Analysis of Circulating Biomarkers for Minimally Invasive Early Detection of Breast Cancer

UNIVERSITÀ
DI TORINO

Emir Šehović
PhD Thesis

Department of Life Sciences and Systems Biology
University of Turin

Supervisor:
Dr. Giovanna Chiorino

Co-supervisors:
Prof. Michele De Bortoli
Prof. Jaakko Kaprio

February 2024

ANALYSIS OF CIRCULATING BIOMARKERS FOR MINIMALLY INVASIVE EARLY DETECTION OF BREAST CANCER

Emir Šehović

Dissertation Submitted in Fulfilment of the Requirements for the
Degree of Doctor of Philosophy

To the PhD school
Complex Systems for Quantitative Biomedicine (35th cycle)
at the department of
Life Sciences and Systems Biology

Dissertation was externally reviewed by two referees:

1. Professor Marike Gabrielson – Department of Medical Epidemiology and Biostatistics at Karolinska Institutet, Stockholm

2. Doctor Giovanna Ambrosini – School of Life Sciences, Philipp Bucher Group, École Polythechnique Fédérale de Lausanne, Lausanne

University of Turin
February 2024

In the name of God
For humanity

Abstract

Breast cancer (BC) is the malignancy with the highest incidence and mortality rates among women. Numerous studies explored cell-free circulating (cfc) microRNAs (miRNAs) as diagnostic biomarkers of BC. However, their results were inconsistent with few intersecting miRNA panels. In a meta-analysis, we evaluated the overall diagnostic performance as well as the sources of heterogeneity between studies on BC detection using cfc miRNA. The findings on sources of heterogeneity would then be applied to our second project, which aimed to identify circulating miRNA ratios associated with BC in women attending mammography screening.

On 56 studies that investigated diagnostic circulating miRNAs by utilising Real-Time Quantitative Reverse Transcription PCR (RT-qPCR), pooled sensitivity and specificity of 0.85 [0.81 to 0.88] and 0.83 [0.79 to 0.87] were obtained, respectively. Subgroup analysis revealed a comparable pooled diagnostic performance between studies using serum (sensitivity: 0.87 [0.81 to 0.91]; specificity: 0.83 [0.78 to 0.87]) and plasma (sensitivity: 0.83 [0.77 to 0.87]; specificity: 0.85 [0.78 to 0.91]) as specimen type. Additionally, miRNA(s) based on endogenous normalisers tend to have a higher diagnostic performance than miRNA(s) based on exogenous ones.

A nested case–control study was conducted on plasma samples of 65 cases and 66 controls (discovery) and 32 cases and 127 controls (validation). Small-RNA sequencing was carried out on the discovery cohort, and to overcome the normalisation issue in RT-qPCR, we computed miRNA ratios and those associated with BC were selected by two-sample Wilcoxon test and lasso penalised logistic regression. Assessment by RT-qPCR of 20 candidate miRNA ratios was carried out as a platform validation. To identify the most promising biomarkers, penalised logistic regression was further applied to candidate miRNA ratios alone or in combination with non-molecular factors. In the resulting model, LASSO regression selected seven miRNA ratios (miR-199a-3p_let-7a-5p, miR-26b-5p_miR-142-5p, let-7b-5p_miR-19b-3p, miR-101-3p_miR-19b-3p, miR-93-5p_miR-19b-3p, let-7a-5p_miR-22-3p and miR-21-5p_miR-23a-3p), together with the interaction term of centred BMI and menopausal status, lifestyle score and breast density. The ROC AUC of the model was 0.79. After applying the model to the validation cohort and recalibrating the predicted probabilities, an ROC AUC of 0.87 was obtained.

In this project, we reaffirmed the ability of circulating microRNAs to diagnose BC, analysed the sources of heterogeneity and discussed the problems of standardisation and reproducibility of results. Additionally, we identified cfc miRNAs potentially useful for BC detection in a screening setting.

Acknowledgements

First and foremost, my gratitude goes to my parents, Muzafer and Rasema Šehović as well as my sister Emina for their love and support throughout the years. Without them none of my achievements would be possible. I greatly appreciate their motivation and kind advice, which were of great value both for this project and my life in general.

I extend heartfelt thanks to my supervisor Dr. Giovanna Chiorino for her guidance, help and for allowing me and assisting me in expressing my scientific creativity. I would also like to thank my consortium co-supervisor Prof. Jaakko Kaprio for his invaluable advice, guidance and brainstorming sessions in all the monthly supervisor meetings as well as during my 5-month secondment stay in Helsinki. Furthermore, my thanks also go to my University co-supervisor, Prof. De Bortoli for assisting me in result interpretations and integrating me with his research group lead where I met Dr. Giulio Ferrero, who assisted me with the section on further exploring the CpG sites associated with puberty and breast cancer and their link to microRNAs.

I would like to thank my colleagues from the Cancer Genomics Laboratory in Biella for their support and numerous discussions. Specifically, I would like to thank Dr. Ilaria Gregnanin and Dr. Maurizia Mello-Grand, for their assistance in conducting the laboratory experiments. Also, my gratitude goes to Dr. Paola Ostano for her assistance on statistical analyses. Finally, I would like to thank my colleague Eirini Chrysanthou for the countless brainstorming sessions and for all scientific and non-scientific help throughout these years.

Next, I would like to thank Prof. Philipp Doebler, with whom I had a very successful collaboration on meta-analysing diagnostic circulating microRNAs in breast cancer detection. His patience, expert advice and guidance have enabled me to learn a lot about meta-analysis techniques and have led to the completion of an important part of my PhD thesis. Importantly, I would like to thank Sara Urru, as she was an integral part of the meta-analysis project, who worked with me on the statistical analyses of the meta-analysis and with whom I had many important discussions which improved the quality of our work.

I express my appreciation to the CancerPrev consortium, led by Prof. Cathrin Brisken, where I was an early stage researcher, for the excellent meetings and useful scientific interactions. This project has received funding, through the CancerPrev consortium, from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 859860. Additionally, I would like to thank Carlos Venturi Ronchi and Markus Kirolos Youssef, who are my early stage researcher colleagues in the consortium, for the regular brainstorming sessions in biostatistics and bioinformatics, which has helped me in performing numerous analyses. I would also like to thank Hannes Frederik Bode, for sharing with me his preliminary

results on DNA methylation and BC risk, which was important for the completion of describing puberty associated CpG sites linked to BC.

I sincerely thank all the scientists and nurses, volunteers, administrative personnel, biologists and nutritionists involved in the ANDROMEDA project, from which we obtained the cohort on which we identified cell-free circulating microRNAs. Specifically, I would like to thank Dr. Elisabetta Petracci for numerous discussions and her assistance on biomarker discovery which were important for the work presented here. My gratitude goes to Dr. Nereo Segnan, for his valuable advice and enabling me to see the bigger picture when it comes to my project. Additionally, I would like to thank Dr. Alessia Russo for sharing the methylation-sensitive high resolution melting data on the mentioned cohort and helping me interpret the obtained results which are presented in this work.

During my 5-month secondment at the Institute for Molecular Medicine Finland (FIMM) I worked on epigenetics of pubertal timing as early puberty is a risk factor of breast cancer and we have published our work in Clinical Epigenetics. I have met many amazing scientists and wonderful people at FIMM. I would like to thank Dr. Miina Ollikainen who, together with Prof. Jaakko Kaprio, supervised my secondment project and gave me important advice on epigenetics and epigenome-wide association analyses as well as helped with interpreting many of the obtained results during the secondment. I would also like to thank Dr. Stephanie Marie Zellers for teaching me many twin modelling techniques, who also worked on twin modelling in our publication. Last but not least, I would like to thank Aino Heikkinen for teaching me how to manage and pre-process Illumina 450K and EPIC Methylation data and who assisted with data management and pre-processing in our publication.

A part of this work is based on the data generated by the TCGA Research Network: https://www.cancer.gov/tcga. I would like to thank the scientists involved in the generation and pre-processing of the TCGA data as well as all the specimen donors.

Last but not least, I am grateful to the external referees (Prof. Marike Gabrielson and Dr. Giovanna Ambrosini) who reviewed this dissertation. Their comments were invaluable for the quality of this work.

Declaration

I hereby state that this dissertation was assembled and written solely by me. Importantly, the dissertation was internally reviewed and corrected by my supervisors and collaborators and then externally reviewed and graded by two referees. The work described here is, in part, based on previously published articles, and therefore some of the results and conclusions from those articles were adapted and explained in detail here. Additionally, the projects on which this dissertation is based involved collaboration between several institutions, and the contributions of the scientists within the respective institutions are briefly stated in the Acknowledgements section.

Emir Šehović
19/02/2024

Table of Contents

List of Tables

List of Figures

List of Appendices

Additional files

Supplementary Tables

Supplementary Figures

List of Abbreviations

BC	Breast Cancer
IDC-NST	No Special Type Infiltrating Ductal Carcinomas
MRI	Magnetic Resonance Imaging
H&E	Hematoxylin and Eosin
IHC	Immunohistochemically
ER	Oestrogen Receptor
PgR	Progesterone Receptor
HER2	Human Epidermal Growth Factor 2 Receptor
CNA	Copy Number Aberrations
TNM	Tumour, Node and Metastasis
HRT	Postmenopausal Hormone Therapy
SNP	Single Nucleotide Polymorphism
GWAS	Genome-Wide Association Study
PRS	Polygenic Risk Score
BI-RADS	Breast Imaging Reporting & Data System
BMI	Body Mass Index
EDC	Endocrine Disrupting Chemical
miRNA	MicroRNA
NGS	Next Generation Sequencing
ctDNA	Circulating Tumour DNA
RT-qPCR	Real-Time Quantitative Reverse Transcription Polymerase Chain Reaction
lncRNA	Long non-coding RNA
sncRNA	Small non-coding RNA
snoRNA	Small nucleolar RNA
siRNA	Small interfering RNA
snRNA	Small nuclear RNA
piRNA	Piwi-interacting RNA
pri-miRNA	Primary miRNA
pre-miRNA	Precursor miRNA
RISC	RNA Induced Silencing Complex
cDNA	Complementary DNA
AR	Androgen Receptor
cfc	Cell-free Circulating
PRISMA-DTA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy
ROC AUC	Area Under the Curve of The Receiver Operating characteristic
TP	True Positive
FP	False Positive
TN	True Negative

FN	False Negative
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies
DOR	Diagnostic Odds Ratio
PLR	Positive Likelihood Ratio
NLR	Negative Likelihood Ratio
PPV	Positive Predictive Value
NPV	Negative Predictive Value
GLMM	Generalised Linear Mixed Model
SROC	Summary Receiver Operating Characteristic Curve
REML	Restricted Maximum Likelihood
WCRF	World Cancer Research Fund
AICR	American Institute for Cancer Research
DM	Digital Mammography
EDTA	Ethylenediaminetetraacetic Acid
RPM	Revolutions per minute
SD	Standard Deviation
CV	Coefficient of Variation
PCA	Principal Component Analysis
Ct	Cycle Threshold
MS-HRM	Methylation-Sensitive High Resolution Melting
RFU	Relative Fluorescence Unit
LR	Logistic Regression
LASSO	Least Absolute Shrinkage and Selection Operator
NM	Non-molecular
HSM	Hierarchical Shrinkage Model
CI	Confidence Interval
OR	Odds Ratio
LOWESS	Locally Weighted Scatterplot Smoothing
IECV	Internal-External Cross-Validation
IPA	Ingenuity Pathway Analysis
FDR	False Discovery Rate
cCRE	candidate Cis-Regulatory Element
TPM	Transcripts per million
PBMC	Peripheral Blood Mononuclear Cells
FPR	False Positive Rate
vsd	Variance Stabilised Data
SE	Standard Error
FC	Fold Change

Introduction

Cancer is an umbrella term for diseases which involve uncontrolled cell proliferation and invasion. Cancer can occur in any tissue and start from any cell type, making it a very heterogeneous set of diseases with differing medical treatments and prognostic estimates.

Cells with abnormal growth, also called neoplastic cells, start by acquiring several mutations or genomic alterations. Such cells are often under replicative stress, making their DNA even more susceptible to DNA changes such as genomic breaks or additional duplications or deletions [1]. Together with the initially acquired mutations, especially in DNA damage response, the new genomic changes would allow for numerous neoplastic clones to arise, which would, through natural selection, shape their phenotype and enable them to adapt to their microenvironment, leading to cancer formation [1]. However, mutations are often not enough for a malignant tumour to develop, as a favourable environment for tumours is also crucial for it to become a disease of the tissue [2,3]. A common example is an inflamed environment, under which cells that had previously acquired some mutations have a much larger probability of becoming cancerous [2,4,5]. Neoplastic cells can be malignant or benign. Malignant neoplasms generally grow much faster than benign neoplasms and have an invasive characteristic that the benign tumours do not have.

In spite of the heterogeneity observed between different cancer tissues as well as between their subtypes, all invasive tumours share a set of hallmarks that characterise them (**Figure 1**), such as enabling replicative immortality, avoiding immune destruction, genome instability and mutation [6]. The cancer hallmarks are a consequence of genetic and functional changes in pathways essential for tumour proliferation. For example, the mutation of the Tumour Protein P53 (*TP53*) affects the p53 protein which is crucial for guiding the cell into performing DNA repair, cell-cycle arrest, apoptosis, etc [7]. Hence, *TP53* is the most mutated gene across all cancers [8]. Further, a mutation on the Phosphatidylinositol 3-kinase (*PIK3CA*) gene deregulates the PI3K/AKT signalling pathway, which is involved in cancer formation, inhibition of apoptosis and angiogenesis (which is the ability of cancer to form additional blood vessels supplying it with nutrients) [9].

Another important pathway for signal transduction in cells, which is also often deregulated in cancer, is the Janus kinase/signal transducer and activator of transcription (JAK/STAT) signalling pathway. Numerous cytokines and growth factors are found in this pathway, and some of its related downstream functions include haematopoiesis, immune fitness, tissue repair, inflammation, apoptosis, etc [10]. Mutations within the Janus Kinase 2 (*JAK2*) can cause constitutive activity of JAK/STAT, which means that the signalling pathway can be active even without the ligand [11]. Notably, there are numerous other commonly deregulated signalling pathways in cancer, including the TGF Beta, MAPK, mTOR, etc [12–14].

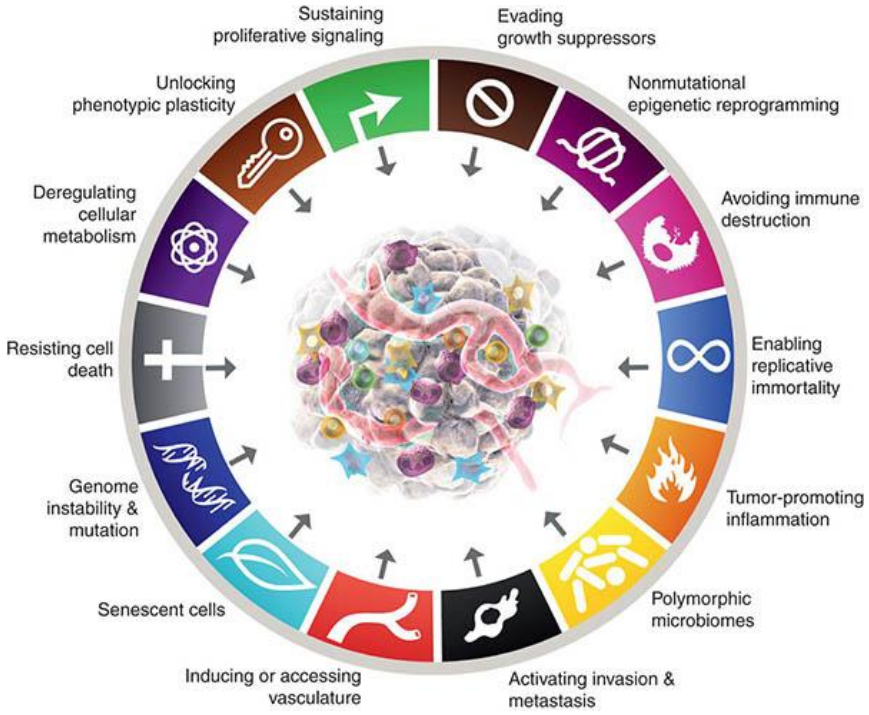


Figure 1. The fourteen cancer hallmarks updated in 2022. Taken from Hanahan 2022 [6].

Breast cancer

Breast cancer (BC) is the most diagnosed cancer in women and the cancer with the highest mortality rate. In 2020, according to the global cancer statistics, 2,260 (all ages, in thousands) new BC cases were reported in females worldwide, with an age-standardised rate per 100,000 of 47.8 and a cumulative risk of 5.20% up to the age of 75 [15]. The age-standardised rate for BC mortality was reported at 13.6, with a 1.49% cumulative risk up to 75 years of age (**Figure 2**).

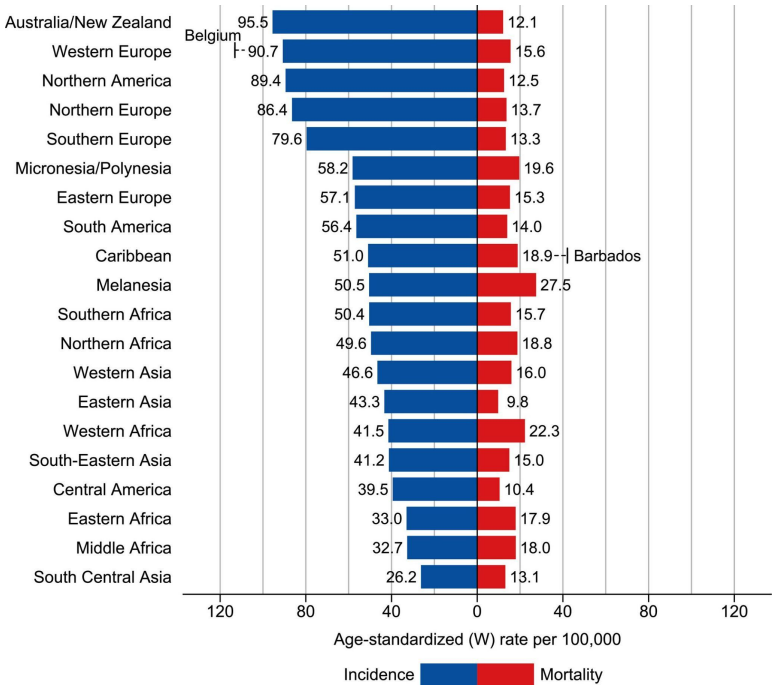


Figure 2. Incidence and mortality age-standardised rates in regions around the world. Taken from the Global Cancer Statistics 2020.

Moreover, according to the Association of the Nordic Cancer Registries, based on all Nordic countries, the prevalence of BC is around 2% (**Figure 3A**), while age-standardised rates of incidence and mortality are 106.4 and 20.6, respectively (**Figure 3B**) [16]. The difference in incidence estimates between the world and the Nordic countries reflects the higher BC incidence in Europe, partly due to more prevalent screening programs or differences in risk factor levels [17].

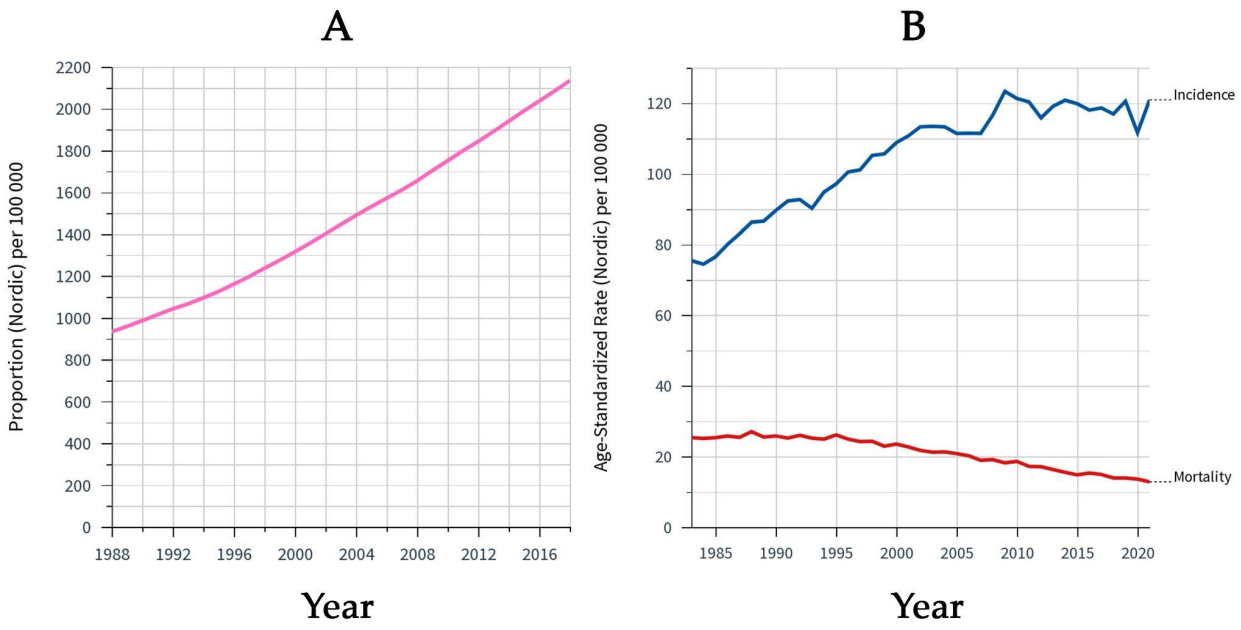


Figure 3. Prevalence estimates of BC in the Nordic countries in the last 35 years reported per 100,000 (A) and the age-standardised incidence and mortality per 100,000 (B). Taken from the NordCan database.

BC tumours are usually carcinomas, which are cancers that form in the epithelial cells in the breast. The most common type of carcinomas in the breast are adenocarcinomas or cancers that form in the milk ducts or glands responsible for making milk (lobules) [18]. Further, the most common adenocarcinomas are also classified as no special type infiltrating ductal carcinomas (IDC-NST) and invasive lobular carcinomas (ILC), which make up around 70% and 10% of all invasive cancers, respectively [19]. Other types of BC carcinomas, which are much rarer, are mucinous, cribriform, micropapillary, papillary, tubular, medullary, metaplastic and apocrine [20]. These classifications are usually referred to as tumour histologic type and will be relevant in the later sections of the thesis. There are some other rare types of BC, such as Angiosarcoma or Paget disease of the breast, which occur in other cell types. Additionally, depending on whether the BC tumour has spread to nearby tissue, the tumour can also be classified as invasive or in situ (see below for details on BC tumour classification and staging).

Diagnosis, prognosis and treatment

After a suspicious lesion in the breast is detected by an imaging technique, such as mammography, ultrasonography or magnetic resonance imaging (MRI), a biopsy sample of the lesion is taken for further examination. Biopsies of nearby lymph nodes are also often taken.

The tissue obtained from the biopsy first needs to be formalin-fixed and paraffin-embedded. Then, a section of the paraffin is cut, usually of thickness ranging from 3 to 5 μm. To visualise the nuclei and cytoplasm, the sample is dyed using haematoxylin and eosin (H&E) [21]. The sample is then

microscopically investigated for the presence of a tumour (**Figure 4**). Afterwards, if the sample is tumour positive, the histotype and grade of the tumour are determined. The grading (ranging from 1 to 3) of the tumour is evaluated by assessing the differentiation of the cells. A low grade indicates that the cells have not substantially dedifferentiated (making them relatively similar to the normal cells), are growing slower and have a lower chance of spreading further. On the other hand, a high grade indicates that the cells have dedifferentiated and that the tumour is much more aggressive [22].

The sample is also usually immunohistochemically (IHC) stained by applying antibodies to the tissue to identify specific antigens. Antigens of interest are usually oestrogen (ER), progesterone (PgR) and human epidermal growth factor 2 (HER2) receptors, as well as Ki-67 protein, which is often used as a cell proliferation marker [23].

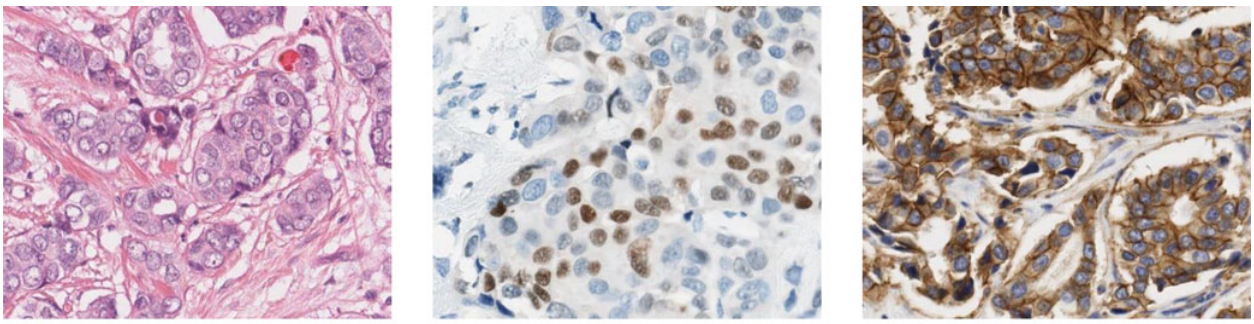


Figure 4. Examples of different staining techniques. Shown are the H&E staining (A), IHC staining for ER (B) and IHC staining for HER2 (C). Taken from Veta et al. 2014 [21].

A molecular classification of the BC tumour is often performed based on the IHC markers [24]. The four main subtypes of BC tumours are Luminal A (75.3%), Luminal B (11.1%), HER2 positive (3.1%) and Triple negative (10.5%). Luminal B can also be divided into HER2 positive and HER2 negative [25].

A 50 gene expression signature was identified which could cluster the four subtypes from the molecular classification in addition to a normal-like subtype which is similar to the luminal A subtype [26]. The classification obtained is referred to as the “intrinsic subtyping”. Another BC classification method, which is based on the copy number aberrations (CNA) of tumour samples, identified ten integrative clusters with distinct disease-specific survival times [27]. CNAs represent the number of times a specific genomic segment has been duplicated and, unlike copy number variants, occur only in the tumour cell.

Routine practice also involves performing tumour, node and metastasis (TNM) staging of the diagnosed BC tumours. Staging can be either clinical or pathological. Clinical staging relies on tests performed before the surgery, such as physical examinations, mammograms, ultrasounds, and MRI scans. On the other hand, pathological staging is based on the surgical findings during the removal of breast tissue and lymph nodes. Generally, the results from pathological staging are

available a few days after surgery, and overall, this method offers the most comprehensive information for assessing a patient's prognosis.

The tumour evaluation and staging (T in TNM) is based on the size and other characteristics of the tumour (e.g., tumour with a direct extension to the chest wall or the skin with macroscopic changes). The nodal evaluation and staging (N in TNM) is based on the presence and location of lymph nodal metastases. Finally, the distant metastasis classification (M in TNM) is based on clinical or imaging evidence of distant metastases. The T, N and M classifications are then used to determine the overall stage of the tumour (**Table 1**) [28]. Details on staging can be found in [29].

Table 1. Breast cancer tumour pathological staging according to the TNM classification, based on the AJCC Cancer Staging Manual.

Stage	TNM
Stage 0	Tis, N0, M0
Stage IA	T1, N0, M0
Stage IB	T0, N1mi, M0
	T1, N1mi, M0
Stage IIA	T0, N1, M0
	T1, N1, M0
	T2, N0, M0
Stage IIB	T2, N1, M0
	T3, N0, M0
Stage IIIA	T0, N2, M0
	T1, N2, M0
	T2, N2, M0
	T3, N1, M0
	T3, N2, M0
Stage IIIB	T4, N0, M0
	T4, N1, M0
	T4, N2, M0
Stage IIIC	Any T, N3, M0
Stage IV	Any T, Any N, M1

The 5-year survival of BC is around 90% [30], and the prognosis usually depends on the stage at diagnosis, how aggressive the cancer is and the success of the treatment. Therefore, the previously mentioned molecular, genomic and transcriptomic subtypes have differing average survival times (**Figure 5**) [31]. For instance, luminal A has the best overall survival time with 90% survival at 5 years, while the triple-negative subtype has the worst survival probability at 5 years with around 30%. Importantly, patients whose BC was detected in an early stage have a much better prognosis than those with later BC stages [32,33].

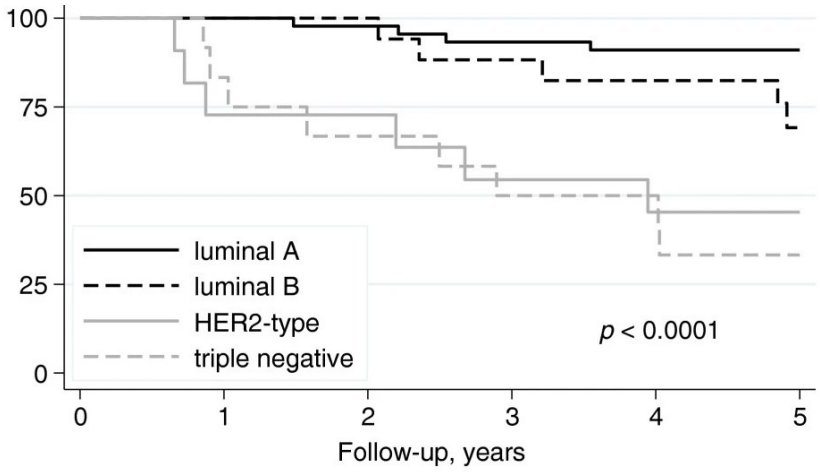


Figure 5. Breast cancer 5-year survival stratified by the St. Gallen molecular subtypes. Taken from Falck et al. 2013 [31].

Depending on the BC subtype, different genetic pathways are deregulated, and some of the pathways are involved in physiological mammary gland development [34]. For instance, a very important pathway is ER signalling, which is relevant in all ER-positive BCs. Through ERα and ERβ, which are transcription factors, the expression of various target genes can be affected. One gene that is affected by ER signalling and promotes BC growth is cyclin D1 (*CDK1*), which is involved in the cell cycle progression [35–37]. Additionally, HER2 signalling is another important pathway for BC, and is found hyperactivated in HER2-positive BCs. Deregulated HER2 activation, through phosphorylation, leads to increased tumour cell proliferation and cancer progression. HER2 activation initiates other tumorigenic signalling pathways, such as the previously mentioned PI3K/AKT pathway [38,39]. Moreover, in the context of BC, PI3K/AKT can also lead to dedifferentiation of luminal or basal mammary progenitor cells, thus allowing them to obtain multiple lineages [40], making them more adaptable to their environment as well as to therapies. Lastly, another notable pathway is the Wnt/β-catenin signalling, which is involved in maintaining the stem cell properties of BC [34], which are important for initiation [41], self-renewal and resistance to apoptosis [42].

BC treatment is usually tailored depending on the receptor status of the cancer. Therefore, three types of treatment strategies exist: treatment for hormone receptor (HR) positive and HER2-negative patients, treatment for HER2-positive patients and treatment for triple-negative patients [43]. Surgical operation or axillary lymph node removal, sometimes followed by postoperative radiation, are the initial local therapies for nonmetastatic BC. Neoadjuvant (before surgery) chemotherapy or immunotherapy (for triple-negative BC) may also be given [43]. Regarding the treatment strategies, all HR+ tumours receive endocrine therapy (inhibiting the binding of oestrogen to ER or inhibiting the conversion of androgens to oestrogens), and for some HR+ patients chemotherapy is introduced as well [43]. Trastuzumab-based HER2-directed antibody therapy and chemotherapy are given to all HER2+ tumours. Additionally, endocrine therapy is given in case of HER2+ tumours being HR+. Finally, for TNBC, chemotherapy is usually the only

therapy administered and a similar therapeutic strategy is used for metastatic BCs [43]. However, neoadjuvant chemotherapy with immunotherapy is currently the preferred approach to treat early-stage (II or III) TNBC [44]. The average cost of treatment across all BC stages was around $85,000 in 2016 in the United States of America, and the cost increased significantly with the TNM stage [45].

The quality of life of a patient is an important aspect when it comes to BC management. Therefore, it is important to consider that most drugs used for BC treatment have unpleasant side effects, often affecting the life quality of the patient [43,46]. Additionally, women with BC can have a reduced body image, worsened family functioning and an increased risk of developing depression [46,47]. The prognostic and treatment cost advantages of detecting BC in an early stage, as well as the decrease in quality of life among patients further emphasise the importance of primary and secondary prevention.

BC risk factors

Like for all types of cancers and diseases, not all individuals are equally susceptible or equally likely to develop BC. Both intrinsic (coming from within the individual) and extrinsic (coming from outside the individual) risk factors have been well-documented epidemiologically in the past decades. More comprehensive BC risk factor reviews can be found elsewhere [34,48,49].

Demographic factors

The most pronounced intrinsic risk factor of BC is sex, as men account for less than 1% of all BC incidences [30]. The next important risk factor is age, which is common across most cancers. BC incidence was shown to be much higher as age increases with the plateau being around after the age of menopause [50,51].

Reproductive factors

Reproductive factors were also found to be associated with BC risk. Namely, earlier age at menarche or later age at menopause were found to be risk factors for BC [52]. These factors regulate the number of menstrual cycles which affects the total exposure of the breast tissue to oestrogen, which is a risk factor for BC. Additionally, a higher number of pregnancies reduces the BC risk, while nulliparous or women with pregnancies at later ages have an increased BC risk [53,54].

Hormonal modulation

Hormone-modulating drugs, such as the use of oral contraceptives or postmenopausal hormone therapy (HRT), were consistently found to be positively associated with BC [55–57]. The increased risk from using contraceptives decreases gradually to average after 5 to 10 years of cessation [58]. Notably, the BC risk increase when using oral contraceptives is not the same for all formulations, as it is believed that the relative risk is higher for contraceptives containing synthetic progesterone receptor agonists [59]. Similar to oral contraceptives, the increased BC risk from HRT usage gradually diminishes to average two years after cessation [56].

Hereditary factors

Numerous genetic variants or mutations are associated with BC risk. Two famous genes associated with hereditary BC (approximately 5% of all BCs), Breast Cancer gene 1 (*BRCA1*) and Breast Cancer gene 2 (*BRCA2*), account for around 40% of hereditary BC incidences [60]. *BRCA1* and *BRCA2* proteins are involved in DNA repair and cell cycle regulation. In the cases of hereditary BC, the individual inherits a mutated *BRCA1* or *BRCA2* gene from one of the parents but will still have one wild-type *BRCA1/2* gene on the other copy of the chromosome. The cells in the human body constantly acquire random mutations. Therefore, it is only a matter of time before the *BRCA1/2* copy on the other chromosome also acquires the mutation. Hence, the so-called two-hit hypothesis of tumour suppressors is made more likely [61]. Mutations of *BRCA1* and *BRCA2* can also cause hereditary ovarian cancer and rarely some other types of cancer, such as pancreatic or prostate cancer [62]. However, the tissue-specific hereditary role of *BRCA1* and *BRCA2* in the breast and ovaries is not fully elucidated yet. Nevertheless, some hypotheses on the interaction of *BRCA1/2* genes and sex hormones, as well as the lack of compensatory proteins for DNA repair have been proposed [63].

Considering the hereditary aspects of BC, family history is one of the key risk factors associated with BC. Even without a *BRCA1* or *BRCA2* mutation, women with BC family history (two or more BC cases in first-degree female relatives younger than 50 years or three or more first-degree relatives with BC at any age) are four times more likely to develop BC than women without [64]. Consequently, there are numerous other genes, such as Phosphatase and tensin homolog (*PTEN*), *TP53*, Cadherin 1 (*CDH1*), etc. [65], as well as single nucleotide polymorphisms (SNPs) [66], which are single nucleotide variations that exist in at least 1% of the population, involved in BC susceptibility.

Numerous genome-wide association studies (GWAS) analysed a large number of SNPs to identify those associated with BC risk [67]. Usually, these SNPs are individually not strongly associated with BC risk. Therefore, their cumulative effect can be combined into a single polygenic risk score (PRS) [68]. Importantly, PRS and family history were found to be relatively independent when it comes to BC risk. This indicates the importance of considering both risk factors [69].

Breast factors

Additional factors which are related to the breast are associated with BC risk. The most important factor is breast density, which is positively associated with BC. Breast density is usually measured with one of the imaging tools mentioned above, and guidelines for density classification have been developed [70,71]. The Breast Imaging Reporting & Data System (BI-RADS) guidelines groups the breast density into four categories:

1) almost fatty
2) scattered fibroglandular densities
3) heterogeneously dense
4) extremely dense

Another breast density classification method is Tabar's classification, which has five categories based on the predominance of tissue type (i.e., fibrous, fat, nodular densities, etc.). Additionally, having previously identified benign biopsies also increases the risk of BC [72]. However, this is most pronounced in premenopausal women as the risk diminishes after menopause [73]. Finally, a longer cumulative duration of breast lactation was found to be associated with reduced BC risk [74].

Lifestyle and other factors

Lifestyle is an important aspect of health, and several lifestyle factors are associated with BC risk. Like in many cancers, obesity was found to be associated with BC [75]. However, the relationship between body mass index (BMI) and BC risk remains to be fully elucidated, as the association between BC risk and BMI is not so clear in premenopausal women but is a significant risk factor in postmenopausal women [76]. One hypothesis is that postmenopausal women with more body fat tend to have higher levels of circulating oestrogen, as oestrogen in postmenopausal women is mainly produced in the fat tissue [77]. Therefore, the higher oestrogen levels increase the BC risk [78]. In contrast, among premenopausal women, oestrogen mainly comes from the ovaries, making the body fat not as impactful when it comes to BC risk among premenopausal women [77].

Both alcohol consumption and smoking were found to be positively associated with BC risk [79,80]. Similarly, some dietary habits, such as red meat or saturated fat consumption, were associated with BC risk [81,82]. Importantly, physical activity was found to be a protective factor of BC [83]. Further, air pollution and previous radiations, either due to treatment or screening, are extrinsic BC risk factors [84,85]. Moreover, external hormone disruptors such as shift work and night work were also found to be associated with BC risk [86]. Lastly, endocrine-disrupting chemicals (EDCs) are also relevant BC risk factors [87,88]. EDCs are substances or synthetic chemical compounds that can deregulate pathways of the endocrine system [88]. They could be found in the human environment as a consequence of industry or agriculture. One example of EDCs relevant to BC risk are

Xenoestrogens, which are external oestrogen-like compounds that can mimic the intrinsic function of oestrogen (e.g., binding to the ER). Xenoestrogens can be found in some types of plastics, pesticides, chemicals, and water systems [89].

Early BC diagnosis and prevention

There is a substantially better prognosis among BC tumours detected in their early stage, which is true for most cancers. Consequently, the quality of life of patients with a better prognosis is drastically better, and the overall healthcare burden and cost are lowered. Hence, primary and secondary prevention are of crucial importance for overall better BC management. Primary prevention, in this case, refers to individuals improving their lifestyle and avoiding substances associated with BC risk. On the other hand, secondary prevention refers to the early detection of BC through self-examinations or mammography screening programs. Both self-examinations and mammography screening can identify suspicious lesions which will then be examined by a pathologist as described previously.

