)	144	(49)	411	(42)
Negative	68	(37)	44	(36)	40	(38)	151	(51)	397	(41)
Unknown	0	(0)	0	(0)	0	(0)	0	(0)	162	(17)

quantitative scoring was done according to the Dako system 0–3 scoring scale, where scores of 0 were given to tissues with no NEDD4 staining, 1+ to weak NEDD4 staining, 2+ to intermediate NEDD4 staining, and 3+ to strong NEDD4 staining (Fig. 1). The IHC scores of 0 and 1+ were then combined and categorized as NEDD4-negative, and scores of 2+ and 3+ were categorized as NEDD4-positive. PTEN IHC results for the Swedish cohort were reported previously [9, 27]. PTEN IHC was performed on the NKI TMAs using methods previously described [13].

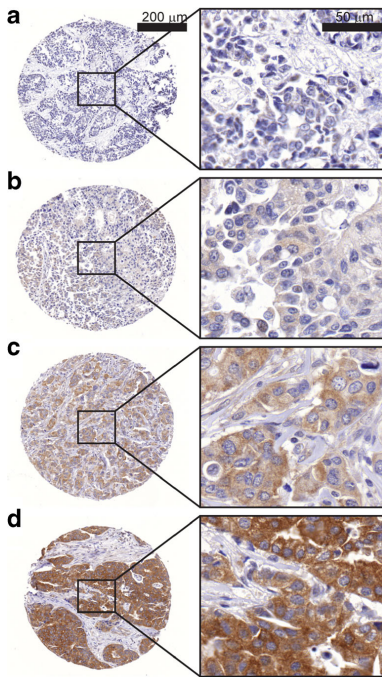
### Statistical Analysis

The chi-squared test was used to test the significance level of correlations between the NEDD4 protein and different breast cancer biomarkers. The Wilcoxon rank-sum test was used for correlation between PTEN and NEDD4 protein levels. The Student's t-test was used for correlations between the PTEN/NEDD4 protein and *PTEN/NEDD4* mRNA levels. The Pearson's correlation test was used for correlations between

*PTEN* and *NEDD4* mRNA levels from gene expression data and RNA-seq data. All tests were two-tailed, and  $P < 0.05$  was considered significant. All statistical analyses were performed with R version 3.1.0 (<http://www.r-project.org>).

### Results and Discussion

Immunohistochemical (IHC) staining was performed for 132 formalin-fixed paraffin-embedded (FFPE) breast tumor specimens (Swedish cohort) using an antibody previously reported to be specific to NEDD4 [14] (see Methods; Fig. 1). Consistent with previous studies in other tissues [14], NEDD4 protein was predominantly cytoplasmic in breast cancer cells (Fig. 1). Among the 132 stained samples, 60 (45 %) had zero or weak NEDD4 protein staining (classified as NEDD4-negative), whereas 72 (55 %) had intermediate to strong expression (NEDD4-positive). NEDD4 protein expression was positively correlated to estrogen receptor status (ER;  $P = 0.0017$ ), but not associated to the other clinical variables



**Fig. 1** NEDD4 immunohistochemistry. 132 breast tumor tissue microarray specimens were immunohistochemically stained with anti-NEDD4 antibody. Shown are representative examples of tumors with NEDD4 IHC scores of **a** 0, **b** 1+, **c** 2+, and **d** 3+. Scores 0/1+ were categorized NEDD4-negative, and 2+/3+ as NEDD4-positive. NEDD4 protein was expressed predominantly in the cytoplasm regardless of the staining intensity

**Table 2** Correlations of NEDD4 protein with biomarkers in the Swedish cohort

	NEDD4-	NEDD4+	N	$\chi^2$ <i>P</i>
Estrogen receptor				
Positive	33	57	129	0.0017
Negative	26	13		
Progesterone receptor				
Positive	21	36	128	0.12
Negative	36	35		
HER2				
Positive	11	6	108	0.12
Negative	40	51		
Nottingham histological grade				
1	0	1	32	0.4
2	9	7		
3	10	5		
Ki-67				
Positive	2	7	37	0.57
Negative	9	19		

positive within the NEDD4-positive group compared to 64 % being PTEN-positive in the NEDD4-negative group (Fig. 2b).

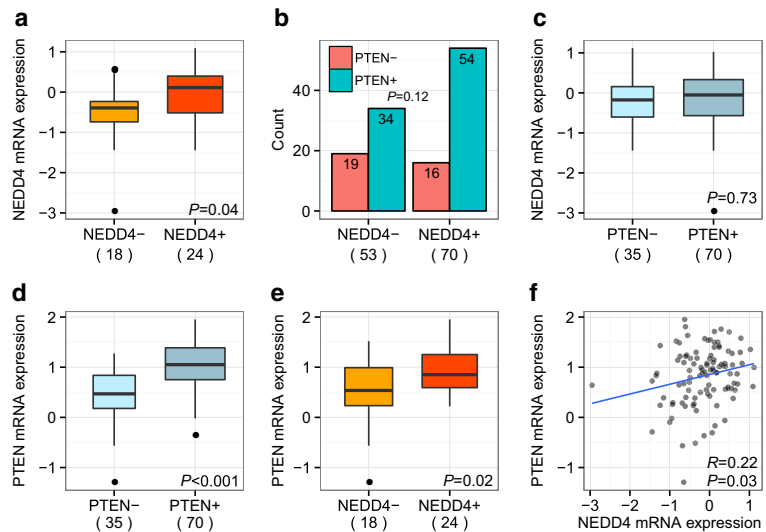
In human bladder carcinoma, Wang et al. reported *NEDD4* mRNA expression and *PTEN* mRNA expression to be uncorrelated, but that *NEDD4* mRNA levels were inversely correlated to PTEN protein levels [14]. To investigate if it is also the case in breast tumors, we next considered the transcript levels of these genes using the GSE5325 microarray dataset of 105 breast tumors previously utilized to develop a gene expression signature for PTEN-loss [27]. In contrast to bladder cancer, we found no correlation between *NEDD4* mRNA and PTEN protein expression ( $P = 0.73$ ; Fig. 2c). *PTEN* mRNA, however, was highly correlated to PTEN protein ( $P < 0.001$ ; Fig. 2d), which has been previously reported [27]. Unexpectedly, we found *PTEN* mRNA levels to be significantly positively correlated to NEDD4 protein expression ( $N = 42$ ,  $P = 0.02$ ; Fig. 2e) as well as to *NEDD4* mRNA levels ( $N = 105$ ,  $P = 0.03$ ; Fig. 2f).

To validate these findings, two independent large-scale breast cancer cohorts from the NKI and TCGA were studied. The NKI cohort contained 295 breast tumor samples with microarray gene expression data [28, 29]. Tissue microarray sections were obtained and immunostained for PTEN protein, of which 267 cases were evaluable. Similar to the Swedish cohort, we found no correlation between *NEDD4* mRNA and PTEN protein ( $P = 0.39$ ; Fig. 3a). The strong positive correlation between *PTEN* mRNA and PTEN protein ( $P < 0.001$ ; Fig. 3b), as well as the association of our previously published PTEN-loss signature [27] with loss of PTEN protein

progesterone receptor (PR;  $P = 0.12$ ), human epidermal growth factor receptor 2 (HER2;  $P = 0.12$ ), Nottingham Histologic Grade ( $P = 0.57$ ), and Ki-67 ( $P = 0.40$ ) (Table 2). Microarray gene expression data were available for 42 of the 132 cases from a previous study [27]. Using this data, we found NEDD4 protein levels to be significantly correlated to *NEDD4* mRNA expression level ( $P = 0.04$ ) (Fig. 2a), supporting the specificity of the antibody and also indicating that *NEDD4* mRNA may be an appropriate surrogate for NEDD4 protein levels in breast cancer.

PTEN protein expression was previously determined by IHC for 123 of the 132 cases [9]. We tested whether NEDD4 protein levels were negatively associated to PTEN protein levels, however no correlation was seen in this Swedish breast cancer material ( $P = 0.12$ ; Fig. 2b). This was inconsistent with the inverse correlation between the two proteins observed in a mouse prostate cancer model [14] and in lung cancers [20]. In fact, in our Swedish cohort the correlation trended positively, with 77 % of cases being PTEN-

**Fig. 2** PTEN and NEDD4 protein and mRNA levels in the Swedish cohort. **a** NEDD4 protein levels were significantly correlated to *NEDD4* mRNA levels ( $N = 42$ ,  $P = 0.04$ ). PTEN protein levels were not significantly correlated to **b** NEDD4 protein levels in breast cancer tissues ( $N = 123$ ,  $P = 0.12$ ) or **c** *NEDD4* mRNA levels ( $N = 105$ ,  $P = 0.73$ ). *PTEN* mRNA levels were significantly correlated to **d** PTEN protein levels ( $N = 105$ ,  $P < 0.001$ ), **e** NEDD4 protein levels ( $N = 42$ ,  $P = 0.02$ ), and **f** *NEDD4* mRNA levels ( $N = 105$ ,  $R = 0.22$ ,  $P = 0.03$ )

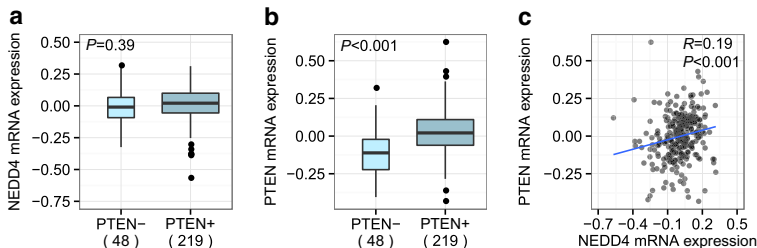


( $P = 0.003$ ; data not shown), were confirmed in this independent dataset. Moreover, the positive association between *NEDD4* mRNA and *PTEN* mRNA found in our Swedish cohort was also validated in the NKI patient material ( $N = 295$ ,  $P < 0.001$ ; Fig. 3c).

These associations were further corroborated in the TCGA breast carcinoma cohort containing RNA-sequencing (RNA-seq) gene expression profiles of primary breast tumors from 970 patients, of which 407 also had available PTEN protein expression data derived from reverse phase protein arrays [10]. In this large cohort the correlation between *NEDD4* mRNA and PTEN protein was also significantly positive ( $P < 0.001$ ; Fig. 4a). Additionally, *PTEN* mRNA and PTEN protein levels were positively correlated ( $P < 0.001$ ; Fig. 4b), as observed in the Swedish and NKI cohorts. Lastly, the

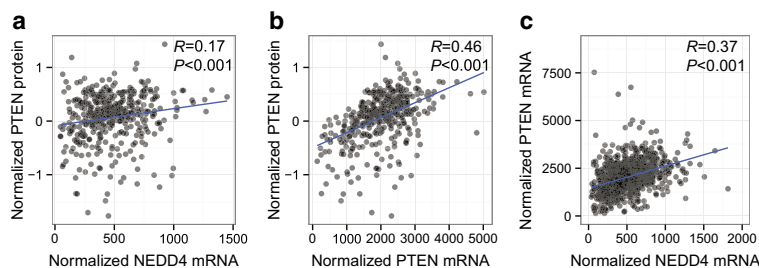
positive correlation between *NEDD4* mRNA and *PTEN* mRNA levels was also confirmed in the TCGA dataset ( $P < 0.001$ ; Fig. 4c).

In conclusion, our study investigated whether PTEN was associated to NEDD4 in three large independent breast cancer sample cohorts. Contrary to reports in some other cancer forms, no inverse relationship was seen between *NEDD4* transcript and PTEN protein levels. Rather, there was no correlation between NEDD4 protein and PTEN protein, and the correlation between *NEDD4* mRNA/protein and *PTEN* mRNA was significantly positive. NEDD4-mediated poly-ubiquitination of PTEN may be an important mechanism that contributes to PTEN protein loss in bladder cancer [14] and non-small cell lung carcinoma [20]; whereas the results in gastric and colorectal cancers have been discrepant [25, 26].



**Fig. 3** PTEN mRNA/protein levels and *NEDD4* mRNA levels in the NKI cohort. PTEN IHC scores were not associated to **a** *NEDD4* mRNA levels ( $N = 267$ ,  $P = 0.39$ ), but were significantly correlated to

**b** *PTEN* mRNA levels ( $N = 267$ ,  $P < 0.001$ ). **c** *PTEN* mRNA and *NEDD4* mRNA levels were also significantly correlated ( $N = 295$ ,  $R = 0.19$ ,  $P < 0.001$ )



**Fig. 4** *PTEN* and *NEDD4* mRNA levels in the TCGA cohort. *PTEN* protein levels were significantly correlated to **a** *NEDD4* mRNA levels ( $N = 407$ ,  $R = 0.17$ ,  $P < 0.001$ ), and **b** *PTEN* mRNA levels ( $N = 407$ ,

$R = 0.46$ ,  $P < 0.001$ ). **c** *PTEN* mRNA levels were significantly correlated to *NEDD4* mRNA levels in the 970 primary breast tumors ( $R = 0.37$ ,  $P < 0.001$ )

Interestingly, in ovarian cancer HeLa cells, *PTEN* has also been reported to negatively regulate *NEDD4* expression via the PI3K/AKT pathway, forming a potential negative feedback loop [30]. Our present study does not support *NEDD4* as a major negative regulator of *PTEN* levels in human breast cancer. Additional studies are necessary to better delineate the underlying mechanisms of *PTEN* loss in this poor-prognosis subgroup.

**Acknowledgments** We thank Kristina Lövgren for laboratory assistance, Björn Frostner and Susanne André for administrative support, and members of the Translational Oncogenomics Unit, Division of Oncology and Pathology, for valuable discussion. This study was funded in part by the Swedish Research Council, Swedish Cancer Society, Governmental Funding of Clinical Research within National Health Service, Crafoord Foundation, Mrs. Berta Kamprad Foundation, Lund University Medical Faculty, Gunnar Nilsson Cancer Foundation, Skåne University Hospital Foundation, BioCARE Research Program, King Gustav Vth Jubilee Foundation, and the Krappert Foundation.

**Author contributions** Y.C. and L.H.S. conceived the study and performed the experiments. M.J.vdV. and R.P. provided reagents. Y.C., H.H., and L.H.S. analyzed the data. Y.C. and L.H.S. wrote and revised the manuscript. L.H.S. supervised the project.

**Competing interests** The authors declare no competing financial interests.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Machama T, Dixon JE (1998) The tumor suppressor, *PTEN/MMAC1*, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. *J Biolumin Chemilumin* 273:13375–13378
2. Arch EM, Goodman BK, Van Wesep RA, Liaw D, Clarke K, Parsons R, et al (1997) Deletion of *PTEN* in a patient with bannayan-riley-ruvalcaba syndrome suggests allelism with cowden disease. *Am J Med Genet* 71:489–493
3. Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, Zheng Z, et al (1997) Germline mutations of the *PTEN* gene in cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet* 16:64–67
4. Hobert JA, Eng C (2009) *PTEN* hamartoma tumor syndrome: an overview. *Genitourin Med* 11:687–694
5. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, et al (1997) *PTEN*, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* 275:1943–1947
6. Steck PA, Pershouse MA, Jasser SA, Yung WK, Lin H, Ligon AH, et al (1997) Identification of a candidate tumour suppressor gene, *MMAC1*, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nat Genet* 15:356–362
7. Shaw RJ, Cantley LC (2006) Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature* 441:424–430
8. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43:D805–811
9. Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, et al (2005) PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with *PTEN* loss in human breast carcinoma. *Cancer Res* 65:2554–2559
10. The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70
11. Simpson L, Parsons R (2001) *PTEN*: life as a tumor suppressor. *Exp Cell Res* 264:29–41
12. Wang X, Jiang X (2008) Post-translational regulation of *PTEN*. *Oncogene* 27:5454–5463
13. Saal LH, Gruvberger-Saal SK, Persson C, Lovgren K, Junttilanen M, Staaf J, et al (2008) Recurrent gross mutations of the *PTEN* tumor suppressor gene in breast cancers with deficient DSB repair. *Nat Genet* 40:102–107
14. Wang X, Trotman LC, Koppie T, Alimonti A, Chen Z, Gao Z, et al (2007) *NEDD4-1* is a proto-oncogenic ubiquitin ligase for *PTEN*. *Cell* 128:129–139
15. Drinjakovic J, Jung H, Campbell DS, Strohlich L, Dwivedy A, Holt CE (2010) E3 ligase *Nedd4* promotes axon branching by downregulating *PTEN*. *Neuron* 65:341–357
16. Goh CP, Low LH, Putz U, Gunnarsen J, Hammond V, Howitt J, et al (2013) *Ndfip1* expression in developing neurons indicates a role for protein ubiquitination by *Nedd4* E3 ligases during cortical development. *Neurosci Lett* 555:225–230

17. Guo H, Qiao G, Ying H, Li Z, Zhao Y, Liang Y, et al (2012) E3 ubiquitin ligase Cbl-b regulates pten via Nedd4 in T cells independently of its ubiquitin ligase activity. *Cell Rep* 1:472–482
18. Chung S, Nakashima M, Zembutsu H, Nakamura Y (2011) Possible involvement of NEDD4 in keloid formation; its critical role in fibroblast proliferation and collagen production. *Proc Jpn Acad Ser B Phys Biol Sci* 87:563–573
19. Shi Y, Wang J, Chandrapaty S, Cross J, Thompson C, Rosen N, et al (2014) PTEN is a protein tyrosine phosphatase for IRS1. *Nat Struct Mol Biol* 21:522–527
20. Amodio N, Scrima M, Palaia L, Salman AN, Quintiero A, Franco R, et al (2010) Oncogenic role of the E3 ubiquitin ligase NEDD4-1, a PTEN negative regulator, in non-small-cell lung carcinomas. *Am J Pathol* 177:2622–2634
21. Hong SW, Moon JH, Kim JS, Shin JS, Jung KA, Lee WK, et al (2014) p34 is a novel regulator of the oncogenic behavior of NEDD4-1 and PTEN. *Cell Death Differ* 21:146–160
22. Trotman LC, Wang X, Alimonti A, Chen Z, Teruya-Feldstein J, Yang H, et al (2007) Ubiquitination regulates PTEN nuclear import and tumor suppression. *Cell* 128:141–156
23. Fouladkou F, Landry T, Kawabe H, Neeb A, Lu C, Brose N, et al (2008) The ubiquitin ligase Nedd4-1 is dispensable for the regulation of PTEN stability and localization. *Proc Natl Acad Sci U S A* 105:8585–8590
24. Maddika S, Kavela S, Rani N, Palicharla VR, Pokorny JL, Sarkaria JN, et al (2011) WWP2 is an E3 ubiquitin ligase for PTEN. *Nat Cell Biol* 13:728–733
25. Yang Z, Yuan XG, Chen J, Lu NH (2012) Is NEDD4-1 a negative regulator of phosphatase and tensin homolog in gastric carcinogenesis? *World J Gastroenterol* 18:6345–6348
26. Eide PW, Cekaite L, Danielsen SA, Eilertsen IA, Kjenseth A, Fykerud TA, et al (2013) NEDD4 is overexpressed in colorectal cancer and promotes colonic cell growth independently of the PI3K/PTEN/AKT pathway. *Cell Signal* 25:12–18
27. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, et al (2007) Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc Natl Acad Sci U S A* 104:7564–7569
28. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999–2009
29. van't Veer, LJ, Dai, H, van de Vijver, MJ, He, YD, Hart, AA, Mao, M et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–536
30. Ahn Y, Hwang CY, Lee SR, Kwon KS, Lee C (2008) The tumour suppressor PTEN mediates a negative regulation of the E3 ubiquitin-protein ligase Nedd4. *Biochem J* 412:331–338



## Paper II









# Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease

Eleonor Olsson<sup>1,2,†</sup>, Christof Winter<sup>1,2,†</sup>, Anthony George<sup>1,2</sup>, Yilun Chen<sup>1,2</sup>, Jillian Howlin<sup>1,2</sup>, Man-Hung Eric Tang<sup>1,2</sup>, Malin Dahlgren<sup>1,2</sup>, Ralph Schulz<sup>1,2,3</sup>, Dorthe Grabau<sup>4</sup>, Danielle van Westen<sup>5</sup>, Mårten Fernö<sup>1,2</sup>, Christian Ingvar<sup>6</sup>, Carsten Rose<sup>2,7,8</sup>, Pär-Ola Bendahl<sup>1,2</sup>, Lisa Rydén<sup>2,6</sup>, Åke Borg<sup>1,2,3,8</sup>, Sofia K Gruvberger-Saal<sup>1,2</sup>, Helena Jernström<sup>1,2</sup> & Lao H Saa<sup>1,2,8,\*</sup>

## Abstract

Metastatic breast cancer is usually diagnosed after becoming symptomatic, at which point it is rarely curable. Cell-free circulating tumor DNA (ctDNA) contains tumor-specific chromosomal rearrangements that may be interrogated in blood plasma. We evaluated serial monitoring of ctDNA for earlier detection of metastasis in a retrospective study of 20 patients diagnosed with primary breast cancer and long follow-up. Using an approach combining low-coverage whole-genome sequencing of primary tumors and quantification of tumor-specific rearrangements in plasma by droplet digital PCR, we identify for the first time that ctDNA monitoring is highly accurate for postsurgical discrimination between patients with (93%) and without (100%) eventual clinically detected recurrence. ctDNA-based detection preceded clinical detection of metastasis in 86% of patients with an average lead time of 11 months (range 0–37 months), whereas patients with long-term disease-free survival had undetectable ctDNA postoperatively. ctDNA quantity was predictive of poor survival. These findings establish the rationale for larger validation studies in early breast cancer to evaluate ctDNA as a monitoring tool for early metastasis detection, therapy modification, and to aid in avoidance of overtreatment.

**Keywords** breast carcinoma; circulating tumor DNA; early detection; liquid biopsy; metastasis

**Subject Categories** Biomarkers & Diagnostic Imaging; Cancer

DOI 10.15252/emmm.201404913 | Received 1 December 2014 | Revised 14

April 2015 | Accepted 16 April 2015 | Published online 18 May 2015

EMBO Mol Med (2015) 7: 1034–1047

See also: **TM af Hällström et al** (August 2015)

## Introduction

Breast cancer is the most common malignancy and leading cause of cancer-related death in women worldwide; once the tumor has metastasized, it is essentially an incurable disease (Jemal *et al*, 2011). The difficulty in curing metastatic breast cancer may be in part because metastatic spread is usually detected only after the deposit has grown large enough to be palpable, cause overt clinical symptoms, or be identified by imaging. In patients with primary (non-metastatic) breast cancer at diagnosis, the risk of subsequent metastatic relapse is greatest within 2 years after primary surgery (Cheng *et al*, 2012). However, an estimated 50% of recurrences are diagnosed > 5 years after surgery (Early Breast Cancer Trialists' Collaborative Group, 2005), indicating that occult metastatic dissemination can have a protracted subclinical period. Earlier detection of metastatic breast cancer may be clinically beneficial. A reasonable assumption is that identification of recurrent disease at the earliest moment will allow for initiation of auxiliary therapies against a nominal tumor burden that has accumulated fewer oncogenic events. So far, this assumption has been tested without success, most likely because modalities and biomarkers that lack sufficient sensitivity and/or specificity have been utilized thus far (Lippman & Osborne, 2013). For example, whereas circulating tumor cells (CTCs) may carry additional prognostic information in primary breast cancer (Lucci *et al*, 2012; Rack *et al*, 2014), available evidence does not support the use of imaging, serum protein markers, and CTCs for routine monitoring after primary surgery (Khatcheressian *et al*, 2013; Theriault *et al*, 2013). At the same time, many breast cancer patients are likely being overtreated; that is, they may in fact be cured by locoregional treatment and unnecessarily enduring the side effects of systemic therapies. For

1 Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden

2 Lund University Cancer Center, Lund, Sweden

3 SCIBLU Genomics, Department of Clinical Sciences, Lund University, Lund, Sweden

4 Department of Pathology, Skåne University Hospital, Lund, Sweden

5 Department of Radiology, Skåne University Hospital, Lund, Sweden

6 Department of Surgery, Lund University and Skåne University Hospital, Lund, Sweden

7 Department of Immunotechnology, Lund University, Lund, Sweden

8 CREATE Health Strategic Centre for Translational Cancer Research, Lund University, Lund, Sweden

\*Corresponding author. Tel: +46 46 2220365; Fax: +46 46 147327; E-mail: lao.saa@med.lu.se; Twitter: @LaoSaa

<sup>†</sup>These authors contributed equally to this article and are listed alphabetically

these reasons, improved surveillance methods to determine occult tumor burden (or lack thereof) in the primary breast cancer setting are still highly desirable (Lippman & Osborne, 2013).

Clinical monitoring of minimal residual disease is routinely performed in several hematological malignancies with known pathognomonic chromosomal rearrangements, for example by serial quantification of TEL-AML1 or BCR-ABL fusion-gene chromosomal translocations in acute lymphoblastic leukemia and chronic myelogenous leukemia, respectively (Dolken, 2001). In cancer patients, tumor-derived DNA (termed cell-free circulating tumor DNA; ctDNA) can be found in the blood circulation and usually comprises a small fraction of the total circulating DNA (Jung *et al*, 2010). Circulating DNA is rapidly degraded into short fragments, and the quantity of ctDNA appears to be related to tumor progression (Stroun *et al*, 1989; Diehl *et al*, 2008; Yung *et al*, 2009; Jung *et al*, 2010; Leary *et al*, 2010; McBride *et al*, 2010; Diaz *et al*, 2012; Dawson *et al*, 2013; Murtaza *et al*, 2013; Bettegowda *et al*, 2014; Newman *et al*, 2014). Therefore, ctDNA “liquid biopsy” analysis is an attractive biomarker for noninvasive monitoring of tumor growth, response, and spread (McDermott *et al*, 2011). Until recently, assays for ctDNA have been infeasible for most solid cancers due to a paucity of recurrent mutations for interrogation as well as the practical and economical hurdles of enumerating tumor-specific aberrations on a per-patient basis.

Advances in deep-sequencing technology now enable comprehensive cataloguing of tumor-specific (somatic) chromosomal rearrangements and mutations at an ever-decreasing cost (Meyerson *et al*, 2010). Recent studies have shown that breast cancer genomes may harbor from a few to several hundred rearrangements and mutations per tumor (Shah *et al*, 2009; Stephens *et al*, 2009, 2012; Banerji *et al*, 2012; Cancer Genome Atlas Network, 2012; Ellis *et al*, 2012; Nik-Zainal *et al*, 2012). In contrast to somatic point mutations, in which the identical mutation can be present across many tumors, tumor types, and individuals (for example *PIK3CA* hot-spot mutations), chromosomal rearrangements are inherently highly tumor specific and can serve as unique genetic “fingerprints” of an individual tumor (Leary *et al*, 2012). Serial measurement of ctDNA using various methods has shown encouraging results for several solid cancer types (Diehl *et al*, 2008; Yung *et al*, 2009; Leary *et al*, 2010; McBride *et al*, 2010; Diaz *et al*, 2012; Misale *et al*, 2012; Newman *et al*, 2014), and in the metastatic breast cancer setting, measurement of ctDNA dynamics compares favorably to the serum protein marker CA 15-3 and CTCs (Dawson *et al*, 2013).

Here, we tested in patients with primary breast cancer and long-term follow-up the hypothesis that monitoring of tumor-specific chromosomal rearrangements in cell-free circulating DNA can detect occult metastatic disease following primary surgery and serve as a sensitive, specific, and thus potentially clinically useful noninvasive biomarker in the adjuvant setting (Fig 1).

## Results

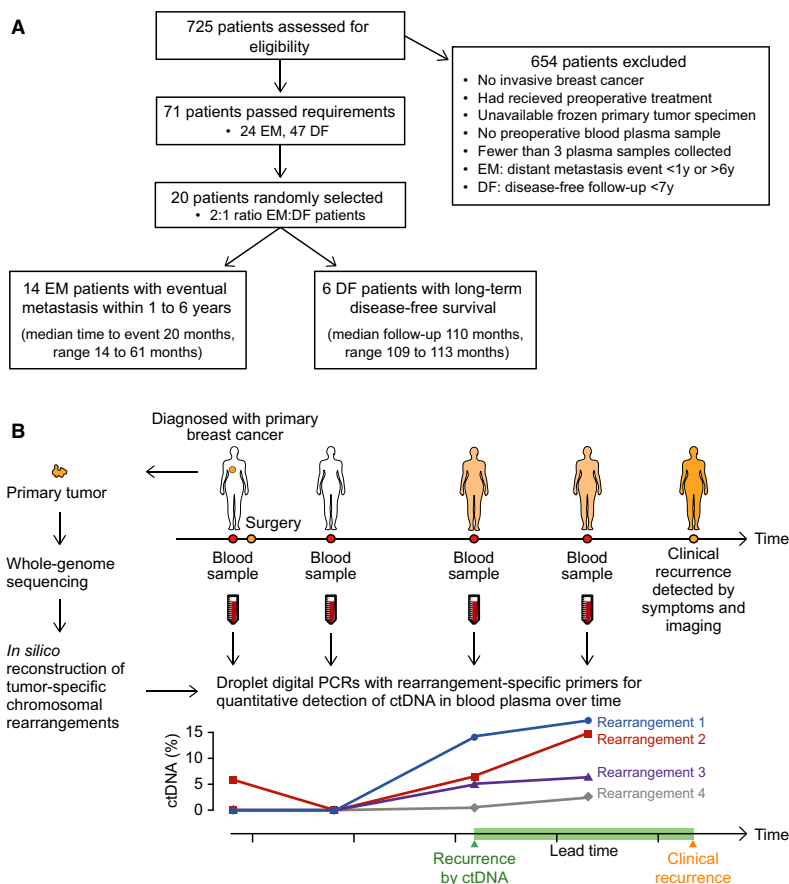
### Enumeration of tumor-specific chromosomal rearrangements

Twenty patients enrolled in the Breast Cancer and Blood Study (BC Blood, Sweden) (Borgquist *et al*, 2013), an ongoing

prospective study at Lund University since 2002, were included in the present investigation for retrospective analysis of ctDNA (Fig 1A). Six patients had long-term disease-free survival (9.2 years median follow-up; termed DF patients), and 14 had eventual diagnosis of clinical metastasis from 1.2 to 5.1 years after primary surgery (termed eventual metastatic [EM] patients) (Table 1). For each patient, a sample of the primary tumor, a normal tissue sample, and 3–6 blood plasma samples that were collected during the clinical course were available. First, to identify tumor-associated chromosomal rearrangements that could serve as biomarkers, whole-genome sequencing (WGS) was performed on DNA isolated from 21 primary breast tumors (patient EM6 had bilateral primary breast cancers). On average, 93 million DNA fragments were sequenced per tumor (range 54–160 million), yielding a mean genome sequence coverage of 5.3-fold (range 1.8–12.9) and mean physical coverage of 15.6 (range 9.2–28.2) (Supplementary Table S1). We developed an analysis pipeline incorporating our SplitSeq computational method to identify inter- and intra-chromosomal rearrangements using an approach that scanned for paired sequence reads where the two reads aligned to discordant positions in the human genome, or individual reads in a read pair that contained juxtaposed sequences from two disparate genomic regions. Chromosomal rearrangements supported by two or more sequenced fragments could be detected in all primary tumors, and on average, 92 rearrangements were identified per tumor (range 21–305) (Fig 2, Supplementary Fig S1 and Supplementary Tables S1 and S2). There was no significant difference in sequence coverage or frequencies of chromosomal rearrangements detected between EM patients and DF patients (Mann-Whitney test), and the numbers of detected rearrangements for these 21 cases are similar to other studies of primary breast tumors (Stephens *et al*, 2009; Banerji *et al*, 2012; Nik-Zainal *et al*, 2012).

### Selection and validation of rearrangements

To account for possible intra-tumoral heterogeneity, and since it is not possible to know *a priori* which rearrangements in the primary tumor will be part of derivative metastatic clone(s), candidate rearrangements were selected such that a range of apparent copy number states (in other words, a range of number of supporting reads) were represented for each patient tumor. Our strategy was to design assays for ~10 rearrangements per primary tumor and select additional rearrangements in the event of assay failure or validation as not somatic. In summary, for each of the 237 selected candidate rearrangements, one assay was designed and tested by conventional PCR across the breakpoint junction in tumor and normal DNA from the same patient. Of 197 informative assays (83%; 7–17 per tumor), 167 (85%) were confirmed to be somatic by PCR (Supplementary Tables S3 and S4). Of these, due to limitations on the available plasma volumes and our desire to perform replicate analyses, four to six rearrangements per tumor were selected (again to reflect a variety of copy number states) and the corresponding probe was synthesized for droplet digital PCR (ddPCR) analysis of patient plasma samples. Probe assay success rate was high, with 113 of 122 (93%) validating for ddPCR (Supplementary Table S4).



**Figure 1. Analysis of personalized ctDNA biomarkers in primary breast cancer.**

**A** Patient flow diagram indicating patient selection criteria. EM = eventual metastasis; DF = long-term disease-free.

**B** Study schema. For 20 women with primary breast cancer, patient- and tumor-specific chromosomal rearrangements were determined through whole-genome sequencing of 21 tumor tissue specimens (one patient had bilateral tumors). Genomic fusion sequences were bioinformatically reconstructed, and selected rearrangements were validated as somatic. Personalized droplet digital PCR assays were used to quantify rearranged DNA sequences in the cell-free circulating DNA isolated from 93 patient blood plasma samples taken serially during the clinical course. ctDNA results were then compared to clinical endpoints.

### Optimization of droplet digital PCR

In ddPCR, the PCR with input DNA and target sequence-specific fluorescent probe and primers is partitioned into thousands of nanoliter-sized reaction droplets. Following thermocycling, successful amplification of the target cleaves the fluorescent molecule from the specific probe, thereby unquenching the fluorophore (Fig 2C). Each droplet is read as either containing amplifiable target sequence (positive fluorescence above a threshold) or not, yielding a binary (digital) readout. Because the distribution of zero, one, two, or more amplifiable targets into droplets is a random process, the fraction of

positive droplets to total droplets can be Poisson-corrected to derive a highly quantitative estimate of the number of amplifiable molecules that were present in the input sample (Hindson *et al*, 2011). We optimized a ddPCR method for measurement of circulating DNA that employs a universal touchdown PCR thermocycling protocol for increased specificity. For quantification of tumor-specific rearrangements, we determined our ddPCR method to be highly linear over at least 3 orders of magnitude and able to discriminate somatic mutant rearranged sequences down to 0.01% tumor DNA content (one rearranged sequence per 10,000 wild-type sequences) (Fig 3A and B). Importantly, zero tumor-specific rearrangements were

**Table 1. Patient and tumor characteristics.**

Patient ID	Age at primary diagnosis (years)	Tumor size (mm)	Lymph node status (positive/total)	Distant metastasis at diagnosis	ER status	PR status	HER2 status	Nottingham Histological Grade	Time to recurrence (months)	Time to last follow-up or death <sup>b</sup> (months)
EM1	42	33	0/2	No	Positive	Positive	Negative	3	20.0	29.4 <sup>b</sup>
EM2	57	28	1/17	No	Positive	Positive	Negative	2	40.0	55.2 <sup>b</sup>
EM3	78	20	6/17	No	Positive	Positive	Negative	3	16.1	17.9 <sup>b</sup>
EM4	34	28	0/2	No	Positive	Positive	Negative	2	31.8	99.0
EM5	61	12	0/5	No	Positive	Positive	Negative	1	48.8	97.1 <sup>b</sup>
EM6	62	Right: 28	2/13	No	Positive	Positive	Negative	2	61.3	87.3
		Left: 55	1/12		Positive	Positive	Negative	3		
EM7	55	22	0/2	No	Positive	Negative	Amplified	2	18.9	59.3 <sup>b</sup>
EM8	67	22	2/12	No	Positive	Negative	Negative	2	13.9	33.2 <sup>b</sup>
EM9	50	18	0/1	No	Positive	Positive	Negative	3	36.0	54.7 <sup>b</sup>
EM10	64	45	1/14	No	Positive	Negative	Negative <sup>a</sup>	2	17.7	33.2 <sup>b</sup>
EM11	59	20	16/18	No	Positive	Positive	Negative <sup>a</sup>	3	13.9	32.5 <sup>b</sup>
EM12	53	37	4/10	No	Positive	Positive	Negative	2	16.2	33.5 <sup>b</sup>
EM13	69	25	0/4	No	Positive	Negative	Negative <sup>a</sup>	2	43.9	58.7 <sup>b</sup>
EM14	47	19	1/18	No	Negative	Positive	Amplified	2	20.0	46.7 <sup>b</sup>
DF1	58	15	0/2	No	Positive	Positive	Negative	3		108.7
DF2	37	20	0/3	No	Positive	Positive	Negative	3		111.7
DF3	56	19	0/1	No	Positive	Positive	Negative <sup>a</sup>	2		109.9
DF4	46	13	0/1	No	Negative	Negative	Negative <sup>a</sup>	2		110.4
DF5	54	15	0/2	No	Positive	Positive	Negative <sup>a</sup>	3		109.5
DF6	58	18	0/2	No	Positive	Negative	Negative <sup>a</sup>	2		113.2

All patients analyzed are women.

<sup>a</sup>Clinical HER2 analysis not performed. HER2 status determined from gene copy number derived from whole-genome sequencing results.

<sup>b</sup>Time from primary diagnosis to death.

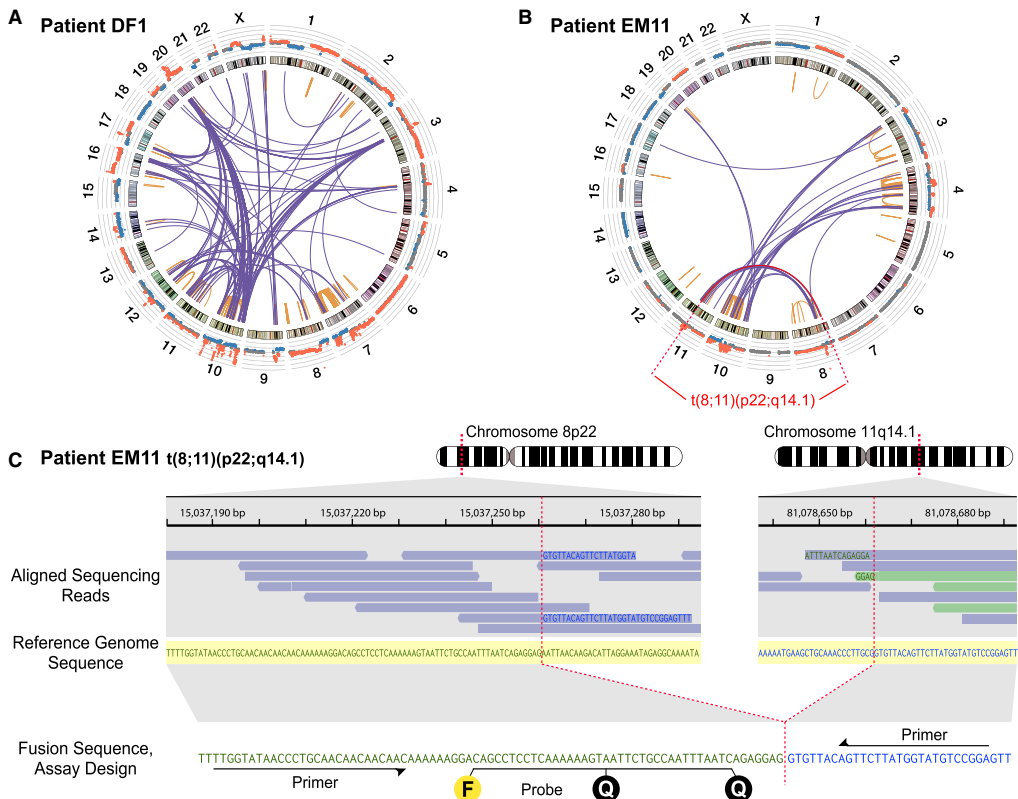
detected by our method in over 2.7 million negative control DNA droplets analysed, corresponding to > 200 control ddPCR reactions that in total interrogated more than 2.5 million normal haploid genome equivalents (i.e. zero rate of false-positive signals).

#### Quantification of ctDNA in serial plasma samples

Circulating cell-free DNA was isolated from 93 plasma samples for the 20 patients. The number of fragments of each tumor-specific chromosomal rearrangement was quantified in the circulating DNA by ddPCR. Each tumor-specific rearrangement assay was run in duplicate and included positive (primary tumor DNA) and negative (matched normal DNA) controls, and on average, 25,704 (SD 2,320) droplets were analyzed per assay per plasma sample. As expected, the relative copy numbers of rearrangements were well correlated between the WGS analysis and ddPCR analysis of primary tumor DNA ( $R^2 = 0.65$ ; Fig 3C). A ddPCR assay targeting a non-rearranged normal region of chromosome 2p14, which rarely undergoes copy number alteration in breast cancer (Jonsson *et al*, 2010), was used to estimate total circulating DNA (both tumor and normal cell derived). The average number of amplifiable 2p14 control region fragments was 1,908 copies/ml plasma (range 280–8,960)

(Supplementary Fig S2 and Supplementary Table S5). There was no significant difference in the number of 2p14 control region fragments per ml plasma between EM and DF patients within the pre-operative time-points nor when comparing across all time-points (Mann–Whitney *U*-test). Tumor-specific rearrangements were detected in 29 plasma samples corresponding to 13 EM patients, and the fractional quantity was calculated as the measured rearrangement divided by the measured 2p14 control region. In these 29 samples, ctDNA levels (taking the maximal value if more than one rearrangement was detected in a sample) ranged from 1.4 to 72.4% (mean 19.3%), and the concentration of rearranged fragments ranged from 38 to 2,617 fragments/ml plasma (mean 552 fragments/ml plasma) (Supplementary Table S5). The lowest ctDNA level detected in our patient material was 0.45%.

Among the 14 EM patients with known eventual clinical recurrence, 13 patients had positive ctDNA levels for one or more follow-up plasma time-points and only patient EM3 had undetectable ctDNA (Fig 4A–C and Supplementary Fig S2). Conversely, none of the patients with long-term disease-free survival had detectable ctDNA at any time-point after surgery (Fig 4E and F and Supplementary Fig S2). Thus, our noninvasive blood test for metastasis during follow-up had a sensitivity of 93% and specificity of 100% (95%



**Figure 2. Identification of chromosomal rearrangements and personalized assay design.**

**A** Low-coverage whole-genome sequencing of the primary tumor was used to enumerate chromosomal rearrangements. Shown are results for patient DF1, with inter- and intra-chromosomal rearrangements plotted as a Circos diagram (Krzywinski *et al*, 2009). Chromosomes 1–22 and X are ordered in the outer circle. From the outside, concentrically, are plotted the DNA copy number estimations from the whole-genome sequencing data and the chromosome ideograms. The orange intra-chromosomal and blue inter-chromosomal arcs in the center indicate chromosomal rearrangements supported by two or more paired-end reads.

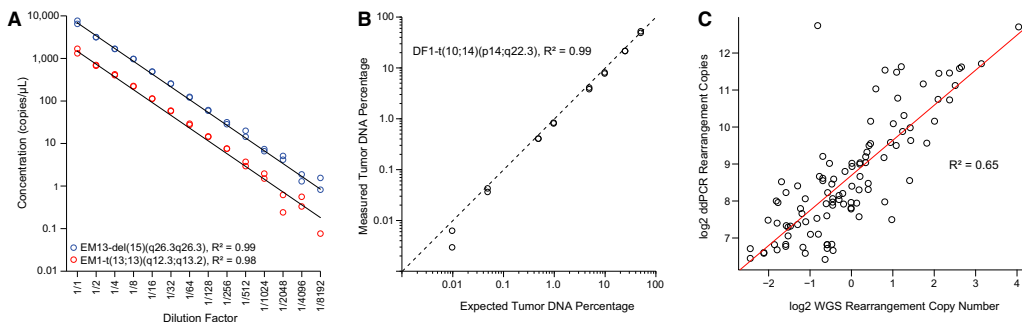
**B** Circos diagram for patient EM11. Plots for all patient tumors are shown in Supplementary Fig S1.