Mammography is the golden standard for BC detection and is utilised in most BC screening programs worldwide. It is a fast, relatively cheap, and simple-to-use imaging tool based on X-rays. Regarding its diagnostic performance in detecting BC, it has a sensitivity and specificity of 85% and 90%, respectively [90]. Nevertheless, mammography screening does have certain drawbacks, such as radiation exposure, higher false positives in women with dense breasts (especially among younger women) and interval cancers [91].

With the introduction of digital mammography (DM), radiation exposure has been reduced significantly over the years, making the added risk of BC when performing mammography minuscule (rising only from 8.8% to 8.9% in women aged 50 to 69 that undergo biannual screening [92,93]) and should not be a deterrent to screening [91]. However, optimising mammography screening scheduling and finding complementary biomarkers could further reduce radiation exposure. Mammography has a false positive rate of around 10%, and a false positive mammography result implies a recall for further investigation where the outcome is a negative status for BC. Notably, the probability of having one false positive result in 10 annual mammograms is around 50% [94]. This drawback is especially relevant in women with higher breast density as they are more likely to have false positive recalls. Consequently, false positive recalls could cause worry and fear among the screened women [95].

Another notable drawback of mammography are interval cancers, which are cancer incidences which occur between the mammography screenings. They can happen either due to the inability to detect cancer during screening or when a fast-progressing cancer develops after the screening. The sensitivity of mammography can strongly be affected by increased breast density. For instance, due to masking, the sensitivity of mammography is only around 50% on breasts with BI-

RADS 4 (extremely dense) density classification [96]. Moreover, BCs without microcalcifications (calcium deposits in breast tissue) are significantly harder to detect in breasts with higher density [97,98]. Additionally, lobular cancers and diffusely growing cancers strongly resemble normal breast tissue and are, therefore, harder to detect using mammography [91]. Therefore, there is a strong need to overcome the mentioned limitations and find biomarkers which could complement mammography.

The screening and prevention programs for individuals with confirmed BRCA mutations are different from those for women without BRCA mutations [99]. In this project we focus on improving secondary prevention of BC in women without *BRCA1* or *BRCA2* mutations.

Diagnostic and risk-assessment biomarkers in BC

Due to the mentioned drawbacks of mammographic screening, studies have explored and assessed tailoring BC screening programs through different imaging technologies and combinations of risk factors [100–102]. The risk scores were based on some of the above-mentioned risk factors (i.e., reproductive history, previous breast biopsies, family history, etc.), which could be obtained through a questionnaire. Despite promising results, risk-stratified mammography is still not widely used and is mainly based on age stratifications of risk. For example, in the Italian region of Piemonte, women aged 45 to 49 with average risk (i.e., no BC-related mutations or close family members who had BC) can perform mammography once a year but are not invited (i.e., they are screened on their own initiative), while women older between 50 and 69 years old perform a mammography every two years, upon invitation. Women above 70 and below 75 are not invited to BC screening programs but can spontaneously adhere every two years. Women aged above 76 and below 45 do not enter the screening programs. Similar guidelines based on age were reported by the European Breast Guidelines [103] and in the United States of America [104].

In addition to biomarkers or risk factors that can be obtained through questionnaires, molecular biomarkers, such as SNPs combined into PRS or DNA methylation profiles, circulating DNA or microRNAs (miRNAs), could also be used for early BC detection or risk stratification.

Genetic biomarkers

Several types of genetic biomarkers are candidates for BC detection or risk identification. Next generation sequencing (NGS) tools have enabled the development of multigene panels for testing whether an individual has a mutation on genes associated with BC onset [105,106]. The predecessor to the NGS for detecting single gene mutations, which are still somewhat in use, are the Sanger sequencing and denaturing high performance liquid chromatography [107]. The panels test for mutations in BRCA genes and other relevant genes such as *TP53*, Partner and Localizer of BRCA2 (*PALB2*), *CDH1*, etc. However, the prevalence of known non-BRCA mutations among individuals

with hereditary BC is around 10%. Therefore, the guidelines for non-BRCA mutation genotyping are still under revision and optimisation [105].

Another type of genetic biomarker for identifying inherent BC risk are SNPs. Each SNP has its own risk estimate and is obtained by multiplying the allele of the SNP (either homozygous for the allele in the reference genome – wild type, homozygous for the variant allele, or heterozygous) by the log odds, usually obtained through logistic regression. Individual SNPs confer low risk, but the risk effects of the individual SNPs found by GWA studies are summed up to form a PRS [108] that has a stronger weight on BC risk. PRS scores in the context of BC were identified using very large cohorts and analysing millions of SNPs [66,109]. In this project we assessed the previously reported PRS on 77 SNPs [110].

Circulating tumour DNA (ctDNA) and whole blood mRNA expression were also found to be promising genetic biomarkers for BC detection. It is hypothesised that ctDNAs originate from cellular breakdown or active secretion by the tumour [111,112], or from circulating tumour cells (CTCs) [111,112]. Usually, digital droplet PCR or NGS tools are used to quantify the ctDNAs [113]. Several clinical trials were performed or are underway to assess the clinical utility of ctDNAs in the context of BC [114]. A meta-analysis on the different quantifying techniques obtained a pooled sensitivity and specificity of 87%, making ctDNAs promising biomarkers [115]. Precancerous and cancerous tumour cells usually change the gene expression and abundance profile of immune cells in their microenvironment. Hence, as the circulatory system is involved in pathological activities and defence, it is also expected that immune changes related to gene expression are observed in peripheral tumour cells [116,117]. Studies have identified blood gene expression signatures which showed potential in discriminating BC patients from healthy controls, with sensitivities and specificities being around 80% [118,119]. Additionally, a study that prospectively sampled the blood of individuals identified gene expression profiles that differed between BC cases and healthy controls many years (up to eight years) after diagnosis [120–122]. Finally, a Real-Time Quantitative Reverse Transcription Polymerase Chain Reaction (RT-qPCR) based 12-gene biomarker panel for early BC detection has been commercially developed and is designed for women aged 25 to 80. The inventors of the mentioned gene panel claim an accuracy of 92% [123]. Other commercial blood gene expression panels were reviewed in [124].

Epigenetic biomarkers

Epigenetics refers to the regulation of DNA transcripts or heritable changes on the DNA, such as methylation, which do not alter the DNA sequence but affect how the information on the DNA is used in a cell. Examples of epigenetic modulators are histone methylation or acetylation, DNA methylation, microRNA regulation of mRNA abundance, etc. Deregulation of the epigenetic profile is a key characteristic of tumour cells [125]. It is exploited by cancer in order to increase the transcriptional accessibility and mRNA quantity of genes necessary for proliferation and decrease them for tumour suppressors [125]. Additionally, changing the epigenetic profile enables the cancer

cells to undergo dedifferentiation [126]. Hence, a plethora of epigenetic biomarkers associated with tumour development, including BC, have been identified, such as the DNA methylation of various genes, non-coding RNAs such as long non-coding RNAs or miRNAs, etc [127–131].

DNA methylation

DNA methylation is the addition of the methyl group (CH_3) to the fifth carbon of the cytosine nucleotide ring. DNA methylation in mammals is usually found on cytosine nucleotides which come before the guanine nucleotide (CpG). DNA methylation is usually catalysed by a family of enzymes called DNA methyltransferases encoded by DNA methyltransferase (*DNMT*) genes (such as *DNMT1* or *DNMT3*) [132,133]. These genes are often deregulated in many cancers [133,134], including BC [135]. Depending on their genomic location, there are different ways in which DNA methylation can affect the function or expression of a gene. For example, DNA methylation of gene enhancers or promoter regions could affect the regulation of the respective gene by affecting the transcription factor binding ability [136]. Further, DNA methylation sites found on the first intron of a gene could also have an impact on the function of the respective gene by affecting transcription factor binding [137].

The methylation of a specific DNA region or a specific CpG site can be quantified in different ways, and this is reflected in the various methods by which DNA methylation biomarkers were identified. Bisulfite conversion is a method through which unmethylated Cytosine is converted to Uracil by denaturing DNA and applying sodium bisulfite. After subsequent amplification, the Uracil is then converted to Thymine. Several DNA methylation analysis methods rely on this nucleotide conversion, such as microarray (e.g., EPIC Illumina Infinium BeadChip microarray), bisulfite sequencing or methylation-sensitive high resolution melting (MS-HRM). In bisulfite sequencing, the sequencing results of sodium bisulfite-treated DNA are compared to those of untreated DNA and the methylated sites are found [138]. In the Infinium BeadChip arrays, specific probes for each locus of interest are designed to determine the proportion of methylated DNA samples through single-base extension and light intensity [139]. Finally, the MS-HRM method exploits the fact that more energy is required to break the cytosine-guanine bonds than the thymine-guanine bonds [140]. Considering the importance of DNA methylation in tumour development, numerous studies have tried to identify prognostic and diagnostic biomarkers for various cancers, including BC [128]. In the context of BC, the promoter methylation of numerous genes was analysed, such as Retinoic Acid Receptor Beta (*RARB*), APC regulator of WNT signalling pathway (*APC*), *BRCA1*, Death-associated protein kinase 1 (*DAPK1*), Ras Association Domain Family Member 1 (*RASSF1A*), etc [130,141]. Additionally, epigenome-wide association studies (EWAS) identified DNA methylation sites associated with BC risk [142,143] as well as DNA methylation sites that, in a panel, could be used for early detection of BC [144,145]. In this project, we investigated the promoter methylation of *RARB*, *APC* and *BRCA1* in the context of BC.

Non-coding RNAs

Non-coding RNAs represent the broad term for all RNAs which do not undergo translation, and approximately 98% of all human DNA transcripts are non-coding RNAs [146]. Their classification is usually based on length, with the cut-off between long and small non-coding RNAs being 200 nucleotides [147,148]. However, some RNA species that are in the grey zone of this cut-off make the classification more complicated, and therefore, dividing non-coding RNAs into three categories was proposed [148]:

1) Small RNAs (< 50 nucleotides)
2) RNA Polymerase III transcripts such as tRNAs, RNA Polymerase V transcripts in plants and small RNA Polymerase II transcripts such as (most) snRNAs and intron-derived snoRNAs (~50 to 500 nucleotides)
3) Long non-coding RNAs, which are mostly transcribed by Pol II (> 500 nucleotides)

Long non-coding RNAs

Long non-coding RNAs (lncRNAs) length ranges up to 100,000 nucleotides and are less conserved among species when compared to mRNAs [148]. Some lncRNAs are spliced and polyadenylated, meaning having a poly-A tail at the 3' of the transcript. This is a property of mRNAs, which makes such lncRNAs "mRNA-like" [148]. However, there are lncRNAs which are not polyadenylated or expressed from Pol I or Pol III promoters [148]. With respect to protein-coding genes, lncRNAs can be 'intergenic', antisense or intronic, but they can also be derived from pseudogenes [148], which are segments of the DNA that are structurally similar to a regular gene but are not able to code for a protein.

The most notable function of lncRNAs is their involvement in cell differentiation and development in both animals and plants [147,149–151]. However, they have been linked to numerous other functions such as p53-mediated response to DNA damage [152], cytokine expression [153], cholesterol biosynthesis and homeostasis [154,155], growth hormone and prolactin production [156], glucose metabolism [157,158], cellular signal transduction and transport pathways [159–161], etc [148].

LncRNAs are very important for cancer development as they have been linked to assisting the tumour cells in acquiring all hallmarks of cancer described previously [149]. Due to lncRNAs being regulated by several oncogenic or tumour-suppressive transcription factors, such as p53 [152,162], MYC [163,164], ER [165], etc. [149], they can be considered as the functional output of the oncogenic or tumour-suppressive pathways.

In the context of BC, it was shown that lncRNA HOTAIR could promote BC epithelial-to-mesenchymal transition and lung metastasis in mice via activating Cyclin-dependent kinase 5 (*CDK5*) signalling. Additionally, Linc-ROR is believed to promote oestrogen-independent growth

of BC cells by regulating the ERK-specific DUSP7 phosphatase, thus enhancing MAPK/ERK signalling with potential implications for tamoxifen resistance [166,167].

Finally, a lncRNA biomarker, RP11-445H22.4, was identified, which could differentiate between BC cases and controls and was significantly upregulated in BC patients [168]. This biomarker was analysed in serum on a cohort of 68 BC patients and 68 controls using the RT-qPCR method. Despite its promising performance, to my knowledge, this lncRNA was not validated in external cohorts, reducing its reliability.

Small non-coding RNAs

There are numerous types of small non-coding RNAs (sncRNAs), all of which have differing average sizes and biological functions [169]. Some of the more common and well-studied sncRNAs are small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), small interfering RNAs (siRNAs), small nuclear RNAs (snRNAs) and Piwi-interacting RNA (piRNAs). In this section I will briefly cover some of the scnRNAs.

snoRNAs are RNAs ranging from 60 to 300 nucleotides [170]. Genomically, they are usually located on the intronic region of coding or non-coding genes (i.e., the snoRNA host genes) and are believed to form through transcription and post-transcriptional regulation (such as splicing) of the mentioned host genes [171]. Additionally, they can be involved in post-transcriptional regulation of rRNA which is involved in protein synthesis [171,172]. snoRNAs can have both oncogenic and tumour suppressor roles as they were found to regulate various signalling pathways in cancer [173]. For instance, SNORD126 activates the PI3K/AKT signalling pathways to increase tumour growth in liver cancer and colon cancer [174], but could also be implicated with BC or other cancers relevant to this pathway. Moreover, copy number deletion of two snoRNAs (SNORD50A and SNORD50B) can be synergistic with the K-Ras signalling pathway activation, thereby promoting tumorigenesis in multiple cancer types [175].

Another type of sncRNAs are snRNAs, which are around 150 nucleotides long and are considered to have a function in RNA splicing [176]. Dysregulation of snRNA can increase oncogenic transcripts and decrease tumour suppressor transcripts in tumour cells [177]. The U1 snRNA expression upregulation deregulates the expression of numerous genes, some of which were enriched in the p53 signalling, cell cycle, and MAPK pathways [178].

Another important type of sncRNAs are piRNAs which usually range from 20 to 30 nucleotides in length and are associated with gene regulatory functions [179]. In humans, piRNAs are usually transcribed from the piRNA genomic clusters, mediated by RNA Polymerase II, or they can be derived from the 3′UTR of mRNAs and lncRNAs [180]. To perform their regulatory function, a complex is formed by piRNA and the Piwi protein in order to be involved in gene silencing during gene transcription or the post-transcription process [179]. Importantly, piRNAs were shown to repress the expression of numerous cancer-related genes in several types of cancers, including

breast, lung and liver cancers [180,181]. In the context of BC, PiR-021285 was found to promote tumour growth and invasion in breast cancer [182].

Finally, a type of sncRNAs which can also be involved in post-transcriptional gene expression regulation are miRNAs. Due to their ability to affect numerous genes and as there is abundant research on them in the cancer context, especially BC, I will dedicate a separate section to describe them.

miRNAs

miRNAs are short RNAs whose length ranges from 18 to 25 nucleotides and are relatively evolutionarily conserved [183]. miRNA genes can be both intergenic and intragenic; however, most of them are intragenic and are found inside introns and exons of genes or in untranslated regions (UTR) and regions of the genome containing repetitive sequences [184]. miRNA genes are often found in clusters within the genome and are, therefore, transcribed as polycistronic transcripts with a single promoter. Nevertheless, some miRNAs are monocistronic and are transcribed from a specific promoter [185]. Some miRNAs are intronic within the host gene and are spliced from the mRNA transcript [186,187].

miRNA genes are transcribed by RNA polymerase II together with various transcriptional factors into primary miRNA transcript (pri-miRNA) of more than 1 kb length [188]. The pri-miRNA consists of a long loop structure containing the stem and the loop region (responsible for encoding one or more mature miRNAs) and single-stranded RNA segments at the ends. The 5′ end is capped with 7-methyl-guanosine (m7G) and the 3′ end is polyadenylated [183]. This pri-miRNA is recognized by an enzyme complex (called Microprocessor complex), consisting of Drosha and DGCR8, and is cleaved to produce a long precursor miRNA (pre-miRNA), usually around 70 nucleotides long. The DGCR8 protein is responsible for recognising the pri-miRNA, while Drosha is responsible for specifically cleaving the pri-miRNA at specific points. These two points are usually between the apical junction linked to the terminal loop and the basal junction found between the single-stranded RNA and stem-loop structure [187]. Hence, the pre-miRNA retains the loop-like property observed in pri-miRNA. Then, the pre-miRNA is exported to the cytoplasm by exportin 5 (EXP5), which is a transporter protein [183]. The pre-miRNA is further processed in the cytoplasm by endoribonuclease (RNase III). Endoribonuclease is a Dicer enzyme which, in this context, is responsible for creating the miRNA duplex from the pre-miRNA. In mammals, Dicer recognizes the 5′ end of the pre-miRNA and cleaves the pre-miRNA at 22 nucleotides away from the 5′ end [189]. The created miRNA duplex, which is around 22 nucleotides long, is then phosphorylated at the 5′ end, and a 3′ overhang is formed. The whole process of miRNA biosynthesis is visualised in **Figure 6**.

The miRNA duplex consists of two strands, the guide strand and the passenger strand, where the guide strand is relevant for the assembly with the Argonaute proteins. miRNAs usually exhibit their regulatory function as a part of the RNA Induced Silencing Complex (RISC). The RISC is

assembled in two steps: first, the miRNA duplex is loaded onto the Argonaute protein (Ago 1-4 in humans). In the second step, the miRNA duplex is unwound, and the guide strand is selected and anchored into the Argonaute protein, while the passenger strand is ejected and degraded [190].

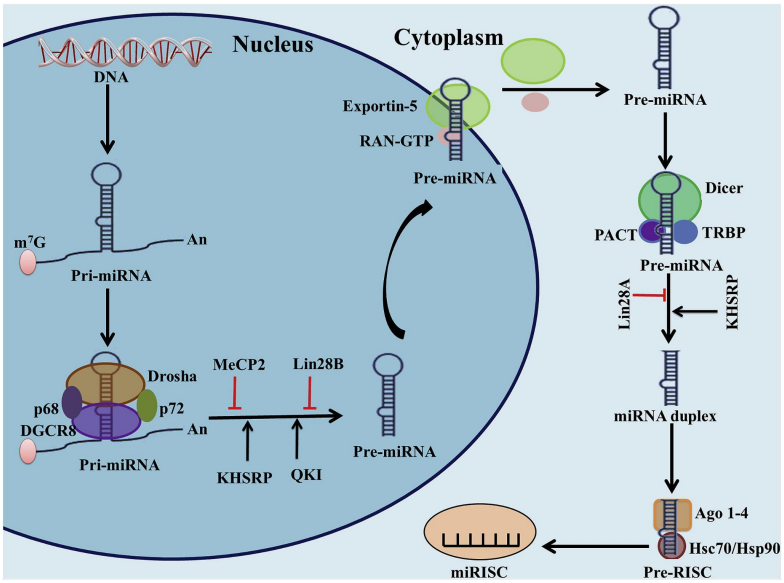


Figure 6. The pathway of biogenesis of miRNAs. Taken from Khan et al. 2019 [183].

The regulatory or silencing role of miRNA is manifested through the miRISC complex. The guide strand of the miRNA is responsible for targeting specific mRNAs. The specific mRNAs are recognised via the 3′ UTR region of mRNAs that have the binding sites for specific miRNAs [191]. The guided miRISC complex then either degrades the target mRNA or represses translation. Interestingly, a single mRNA can be targeted by an individual or multiple miRNAs, and a single miRNA can target several different mRNAs, making the interaction network between miRNAs and mRNAs rather complex [190].

Importantly, numerous miRNAs, both dysregulated and physiological, are reproducibly found in body fluids such as plasma, serum and saliva, where they have a role in intercellular communication or are exported out of cells that want to get rid of them. They are believed to be protected from degradation by association with secreted membrane vesicles (e.g., exosomes) or RNA-binding proteins [192]. This makes miRNAs highly promising candidates for becoming non-invasive diagnostic and prognostic biomarkers for various diseases.

There are several techniques for quantifying the expression of miRNAs in a tissue. The most widely used method, which is robust and relatively cheap, is the RT-qPCR platform. After RNA extraction (exosomal miRNA analysis will include an exosome disruption step) and reverse transcription to create complementary DNAs (cDNAs), fluorescence is emitted and measured after each synthesis of double-stranded nucleic acid by adding complementary bases to denatured single-stranded cDNAs [193]. Usually, this is repeated 40 times (cycles), and a cycle value is

obtained for a sample once the fluorescence intensity surpasses a set threshold (**Figure 7**). The downside of this technique is lower throughput compared to other methods, and it is being quite laborious when many miRNAs are assayed.

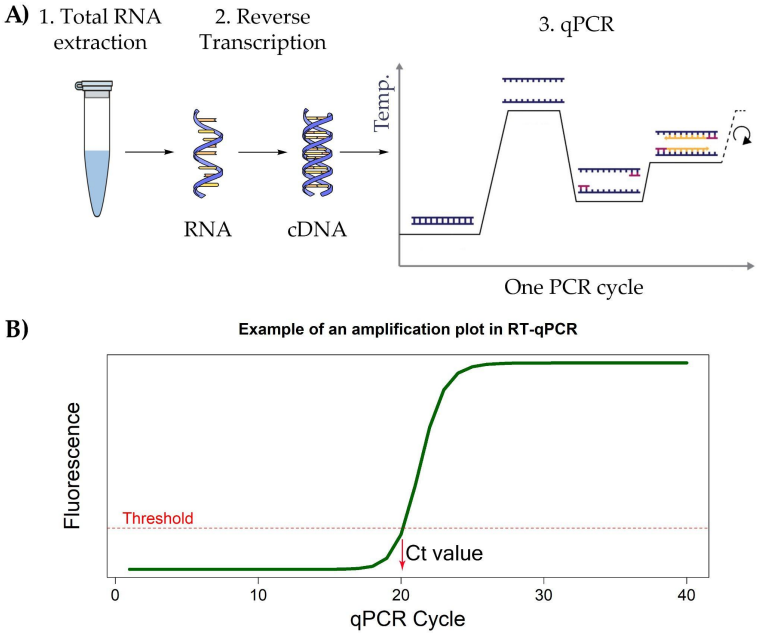


Figure 7. Illustrated are the (A) simplified protocol for miRNA analysis using RT-qPCR platform and (B) an example of a resulting amplification plot from which the cycle threshold value for a specific miRNA in a sample is derived.

Another commonly used technique is the microarray, where thousands of oligonucleotide probes are designed to be complementary to the known miRNAs and are immobilized on a solid substrate (usually glass) in discrete circular areas called spots [194]. Following total RNA dephosphorylation, denaturation and a ligation step, miRNAs in total RNA samples are fluorescently labelled and hybridized to complementary probes. The fluorescence intensity of labelled miRNA bound to microarrays is then used to derive an expression value for each miRNA.

Lastly, an NGS method called small-RNA sequencing is a high-throughput method for quantifying miRNA expression [195]. Following RNA extraction, adapters are added to the 3' and 5' ends of small RNAs, which are then reverse transcribed and amplified to obtain a cDNA library. Libraries are sequenced and reads are aligned to the genome. The number of reads aligned to a specific region of the genome will indicate its expression.

miRNAs and cancer

miRNAs have been found to be differentially expressed between normal and tumour cells and have, therefore, been associated with cancer onset and progression [196–198]. More than 50% of all annotated miRNAs in humans are located at genomic regions that tend to be amplified, deleted or translocated in cancer [199,200]. Additionally, due to their mRNA silencing or degradation role, they

can be important players in regulating tumour suppressor genes or oncogenes in favour of tumour cells. The general idea is that higher expression of miRNAs tends to increase the degradation or translational inhibition of tumour-suppressive or oncogenic mRNAs. Hence, the miRNAs targeting tumour suppressors will tend to be overexpressed in tumours, while the miRNAs targeting oncogenes will tend to be underexpressed in cancer cells.

The tumour cells exploit various mechanisms to dysregulate the expression of miRNAs, including (epi)genetic alterations (e.g., deletions or amplifications of genomic regions where a miRNA is encoded, increasing or decreasing accessibility of DNA for transcription factors, etc.) or defects within the enzymes involved in the miRNA biogenesis [201]. One important example of miRNA dysregulation is by the oncogene c-Myc. It is known to repress the transcriptional activity of tumour-suppressive miRNAs such as miR-15a, miR-26, miR-29, miR-30 and let-7 families [202].

Specifically, two miRNA clusters which are a part of the miR-200 family, miR-200bc/429 and miR-200a/141, are believed to play a role in the epithelial-to-mesenchymal transition, which is an important characteristic of cancers due to the acquisition of invasiveness and migratory ability by the cells [203]. The expression of miR-200 family was found to be dysregulated in numerous cancers including bladder cancer, gastric cancer, nasopharyngeal carcinomas, ovarian cancer, pancreatic cancer, and prostate cancer [204]. Furthermore, miR-32 is a miRNA regulated by the androgen receptor (AR), and it was found to promote prostate cancer cell growth and progression by inhibiting the expression of tumour suppressor genes (phosphoinositide-3-kinase interacting protein 1 (*PIK3IP1*) and B-cell translocation gene 2 (*BTG2*)) and favouring the PI3K/AKT/mTOR pathway [205].

Lastly, as they can stably be found in blood and other body fluids, miRNAs were investigated as prognostic and diagnostic biomarkers for various types of cancer [206]. For instance, various miRNAs were found as candidates for the early detection of prostate [207,208], lung [209,210], as well as numerous other cancers [211]. It is believed that the deregulated miRNAs found in the blood are mainly secreted by the tumour cells for reasons yet to be fully clarified [212], but a considerable proportion of the miRNAs could also be excreted by the red or white blood cells [213]. Another hypothesis explaining cell-free miRNAs in blood is that they are released by dying cells (both tumour and normal). Hence, deciphering the functional role of circulating biomarkers is a more complicated task than for solid tissue biomarkers.

miRNAs and BC

Numerous miRNAs were found to be involved in BC progression and onset [214]. Both tumour-suppressive and oncogenic miRNAs were identified. Examples of well-studied oncomiRs are miR-10b, miR-21, miR-155, miR-373 and miR-520c. For instance, through translational inhibition, miR-10b targets the tumour suppressor Homeobox D10 (*HOXD10*) [215], which inhibits the RHOC/AKT/MAPK pathways [216]. Further, miR-155 targets the tumour suppressor gene suppressor of cytokine signalling 1 (*SOCS1*). *SOCS1* plays a role in several cytokine signal

transduction pathways [217] and is believed to regulate the JAK/STAT signalling pathway [218]. Hence, overexpression of miR-155 in BC cells leads to the activation of signal transducer and activator of transcription 3 (*STAT3*) through the JAK pathway [219], which could produce immune tolerance within the cancer cells as they will be able to release factors that block antigen presenting cells' maturation or activation and inhibit the generation of antigen-specific T cells [220]. Lastly, miR-21, one of the most comprehensively described oncomiRs and one of the most frequently deregulated miRNAs in BC, targets several tumour suppressor genes, such as *PTEN* and TIMP Metallopeptidase Inhibitor 3 (*TIMP3*), and is associated with BC growth and progression [214,221]. *PTEN* has numerous tumour-suppressive functions, but the most notable one is that it blocks PI3K signalling by inhibiting PIP3-dependent processes such as the membrane recruitment and activation of AKT. Hence, the cell survival, growth, and proliferation are stopped [222].

The tumour-suppressive miRNAs that were also well studied in the context of BC are miR-125b, miR-205, miR-200, miR-146b, miR-126, miR-335, etc [214]. For example, miR-125b targets Erythropoietin (*EPO*), Erythropoietin Receptor (*EPOR*) and Erb-B2 Receptor Tyrosine Kinase 2 (*ERBB2*) [214]. In addition to activating *EPOR*, the protein encoded by *EPO* was found to induce PI3K/AKT and MAPK pathways in human breast cancer cell lines [223]. This is of relevance as the knockdown of *EPOR* reduced human tumour cell growth, induced apoptosis and reduced the invasiveness of the tumour [223]. Hence, miR-125b could be responsible for inhibiting cell proliferation and migration. Additionally, miR-205 was found to suppress proliferation and invasion by targeting HMGB3 [214].

Like many other cancers, circulating miRNAs were identified as biomarkers for diagnosis, prognosis and treatment of BC [224]. Some of the identified miRNAs' expression levels would change before the routinely applied diagnostic tools could detect the tumour [225]. Additionally, as was mentioned earlier, miRNAs are abundant in body fluids, stable and relatively cheap to analyse. Hence, circulating cell-free (cfc) miRNAs are potentially more effective in detecting early-stage BC when compared to the other mentioned biomarkers. Exosomal miRNAs found in the blood are also promising biomarkers for BC detection [226]. However, due to simpler and more standardised extraction protocol for cfc miRNAs (as multiple protocols were suggested for exosome isolation and analysis, and it is not always clear whether exclusively RNA from exosomes is quantified) [227], we opted to investigate only the latter.

A myriad of miRNAs or diagnostic models based on cfc miRNAs were reported in the context of BC [224]. Nevertheless, the published results were often non-intersecting or sometimes contradictory, as there have been many reported candidate miRNAs or panels of miRNAs, but a common significant panel of miRNA(s) as a clinically viable tool still needs to be identified [224]. One reason for this is the lack of experimental and methodological standardisation between the studies (e.g., normaliser or specimen type) [224,228]. Two meta-analyses from 2014 reviewed studies which reported diagnostic cfc miRNAs for BC and concluded that miRNAs have promising

diagnostic performance but also stated that a substantial degree of heterogeneity between the studies exists [229,230], partly due to the lack of standardisation.

Considering the importance of early BC detection and the fact that no common circulating miRNA panels for BC detection have been reported, there is a need to standardise the laboratory and research design protocols for identifying diagnostic circulating miRNAs. Hence, in this project, our goal is to identify the issues with standardisation and apply the findings to our biomarker discovery pipeline, which aims to identify cfc diagnostic miRNAs in a BC screening setting.

Research objectives

The key objective of this project was to identify robust and reliable blood circulating biomarkers, with the focus on cfc miRNAs, associated with BC in a screening setting. Ideally, these candidate biomarkers would complement or be an alternative to the current golden standard for BC detection, mammography. To achieve this goal, the three following research objectives were set:

1) As inconsistent and rarely intersecting diagnostic panels of cfc miRNAs for BC detection have been reported in the studies published thus far, we aimed to evaluate the overall diagnostic performance as well as the sources of heterogeneity between studies. This would be done by performing a meta-analysis where we would seek to include all high-quality evidence on the diagnostic performance of circulating diagnostic miRNA(s) for detecting BC using any RT-qPCR platform. Pooled diagnostic performance, heterogeneity analysis in the context of lack of standardisation, publication bias and the general risk of bias in individual studies were the main interests of the meta-analysis. Additionally, we wanted to assess the within and between-study preference for sensitivity over specificity and to stress how factors relevant to sensitivity or specificity preference should be considered and discussed in research papers. Based on a thorough review of the meta-analysed studies and the causes of inconsistency, we would use this knowledge to adapt our own methods for identifying circulating microRNAs associated with BC. Finally, the most consistent microRNAs among the meta-analysed studies would be prioritised in the variable selection analyses.

2) Identify cfc miRNAs associated with BC through small-RNA sequencing in a nested case–control study within a large cohort of women attending the BC screening program. This would be followed by a platform validation using RT-qPCR in the same cohort. A logistic regression model would be created on these miRNAs validated in RT-qPCR and the model would then be further validated in a separate cohort.

3) To functionally understand the candidate biomarkers and the limits of their applicability [231], we performed enrichment and network analyses, followed by extensive literature searches to see whether the function of the biomarker has been described before or whether it was found to be associated with any other disease. Finally, candidate miRNAs relevant to genes mapped to DNA methylation sites associated with pubertal timing or development, which was a research secondment project I performed in Finland, would be studied in depth and linked to increased risk of BC among women with earlier pubertal timing.

Material and Methods

The material and methods section, as well as the results and discussion, will provide a more detailed description of our two peer-reviewed publications [232,233]. I will first describe all the methods and statistical techniques used in the meta-analysis of cfc miRNAs in BC detection, followed by the materials and methods used for identifying novel circulating biomarkers associated with BC in a screening setting. Finally, a part was dedicated to investigating CpG sites associated with puberty that are linked to BC and miRNAs significantly targeting the genes mapped to the CpG sites. The CpG sites associated with puberty were identified in a cohort of Finnish young adult twins and an enrichment analysis was performed to identify CpGs enriched in various functions and diseases [234].

All statistical analyses were performed in the R software version 4.1.1 or 4.1.2. The packages and functions used will be mentioned in the relevant sections.

Meta-analysis on cfc miRNAs

Despite there being many studies that have reported cfc miRNAs associated with BC onset, there is a lack of consistent or overlapping microRNAs between the studies. To investigate this phenomenon, to assess the overall diagnostic ability of miRNAs as well as to help guide some of the decisions relevant to the project, there was merit in performing a diagnostic meta-analysis. The complete R analysis script and dataset of the meta-analysis can be found in the GitHub repository of the project (https://github.com/saraurru/Meta-analysis-of-diagnostic-cell-free-circulating-miRNAs-for-BC-detection).

Search strategy and inclusion criteria

The methodology was pre-registered in the international database of prospectively registered systematic reviews (PROSPERO; CRD42021229910). The workflow and methodology of the meta-analysis were based on the guidelines of Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy (PRISMA-DTA) [235]. Publications were searched in PubMed and PubMed Central (NCBI PMC) databases as well as the Google Scholar search engine. The search was performed up to March 21st, 2022. The full search strategy, with the keywords, is documented in the pre-registration. Only peer-reviewed journal articles published in English were considered. Abstracts and other types of publications were excluded. Eligible articles for inclusion were studies which analysed the diagnostic performance of cfc miRNAs in (early stage) breast cancer patients compared to healthy controls or to healthy controls plus patients with benign breast lesions. Therefore, any prognostic studies, studies that analysed exosomal miRNAs,

studies that did not have a miRNA model based on RT-qPCR data and studies that did not have a model with healthy controls were excluded.

The study designs included in this meta-analysis are retrospective or prospective case–control studies. Studies which included more than 4.5% metastatic (TNM Classification of Malignant Tumours stage IV- **Table 1**) breast cancer patients were also excluded because it is not expected to have more than 4.5% metastatic patients within a screening population [236]. Additionally, having too many metastatic patients when constructing a diagnostic model might create a bias in selecting and evaluating potential biomarkers due to the overall bigger biological differences between metastatic and healthy patients compared to non-metastatic patients and healthy controls. The studies were also required to report diagnostic performance data (sensitivity, specificity, area under the curve of the receiver operating characteristic (ROC AUC), etc.). Studies from which the frequencies of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) could not be directly or indirectly extracted were excluded. If studies had unclear but existing patient data, they were included in the analysis, but the authors were contacted for clarification. However, studies which did not specify whether stage IV cases were included and did not specify their number or percentage were excluded from the study if the authors did not reply to our inquiry. In addition, since the Google Scholar search engine was used, we checked whether all article's journals were peer-reviewed and indexed before inclusion in the full-text eligibility evaluation.

Data extraction and synthesis

The items and research publications obtained from the mentioned search sources were collected as a list in one spreadsheet. All duplicate hits were removed. The publication type, title and keywords were evaluated by myself (E.S) and my supervisor Giovanna Chiorino (G.C). Then, the abstracts of all articles not excluded in the initial evaluation were read. In case of any disagreements, a third reviewer, Philipp Doebler (P.D), was the arbiter. Afterwards, the articles that satisfied inclusion criteria based on the screening of abstracts were selected for the full-text evaluation, which was performed thoroughly, again by E.S and G.C, in order to decide on inclusion or exclusion. In all three steps, the reasons for exclusion were documented. Lastly, a list of articles fully eligible for this meta-analysis was compiled.

Using the same data extraction protocol and data structure, data from the selected articles was independently extracted by E.S and G.C. In case disagreements occurred between the two reviewers, P.D was the arbiter. From each study, the country, bibliometric data (author, year and journal), patients' average or median age, patients' BC stage distribution (from stage 0 to stage IV), diagnostic performance data (TP, FP, TN, FN; potentially several miRNA models were reported and if a study had a train as well as test/validation cohort the performance data were extracted only for test/validation cohorts), ROC AUC value(s), normalisation method, cut-off value(s), sample size of all groups, miRNA(s) profiled, specimen type, platform information and statistical model information were extracted. In addition, from the reported ROC curves, the q-

Point of the ROC (intersection of the anti-diagonal line on the ROC plot with the ROC curve) and three other points, aiming for equal distance between them, which were not on the extremities were extracted. As some studies only reported an ROC curve, the q-Point was extracted to obtain a uniform performance statistic from all the models. This enabled a complementary analysis because there were more studies which reported an ROC curve than studies with diagnostic performance data. The three additional points were extracted to fit a parametric ROC curve, which would then be used for analyses on sensitivity or specificity preference. The mentioned points were extracted from the ROC graphs using the "digitize" function from the *digitize* package in R software [237].

Risk of bias analysis

All the included studies were evaluated independently by two reviewers, E.S and G.C, using the revised tool for Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) [238] in order to evaluate the potential risks of bias (in four key domains: patient selection; index test; reference standard; flow and timing). The QUADAS-2 was tailored to be more suitable for studies which dealt with diagnostic performance of miRNAs for early BC diagnosis (*Additional file 1* – Appendix A). The main changes were made in Domain 2 (Index test) and Domain 4 (Flow and Timing). For each variable in QUADAS-2, the percentage of agreement between the two reviewers was determined. Discrepancies in coding or QUADAS-2 evaluations were resolved by trying to reach a consensus. In case no consensus could be reached, a third reviewer, P.D, was the arbiter.

Statistical analysis

Primary studies use a wide range of computational methods to obtain estimates of diagnostic performance and ROC curves, including classification methods like logistic regression and machine learning when the screening result depends on more than one variable. In this thesis, I will refer to the study-level computations as models, even if the computations are relatively simple. By utilising the diagnostic performance data (TP, TN, FP, FN) of the models, the sensitivity, specificity and diagnostic odds ratio (DOR) were calculated. In addition, other diagnostic performance parameters of the model, such as positive likelihood ratio (PLR), negative likelihood ratio (NLR), positive predictive value (PPV), negative predictive value (NPV), accuracy, etc., were calculated. Confidence intervals of PPV and NPV were calculated using the formula from [239] if sensitivity or specificity were equal to 1, otherwise logit transformation from [240] was applied. A formula from [241] was used to calculate the confidence intervals of PLR and NLR.

Descriptive statistics on diagnostic performance data were calculated using the "madad" function from the *mada* package in R software [242]. The equality of sensitivities and specificities, as well as the DOR and their confidence intervals were calculated. In addition, the correlation of sensitivities and false positive rates was calculated. Forest plots of sensitivities and specificities, the crosshair and ROC ellipse plots were based on those models labelled as the preferred model

by primary study authors or, if no preferred model was specified, on the best performing model (from now on 'most important model per study').