**C** One example rearrangement from patient EM11, indicated in red in (B), with identification of the exact fusion sequence between chromosomes 8p22 and 11q14.1. Aligned sequencing reads are highlighted in blue when its read pair aligns concordantly on the same chromosome or in light green if its read pair aligns on another chromosome. Within each sequencing read, nucleotide bases with exact match to the reference sequence (shown in the middle with yellow shading) are not printed. Mismatching bases are shown in blue if matching to 11q14.1 and green if matching to 8p22. At the bottom, the personalized dual-labeled probe and primers designed for this validated rearrangement are illustrated. F denotes the fluorescent molecule and Q the two quenching molecules.

confidence intervals [CI] 66–100% and 61–100%, respectively) for discrimination of EM versus DF status. Of note, ctDNA was detected in the presurgical plasma sample for four of 20 patients (20%; EM2, EM8, EM12, EM14); all four of these patients had eventual recurrent disease.

For each ddPCR assay, a uniform threshold of 0.5, i.e. at 50% of the normalized range of intensity values between the positive and negative control droplets, was used. Because the discriminatory accuracy of our ctDNA test could be influenced by the fluorescent intensity threshold used in dichotomizing a

ddPCR droplet as positive or negative, we performed receiver operating characteristic (ROC) curve analysis wherein the intensity threshold was varied incrementally (see Supplementary Methods; Supplementary Fig S3). This analysis indicated our test to have a high accuracy for postoperative discrimination of EM versus DF patients with an area under the curve of 0.98 (95% CI 0.75–1.00;  $P = 0.001$ , Mann–Whitney  $U$ -test) and an equivalent performance across a wide range of fluorescence intensity thresholds, from 0.35 to 0.95 (Fig 5A). Thus, the ddPCR signals were robust and distinct.



**Figure 3. Performance of droplet digital PCR (ddPCR) method.**

- A** Dilution series for two tumor-specific rearrangements, patient EM13-del(15)(q26.3q26.3) and patient EM1-t(13;13)(q12.3q13.2), starting with input of 20 ng of the respective patient's primary tumor DNA in each ddPCR, and diluting twofold in the series as indicated (x-axis). Experiments were performed in duplicate. Linear regression lines are plotted in black, and goodness of fit statistics ( $R^2$ ) were calculated.
- B** Observed percentages by ddPCR of a tumor-specific chromosomal rearrangement, patient DF1-t(10;14)(p14;q22.3), in admixtures of tumor and normal DNA of varying amounts from 50% down to 0.01% tumor DNA content (total DNA input fixed at 200 ng). Concentrations of the tumor-specific rearrangement and the control region in chromosome 2p14 were used in the calculations for amounts of tumor and total DNA, respectively. The black diagonal dashed line indicates the ideal correlation line ( $y = x$ ). The  $R^2$  was calculated for the linear regression line (not plotted). All axes are on log scales.
- C** Correlation between whole-genome sequencing (WGS) rearrangement copy number estimates and the number of copies in 40 ng primary tumor DNA as measured by ddPCR. Axes on  $\log_2$  scales. The  $R^2$  was calculated for the linear regression line (drawn in red).

### ctDNA and clinical course

Circulating tumor DNA showed dynamic changes across serial plasma samples for 13 of 20 patients (65%) (Supplementary Fig S2 and Supplementary Table S5). Three representative examples of changes in ctDNA levels during the clinical course are highlighted below. Due to the favorable clinicopathological features of patient EM5 (12 mm primary invasive ductal carcinoma, wide margins, no positive lymph nodes, histological grade 1, estrogen and progesterone receptor positivity, and HER2 negativity), she received postoperative radiotherapy and no systemic adjuvant therapy. Clinical metastasis was detected at 49 months after primary surgery; however, our ctDNA-based method detected molecular recurrence at 13 months, providing a potential earlier diagnosis of metastatic cancer by 3 years (Fig 4A; Supplementary Table S6). Analysis of ctDNA identified one tumor-specific rearrangement between chromosomes 10q and 13q in the plasma sample from 12-month follow-up; this and a 13q-16q rearrangement were detected at 24 months, and 10q-13q at 3 years. Two additional rearrangements (5q-22q and 10q-16q) were not detected at any time-point, indicating that these may not have been present in the cancer clone(s) that seeded the metastasis or that they were present below our level of detection.

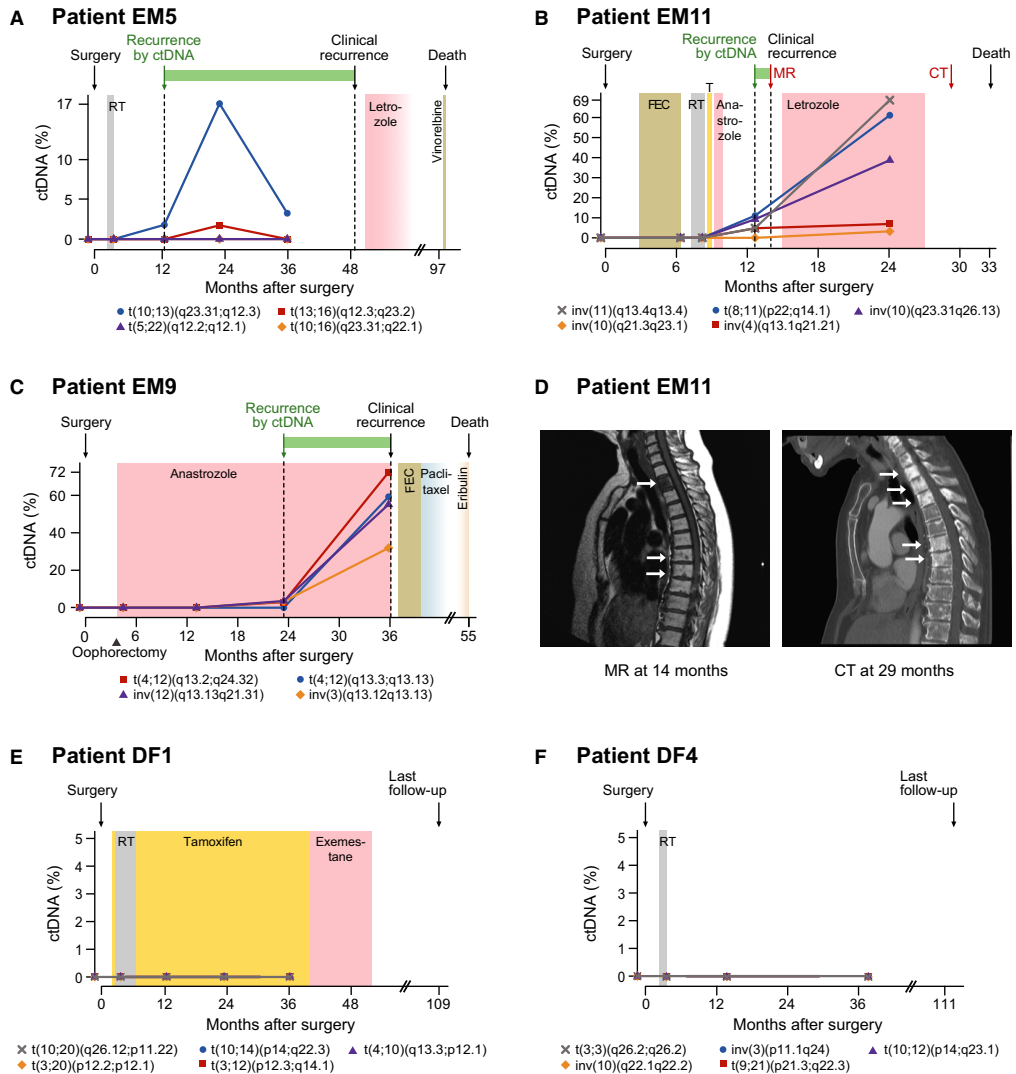
Patient EM11 displayed complex circulating tumor DNA dynamics. She was diagnosed with stage III invasive ductal carcinoma, hormone receptor-positive and high-grade histopathology, and received radiotherapy as well as several adjuvant systemic therapies due to intolerance (Fig 4B). Molecular recurrence was detected at 13 months via positive detection of four out of five rearrangements in her circulating DNA. Clinical recurrence was diagnosed at 14-month follow-up due to bone pain and confirmed by magnetic resonance imaging (Fig 4D), and she received letrozole therapy. At the 24-month follow-up time-point, however,

three of five rearrangements increased in abundance by fourfold to 14-fold, one chromosome 4q inversion remained stably low, and a fifth rearrangement (inversion on 10q) could be detected. This is consistent with partial response of the tumor clone containing the 4q inversion but inherent or acquired resistance to letrozole by one or more subclones containing the other four rearrangements. Computed tomography (CT) of the spine at 14 months showed progressive disease, consistent with ctDNA quantification (Fig 4D). Similarly, for patient EM9, three out of four tumor-specific chromosomal aberrations indicated molecular recurrence at 23-month follow-up, preceding clinical detection by 13 months, and during ongoing anastrozole therapy (Fig 4C). All four rearrangements increased dramatically at the 36-month follow-up time-point, coincident with confirmed distant metastases in the brain and liver by CT.

Interestingly, our sequencing analysis of the bilateral tumors of patient EM6 confirmed that they were two independent primaries with no clonal relatedness (Supplementary Table S2). Furthermore, ctDNA analyses indicated that the right-side tumor gave rise to the occult metastatic disease that was detectable by ddPCR at 2-year follow-up (37 months prior to clinical recurrence; Supplementary Table S6), whereas there was no ctDNA evidence of metastatic disease arising from the left primary tumor (Supplementary Fig S2).

### ctDNA as a predictive factor

In patients with known eventual clinical metastasis, ctDNA-based molecular detection of occult metastasis preceded the clinical diagnosis in 12 of 14 patients (86%), with an average lead time window of 11 months (range 0–37 months) (Fig 5B; Supplementary Table S4). Furthermore, a positive ctDNA blood test was always eventually



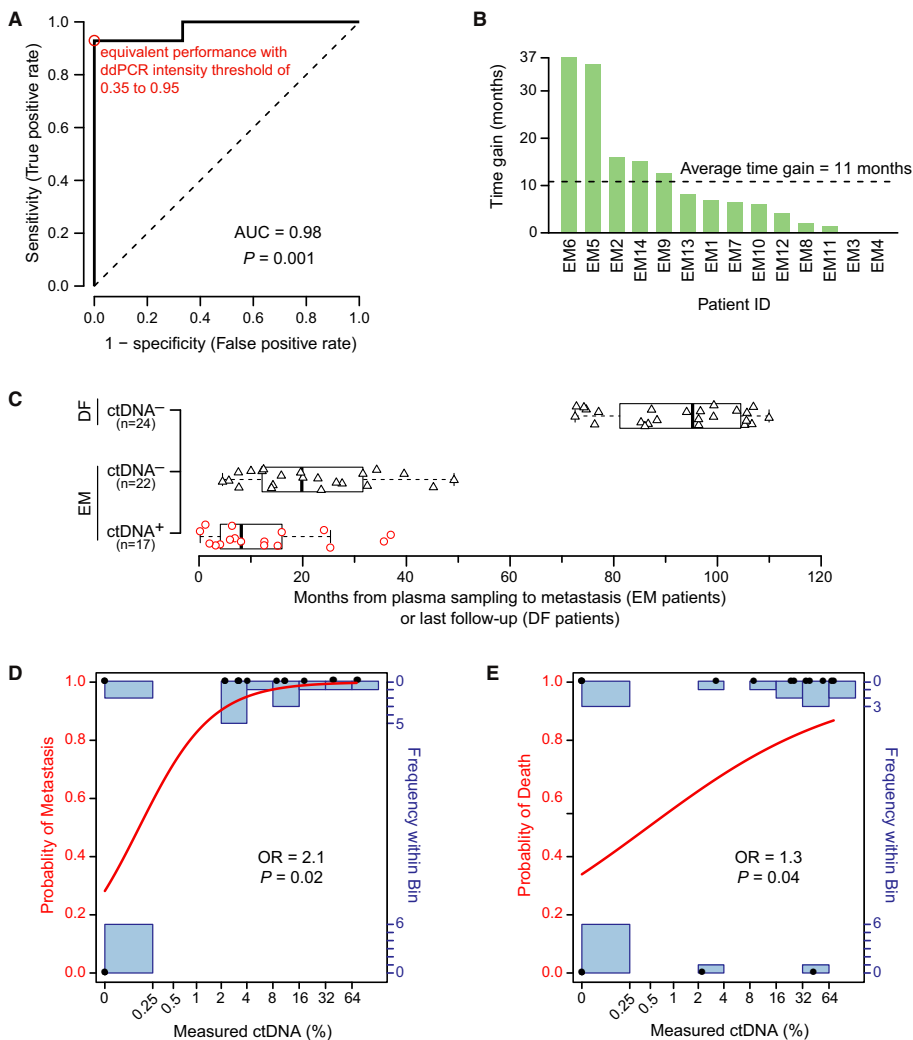
**Figure 4. Monitoring multiple tumor-specific chromosomal rearrangements in circulating DNA.**

**A–C** Plasma levels of circulating tumor DNA (ctDNA), quantified using ddPCR, for three patients with known eventual recurrence. Specific rearrangements are indicated by colored markers and labeled according to cytogenetic nomenclature (t denotes translocation, inv is inversion, and del is deletion). The recurrence by ctDNA time-point is defined as the earliest follow-up plasma sample (after surgery) with ctDNA detected at a level greater than 0% (compared to total cell-free circulating DNA) for at least one rearrangement. All relevant clinical events are indicated above by arrows, time gain by ctDNA-based detection is indicated by a green horizontal bar, and radiation (RT), endocrine, and cytotoxic treatments are indicated by colored shading. T = tamoxifen; FEC = fluorouracil, epirubicin, and cyclophosphamide. See Supplementary Fig S2 for ctDNA time-course plots with clinical annotations for all patients.

**D** Correlative magnetic resonance (MR; T1 weighted) and computed tomography (CT) imaging for patient EM11 corresponding to the red arrows in (B). In the MR, low T1 signal (dark) is present in the entire second thoracic vertebra and as punctate lesions in several vertebrae in the middle thoracic spine. The CT 15 months later shows sclerosis (white) in multiple additional thoracic vertebrae, consistent with progression of metastatic disease.

**E, F** ctDNA plots for two patients with long-term disease-free survival.





**Figure 5. ROC analysis, time gain, and clinical outcome.**

**A** Receiver operating characteristic (ROC) curve analysis. The area under the curve (AUC) as a measure of the postsurgery classification accuracy to discriminate between 6 long-term disease-free (DF) and 14 eventual metastasis (EM) patients based on ctDNA is 0.98 (95% CI 0.75–1.00;  $P = 0.001$ , two-sided Mann–Whitney  $U$ -test). The sensitivity and specificity were maximal (red circle) at all ddPCR relative fluorescence intensity thresholds between 0.35 and 0.95 (on a normalized scale from 0 to 1). The dashed line indicates a hypothetical test with performance no better than random.

**B** Time gained by ctDNA-based detection of recurrence in advance of clinically detected recurrence for all patients with clinical recurrence. For 12 out of 14 EM patients, ctDNA-based recurrence preceded clinical recurrence (time gain greater than zero).

**C** Boxplots indicating the time from a positive (red circles) or negative ctDNA plasma sample (black triangles) until an event, metastasis or last follow-up, for EM and DF patients. Box indicates the interquartile range (IQR), thick bar indicates the median, and whiskers extend to values within 1.5 times the IQR.

**D** Fitted curve from logistic regression with metastasis as endpoint. Measured ctDNA percentage and actual outcomes are indicated by black dots, the modeled probability is given by the red curve (left axis), and the number of measured data points in each bin is indicated by the blue bar graphs (right axis). Logistic regression odds ratio (OR) of 2.1 (95% CI 1.3 to infinity;  $P = 0.02$ , Wald test) is for each doubling of ctDNA.

**E** Fitted curve from logistic regression with death as endpoint. OR of 1.3 (95% CI 1.03–1.9;  $P = 0.04$ , Wald test) is for each doubling of ctDNA.

followed by clinical detection of metastasis with a median time from a positive ctDNA test to clinical metastasis of 8 months (Fig 5C). Among EM patients, a negative ctDNA test occurred for 13 of 14 patients at least once (patient EM14 has only positive time-points), with median time from a negative ctDNA blood test to a clinical metastasis of 20 months (Fig 5C). However, for all patients but one, the negative ctDNA tests were followed by a positive test (patient EM3 had undetectable ctDNA at all time-points). For the two EM patients (EM3, EM4) with exclusively negative ctDNA results prior to clinical metastasis (Supplementary Fig S2), the time interval from the preceding negative ctDNA test to clinical metastasis was 4.5 and 12 months, respectively (Fig 5C), indicating that narrower time intervals of ctDNA testing could be considered in future prospective studies.

Finally, we found ctDNA level to be quantitatively predictive of poor clinical outcome. Whereas none of the conventional univariable biomarkers (tumor size T, node status N, histological grade, ER, PR, HER2, or Nottingham Prognostic Index) were associated with outcome using logistic regression in this limited patient series, ctDNA level was a significant predictor of poor disease-free survival (odds ratio (OR) of 2.1 for each doubling of ctDNA level, 95% CI 1.3 to infinity;  $P = 0.02$ ; Fig 5E) as well as poor overall survival (OR 1.3 for each ctDNA doubling, 95% CI 1.03–1.9;  $P = 0.04$ ; Fig 5F) (Supplementary Table S7).

## Discussion

We studied cell-free circulating DNA in patients with primary breast cancer and show that ctDNA monitoring is accurate for the detection of occult metastasis. Metastasis could be detected by ctDNA in plasma for 13 of 14 patients and in none of the 6 patients with long-term disease-free survival. Moreover, ctDNA-based detection preceded clinical detection of metastasis for 86% patients with an average lead time of 11 months, and ctDNA was found to be a significant predictor for poor disease-free and overall survival. As far as we are aware, this study is the first to demonstrate that ctDNA monitoring can herald clinical detection of metastasis by months to several years and that ctDNA level, even when measured in the setting of primary breast cancer, is associated with significantly increased risk of poor outcome. Our results are in line with a recent report analyzing ctDNA in metastatic breast cancer patients using similar methods (Dawson *et al*, 2013) and are a significant finding given that ctDNA levels are considerably lower in earlier stage disease and thus inherently more difficult to detect than after clinical diagnosis of metastatic disease (Bettegowda *et al*, 2014). Together, these data provide support for the evaluation of ctDNA in the adjuvant setting in larger prospective studies to address several important questions. For example, it should be ascertained in clinical trials whether tailoring secondary adjuvant therapy by ctDNA monitoring can increase the rate of long-term breast cancer cure. Second, although other modalities have not shown a clinical benefit of early detection of occult metastasis, our results suggest that ctDNA may have the performance characteristics needed for earliest and accurate detection. This prompts for evaluation of whether and to what extent detection of occult metastasis by a ctDNA monitoring can improve outcomes. Furthermore, as part of a “watchful waiting” approach, additional inexpensive yet sensitive and specific molecular

surveillance by liquid biopsies could help enable a reduction in the “overtreatment” of patients with low-risk breast cancer.

Our method combines low-pass whole-genome sequencing with quantitative ddPCR-based personalized rearrangement analysis of plasma ctDNA and can be performed across dozens of liquid biopsies per patient for < €1,000 in reagents and < €50 per time-point, currently making it more much cost effective than approaches where sequencing of each liquid biopsy time-point is performed. Our analysis can also be achieved within a clinically useful time frame. In practice, candidate rearrangements could be identified and personalized ddPCR assays validated within 1 month of tumor biopsy, and a panel of ddPCR tests on patient plasma samples can be performed within 1 day. Multiple chromosomal rearrangements, supported by variable numbers of sequencing reads (including those nearby copy number aberrations which may be under positive selection), were chosen for plasma analysis to overcome the potential issue of intra-tumoral heterogeneity, where only a subclone comprising a varying fraction of the primary tumor gives rise to the metastatic growth(s). The chance for a false-negative result, where metastatic disease is present but never detected by ctDNA analysis, will decrease with each additional genomic aberration tested. In a WGS analysis of matched primary and metastatic breast cancers from the same patients, typically over 50% of chromosomal rearrangements present in the primary tumor can be found in its distant metastatic tumor, indicating that most genomic rearrangements occur relatively early during tumorigenesis and can be stable fingerprints for an individual's breast cancer (Tang and Gruvberger-Saal, manuscript in preparation). Determining the optimal criterion for candidate rearrangement selection and how many to monitor per patient/tumor are matters deserving additional study. Here, we chose to monitor four to six selected rearrangements per tumor due to limited volumes of plasma, which nevertheless was sufficient to detect metastatic disease in 13 out of 14 patients. Patient EM3 (Supplementary Fig S2), the only EM patient where we did not detect any ctDNA, had the fewest number of plasma samples (three compared to a median number of five samples per patient); therefore, we believe that increasing volume and frequency of plasma samples would be more beneficial than increasing the number of rearrangements tested per case. Our patient results for time to an event following a positive or negative ctDNA plasma sample suggests an interval of ~4–6 months between sampling may be reasonable, at least during the first few years of follow-up.

Our ddPCR-based method has similar analytical performance characteristics to other recently described methods for the analysis of circulating DNA, such as nested real-time PCR (McBride *et al*, 2010), digital PCR (Dawson *et al*, 2013), personalized analysis of rearranged ends (Leary *et al*, 2010), targeted deep sequencing of mutated genes of interest (Dawson *et al*, 2013), or direct deep sequencing of circulating DNA (Leary *et al*, 2012). We show our ddPCR method to be highly reproducible, linear, and able to detect 1 mutant target within 10,000 wild-type sequences. Importantly, our method capitalizes on the unique juxtaposition of sequences formed by chromosomal rearrangements and thus is less prone to false-positive signals compared to methods that use a preamplification step of the circulating DNA and/or assays that must discriminate between single-base differences amid wild-type and mutated alleles (Beaver *et al*, 2014). Our method's zero false-positive rate for the detection of somatic rearrangements in over 2.5 million control

normal haploid genomes compares exceedingly well to other methods and is a critical feature needed for clinically useful monitoring of patients with primary cancer where ctDNA fractions are low.

Although our results demonstrate the promise and benefits of ctDNA monitoring in primary breast cancer, there are several limitations. One, this proof-of-principle study was limited to 20 patients. Larger validation studies will be important to further clarify the utility of ctDNA monitoring in early-stage breast cancer and within the molecular subtypes. The availability and quantity of archival frozen plasma as well as the specific time-points of their collection also limited us. In the current configuration, in which cell-free DNA was isolated from 0.5 ml plasma and 4% of this was input per replicated ddPCR reactions (for four to six rearrangements per case), we estimate our method to be sensitive to detect one amplifiable target DNA molecule in 40  $\mu$ l of plasma (approximately 25 targets per ml of plasma). The sensitivity of our method to detect exceedingly low counts of target ctDNA could be improved linearly by increasing the amount of input DNA into ddPCR reactions, by multiplexing, by preamplification, and/or by isolating circulating DNA from a larger volume of plasma. For example, greater amounts of analytical material from 5 to 50 ml plasma would allow for an improved limit of detection of our method by at least one order of magnitude and to 1 target ctDNA molecule per 5 ml plasma or better. In the prospective setting, there would be the opportunity to better control the blood plasma collection procedures and time-points and take larger volume samples. Therefore, the sensitivity and apparent lead time advantage for occult metastasis detection reported herein may in fact be an underestimation.

Recently, Bettgeowda and colleagues reported that ctDNA was detected in a single time-point for 10 of 19 patients with localized breast cancer when inputting cell-free DNA isolated from 2 to 5 ml plasma; but no association with outcome was possible (Bettgeowda *et al*, 2014). In our study of patients with primary breast cancer, ctDNA could be detected in the presurgical plasma sample for 4/20 patients and all four had eventual recurrent disease. Although the sample size is small, and given the limitation of available plasma discussed above, the variation between patients in presurgical levels of ctDNA is intriguing and suggests that presurgical levels could serve as a potential prognostic factor deserving further study. In theory, ctDNA should be present in all patients prior to primary surgery. The limited plasma availability, and desire to analyze four to six rearrangements per time-point, likely impacted our preoperative detection rate. Indeed, oversampling for 17 patients with remaining presurgery cell-free DNA was possible using a single assay tested in at least 3 additional ddPCR reactions, which increased the presurgery detection rate to 9/20 (45%). Circulating tumor DNA monitoring might be feasible for the measurement of minimal residual disease at a time-point shortly after primary surgery; prospective studies with optimized plasma collection schedule and much larger plasma volumes will be required to evaluate this important question.

We have shown that ctDNA monitoring can herald clinical metastasis by months to years and that ctDNA is a quantitative predictive factor for poor outcome in the primary breast cancer setting. The future of breast cancer medicine is personalized therapies and precision care. For this to become a reality, noninvasive and accurate methods for monitoring of breast cancer progression and response to treatment will be necessary within the neoadju-

vant, adjuvant, and metastatic settings. Patient monitoring using noninvasive assays for ctDNA is proving to be a realistic means to discern biologically and clinically relevant information and shows great promise for incorporation into routine clinical management.

## Materials and Methods

### Ethics statement

The study was approved by the Regional Ethics Committee at Lund University including permission to publish de-identified clinical images (DNR 75-02, 37-08, 658-09, 58-12, 379-12, and 227-13). Trained health professionals provided written and oral information and all patients signed written informed consent in accordance with the WMA Declaration of Helsinki and the U.S. Department of Health and Human Services Belmont Report.

### Patients

Patients enrolled in the Breast Cancer and Blood Study (BC Blood, Sweden) (Borgquist *et al*, 2013), an ongoing prospective study at Lund University since 2002, were included in the present investigation for retrospective analysis of ctDNA. As shown in Fig 1A, patients were identified based on the following criteria: non-metastatic (stage I–III) breast cancer at initial diagnosis who received no neoadjuvant therapy, availability of frozen primary tumor specimen, frozen presurgery and two or more follow-up plasma samples collected during clinical course, and either clinically detected distant metastasis 1–6 years after diagnosis (termed eventual metastatic [EM] patients) or long-term disease-free survival > 7 years at last follow-up (termed DF patients). Out of 725 patients assessed, 24 EM and 63 DF patients passed eligibility requirements. From these, 20 patients were randomly selected 2:1 with respect to EM:DF categories. This sample size with multiple time-points per patient was considered to be sufficient to demonstrate the feasibility of ctDNA monitoring and test the hypothesis that occult metastasis can be detected by ctDNA analysis. Fourteen EM patients (first metastasis detected clinically at 14–61 months following diagnosis, median 20 months) and 6 DF patients (disease free at last follow-up, 109–113 months after diagnosis, median 110 months) were studied (Table 1 and Fig 1). The 20 patients were diagnosed between November 2002 and May 2007, received the standard of care, and were followed according to Swedish National Guidelines as well as additional structured follow-up as part of the BC Blood Study: patients met with a research nurse for study questionnaires (aimed at assessing symptoms and change in medication) and serial blood collection at specified time-points: prior to primary surgery and at approximately 3- to 8-, 12-, 24-, and 36-month follow-up time after primary surgery, and for biennial questionnaires thereafter. This was in addition to the routine clinical follow-up, which for patients not receiving chemotherapy consisted of clinical visits and mammography at follow-up years 1, 2, and 3 after primary surgery, and then by mammographic surveillance in the national screening program; and for patients receiving chemotherapy consisted of a clinical evaluation after completing chemotherapy

and followed by yearly clinical visits up through year 5, and then by mammographic surveillance. If any of the follow-up modalities indicated symptoms or signs of metastatic disease, appropriate imaging and confirmatory workup was performed per standard clinical practice. All cancer therapies are indicated for each patient in Supplementary Fig S2. For all patients included herein, all collected blood sample time-points were analyzed, and study results were blinded to the clinic. In all parts (sequencing, circulating DNA isolation, and ddPCR), patients were analyzed in random order without regard to clinical parameters and the ddPCR data were analyzed in an automatic fashion blinded to outcome and operator (detailed below).

### Whole-genome sequencing analysis

Primary tumor specimens were snap-frozen immediately after surgery and stored at  $-80^{\circ}\text{C}$  in the South Swedish Breast Cancer Group tumor bank. The tumor DNA isolation method is described in the Supplementary Methods. Whole-genome paired-end Illumina sequencing libraries were constructed from tumor DNA sheared to a median insert size of 500 bp, sequenced on our laboratory HiSeq 2000 instruments, and aligned to the human reference genome GRCh37 (Supplementary Table S1). Matched normal genomic DNA was isolated for all patients from whole blood. For three of the included patients as well as seven unrelated patients, normal genomic DNA samples were also sequenced and used to filter germline and false-positive rearrangements arising from errors in the human reference genome sequence and from regions of unreliable mappability. Chromosomal rearrangements were identified (Supplementary Fig S1 and Supplementary Table S2) and the exact rearrangement fusion sequence reconstructed using our bioinformatics pipeline SplitSeq (Supplementary Methods). PCR validation is described below and in the Supplementary Methods.

### Plasma DNA isolation and ddPCR

Blood samples were collected from patients in EDTA tubes and were centrifuged to separate plasma from peripheral blood cells within 2 h of collection, and the fractions were frozen at  $-80^{\circ}\text{C}$ . Total cell-free circulating DNA was isolated from 0.5 ml plasma using the QIAamp UltraSens Virus DNA kit (Qiagen) with protocol modifications. For selected rearrangements, polymerase chain reaction (PCR) primers and a double-quenched fluorescent 5'-3'-exonuclease hydrolysis probe were designed (mean amplicon size, 101 bp; range 63–155 bp) (Supplementary Table S3). For a subset of the rearrangements confirmed somatic using touchdown PCR with rearrangement-specific primers and primary tumor DNA or matched normal DNA as input (Supplementary Methods; Supplementary Table S4), the probe was synthesized (Integrated DNA Technologies) and the quantitative assay validated using a Bio-Rad QX100 droplet digital PCR (ddPCR) instrument using primary tumor DNA and matched normal DNA as controls. A ddPCR assay (Supplementary Table S3) targeting a 132-bp non-rearranged normal region of chromosome 2p14, which rarely undergoes copy number alteration in breast cancer (Jonsson *et al*, 2010), was used to estimate total circulating DNA (both tumor- and normal cell derived). For ddPCR, four to six tumor-specific rearrangement assays were analyzed,

wherein 4% (4  $\mu\text{l}$ ) of the isolated cell-free DNA (corresponding to 20  $\mu\text{l}$  plasma) was input in each assay reaction and the absolute count of the target sequence was measured (Hindson *et al*, 2011). Primary tumor DNA and matched normal DNA were used as positive and negative controls, respectively, for every personalized rearrangement assay in every ddPCR run, and a no-template control (water) was used as a negative control for the 2p14 control assay. All rearrangement reactions were run in duplicate. Detailed methods are presented in the Supplementary Methods.

### ddPCR data normalization

To enable an unbiased, uniform, and outcome- and operator-blinded automatic evaluation of ddPCR data, droplet fluorescent intensity measurements of each assay were normalized to a relative scale ranging from 0 to 1 by scaling to the negative control and positive control droplet intensities, for each assay, using custom scripts (see Supplementary Methods and Supplementary Fig S3). Droplets with a relative intensity  $\geq 0.5$  were defined positive (receiver operating curve characteristic analyses were performed to assess discriminatory accuracy at all thresholds; see below). The number of fragments per  $\mu\text{l}$  input purified circulating DNA ( $C_{Vi}$ ) was calculated from the number of positive droplets  $P$ , total number of droplets analyzed  $T$ , droplet volume  $V_d$  ( $0.91 \times 10^{-3}$   $\mu\text{l}$ ), ddPCR volume  $V_r$  (including PCR mix, primers, probe, input DNA), and volume of purified circulating DNA input into the reaction  $V_i$ , using the formula  $C_{Vi} = \left( \frac{-\ln(1 - \frac{P}{T})}{V_d} \right) \left( \frac{V_r}{V_i} \right)$ . A plasma sample was defined to be positive for ctDNA if one or more of the target tumor-specific rearrangements in the sample had a molecular count greater than zero by ddPCR analysis. To control for possible variability in the efficiency of plasma DNA isolation or degradation of cell-free circulating DNA during long-term storage of plasma, for each rearrangement, ctDNA level was estimated as a percentage of total circulating DNA by dividing the quantity of measured rearrangement by the quantity of the 2p14 control region.

### Receiver operating characteristic (ROC) curve analysis

Because the fluorescent intensity threshold used in calling ddPCR droplets positive or negative may influence the accuracy of ctDNA-based monitoring for occult disease, we applied a ROC curve analysis. In this analysis, the droplet intensity threshold, for every assay, was incrementally varied from 0 to 1 in 0.1 steps and applied to the normalized data for all samples, defining negative droplets (below threshold) and positive droplets (above threshold). At each threshold, the concentration of each rearrangement was calculated across all time-points and the rearrangement with the highest concentration was used to represent each time-point as this was thought to be most clinically relevant. Thus, ctDNA was represented and analyzed using a single covariate. Based on this, a patient was classified either as recurrence positive if one or more plasma samples during the follow-up period were positive for ctDNA, or as recurrence negative if all plasma samples during the follow-up period were negative for ctDNA. The predicted recurrence state was then compared with the known true recurrence state obtained from the clinical records in order to determine true-positive (TP), true-negative (TN), false-positive (FP), and

false-negative predictions (FN). Sensitivity was calculated as  $TP/(TP+FN)$ , and specificity was calculated as  $TN/(TN+FP)$ . Sensitivity was plotted against 1-specificity for each threshold, producing a ROC curve. The area under the curve was calculated using the R package *ROCR* (Sing *et al*, 2005).

### Statistical analyses

All statistical calculations were done in R v2.14.1. Confidence intervals for sensitivity, specificity, and area under the ROC curve were calculated based on the Clopper–Pearson exact binomial distribution method using the R package *binom* v1.1-1 (see Supplementary Methods). Except for the logistic regression odds ratios (see below), the Mann–Whitney test for significance was utilized throughout because the data types are not normally distributed and this test makes no assumption on the distribution. All *P*-values and confidence intervals calculated are two-sided except for the confidence interval for specificity (one-sided 95% confidence interval since the proportions are estimated to 1).

### Logistic regression

To determine the influence of ctDNA level and primary diagnosis clinical parameters (Table 1) on the risk of clinical metastasis and of death, we carried out univariable logistic regression analyses. Postsurgical plasma ctDNA percentage levels were used as a continuous covariate by taking, for each patient, the most recent plasma sample time-point prior to an outcome event, and for each time-point, using the rearrangement with the maximal ctDNA percentage value as this was thought to be most clinically relevant. Due to quasi-complete separation of ctDNA level between DF patients (Fig 5D, lower black dots) and EM patients (Fig 5D, upper black dots), we employed Firth's penalized likelihood approach (Firth, 1993) that allows reliable estimation also for separated data (Heinze, 2006). Since we assumed that, for example, a 10-unit increase in ctDNA percentage from 0 to 10% may have a different prognostic implication than an increase of the same magnitude from 50 to 60%, we allowed for nonlinear effects of ctDNA levels on the risk. Log<sub>2</sub>-transformation minimized the summed Akaike information criteria (see Supplementary Methods); therefore, log<sub>2</sub>-transformed ctDNA percentage was used as covariate, and accordingly, the resulting odds ratios are for each twofold increase in percentage ctDNA (e.g., from 1 to 2%, or 3 to 6%). The primary diagnosis clinical parameters of tumor size (T3, > 5 cm, versus T1, ≤ 2 cm, and T2, 2–5 cm), number of positive lymph nodes (N1, 1–3 positive, N2, 4–9 positive, and N3 > 9 positive nodes versus N0, none), Nottingham histological grade (G3 versus G1 and G2), estrogen receptor status (ER negative versus ER positive), progesterone receptor status (PR negative versus PR positive), and HER2 status (HER2 positive versus HER2 negative) were each used as single covariates in univariable logistic regression analyses with respect to the outcome variables, clinical recurrence, and vital status at last follow-up. No other candidate variables were considered. For patient EM6 with bilateral breast cancer, the variables for the left-side tumor with worse clinical prognostic features were used (Table 1). Analyses were carried out using the R package *brglm* (Kosmidis, 2013), with the statistical significance of estimated odds ratios evaluated by the Wald test.

### The paper explained

#### Problem

Breast cancer is the most common cancer affecting women and a leading cause for cancer-related death. Despite our best medical treatment, late metastatic recurrences are common. Unfortunately, metastatic breast cancer is usually diagnosed only after it has become symptomatic, and by this time, it is essentially incurable. On the other hand, a significant number of patients with non-metastatic breast cancer may be "overtreated" and subject to unnecessary side effects of systemic therapies when they are in fact cancer free. It may be possible, with a highly sensitive and specific molecular method to quantify circulating tumor DNA, to detect asymptomatic metastatic recurrences early or to determine a cancer-free state, noninvasively, using a blood test.

#### Results

We have used whole-genome sequencing of breast cancers to identify tumor-specific chromosomal rearrangements that serve as molecular "fingerprints" of each patient's cancer. This is followed by droplet digital PCR-based quantification of tumor-specific rearranged DNA molecules in patient blood samples collected at various time-points during their clinical follow-up. Here, we identify for the first time that ctDNA monitoring provides a sensitive method for early detection of asymptomatic metastatic recurrence in patients diagnosed with primary breast cancer and that the presence and quantity of ctDNA is predictive of poor outcome in this key patient group. Patients with long-term disease-free survival had no detectable ctDNA at any time-point after surgery.

#### Impact

Our study shows that ctDNA monitoring is a highly accurate method for early detection of asymptomatic metastatic recurrence in patients diagnosed with non-metastatic breast cancer and that ctDNA-based metastasis detection can precede symptoms and clinical detection by wide margins. These results provide the rationale for clinical trials in early breast cancer to test the clinical utility and benefit of ctDNA monitoring.

### Data deposition

The raw unprocessed droplet digital PCR data and normalized data have been deposited in the Dryad Digital Repository (<http://datadryad.org>) with identifier doi: 10.5061/dryad.b6928 (<http://dx.doi.org/10.5061/dryad.b6928>). Due to patient privacy, the whole-genome sequencing data, which may contain personally identifiable genetic variation and disease-associated alleles, are not publicly available.

**Supplementary information** for this article is available online:

<http://embomolmed.embopress.org>

### Acknowledgements

We thank the patients for participation in this study; the surgeons, oncologists, pathologists, and nursing staff at the Skåne University Hospital (SUS) Breast Cancer Clinic and the South Swedish Breast Cancer Group for their support; Anders Kvist, Therese Törnqvist, Daniel Filipazzi, Christel Reuterswärd, Katja Harbst, Martin Lauss, Gabriella Honeth, Kristina Lövgren, Annette Möller, Karin Henriksson, Anna Weddig, Linda Ågren, Maj-Britt Hedenblad, and Sol-Britt Olsson, at the Division of Oncology and Pathology and SUS, for assistance and discussion; and Jeanette Valcich, Ulrika Åström, Ingrid Wilson, Björn Frostner,

and Susanne André at the Division of Oncology and Pathology for administrative assistance. This study was funded by the Swedish Cancer Society, Swedish Research Council, Swedish Foundation for Strategic Research, Knut and Alice Wallenberg Foundation, VINNOVA, and Governmental Funding of Clinical Research within National Health Service, Swedish Breast Cancer Group, Crafoord Foundation, Lund University Medical Faculty, Gunnar Nilsson Cancer Foundation, Skåne University Hospital Foundation, BioCARE Research Program, King Gustav Vth Jubilee Foundation, the Krappereup Foundation, and the Mrs. Berta Kamprad Foundation. The funders had no role in study design, data gathering, data analysis, data interpretation, decision to publish, or writing of the report.

### Author contributions

EO, AB, SKG-S, and LHS conceived the study. EO optimized whole-genome sequencing and ddPCR. CW performed all bioinformatics analyses. CW performed statistical analyses with input from P-OB. EO, AG, YC, RS, and LHS performed DNA purification and ddPCR experiments. EO, CW, AG, YC, JH, M-HET, MD, P-OB, LR, AB, SKG-S, and LHS designed research. EO, CW, AG, YC, P-OB, SKG-S, and LHS analyzed data. DG, DvW, MF, CI, CR, LR, and HJ provided clinical information, and CI, CR, and HJ provided blood samples. LHS supervised the project and wrote the report with assistance from EO and CW. All authors discussed, critically revised, and approved the final version of the report for publication.

### Conflict of interest

The authors declare that they have no conflict of interest.

### For more information

Author's Web site: <http://tinyurl.com/saalgroup>.

## References

- Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L *et al* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* 486: 405–409
- Beaver JA, Jelovac D, Balukrishna S, Cochran R, Croessmann S, Zabransky D, Wong HY, Valda Toro P, Cidado J, Blair BG *et al* (2014) Detection of cancer DNA in plasma of early stage breast cancer patients. *Clin Cancer Res* 20: 2643–2650
- Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM *et al* (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* 6: 224ra224
- Borgquist S, Hjertberg M, Henningson M, Ingvar C, Rose C, Jernstrom H (2013) Given breast cancer, is fat better than thin? Impact of the estrogen receptor beta gene polymorphisms. *Breast Cancer Res Treat* 137: 849–862
- Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490: 61–70
- Cheng L, Swartz MD, Zhao H, Kapadia AS, Lai D, Rowan PJ, Buchholz TA, Giordano SH (2012) Hazard of recurrence among women after primary breast cancer treatment—a 10-year follow-up using data from SEER-Medicare. *Cancer Epidemiol Biomarkers Prev* 21: 800–809
- Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, Dunning MJ, Gale D, Forshew T, Mahler-Araujo B *et al* (2013) Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 368: 1199–1209
- Diaz LA Jr, Williams RT, Wu J, Kinde I, Hecht JR, Berlin J, Allen B, Bozic I, Reiter JG, Nowak MA *et al* (2012) The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 486: 537–540
- Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, Thornton K, Agrawal N, Sokoll L, Szabo SA *et al* (2008) Circulating mutant DNA to assess tumor dynamics. *Nat Med* 14: 985–990
- Dolken G (2001) Detection of minimal residual disease. *Adv Cancer Res* 82: 133–185
- Early Breast Cancer Trialists' Collaborative Group (2005) 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
- Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, Van Tine BA, Hoog J, Goiffon RJ, Goldstein TC *et al* (2012) Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* 486: 353–360
- Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* 80: 27–38
- Heinze G (2006) A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med* 25: 4216–4226
- Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright JJ, Lucero MY, Hiddessen AL, Legler TC *et al* (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* 83: 8604–8610
- Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. *CA Cancer J Clin* 61: 69–90
- Jonsson G, Staaf J, Vallon-Christersson J, Ringner M, Holm K, Hegardt C, Gunnarsson H, Fagerholm R, Strand C, Agnarsson BA *et al* (2010) Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res* 12: R42
- Jung K, Fleischhacker M, Rabien A (2010) Cell-free DNA in the blood as a solid tumor biomarker—a critical appraisal of the literature. *Clin Chim Acta* 411: 1611–1624
- Khatcheressian JL, Hurley P, Bantug E, Esserman LJ, Grunfeld E, Halberg F, Hantel A, Henry NL, Muss HB, Smith TJ *et al* (2013) Breast cancer follow-up and management after primary treatment: American Society of Clinical Oncology clinical practice guideline update. *J Clin Oncol* 31: 961–965
- Kosmidis I. (2013) brglm: Bias reduction in binary-response Generalized Linear Models
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19: 1639–1645
- Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, Duncan C, Antipova A, Lee C, McKernan K, De La Vega FM *et al* (2010) Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med* 2: 20ra14
- Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, O'Shaughnessy J, Kinzler KW, Parmigiani G, Vogelstein B *et al* (2012) Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* 4: 162ra154
- Lippman M, Osborne CK (2013) Circulating tumor DNA—ready for prime time? *N Engl J Med* 368: 1249–1250
- Lucci A, Hall CS, Lodhi AK, Bhattacharyya A, Anderson AE, Xiao L, Bedrosian I, Kuerer HM, Krishnamurthy S (2012) Circulating tumour cells in non-metastatic breast cancer: a prospective study. *Lancet Oncol* 13: 688–695
- McBride DJ, Orpana AK, Sotiropoulos C, Joensuu H, Stephens PJ, Mudie LJ, Hamalainen E, Stebbings LA, Andersson LC, Flanagan AM *et al* (2010) Use of cancer-specific genomic rearrangements to quantify disease burden in

- plasma from patients with solid tumors. *Genes Chromosom Cancer* 49: 1062–1069
- McDermott U, Downing JR, Stratton MR (2011) Genomics and the continuum of cancer care. *N Engl J Med* 364: 340–350
- Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 11: 685–696
- Misale S, Yaeger R, Hobor S, Scala E, Janakiraman M, Liska D, Valtorta E, Schiavo R, Buscarino M, Siravegna G et al (2012) Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* 486: 532–536
- Murtaza M, Dawson SJ, Tsui DW, Gale D, Forshew T, Piskorz AM, Parkinson C, Chin SF, Kingsbury Z, Wong AS et al (2013) Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* 497: 108–112
- Newman AM, Bratman SV, To J, Wynne JF, Eclow NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE et al (2014) An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* 20: 548–554
- Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA et al (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149: 979–993
- Rack B, Schindlbeck C, Juckstock J, Andergassen U, Hepp P, Zwingers T, Friedl TW, Lorenz R, Tesch H, Fasching PA et al (2014) Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *J Natl Cancer Inst* 106: dju066
- Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J et al (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 461: 809–813
- Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21: 3940–3941
- Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ et al (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 462: 1005–1010
- Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR et al (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486: 400–404
- Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M (1989) Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* 46: 318–322
- Theriault RL, Carlson RW, Allred C, Anderson BO, Burstein HJ, Edge SB, Farrar WB, Forero A, Giordano SH, Goldstein LJ et al (2013) Breast cancer, version 3.2013. *J Natl Compr Canc Netw* 11: 753–761
- Yung TK, Chan KC, Mok TS, Tong J, To KF, Lo YM (2009) Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non-small cell lung cancer patients. *Clin Cancer Res* 15: 2076–2084



**License:** This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## Paper III









# Detection of circulating tumor cells and circulating tumor DNA before and after mammographic breast compression in a cohort of breast cancer patients scheduled for neoadjuvant treatment

Daniel Förnvik<sup>1</sup> · Kristina E. Aaltonen<sup>2</sup> · Yilun Chen<sup>3</sup> · Anthony M. George<sup>3</sup> · Christian Brueffer<sup>3</sup> · Robert Rigo<sup>3</sup> · Niklas Loman<sup>3,4</sup> · Lao H. Saal<sup>3</sup> · Lisa Rydén<sup>5,6</sup>

Received: 25 April 2019 / Accepted: 17 June 2019 / Published online: 24 June 2019  
© The Author(s) 2019

## Abstract

**Purpose** It is not known if mammographic breast compression of a primary tumor causes shedding of tumor cells into the circulatory system. Little is known about how the detection of circulating biomarkers such as circulating tumor cells (CTCs) or circulating tumor DNA (ctDNA) is affected by breast compression intervention.