To estimate pooled sensitivity and specificity, two generalised bivariate linear mixed models (in this case referred to as statistical analysis models) were performed: one including all the models and one considering only one model per study. A generalised linear mixed model (GLMM) is a type of generalised linear model that incorporates fixed and random effects for instances where the data comes from different groups such as the diagnostic performance from different studies [243]. A bivariate GLMM model is used when the analysed variables of interest, as is the case for sensitivity and specificity, are not independent of each other.

In GLMM on all reported models, random effects on models and studies were added to take into account the between- and within-study variance. In the model on most important models per study, only the random effect on study was considered, resulting in the bivariate model from [244]. The approach was implemented with the "glmer" function in the *lme4* package [245], recommended by [246], and the summary receiver operating characteristic curve (SROC) was plotted for both models. The analyses were repeated on subgroups to detect possible differences in the performance measures. Subgroup analyses were based on normaliser type, specimen type, miRNA profiles (single or multiple miRNA panel) and presence of stage III and/or stage IV cases (< 4.5% as previously described). In addition, a subgroup analysis was performed on three subsets of studies depending on their QUADAS-2 score. Specifically, the score was determined by the number of "low" classifications (indicating a low probability of bias) among the seven key QUADAS-2 questions. The cut-points of the three subsets were set at > 3, > 4 and > 5 "low" classifications.

To assess the performance of models that did not report performance data, we used the extracted q-Points from the ROC graphs. From the extracted q-Points, we calculated the log-DOR on which we performed univariate analyses. The univariate analysis was performed on all models and the most important model per study. Forest plots were generated on the calculated log-DOR. The univariate analysis based on the log-DOR was also performed on the subgroups, both on all models and the most important model per study within the subgroups.

The univariate analysis was performed using the *metafor* package [247], with functions "escalc", "rma.mv" and "rma.uni". The "escalc" function was used to calculate the effect sizes and sampling variances for the log odds ratio. Then, on the calculated effect sizes and sampling variances "rma.mv" was applied when analysing multiple models from one study, allowing us to account for this fact, while the function "rma.uni" was used when analysing the most important model from studies. For both functions we used the restricted maximum likelihood (REML) method.

Sensitivity analysis

The outlier analysis was performed on all the models that reported diagnostic performance data. It was calculated based on the odds ratio. After the odds ratio for all models had been calculated, the

z-scores were calculated and a cut-off of z-score > 2 was selected to classify outliers. Additionally, influence analysis was performed on all models and the most important model per study. To identify the most influential studies, influence analysis was performed on the study level by taking into account all extracted models per study. Cook's distance of the bivariate mixed models was calculated using the "influence" function from the *influence.ME* package [248]. The z-scores were calculated based on Cook's distance and models with a z-score > 2 were deemed influential.

Imbalance of proportions

Different research designs are also reflected in the proportion of cases to controls, which might have an effect on the resulting performance measurements (i.e., sensitivity and specificity). To compare the performance of models with the imbalance of proportions of cases to controls or predicted positive to predicted negative screens, all reported models were divided into three groups. The cut-points for imbalance of proportions were set at < 0.7, between > 0.7 and < 1.3 and > 1.3. A graphical technique was utilised where the models were plotted on an ROC plane and marked according to the imbalance of proportions group they belonged to. Additionally, to further test the sensitivity of the mentioned graphical technique, the models were also divided into five groups with the cut-points for the imbalance of proportions set at ≤ 0.4, between > 0.4 and ≤ 0.8, between > 0.8 and ≤ 1.2, between >1.2 and ≤ 1.6 and > 1.6.

Implicit cost of misdiagnosis

Despite similar accuracy in terms of statistics like the ROC AUC, study-level ROC curves can have very different shapes. Assuming authors consciously or intuitively balance the shape of the study level ROC curve in accordance with the primary screening purpose, the study level ROC reflects a preference or compromise between sensitivity and specificity in the context of a population-level prevalence. Based on a method of [249], we include two statistics explained subsequently: (i) The shape parameter α that quantifies the (a)symmetry of the study level ROC curve. A value of $\alpha = 1$ indicates an ROC curve symmetric around the anti-diagonal on ROC space. Low values of α indicate a preference for specificity over sensitivity at the same overall accuracy, while high values lead to a preference for sensitivity over specificity. (ii) The cost parameter c_1 is a measure of the (implicit) author's perceived cost of a false negative misdiagnosis in relation to the cost of a false positive misdiagnosis. A value of $c_1 = 1$ indicates that for the prevalence at hand, authors chose a cut-off value for the primary study's ROC curve that assumes equal cost for both types of misdiagnoses. Values lower/higher than 1 correspond to lower/higher cost of a false negative case in relation to a false positive case. To assess the preference, the shape of the ROC curve was analysed, adapting a parametric method of [249].

Assuming that for every study the following relationship holds:

$$t_\alpha(sens) = t_\alpha(spec) + \theta$$

where

$$t_\alpha(x) = \alpha \log(x) - (2-\alpha)\log(1-x),$$

$$x \in (0,1), \alpha \in (0,2)$$

so that α is a shape parameter and θ is an accuracy parameter [249]. For a constant accuracy, the parameter α governs the asymmetry of the ROC curve. Hence, low values of α lead to a preference for specificity, while high values lead to a preference for sensitivity. We have used estimates of α to evaluate if an individual model has an inherent preference for sensitivity or specificity as well the general preference characteristics of meta-analysed models.

The t_α transformation was chosen because it has been shown to be more suitable than the logit transformation [250]. Based on the extracted three points from the ROC curve, three pairs of sensitivity and specificity values, and for a set of α values, t_α and θ were calculated for each point. By minimising the heterogeneity statistic Q, estimates of α result for each model in each study (cf. Eq. 23 in [249]).

We assume that authors base their decision about the study-level cut-off on study-specific (perceived) costs c_1 for not detecting a BC patient and c_0 for a positive screen of a healthy woman. The cost c_1 is represented in units of c_0, and by setting c_0=1, we simplified the calculation to this one parameter. Similar to the α parameter, the c_1 cost was used to evaluate if an individual model was affected by an inherent author preference as well as to evaluate the general author preference among the meta-analysed models.

For a prevalence π the expected cost is therefore

$$\mathbb{E}(cost) = c_1\pi(1-p) + (1-\pi)q$$

where p and q are short for sensitivity and false positive rate. Without loss of generality, the ROC curve is parametrized in q, so that p is a function of q. We can differentiate by q to obtain

$$\frac{\partial}{\partial q}\mathbb{E}(cost) = -c_1\pi\frac{\partial}{\partial q}p(q) + (1-\pi)\frac{\partial}{\partial q}q = -c_1\pi\frac{\partial}{\partial q}p(q) + (1-\pi)$$

The minimum cost is found at the q with

$$\frac{\partial}{\partial q}\mathbb{E}(cost) = 0$$

and we obtain

$$c_1 = \frac{1-\pi}{\pi}\frac{1}{p'(q)} = \frac{1-\pi}{\pi}\frac{1}{g^{-1}(g(q)+\theta)'} = \frac{1-\pi}{\pi}\frac{g'(p)}{g'(q)}$$

where g represents the t_α transformation. Since t_α has a closed-form first derivative [251], an explicit formula for c_1 results by plugging in the derivative. Also, note that the value of c_1 depends on the prevalence, but when the same prevalence can be assumed for the target populations of all studies, the prevalence factor $(1-\pi) / \pi$ is the same for all studies. This means c_1 values can be compared even when there is uncertainty about the prevalence. Hence, here we ignore the prevalence factor, so that

$$c_1 = \frac{t'_\alpha(p)}{t'_\alpha(q)}$$

It is important to note that as we included studies from all over the world, the BC incidence rate was probably not the same for all target populations. This might also imply that the prevalence varies across the target populations of the meta-analysed studies, which could be a limitation of the method. Nevertheless, it was not possible to retrieve the prevalence for all the meta-analysed target populations, therefore we opted to assume the same prevalence for all studies. The implicit cost of misdiagnosis was assessed among all reported models, as well as the most important models for each study.

Publication bias

The "escalc" function from the *metafor* package in R was used to calculate the effect sizes and sample variances of the models, which were then used to generate a funnel plot. In order to test for publication bias, Egger's test using the "rma.mv" function was performed.

Identifying new cfc miRNAs associated with BC detection

Cohort and questionnaire data

This project was based on a prospective cohort study (ANDROMEDA) on women of the city of Turin and the province of Biella who attended breast cancer screening in two clinical centres. The target population of the study included women aged between 46 and 67 who were invited to breast screening. The enrolment of women started in July 2015 for the clinical centre of Turin and in May 2016 for the clinical centre Biella, and it lasted until March 2018 for both centres. A total of 26,640 women were included in this study, and the cohort has been followed to date through screening archives and hospital discharge cards to document the onset of potential new BC cases after enrolment. At the time of the screening appointment, all eligible women were offered to participate in the mentioned study. A detailed explanation of the study protocol was given to each participant, who signed a written informed consent form. Women with a personal history of BC, with a severe

disease or who were unable to give informed consent were excluded from the study. Recruited women were also invited to undergo anthropometric measurements (height, weight, waist circumference and body composition) and to provide a blood sample from which the serum, plasma and buffy coat would be extracted and stored. Ethical approval for the study was obtained from the Ethics Committee of each participating centre (Ethical and deontological institutional review board of the A.O.U Città della Salute e della Scienza of Turin with the protocol number 78326 on 11.07.2013, and Ethical Committee of Novara with the protocol number 248/CE and study number CE 27/15). The research was performed in accordance with the Declaration of Helsinki guidelines, and the study was registered in ClinicalTrials.gov with the number NCT02618538 on November 27th, 2015.

Women who agreed to participate were asked, immediately at the enrolment desk, to complete a short questionnaire on general BC risk factors such as reproductive and BC family history, previous breast biopsies, basic physical activity level, BMI and alcohol consumption. Additionally, at a later time-point, they were asked to complete a more detailed questionnaire on diet, physical activity, smoking habits, general state of health and psychological distress.

Lifestyle information was gathered and employed to build a comprehensive lifestyle score, as proposed by Romaguera and colleagues on the EPIC cohort [252], based on the adherence to the World Cancer Research Fund (WCRF)/American Institute for Cancer Research (AICR) recommendations [253]. The lifestyle score ranges from 0 to 8 and sums up the scores of the following eight items, which all have a score of 0, 0.5 or 1:

1) BMI (18.5 to 24.9 = 1; 25 to 29.9 = 0.5; < 18.5 or ≥ 30 = 0)
2) Level of physical activity (manual/heavy manual job, or > 2 h/week of vigorous physical activity, or > 30 min/day of cycling/sports = 1; 15 to 30 min/day of cycling/sports = 0.5; < 15 min/day of cycling/sports = 0)
3) History of breastfeeding (cumulative breastfeeding: ≥ 6 month = 1; cumulative breastfeeding: > 0 to < 6 month = 0.5; no breastfeeding = 0)
4) Consumption of high energy-density foods (energy density ≤ 125 kcal · 100 g^{-1} · day^{-1} = 1; energy density > 125 to < 175 kcal · 100 g^{-1} · day^{-1} = 0.5; > 175 kcal · 100 g^{-1} · day^{-1} = 0)
5) Plant-based foods such as whole grains, vegetables, fruits, and beans (sum of fruit and vegetable intake and dietary fibre intake: F&V intake: ≥ 400 g/day = 0.5; F&V intake: 200 to < 400 g/day = 0.25; F&V intake: < 200 g/day = 0; dietary fibre intake: ≥ 25 g/day = 0.5; dietary fibre intake: 12.5 to < 25 g/day = 0.25; dietary fibre intake: < 12.5 g/day = 0)
6) Red or processed meat (red and processed meat < 500 g/week and processed meat intake < 3 g/day = 1; red and processed meat < 500 g/week and processed meat intake 3 to < 50 g/day = 0.5; red and processed meat ≥ 500 g/week or processed meat intake ≥ 50 g/day = 0)

7) Alcoholic drinks (standards for women: ethanol intake ≤ 10 g/day = 1; ethanol intake > 10 to 20 g/day = 0.5; ethanol intake > 20 g/day = 0)
8) Added salt to food (salt adding to food based on the questionnaire: never = 1; sometimes = 0.5; very often = 0)

We also tested the lifestyle score based on the newer WCRF guidelines from 2020 [254] with two key differences compared to the previous score. One is the addition of consumption of sugar-sweetened drinks (sugary drink intake 0 g/day = 1; sugary drink intake ≤ 250 g/day = 0.5; sugary drink intake > 250 g/day = 0) instead of the score of added salt to food. The other difference is that the BMI score was added to the circumference score (< 80 cm = 1; > 80 cm to < 88 cm = 0.5; ≥ 88 cm = 0) and divided by 2.

Breast density calculation

Standard DMs were performed and read by two expert radiologists. Before the screening program, the radiologists took part in an internally organised workshop to homogenise their breast density assessments. Hence, each mammogram was read by one radiologist and there was no consensus protocol. Additionally, artificial intelligence was not used to read the mammograms. Breast density was calculated during breast examination through two different algorithms: BI-RADS [70] and Tabar [71]. The BI-RADS classified the breast density into category 1– almost fatty (< 25% glandular component), category 2– scattered fibroglandular densities (25 to 50% glandular); category 3– heterogeneously dense (51 to 75% glandular); and category 4– extremely dense (> 75% glandular) [70]. Similarly, Tabar classification was adopted as follows: I (balanced proportion of all components of breast tissue with a slight predominance of fibrous tissue), II (predominance of fat tissue), III (predominance of fat tissue with retroareolar residual fibrous tissue), IV (predominantly nodular densities), V (predominantly fibrous tissue) [71]. For subsequent analyses, considering sample distribution and risk classification, the IV and V categories of Tabar's classification were grouped in a unique category.

The mean and standard deviation were calculated, and a univariate logistic regression analysis between cases and controls was applied on the variables of interest within the questionnaire data as well as on the breast density classifications. The "glm" function in R was used, with the family set to "binomial" to perform the univariate logistic regressions (family set to binomial).

Cases and controls

Incident BC cases were identified, and histopathological information was obtained through record linkage with screening archives, cancer registries and hospital discharge cards. From the large cohort of women enrolled in the ANDROMEDA study, cases and controls were selected among participants who agreed to provide blood samples at the time of recruitment in order to conduct a nested case–control study.

Cases in the discovery cohort were restricted to women with incident BCs diagnosed within June 2018, for whom blood was collected before any treatment (n = 70). Moreover, due to the relatively short time between blood storage and cases/controls extraction, random sampling, without variable matching, of 70 controls from women who did not experience any BC event before June 2018 was performed. No interval cancers were observed among the controls. This cohort of 70 cases and controls was used as a discovery cohort.

A validation cohort, also nested into the study cohort, was extracted as follows. BC cases were obtained from cases diagnosed after June 2018 (these were either women who were recruited in the later window of screening or women who were negative at the time of blood sampling and were diagnosed at a later point in time). Unfortunately, there was a power issue with a freezer that stored the Torino samples, and therefore, the number of validation cases was drastically reduced, relying only on cases observed from the Biella samples. Therefore, all the cases (n = 32) in the validation cohort came from the hospital of Biella, and their details were obtained from the pathology reports. The range of time to diagnosis after blood sampling was from 21 days to 4.3 years (mean: 2.1 years; standard deviation: 1.3 years). A total of 127 healthy controls to be included in the validation cohort were randomly selected in the same way as in the discovery cohort. The validation healthy controls were made up of two subtypes:

1) Women classified as negative for BC during the mammography screening.
2) Women who had a suspicious mammography and underwent a biopsy for a histological examination of breast tissue but were in the end confirmed to be negative for BC.

The selection and number of healthy controls were determined such that for each BC case there are three type 1 and one type 2 healthy controls. The reasoning behind such a selection and sample distribution was to resemble the screening cohort to the extent feasible in this nested case–control study. Intrinsic subtypes of BC in the discovery and validation cohorts were defined using the clinicopathologic surrogate definition reported at the 13th St. Gallen International BC Conference [24].

Missing data on selected clinical, demographic, lifestyle and other non-molecular variables was imputed using the *mice* package from R [255], which allows for selecting variables from which the imputation can be inferred. Imputation had to be performed, as due to the limited sample size, it is a reasonable trade-off to have the already characterised questionnaire data in the context of BC with some errors while keeping the miRNA information and sample size as high as possible within this study.

Blood handling

At the time of blood collection, 6 ml of blood was sampled in two tubes, one with ethylenediaminetetraacetic acid (EDTA) and one with lithium heparin. Both tubes were centrifuged at 2500 revolutions per minute (RPM) and 4 degrees for 10 minutes. The tube

containing EDTA was used for plasma and buffy coat extraction, while the other was for serum extraction, which will not be covered in this thesis as no analyses were performed on it. After centrifugation, without disrupting the part with the buffy coat, the supernatant was then carefully transferred to a 15 ml falcon tube (details on plasma reported below). To transfer the buffy coat, we used a 1000 µl tip, which had a part of its end horizontally cut off. Both tubes, including plasma and buffy coat, had a barcode on them and were stored at −80 °C degrees.

In order to prevent haemolysis, plasma was isolated from EDTA blood tubes within 1h from collection. Blood was centrifuged at 2500 RPM (1250 g) at 4°C for 10 minutes. The supernatant was transferred into new tubes and was centrifuged again at 2500 RPM (1250 g) at 4°C for 10 minutes to remove cell debris and fragments. Plasma was stored in 4.5 ml cryovials at −80°C until its transfer to the Cancer Genomics Lab. For each sample, we calculated the haemolysis score by centrifuging 10 µl of plasma at 1000 g for 5 minutes at room temperature and measuring the absorbance at 385 and 414 nm using the NanoDrop spectrophotometer (Thermo Fisher) with the UV-VIS program [256]. Lastly, 220 µl aliquots were created for each sample and stored in 1.5 ml tubes at −80°C. Samples with haemolysis score < 0.057 or 414 nm / 385 nm absorbance ratio below 2 were kept for further processing.

RNA extraction

Centrifuged aliquots of 220 µl at 1000 g at 4°C for 5 minutes were used for total RNA extraction. The extraction was performed using the miRNeasy serum/plasma kit (Qiagen) following the Exiqon protocol, with the bacteriophage MS2-RNA carrier (Roche Diagnostics) inserted to promote RNA precipitation and purification on membranes. Additionally, the *Caenorhabditis elegans* cel-miR-39-3p miRNA mimic spike-in (Qiagen) was added, although ultimately, it was not used for normalising purposes, as will be clear in the upcoming sections. RNA samples were eluted in 30 µl of nuclease-free water and stored at −80°C.

DNA extraction

For the purposes of PRS calculation and promoter methylation analysis, genomic DNA was isolated from 200 ul of buffy coat utilising the MagMAX DNA Multi-sample Ultra 2.0 kit (Thermo Fisher Scientific, Waltham, MA, USA). DNA concentration and purity were checked by Nanodrop Spectrophotometer (Thermo Fisher).

Small-RNA sequencing

For library preparation, the Ion Total RNA-Seq kit v2 protocol (Thermo Fisher) with the recommendations for low input RNA quantity was followed. First, RNA samples were enriched for small-RNA fraction using a magnetic bead-based technology. After hybridisation and ligation steps, through which 3' and 5' adaptors are directionally and simultaneously attached, reverse transcription was performed to obtain cDNA. cDNA was subsequently purified, size selected and

then amplified using barcoded primers (obtained from Ion Xpress™ RNA-Seq Barcode 01-16 Kit, Thermo Fisher or synthesised by Eurofins Genomics as custom oligonucleotides (barcodes 17-24)). Finally, libraries were purified with beads. All the purification steps were performed by means of the Magnetic Bead Cleanup Module included in the kit. Libraries were checked for yield and size distribution by Bioanalyzer System and DNA 1000 Kit (Agilent Technologies), and differentially barcoded small-RNA libraries were pooled. Pools were checked by Bioanalyzer System and DNA 1000 Kit (Agilent Technologies) to determine the library dilution required for template preparation. Ion Chef™ System (Thermo Fisher) was used for automated templated Ion Sphere Particles preparation and chip loading. Ion 540 chips (Thermo Fisher) were sequenced using the Ion GeneStudio S5 Plus System (Thermo Fisher). Raw sequence reads were processed using the small-RNA plugin available within the Torrent Suite Software version 5.10 (Thermo Fisher). The reads were aligned to mature miRNAs using the bowtie2 alignment software [257], bundled with the plugin. miRNA counts were generated using the featureCounts [258] software from the Subread package 1.5.3.

In order to assess the best normalisation technique for the raw count miRNA data, the *DANA* package from R was used [259]. *DANA* compares several normalisation methods (Total Count, Upper Quartile, Median, Trimmed Median of Means, DESeq, PoissonSeq, Quantile Normalisation, Remove Unwanted Variation) and reports which of them is the most optimal for the given dataset. The "Remove Unwanted Variation" method was not assessed in this case as it was not applicable to our data (output errors despite several debugging efforts, including contacting the authors of the package). In brief, this method exploits the fact that miRNAs with low counts (marked as interval between: $t_{zero} - t_{poor}$) are probably due to handling effects and are positively correlated with each other. These miRNAs are considered negative controls. The t_{zero} and t_{poor} values are selected manually but should usually be around 1-2 and 3-7, respectively [259]. Additionally, the method exploits the fact that only a subset of miRNAs are expressed in a given sample and that a subset of those miRNAs are often organised into polycistronic clusters that tend to be co-regulated and thereby co-expressed. The software uses the miRNA names and defines which miRNAs are found in a mutual polycistronic cluster based on the miRBase database [260] (hairpins separated by < 10kb on the chromosome). Such miRNAs that have a count > t_{well} are considered positive controls. The t_{well} represents the mean count cut-off above which miRNAs are considered well-expressed. For negative controls, the overall strength of inter-marker correlations is quantified, while for positive controls, the direct co-expression relation for each marker pair is assessed. The goal is to test the ability of the mentioned normalisation methods to remove minor correlations among negative controls while minimally affecting the co-expression of the positive control miRNA pairs located in a mutual polycistronic cluster.

For this project, the t_{zero} was set at 2 counts, the t_{poor} at 5, while t_{well} was set at 20 counts. We used a histogram of the normalisation assessment to compare the relative reduction of handling effects and biological signal preservation among the methods. Further, a histogram of the $\log_2$ counts was plotted as well as a plot of mean and standard deviation of the raw counts (marking the t_{zero}, t_{poor}

and t_{well}). Finally, miRNAs with mean raw counts larger than 20 were selected for all further analyses.

Descriptive statistics were performed on the filtered raw counts, and the variance stabilising transformation of the dataset was performed by the *Deseq2* package in R. This included the mean, standard deviation (SD) and coefficient of variation (CV) of the miRNAs. Additionally, a principal component analysis (PCA) was performed, using the "princomp" function in R, while labelling the samples according to their status (tumour or normal). Moreover, using a generalised linear model we performed class comparison between cases and controls on filtered miRNAs using the *Deseq2* package.

RT-qPCR assaying

The expression of selected miRNAs was evaluated by RT-qPCR on a Bio-Rad CFX-96 machine with TaqMan probes. Four µl of RNA from each sample were reverse transcribed using TaqMan MicroRNA Reverse Transcription kit (Thermo Fisher), with a custom pool of selected microRNA primers (Thermo Fisher). Then, 2.5 µl of reverse transcription reaction product was pre-amplified with TaqMan PreAmp Master Mix (Thermo Fisher) and a second pool of selected MicroRNA primers (Thermo Fisher). Preamplified samples were diluted with TE buffer and stored at -20°C for up to one week. A volume of 0.10 µl of diluted preamplified sample was mixed with PCR Master Mix (Thermo Fisher) and water and then transferred in a well of the microRNA plate (10 µL of reaction volume). Custom 96 well plates (Thermo Fisher) with 24 miRNA assays spotted in triplicate were used, allowing for the analysis of one sample per plate. One of the plates for each run also assayed the negative control of the reverse transcription, while all the plates ran a blank negative control on all the miRNAs assayed.

The mean Cycle threshold (Ct) was calculated from the obtained triplicates. Non-detects were replaced with the Ct value of 40. If a replicate within the triplicate was 1 standard deviation away from the mean, it was excluded, and a new Ct mean was calculated.

The RT-qPCR protocol on miRNAs analysed in the validation cohort was the same as the discovery cohort, with the key difference being that miRNAs were assayed in duplicates instead of triplicates. This was done due to the assumption led by results from the discovery cohort that the miRNAs expression would be stable, therefore saving time and materials. Hence, the SD was utilised to remove one replicate, and the average Ct was calculated using all duplicates. Imputation was done in the same way as in the discovery cohort.

For RT-qPCR results in the two cohorts, descriptive statistics were performed on raw Ct values. To investigate the raw Cts, a plot of means on the x-axis to SDs on the y-axis was created for each miRNA, with points coloured based on the status (case or control).

SNP genotyping and polygenic risk score calculation

The SNP genotyping was done on a larger ANDROMEDA sample (384: 115 cases and 269 controls), which included the 131 samples from our discovery cohort [261]. Libraries were prepared starting from 15 ng of DNA and according to the Ion AmpliSeq Library Kit 2.0 protocol for sequencing on the Ion PGM system. The custom panel (Ion AmpliSeq Custom Panel) that selectively covered 77 SNP target sequences was designed through AmpliSeq Designer (www.ampliseq.com). Ion Xpress Barcodes kit (1-16, 17-32 and 33-48), Ion Ampliseq custom Primer Pool and Ion AmpliSeq Library Kit 2.0-384LV were used in conjunction to obtain libraries. The Ion Library Equaliser kit was used to normalise for DNA concentration. Equalised barcoded libraries were pooled and sequenced using Ion PGM Hi-Q OT2 kit and Ion PGM Hi-Q Sequencing Kit on Ion PGM 318 chip V2 on an Ion Torrent PGM (Thermo Fisher Scientific).

Variant calling was performed utilising the Variant Caller plugin within the Torrent Suite Software version 5.10 (Thermo Fisher). The polygenic risk score (PRS) was calculated by adding the multiplications of the log odds ratio of each of the 77 SNPs [109] by the genotype at respective loci (0 for wild type, 1 for heterozygous variant and 2 for homozygous variant). Details on the SNP genotyping, imputation and additional analyses can be found elsewhere [261]. For the PRS on four samples which were not successfully genotyped, we imputed the PRS by taking the mean of the PRS (from the larger cohort) in the respective case or control group.

Methylation profiling of gene promoters

On the 70 cases and 70 controls, we performed methylation-sensitive high resolution melting (MS-HRM) analysis on promoters of three genes: *BRCA1*, *RARB* and *APC*. This was done to determine if any of the methylation profiles were different between cases and controls and whether they could be used in the final logistic regression model (with miRNAs and non-molecular variables). The three genes were selected according to information from the literature on promoter methylation associated with BC diagnosis.

The MS-HRM method exploits the fact that it takes more energy to break the cytosine-guanine bonds than the thymine-guanine bonds, which would happen on loci where the cytosine was not methylated, as the bisulfite treatment of DNA preserves methylated cytosine and converts unmethylated cytosine to uracil which will then be replaced by thymines after the replication by PCR. During the PCR, after the heteroduplex formation, the temperature will gradually be increased, and the bound fluorescent dye will be released, creating an overall decrease in fluorescence. Hence, amplicons from samples with high methylation will melt at higher temperatures than those with lower methylation [140].

The experiment on MS-HRM was performed in the Laboratory of Genomics at the Department of Medical Sciences of Turin and at the Italian Institute for Genomic Medicine in Candiolo by Dr.

Alessia Russo. Briefly, aliquots of the extracted DNA used for the SNP analysis were used for MS-HRM. The DNA integrity was assessed using agarose electrophoresis gel and quantity was measured using a fluorometric method (Qubit broad range assay, Thermo Fisher). A total of 500 ng of DNA was bisulfite converted using EZ-96 DNA Methylation-Gold™ Kit, Zymo.

In each MS-HRM run, a series of six methylation standards (0%, 10%, 25%, 50%, 75%, 100%) were included to estimate the methylation levels in the samples together with an unconverted control DNA sample to check primers specificity for bisulfite converted DNA. This was done by mixing 100% and 0% methylated DNA in different proportions. For each gene, due to the sample size, the analysis was performed on two plates. Further, in the second plate of each gene, an additional methylation standard was added (5%) for a more precise characterisation of smaller methylation values, as we observed relatively low methylation in the first plates. MS-HRM was performed on a 7900HT Fast Real-Time PCR System (Applied Biosystems). Each sample was run in triplicate. Primer sequences were chosen as reported in [262] and the MeltDoctor HRM. Master Mix guide was followed for PCR reactions and conditions. The temperature values at which the fluorescence was measured were the same for all the samples and standards. Hence, the only parameter changing was the aligned relative fluorescence unit (RFU). Data quality analysis was performed using SDS and HRM software (Applied Biosystems).

As the melting data was obtained in triplicates, the average of the triplicates was calculated for both standards and samples. Descriptive plots and statistics (including the melting point) were performed on the replicates using the "meltcurve" function from *qpcR* package [263]. If the melting point of the replicates is 0.2 °C away from the other 2, then it was excluded, and the average was calculated on the other two replicates [264]. From the melting curve data, the triplicate derivative curves were plotted on methylation 0% and methylation 100% standards, to assess the difference in melting curve and melting point between them. Further, a difference plot was created where the fluorescence at each temperature point of methylation standard 0% was subtracted from all other standards. The variability of fluorescence at each temperature point was assessed for standards and samples using the "MFIerror" function from *MBmca* package [265].

Interpolation of methylation

In order to derive the methylation level of a sample it was compared to the methylation standards. First, splines were computed on the standards, to which the samples' values were compared, using the "spline" function in R with the method set to "natural" at 1000 equally spaced points. The splines were then used to generate an interpolation curve. The splines were obtained in two ways, depending on which approach was better for a given dataset:

1) Interpolation based on the maximum relative fluorescence difference of a standard to 0% methylation standard.
2) Interpolation based on the unweighted average fluorescence value across all temperature points.

After obtaining the interpolation curve for the standards, the methylation values for the samples were estimated by either calculating the maximum difference between fluorescence on a given sample and 0% methylation standard or the average fluorescence value. The methylation estimates were then calculated by determining the position of each patient sample's average or maximum RFU among the methylation standard splines. This was done using the "findInterval" function. Then, by dividing this value by the number of interpolations and multiplying by 100, we would get the methylation percentage for that specific sample.

Biomarker screening and validation strategy

In this section, I will report the statistical analyses and research strategies employed to obtain candidate biomarkers associated with BC detection. I will also describe the initial selection procedure in the discovery cohort and the subsequent assessment in the validation cohort.

Discovery cohort

This section will cover the methods used to select the most promising minimally invasive biomarkers within the discovery cohort.

PRS analysis

The normality of the distribution of PRS on the 77 SNPs was tested using the Shapiro–Wilk test ("shapiro.test" function). Summary statistics were calculated, and a density plot of the PRS was created on all samples, as well as samples stratified by BC status. We also tested whether the PRS means significantly differ between cases and controls using a two-sample t-test. In addition, we tested the variance and distribution of PRS using an F-test and Kolmogorov–Smirnov test, respectively. Finally, a logistic regression, using the "glm" function in R (family set to binomial), was created on the PRS score with the BC status as the dependent variable. An ROC AUC was also created based on an apparent validation within the discovery cohort.

MS-HRM analysis

The methylation estimates from the MS-HRM plates were merged for further analyses, and general descriptive statistics were performed. To test for normality, we performed the Shapiro–Wilk test. Class comparison between cases and controls was performed using the Mann–Whitney U test. In order to complement the class comparison, using the "glm" function and setting the family to binomial, we evaluated the classification performance of the methylation data using logistic regression. The performance was estimated by inspecting the model coefficient of the predictor (in this case methylation estimate) or by inspecting the ROC AUC of the prediction for which we separated the original dataset in train (70%) and test set (30%). Bootstrap (n = 2000) was performed on the coefficient and AUC estimates, and their respective histogram depicting values

at each n, as well as the quantiles of standard normal were plotted. Package *boot* was used for all bootstraps with default settings unless specified otherwise [266].

In cases where data contained many zeroes, or in this case many samples with an estimated 0% methylation on a given gene promoter, we employed additional statistical analyses. The two-part statistic separates the non-zero from zero data, performs separate statistical analyses on them and finally merges them for a combined p-value that determines whether the cases and controls differ based on this data [267,268]. Firstly, we compared the proportion of zeros in cases versus controls using the following equation [267]:

$$B_{obs}^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1-\hat{p})\frac{n_1+n_2}{n_1 n_2}}$$

where n1 and n2 the number of samples in cases and controls and p1 = m1/n1, p2 = m2/n2 and p = (m1+m2)/(n1+n2). M1 and m2 represent the number of zero values in the two groups.

Since this statistic follows the chi-squared distribution, we used the "pchisq" function in R at one degree of freedom to obtain the p-value for only this test. Additionally, the continuous part can be analysed using the Wilcoxon rank sum test, Student's t-test, and the Kolmogorov–Smirnov test. In this study we used the Wilcoxon two-sample test using the "wilcox.test" function. However, for the two-part analysis method, we needed to extract the standardised rank sum statistic (W), which is done as follows:

$$W = \frac{R1 - [(n_1 - m_1)(n_1 - m_1 + n_2 - m_2 + 1)/2]}{\sqrt{(n_1 - m_2)(n_2 - m_2)(n_1 - m_1 + n_2 - m_2 + 1)/12}}$$

where R1 is the rank sum, i.e., the sum of the ranks in group 1. Details on the method can be found in [268].

After calculating W, we squared it and added it to the previously obtained B^2 to obtain X^2, which is the statistic for the two-part test. X^2 also follows a chi-squared distribution at 2 degrees of freedom. Hence, we employed the "pchisq" function in R at 2 degrees of freedom to obtain the p-value of the two-part test. We also performed a permutation on the same test statistic and obtained the p-value by counting the number of results as large or larger than the observed X^2 and dividing by the number of permutations. Finally, we also computed a zero-inflated regression model (which fits a generalised additive model for location, scale and shape model for the positive value part, and a logit model for the zero part versus the non-zero part) and tobit regression model (which assumes the data is normally distributed but that the values get censored at 0). The zero-inflated model was created using the "gamlssZadj" function from the *gamlss.inf* package [269], while the tobit regression model was created using the *censReg* package [270]. A bootstrap was also performed on the predictor coefficient for both models.

miRNA analysis

The chief goal of this study was to find potential miRNAs associated with BC onset that could be utilised in the clinics via RT-qPCR, and as there are no optimal RT-qPCR normalisers for small non-coding RNAs [271], we decided to focus on miRNA ratios which will eliminate all laboratory systematic biases (**Figure 8**). Since the initial discovery of miRNAs as biomarkers was carried out utilising small-RNA sequencing, to make the two platforms comparable, we computed the pairwise ratios of filtered miRNAs already within the small-RNA sequencing data.

For any miRNA pairs with identical count profiles (either due to being clustered or to one miRNA having two different names), one was removed, and a unique identifier was assigned to represent the two miRNAs. Some miRNAs with the same name but different chromosomal origin showed different count profiles and were therefore considered as separate. Nevertheless, because the mature sequences of the same miRNA originating from different genomic loci are the same, in the RT-qPCR validation, such occurrences were considered as one single mature miRNA. In order to calculate the ratio between miRNA X and miRNA Y using RT-qPCR data, the following equation was used as explained by Deng and colleagues in 2019 [271]: $Ct_{mean(Y)} - Ct_{mean(X)}$.

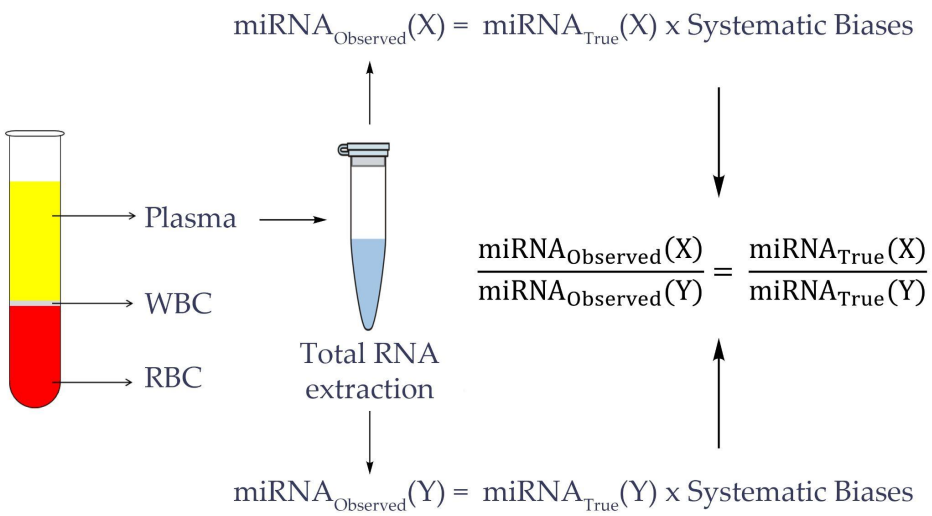


Figure 8. Ratio computation of individual miRNAs to eliminate experimental systematic biases.

miRNA ratios computed based on small-RNA sequencing and RT-qPCR data were descriptively investigated in the same way as the individual small-RNA sequencing miRNAs (see last paragraph of Small-RNA sequencing section in the methods). Additionally, the association between demographic, lifestyle, anthropometric and reproductive factors, as well as cancer characteristics and the RT-qPCR computed ratios, were performed using the Mann–Whitney U test or the Kruskal Wallis test, as appropriate, for categorical covariates and using the Spearman correlation coefficient for continuous covariates. Continuous variables were reported by mean ± SD or median

and I and III quartiles, as appropriate, whereas categorical variables were reported as natural frequency and percentage.

On miRNA ratios computed from small-RNA sequencing raw counts, two-sample Wilcoxon test was performed, using the "wilcox.test" function, to compare miRNA ratios between cases and controls and p-values were corrected using the Benjamini–Hochberg method. The fold-change was calculated based on the median in cases and controls. Ratios significantly different between cases and controls (two-sample Wilcoxon test adjusted p-value ≤ 0.01) with a fold change > 2 or < 0.5 were selected as strategy 1. In contrast, ratios with a coefficient of variation < 0.5 within controls and significantly different between cases and controls, without setting any criteria on fold change, were selected as strategy 2. The ratios from the two strategies were further analysed, and the most promising were selected by a Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression. Five-fold cross-validation was used to preliminarily assess the performance of the model-selected ratios, separately for the two strategies defined above. Thus, the sample was randomly divided into five groups, called folds, and the LASSO logistic model was trained on five minus one folds using the "cv.glmnet" function from the *glmnet* package [272]. Then, the performance of the resulting model was evaluated in the remaining part of the data. This procedure was repeated for each fold, and the performances were averaged across folds. The following performance measures were considered in the five-fold cross-validation: calibration intercept, Cox's measure of spread (often called "calibration slope") [273], scaled Brier score, and ROC AUC. The first three measures mainly relate to the agreement between the observed outcomes and the outcomes predicted from the model. For the intercept and scaled Brier, ideal values should be as close to zero as possible, whereas for the Cox calibration slope, as close to one as possible. The AUC refers to the model's ability to discriminate between individuals with a different outcome and the ideal values should be close to one.