**Methods** CTCs and ctDNA were analyzed in blood samples collected before and after breast compression in 31 patients with primary breast cancer scheduled for neoadjuvant therapy. All patients had a central venous access to allow administration of intravenous neoadjuvant chemotherapy, which enabled blood collection from superior vena cava, draining the breasts, in addition to sampling from a peripheral vein.

**Results** CTC and ctDNA positivity was seen in 26% and 65% of the patients, respectively. There was a significant increase of ctDNA after breast compression in central blood ( $p=0.01$ ), not observed in peripheral testing. No increase related with breast compression was observed for CTC. ctDNA positivity was associated with older age ( $p=0.05$ ), and ctDNA increase after breast compression was associated with high Ki67 proliferating tumors ( $p=0.04$ ). CTCs were more abundant in central compared to peripheral blood samples ( $p=0.04$ ).

**Conclusions** There was no significant release of CTCs after mammographic breast compression but more CTCs were present in central compared to peripheral blood. No significant difference between central and peripheral levels of ctDNA was observed. The small average increase in ctDNA after breast compression is unlikely to be clinically relevant. The results give support for mammography as a safe procedure from the point of view of CTC and ctDNA shedding to the blood circulation. The results may have implications for the standardization of sampling procedures for circulating tumor markers.

**Keywords** Circulating tumor cells · Circulating tumor DNA · Breast compression · Breast cancer · Mammography · Neoadjuvant

Lao H. Saal and Lisa Rydén shared senior authorship.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10549-019-05326-5>) contains supplementary material, which is available to authorized users.

✉ Daniel Förnvik  
daniel.fornvik@med.lu.se

<sup>1</sup> Department of Translational Medicine, Medical Radiation Physics, Lund University, Malmö, Sweden

<sup>2</sup> Department of Laboratory Medicine, Division of Translational Cancer Research, Lund University, Lund, Sweden

<sup>3</sup> Department of Clinical Sciences Lund, Division of Oncology and Pathology, Lund University, Lund, Sweden

## Abbreviations

CTC	Circulating tumor cells
ctDNA	Circulating tumor DNA
RNA-seq	RNA sequencing

<sup>4</sup> Department of Oncology, Skåne University Hospital, Lund, Sweden

<sup>5</sup> Department of Clinical Sciences Lund, Division of Surgery, Lund University, Lund, Sweden

<sup>6</sup> Department of Surgery and Gastroenterology, Skåne University Hospital, Malmö, Sweden

MAF      Mutant allele frequency  
TNBC      Triple-negative cancers

## Introduction

Circulating tumor markers such as circulating tumor cells (CTCs) and circulating tumor DNA (ctDNA) can be found in the blood of cancer patients. As a liquid biopsy, these markers complement solid biopsies and have the advantage of being physically more accessible and patient-friendly than traditional tissue biopsies. This provides a possibility for prognosis prediction, closer monitoring of treatment response and disease progression, identification of drug targets, as well as an opportunity for early detection of recurrence. The presence of CTCs in the blood of patients with primary breast cancer has been shown to be an independent predictor of decreased disease-free and overall survival [1, 2], but the treatment predictive value of the cells is still under debate [3, 4]. The CTC methodology in primary breast cancer is also limited by the low number of detected cells, which makes enumeration and evaluation statistically challenging [5]. ctDNA, the cell-free DNA that originates from cancer cells, is a promising biomarker whose prognostic and treatment predictive power is emerging [6, 7]. Recent studies have shown that quantification of specific mutations in ctDNA can be associated with early detection of metastases and therapy resistance in breast cancer as well as in other diagnoses [8–12].

The risk that tumor cells are released into the bloodstream from a primary tumor during surgical interventions has been addressed in a few studies [13–17], although how and when tumor cells are shed as well as the clinical importance of this release is poorly understood [18]. Animal studies have also found that physical manipulation of a primary tumor by applying pressure to it causes tumor cell dissemination [19–21]. We have previously investigated if mammographic breast compression in patients with an already present breast tumor could cause shedding of tumor cells to the peripheral circulation [22]. We found no indications that this would be the case in a pilot study of 24 patients with primary breast cancer.

However, the configuration of the human blood circulation can cause tumor cells released from the breast to pass through the capillary vasculature of the lungs before reaching the peripheral blood vessels. In our previous study [22], CTCs captured only in the peripheral blood might have resulted in an underestimation of CTC number. It has been shown that a higher number of CTCs can be found in central compared to peripheral venous blood in patients with metastatic breast cancer [23] as well as in other diagnoses [24–26]. Animal studies of colon carcinoma cells have shown that the majority (80–100%) of tumor cells could be

trapped in the capillary bed of the first organ they encounter [27]. In breast cancer, an autopsy study by Peeters et al. [28] showed that CTCs were trapped in the lung microvasculature in four of the nine patients who all had high CTC counts (> 100). Thus, it is likely that the number of CTCs found in the peripheral blood system is not representative of a possible release of tumor cells from the primary tumor during manipulation such as breast compression during mammography or surgery. The difference in CTC number between central and peripheral blood is possibly even more pronounced after specific interventions compared to a more steady-state-like condition of metastatic disease [24, 26].

To our knowledge, ctDNA levels have not been used to study a possible release of tumor cells or tumor cell debris after breast compression or any other mechanical intervention in breast cancer. Relatively few studies have so far compared the levels of both CTCs and ctDNA in the same clinical patient cohort at identical time points and our understanding of the relationship between the two liquid tumor markers is limited. However, both the level of ctDNA and the number of CTCs have been shown to have a prognostic value in mainly metastatic breast cancer cohorts [29, 30]. Mutation analysis of ctDNA and single CTCs suggests that ctDNA reflects the heterogeneity of mutations found in individual CTCs [30], but ctDNA levels have been found to have a higher correlation with tumor burden than CTCs [29].

The aim of this study was to investigate how the presence of CTCs and ctDNA are affected by breast compression during mammography in patients with primary breast cancer. Special emphasis was made on comparing circulating tumor marker burden between the central and peripheral blood circulation.

## Materials and methods

### Patient cohort and clinical parameters

The patient cohort comprises preoperative patients within the ongoing SCAN-B trial (Clinical Trials ID NCT02306096) at Lund University and Skåne University Hospital, Sweden [31, 32]. During 2015–2016, 31 patients scheduled for neo-adjuvant therapy volunteered to do an extra mammography after diagnosis and were included in the present study. The patient mean age was 51.9 years (range 33–74 years) and the mean compressed breast thickness and applied compression force during the examination were 55.8 mm (range 26.5–77.0 mm) and 103.4 N (range 71.5–123.1 N), respectively, as indicated by the mammography system (Mammomat Inspiration, Siemens Healthineers, Erlangen, Germany). All patients gave written informed consent and the study was approved by the Regional Ethical Review Board in Lund, Sweden (diary number 2014/521).

Clinical data including biomarker expression, histological subtype, and nodal status were retrieved from pathology reports and the patient's clinical charts. Information on biomarker expression was based on analysis from the core needle biopsy before initiation of neoadjuvant therapy. Estrogen receptor (ER) positivity was defined as  $\geq 10\%$  positive cancer cells, human epidermal growth factor receptor 2 (HER2) positivity was defined by immunohistochemistry (IHC) or in situ hybridization (ISH) as (IHC3+) or ISH-positive cells, and Ki67 positivity was defined as  $> 20\%$  positive cancer cells. Information from mammograms, ultrasound images, and breast tomosynthesis was compiled into one measure of tumor size.

### Blood sampling

All patients had a central venous access to allow administration of intravenous neoadjuvant chemotherapy, which enabled blood collection from superior vena cava, draining the breasts. A dedicated research nurse attended the patient during the mammography examination and acquired blood samples before and after mammographic breast compression, first from central venous access and secondly from a peripheral vein at both occasions. The median time for blood sampling after compression was 2 min (range 0–5 min) for central blood samples and 7 min (range 5–23) for peripheral blood samples. At each time point, 10 ml whole blood was collected in CellSave tubes (Menarini Silicon Biosystems, Bologna, Italy) for CTC analysis and 10 ml whole blood was collected in Cell-Free DNA Blood Collection Tubes (Streck Inc., Omaha, USA) for ctDNA analysis. The blood samples were transported at room temperature and subsequent analyses were performed within 96 h after sample taking.

### CTC analysis

The blood samples were analyzed for CTC number using the FDA-approved CellSearch<sup>®</sup> system (Menarini Silicon Biosystems). Briefly, a ferrofluid-conjugated epithelial cell adhesion molecule (EpCAM)-directed antibody was used to separate CTCs from the majority of white blood cells. Fluorescent staining with DAPI (nuclear staining), cytokeratin (CK) 8, 18, 19-directed PE-conjugated antibodies, and CD45-directed APC-conjugated antibodies were applied to identify CTCs (DAPI+/CK+/CD45–). Two independent and accredited technicians manually evaluated images of CK+ events selected automatically by the CellTracks II system (Menarini Silicon Biosystems). The method has been described in detail elsewhere [33]. Cut-off for CTC positivity was  $\geq 1$  CTC/7.5 ml blood as suggested by a recent review of primary breast cancer [1].

### ctDNA analysis

Candidate somatic mutations for ctDNA measurement were obtained from RNA sequencing (RNA-seq) data generated within SCAN-B [31, 34]. Twenty-eight of the 31 patients had available tumor RNA-seq data. Sequencing, base calling, FASTQ file processing, and filtering were performed as previously described [34]. Using a Snakemake workflow, reads in FASTQ format were aligned to the human reference genome GRCh38.p8 (including alternative sequences and decoys, and patched with dbSNP Build 147) using HISAT2 2.0.5 [35] (with default options except `--rna-strandness RF`, `--rg-id ${ID_NAME}`, `--rg PL:illumina`, `--rg PU:${UNIT}`, `--rg SM:${SAMPLE}`), and duplicate reads were marked using SAMBLASTER 0.1.24. Variants were called using VarDict-Java 1.5.0 [36] (with default options except `-f 0.02`, `-N ${SAMPLE}`, `-b ${BAM_FILE}`, `-c 1`, `-S 2`, `-E 3`, `-g 4`, `-Q 10`, `-r 2`, `-q 20`), and annotated with dbSNP build 150 and COSMIC v84 using vcfanno 0.2.8 [37].

From the RNA-seq mutation calling, one somatic mutation for each patient was selected for IBSAFE assay design for ultrasensitive mutation detection. IBSAFE<sup>®</sup> (SAGA Diagnostics AB, Lund, Sweden) is an enhanced droplet digital PCR technology with significantly improved sensitivity and specificity, allowing for quantification of alleles to 0.001% mutant allele frequency (MAF) [George et al. manuscript in preparation]. IBSAFE assays targeting a somatic mutation were designed for 20 patients and the assays validated using 6 ng of corresponding tumor DNA as positive control and 180 ng of human normal genomic DNA (Promega, Madison, USA) as negative control, confirming a lower limit of detection of at least 0.0017% MAF for each assay.

Whole blood collected in Streck tubes were centrifuged at  $2000\times g$  for 15 min at room temperature to fractionate plasma, followed by clearing of the plasma fraction by centrifugation at  $10,000\times g$  for 15 min at 4 °C. Cell-free DNA was isolated using the QIAamp Circulating Nucleic Acid Kit or the QIAamp MinElute ccfDNA Midi Kit (Qiagen, Hilden, Germany), of which 20% of the eluate used for IBSAFE reactions and measurement of mutant and wild-type ctDNA copies and calculation of MAF.

### Statistical analysis

Clinical and patient-specific characteristics were compared between patients that had  $\geq 1$  CTC/ $\geq 0.01\%$  MAF present in any sample and patients with 0 CTCs/0% MAF in all samples. Agreement between CTC- and ctDNA-positive patients was analyzed using Cohen's kappa statistics. The Mann–Whitney U-test was used to compare the distribution of continuous variables. For categorical variables, Fisher's exact test was used in all comparisons due to less than five

expected cases in at least one of the groups in all cross-tables. Statistical analysis of all characteristics was also performed between patients that had an increase in CTC number/% MAF after compression with patients that did not have an increase in CTC number/% MAF after compression. A non-parametric Wilcoxon signed-rank test was applied to test for CTC differences/% MAF changes between before and after compression. For comparison between central and peripheral CTC/ctDNA measurements, a sign test was used.

All statistical analyses were performed in IBM SPSS Statistics (version 24, IBM, Armonk, NY, USA) and  $p$  values  $<0.05$  were considered significant.

## Results

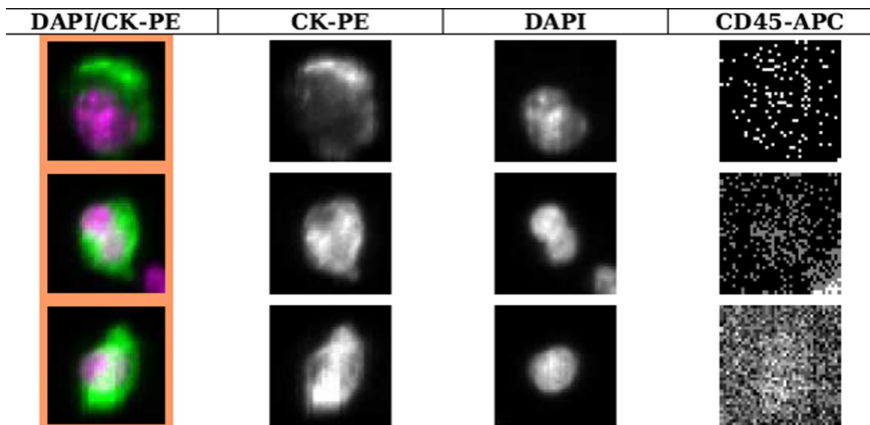
In total, 8/31 patients (26%) had  $\geq 1$  CTC in at least one of the blood samples taken before or after mammographic breast compression (Fig. 1). Correspondingly, 13/20 patients (65%) had  $\geq 0.01\%$  MAF and were defined as ctDNA positive. No agreement was found between CTC- and ctDNA-positive patients ( $\kappa=0.02$ ,  $p=0.92$ ) (A plot of CTC count versus % MAF can be found in supplementary Fig. S1).

Patient and tumor characteristics of the whole cohort, as well as of CTC/ctDNA-positive and CTC/ctDNA-negative patients separately, are shown in Table 1. No patient or pathologic characteristics were statistically associated with CTC positivity. Larger tumor size, non-ductal histological subtype, and older age were more predominant in the CTC-positive group but the difference was not statistically significant. ctDNA positivity was associated with higher age ( $p=0.05$ ). Higher Ki67, ductal histological type, and triple-negative breast cancer were more predominant in ctDNA-positive

patients, without reaching statistical significance (Table 1). Notably, 4/4 triple-negative breast cancers (TNBC) and 4/4 T4 staged cancers were all ctDNA positive.

Thirty patients had CTC results from the central blood sample before and after breast compression and 22 of these patients had 0 CTCs at both time points. Five of eight patients with detectable CTCs had an increased number of CTCs after compression ( $p=0.19$ ) (Fig. 2a). The average CTC increase was 3.2 cells (median 1.0 cell). Only two evaluable patients had detectable CTCs in the peripheral blood sample (Fig. 2b). Both central and peripheral % MAF generally increased after compression with the latter reaching significance ( $p=0.08$  and  $p=0.01$ ) (Fig. 2c, d). The average increase of % MAF was relatively small, 0.77 and 0.35 (median 0.35 and 0.22% MAF) for central and peripheral, respectively. Of the 20 patients with assessable ctDNA samples before and after breast compression, eleven and eight patients had 0% MAF in central and peripheral plasma samples, respectively, at both time points.

The median fraction of Ki67-positive cells was 66% (range 30–90%) in the five patients that had an increase in CTC number after compression, compared to 45% (range 15–95%) in patients with no increase ( $p=0.31$ ) (Supplementary Table 1). Also, Ki67 fraction was significantly higher in the group with increasing ctDNA after compression, 45% versus 30% ( $p=0.04$ ) (Supplementary Table 2). No other factors were differentially expressed between patients with an increase in CTCs/ctDNA levels and patients with a stable or a decrease in CTCs/ctDNA levels after compression. However, the histological type of the primary tumor seemed to differ between patients with an increase in the number of CTCs and patients with an increase in the levels of ctDNA after compression (Supplementary Tables 1 and 2).



**Fig. 1** Examples of CTCs detected with the CellSearch system from a patient in the study

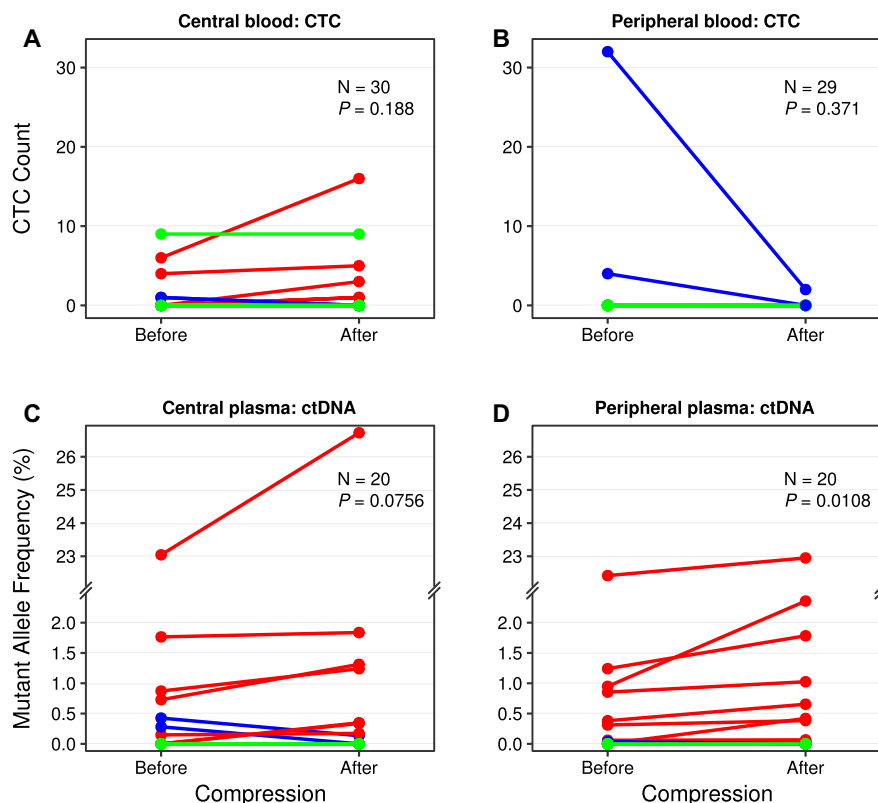
**Table 1** Comparison of patient and tumor characteristics between patients positive for CTCs and ctDNA ( $\geq 1$  CTC/ $\geq 0.01\%$  MAF) and patients with no CTCs/% MAF

	Total (N=31)	CTC negative (N=23)	CTC positive (N=8)	p value	Total (N=20)	ctDNA negative (N=7)	ctDNA positive (N=13)	p value
Age (years)								
Median (range)	50 (33–74)	47 (33–74)	56 (43–71)	0.21 <sup>a</sup>	51 (35–74)	46 (35–62)	58 (40–74)	0.05 <sup>a</sup>
< 50	16	13	3	0.43 <sup>b</sup>	10	5	5	0.35 <sup>b</sup>
$\geq 50$	15	10	5		10	2	8	
Tumor size and stage								
Median size, mm (range)	30 (4–90)	30 (4–90)	38 (16–80)	0.21 <sup>a</sup>	30 (4–80)	24 (8–80)	30 (4–80)	0.60 <sup>a</sup>
T1 (< 20 mm)	8	7	1	0.64 <sup>b</sup>	6	3	3	0.61 <sup>b</sup>
T2–T4 (20 mm or higher)	23	16	7		14	4	10	
Nodal stage								
N0	4	3	1	1.0 <sup>b</sup>	2	0	2	0.52 <sup>b</sup>
N+	27	20	7		18	7	11	
ER								
Negative (10% or lower)	8	6	2	1.0 <sup>b</sup>	4	0	4	0.25 <sup>b</sup>
Positive (> 10%)	23	17	6		16	7	9	
HER2								
Negative	25	18	7	1.0 <sup>b</sup>	17	5	12	0.27 <sup>b</sup>
Positive	6	5	1		3	2	1	
Ki67								
Median % of cells stained (range)	45 (15–95)	45 (20–95)	49 (15–90)	0.61 <sup>a</sup>	40 (15–90)	30 (15–90)	45 (20–90)	0.19 <sup>a</sup>
Low (20% or lower)	3	2	1	1.0 <sup>b</sup>	3	1	2	1.0 <sup>b</sup>
High (> 20%)	28	21	7		17	6	11	
Breast cancer subtype								
ER+	18	12	6	0.63 <sup>b</sup>	13	5	8	0.20 <sup>b</sup>
HER2+	6	5	1		3	2	1	
TNBC	7	6	1		4	0	4	
Multifocality								
No	22	17	5	0.64 <sup>b</sup>	15	4	11	0.29 <sup>b</sup>
Yes	8	5	3		5	3	2	
Missing	1	1						
Histological subtype								
Ductal	23	19	4	0.15 <sup>b</sup>	13	3	10	0.17 <sup>b</sup>
Other	8	4	4		7	4	3	
Detection mode								
Screening	9	7	2	1.0 <sup>b</sup>	8	2	6	0.64 <sup>b</sup>
Symptomatic	22	16	6		12	5	7	

<sup>a</sup>Mann–Whitney U-test<sup>b</sup>Fisher's exact test

CTCs were more abundant in central compared to peripheral blood in 8/10 positive samples ( $p = 0.04$ ) (Fig. 3a). Forty-nine comparisons between central and peripheral blood contained 0 CTCs in both samples. There was no significant difference in % MAF levels between

central and peripheral sampling (8/20 favoring higher % MAF in the central blood sample,  $p = 0.50$ ) (Fig. 3b). Twenty comparisons between central and peripheral blood contained 0% MAF in both samples.



**Fig. 2** The number of CTCs found before and after mammographic breast compression in central venous access (a), where two patients are represented by a line going from 0 to 1 CTC and from 1 to 0 CTCs, respectively. The corresponding number of CTCs before and after compression in peripheral blood (b). Figures for mutant allele frequency before and after breast compression in central (c), where two patients are represented by a line from 0 to approximately 0.35,

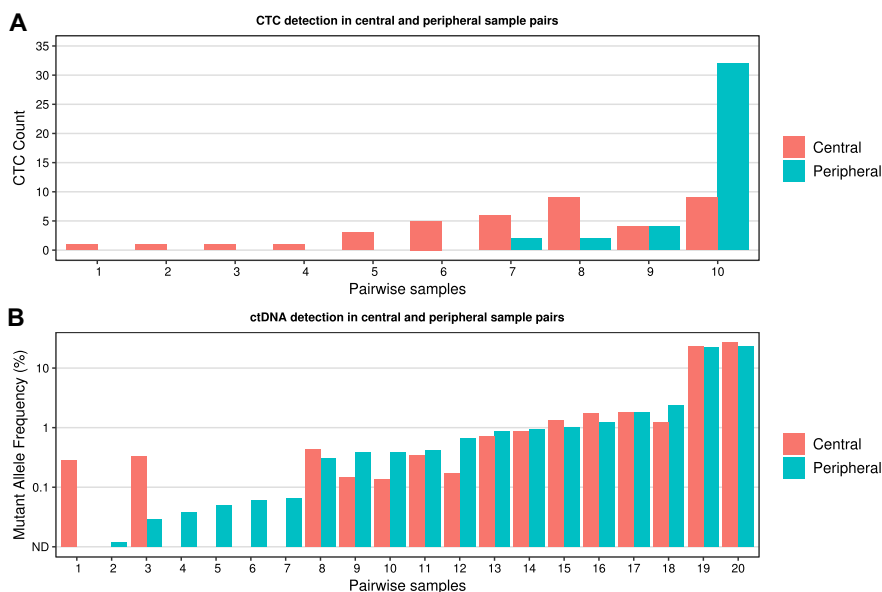
and peripheral (d) plasma, where two patients are represented by a line going from 0 to approximately 0.04 and from approximately 0.03 to 0, respectively. Patients with an increasing value are plotted with red lines, decreasing values in blue, and constant values in green. All patients that did not have any circulating tumor markers are summarized in one line at number/frequency = 0

## Discussion

CTCs were detected in 26% of the patients before start of neoadjuvant therapy for primary breast cancer. This is in line with a recent meta-analysis where 25.2% of breast cancer patients had CTCs before onset of neoadjuvant chemotherapy (deemed independent of blood sampling volume) [2]. ctDNA was positive in 65% of the patients. The concordance between CTCs and ctDNA has been shown to be higher in metastatic breast cancer patients [29] as compared to what was found in this study, which is most likely due to the lower rate of CTCs in primary breast cancer (Supplementary Fig. S1). ctDNA positivity, defined as  $\geq 0.01\%$  MAF in this

study, was associated with higher age ( $p=0.05$ ) and a trend was noted that a more aggressive tumor phenotype, including high Ki67, TNBC, and T4 staged cancers, favors ctDNA positivity (not statistically significant).

For CTCs, no increase was seen in either central or peripheral blood after mammographic breast compression ( $p=0.19$  and  $p=0.37$ , respectively). However, both central and peripheral ctDNA levels increased after breast compression ( $p=0.08$  and  $p=0.01$ , respectively). Only one patient had a CTC count difference of  $> 5$  cells/7.5 ml between samples taken before and after compression (Fig. 2a). This suggests a lack of a larger bolus release of cells during breast compression of women with primary breast cancer.



**Fig. 3** CTC (a) and ctDNA (b) detection in central and peripheral sample pairs. In 8/10 CTC-positive pairwise samples, a higher number of CTCs was detected in the central compared to the peripheral blood sample ( $p=0.04$ ). In pairwise samples 1–6, no CTCs were found in the peripheral blood sample. In 12/20 ctDNA pairwise

samples, a higher mutant allele fraction was found in the peripheral plasma sample ( $p=0.50$ ). In pairwise samples 2, 4–7, no ctDNA was detected centrally, and in sample 1, no ctDNA was detected peripherally

The central blood samples were drawn on average 2 min after compression and, according to an animal study, the release of malignant cells starts at the manipulation procedure and stays elevated up to 60 min [19]. To the best of the authors' knowledge, no published study has investigated how ctDNA levels vary with manipulation of a primary tumor with regard to applied pressure. The increase of ctDNA after breast compression found in this study can be considered relatively small, with only one case going from % MAF 0.95 to 2.36 which possibly could affect prognostication (Fig. 2d) [8]. High Ki67 were associated with increased ctDNA levels ( $p=0.04$ ) (Supplementary Table 2).

As hypothesized, CTCs were in general significantly more likely to be present in central than in peripheral blood samples ( $p=0.04$ ). Six patients presented CTCs only in the central samples (Fig. 3a) suggesting a differential CTC yield depending on sample location. The results are comparable to the work by Peeters et al. [23] in metastatic breast cancer but no data from studies involving differential blood sampling of primary breast cancer are hitherto available. This differential yield was not seen for ctDNA ( $p=0.50$ ). Since ctDNA is a much smaller moiety

and soluble in the blood, we speculate that ctDNA is much less affected by physical hindrance in the capillaries as compared to CTCs. Hence, ctDNA blood sampling is independent of blood drawing location, a finding that could contribute to the definition of clinical sampling routines in primary breast cancer for ctDNA.

The major limitation of this study was the small sample size and low count of CTCs, despite that a total of up to 40 ml whole blood was drawn from each patient. The statistical nature of CTC sampling has been described by Tibbe et al. [5]. Due to the low sample size, a possible difference between CTCs detection before and after breast compression may have been underestimated. Similarly, the ctDNA analysis was limited by a relatively low plasma input volume, and therefore a limited number of genome equivalents being analyzed for the presence of mutations.

When CTCs and ctDNA markers are implemented into clinical routine, our understanding of how the concentrations fluctuate during different interventions should be better understood. The women in this cohort are continuously being monitored and follow-up data will be available and presented in future publications.



## Conclusion

In summary, there was no significant release of CTCs after mammographic breast compression but more CTCs were present in central compared to peripheral blood. There was a small average increase in ctDNA levels after breast compression, unlikely to be clinically relevant, and no difference between central and peripheral levels was found.

**Acknowledgements** The authors would like to acknowledge all patients who participated in this study, the work of the SCAN-B community and the South Sweden Breast Cancer Group, and Sergii Gladchuk for bioinformatics assistance.

**Author contributions** All authors (DF, KEA, YC, AMG, CB, RR, NL, LHS, LR) contributed to the study conception and design. Material preparation: DF, KEA, NL, LHS, LR; data collection: DF, KEA, NL, LHS, LR; sample analysis: KEA, YC, AMG, CB, RR; and statistical analysis: DF, YC. The first draft of the manuscript was written by DF, KEA, LHS, LR, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** This work was supported by the Swedish Cancer Society, Swedish Research Council, VINNOVA, Mrs. Berta Kamprad Foundation, Governmental Funding of Clinical Research within National Health Service, Lund University Medical Faculty, Cancer Research Foundation at the Department of Oncology Malmö University Hospital, Gunnar Nilsson Cancer Foundation, BioCARE Research Program, King Gustav Vth Jubilee Foundation, and the Krappertup Foundation.

## Compliance with ethical standards

**Conflict of interest** YC, AMG, CB, RR, and LHS are shareholders and employees of SAGA Diagnostics AB. LHS has received honorarium from Novartis AG. All remaining authors have declared no conflicts of interest.

**Ethical approval** All procedures performed were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards and ethical approval was obtained from the Regional Ethical Review Board in Lund, Sweden (diary number 2014/521). The experiments comply with the current laws of Sweden.

**Informed consent** Written informed consent was obtained from all individual participants included in the study.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Janni WJ, Rack B, Terstappen LW et al (2016) Pooled analysis of the prognostic relevance of circulating tumor cells in primary breast cancer. *Clin Cancer Res* 22:2583–2593
- Bidard FC, Michiels S, Riethdorf S et al (2018) Circulating tumor cells in breast cancer patients treated by neoadjuvant chemotherapy: a meta-analysis. *J Natl Cancer Inst* 110:560–567
- Bardelli A, Pantel K (2017) Liquid biopsies, what we do not know (yet). *Cancer Cell* 31:172–179
- Yan WT, Cui X, Chen Q et al (2017) Circulating tumor cell status monitors the treatment responses in breast cancer patients: a meta-analysis. *Sci Rep* 7:43464
- Tibbe AG, Miller MC, Terstappen LW (2007) Statistical considerations for enumeration of circulating tumor cells. *Cytometry A* 71:154–162
- Wan JC, Massie C, Garcia-Corbacho J et al (2017) Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* 17:223–238
- Beddowes E, Sammut SJ, Gao M, Caldas C (2017) Predicting treatment resistance and relapse through circulating DNA. *Breast* 34:S31–S35
- Olsson E, Winter C, George A et al (2015) Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. *EMBO Mol Med* 7:1034–1047
- Murtaza M, Dawson SJ, Tsui DW et al (2013) Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* 497:108–112
- Spindler KL, Pallisgaard N, Andersen RF, Jakobsen A (2014) Changes in mutational status during third-line treatment for metastatic colorectal cancer—results of consecutive measurement of cell free DNA, KRAS and BRAF in the plasma. *Int J Cancer* 135:2215–2222
- Scholer LV, Reinert T, Orntoft MW et al (2017) Clinical implications of monitoring circulating tumor DNA in patients with colorectal cancer. *Clin Cancer Res* 23:5437–5445
- Loman N, Saal LH (2016) The state of the art in prediction of breast cancer relapse using cell-free circulating tumor DNA liquid biopsies. *Ann Transl Med* 4:S68
- Sandri MT, Zorzino L, Cassatella MC et al (2010) Changes in circulating tumor cell detection in patients with localized breast cancer before and after surgery. *Ann Surg Oncol* 17:1539–1545
- Papavasiliou P, Fisher T, Kuhn J et al (2010) Circulating tumor cells in patients undergoing surgery for hepatic metastases from colorectal cancer. *Proc (Bayl Univ Med Cent)* 23:11–14
- Koch M, Kienle P, Hinz U et al (2005) Detection of hematogenous tumor cell dissemination predicts tumor relapse in patients undergoing surgical resection of colorectal liver metastases. *Ann Surg* 241:199–205
- van Dalum G, van der Stam GJ, Tibbe AG et al (2015) Circulating tumor cells before and during follow-up after breast cancer surgery. *Int J Oncol* 46:407–413
- Hashimoto M, Tanaka F, Yoneda K et al (2014) Significant increase in circulating tumour cells in pulmonary venous blood during surgical manipulation in patients with primary lung cancer. *Interact Cardiovasc Thorac Surg* 18:775–783
- Martin OA, Anderson RL, Narayan K, MacManus MP (2017) Does the mobilization of circulating tumour cells during cancer therapy cause metastasis? *Nat Rev Clin Oncol* 14:32–44
- Juratli MA, Sarimollaoglu M, Siegel ER et al (2014) Real-time monitoring of circulating tumor cell release during tumor manipulation using in vivo photoacoustic and fluorescent flow cytometry. *Head Neck* 36:1207–1215
- Juratli MA, Siegel ER, Nedosekin DA et al (2015) In vivo long-term monitoring of circulating tumor cells fluctuation during medical interventions. *PLoS ONE* 10:e0137613
- Nishizaki T, Matsumata T, Kanematsu T et al (1990) Surgical manipulation of VX2 carcinoma in the rabbit liver evokes enhancement of metastasis. *J Surg Res* 49:92–97

22. Förnvik D, Andersson I, Dustler M et al (2013) No evidence for shedding of circulating tumor cells to the peripheral venous blood as a result of mammographic breast compression. *Breast Cancer Res Treat* 141:187–195
23. Peeters DJE, Van den Eynden GG, van Dam PJ et al (2011) Circulating tumour cells in the central and the peripheral venous compartment in patients with metastatic breast cancer. *Br J Cancer* 104:1472–1477
24. Jiao LR, Apostolopoulos C, Jacob J et al (2009) Unique localization of circulating tumor cells in patients with hepatic metastases. *J Clin Oncol* 27:6160–6165
25. Deneve E, Riethdorf S, Ramos J et al (2013) Capture of viable circulating tumor cells in the liver of colorectal cancer patients. *Clin Chem* 59:1384–1392
26. Reddy RM, Murlidhar V, Zhao L et al (2016) Pulmonary venous blood sampling significantly increases the yield of circulating tumor cells in early-stage lung cancer. *J Thorac Cardiovasc Surg* 151:852–857
27. Mizuno N, Kato Y, Izumi Y et al (1998) Importance of hepatic first-pass removal in metastasis of colon carcinoma cells. *J Hepatol* 28:865–877
28. Peeters DJ, Brouwer A, Van den Eynden GG et al (2015) Circulating tumour cells and lung microvascular tumour cell retention in patients with metastatic breast and cervical cancer. *Cancer Lett* 356:872–879
29. Dawson SJ, Tsui DW, Murtaza M et al (2013) Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 368:1199–1209
30. Shaw JA, Guttery DS, Hills A et al (2016) Mutation analysis of cell-free DNA and single circulating tumor cells in metastatic breast cancer patients with high CTC counts. *Clin Cancer Res* 23:88–96
31. Saal LH, Vallon-Christersson J, Hakkinen J et al (2015) The Sweden cancerome analysis network-breast (SCAN-B) initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med* 7:20
32. Rydén L, Loman N, Larsson C et al (2018) Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. *Br J Surg* 105:e158–e168
33. Cristofanilli M, Budd GT, Ellis MJ et al (2004) Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med* 351:781–791
34. Brueffer C, Vallon-Christersson J, Grabau D et al (2018) Clinical value of rna sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden cancerome analysis network—breast initiative. *JCO Precis Oncol* 2:1–18
35. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12:357–360
36. Lai Z, Markovets A, Ahdesmaki M et al (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 44:e108
37. Pedersen BS, Layer RM, Quinlan AR (2016) Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol* 17:118

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



## Paper IV






ORIGINAL ARTICLE



## Subclonal patterns in follow-up of acute myeloid leukemia combining whole exome sequencing and ultrasensitive IBSAFE digital droplet analysis

Louise Pettersson<sup>a,b\*</sup>, Yilun Chen<sup>c\*</sup>, Anthony M. George<sup>c</sup>, Robert Rigo<sup>c</sup>, Vladimir Lazarevic<sup>d</sup>, Gunnar Juliusson<sup>d,e</sup>, Lao H. Saal<sup>f,†</sup>  and Mats Ehinger<sup>b,†</sup>

<sup>a</sup>Department of Pathology, Halland Hospital Halmstad, Region Halland, Halmstad, Sweden; <sup>b</sup>Department of Clinical Sciences, Division of Pathology, Lund University, Skane University Hospital, Lund, Sweden; <sup>c</sup>Department of Clinical Sciences, Division of Oncology, Faculty of Medicine, Lund University, Lund, Sweden; <sup>d</sup>Department of Hematology, Oncology and Radiation Physics, Lund University, Skane University Hospital, Lund, Sweden; <sup>e</sup>Department of Laboratory Medicine, Stem Cell Center, Lund University, Skane University Hospital, Lund, Sweden; <sup>f</sup>Lund University Cancer Center, Medicon Village, Lund, Sweden

### ABSTRACT

We studied mutation kinetics in ten relapsing and four non-relapsing patients with acute myeloid leukemia by whole exome sequencing at diagnosis to identify leukemia-specific mutations and monitored selected mutations at multiple time-points using IBSAFE droplet digital PCR. Five to nine selected mutations could identify and track leukemic clones prior to clinical relapse in 10/10 patients at the time-points where measurable residual disease was negative by multicolor flow cytometry. In the non-relapsing patients, the load of mutations gradually declined in response to different therapeutic strategies. Three distinct patterns of relapse were observed: (1) one or more different clones with all monitored mutations reappearing at relapse; (2) one or more separate clones of which one prevailed at relapse; and (3) persistent clonal hematopoiesis with high variant allele frequency and most mutations present at relapse. These pilot results demonstrate that IBSAFE analyses detect leukemic clones missed by flow cytometry with possible clinical implications.

### HIGHLIGHTS

- The IBSAFE ddPCR MRD method seems applicable on virtually all newly diagnosed AML patients and was more sensitive than flow cytometry.
- Monitoring a few mutations captured the kinetics of the evolving recurrent leukemia.
- *NPM1*-mutation alone may not be a reliable MRD-marker.

### ARTICLE HISTORY

Received 10 January 2020  
Revised 1 April 2020  
Accepted 5 April 2020





### KEYWORDS

AML; MRD; ddPCR; clonal patterns; subclones; genetic evolution

## Introduction


Optimal management of patients with acute myeloid leukemia (AML) depends on accurate monitoring of measurable residual disease (MRD), after treatment. The presence of MRD predicts outcome, and guides treatment decisions [1–3]. Most patients achieve complete remission (CR), but a significant number of patients nevertheless eventually relapse despite having one and often multiple instances of MRD-negativity. Therefore, improvements of MRD determination may optimize treatment and result in more cures. Current MRD methods include multicolor flow

cytometry (MFC), real-time quantitative polymerase chain reaction (qPCR) with or without preceding reverse transcription (RT-qPCR), and more recently next-generation sequencing (NGS) and droplet digital PCR (ddPCR) [1,3–5]. MFC can be applied on most patients but suffers from limited sensitivity, around 0.1% leukemic cells among nucleated bone marrow cells, depending on the phenotypic aberrancies of the leukemic blasts as compared to the background phenotype of normal or regenerating bone marrow cells [6]. RT-qPCR and qPCR are more sensitive than MFC [1,7] but can only be applied on leukemias

**CONTACT** Lao H. Saal  [lao.saal@med.lu.se](mailto:lao.saal@med.lu.se)  Department of Clinical Sciences Lund, Division of Oncology, Lund University Cancer Center, Medicon Village 404-B2, Lund, SE-22381, Sweden; Louise Pettersson  [louise.pettersson@regionhalland.se](mailto:louise.pettersson@regionhalland.se)  Department of Pathology, Halland Hospital Halmstad, Region Halland, SE-30185, Halmstad, Sweden

\*These authors contributed equally.

†These authors share senior authorship.

 Supplemental data for this article can be accessed [here](#).

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

carrying specific fusion genes such as *t(8;21)(q22;q22); RUNX1-RUNX1T1* (8) or a specific mutation such as those occurring in the *nucleophosmin 1 (NPM1)* gene [9]. NGS has tremendously increased our knowledge about the molecular heterogeneity of AML [10–13]. In the MRD setting, NGS has the advantage of tracking several mutations simultaneously and thus being applicable on nearly all AML patients; however, NGS has limited sensitivity and specificity with standard platforms. Computational error-correction and/or utilization of unique molecular indexes help to improve NGS limit of detection (LoD), however the approach is not widely used clinically [14]. Another approach is to use the knowledge of the mutational content of an AML derived from standard NGS (e.g. panel sequencing, whole exome or whole genome sequencing), to choose mutations for follow-up by methods such as ddPCR. This method provides a new way to monitor several mutations simultaneously with higher sensitivity, indeed suitable for MRD assessment [15–17].

The aim of this study was to investigate if pending relapses in AML can be identified by using an ultrasensitive molecular MRD approach targeting several mutations, thereby producing information on multiple putative subclones. We employed IBSAFE, an innovative method using a ddPCR platform with an alternative chemistry that allows for a lower LoD to 0.001% variant allele frequency (VAF) [18–20]. To demonstrate the applicability of IBSAFE for MRD in AML, we analyzed ten relapsing and four non-relapsing AML patients.

## Materials and methods

### Patients and samples

Ten relapsing, defined by standard relapse definitions by the 2016 World Health Organization's (WHO) criteria for AML, and four non-relapsing AML patients were selected and retrospectively tested for molecular MRD in bone marrow (BM) aspirates taken at two to twelve follow-up time-points between 145 and 2607 days after the diagnosis for the relapsing patients, and at three to five time-points between 176 and 895 days after the diagnosis for the non-relapsing patients (Table 1). Time-points were chosen after guidelines and clinical needs, and not by research purpose. End of follow-up was September 2019. The mean time from diagnosis to the first relapse was 498 days (range 145–2054). For the non-relapsing patients, the mean time from diagnosis to end of follow-up was 1630 days (range 1450–1750).