For exploratory purposes, we also employed the hierarchical shrinkage model (HSM) method based on the horseshoe priors, using the *hsstan* package in R [274]. The horseshoe priors are implemented to obtain a reduced list of informative predictive biomarkers. Four chains and 2000 iterations were performed for each HSM run. This was performed on all computed ratios and ratios in strategy 1 and strategy 2. This approach could be considered a useful alternative when the event-to-variable ratio in a study is low, as is the case in this project. The performance measures explained above were also used on the set of ratios selected by the horseshoe priors.

Biomarker panels

On RT-qPCR data (i.e., the selected miRNA ratios from small-RNA sequencing analysis), univariate odds ratios (OR) and corresponding 95% confidence intervals (CIs) were obtained using standard logistic regression. The linearity assumption between a continuous predictor and the logit of risk was inspected through the Locally Weighted Scatterplot Smoothing (LOWESS) and restricted cubic splines, whereas for ordinal variables the Cochran–Armitage trend test was used

to assess the presence of a linear trend. To derive a ratio-based signature as well as to preliminarily investigate the potential added value of miRNA ratios over more conventional BC risk factors and their potential independent role in predicting BC risk, the LASSO logistic regression was used. Three models were then fitted: one using miRNA ratios only, one combining the ratios with other potential BC risk factors and one on BC risk factors alone. To select BC-associated factors for inclusion in the model together with the miRNA ratios, we assessed the association between BC detection and other factors such as PRS, methylation profile of *RARB*, *APC* and *BRCA1* promoters, demographic, family, reproductive and screening history, lifestyle, and breast density information, as well as any interaction between them relevant to BC. The discriminatory ability of the models was assessed using ROC AUC (with reported 95% CIs), whereas the Youden index was used as the criterion to derive a cut-off point on the predicted probabilities and compute sensitivity and specificity. The paired Delong test was used to compare the discrimination among different models.

The correlation of the same ratio between platforms was estimated using the Spearman correlation within the "cor.test" function. We also compared the class comparison and OR results for each ratio between the platforms.

Validation cohort

The LASSO logistic regression coefficients of the selected variables obtained from the discovery cohort were applied in the validation cohort. Notably, the RT-qPCR miRNA ratios in the validation cohort were computed in the same way as in the discovery cohort. This was performed on the model including non-molecular variables only, the model with only miRNA ratios and the combined model. To obtain the probability of BC diagnosis for each sample in the validation cohort, the coefficients were applied using the following sigmoid function used for calculating the probability of an event. For a number of predictors in the model with a coefficient:

$$y = \frac{1}{1 + e^{-(B_0 + B_1X_1 + B_1X_2 + \cdots + B_1X_n)}}$$

where y is the predicted probability, B_0 is the intercept and B_1X_n is the coefficient of the n^{th} predictor. To evaluate the calibration of the model, we computed a logistic calibration curve, which is obtained by creating a new logistic regression model where the dependent variable is the outcome and the independent variable is the log odds of the predicted probabilities. This was done using the "val.prob.ci.2" function in the *CalibrationCurves* package [275–277]. We then investigated the calibration intercept ('calibration-in-the-large') and the calibration slope as explained in the vignette of the package. Additionally, just as in the discovery cohort, we assessed the ROC AUC and the Brier score of the predictions.

To calibrate any potentially miscalibrated models, we used the so-called closed-testing procedure to select the optimal model updating method [278]. This approach is based on a series of likelihood

ratio tests of updated models compared to the original model. After selecting a p-value for the hypothesis that the model we are assessing does not need updating, complete model revision (i.e., the coefficients are re-estimated) is tested against the original model. If the model revision is significantly better than the original model, we proceed to test the model revision against only recalibrating the intercept. If this test is not significant, intercept calibration will be selected. However, if it is, the model revision will be selected and compared against calibrating both the intercept and slope. If it is significant, model revision is the final selected updating procedure, and if not, then the recalibration of intercept and slope is selected. The reason for performing such a closed testing approach is to avoid increasing the Type I error if all the model updating options were assessed separately. To determine the best model calibration approach based on the closed-testing approach, a function was created in R, as reported in the appendix of Vergouwe and colleagues [278]. For complete model re-evaluation (including the coefficients, intercept and slope), we used the ridge regression to reduce some of the overoptimism and overfitting that would arise from a regular logistic regression. The function used was "cv.glmnet" from the *glmnet* package, with alpha set at 0. Further, to additionally assess overfitting, we performed a bootstrap on the ROC AUC of the ridge regression to gain insight into the performance range on the re-evaluated model.

The alternative model calibration method, in this case used as a comparison to the frequentist ridge regression, was the Bayesian approach where the coefficients of the discovery would be used as means of prior probabilities and a constant of log(4)/2 would be used for the standard deviation of the priors [279,280]. The Bayesian approach is especially recommended for smaller sample sizes [281].

Finally, we used the internal-external cross-validation (IECV) approach to merge the discovery and validation cohort data to identify generalisable predictors and to identify whether a model constructed on the merged cohorts would be more informative. IECV merges the individual participant data and then trains the model on K-1 cohorts and validates it on the remaining one [282]. This is done iteratively until the most optimal model characteristics have been obtained for the given data. Heterogeneity between the cohorts (based on the Brier score) of the IECV model was assessed, and in case of large heterogeneity, it is not advisable to merge the cohorts for a combined model. Nevertheless, the same approach also allows for reducing the heterogeneity between the cohorts by starting from an intercept-only model and iteratively adding predictors until the heterogeneity is optimised. Additionally, for the IECV we used the restricted maximum likelihood (REML) random effect meta-analysis, with the model being estimated in the first stage using Firth's correction (similar to a generalised linear model in performance but recommended by authors in cases of small sample size).

Subgroup and sensitivity analyses

As there are many different subgroups within the analysed cohort of this study (such as different molecular subtypes, differences based on the PRS score or family history), the identified

biomarkers are assumed to be generalisable across these subgroups. To test this, I analysed the distribution and variances of the identified biomarkers. When only two subgroups were compared, this was done using the Kolmogorov–Smirnov test ("ks.test" function), Mann–Whitney U test, and F test ("var.test" function). When more than two subgroups were tested, I used the Levene test ("leveneTest" function in R) for testing the equality of variances, the Kruskal–Wallis ("kruskal.test" function) for testing the mean ranks between the groups and Anderson–Darling ("ad.test" function). Lastly, I performed various sensitivity analyses by removing variables and testing the performance of the underlying models. The sensitivity and subgroup analyses were performed on both the discovery and validation cohorts.

Validation in TCGA

We downloaded the TCGA processed raw counts of microRNAs on BC tissue samples and adjacent healthy tissue (in February 2023). This was done using the "gdcRNADownload" function from the *GDCRNATools* package [283]. The project.id was set to "TCGA-BRCA" and the data.type to "miRNAs". Metadata was then obtained and merged using "gdcParseMetadata" and "gdcRNAmerge" functions, respectively. The metastatic samples and duplicate replicates from Formalin-Fixed Paraffin-Embedded (FFPE) blocks were excluded from all analyses, for a total of 1,078 cases and 104 controls. There were 103 paired tumour and healthy tissue samples. Additionally, we filtered out all miRNAs with a raw mean count of ≤ 20 across the samples.

On the filtered list of miRNAs, using the *Deseq2* package, we performed a paired class comparison with the focus on the miRNAs which were selected in the discovery cohort. Then, we computed pairwise ratios to obtain the same list of ratios selected in the discovery cohort and performed a paired Wilcoxon two-sample test on the paired tumour and adjacent healthy tissues. Additionally, we computed a conditional logistic regression, used for paired samples, on the selected ratios using the "clogit" function from the *survival* package [284,285]. The function "strata" is used within the formula of the "clogit" function to indicate the variable which determines the paired samples.

Target enrichment and network analysis

Target enrichment and functional enrichment analyses were performed on miRNAs of interest using the Mienturnet online software [286]. The miRTarBase was used to obtain the list of targets and for the functional enrichment as it is based on experimental validation. The functional enrichment output included the KEGG and WikiPathways databases as well as the Reactome and disease enrichments.

The network analysis on miRNAs of interest was performed utilising the MetaCore database from Clarivate. Through the MetaCore software we looked at the transcription factors, canonical pathways as well as significant network (z-score > 60) interactions of the input miRNAs/genes and related molecules added by the software.

Puberty-associated CpGs linked to BC and miRNAs

Since early pubertal timing is a risk factor for BC [287], in this section I report the methods of investigating DNA methylation sites (CpG sites) associated with puberty and linked to BC processes or BC risk. This was a part of my secondment project at FIMM, where we identified DNA methylation sites associated with puberty and tried to functionally describe them through enrichment analyses. Additionally, we employed twin modelling techniques to identify heritable CpG sites or CpG sites whose association with puberty was non-genetically driven [234].

I investigated which microRNAs are relevant to the genes mapped on the CpGs associated with puberty that are linked to BC. This would result in a set of epigenetic biomarkers associated with puberty and BC. Additionally, a comprehensive analysis, in the context of genetics and epigenetics, was performed on DNA methylation sites to elucidate how they are relevant to breast cancer in the context of puberty and if some of them are linked to BC risk, onset or progression. The CpG sites which were investigated here were retrieved in two ways:

1) A total of 2,711 CpG sites associated with puberty which were linked to BC processes through the Ingenuity Pathway Analysis (IPA) knowledgebase [234], referred to from now on as set 1.
2) CpGs which were associated with puberty and BC risk, referred to from now on as set 2:
    a) CpGs associated with risk based on the EPIC study [143].
    b) CpGs which are differentially methylated between monozygotic twins discordant for BC – non-genetic drivers of BC risk manifested on those CpG sites [288].

miRNAs targeting genes mapped to CpGs linked to BC

To determine which miRNAs significantly target the genes mapped to CpGs of set 1 or set 2, the Mienturnet software, introduced above, was used. The miRTarBase database was used, and the number of minimum interactions was set at 2, with a false discovery rate (FDR) cut-off for target enrichment set at 0.05 and 0.2 for set 1 and set 2, respectively. The downloaded output file included the miRNA, p-value as well as FDR of target enrichment, OR, number of targeted genes and the list of genes.

Datasets used

Two datasets were used to assess the methylation status of the CpG sites associated with BC in tissue and blood: TCGA and EPIC methylation data. The TCGA tissue DNA methylation data was downloaded (in February 2023) using the "GDCquery" followed by the "GDCdownload" functions (project = "TCGA-BRCA" and data.type = "Methylation Beta Value") from the *TCGAbiolinks* package [289] and the dataset consisted of tumour tissue and adjacent healthy tissue samples. The summarised experiment object was obtained using the "GDCprepare" function, followed by the assay function which gave us the methylation Beta values for the mentioned

samples. Metadata was obtained from the initial object obtained from the "GDCquery" function. We removed FFPE sample duplicates and metastatic samples for a total of 784 cases and 97 controls. There were a total of 57 tumour and adjacent healthy tissue pairs.

The EPIC data was obtained from the gene expression omnibus database (GSE51057) and consisted of 152 cases and 177 controls, all of which were prospectively sampled with a follow-up until diagnosis. Like the TCGA dataset, the EPIC methylation data consisted of pre-processed Beta-values.

We were also interested in the gene expression of the genes mapped to CpG sites of interest. Therefore, the TCGA tissue gene expression data was downloaded (in February 2023) using the GDCquery followed by the GDCdownload functions (project = "TCGA-BRCA" and data.type = "Gene Expression Quantification") from the *TCGAbiolinks* package, and the dataset consisted of tumour tissue and adjacent healthy tissue samples. The summarised experiment object was obtained using the "GDCprepare" function, followed by the assay function which gave us the raw mRNA counts for the mentioned samples. Metadata was obtained from the initial object obtained from the "GDCquery" function, and we removed FFPE sample duplicates and metastatic samples for a total of 1,095 cases and 113 controls. The downloaded data included raw gene counts, and there were 58 tumour and adjacent healthy tissue pairs.

We also explored relevant mature miRNAs, in both tissue and blood, to the identified CpG sites and their genes. For tissue miRNAs, we used the already described TCGA miRNA data (see section Validation in TCGA), and for plasma miRNA data, we used our small-RNA sequencing data, which was also described above.

Dataset analysis

miRNAs found to significantly target genes mapped to either of the two CpG sets were tested for differential expression in tissue and blood. For miRNAs in tissue, a paired class comparison design was created using the *Deseq2* package, while for miRNAs in blood we used the regular class comparison design using the *Deseq2* package, which was described in the sections above.

Using a paired Wilcoxon two-sample test, we determined the differentially methylated CpG sites in TCGA tissue data, which included tumour tissue and adjacent healthy tissue samples. On the other hand, we tested for differential methylation in blood CpGs using a regular two-sample Wilcoxon test, as the samples were not paired. The p-values were corrected for multiple testing using the Benjamini–Hochberg method.

For count data of miRNAs and genes, using the *Deseq2* package, we performed paired class comparisons between tumour and adjacent healthy tissues. In all analyses, the p-values were corrected for multiple testing using the Benjamini–Hochberg method and the significance threshold was set at adjusted p-value < 0.05.

Since some samples available on TCGA have data on small-RNA sequencing, RNA sequencing and methylation, we performed a correlation analysis (using the Pearson method) between differentially expressed genes and miRNAs significantly targeting the list of genes which are mapped to CpGs associated with puberty and linked to BC processes. Further, for all CpG sites associated with puberty and linked to BC processes (set 1) or associated with BC risk (set 2), we tested the correlation (using the Pearson method) between their methylation value and gene expression of their respective mapped gene. The correlation p-values were corrected for multiple testing using Benjamini–Hochberg method.

Super-enhancers

Enhancers are genomic elements, usually regulated by numerous transcription factors, that can activate gene transcription regardless of its orientation on the DNA strand [290]. Super-enhancers, on the other hand, are usually made up of multiple "stitched" enhancers, for which transcriptional coactivators, most commonly Mediator (*Med1*), have a strong binding affinity [290,291]. Due to their importance for gene regulation, I aimed to identify which CpG sites associated with puberty and BC are located within super-enhancers. Super-enhancers were obtained from a super-enhancer database (SEdb 2.0) and were based on one breast epithelium sample [292]. According to the database, the sample's data source is ENCODE. After downloading the list of super-enhancers, their genomic coordinates were converted from GRCh38/hg38 to GRCh37/hg19, and I checked which CpG sites have their genomic locations within one of the super-enhancers of the breast epithelium sample. We then matched those CpG sites with the corresponding super-enhancer ID.

Network Analysis

The network analysis on genes mapped to CpG sites of interest was also performed using the MetaCore database from Clarivate. In addition to the parameters explained in the network analysis on miRNAs, due to the larger number of input parameters (genes), we also looked at the direct interaction networks.

Regulatory functions of CpG sites of interest

The identified lists of CpG sites were filtered based on their relevance to regulatory elements or super-enhancers, as well as being located on exons of a gene or having a high correlation between CpG methylation and respective gene expression. The potential regulatory elements of a locus were determined using the "SCREEN: Search Candidate cis-Regulatory Elements by ENCODE" [293]. The genomic locations of the CpGs had initially to be converted from GRCh37/hg19 to GRCh38/hg38 utilising the "Lift Genome Annotations" within the UCSC genome browser website. The DNase, H3K4me3, H3K27ac and CTCF max z-scores were explored as an average across all tissues, as well as in breast epithelium and peripheral blood. The chromatin states of the HMEC and GM12878 cell lines were also assessed using the "Chromatin State Segmentation by HMM from ENCODE/Broad". The HMEC and GM12878 cell lines were of focus as they were

supposed to represent the mammary and blood cells relevant to this project, respectively. Further, the intron or exon location of the CpG and the potential expression of the underlying transcript (in transcripts per million – TPM) in mammary tissue was identified utilising the UCSC genome browser (Human genome GRCh37/hg19). The subset of CpGs or genomic loci, deemed to be more likely to have functional relevance, were selected based on the following criteria (minimum one needs to be satisfied):

1) If the CpG is found within a super-enhancer.
2) If the locus at both HMEC and GM12878 cell lines has a relevant regulatory function (active promoter, strong enhancer, transcriptional elongation or insulator), and the mapped gene has a significant correlation with the DNA methylation of the found CpG (adjusted $p < 0.01$) in TCGA data.
3) If the underlying CpG is found on the exon of the mapped gene and the mapped gene has a significant correlation with the DNA methylation of the found CpG (adjusted $p < 0.01$) in TCGA data.
4) If all four regulatory markers are found across the average of all tissues as well as breast epithelium, and one of the markers has a z-score > 1.64 (that is the cut-off for high presence of that marker). According to ENCODE, if all four markers are available, it is possible to infer the group of the candidate cis-Regulatory Elements (cCREs).

We performed an additional network analysis on this final list of filtered genes for potentially more specific results. Additionally, a comprehensive literature search (using PubMed, Google Scholar and GeneCards) was performed on the filtered lists of loci with the focus on their mapped genes. The keywords in the search were: "Gene name" AND puberty or "Gene name" AND breast cancer. This was done to identify the genes already associated with puberty and BC onset or progression. The GeneCards database was investigated for functional description and identifying relevant SNP associations at the target genes. We also used the Harmonizome gene knowledgebase to identify genes associated with precocious puberty [294]. Lastly, these genes were also examined in the human protein atlas [295,296] (https://www.proteinatlas.org), within which the following information was obtained:

1) Protein and RNA expression in breast tissue.
2) Cell types expressing the mRNA in breast tissue or peripheral blood mononuclear cells (PBMC) based on single-cell RNA sequencing data.

Results

Diagnostic meta-analysis on cfc miRNAs

A total of 1,165 publication hits were obtained after performing a search in two databases (PubMed and NCBI PMC) and the Google Scholar search engine (**Figure 9**). PubMed and NCBI databases yielded 449 and 235 publications, respectively. The Google Scholar engine yielded 481 hits. After removing duplicates (n = 443), 722 unique publications were obtained. The type of publication, title and keywords were evaluated in the initial eligibility assessment, while the abstract was evaluated in the secondary eligibility assessment. In the initial and secondary eligibility assessment, 397 and 145 publications were excluded, respectively. The final, full-text, eligibility evaluation was performed on 180 articles, of which 124 were excluded. Hence, a total of 56 articles remained eligible for the meta-analysis. A generalised summary of the exclusion reasons for all three eligibility evaluation steps is shown in **Table 2**. The supplementary table with the complete list of reasons and their frequencies can be found in the published meta-analysis [232].

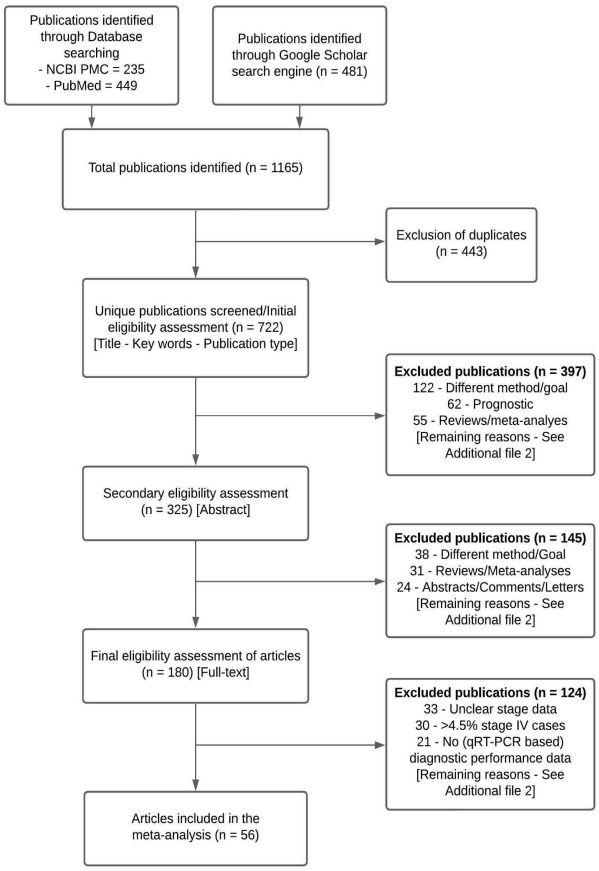


Figure 9. Flow diagram of the selection procedure for the inclusion of studies in the meta-analysis.

Table 2. Summary of the exclusion reasons for all three eligibility evaluation steps.

Reason for exclusion	Number
Abstracts/Comments/Letters	60
Metastatic focus	17
Dubious article/Language/Not found	45
Different method/Goal	162
No performance data	54
Too specific subtype of BC	11
Unclear stage data	33
> 4.5% stage IV samples	31
Review/Meta-analysis	86
Prognostic	68
Not related to BC	3
Exosomal miRNAs	23
Therapeutic	47
Not biomarker focused	26
Total excluded publications	**666**

Included studies

Within the 56 studies that analysed the performance of circulating miRNAs in diagnosing BC using RT-qPCR, a total of 3,894 cases, 2,948 controls and 647 benign patients were included. The sample size range of BC patients in the studies was from 15 to 180 with a mean of 69.5 (median = 58), while the range of controls was from 10 to 199 with a mean of 52.6 (median = 40). Additionally, the range of benign patients was from 0 to 196, with a mean of 11.6 (median = 0).

The recorded case and control number of each study is based on the model within each study with the largest case/control number. The studies were conducted in 15 different countries: Belgium (n = 1), China (n = 21), Egypt (n = 7), Germany (n = 3), Indonesia (n = 1), Iran (n = 6), Iraq (n = 1), Kazakhstan (n = 1), Lebanon (n = 1), Mexico (n = 2), Rwanda (n = 1), Singapore (n = 1), South Korea (n = 2), Spain (n = 4), USA (n = 3) and one included samples from multiple institutions. Hence, 8 studies were conducted in Africa, 34 in Asia, 8 in Europe, 5 in North America and one study was multicontinental. The publishing date for the studies ranged from 2010 to 2022, with the majority of the studies published in 2021 (**Figure 10**).

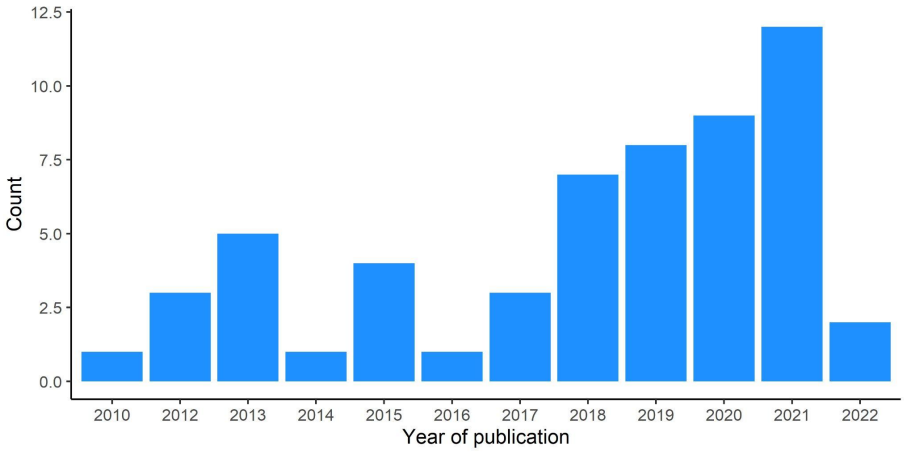


Figure 10. Frequencies of years of publication within the meta-analysed studies.

Seven of the 56 studies included stage IV breast cancer patients, with a proportion of 4.5% or less of the total cancer patient cohort. The remaining 49 studies did not include any stage IV cases. The proportion of stages for all the studies, also stratified based on the inclusion of stage IV cases, can be seen in **Table 3**. More than 75% of the cases were stage 0, I or II. Ten of the 56 studies did not report diagnostic performance data but reported ROC graphs with AUC values, while three studies did not report ROC graphs with AUC values but reported only diagnostic accuracy in terms of sensitivity and specificity. Key information about the included studies can be seen in *Supplementary Table 1* (Appendix A). The 56 studies reported a total of 173 different models. Among them, 121 analysed single miRNA performance, which covered a total of 68 unique miRNAs. On the other hand, 52 models analysed panels of miRNAs and their performance, covering 55 unique miRNAs. Moreover, 82 models had plasma as the specimen type, 81 had serum, and 10 had whole blood. It is worth restating that, in addition to the analyses performed on all the reported models, this meta-analysis also evaluates one model per study (n = 56), the most important model per study.

Table 3. Average percentage of TNM stages within the meta-analysed studies. Stages 0, I and II were grouped together because they are commonly referred to as early stages.

		Stage 0-I-II %		**Stage III %**		**Stage IV %**	
	N	**Mean**	**Missing**	**Mean**	**Missing**	**Mean**	**Missing**
Overall	56	77.36	15	20.67	15	0.15	1
With stage IV	7	76.35	2	17.34	2	1.14	0
Without stage IV	49	77.50	13	21.13	13	0	1

QUADAS-2 risk of bias assessment

The QUADAS-2 assessment was performed on the 56 included studies. More than 75% of studies had a low probability of having an index test and patient selection applicability concern, while

82.1% of studies had a low probability of having a reference standard applicability concern. On the other hand, 41.1% of the studies had a low risk of bias within the patient flow and timing category. Despite the low probability of applicability concern for the index tests in most of the meta-analysed studies, only 44.6% of studies had a low probability of risk of bias coming from the index test. Nevertheless, in the index test category, only 16.1% of the studies had a high probability of bias (**Figure 11A**). Interestingly, only 8.9% of the studies performed or explicitly stated that prospective sampling was performed. This is also associated with the fact that in most meta-analysed studies, blood was collected after the biopsy was performed on the patient. Additionally, 50% of studies explicitly stated that blood collection was performed before surgery (**Figure 11B**).

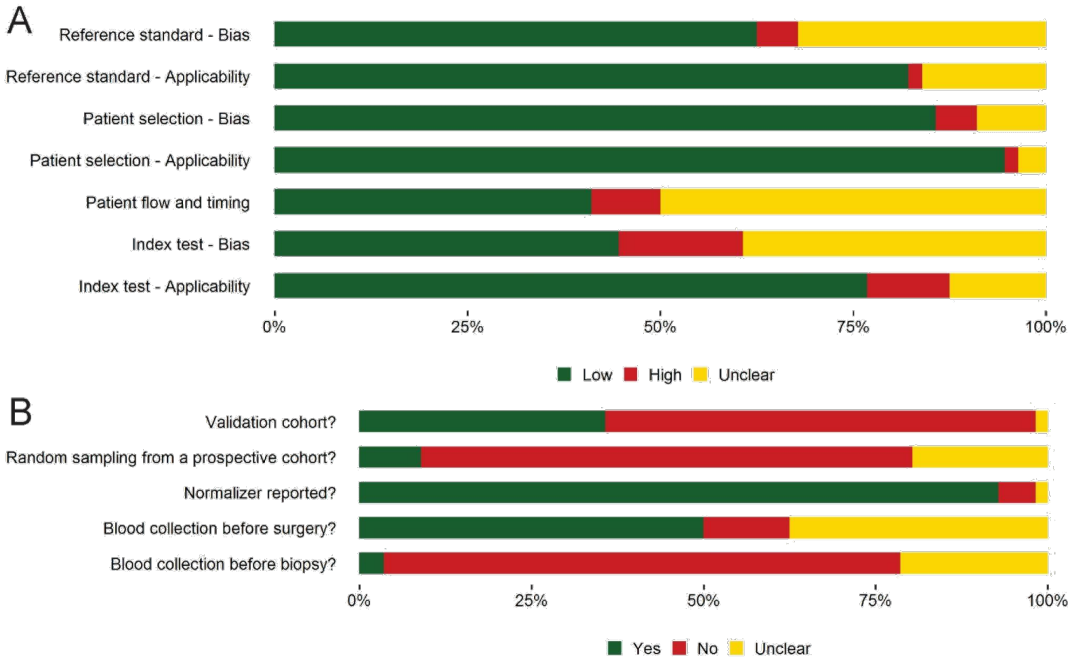


Figure 11. Summary of the QUADAS-2 evaluation performed on 56 articles. Proportions of Low risk of bias (Yes), Unclear and High risk of bias (No) are shown for A) key questions on applicability and bias and B) most important signalling questions.

Descriptive statistics

Both sensitivity and specificity reports were heterogeneous across models (sensitivity: $X^2 = 1171.8$, $p < 0.001$; specificity: $X^2 = 1019.3$, $p < 0.001$). The X^2 estimate of equal proportions describes whether the observed differences in sensitivity or specificity are only due to chance. For both sensitivity and specificity, the X^2 was statistically significant, indicating that factors such as miRNA tested, population, research design, etc., were the cause for the differences in the observed sensitivities and specificities. In addition, in the same group of models, a negligible positive correlation r = 0.09 [-0.08 to 0.25] of sensitivities and false positive rates (FPRs) was found. Forest

plots of sensitivity and specificity were based on the most important models per study and can be seen in **Figure 12A** and **Figure 12B**, respectively.

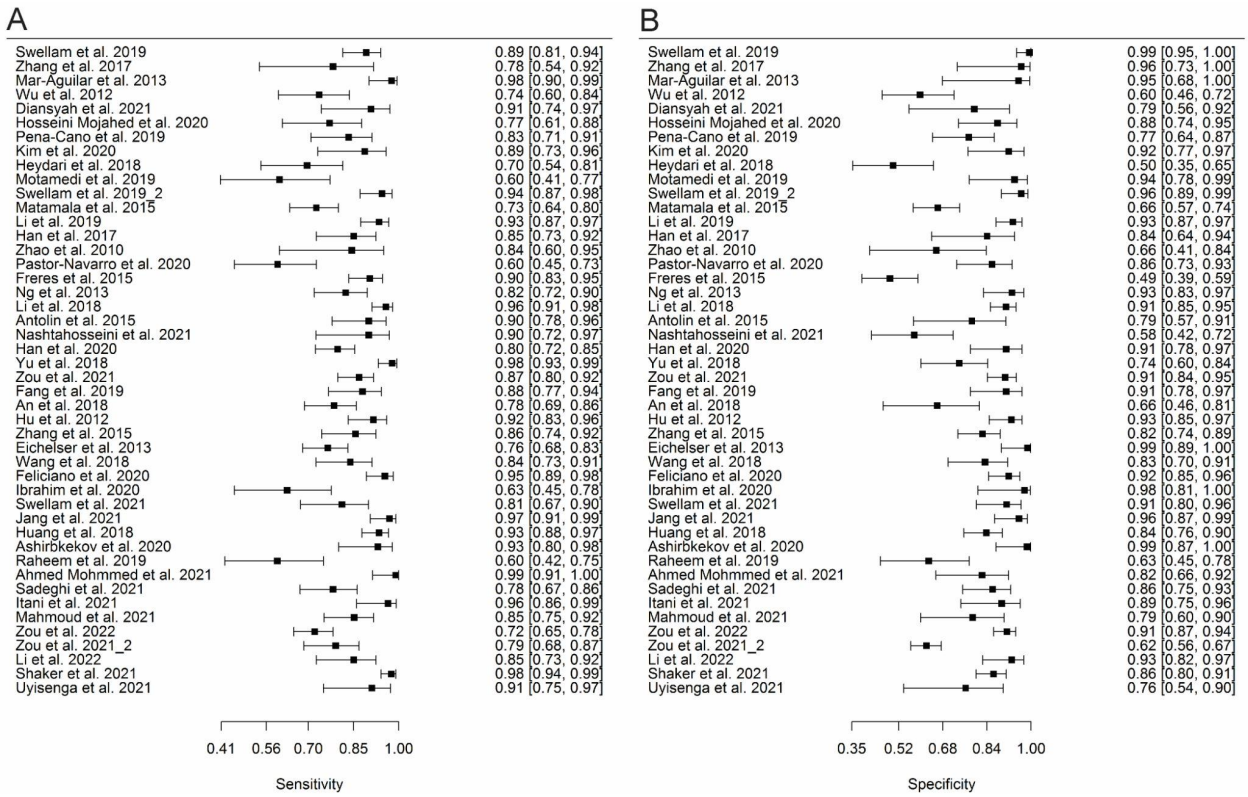


Figure 12. Forest plot of A) sensitivities and B) specificities of the most important model from each study. The respective values and their confidence intervals can be seen on the right side of each plot.

Bivariate analysis

A pooled estimate of 0.85 was obtained for sensitivity and 0.83 for specificity on all the reported models with performance data (146 models). For the most important model per study (46 models), slightly better pooled sensitivity (0.88) and specificity (0.88) were obtained. Confidence intervals, variances of logit transformed sensitivity and FPR as well as the correlation estimates for both bivariate models can be found in **Table 4**. The SROCs of the two models are shown in **Figure 13A** and **Figure 13B**.

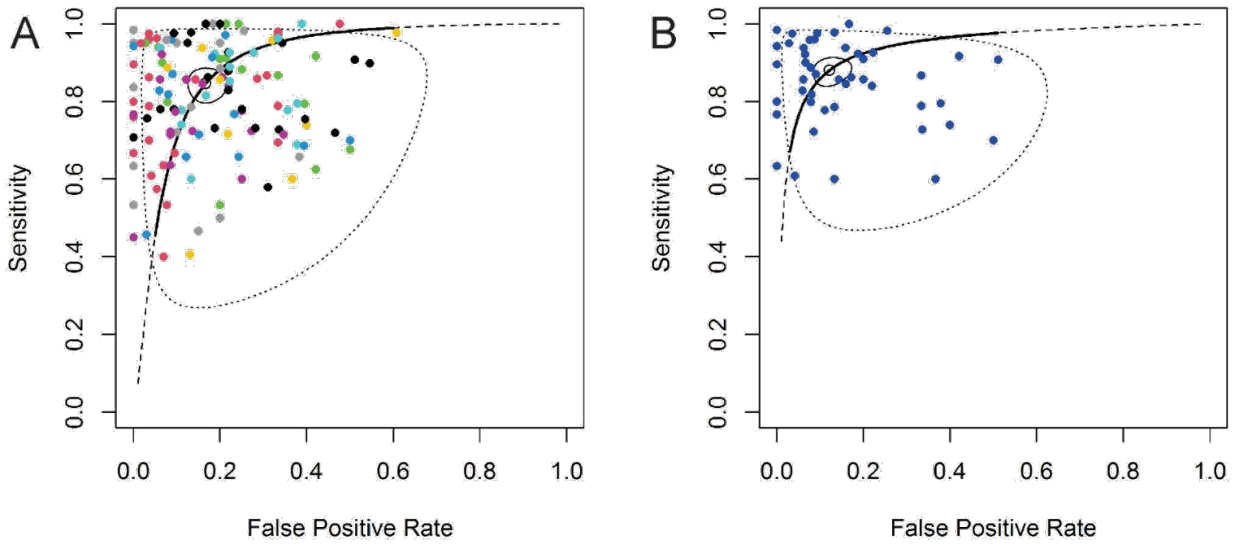


Figure 13. SROCs of the bivariate models. A) SROC of all reported models. Points with the same colour in the graph represent models originating from the same study. B) SROC of the most important model from each study.

Table 4. Summary of the bivariate analyses on all reported models and on the most important model per study.

		Fixed Effects		**Random Effects**					
				Model			**Study**		
		Estimates	CI	SD	Corr.	n	SD	Corr.	n
All reported models	Sens.	0.85	[0.81, 0.88]	0.85	-0.17	146	0.70	0.06	46
	Spec.	0.83	[0.79, 0.87]	0.60	-0.17	146	0.74	0.06	46
Most important models	Sens.	0.88	[0.85, 0.91]	0.86	0.23	46			
	Spec.	0.88	[0.84, 0.91]	1.00	0.23	46			

To account for the experimental and study-design differences among studies, fixed effects were added to the bivariate mixed models (specimen type, normaliser, single or multiple miRNA panel and inclusion of stage III and/or stage IV cases). The significant fixed effects for all models were the single or multiple panel type as well as the normaliser type, whereas for the most important models there were no significant fixed effects. Details on the fixed effect models can be found in **Table 5** and **Table 6**.

Table 5. Bivariate generalised linear mixed effect model on all reported models adjusted for covariates.

Fixed effects	Estimate	SE*	Z	p-value
Sensitivity	2.01	0.34	5.86	< 0.001
Specificity	2.03	0.33	6.18	< 0.001
Specimen type: Serum	0.14	0.27	0.52	0.60
miRNA panel: Single	-0.61	0.18	-3.48	< 0.001
Normaliser: Exogenous	-0.68	0.26	-2.58	0.01
Inclusion of stage III: True	0.07	0.20	0.33	0.74
Inclusion of stage IV: True	0.03	0.32	0.10	0.92

*Standard error

Table 6. Bivariate generalised linear mixed effect model on the most important model of each study adjusted for covariates.

Fixed effects	Estimate	SE	Z	p-value
Sensitivity	1.89	0.49	3.88	< 0.001
Specificity	1.94	0.49	3.99	< 0.001
Specimen type: Serum	-0.05	0.33	-0.16	0.87
miRNA panel: Single	-0.40	0.33	-1.21	0.23
Normaliser: Exogenous	-0.22	0.50	-0.43	0.67
Inclusion of stage III: True	0.38	0.34	1.14	0.25
Inclusion of stage IV: True	-0.08	0.44	-0.18	0.86

Influence analysis and outliers

Outlier analysis was performed on the complete set of models and was based on the odds ratio. Models with an odds ratio of 2 SDs away from the mean were considered outliers. A total of five models were identified as outliers. In order to detect influential models in the two generalised linear multilevel models mentioned above, Cook's distances of the included models were calculated (**Figure 14A** and **Figure 14B**). Models with a Cook's distance more than 2 SDs away from the mean were deemed as very influential.

On all reported models, eight of them were influential. Interestingly, none of the models from the outlier analysis matched the ones obtained from the influence analysis. Generalised linear multilevel models without the influential models were fit to determine statistical robustness; a pooled estimate of 0.84 [0.80 to 0.87] was obtained for sensitivity and 0.84 [0.80 to 0.88] for specificity. On the most important model per study, three models were found to be influential. After repeating the generalised linear multilevel model, pooled sensitivity and specificity were 0.87 [0.84 to 0.90] and 0.86 [0.82 to 0.89], respectively. A very modest discrepancy was observed between the bivariate analyses with and without the influential models. This was observed for estimates on all and most important models, indicating the robustness of the pooled estimates.