For all patients, the diagnostic and follow-up samples were evaluated with morphology, flow cytometry, qPCR (if a *NPM1* type A mutation was present at diagnosis) and IBSAFE ddPCR (except for patient #1; no immunophenotyping performed on the follow-up samples before relapse). In addition, whole exome sequencing (WES) was performed on all diagnostic samples and all relapse samples with one exception (#8; no WES at relapse). Follow-up samples were collected after two courses of cytoreductive chemotherapy, after completion of therapy, before stem cell transplantation (SCT), at suspicion of relapse and at various additional time-points.

### Whole exome sequencing (WES)

The mutational profile of each leukemia was determined at diagnosis and at first relapse by WES using cultured skin fibroblasts as germline controls as previously described [21]. Cutoff VAF for somatic variants was in general 5% in either the diagnostic or the relapse sample. The assignment of mutations to genes known to be recurrently mutated in AML or non-recurrently mutated genes was in accordance with recurrently mutated genes in The Cancer Genome Atlas Research Network data for AML [11].

### IBSAFE ddPCR

For molecular MRD detection we used the recently developed ultrasensitive mutation detection method IBSAFE, with an effective lower LoD down to approximately 0.001% VAF based on the amount of DNA analyzed per sample [18–20]. In short, IBSAFE marries a two phase chemistry, linear copying and exponential signal generation, within the reaction droplet, thereby greatly enhancing true-positive signals and simultaneously reducing false-positive signals (described in the [Supplementary Methods](#)). An example dilution series from 10% VAF to 0.001% VAF as well as pure wild-type 0% VAF is shown in [Figure 1\(A\)](#).

For the 14 patients, a total of 86 mutations (SNPs and small indels) were selected from WES data, and IBSAFE assays were developed for between 5–9 mutations for each patient. Candidate mutations were selected with priority toward mutations in genes known to be recurrently mutated in AML. In addition, some mutations present at both diagnosis and relapse as determined by WES were chosen. Finally, for a few patients, some mutations present only at relapse were selected to backtrack potential emerging clones.

Table 1. Clinical information at diagnosis, NGS at first relapse, MFC-MRD and MRD for *NPM1* (ddPCR and qPCR) until the first relapse.

Patient	Age at diagnosis/sex	Diagnosis (WHO)	% blasts at diagnosis (morphology)/MFC	Karyotype/ <i>FLT3</i> status at diagnosis	Genetic riskgroup	<i>NPM1</i> status at diagnosis (VAF% by IBSAFE/ MRD% by qPCR for the type A mutation/ VAF% by NGS)	<i>NPM1</i> - mutations at follow-up before first relapse, by qPCR	<i>NPM1</i> - mutations at follow-up before first relapse, by IBSAFE (days after diagnosis, VAF%)	<i>NPM1</i> - mutations at follow-up before first relapse, by qPCR for the type A mutation (days after diagnosis, MRD%)	<i>NPM1</i> status at first relapse (VAF% by IBSAFE/ MRD% by qPCR for the type A mutation/ VAF% by NGS)	MFC-MRD (days after diagnosis, % leukemic cells)	First relapse time point (days after diagnosis)/ SCT (days after diagnosis)	Remarks	Alive† (days after diagnosis)
1	29/M	Acute myelo-blastic leukemia with maturation	58/60	normal/wt	favourable	wt	–	–	–	wt	N.D.	20/54/auto (2160)	Alive	
2	64/F	AML with mutated <i>NPM1</i>	70/73	+8/wt	intermediate	type A (29.3/89/26)	36: N. D. 73: N. D. 233: neg 352: neg 409: neg 498: neg	36: 0.015 73: 0.0057 233: neg 352: neg 409: neg 498: neg	–	wt (neg/neg/neg)	36: <0.1% 73: <0.1% 233: <0.1% 352: <0.1% 409: <0.1% 498: <0.1% 17: N. D. 83: <0.1% 122: pos 0.2% 30: <0.1% 58: <0.1% 80: <0.2% 108: <0.1% 158: <0.1% 341: pos 0.8% 355: pos 0.8%	736/alto (911)	pre leukemic hematopoiesis, relapse <i>NPM1</i> wt	† (1051)
3	55/F	AML with mutated <i>NPM1</i>	60/60	normal//TD	intermediate	type D05 (57.9/ N. D./24)	17: 10.8 83: 0.08 122: 0.3	–	–	type D05 (42.7/ N. D./31)	161/no		† (220)	
4	43/F	AML with mutated <i>NPM1</i>	30/26	normal//TD	intermediate	type A (26.8/ 79.4/25)	30: 0.02 58: neg	30: 0.013 58: 0.0028	–	type A (32.3/ 57.0/28) wt	241/no 388/alto (453)		† (389)	
5	41/M	AML with inv(16)	48/48	inv(16)/wt	favourable	wt	–	–	–	wt			Alive	
6	61/M	AML with MDS-related changes	21/21	complex/wt	adverse	wt	–	–	–	wt	17: 5% 32: <0.1% 40: <0.1% 90: <0.1% 109: <0.01% 136: <0.1% 71: <0.1% 119: <0.1% 191: <0.1% 93: pos 0.2-0.3% 133: <0.1%	189/no	hemophagocytosis at diagnosis remission. monitored with RT-qPCR for inv(16), molecular relapse +22.	† (191)
7	60/F	AML with mutated <i>NPM1</i>	19/20	normal/wt	favourable	type A (31.6/60/31)	71: neg 119: neg 191: neg	71: pos <0.0024 119: pos <0.0024 191: 0.011	–	type A (9.5/12/3.7)	339/alto (438)		† (1067)	
8	37/F	AML with t(3,3), MECOM+	82/86	complex/wt	adverse	wt	–	–	–	wt	179/alto (114)		† (251)	
9	48/F	AML with mutated <i>NPM1</i>	60/57	normal/wt	favourable	type A (40.3/ 47.9/17)	29: 0.08 63: 0.006 102: 0.03	29: 0.059 63: 0.021 102: 0.021 149: pos <0.0024	–	type A (40.3/ 33.2/26) type D (43.6/ N. D./18)	553/alto (647)		† (927)	
10	71/F	AML with mutated <i>NPM1</i>	34/12	normal//TD	intermediate	type D (37.8/ N. D./31)	63: 0.08	–	–	–	145/no		† (304)	
11	66/M		92/80	normal//KO	favourable					–	no/no		Alive	

(continued)



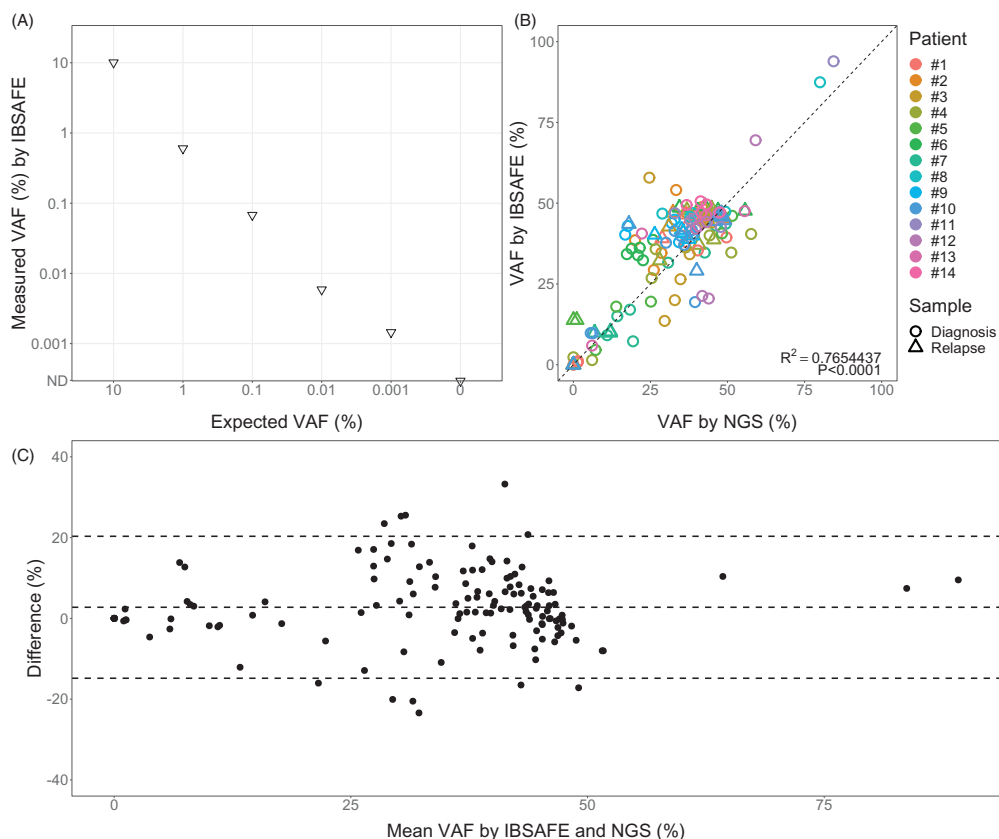


Table 1. Continued.

Patient	Age at diagnosis/sex	Diagnosis (WHO)	% blasts at diagnosis (morphology/MFC)	Karyotype/ <i>FLT3</i> status at diagnosis	Genetic riskgroup	<i>NPM1</i> status at diagnosis (VAF% by IBSAFE/MRD% by qPCR for the type A mutation/VAF% by NGS)	<i>NPM1</i> - mutations at follow-up before first relapse, by qPCR for the type A mutation (days after diagnosis, MRD%)	<i>NPM1</i> - mutations at follow-up by IBSAFE (days after diagnosis, VAF%)	<i>NPM1</i> status at first relapse (VAF% by IBSAFE/MRD% by qPCR for the type A mutation/VAF% by NGS)	MFC-MRD (days after diagnosis, % leukemic cells)	First relapse time point (days after diagnosis)/ SCT (days after diagnosis)	Remarks	Alive/t (days after diagnosis)
AML with mutated <i>NPM1</i>													
12	47/F	AML with t(821)	33/16	t(821) + additional chromosome changes/wt	favourable	type A	14: N, D.*	14: N, D.*	14: N, D.*	14: <0.1%	no/no	Alive	
						(47/62/33)	28: 0.006	28: 0.0045	28: 0.0045	28: <0.1%			
						wt	73: neg	73: neg	73: neg	73: <0.1%			
13	71/M	Acute myelo-monocytic leukemia	70/77	trisomy 13/wt	adverse	wt	–	–	–	176: neg	no/allo (189)	Alive	
						wt	–	–	–	176: neg			
						wt	–	–	–	176: neg			
14	51/F	Acute myelo-blastic leukemia with minimal differentiation	90/90	normal/wt	intermediate	wt	–	–	–	–	no/allo (58)	Alive	
						wt	–	–	–	–			
						wt	–	–	–	–			

WHO: World Health Organization classification; MFC: multicolor flow cytometry; *FLT3*: *FMS-like tyrosine kinase 3*; *NPM1*: *nucleophosmin 1* gene; VAF: variant allele frequency; IBSAFE: in house digital droplet polymerase chain reaction; MRD%: %leukemic cells; qPCR: real-time quantitative polymerase chain reaction; NGS: next generation sequencing; MRD: measurable residual disease; SCT: stem cell transplantation; t: dead; M: male; F: female; AML: acute myeloid leukemia; inv: inversion; MDS: myelodysplastic syndrome; MECOM+: "MDS1 and EVI1 complex locus"; wt: wildtype; ITD: internal tandem duplication; AKD: activated kinase domain; N/D: not determined; neg: negative; pos: positive; auto: autologous stem cell transplantation; allo: allogeneic stem cell transplantation; RT-qPCR: reverse transcriptase PCR; molecular remission: morphological remission and two negative MRD-samples.

\*Mutational status not determined due to insufficient DNA.



**Figure 1.** Performance of IBSAFE and comparison of IBSAFE and WES measured variant allele frequencies (VAFs) in diagnosis and relapse samples. (A) Dilution series for IBSAFE assay *NPM1* type A for constructed samples with known VAFs at 10%, 1%, 0.1%, 0.01%, 0.001%, and 0%. (B) Scatterplot for agreement between the methods with Pearson's coefficient of determination  $R^2=0.77$  and  $p$ -value  $<.0001$ ,  $N=144$ , (C) Bland Altman plot.

Each IBSAFE assay was confirmed to have zero false-positive droplets using at least 59,000 haploid genome copies (180 ng) of negative control DNA (Promega), demonstrating an assay LoD of at least 0.0017% VAF. An example assay is shown in [Supplementary Figure 1](#). IBSAFE analyses were performed on all diagnostic, follow-up, and relapse samples, using 60 ng of DNA per reaction and each reaction performed in duplicate thus enabling an effective LoD down to 0.003% VAF. In every IBSAFE run, positive (diagnostic or in a few cases relapse sample) and negative control (human male normal) DNA were used and confirmed test reliability.

### Flow cytometry and qPCR for *NPM1* type a mutations

Immunophenotyping and quantification of *NPM1* type A mutations with qPCR was performed as previously described [7].

### Statistical analyses

To investigate the correlations between measuring the VAF of the mutations in the diagnostic and relapse samples with IBSAFE, WES and qPCR, Pearson's correlation test and Bland-Altman plots were applied [22].

**Table 2.** Number of mutations at diagnosis and first relapse.

Patient	Number of genes known to be recurrently mutated in AML at diagnosis	Number of non-recurrently mutated genes in AML at diagnosis	Number of recurrently mutated genes at first relapse (lost/new)	Number of non-recurrently mutated genes at first relapse (lost/new)
1	5	9	5 (1 / 1)	10 (2 / 3)
2	6	24	5 (2 / 1)	12 (17 / 5)
3	2	15	2 <sup>a</sup> (0 / ND <sup>a</sup> )	7 <sup>a</sup> (8 / ND <sup>a</sup> )
4	3	13	4 (1 / 2)	13 (6 / 6)
5	2	11	0 <sup>a</sup> (2 / ND <sup>a</sup> )	8 <sup>a</sup> (3 / ND <sup>a</sup> )
6	1	14	1 (0 / 0)	15 (1 / 2)
7	7	10	2 (5 / 0)	28 (4 / 22)
8	0	10	ND <sup>b</sup>	ND <sup>b</sup>
9	4	9	4 (0 / 0)	15 (0 / 6)
10	3	14	3 (0 / ND <sup>a</sup> )	10 (4 / ND <sup>a</sup> )
11	4	7	—	—
12	0	24	—	—
13	6	21	—	—
14	4	18	—	—

ND: not determined.

<sup>a</sup>NGS analysis not possible to interpret for new mutations due to a high background noise.<sup>b</sup>No NGS analysis performed on the relapse sample because germline variants could not be determined after allo-SCT before the relapse.

## Results

### Diagnostic and monitored mutations

WES of the BM from 10 relapsing and four non-relapsing patients identified 10–30 somatic mutations at diagnosis (mean 18) in each patient, of which 0–7 (mean 3) were in genes known to be recurrently mutated in AML (Table 2). In total 12 of the patients had mutations in genes known to be recurrently mutated in AML. For 11 of the patients at least one of these recurrent mutations could be monitored by IBSAFE. In six of the relapsing patients, WES of the relapse sample detected between 2–22 new mutations (Table 2). For one patient, WES was not performed at relapse after allogeneic-SCT (allo-SCT) since the presence of donor cells prohibited detection of new mutations. For the remaining three patients, no new mutations could be identified due to poor sequence quality and therefore the WES data was only used to confirm presence of mutations identified at diagnosis.

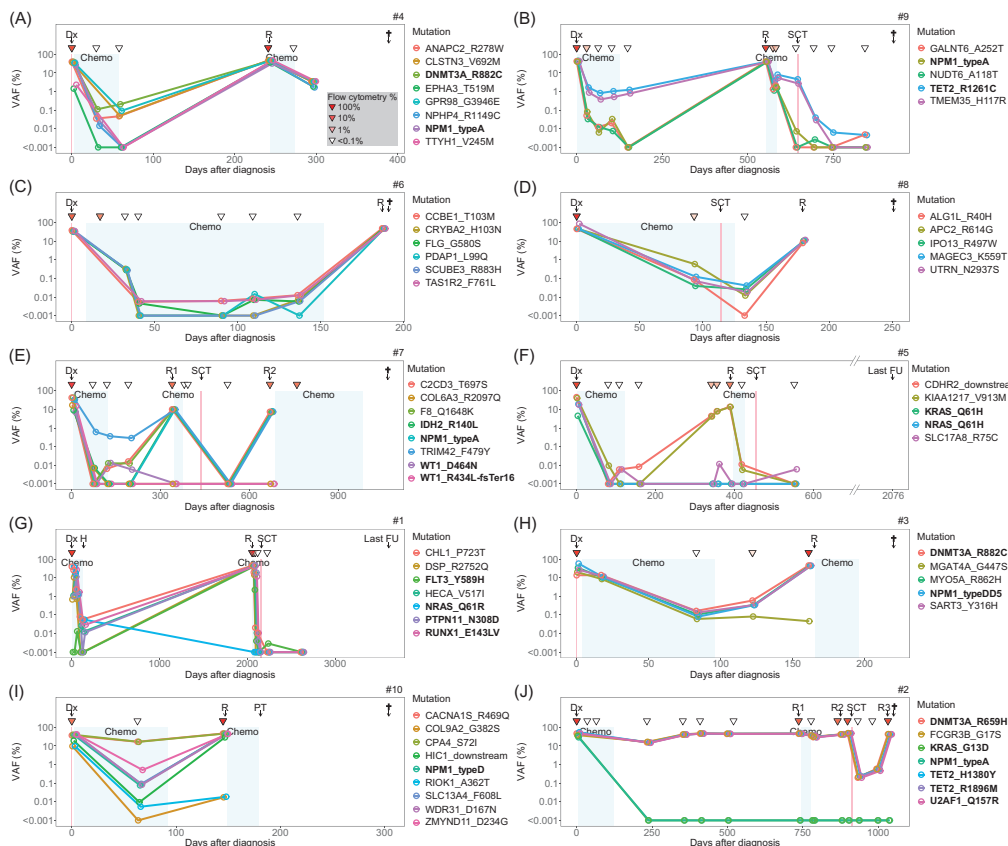
On the basis of the WES results, 86 mutations, 81 of which were unique (*NPM1* type A mutation was monitored in five patients, *DNMT3A* R882C in two), were monitored (Supplementary Table 1). Of these, 31 were mutations in genes known to be recurrently mutated in AML, including seven *NPM1* mutations. Four mutations in genes known to be recurrently mutated in AML were identified by WES at relapse but absent at diagnosis in three patients (*RUNX1*, *IDH1* and two different *FLT3* mutations). Of these, *FLT3* Y589H was monitored. Of note, scatterplot and Bland–Altman plot of VAFs as measured by IBSAFE and WES on the diagnostic and relapse samples displayed excellent

agreement across a range of allele frequencies and considering the WES sequencing depth ( $R^2=0.77$ ,  $N=144$ ,  $p<.0001$ ; Figure 1(B,C)). A comparison of qPCR and IBSAFE data for the *NPM1* type A mutation is shown in Supplementary Figure 2.

For all 10 relapsing patients, IBSAFE analysis revealed molecular evidence of persisting or emerging mutations at time-points prior to the clinical relapse (Figure 2). Of all 66 follow-up time-points tested across the 10 patients, 35 time-points from 9 patients (2–7 time-points per patient, or 50–100% of all time-points tested for the patient) exhibited at least one IBSAFE-detected mutation detected at VAFs between 0.1% and 0.003%. In addition, all relapsing patients had at least one follow-up time-point with VAF >0.1% for at least one mutation detected before clinical relapse. Moreover, IBSAFE-based molecular MRD was more sensitive to identify residual disease as compared to MFC, with no time-point being MFC-positive and IBSAFE-negative (Supplementary Table 2).