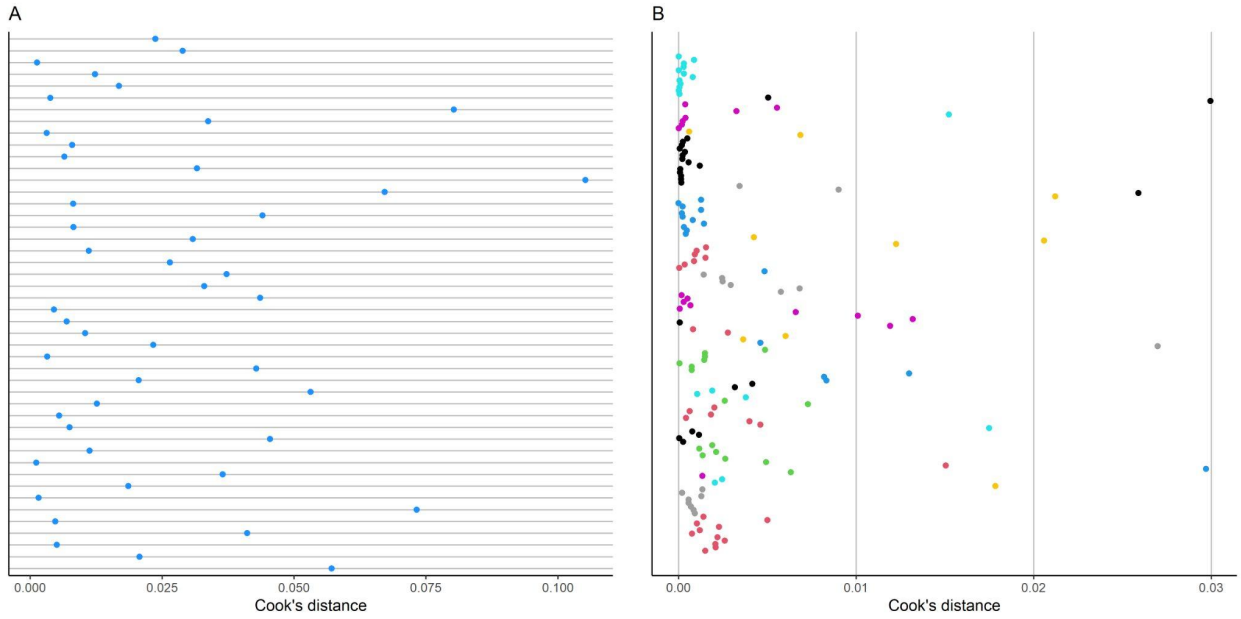


Figure 14. The calculated influence analysis (represented in Cook's distance units) on the included models. A) Influence analysis of the most important models from each study. B) Influence analysis of all reported models where the points with the same colour represent models originating from the same study.

We also identified the most influential studies while accounting for all reported models. There were three highly influential studies, as can be seen in **Figure 15**.

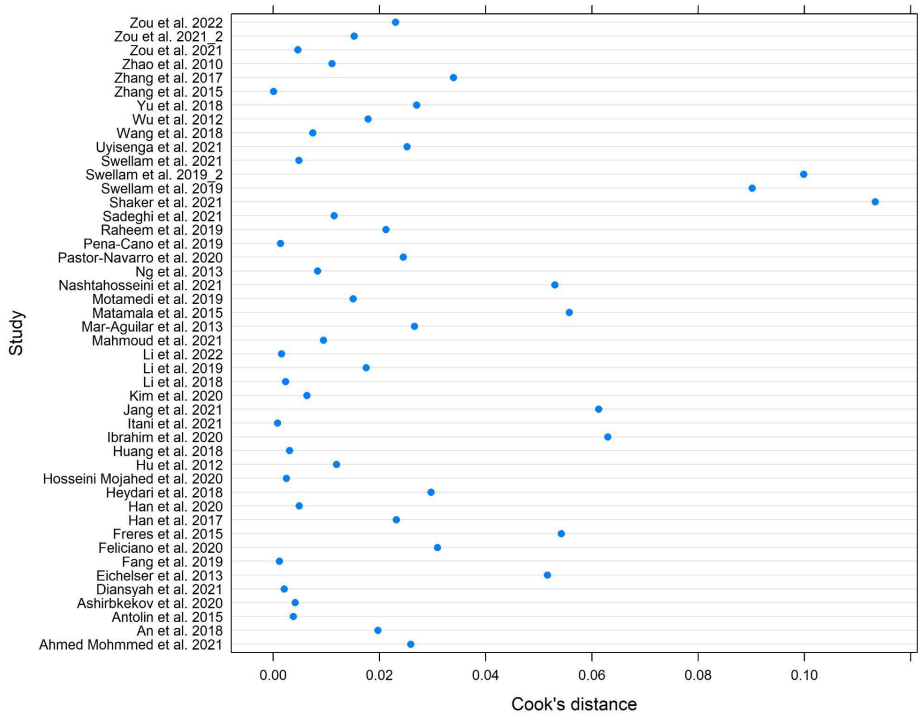


Figure 15. The calculated influence analysis (represented in Cook's distance units) of the included studies by taking into account all reported models.

Publication bias

Publication bias was evaluated for all the reported models. A funnel plot was generated on the log odds ratio and standard error (**Figure 16**). Egger's test, in which a random effect on the studies was added, was used to test for publication bias. A p-value of < 0.001 indicated a potential publication bias.

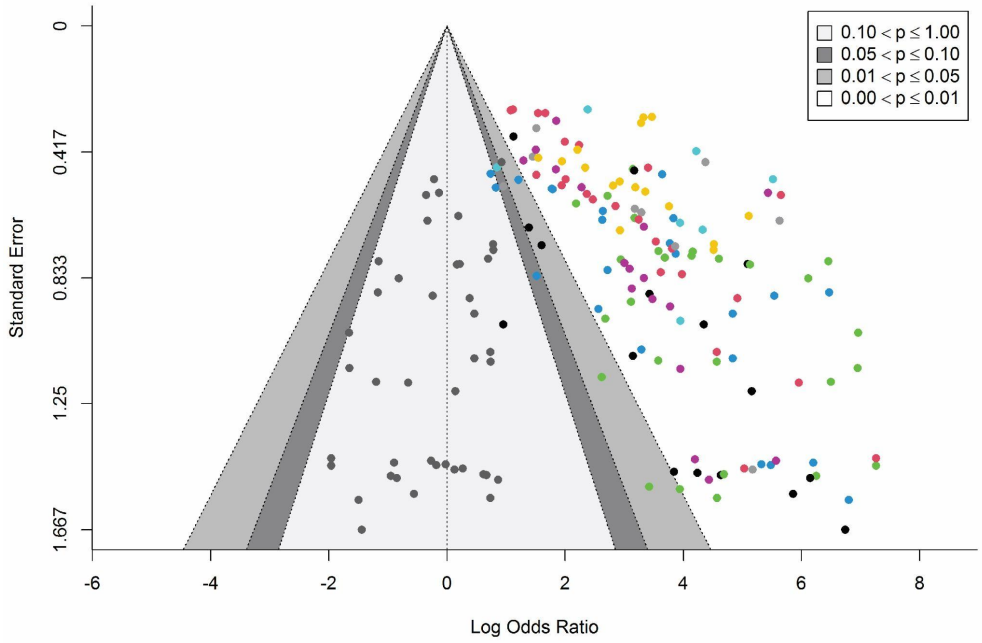


Figure 16. Publication bias was performed on all reported models. Points with the same colour in the graph represent models originating from the same study. The cluster of grey points on the left-hand side of the graph represents the missing models which would be required in order not to have a publication bias.

Subgroup bivariate analysis

In order to determine performance differences between methodological variations in the studies as well as to evaluate some potential candidate sources of between-study heterogeneity, subgroup analyses were performed. The main subgroups considered were single vs. multiple (panel) miRNAs, plasma vs. serum specimen type, studies including stage III and/or IV BC cases vs. studies not including stage III and/or IV BC cases, exogenous vs. endogenous normaliser and stratification of studies by QUADAS-2 performance. The subgroup analyses based on all reported models were performed utilising generalised linear multilevel models with random effects on the study and model. Pooled sensitivity and specificity on plasma models were 0.83 [0.77 to 0.87] and 0.85 [0.78 to 0.91], respectively, while for serum, the pooled sensitivity and specificity were 0.87 [0.81 to 0.91] and 0.83 [0.78 to 0.87], respectively (**Figure 17A**).

On average, models based on miRNA panels perform better than models based on a single miRNA. The former subgroup had a pooled sensitivity and specificity of 0.90 [0.86 to 0.93] and 0.86 [0.80 to 0.90], respectively, while the latter subgroup had a pooled sensitivity and specificity of 0.82 [0.77 to 0.86] and 0.83 [0.78 to 0.87], respectively (**Figure 17B**).

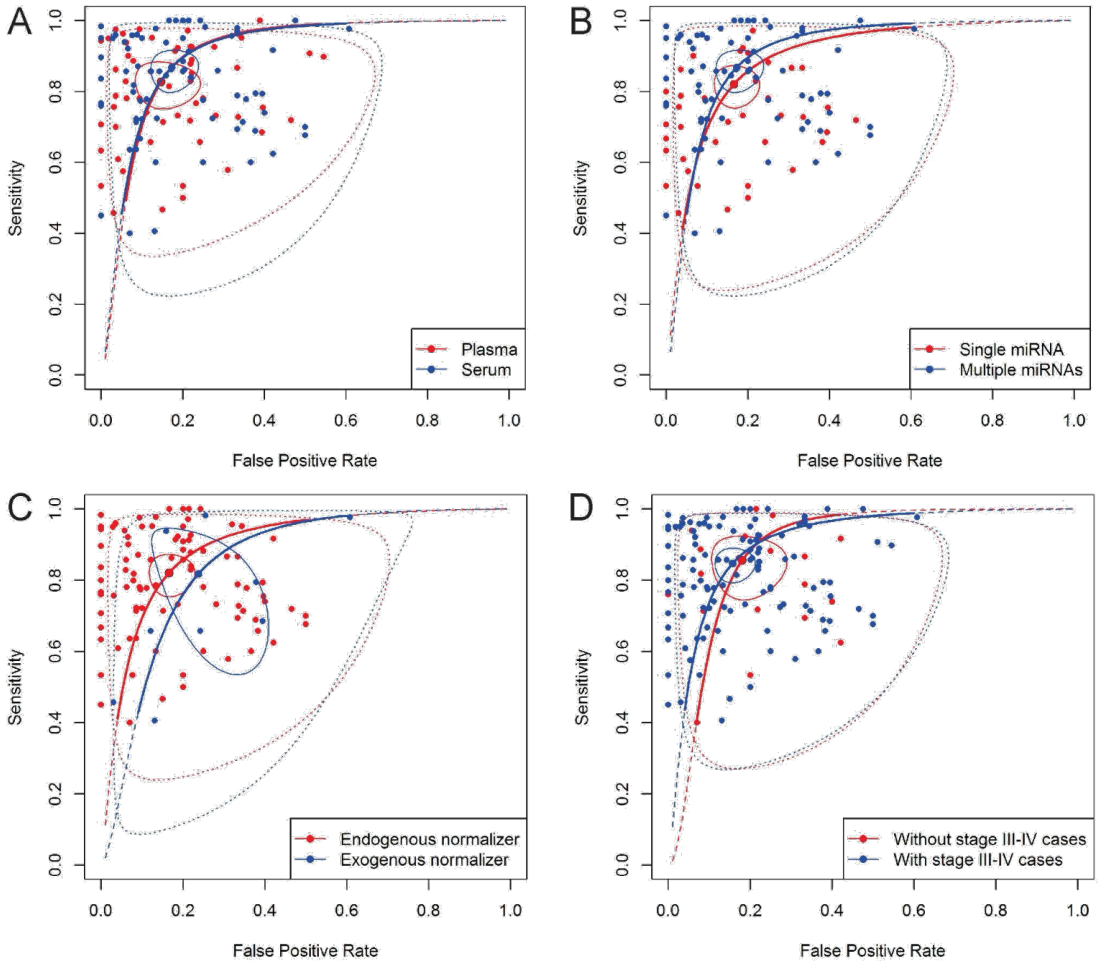


Figure 17. SROCs of the subgroup bivariate models based on all reported models. A) Plasma vs. serum B) single vs. multiple panel miRNAs C) endogenous vs. exogenous normaliser D) with vs. without stage III and stage IV cases.

Considering the sample size disparity between models that used exogenous and endogenous normalisers, the performance between the two groups is quite similar, with the models based on endogenous normalisers having a higher specificity (**Figure 17C**). For models with an exogenous normaliser, the pooled sensitivity and specificity were 0.82 [0.60 to 0.93] and 0.76 [0.63 to 0.86], respectively, while the pooled sensitivity and specificity for models with an endogenous normaliser were 0.82 [0.77 to 0.86] and 0.83 [0.78 to 0.87], respectively. Expectedly, models without stage IV BC samples and models with < 4.5% stage IV BC samples performed similarly when the pooled sensitivities and specificities were compared. The models without stage IV cases had a pooled sensitivity of 0.85 [0.81 to 0.88] and specificity of 0.84 [0.80 to 0.88], while models

with stage IV cases had a slightly better pooled estimate where the sensitivity was 0.87 [0.61 to 0.97] and specificity was 0.86 [0.80 to 0.90]. This slight difference could be attributed to the difference in model numbers analysed in the two groups, as seen from the confidence interval for the sensitivity estimate for models with stage IV cases. Thus, since low between-study heterogeneity was observed in this subgroup analysis, the total cohort of models, which includes both with (< 4.5%) and without stage IV BC samples, can be considered reliable for assessing the general ability of circulating miRNAs to diagnose BC, with the condition that the models assessed do not have a higher percentage of stage IV cases than would be observed in community screening for BC.

To further investigate the impact of stages on diagnostic performance, a subgroup analysis of the models with and without stage III and IV samples was performed. Pooled sensitivity and specificity of 0.84 [0.80 to 0.88] and 0.85 [0.80 to 0.88], respectively, were obtained for the former group, while 0.86 [0.77 to 0.91] and 0.82 [0.74 to 0.88], respectively, for the latter (**Figure 17D**). As observed in the previous subgroup analyses, models that include later BC stages (III and IV) have slightly better diagnostic performance than models that include only earlier stages (0, I and II). The SROCs of the same subgroup analyses were performed on the most important model of each study, as can be seen in **Figure 18**.

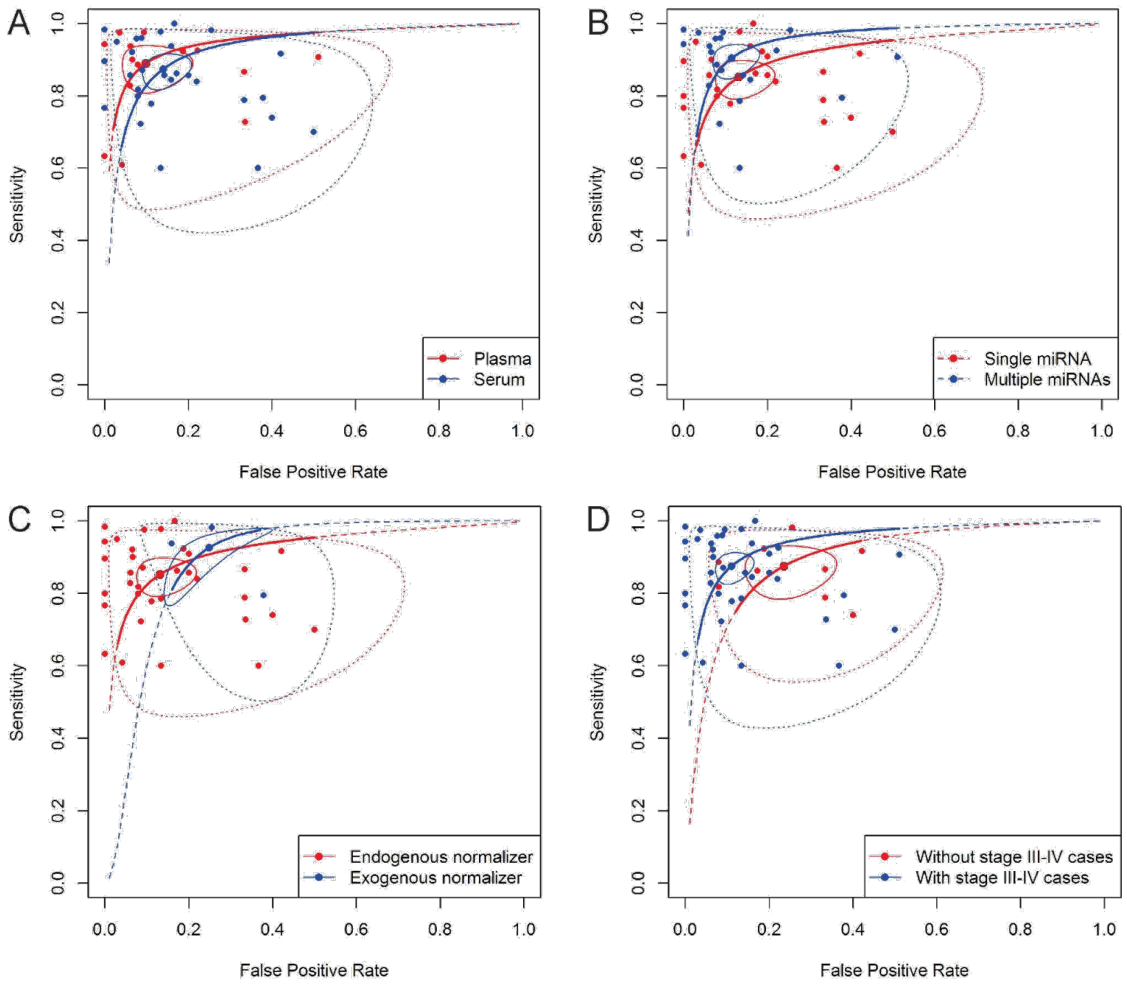


Figure 18. SROCs of the subgroup bivariate models based on the most important model of each study. A) Plasma vs. serum B) single vs. multiple panel miRNAs C) endogenous vs. exogenous normaliser D) with vs. without stage III and stage IV cases.

The results are concordant with the subgroup analyses on all reported models, with slightly more pronounced differences between endogenous and exogenous normalisers and between with and without stage III/IV cases. Interestingly, when studies were stratified based on the QUADAS-2 performance cut-points (no cut-point, > 3, > 4 and > 5 "low" on the seven key questions), increasing QUADAS-2 score corresponded to decreasing pooled diagnostic performance, chiefly reflected in specificity. This was observed on all reported models as well as on the most important model per study. Details on the results of subgroup analysis on all reported models and on the most important model per study can be found in *Supplementary Table 2* and *Supplementary Table 3*, respectively (Appendix A). Lastly, we estimated the pooled sensitivity and specificity on all reported models for each year to assess if there was a diagnostic performance trend throughout the years. A linear regression was performed on pooled sensitivities and specificities, and no significant linear association was found (**Figure 19**).

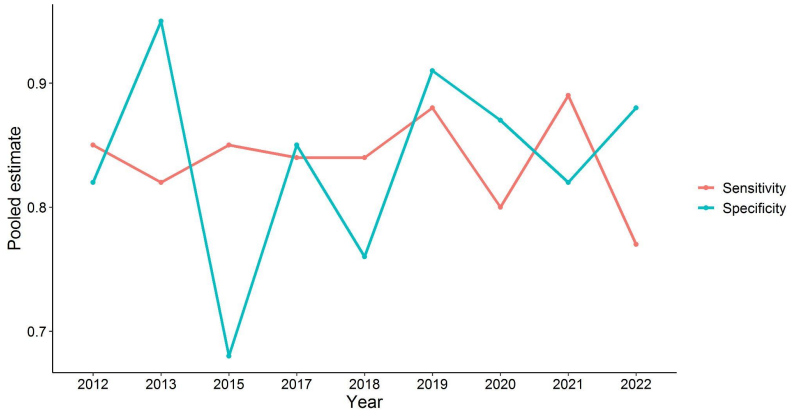


Figure 19. Pooled estimates of sensitivity and specificity were calculated on all models of studies stratified by year of publication. Linear regression was performed on both sensitivity and specificity across the years, and no significant linear trend was observed. For both sensitivity and specificity, the linear regression estimates were around 0.

miRNA-21-5p

miRNA-21-5p is the most analysed miRNA among the included studies in this meta-analysis and is a miRNA that was often reported as dysregulated in the breast but also in many other cancers. Therefore, we performed a bivariate analysis using the generalised linear multilevel model to meta-analyse the diagnostic ability of circulating cell-free miRNA-21-5p in BC. The pooled sensitivity and specificity for models evaluating only miRNA-21-5p were 0.74 [0.64 to 0.83] and 0.81 [0.70 to 0.89], respectively. The SROCs are shown in **Figure 20**, while the details on the model are found in *Supplementary Table 2* and *Supplementary Table 3* (Appendix A).

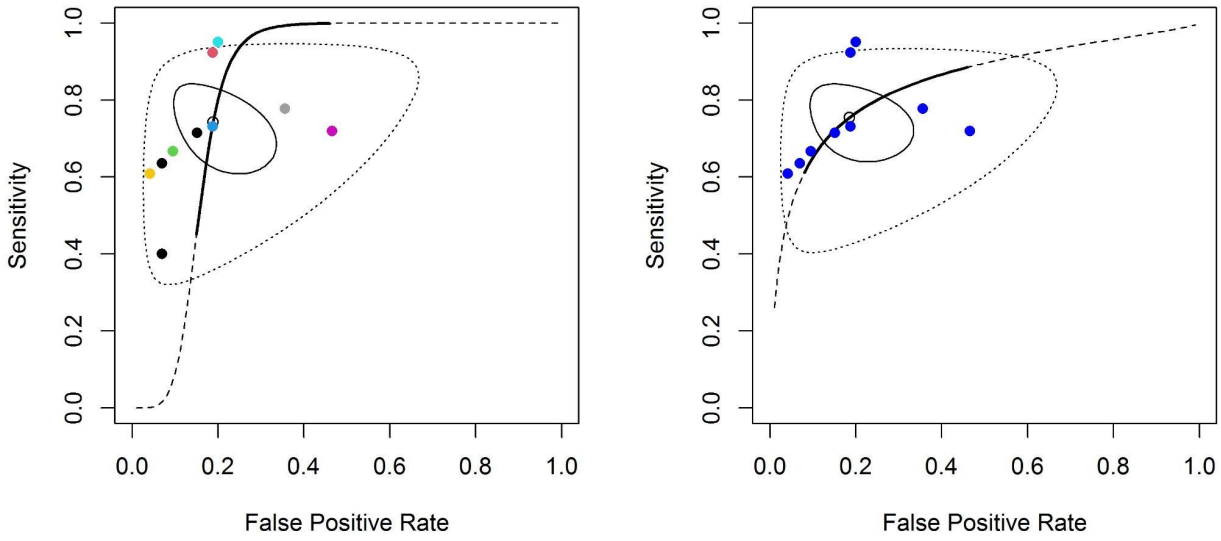


Figure 20. SROCs on miRNA-21-5p bivariate models. A) miRNA-21-5p SROC of all reported models. Points with the same colour in the graph represent models originating from the same study. B) miRNA-21-5p SROC of the most important model from each study.

Univariate analysis on log-DOR

In order to include studies that did not report diagnostic accuracy in terms of sensitivity and specificity we performed a univariate analysis on log-DOR using the q-Point data from the reported ROC graphs. The q-Point was extracted for all models with an ROC curve. A pooled log-DOR based on all reported models of 2.48 [2.15 to 2.81] resulted. Significant heterogeneity was observed in the model (Cochran's $Q = 978.9$, $p < 0.001$). For the most important models, a pooled log-DOR of 2.99 [2.56 to 3.41] was observed with a significant heterogeneity (Cochran's $Q = 402.6$, $p < 0.001$).

As there was a large difference in the number of models that used endogenous and exogenous normalisers, we complemented the bivariate subgroup analysis on endogenous versus exogenous models with the log-DOR univariate analysis, where the difference in model numbers was smaller. The estimate of pooled log-DOR for endogenous models was 2.58 [2.22 to 2.94], while for the exogenous models it was 1.45 [0.86 to 2.04], confirming the discrepancy in diagnostic accuracy found with bivariate models. The log-DOR estimate details of all reported models and most important models per study, as well as all their subgroups, are found in **Table 7** and **Table 8**, respectively.

Table 7. Summary of the univariate (log-DOR) analysis on all the reported models and its corresponding subgroup analysis.

Subgroup	Pooled log-DOR	Cochran's Q (p-value)
All models	2.48 [2.15, 2.81]	978.91 (< 0.001)
Plasma	2.48 [1.82, 3.14]	412.48 (< 0.001)
Serum	2.64 [2.22, 3.06]	496.29 (< 0.001)
Single miRNA panel	2.16 [1.80, 2.53]	669.47 (< 0.001)
Multiple miRNA panel	3.20 [2.76, 3.64]	126.68 (< 0.001)
Endogenous normaliser	2.58 [2.22, 2.94]	501.87 (< 0.001)
Exogenous normaliser	1.45 [0.86, 2.04]	115.10 (< 0.001)
With stage III & IV cases	2.52 [2.16, 2.88]	866.86 (< 0.001)
Without stage III & IV cases	2.33 [1.69, 2.97]	111.92 (< 0.001)
With stage IV cases	2.22 [1.39, 3.06]	210.43 (< 0.001)
Without stage IV cases	2.54 [2.19, 2.89]	767.49 (< 0.001)

Table 8. Summary of the univariate analysis (log-DOR) on the most important model of each study and its corresponding subgroup analysis.

Subgroup	Pooled log-DOR	Cochran's Q (p-value)
Most important models	2.99 [2.56, 3.41]	402.58 (< 0.001)
Plasma	2.82 [1.98, 3.67]	153.72 (< 0.001)
Serum	3.13 [2.61, 3.66]	208.05 (< 0.001)
Single miRNA panel	2.64 [2.07, 3.21]	248.10 (< 0.001)
Multiple miRNA panel	3.51 [2.96, 4.06]	109.27 (< 0.001)
Endogenous normaliser	3.08 [2.58, 3.58]	229.96 (< 0.001)
Exogenous normaliser	1.86 [0.89, 2.84]	70.38 (< 0.001)
With stage III cases	3.18 [2.68, 3.67]	327.56 (< 0.001)
Without stage III cases	2.36 [1.61, 3.11]	60.71 (< 0.001)
With stage IV cases	2.87 [1.79, 3.95]	50.92 (< 0.001)
Without stage IV cases	3.01 [2.54, 3.47]	347.02 (< 0.001)

Preference for sensitivity or specificity

To investigate whether a preference of a model for sensitivity or specificity is related to an imbalance of proportions between cases and controls or to predicted positive (TP + FP) and predicted negative (TN + FN) samples, a graphical technique was employed: models were divided into three groups according to the proportion of cases to controls or of predicted positive to predicted negative samples, coloured and plotted on an ROC plane (**Figure 21**). Differences in model designs based on the proportion of cases to controls are mainly reflected in the FPR (**Figure 21A**), as models with fewer cases than controls tend to have a larger FPR.

Overall, models with a balanced case–control design or a design with more cases than controls are far more abundant than models with fewer cases than controls. A clearer performance trend can be seen when the proportion of the positive screens and negative screens is taken into account (**Figure 21B**). Models with fewer positive screens than negative usually tend to have a smaller FPR and sensitivity. Conversely, models with more positive screens than negative have the tendency for a larger FPR and sensitivity. Those models with balanced positive and negative screens have more balanced FPR and sensitivity when compared to the previous two groups.

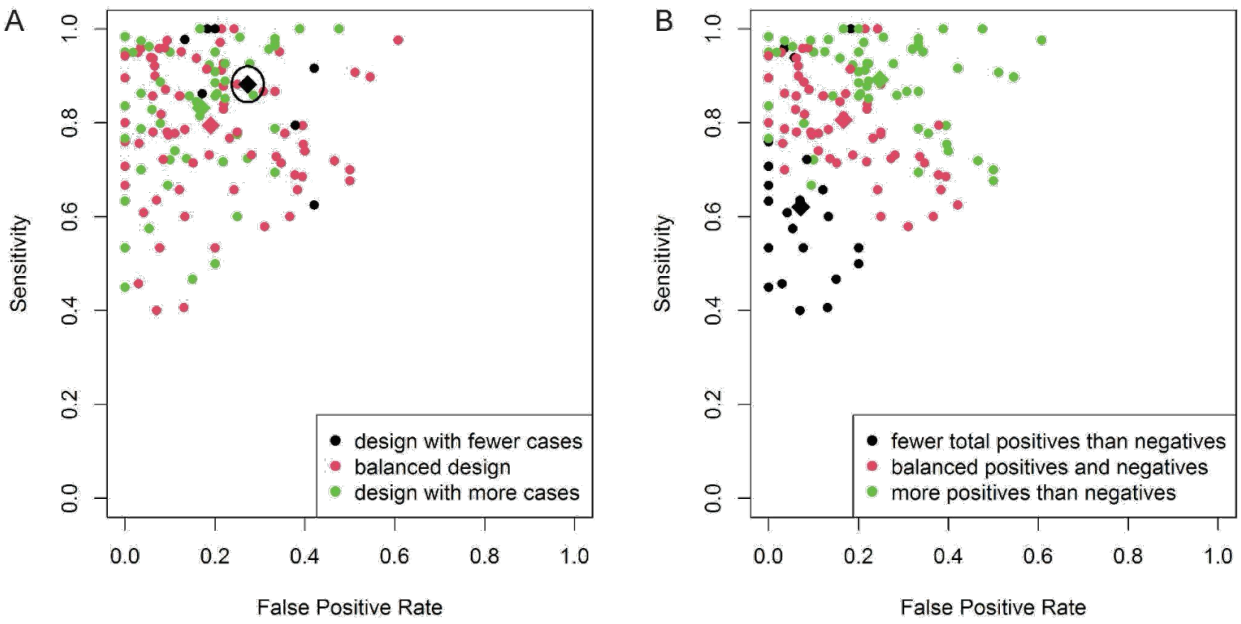


Figure 21. Comparison of diagnostic performance of models to their imbalance of proportions of A) cases to controls or B) predicted positive to predicted negative screens, represented by a colour corresponding to one of the three imbalance of proportions cut-point groups. Diagnostic performance means (with confidence intervals) of the three ratio groups are represented by diamonds. Mean points without confidence intervals indicate very narrow ranges.

An alternative plot was also created by dividing models into five groups instead of only three (**Figure 22**) and the same conclusions can be drawn as in the previous figure. In sum, sample composition, i.e., the ratio of cases to controls, seems to influence diagnostic accuracies, probably via study-level model tuning. Moreover, the predicted positive and predicted negative ratio is most likely influenced by the compromise or preference between sensitivity and specificity.

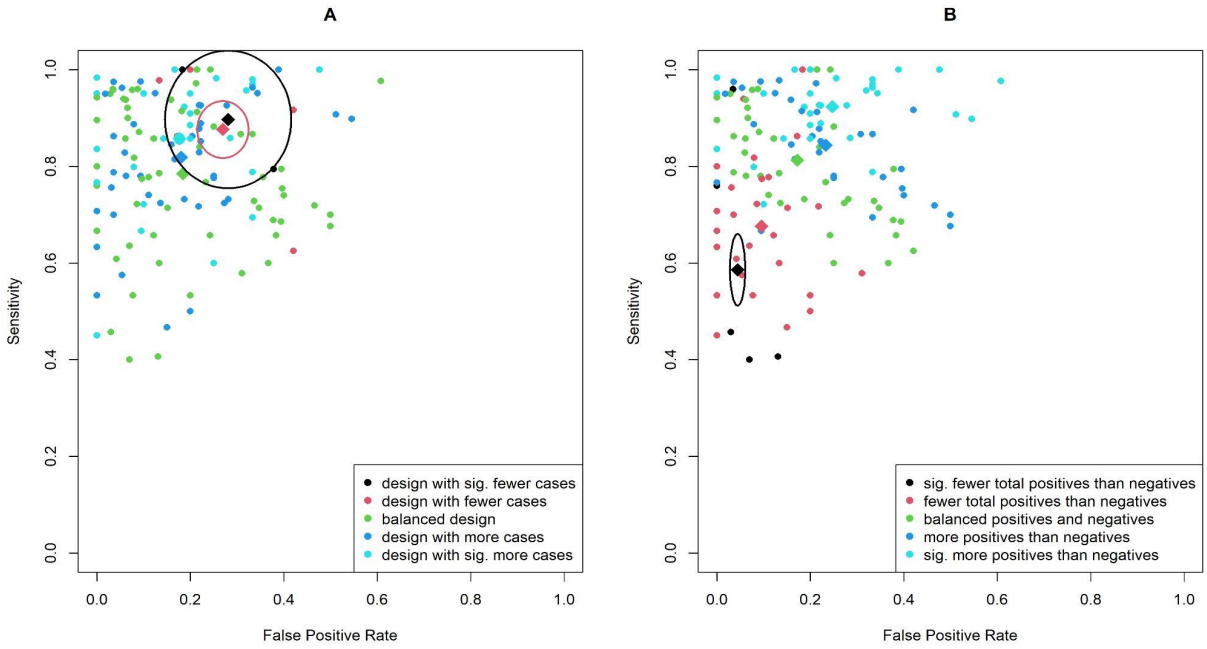


Figure 22. Comparison of diagnostic performance of models to their imbalance of proportions of A) cases to controls or B) predicted positive to predicted negative screens, represented by a colour corresponding to one of the five imbalance of proportions cut-point groups. Diagnostic performance means (with confidence intervals) of the five ratio groups are represented by diamonds. Mean points without confidence intervals indicate very narrow ranges.

Quantifying the author or model preference for sensitivity or specificity

By utilising the α parameter from the ROC shape, we assessed whether the meta-analysed models preferred sensitivity or specificity (**Figure 23A**). A general trend of preference can be seen in the plot with all reported models. However, since the trend is not strong enough, only the models with an α z-score > 0.8 SDs away from the mean were considered as studies with some kind of preference. Based on the cut-off value, 25 of the 117 analysed models had a preference for sensitivity, while 24 had a preference for specificity. The preference is derived from ROC curve shape, so a preference in shape does not necessarily imply that the pair of sensitivity and specificity at the authors' preferred cut-off value reflects this preference: 22 out of the 25 models considered to prefer sensitivity had a higher sensitivity than specificity, while 18 out of the 24 models considered to prefer specificity had a higher specificity.

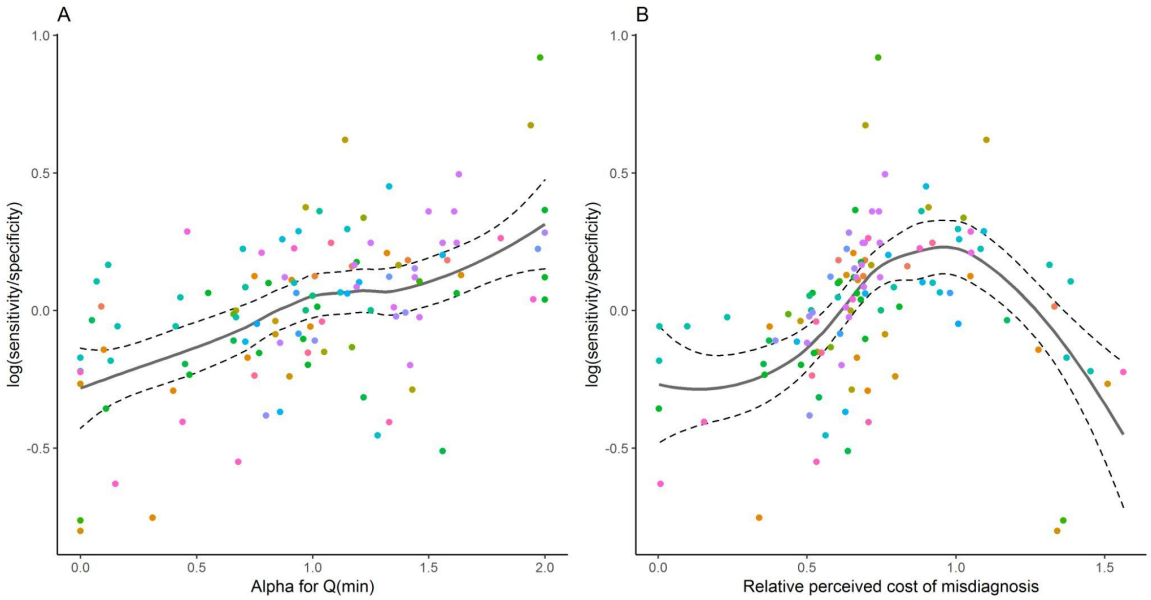


Figure 23. Preference estimates based on log (sensitivity/specificity) for all reported models using A) alpha for minimum Q and B) relative perceived cost of misdiagnosis (c_1). Points with the same colour in the graph represent models originating from the same study.

Based on the plot on the most important models, the α for Q(min) is not able to catch a direct trend of preference (**Figure 24**). We do observe, however, that the α for Q(min) is not evenly distributed and that, based on the ROC shape, there tends to be an overall preference for sensitivity, which does not necessarily have to be reflected in the outcome values due to different factors (e.g., biology of the predictor, measurement tools, statistical modelling, population, etc.).

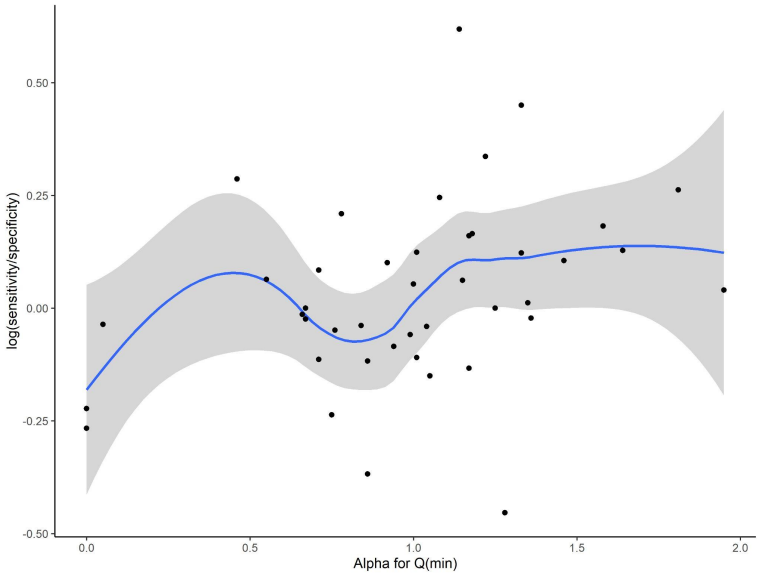


Figure 24. Preference estimates based on log (sensitivity/specificity) for all reported models using alpha for minimum Q in the most important models for each study.

In addition to the assessment of preference of the model by the α parameter, we assumed that in all the models, the study authors base their decision about the cut-off value on a perceived cost c_1 for not detecting a BC patient and a cost c_0 for a positive screen on a healthy person. Recall that the perceived cost c_1 is calculated in units of $c_0 = 1$ (**Figure 23B**) and note that the prevalence factor was omitted. The strength of the preference trend is similar to that of the previous plot. Hence, models with a c_1 z-score of > 0.8 SDs away from the mean were considered as studies with some kind of author preference.

Based on the c_1 value, 10 of the 117 analysed models had a preference for sensitivity, while 80 had a preference for specificity. Of the 80 models considered to prefer specificity, 41 had a higher specificity than sensitivity. Interestingly, most of the models with a high c_1 value (> 0.8 SD) did not have a higher sensitivity than specificity, a consequence of the underlying ROC curve shapes. In this sense, most of the ten models did not have a preference for sensitivity in the naive sense. Until the c_1 starts surpassing the value of 1, the plot seems to be linear and in concordance with the plot in **Figure 23A**. Importantly, for the models that did not report the sensitivity and specificity measures but reported an ROC curve, we chose the cut-offs and obtained a sensitivity and specificity pair using the q-Points, which might have affected the robustness of the mentioned preference quantification methods.

When investigating the c_1 plot on the most important models (**Figure 25**), as in the α (at Q min) plot, we cannot observe a clear trend of preference, but we can see that the distribution of c_1 tends to be centred below 1, which could indicate that authors tend to give more importance in reducing false negatives. Nevertheless, due to the non-linear trend of preference using the c_1 statistic, the α parameter preference method has shown more robust results, while the c_1 metric could be more successful in adequately designed models due to its better ability to catch preference in individual models. By “adequately designed models”, we refer to diagnostic models with a large enough sample size as well as the reported reasoning why the chosen sample size was selected and a clear predictor selection strategy. Having said that, it is worth noting that between the two preference assessment methods, there were 12 common models (from all reported models) which preferred specificity and 23 common models which did not have a significant preference. No common models were found for sensitivity preference.

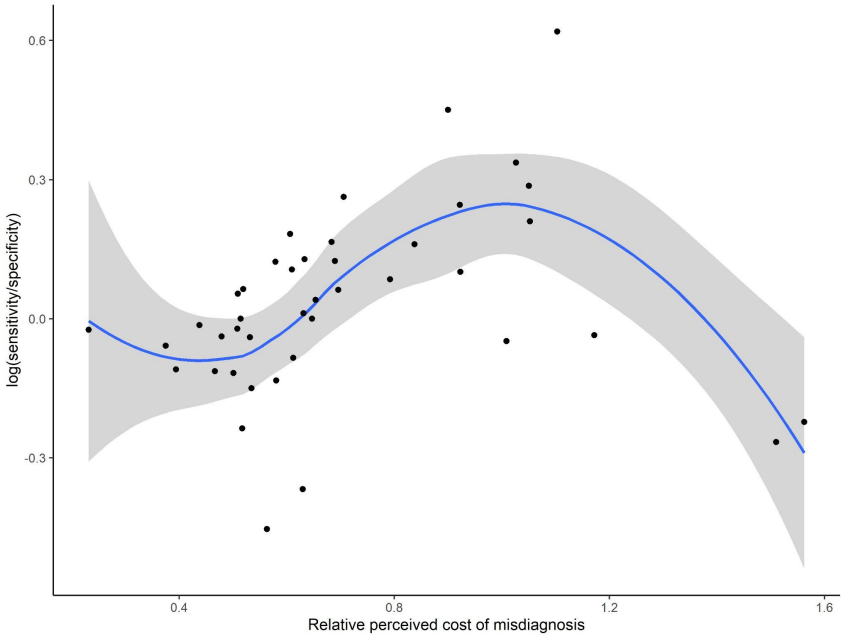


Figure 25. Preference estimates based on log (sensitivity/specificity) for all reported models using relative perceived cost of misdiagnosis (c_1) in most important models for each study.

Circulating biomarkers for early BC detection

Population characteristics

The discovery cohort on which we performed targeted SNP and methylation analysis or miRNA profiling included 70 cases and 70 controls. All samples had successful DNA extraction from the buffy coat, while after RNA extraction from plasma of 70 cases and 70 controls and library preparation, nine samples were excluded due to poor quality. Thus, the final discovery cohort for cfc miRNA analysis consisted of 65 cases and 66 controls. The general characteristics of the study population, including the samples available for all analyses (i.e., miRNAs, SNPs and methylation), are reported in *Supplementary Table 4* (Appendix B). The only variables that showed a significant association with BC detection in this cohort were: BMI, breast density (Tabar's scale) and WCRF lifestyle score. The characteristics of cases are reported in *Supplementary Table 5* (Appendix B), separately for invasive and in situ tumours. Cases were diagnosed on average 3 ± 2 months after blood collection. Fifty-five women were diagnosed with invasive breast tumours and eight with in situ lesions. The most frequent histotype was ductal (56.0% of invasive and 37.5% of in situ BCs), and the majority of cancers were stage IA (87.5%), ER positive (84.9%), PgR positive (69.9%), Her2 negative (86.5%) and Ki-67 negative (76.5%).

The cohort on which we validated the biomarkers selected in the discovery cohort included 32 cases (all from the Biella hospital) and 127 controls. All of the samples had good RNA quality. The general characteristics of the validation sample can be seen in *Supplementary Table 6* (Appendix B). The variables which were found to be associated with BC in the validation cohort were the presence of previous benign biopsies (OR: 3.28, P: 0.04), breast density based on Tabar scale (Tabar 3 vs. reference – OR: 16.67, P: 0.00004) and breastfeeding status (OR: 0.23, P: 0.01).

The tumour characteristics of the 32 samples can be seen in *Supplementary Table 7* (Appendix B). Unlike the discovery cases, the cases in the validation cohort were diagnosed on average 2.1 ± 1.3 years after blood collection. This implies that, in a way, we were also testing the predictive ability of the selected biomarkers. Most of the tumour samples were invasive, with only one in situ sample. The most frequent histotype was ductal (67.7% of invasive tumours). Like in the discovery cohort cases, the majority of invasive tumours were stage IA (45.1%), ER positive (74.2%), PgR positive (67.6%) and Her2 negative (74.2%). The key difference between the cases in the discovery and validation cohort was in the Ki-67 status where the majority of validation cohort cases were Ki-67 positive (74.2%).

Polygenic risk score

We examined the PRS score on the 131 samples from the discovery cohort. The PRS score was normally distributed according to the Shapiro–Wilk normality test (p-value = 0.289). The PRS average across all 131 samples was 0.98, with a standard deviation of 0.41. The PRS did not differ

significantly between cases and controls based on the mean (two-sample t-test p-value = 0.784), variance (F-test p-value = 0.923) or distribution (two-sample Kolmogorov–Smirnov test: p-value = 0.464). The density plot of the PRS across all samples, as well as stratified by BC status, can be seen in **Figure 26**.

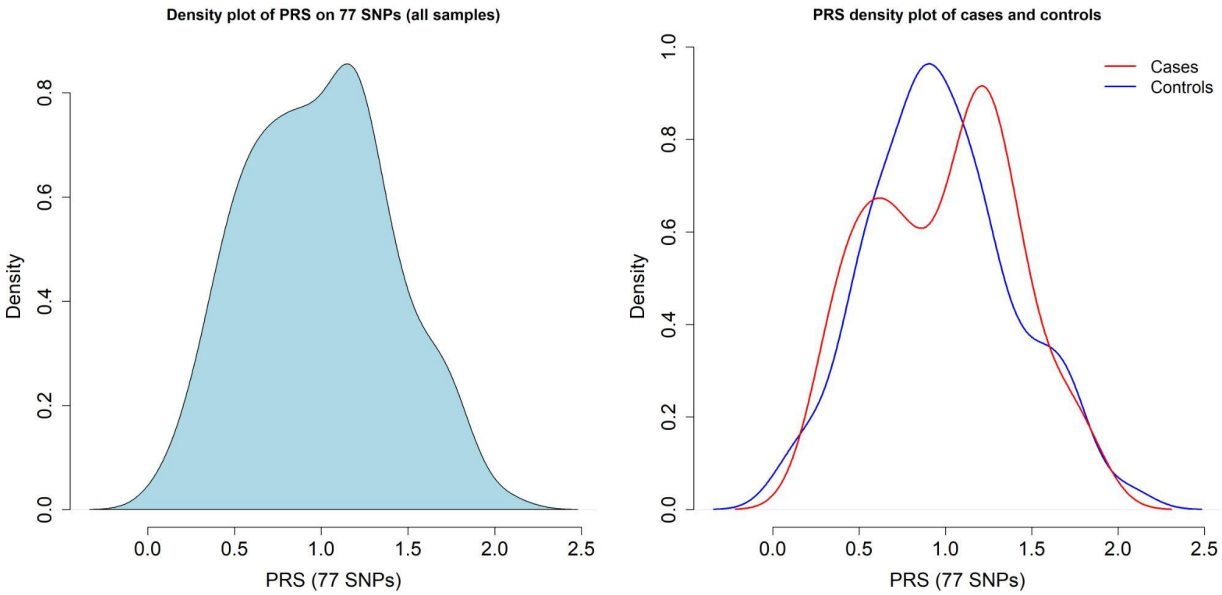


Figure 26. Density plots of PRS scores in all samples and stratified by cases and controls.

We also performed a logistic regression with the status as dependent and PRS as independent variable. PRS on the 77 SNPs is not associated with BC in our cohort **Table 9**. After computing the predictions based on the PRS, a poor AUC of 0.52 was obtained (**Figure 27**). Considering that the PRS calculated in this project could not differentiate between cases and controls, we did not include it as a predictor in the final model consisting of miRNA and other variables.

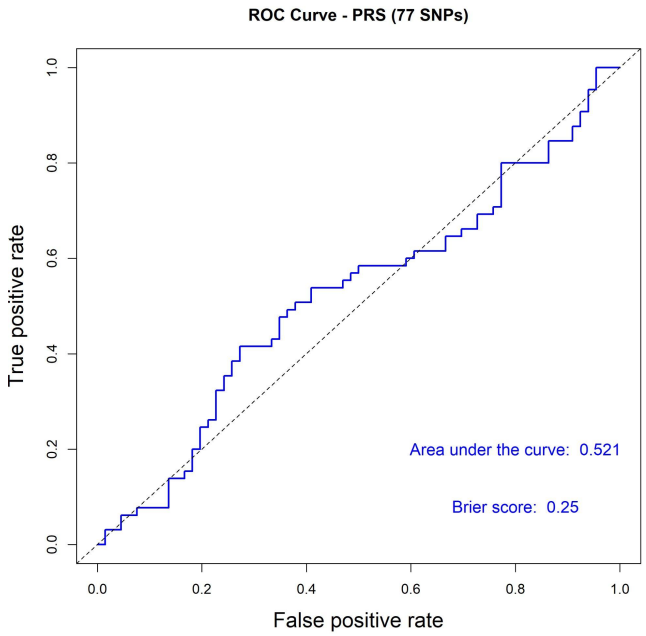


Figure 27. The ROC curve of the PRS score used to discriminate between BC cases and controls.

Table 9. Logistic regression on PRS based on 77 SNPs to discriminate between cases and controls.

	Estimate	**SE**	**Z**	**p-value**
Intercept	-0.13	0.45	-0.29	0.773
PRS (77 SNPs)	0.12	0.42	0.28	0.782

Methylation of promoter regions

The methylation of promoter regions of the *RARB*, *APC* and *BRCA1* genes was measured using the MS-HRM method. Across two plates for each gene there were 140 (70 cases and 70 controls), 100 (54 cases and 46 controls) and 135 (67 cases and 68 controls) successfully evaluated samples for *RARB*, *APC* and *BRCA*, respectively. The smaller number of samples for *APC* genes was due to a mistake in the MS-HRM instrument setup. The MS-HRM results are affected by the methylation of CpG sites within the region and are measured in a collective/additive manner. The derivative curves and their melting points at 0% and 100% methylation standards can be seen for all genes on both plates in **Figure 28**.

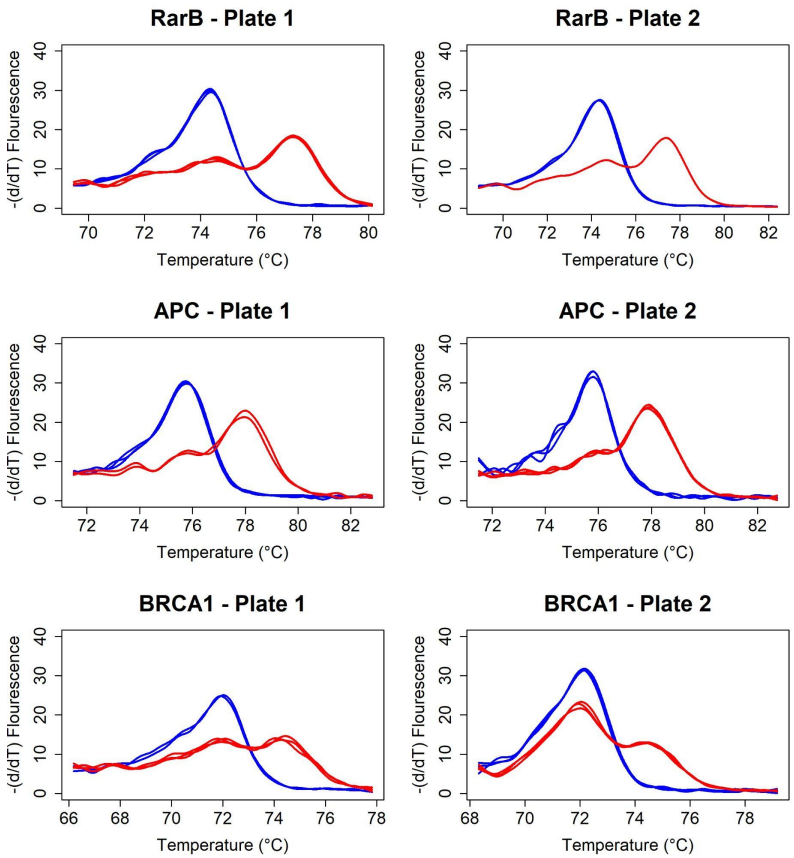


Figure 28. Derivative curves of methylation 0% and methylation 100% standards for the three gene promoters at each of the two plates (except for BRCA1 plate 2, which had a methylation 75% standard). The peaks of the curves represent the melting points.

Differing melting points between the methylation 0% and methylation 100% standards, indicated by curve peaks at different temperatures, can be seen in all genes and for all plates except for plate 2 of the *BRCA1* gene, which had a technical issue with the methylation standards. Additionally, for some plates, the standard replicates were not successful, and therefore, in the derivative plots for some of them there were less than three curves. Nevertheless, for all plates except the *BRCA1* plate 2, a clear difference in melting points between the standards was observed.

A difference plot was created for each plate where we subtracted the relative fluorescence at each temperature point at methylation 0% from the other methylation standards. A clear separation between the standards can be seen in all genes and all plates (**Figure 29**).

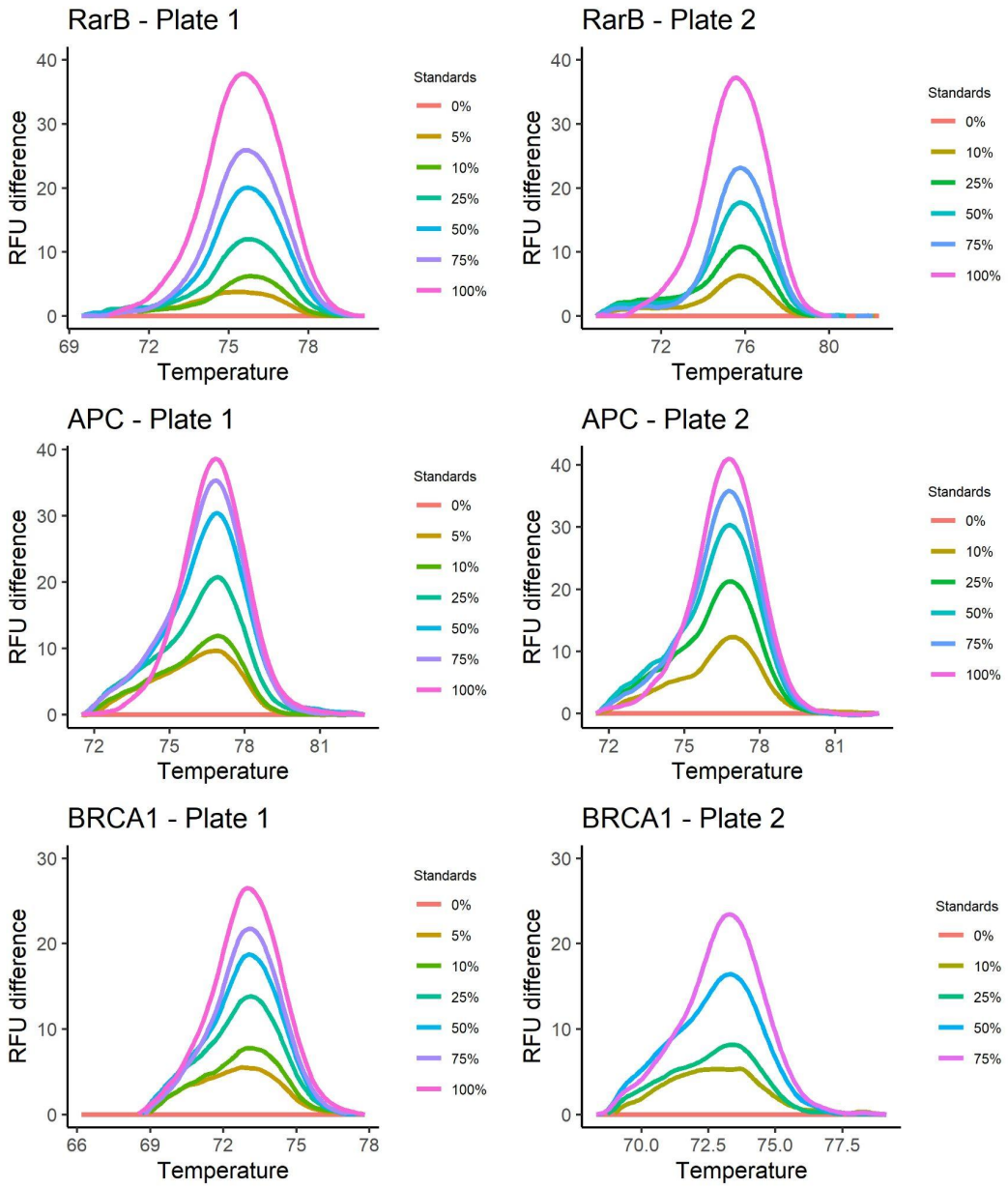


Figure 29. Difference plots where the relative fluorescence at each time point from methylation 0% standard was subtracted from the other methylation standards.

We then investigated at which temperature points the highest variability of RFU occurs, which in this case indicated the highly informative points. We observed that the standards tend to vary the most in the temperature range from 75°C to 79°C (**Figure 30**). On the other hand, the samples had low variability across the whole temperature range, indicating a small degree of methylation difference in the three genes between the samples (data not shown).

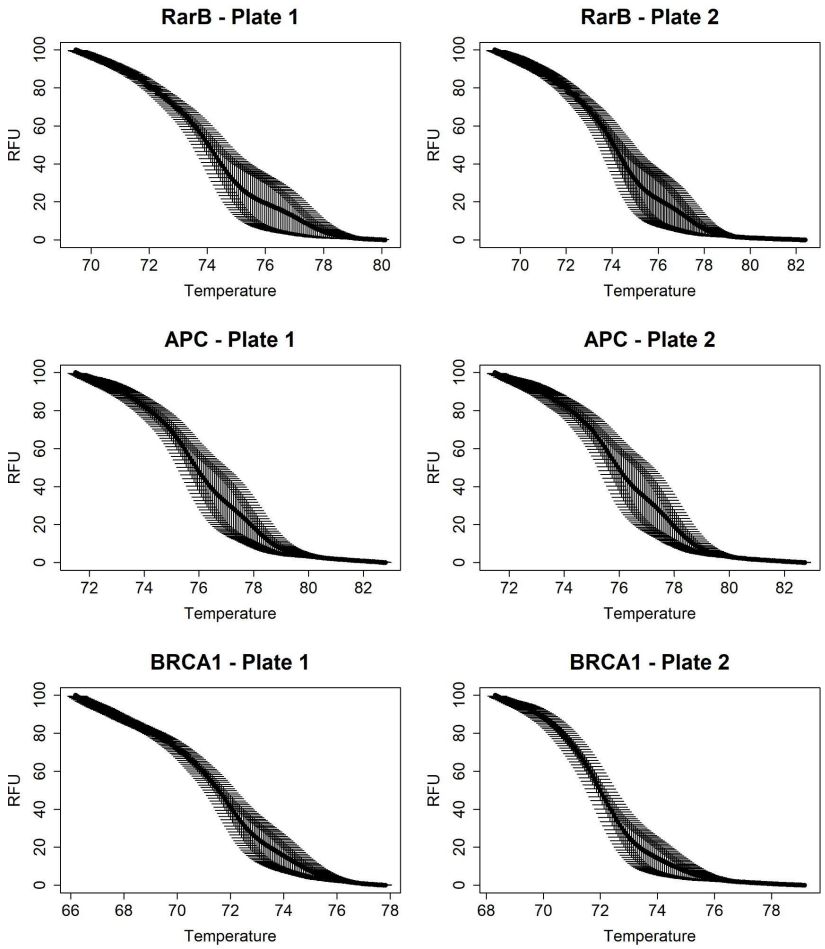


Figure 30. Variability of RFU at different temperature units. Results for all three genes are shown.

To infer the methylation values of each sample on the three genes, an interpolation curve was performed on each plate (n = 1000) and based on where the samples would fall in the curve, they would be assigned a methylation value. The interpolation curves for all three genes and their plates are shown in **Figure 31**.

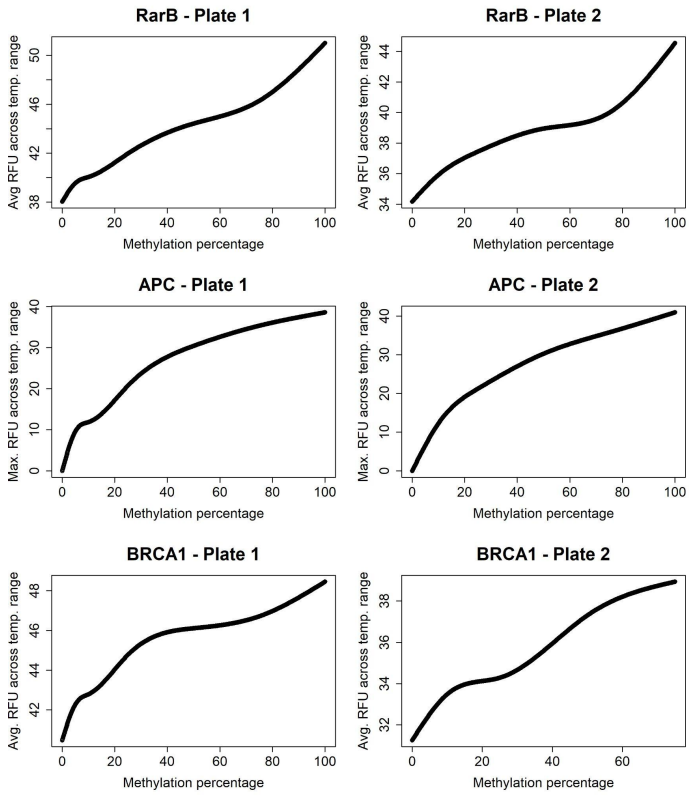


Figure 31. Interpolation curves based on the methylation standards for the three gene promoters.

The optimal interpolation curve would have a slope of 1, and as can be seen in **Figure 31**, the interpolation was suboptimal for some of the plates. The histograms of methylation values of the samples for the three genes and their two plates can be seen in **Figure 32**.

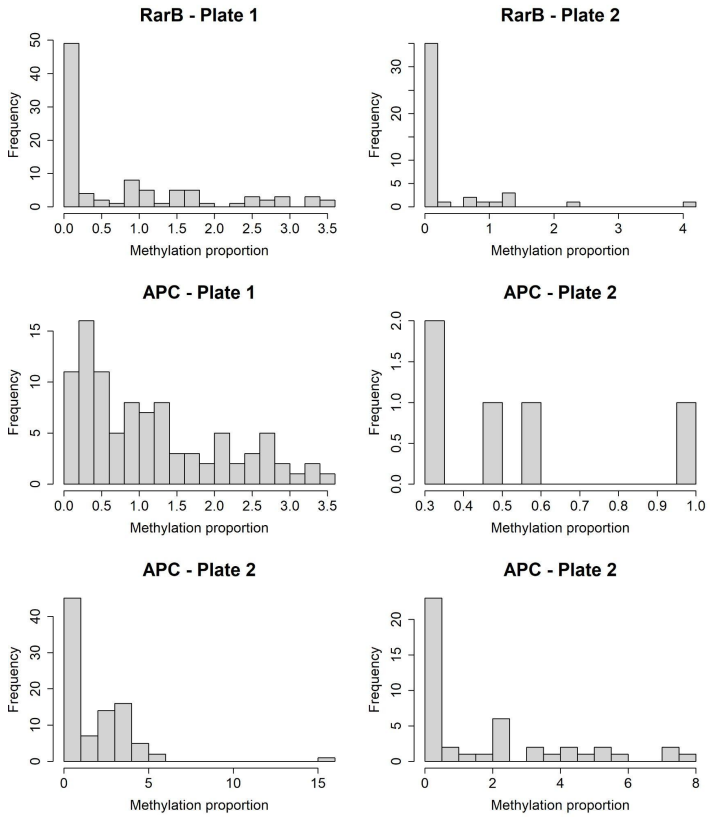


Figure 32. Histogram plots of the methylation estimates for the three gene promoters.

The overall methylation of the gene promoters in the three genes is very low and contains a large proportion of samples with an estimated methylation of 0%. A slightly higher proportion of non-zero methylation values can be seen for the samples analysed for the *APC* gene. This is because for this gene we had to use the maximum RFU difference between the methylation 0% standard and the other standards or samples. The methylation values are still very close to zero. The summary statistics of the methylation values for each gene on both plates can be seen in **Table 10**.

Table 10. Summary statistics of the methylation estimates for the three gene promoters.

Gene	Min.	Median	Mean	Max.	SD
RARB – Pt.1	0	0.1	0.82	3.6	1.08
RARB – Pt.2	0	0	0.32	4.1	0.77
APC1 – Pt.1	0.1	1	1.19	3.5	0.92
APC1 – Pt.2	0.3	0.5	0.54	1	0.29
BRCA1 – Pt.1	0	1.2	1.81	15.7	2.17

After combining the methylation values from the two plates, we assessed the distribution of methylation estimates, and none of the three genes had a normal distribution. Thereafter, we performed a class comparison, using the Mann–Whitney U test, between tumour and healthy

control samples. None of the genes had a significantly different methylation profile in cases compared to controls (**Figure 33**).

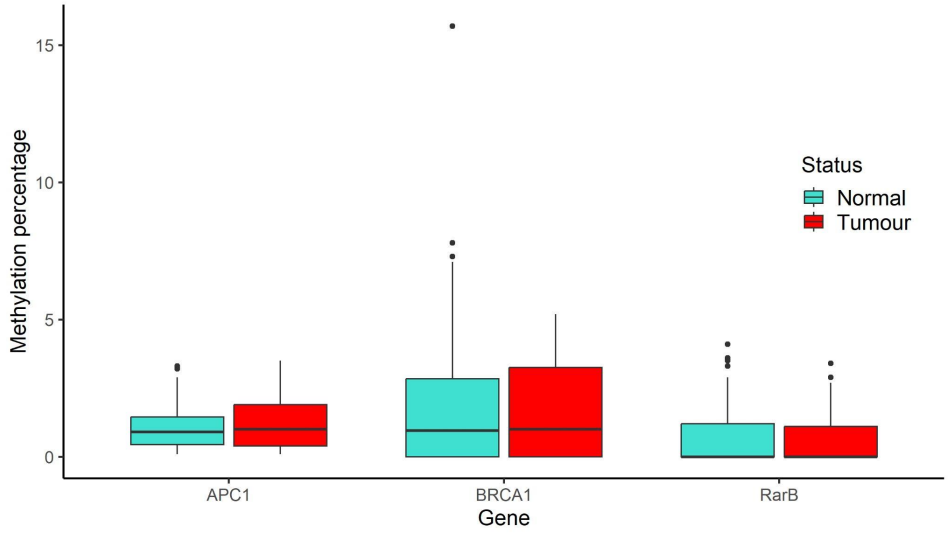


Figure 33. Boxplots of the methylation estimates of the promoters of the three genes stratified by BC status.

To correct for the samples' plate of origin, we complemented the class comparison with a logistic regression analysis where the class was the dependent variable, while estimated methylation and plate were the independent variables. A bootstrap (n = 2000) was performed on the Beta coefficients of the mentioned logistic regressions, and no evidence was found that the methylation of the three genes was different between cases and controls (**Figure 34**).

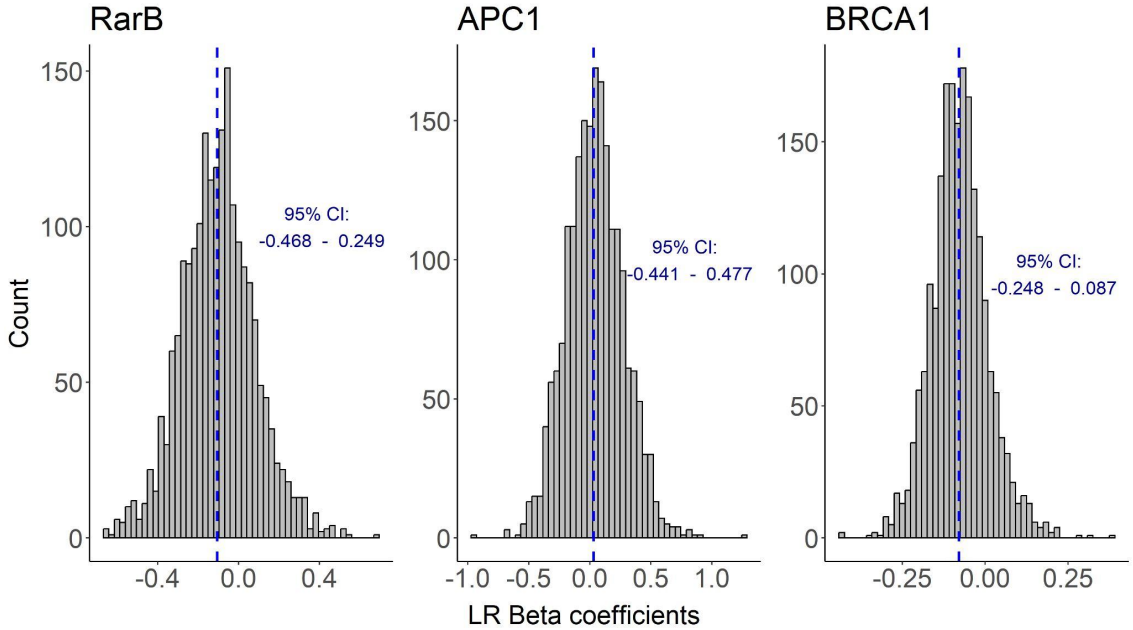


Figure 34. Bootstrap frequencies of the logistic regression coefficients of association with BC status for the three gene promoters.

Additionally, we also calculated the ROC AUC of the genes by training the model on 70% of the data and testing the AUC in 30% of the data. This analysis was bootstrapped as well and expectedly the AUC scores were quite poor (**Figure 35**).

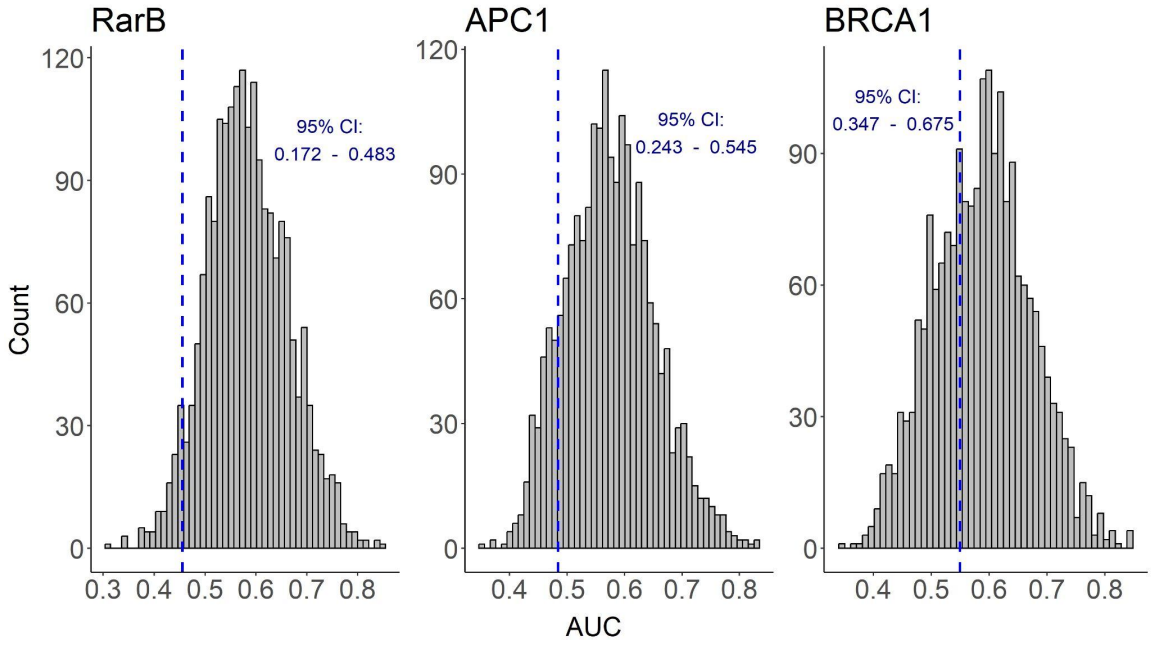


Figure 35. Bootstrap frequencies of the ROC AUCs for three gene promoters.

A proportion analysis of the zero data, using the B^2 statistic and Wilcoxon rank sum exact test on the non-zero data was performed in a two-part analysis to determine whether there is a difference in methylation between cases and controls among the three genes. The combined X^2, statistic which sums the B^2 statistic and W^2 from the Wilcoxon test, was computed and could be an alternative to the Mann–Whitney U test when there are many zero data points, as is the case here. Based on the chi-squared distribution at two degrees of freedom, for all three genes, there was no significant difference in methylation between cases and controls. The X^2 p-value was 0.959 and 0.531 for *RARB* and *BRCA1*, respectively. The X^2 statistic could not be computed for the *APC1* gene as we used the maximum RFU across the temperature ranges for interpolation, and hence, there were no zero values. The results using the permutation method on the X^2 showed highly similar values to the p-values obtained by looking at the chi-squared distribution (**Figure 36**).

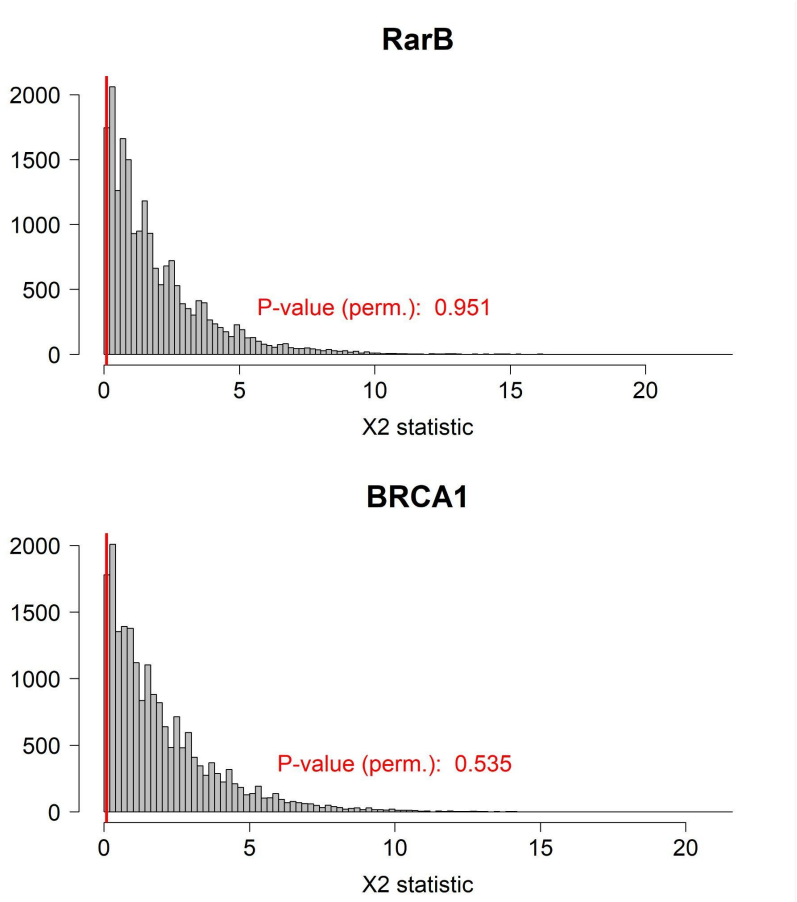


Figure 36. Permutation of the X^2 statistic which was used to determine whether there was a difference in methylation between cases and controls among the three genes. The red line represents the X^2 value in the original sample.

We also performed a zero-inflated model, which is also a two-part statistical analysis method, in which we can account for the plate of origin for the analysed samples. Again, this type of analysis could not be applied to the methylation data of *APC1* for the reasons mentioned above. Nevertheless, neither *BRCA1* nor *RARB* had a significantly different methylation between cases and controls. The same outcome was achieved when we performed the tobit regression analysis, which assumes that the data is normally distributed but that the values are censored at 0. Results of the tobit regression and zero-inflated regression model are reported in **Table 11**.

Table 11. Zero-inflated (ZI-model) and tobit regression model results for the RARB and BRCA1 gene promoters.

Model	Parameters	Estimate	SE	t-value	p-value	Gene
ZI-model	(Intercept)	-0.28	0.25	-1.10	0.275	
	Methylation	0.11	0.18	0.60	0.551	
	Plate	0.74	0.38	1.95	0.054	RARB
Tobit	(Intercept)	0.30	0.13	2.38	0.018	
	Methylation	-0.06	0.09	-0.64	0.523	

Model	Parameters	Estimate	SE	t-value	p-value	Gene
	Plate	-0.38	0.19	-1.99	0.046	
	logSigma	-0.10	0.10	-0.96	0.338	
ZI-model	(Intercept)	-0.36	0.26	-1.40	0.164	
	Methylation	0.08	0.08	0.94	0.349	
	Plate	0.73	0.38	1.94	0.055	BRCA1
Tobit	(Intercept)	0.35	0.13	2.73	0.006	
	Methylation	-0.04	0.04	-1.02	0.308	
	Plate	-0.37	0.19	-1.99	0.047	
	logSigma	-0.11	0.10	-1.05	0.296	

Like the PRS score, the methylation values on promoters of the three analysed genes were not found to be associated with BC and were hence not included among the predictors in the models.

Small-RNA sequencing

Before generating the raw counts of miRNAs, we performed a quality control of the small-RNA sequencing chips. The small-RNA sequencing was performed on 8 IonTorrent Chips, and the number of samples included on each chip ranged from 18 to 24. The percentage of chip wells that contained the Ion Sphere Particle (ISP) ranged from 75% to 94% (**Figure 37**). The percentage of reads which passed all the filters and were recorded for future processing ranged from 14% to 32%. Overall, the results were good for all chips with minor coverage issues on chip 4.