### Patterns of mutations for relapsing patients

Interestingly, distinct patterns of emerging and retreating mutations could be discerned from the IBSAFE results. In Pattern 1, featuring four patients, one or several clones were apparent during follow-up with all the monitored mutations reappearing at relapse (Figure 2(A–D); Table 2). Some mutations were undetectable at certain time-points, whereas others were present at all time-points at low levels, possibly representing minor pre-leukemic clones. Two of these patients (#4 and #9) displayed mutations in genes

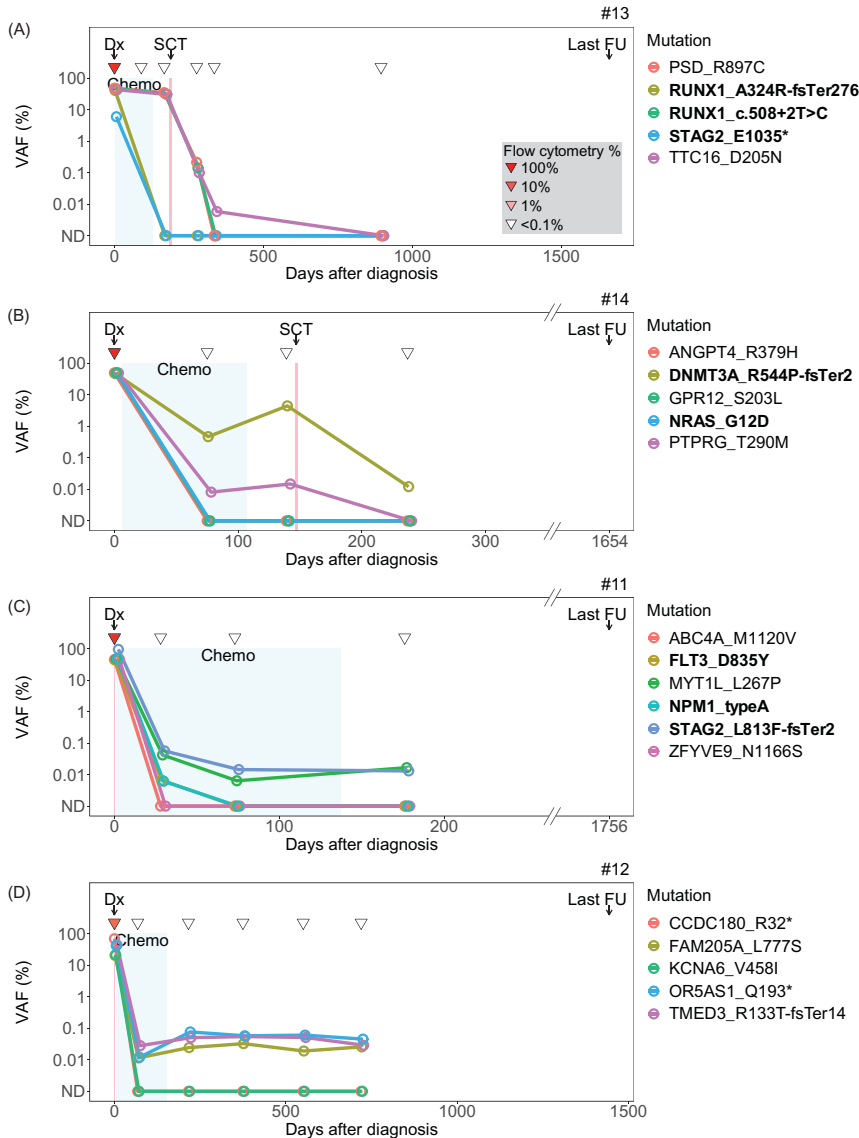


**Figure 2.** Monitoring leukemic mutations using ultrasensitive IBSAFE for ten relapsing AML patients (A) #4, (B) #9, (C) #6, (D) #8, (E) #7, (F) #5, (G) #1, (H) #3, (I) #10 and (J) #2. In each plot, the y-axis represents the detected variant allele frequency (VAF %) for each tracked mutation (key to the right; genes known to be recurrently mutated in AML in bold), and the x-axis indicates the days after diagnosis with therapies indicated by shading (chemotherapy) and clinical events indicated along the top of each plot. The inverted triangles indicate the flow cytometry MRD results, with the color-key indicated in the lower-right of plot (A). Dx: diagnosis; R: relapse; †: dead; SCT: stem cell transplantation; FU: follow-up; H: harvest; ND: not detected (VAF below lower effective limit of detection of 0.003% determined by input DNA quantity).

known to be recurrently mutated in AML (*DNMT3A*, *NPM1* or *TET2*) (Figure 2(A,B)). In these two patients, selection of clones containing mutations in *DNMT3A*, *ANAPC2*, *CLSTN3* and *GPR98* (Figure 2(A)) and *TET2* and *TMEM35* (Figure 2(B)), respectively, was evident after induction therapy. For patient #4, a subclone containing two additional mutations (*EPHA3* and *TTYH1*) at VAFs 1–2% was present at diagnosis (Figure 2(A)). Except for patient #8 (Figure 2(D)), MFC-MRD was not detected before the relapse in this group. This patient had a positive MFC-MRD at day 92 (0.2–0.3%), 18 days before the SCT with all monitored mutations detectable. After SCT, MFC-MRD was negative but 4/5

mutations were still detectable, and the patient soon exhibited clinical relapse about two months later.

In Pattern 2 containing four patients (Figure 2(E–H)), the emerging relapsing leukemia carried only some of the mutations monitored, both in recurrently as well as non-recurrently mutated genes (Table 2). All patients in this group displayed at least two mutations in known AML-associated genes including *NPM1*, *IDH2*, *WT1*, *NRAS*, *KRAS*, *FLT3*, *PTPN11*, *RUNX1*, or *DNMT3A*. After three courses of chemotherapy, at least one mutation was detectable by IBSAFE in all four patients despite negative MFC-MRD when available, with *TRIM42* in patient #7



**Figure 3.** Monitoring leukemic mutations using ultrasensitive IBSAFE for four non-relapsing AML patients (A) #13, (B) #14, (C) #11 and (D) #12. In each plot, the y-axis represents the detected variant allele frequency (VAF %) for each tracked mutation (key to right; genes known to be recurrently mutated in AML in bold), and the x-axis indicates the days after diagnosis with therapies indicated by shading (chemotherapy) and clinical events indicated along the top of each plot. The inverted triangles indicate the flow cytometry MRD results, with the color-key indicated in the lower-right of plot (A). Dx: diagnosis; FU: follow-up; SCT: stem cell transplantation; ND: not detected (VAF below 0.003%).

consistently higher than 0.1% VAF. MFC-MRD was positive (>0.1%) in two of the patients after completion of chemotherapy with concomitant VAF >0.1% for several mutations (patient #5, Figure 2(F); patient

#3, Figure 2(H)). At least one clone disappeared in all four patients in response to therapy as evidenced by the diminishing or undetectable amounts of mutant DNA.

In Pattern 3 with the remaining two relapsing patients (Figure 2(I,J)), there was no distinct decrease of the mutation allele frequencies before SCT for two or more of the mutations despite morphological remission and negative MFC-MRD, demonstrating the pre-leukemic nature of the regenerating hematopoiesis. In patient #10 (Figure 2(I)), nine mutations were monitored. Two mutations were detected at 17% VAF (*CACNA9* and *CPA4*) after induction therapy, preceding the relapse about two months later, after completion of therapy despite MFC-MRD negativity <0.1% and morphological remission. Patient #2 (Figure 2(J)) showed clonal hematopoiesis and an *U2AF1* mutation with high residual VAF, apparently unresponsive to chemotherapy, but in complete morphological and immunophenotypical (<0.1%) remission until the relapse at day 736. At the first relapse, both the mutational profile (loss of 19 mutations and gain of six new mutations by WES) and the immunophenotype changed significantly (Table 2). Only after the second relapse (150 days after the first relapse) treated with SCT, a significant decrease of the mutational load for the monitored mutations was seen, but the VAFs of monitored mutations never fell below 0.2%.

### **Patterns of mutations for patients in complete remission**

Among the non-relapsing patients (Figure 3(A–D)), patients #13 and #14 underwent allo-SCT and displayed a similar pattern of monitored mutations before transplantation with some persistent mutations with high VAF despite morphological and immunophenotypical (<0.1%) remission, suggesting pre-leukemic hematopoiesis refractory to conventional cytoreductive therapy (Figure 3(A,B)). This pattern was reminiscent of that of the relapsing patients #10 and #2 (Figure 2(I,J)) described above. After allo-SCT the mutations gradually disappeared and were unmeasurable at the last follow-up time-point except for 0.01% VAF of the *DNMT3A* mutation in patient #14 (Figure 3(B)). In patient #11 and #12 all monitored mutations declined after conventional cytoreductive therapy; some of them disappeared completely whereas others seemed to stabilize at low VAF levels between 0.01% and 0.08% (Figure 3(C,D)).

### ***NPM1*-mutations in relapsing and non-relapsing patients**

Six out of ten patients in the relapsing group had an *NPM1* mutation (type A in patients #2, #4, #7 and #9;

type D in #10; and type DD5 in #3). All of these but patient #2 experienced an *NPM1*-positive relapse. For patient #2 the *NPM1* type A mutation was undetectable by IBSAFE and qPCR at all follow-up time-points after day 73, including the complete morphological and immunophenotypical remission time-points and the relapse time-points (Figure 2(J)) and (Table 1). Patient #7 (in complete morphological and immunophenotypical remission after completion of therapy) had no detectable *NPM1* type A mutation by IBSAFE at any time-point before the *NPM1* positive relapse (Figure 2(E)) but a quantifiable signal by qPCR at the last time-point (0.011%) and a detectable but not quantifiable signal for the remaining two time-points. Patient #4 and #9 (Figure 2(A,B)) had detectable *NPM1* mutations by IBSAFE as well as qPCR at all follow-up time-points except the last time-point, when it was undetectable by IBSAFE and weakly positive (0.0028%, #4) or detectable but not quantifiable (<0.0024%, #9) by qPCR. Both these patients were in complete morphological and immunophenotypical remission after completion of therapy. For the type D and DD5 *NPM1* mutation-positive patients (#10 and #3) no qPCR data exist. For the type D *NPM1* mutation-positive patient (#10) only one follow-up sample exists between diagnosis and relapse, day 63. At this time-point MFC-MRD was negative, but IBSAFE-MRD was positive (0.08% VAF). For the type DD5 *NPM1* mutation-positive patient (#3), three time-points between the diagnosis and relapse were tested. IBSAFE detected MRD in all these three time-points, whereas MFC was only positive at the last time-point before the relapse.

In the non-relapsing group, one patient (#11) had a *NPM1* mutation (type A; Table 1) detectable at low VAF <0.1% with IBSAFE and qPCR after induction therapy, day 28, that disappeared during follow-up (Figure 3(C)). MFC-MRD was negative at all time-points (Figure 3(C) and Table 1).

### **Discussion**

Because of the clonal complexity of AML, the assessment of MRD in routine practice is difficult. Minor subclones present at diagnosis may evolve and escape detection by MFC or targeted qPCR-MRD. The aim of this study was to investigate if relapses in AML can be identified and predicted by using a sensitive molecular MRD approach (IBSAFE) targeting several mutations, thereby producing information on multiple putative subclones. Our proof-of-concept study demonstrates the ability of the IBSAFE method to do this. For all ten relapsing patients, a few selected mutations were able

to track early recurrence of leukemic clones. In a clinical routine setting, the most commonly used MRD-method is MFC except for RNA based methods for *NPM1* mutated leukemias and for some fusion genes [23]. Currently, a cutoff of 0.1% residual leukemic blasts after two courses of therapy, as determined by MFC, has important implications for risk stratification and therapy decisions [24]. Three of the relapsing patients in this study (#3, #5 and #8) displayed MFC-MRD levels >0.1% after completion of therapy. In concordance, these three patients also had persisting mutations >0.1% VAF as determined by IBSAFE. IBSAFE also showed a better sensitivity than standard MFC-MRD for all patients in this study. Among the ten relapsing patients, MFC-MRD was positive at 4/30 follow-up time-points (from day 25 until the last time-point before the first relapse) while IBSAFE-MRD was detectable for at least one mutation at 31/31 follow-up time-points (Figure 2). These results are important, because many patients eventually relapse despite MFC-MRD levels below 0.1%, emphasizing the problem of false-negative MFC-MRD and therefore the need for new MRD-strategies to more accurately predict recurrence.

For comparison, we also monitored four non-relapsing patients without clinical, morphological or immunophenotypical signs of residual disease. Two of these patients showed a persisting but stabilized subclone with low VAF at the last follow-up (Figure 3(C,D)) of which one patient (Figure 3(D)) exclusively in non-recurrently mutated genes. The significance of the persisting mutations in the non-relapsing patients is unclear. They could represent pre-leukemic stem cells or progenitors that have not yet acquired all mutations needed for progression, or sub-clinical leukemia where overt relapse has not yet occurred. Alternatively, they are passenger mutations of clonal hematopoiesis [25].

Our approach allowed for the deciphering of three different mutational patterns in the follow-up samples from the relapsing patients. In Pattern 3 ( $N=2$ ; Figure 2(I,J)), no distinct decrease of the VAF before relapse was observed for two or more of the mutations despite morphological remission and negative MFC-MRD, indicating the pre-leukemic nature of the regenerating hematopoiesis. It is tempting to speculate that patients with this pattern are at risk for relapse. However, the presence of some of these mutations do not necessarily signalize impending relapse. For example, *DNMT3A* and *TET2* mutations, as in patient #2, can persist at variable VAF levels in the follow-up samples in patients in CR without impact on

prediction of relapse [24,26]. Moreover, it is well known that the number of somatic mutations rises with increasing age. In some individuals, these mutations form clonal hematopoiesis that arises when a single hematopoietic stem cell contributes disproportionately to the population of mature blood cells [27,28]. The presence of clonal hematopoiesis is associated with an increased risk of developing a myeloid neoplasm, but the vast majority of individuals with age-related clonal hematopoiesis do not develop AML [29–33]. Such clonal mutations are sometimes called pre-leukemic and include genes such as *DNMT3A*, *TET2*, *IDH1/2*, *ASXL1*, and *IKZF1* (29). Other recurrent mutations such as those occurring in spliceosome genes (e.g. *U2AF1*; patient #2) are more often predictable of evolution to AML, [34]. Hence, with respect to MRD-assessment, the presence of mutations such as those in *DNMT3A*, *ASXL1*, and *TET2* are often of limited value for prediction of relapse. Nevertheless, even after exclusion of *DNMT3A* and *TET2* mutations, MRD was detectable at all follow-up time-points. Likely, both the nature (e.g. preleukemic versus non-preleukemic mutations), and the kinetics of the monitored mutations are important biological determinants for reemerging AML [35].

In 2011 Krönke *et al.* showed that 9% of all *NPM1* mutated patients at diagnosis had no detected transcript levels at relapse [36]. In a recently published study from the Munich Leukemia Laboratory, 13% of leukemias harboring a *NPM1*-mutation at diagnosis relapsed with a *NPM1*-wildtype leukemia [37]. More recent studies have also indicated that markers such as mutated *NPM1* may not always be stable over the course of disease and that relapses sometimes emanate from *NPM1*-wildtype clones [29,32]. In a recently published case report, featuring a case resembling our patient #2 (Figure 2(J)), the authors describe loss of the *NPM1* mutation found at diagnosis, a persisting *DNMT3A* mutation, and a late relapse [38].

Our results support the limitations of employing *NPM1* mutation as the sole marker of disease. In two of our six *NPM1* positive relapsing patients (33%) the *NPM1* mutation was not a reliable marker of residual disease. Patient #7, (Figure 2(E)) had *NPM1* negative MRD and patient #2 (Figure 2(J)) had a *NPM1* negative relapse.

The difference between the results from qPCR and IBSAFE might at least partly be explained by the amount of input DNA (600–1000 ng/test compared to 120 ng/test). It is important to point out that although several mutations were monitored for each patient, there may have been additional relevant subclones to

follow. In addition, acquired new mutations after therapy will not be detected with the approach of the present study. Indeed, backtracking of mutations from the relapse of two patients, #1 and #4 showed emerging clones containing these mutations (Figure 2(A,G)). Nevertheless, the pattern of persisting and emerging clones in the relapse group suggest that it may suffice to select a limited number of mutations for powerful MRD-assessment.

In conclusion, this pilot study demonstrates the feasibility of the IBSAFE method to measure MRD with high sensitivity and on essentially any newly diagnosed adult with AML where there are no fusion genes that are recommended for MRD follow-up. The method allows for a lower LoD to 0.003% VAF, based on available input DNA, to follow several mutations and track different emerging clones. Developed IBSAFE assays can rapidly be applied on follow-up samples and easily utilized for other patients carrying the same mutation. In addition to the established recurrent mutations, personalized assays (due to the mutational heterogeneity of AML) may also be developed for individual AML patients based on their specific mutational profiles. The prognostic relevance of such monitoring should be evaluated in large prospective studies.

## Acknowledgements

The authors would like to thank Christina Orsmark-Pietras, Henrik Lilljebjörn and Thoas Fioretos for performing the sequencing analyses and analyzing the WES data. The authors would also like to thank Li Zhou, Heike Kotarsky, Kerstin Torikka and Marianne Rissler for excellent technical support.

## Author contributions

LP, YC, ME and LHS conceived the study. LP and ME provided bone marrow samples. GJ and VL provided clinical information. LP and ME analyzed the MFC-MRD and qPCR data. LP, YC, RR, AMG, and LHS performed IBSAFE analyses. LP, YC, ME and LHS wrote the report. All authors approved the final version of the report for publication.

## Disclosure statement

YC, AMG, RR, and LHS have ownership interest (including stock, patents, etc.) in SAGA Diagnostics AB. LP, VL, GJ and ME report no conflict of interest.

## Funding

This work was supported by the Olle Engkvist Foundation, Skåne University Hospital Research Grants, Region Skåne

UFo Grants, Swedish Cancer Society, Swedish Research Council, VINNOVA, Governmental Funding of Clinical Research within National Health Service, Lund University Medical Faculty, Gunnar Nilsson Cancer Foundation, Mrs. Berta Kamprad Foundation, and the Krapperup Foundation. LP was supported by the Regional Scientific Council of Halland.

## ORCID

Lao H. Saal  <http://orcid.org/0000-0002-0815-1896>

## References

- [1] Grimwade D, Freeman SD. Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for "prime time"? *Blood*. 2014;124(23):3345–3355.
- [2] Ivey A, Hills RK, Simpson MA, et al. Assessment of minimal residual disease in standard-risk AML. *N Engl J Med*. 2016;374(5):422–433.
- [3] Kayser S, Walter RB, Stock W, et al. Minimal residual disease in acute myeloid leukemia—current status and future perspectives. *Curr Hematol Malig Rep*. 2015;10(2):132–144.
- [4] Ommen HB. Monitoring minimal residual disease in acute myeloid leukaemia: a review of the current evolving strategies. *Ther Adv Hematol*. 2016;7(1):3–16. Feb
- [5] Sung PJ, Luger SM. Minimal residual disease in acute myeloid leukemia. *Curr Treat Options Oncol*. 2017;18(1):1.
- [6] Jaso JM, Wang SA, Jorgensen JL, et al. Multi-color flow cytometric immunophenotyping for detection of minimal residual disease in AML: past, present and future. *Bone Marrow Transplant*. 2014;49(9):1129–1138.
- [7] Pettersson L, Leveen P, Axler O, et al. Improved minimal residual disease detection by targeted quantitative polymerase chain reaction in Nucleophosmin 1 type a mutated acute myeloid leukemia. *Genes Chromosomes Cancer*. 2016;55(10):750–766. Oct
- [8] Leroy H, de Botton S, Gardel-Duflos N, et al. Prognostic value of real-time quantitative PCR (RQ-PCR) in AML with t(8;21). *Leukemia*. 2005;19(3):367–372. Mar
- [9] Falini B, Mecucci C, Tiacci E, et al. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med*. 2005;352(3):254–266.
- [10] Bullinger L, Dohner K, Dohner H. Genomics of acute myeloid leukemia diagnosis and pathways. *JCO*. 2017;35(9):934–946.
- [11] Ley TJ, Cancer Genome Atlas Research Network, Miller C, Ding L, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059–2074.
- [12] Grimwade D, Ivey A, Huntly BJ. Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance. *Blood*. 2016;127(1):29–41.



- [13] Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. *N Engl J Med*. 2016;374(23):2209–2221.
- [14] Roloff GW, Lai C, Hourigan CS, et al. Technical advances in the measurement of residual disease in acute myeloid leukemia. *J Clin Med*. 2017;6(9):pii: E87.
- [15] Cruz NM, Mencia-Trinchant N, Hassane DC, et al. Minimal residual disease in acute myelogenous leukemia. *Int J Lab Hem*. 2017;39(Suppl 1):53–60.
- [16] Mencia-Trinchant N, Hu Y, Alas MA, et al. Minimal residual disease monitoring of acute myeloid leukemia by massively multiplex digital PCR in patients with NPM1 mutations. *J Mol Diagn*. 2017;19(4):537–548.
- [17] Wertheim GBW, Bagg A. NPM1 for MRD? Droplet like it's hot! *J Mol Diagn*. 2017;19(4):498–501.
- [18] Arildsen NS, Martin de la Fuente L, Masback A, et al. Detecting TP53 mutations in diagnostic and archival liquid-based Pap samples from ovarian cancer patients using an ultra-sensitive ddPCR method. *Sci Rep*. 2019;9(1):15506.
- [19] Fornvik D, Aaltonen KE, Chen Y, et al. Detection of circulating tumor cells and circulating tumor DNA before and after mammographic breast compression in a cohort of breast cancer patients scheduled for neoadjuvant treatment. *Breast Cancer Res Treat*. 2019;177(2):447–455.
- [20] Isaksson S, George AM, Jonsson M, et al. Pre-operative plasma cell-free circulating tumor DNA and serum protein tumor markers as predictors of lung adenocarcinoma recurrence. *Acta Oncologica*. 2019;58(8):1079–1086.
- [21] Lazarevic V, Orsmark-Pietras C, Lilljebjorn H, et al. Isolated myelosarcoma is characterized by recurrent NFE2 mutations and concurrent preleukemic clones in the bone marrow. *Blood*. 2018;131(5):577–581.
- [22] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327(8476):307–310.
- [23] Ehinger M, Pettersson L. Measurable residual disease testing for personalized treatment of acute myeloid leukemia. *APMIS*. 2019;127(5):337–351.
- [24] Schuurhuis GJ, Heuser M, Freeman S, et al. Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. *Blood*. 2018;131(12):1275–1291.
- [25] Parkin B, Londono-Joshi A, Kang Q, et al. Ultrasensitive mutation detection identifies rare residual cells causing acute myelogenous leukemia relapse. *J Clin Investig*. 2017;127(9):3484–3495.
- [26] Jongen-Lavrencic M, Grob T, Hanekamp D, et al. Molecular minimal residual disease in acute myeloid leukemia. *N Engl J Med*. 2018;378(13):1189–1199.
- [27] Steensma DP, Bejar R, Jaiswal S, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood*. 2015;126(1):9–16.
- [28] Zink F, Stacey SN, Norddahl GL, et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*. 2017;130(6):742–752.
- [29] Corces-Zimmerman MR, Hong WJ, Weissman IL, et al. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *PNAS*. 2014;111(7):2548–2553.
- [30] Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med*. 2014;371(26):2488–2498.
- [31] Shlush LI, Mitchell A, Heisler L, et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*. 2017;547(7661):104–108.
- [32] Shlush LI, Zandi S, Mitchell A, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*. 2014;506(7488):328–333.
- [33] Sykes SM, Kokkalis KD, Millsom MD, et al. Clonal evolution of preleukemic hematopoietic stem cells in acute myeloid leukemia. *Exp Hematol*. 2015;43(12):989–992.
- [34] Abelson S, Collord G, Ng SWK, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*. 2018;559(7714):400–404.
- [35] Rothenberg-Thurley M, Amler S, Goerlich D, et al. Persistence of pre-leukemic clones during first remission and risk of relapse in acute myeloid leukemia. *Leukemia*. 2018;32(7):1598–1608.
- [36] Kronke J, Schlenk RF, Jensen KO, et al. Monitoring of minimal residual disease in NPM1-mutated acute myeloid leukemia: a study from the German-Austrian acute myeloid leukemia study group. *J Clin Oncol*. 2011;29(19):2709–2716.
- [37] Hollein A, Meggendorfer M, Dicker F, et al. NPM1 mutated AML can relapse with wild-type NPM1: persistent clonal hematopoiesis can drive relapse. *Blood Adv*. 2018;2(22):3118–3125.
- [38] Bacher U, Porret N, Joncourt R, et al. Pitfalls in the molecular follow up of NPM1 mutant acute myeloid leukemia. *Haematologica*. 2018;103(10):e486–e8.