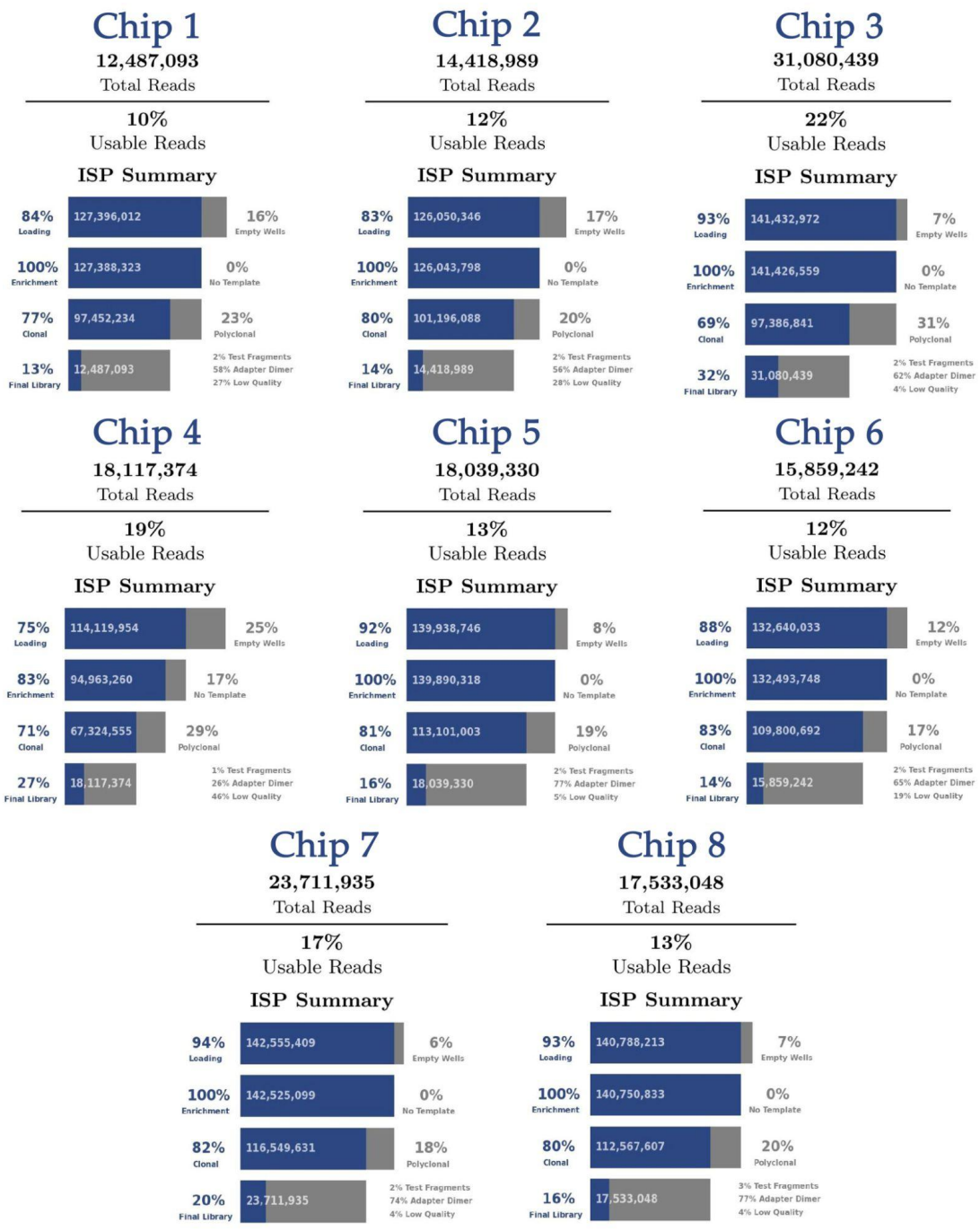


Figure 37. Quality metrics report of small-RNA sequencing reads based on the Ion Torrent Software. Results are shown for all eight chips on which the 131 samples were analysed.

Individual miRNA analysis

From the seven normalisation methods evaluated, the Deseq normalisation performed the best for the miRNA data, while the Poisson and Quantile normalisations were quite close in performance as well (**Figure 38**). Hence, the Deseq normalisation was chosen for the miRNA analysis.

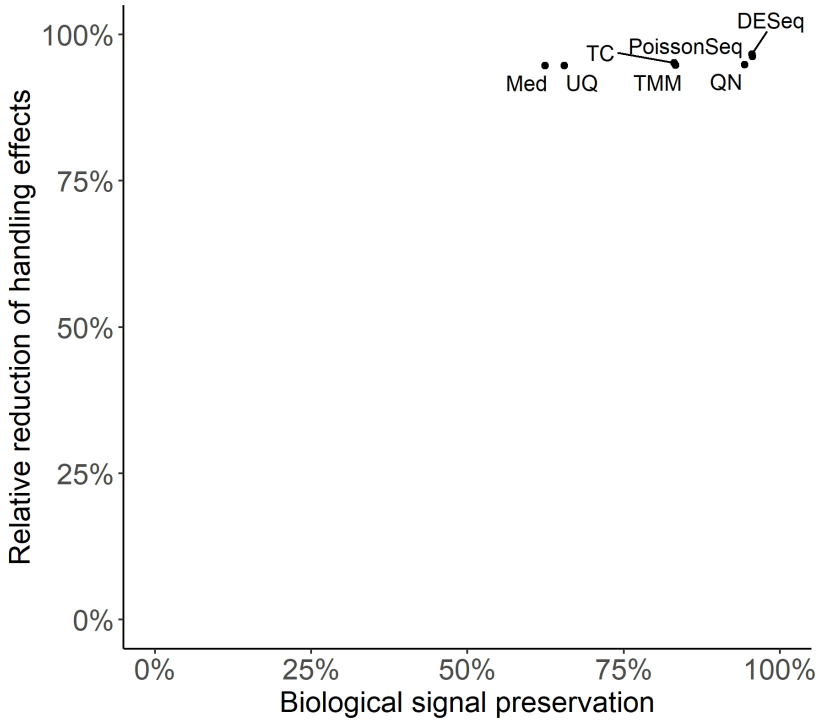


Figure 38. Summary metrics of the DANA normalisation assessment tool where the reduction of handling effects and biological signal preservation are plotted.

To visually inspect our miRNA raw count data and the cut-offs selected for poor/well expressed miRNAs used for normalisation assessment, the log expression histogram plot as well as the mean of the log counts and their standard deviation are shown in **Figure 39**. We utilised the Deseq normalisation provided by the *DESeq2* R package for all the analyses on individual miRNAs. Additionally, all miRNAs with raw counts < 20 were excluded for a total of 104 unique miRNAs.

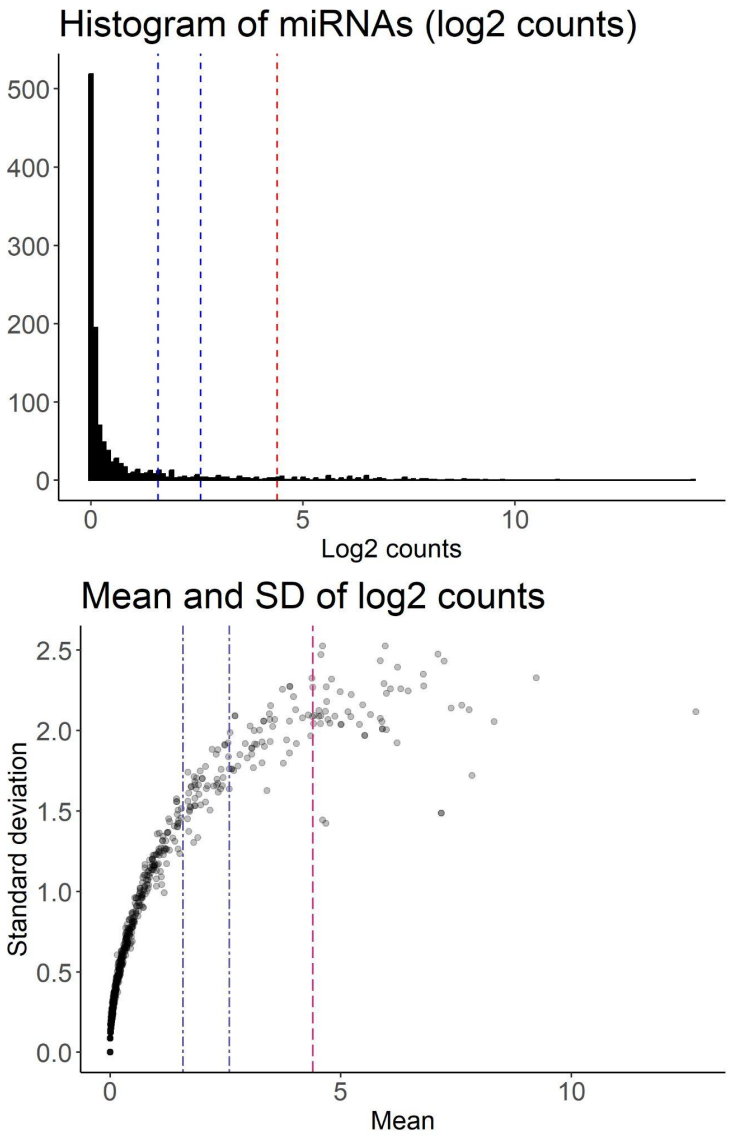


Figure 39. Histogram of log2 miRNA counts as well as the mean and SD of the $\log_2$ counts are shown. The blue vertical lines indicate the lower and upper cut-off for the poorly expressed miRNAs while the red vertical line represents the cut-off for the well-expressed miRNAs.

The descriptive statistics of individual miRNAs were performed on the variance stabilising transformation output which we will refer to as variance stabilised data (vsd) from now on. The coefficient of variation is relatively stable, ranging from 0 to 0.4, with the most frequent CV being around 0.28 (**Figure 40**).

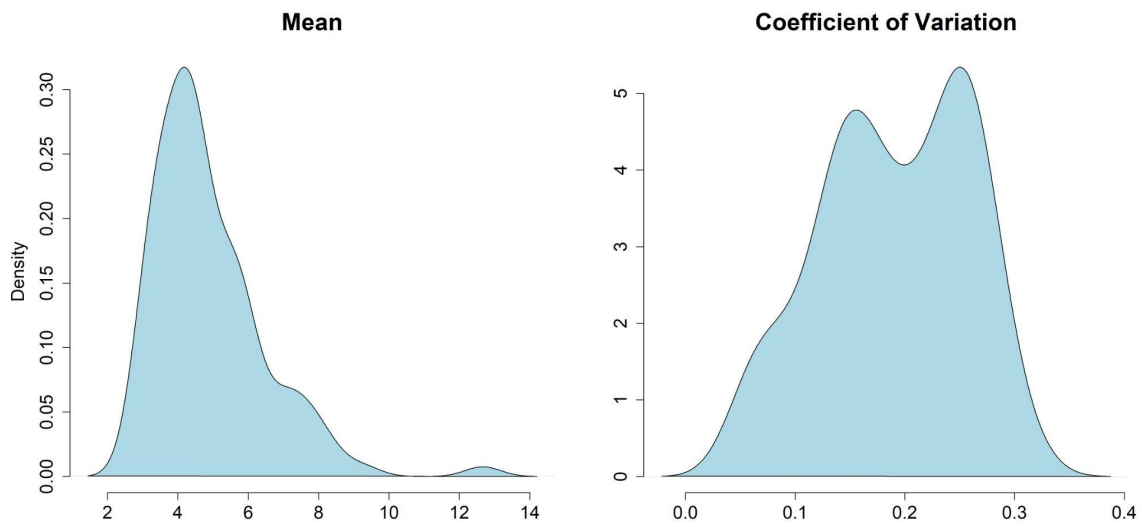


Figure 40. Density plots of mean and CV of vsd miRNAs.

Unsupervised hierarchical clustering on the analysed miRNAs did not create any apparent sample subdivisions associated with BC status, while the miRNAs were grouped according to their expression levels (**Figure 41**).

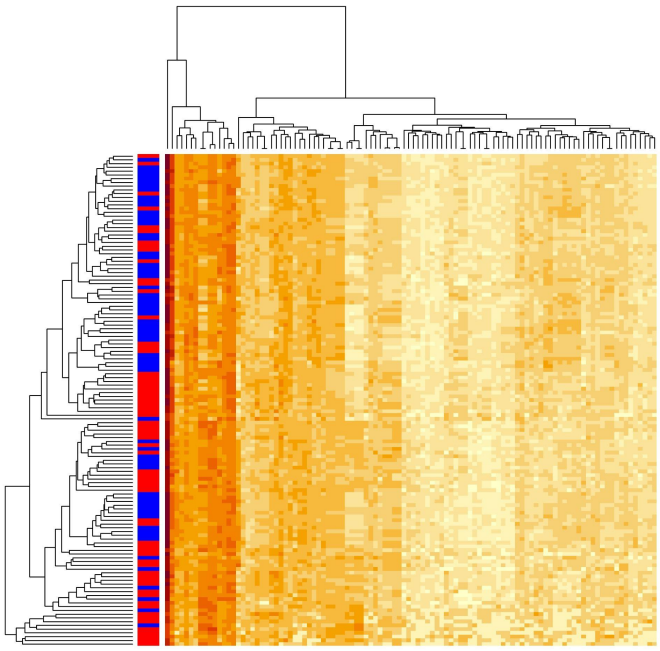


Figure 41. Heatmap on the complete set of clustered individual miRNAs (vsd). The vertical column on the far left indicates BC cases in red and controls in blue.

In the PCA of the variance stabilised miRNAs, PC 1 and PC 2 explained 26.4% and 10.9% of the variance, respectively. These two principal components were visualised in **Figure 42**, and a separation of cases and controls was observed to some extent.

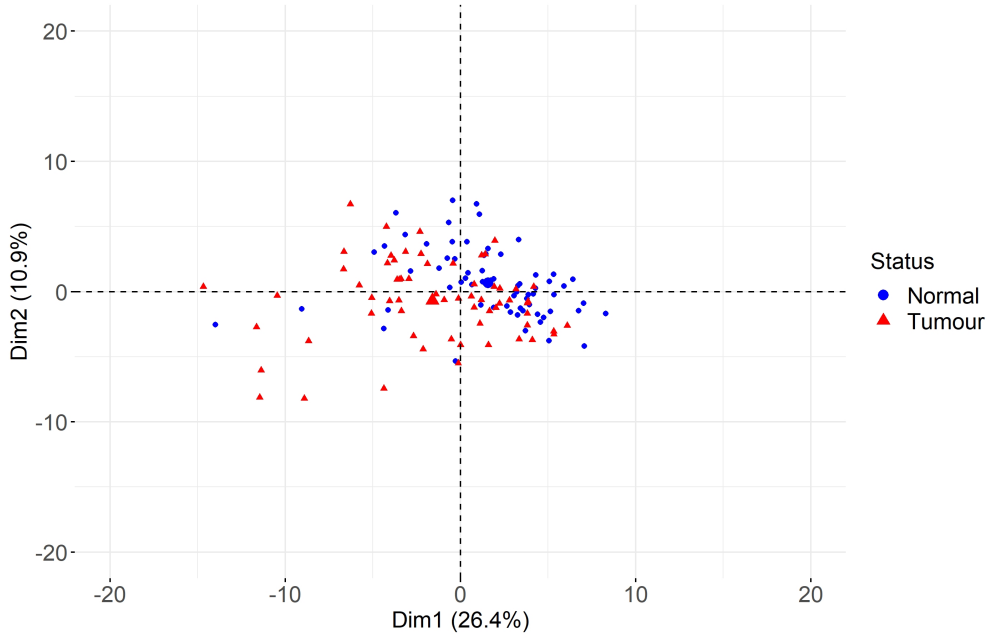


Figure 42. PC1 and PC2 plot from the PCA on individual miRNAs.

In order to investigate and compare the ranges and variability between the miRNAs, a boxplot of all miRNAs was also plotted and the outlying miRNA (miR-451) can be seen on the far right of the plot (**Figure 43**).

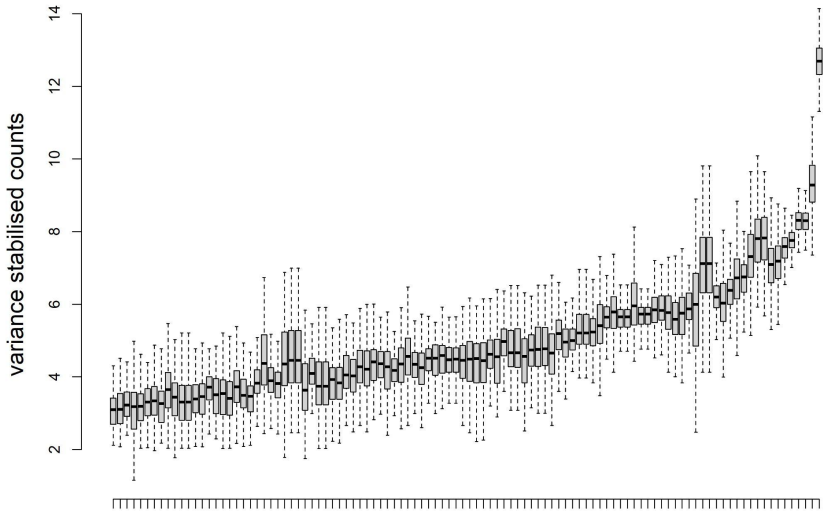


Figure 43. Boxplots of all analysed individual miRNAs.

To determine which cfc miRNAs are differentially expressed between cases and controls, we performed a class comparison on the filtered miRNA counts using the *DESeq2* package. Within the package pipeline, the data would be Deseq normalised followed by a class comparison. Of 104 miRNAs, 27 were differentially expressed between healthy and tumour samples (**Figure 44** and *Supplementary Table 8* – Appendix B).

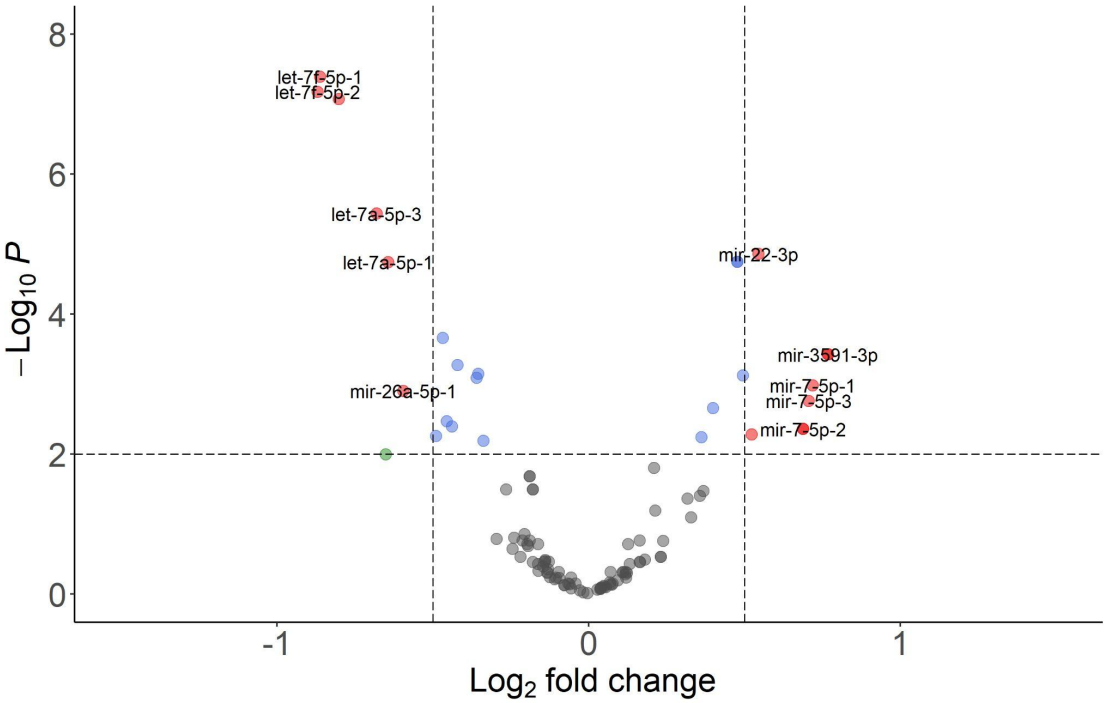


Figure 44. Volcano plot of the plasma miRNA differential expression analysis results between BC cases and controls. The p-value shown in the plot is the Benjamini–Hochberg adjusted p-value. Vertical dotted lines indicate 1.5 $\log_2$ fold deregulation. The red points indicate miRNAs above the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the blue points indicate miRNAs below the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the green points are miRNAs above the $\log_2$ fold deregulation cut-off and above the p-value cut-off, while the grey points indicate miRNAs that do not meet either of the criteria.

Some of the miRNAs, such as let-7f-1 and let-7f-2, which are located at different chromosomes and have a highly correlated count profile, have identical 5p mature sequences. When comparing invasive vs normal samples, there were 28 differentially expressed miRNAs. All 27 miRNAs that were differentially expressed in tumour vs normal were also differentially expressed in invasive vs normal patients. The added miRNA in the latter comparison was mir-15b-5p, which was also close to being significant in the analysis on all tumour vs normal samples ($\log_2$ fold change: -0.65 and adjusted p= 0.0102). No differentially expressed miRNAs were found when comparing in situ to either invasive or normal samples.

miRNA ratios

Since there are no suitable normalisers for small non-coding RNAs when using the RT-qPCR technique, for the discovery of potential biomarkers associated with BC detection, we computed ratios on miRNAs followed by the filtering and biomarker discovery techniques. The reasoning was that all potential biomarkers that would be usable in clinics should be based on the cheap and well-known RT-qPCR platform. Therefore, the miRNA ratios were first computed based on small-RNA sequencing data, and promising ratios were then tested using RT-qPCR.

From the 104 miRNAs with a mean count larger than 20 across the 131 samples, 97 miRNAs with unique count profiles remained. From these 97 miRNAs, we computed 4,656 miRNA ratios. In hindsight, following the ratio computation, $\log_2$ transformation would have been optimal for the optimal comparison with RT-qPCR data ratios obtained subsequently. Finally, the flowchart of the miRNA selection in the discovery cohort can be seen in **Figure 45**.

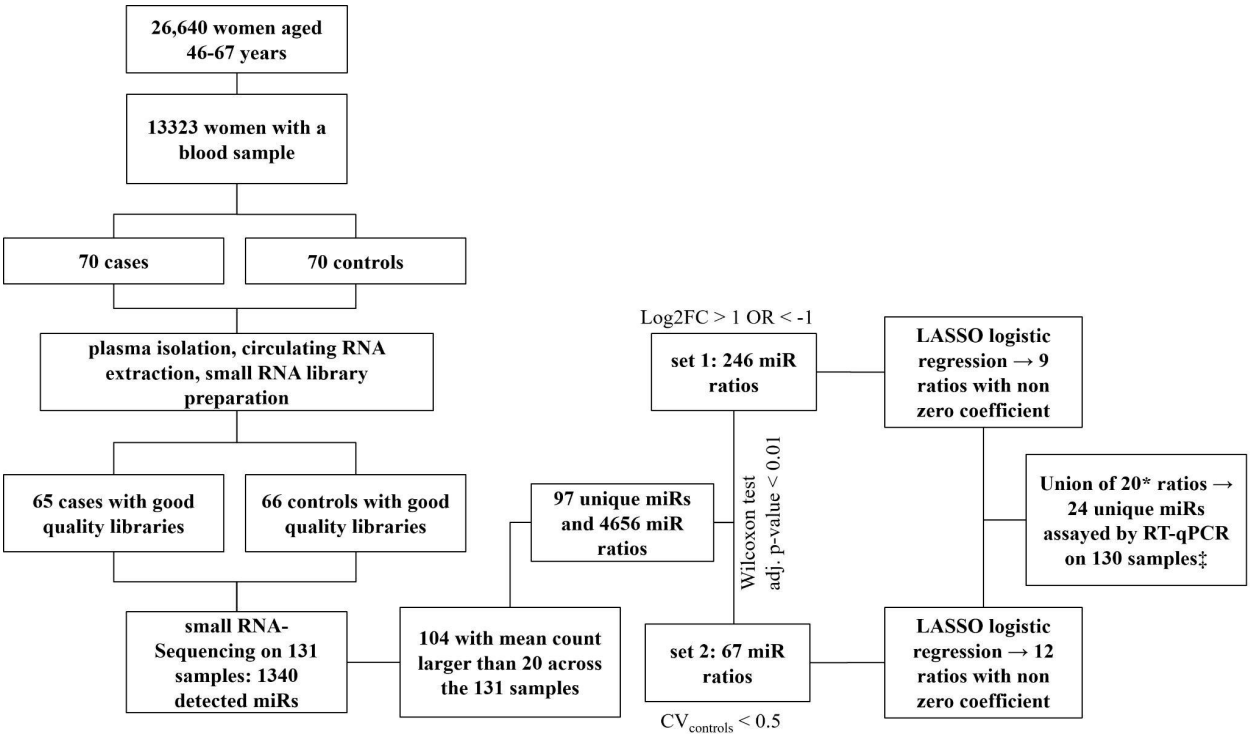


Figure 45. Flowchart of the discovery cohort pipeline. * The let-7f-5p-2_miR-103a-3p-2 ratio was removed as miR-103a-3p-2 and miR-103a-3p-1, found in the let-7f-5p-1_miR-103a-3p-1 ratio, had almost identical counts and their ratio partners had identical mature miRNA sequences. ‡ One sample had to be excluded in the RT-qPCR step due to insufficient plasma volume.

Descriptives statistics

The mean and coefficient of variation density plots of the computed miRNA ratios can be seen in **Figure 46**. Compared to individual miRNAs, the range of the CV within the ratios is much larger, with the most commonly observed CV being around 1.

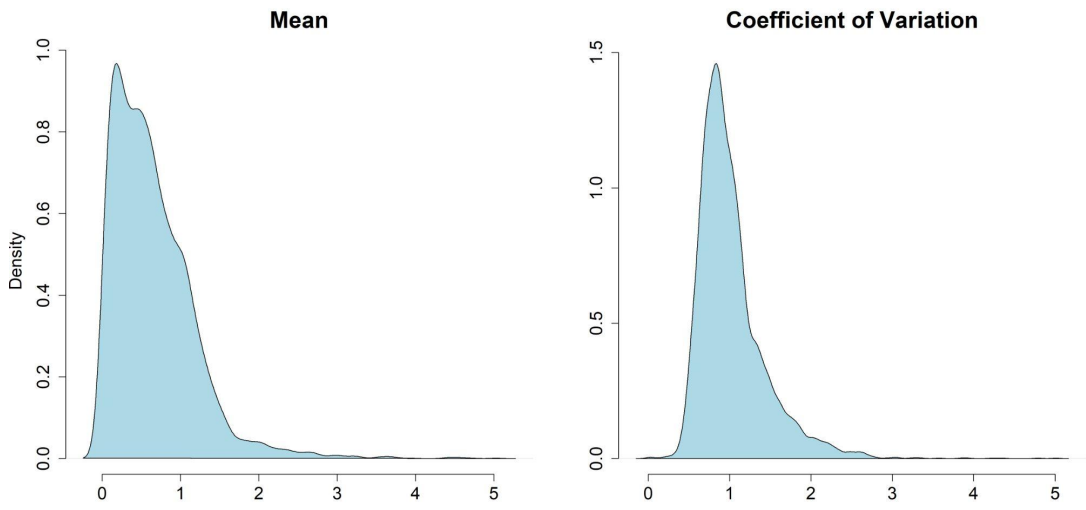


Figure 46. Density plots of mean and CV of miRNA ratios computed from small-RNA sequencing data.

To investigate the miRNA ratio data, a PCA was created on the total matrix of miRNA ratios and 30.1% and 9.6% variance was explained by PC 1 and PC 2, respectively. A separation of cases and controls was observed when plotting the PC 1 and PC 2. Additionally, the top 50 ratios with highest absolute loading values in PC1 had an overall high correlation (**Figure 47**), which might somewhat explain the low variance rates explained by the PCs.

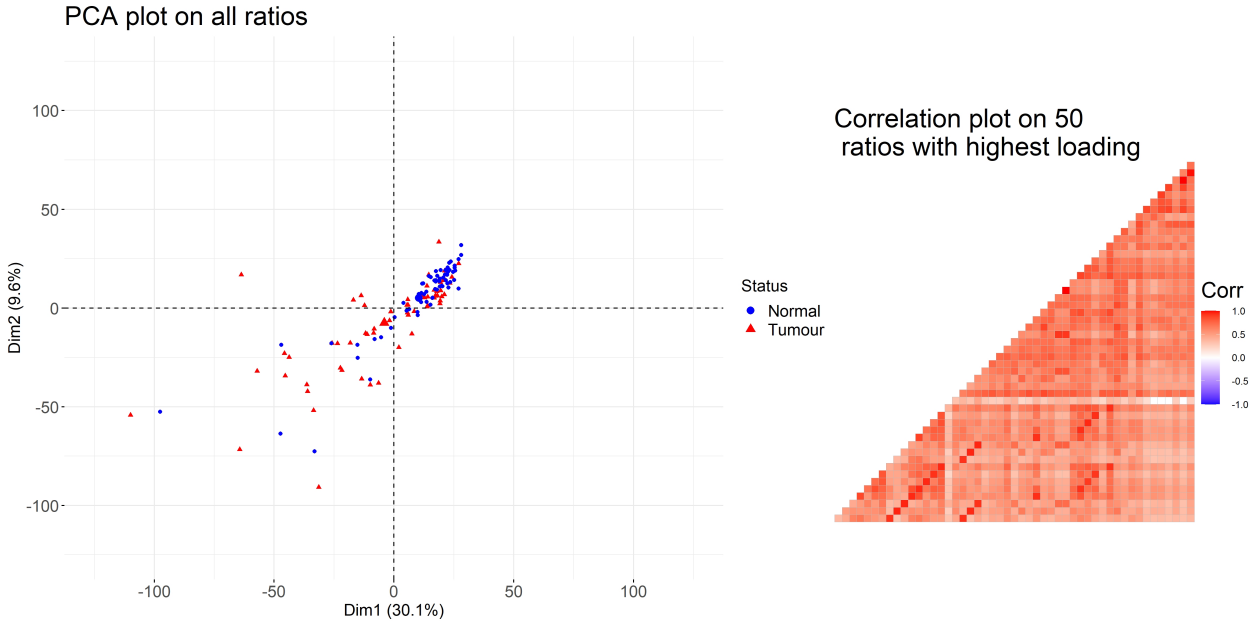


Figure 47. A plot of PC1 and PC2 from the PCA on miRNA ratios can be seen on the left, while on the right is shown the correlation plot of the top 50 miRNA ratios with the highest loading values in PC1.

Variable selection

As most ratios were not normally distributed, we performed the Mann–Whitney U test between cases and controls. This was the initial variable filtering step. Based on the Benjamini–Hochberg adjusted p-value < 0.01 cut-off, 886 miRNA ratios were significantly different between cases and controls (**Figure 48**). We filtered these differentially expressed miRNA ratios based on two distinct strategies for obtaining candidate biomarkers associated with BC. The first strategy filtered the miRNA ratios based on the fold change (FC > 2 or FC < 0.5) and resulted in 246 ratios, while the second strategy only included the differentially expressed miRNA ratios which had a coefficient of variation < 0.5 in controls for a total of 67 ratios. The second strategy was supposed to represent the more stable miRNA ratios.

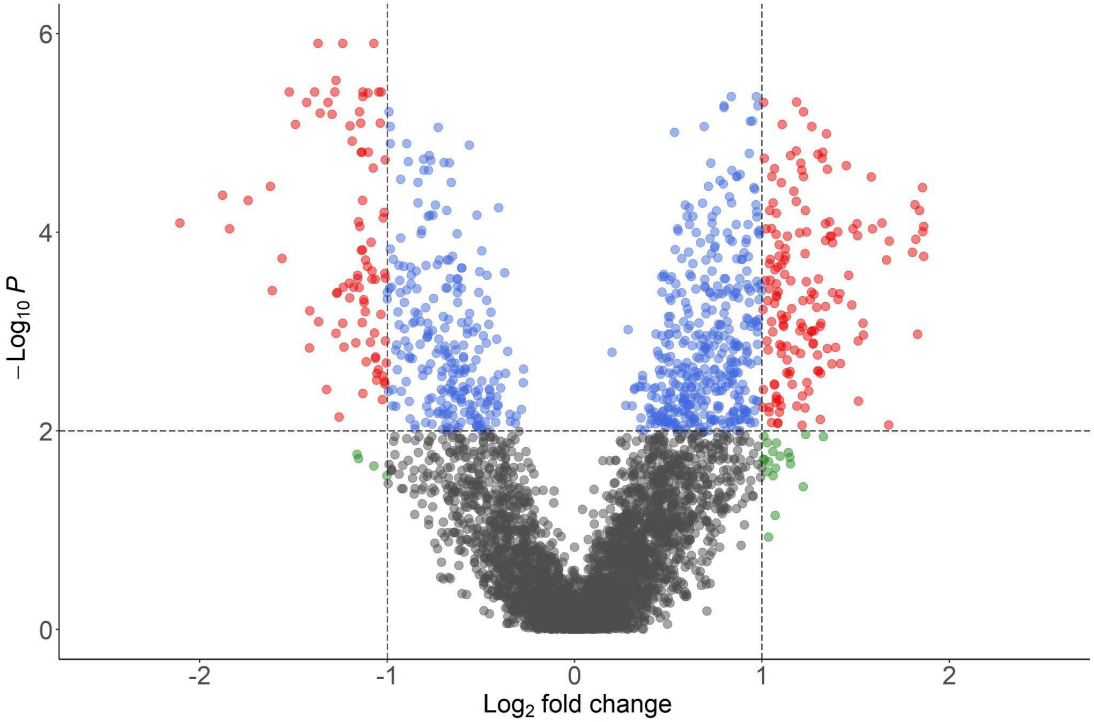


Figure 48. Volcano plot of the Mann–Whitney U test results on the miRNA ratios. The p-value shown in the plot is the Benjamini–Hochberg adjusted p-value. The red points indicate miRNAs above the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the blue points indicate miRNAs below the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the green points are miRNAs above the $\log_2$ fold deregulation cut-off and above the p-value cut-off while the grey points indicate miRNAs that do not meet either of the criteria.

To select candidate miRNA ratios associated with BC detection, we performed penalised logistic regression analysis. The optimal penalty parameter (λ) was selected by a cross-validation LASSO logistic regression performed on the ratios from the two strategies separately. From strategy 1, nine miRNA ratios were selected, while from strategy 2, twelve miRNA ratios were selected. The log(λ) from the two cross-validation LASSO logistic regressions can be seen in **Figure 49**, while the coefficients of the two miRNA ratio sets can be seen in **Table 12**.

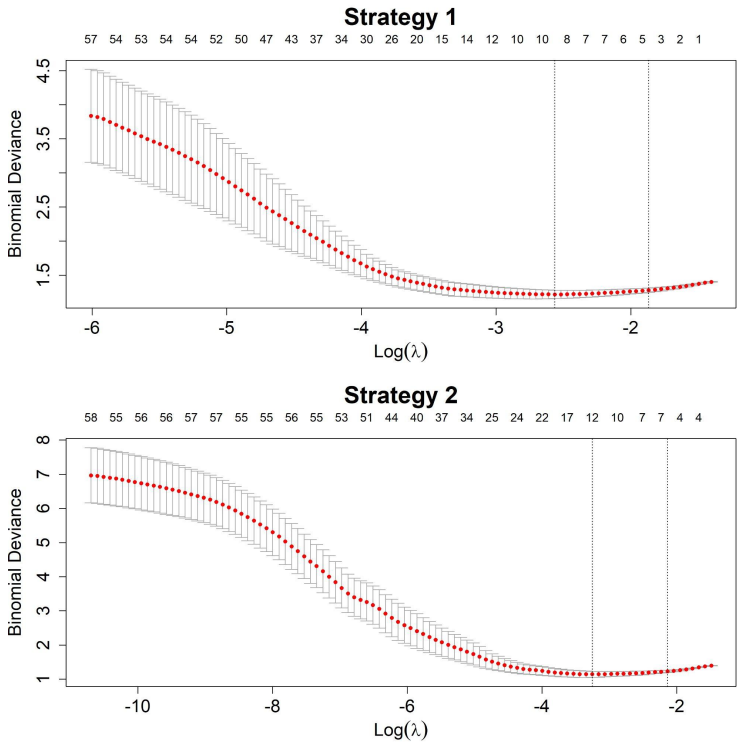


Figure 49. The log(λ) plots for variable selection of strategies 1 and 2 using cross-validation LASSO logistic regression. Lambda minimum and 1 standard error from minimum lambda are represented by the left and right vertical lines, respectively.

For descriptive purposes, a univariate logistic regression was performed on these 21 miRNA ratios, the ROC AUC was calculated, and the results are reported in *Supplementary Table 9* (Appendix B). All the ratios were significantly associated with BC, where 16 ratios had an OR less than 1 and five had an OR larger than 1. The individual miRNA ratio ROC AUC ranged from 0.66 to 0.88. These 21 miRNA ratios were chosen to be validated using the RT-qPCR platform.

Table 12. LASSO logistic regression coefficients of the selected miRNA ratios with non-zero coefficients in strategy 1 and strategy 2.

Strategy 1		**Strategy 2**	
Intercept	1.124	Intercept	2.950
miR-335-5p_let-7f-5p-2	0.002	miR-26b-5p_miR-142-5p	-0.129
miR-199a-3p-2_let-7a-5p-2	0.259	let-7a-5p-2_miR-106b-5p	-0.816
miR-199a-3p-2_let-7f-5p-2	0.001	let-7f-5p-1_miR-103a/b*1	-1.107
let-7a-5p-2_miR-22-3p	-0.535	let-7f-5p-2_miR-103a/b*2	-0.652
let-7a-5p-2_miR-320a	-0.373	miR-93-5p_miR-19b-3p-1	-2.803
let-7f-5p-1_miR-19b-3p-1	-5.019	miR-22-3p_miR-19b-3p-2	2.360
miR-27a-3p_miR-122-5p	-0.473	miR-101-3p-2_miR-19b-3p-1	-3.031
let-7f-5p-2_miR-146a-5p	-0.199	miR-30d-5p_miR-20a-5p	0.147
miR-15b-5p_miR-16-5p-1	-0.067	let-7b-5p_miR-19b-3p-1	-0.836
		miR-15a-5p_miR-16-5p-2	-0.073
		miR-20a-5p_miR-19b-3p-1	-2.085

Strategy 1	Strategy 2	
	miR-21-5p_miR-23a-3p	-0.271

Finally, boxplots stratified by BC status and a heatmap together with the correlation plot of the 21 ratios were created and are shown in **Figure 50** and **Figure 51**, respectively. Notably, the plots were made on the $\log_2$ transformed ratios from **Table 12**. Based on the selected miRNA ratios, the clustering of samples based on their BC status can be observed.

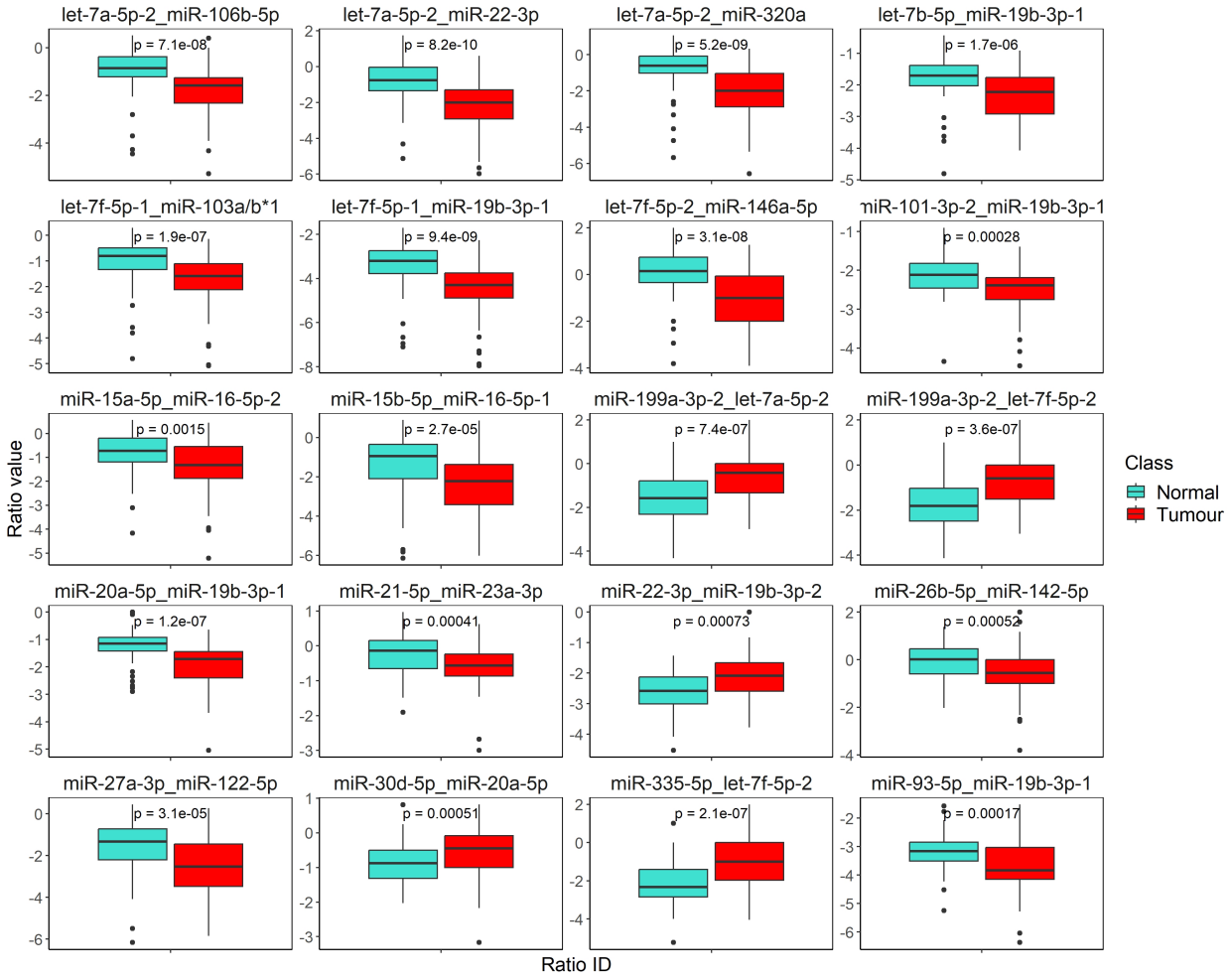


Figure 50. Boxplot of the log2 transformed miRNA ratios selected by the LASSO logistic regression in strategy 1 and strategy 2. Plotted are the 20 miRNA ratios that were validated in RT-qPCR (see the legend of **Figure 45** above, with the study flowchart).

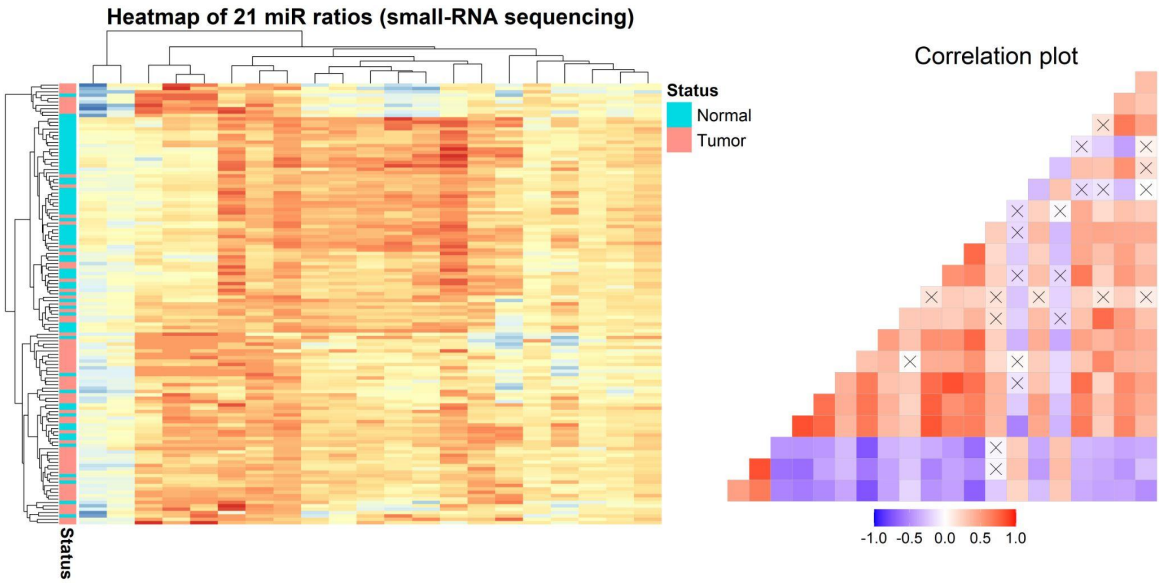


Figure 51. Heatmap of the 21 selected miRNA ratios based on small-RNA sequencing data (left) and their correlation plot (right).

To evaluate the performance of the two strategies and their selected ratios, a five-fold cross-validation was performed on the whole procedure starting from a Mann–Whitney U test followed by a LASSO regression. For both strategies, we evaluated the prediction error, ROC AUC, calibration of the intercept and slope based on the calibration-in-the-large as well as the scaled brier score (**Table 13**). It is important to mention that the selected miRNA ratios were not exactly the same across the cross-validation folds, but that the most impactful ratios (i.e., with the largest absolute beta coefficient) were consistently selected across folds and were found in the original list of ratios based on the complete sample.

Table 13. Performance of the two strategies for selecting miRNA ratios based on averaged values from the cross-validation.

Statistic	Strategy 1	Strategy 2
Pred. Error	0.30	0.30
AUC	0.80	0.77
Intercept	0.12	0.08
Slope	1.13	0.96
Scaled Brier	0.02	0.04

Bayesian variable selection

As an alternative to the frequentist variable selection and evaluation approach, we performed biomarker selection using hierarchical shrinkage models based on the horseshoe priors. This Bayesian approach was performed on the total set of ratios (n = 4656), as well as the ratios from strategy 1 and strategy 2, although this method is best used in high-dimensional contexts. In all three instances, three miRNA ratios were selected. In **Table 14**, summarised are the Kullback–

Leibler (KL) divergence between the full model and the submodel as well as the explanatory power of predictors (ELPD) among the three runs (all ratios, strategy 1 and strategy 2). Across the three runs, eight unique ratios were selected.

Table 14. Model characteristics on variable selection using hierarchical shrinkage model. The Kullback–Leibler divergence and explanatory power of predictors are shown for all miRNA ratios as input or miRNA ratios from the two strategies mentioned above.

Model	KL	ELPD
All miRNA ratios		
Intercept only	0.24	-91.11
let-7f-5p-2_miR-22-3p	0.13	-71.57
miR-1260b_miR-20a-5p	0.11	-67.49
miR-3529-7_miR-26a-5p-1	0.10	-66.12
Strategy 1 (246 ratios)		
Intercept only	0.17	-91.15
let-7f-5p-2_miR-22-3p	0.05	-71.33
miR-122-5p_miR-21-5p	0.04	-69.90
miR-15b-5p_miR-122-5p	0.03	-68.47
Strategy 2 (67 ratios)		
Intercept only	0.09	-91.23
miR-425-5p_miR-20a-5p	0.07	-86.25
let-7a-5p-2_miR-106b-5p	0.06	-83.07
miR-26b-5p_miR-19b-3p-1	0.05	-82.43

A ridge logistic regression cross-validation was performed on these sets of ratios (grouped as in the outputs of the three runs), and we evaluated all the parameters in the cross-validation mentioned above (**Table 15**). A cross-validation on the whole Bayesian procedure was not possible due to extremely high computational time. Additionally, we averaged the coefficients and intercept of the ridge regression models across the folds, which are reported in **Table 16** for each of the three sets.

Table 15. Performance based on cross-validation of the ridge regression on miRNAs selected by the three hierarchical shrinkage models.

	Pred. Error	AUC	Intercept	Slope	Scaled Brier
All miRNA ratios	0.25	0.85	-0.03	1.86	0.11
Strategy 1 (246 ratios)	0.23	0.84	-0.81	4.42	0.19
Strategy 2 (67 ratios)	0.22	0.81	0.05	1.35	0.05

Table 16. Averaged ridge logistic regression coefficients based on the 5-fold cross-validation on the miRNA ratios selected by the three hierarchical shrinkage models.

All miRNA ratios	
Intercept	0.126
let-7f-5p-2_miR-22-3p	-2.161
miR-1260b_miR-20a-5p	3.854
miR-3529-7_miR-26a-5p-1	0.117

Strategy 1 (246 ratios)	
Intercept	1.444
let-7f-5p-2_miR-22-3p	-2.272
miR-122-5p_miR-21-5p	0.065
miR-15b-5p_miR-122-5p	-0.547
Strategy 2 (67 ratios)	
Intercept	1.182
miR-425-5p_miR-20a-5p	3.162
let-7a-5p-2_miR-106b-5p	-2.276
miR-26b-5p_miR-19b-3p-1	-14.755

Not all the miRNAs found in the ratios obtained with the Bayesian approach were tested by RT-qPCR; however, as will be seen later, there were some common ratios between the hierarchical shrinkage model and the frequentist approach.

RT-qPCR assaying of miRNAs

By combining the miRNA ratios from the two above-mentioned strategies, a total of 20 ratios, which included 24 unique miRNAs, were further analysed by RT-qPCR on 130 samples. This was done because RT-qPCR is more robust and more clinically used than small-RNA sequencing. One ratio (let-7f-5p-2_miR-103a-3p-2) was removed as miR-103a-3p-2 and miR-103a-3p-1, found in the let-7f-5p-1_miR-103a-3p-1 ratio, had identical counts in all but two samples (with a negligible difference) and their ratio partners had identical mature miRNA sequences. In addition, one control sample had to be excluded from the RT-qPCR step due to insufficient plasma volume.

For each sample, the miRNAs were analysed in triplicates. The mean Ct and SD across the replicates for each miRNA, stratified by BC status, can be seen in **Figure 52**. Most of the miRNAs were rather stably expressed with a relatively small overall SD across replicates, as the majority of SDs were around or below 1. Furthermore, most of the Cts were within the 20 to 35 range.

The miRNAs that had poor RT-qPCR results were miR-15a-5p and miR-22-3p due to their high variability within the triplicates. Importantly, in a large percentage of samples, miR-16-5p had a mean Ct below 20, which is a low Ct value and could be explained by the fact that miR-16 is highly and consistently abundant in blood. Considering the precautions we undertook regarding haemolysis, we believe that it played a minor role in affecting the expression values of the analysed miRNAs.

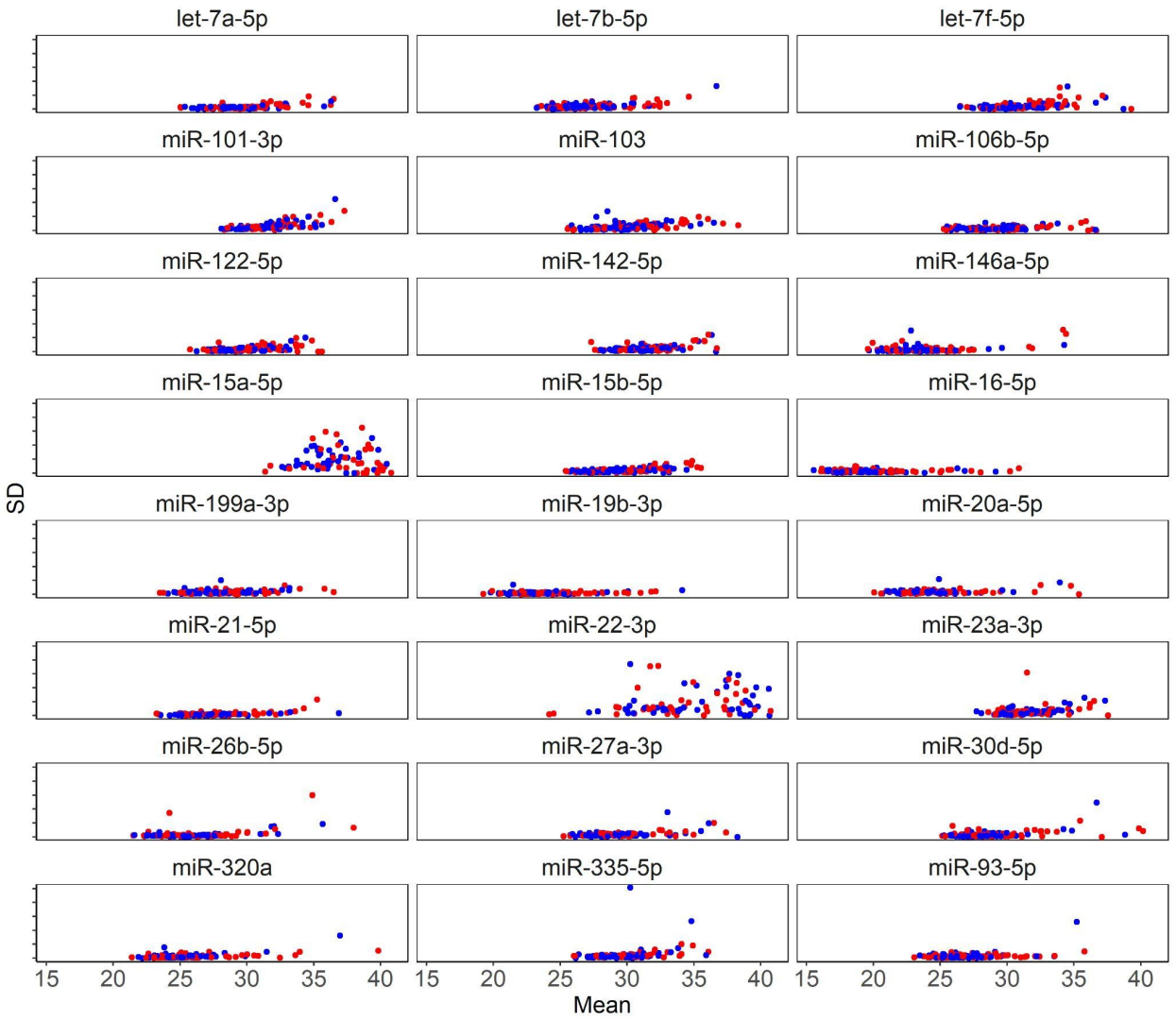


Figure 52. Mean and SD of Cts for each miRNA assayed by RT-qPCR in the discovery cohort. The red points represent BC cases, while the blue points represent controls. The SD ranges from 0 to 5 in every miRNA subplot.

The miRNA ratios identified as candidates for discriminating between BC cases and controls using small-RNA sequencing were then computed using RT-qPCR data by Ct $miRNA_{(Y)}$ - Ct $miRNA_{(X)}$, where $miRNA_{(Y)}$ and $miRNA_{(X)}$ are the denominator and numerator of the original NGS ratios, respectively. The mean, SD and coefficient of variation were calculated for each miRNA ratio constructed from the RT-qPCR data and their density plots can be seen in **Figure 53**. Due to the extremely wide range of CV, we only plotted the standard deviation. Additionally, a boxplot of the 20 ratios assayed in the qPCR platform is shown in **Figure 54**.

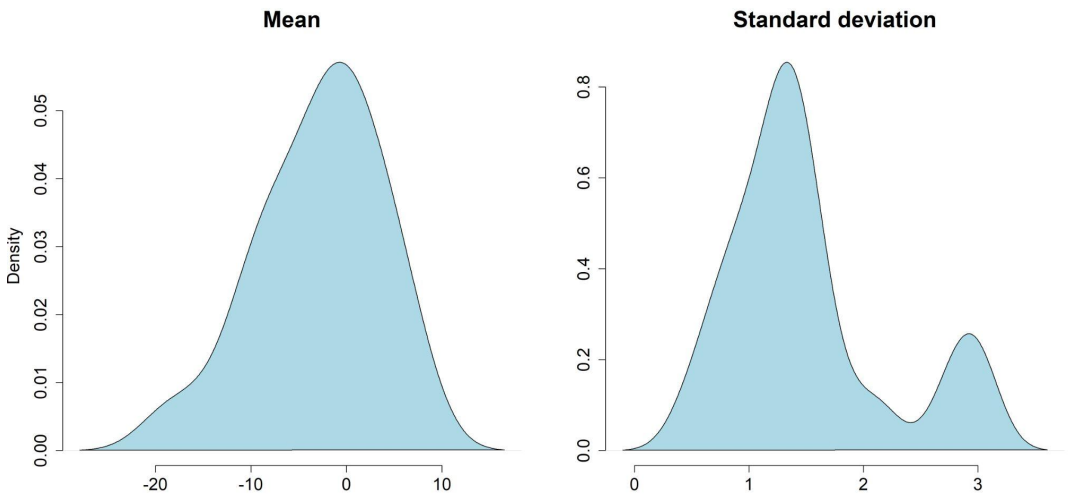


Figure 53. Density plot of mean and SD of the 20-miRNA ratio signature based on RT-qPCR data.

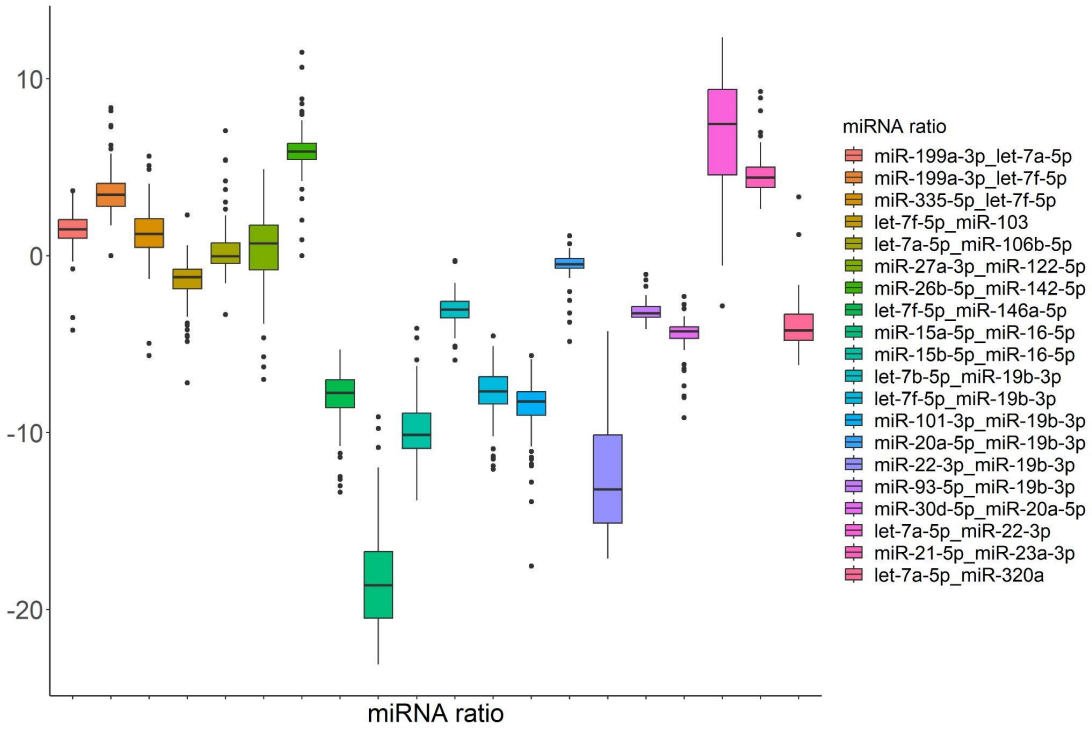


Figure 54. Boxplot of the 20-miRNA ratio signature based on RT-qPCR data.

A PCA was constructed on the 20 miRNA ratios, and 40.1% of the variance was explained in PC 1 and 15.1% in PC2. The PC1 and PC2 axes are visualised in **Figure 55** and the separation of cases and controls was not as clear as in the previously shown instances.

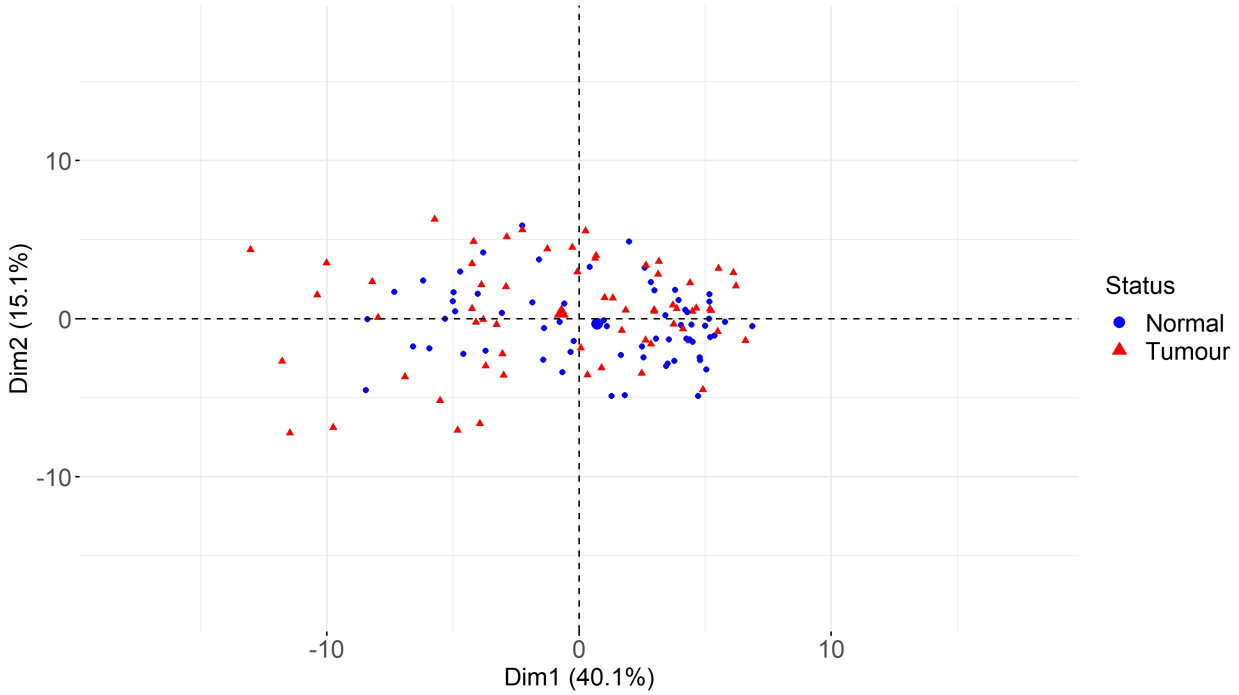


Figure 55. PC1 and PC2 of the PCA on 20-miRNA ratio signature based on RT-qPCR data.

We then created a heatmap and correlation plot of the mentioned ratios to determine how the miRNA ratio expression clusters and if any of the ratios are correlated. Several miRNA ratios were correlated, but the correlation among the predictors was not as prevalent as in the NGS data (**Figure 56**). No apparent clustering based on BC status was observed.

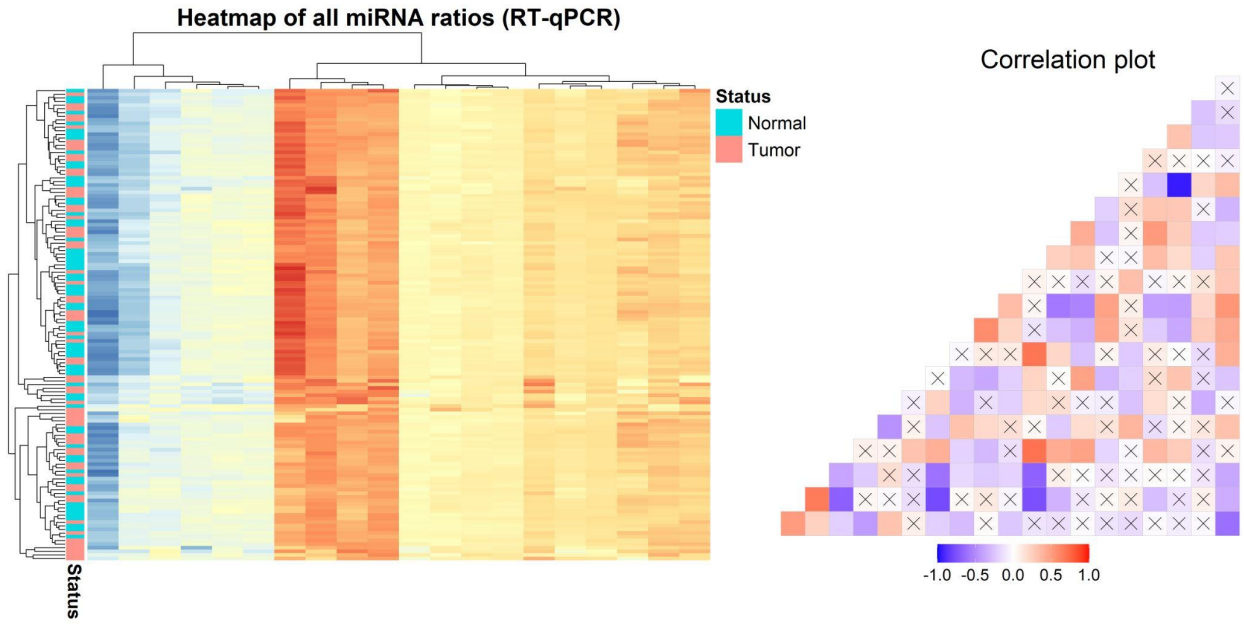


Figure 56. Heatmap and correlation plot of 20 miRNA ratio signature based on RT-qPCR data, where the squares without x refer to significantly positively (red) or negatively (blue) correlated pairs.

Based on the median ratio values, small-RNA sequencing and RT-qPCR had concordant values in cases and controls. Nevertheless, four ratios did show an opposite trend (**Table 17**). Seven ratios had a significantly positive Spearman rank correlation coefficient (p-value < 0.01) between the two platforms (miR-26b-5p_miR-142-5p, miR-101-3p_miR-19b-3p, let-7b-5p_miR-19b-3p, let-7f-5p_miR-19b-3p, let-7a-5p_miR-320a, miR-27a-3p_miR-122-5p, miR-199a-3p_let-7a-5p), with the coefficients ranging from 0.23 to 0.34 (**Table 17**). Albeit not significantly correlated, nine ratios had positive correlation coefficients < 0.20 and four had negative coefficients between the compared platforms.

Univariable logistic regression and AUC of the 20 ratios based on RT-qPCR data are reported in *Supplementary Table 10* (Appendix B). Overall, the individual ROC AUC ranged from 0.48 to 0.65, and three ratios were associated with BC at a nominal 5% level of significance: miR-26b-5p_miR-142-5p, let-7a-5p_miR-22-3p, and miR-199a-3p_let-7a-5p. Additionally, boxplots of the 20 ratios can be seen in **Figure 57**.

Table 17. Spearman correlation of the same miRNA ratio when comparing the NGS and RT-qPCR data.

miRNA ratio	Coefficient	p-value
let-7a-5p_miR-106b-5p	0.09	0.294
let-7a-5p_miR-22-3p	0.09	0.337
let-7a-5p_miR-320a	0.24	0.006
let-7b-5p_miR-19b-3p	0.25	0.005
let-7f-5p_miR-103	-0.08	0.380
let-7f-5p_miR-146a-5p	0.12	0.164
let-7f-5p_miR-19b-3p	0.24	0.006
miR-101-3p_miR-19b-3p	0.28	0.001
miR-15a-5p_miR-16-5p	0.07	0.457
miR-15b-5p_miR-16-5p	-0.12	0.168
miR-199a-3p_let-7a-5p	0.23	0.009
miR-199a-3p_let-7f-5p	0.07	0.417
miR-20a-5p_miR-19b-3p	0.15	0.087
miR-21-5p_miR-23a-3p	-0.11	0.210
miR-22-3p_miR-19b-3p	0.13	0.150
miR-26b-5p_miR-142-5p	0.35	< 0.001
miR-27a-3p_miR-122-5p	0.24	0.007
miR-30d-5p_miR-20a-5p	0.18	0.041
miR-335-5p_let-7f-5p	0.14	0.106
miR-93-5p_miR-19b-3p	-0.15	0.085

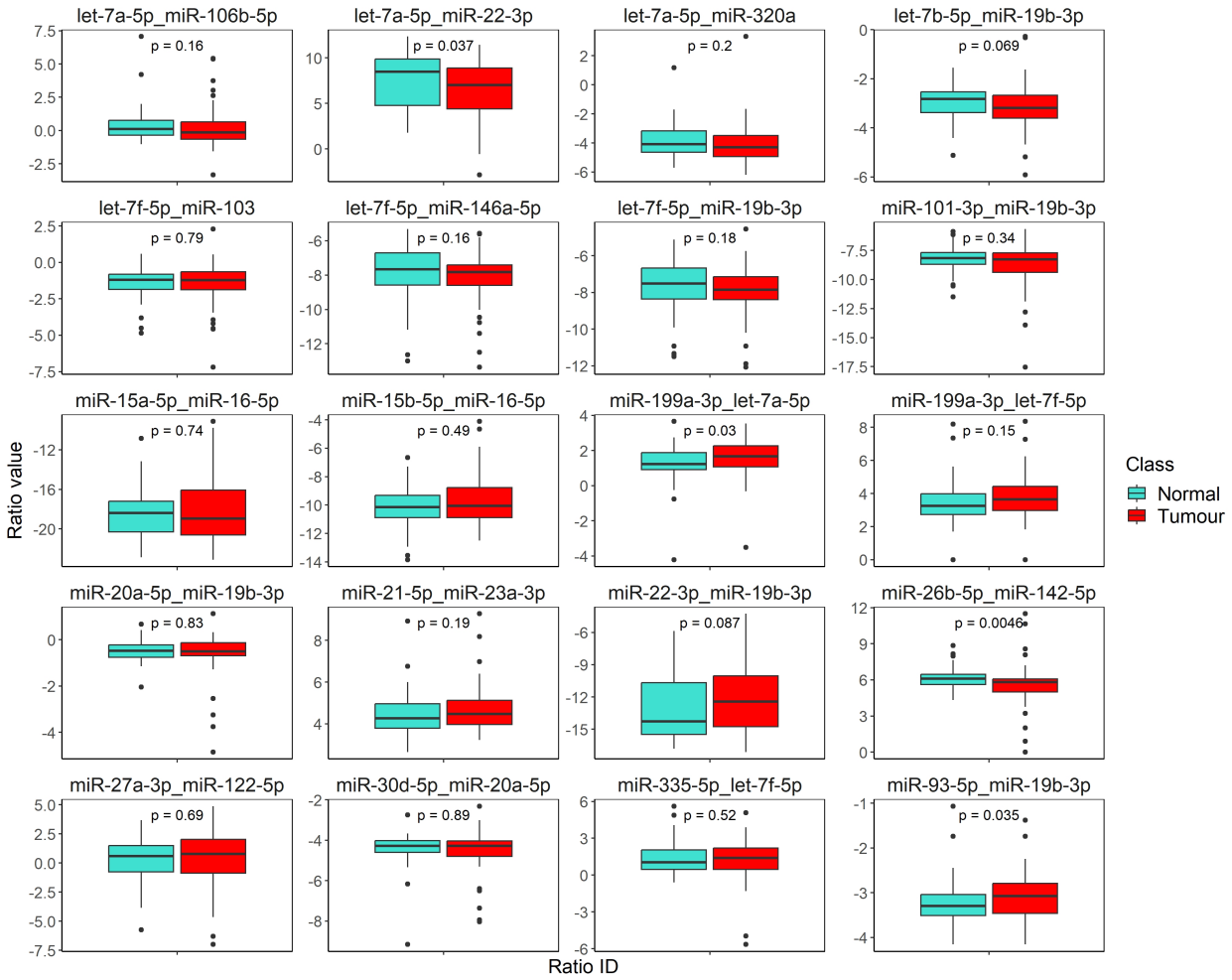


Figure 57. Boxplots of the 20-miRNA ratio signature on RT-qPCR stratified by BC status.

To identify the most promising miRNA ratios among the initial 20 in the RT-qPCR setting, a cross-validation penalised LASSO logistic regression was performed. Then, a cross-validation LASSO was performed on a set of predictors which included the 20 miRNA ratios and non-molecular variables associated with BC in our cohort. Those non-molecular variables were BMI (centred BMI was included in the model), breast density as classified by Tabar and WCRF lifestyle score. Additionally, menopause and the interaction term with centred BMI were included due to the known different effects of BMI in pre- and post-menopausal women. Finally, a cross-validation LASSO logistic regression was performed on only the non-molecular variables associated with BC in our cohort. The cross-validation log(λ) plots for the three models can be seen in **Figure 58**.

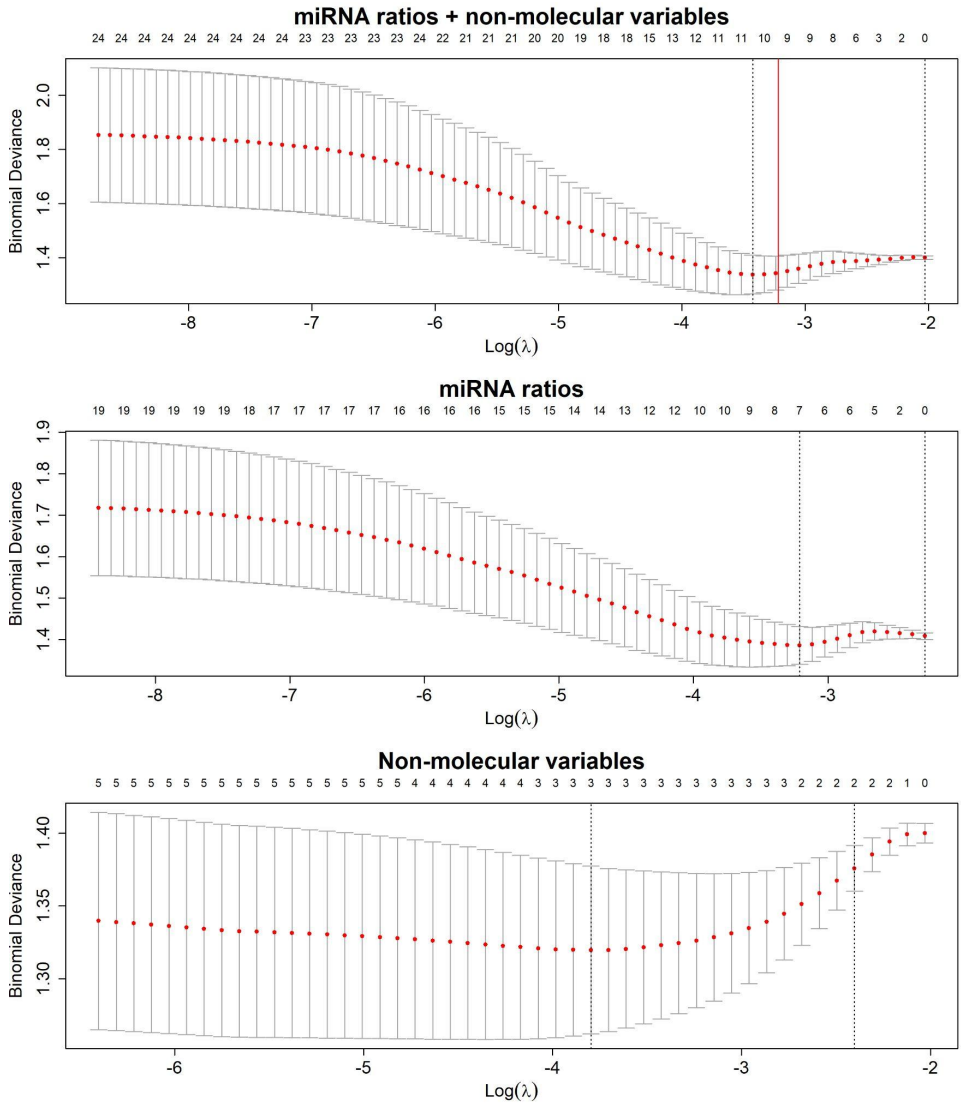


Figure 58. The log(λ) plot of the cross-validation LASSO logistic regression on miRNA ratios + non-molecular predictors, miRNA ratios only and non-molecular predictors only. Lambda minimum and 1 standard error from minimum lambda are represented by the left and right vertical lines, respectively. In the cv LASSO run with miRNA ratios and non-molecular variables we used a slightly higher λ cut-off (vertical line labelled in red) to obtain a more parsimonious model with at most 10 predictors selected.

The penalised LASSO logistic regression on miRNA ratios selected seven non-zero coefficient ratios, while the LASSO logistic regression on miRNAs and non-molecular variables selected the same seven miRNA ratios in addition to breast density (Tabar's scale), interaction term between BMI and menopause and WCRF lifestyle score. Finally, LASSO logistic regression on only non-molecular variables selected breast density, the interaction term between BMI and menopause and WCRF lifestyle score. The intercept and coefficients of the selected variables from the three models are found in **Table 18**.

Table 18. Predictors with non-zero coefficients from the three penalised LASSO logistic regressions.

	Combined	miRNA only	NM only
Intercept	0.958	1.089	-0.230
miR-199a-3p_let-7a-5p	0.173	0.157	-
miR-26b-5p_miR-142-5p	-0.103	-0.164	-
miR-101-3p_miR-19b-3p	-0.061	-0.077	-
miR-93-5p_miR-19b-3p	0.442	0.447	-
miR-21-5p_miR-23a-3p	0.0002	0.018	-
let-7b-5p_miR-19b-3p	-0.195	-0.184	-
let-7a-5p_miR-22-3p	-0.034	-0.034	
Breast density (Tabar)	0.304	-	0.398
BMI*Menopause	0.236	-	0.410
WCRF lifestyle score	-0.156	-	-0.141

Since the same cohort was used as in small-RNA sequencing, we performed an apparent validation (i.e., applying the coefficient to the original dataset) to assess the previously mentioned parameters (i.e., calibration, ROC AUC, Brier score, etc.). The ROC curves, calibration assessment and scaled Brier score can be seen in **Figure 59**.

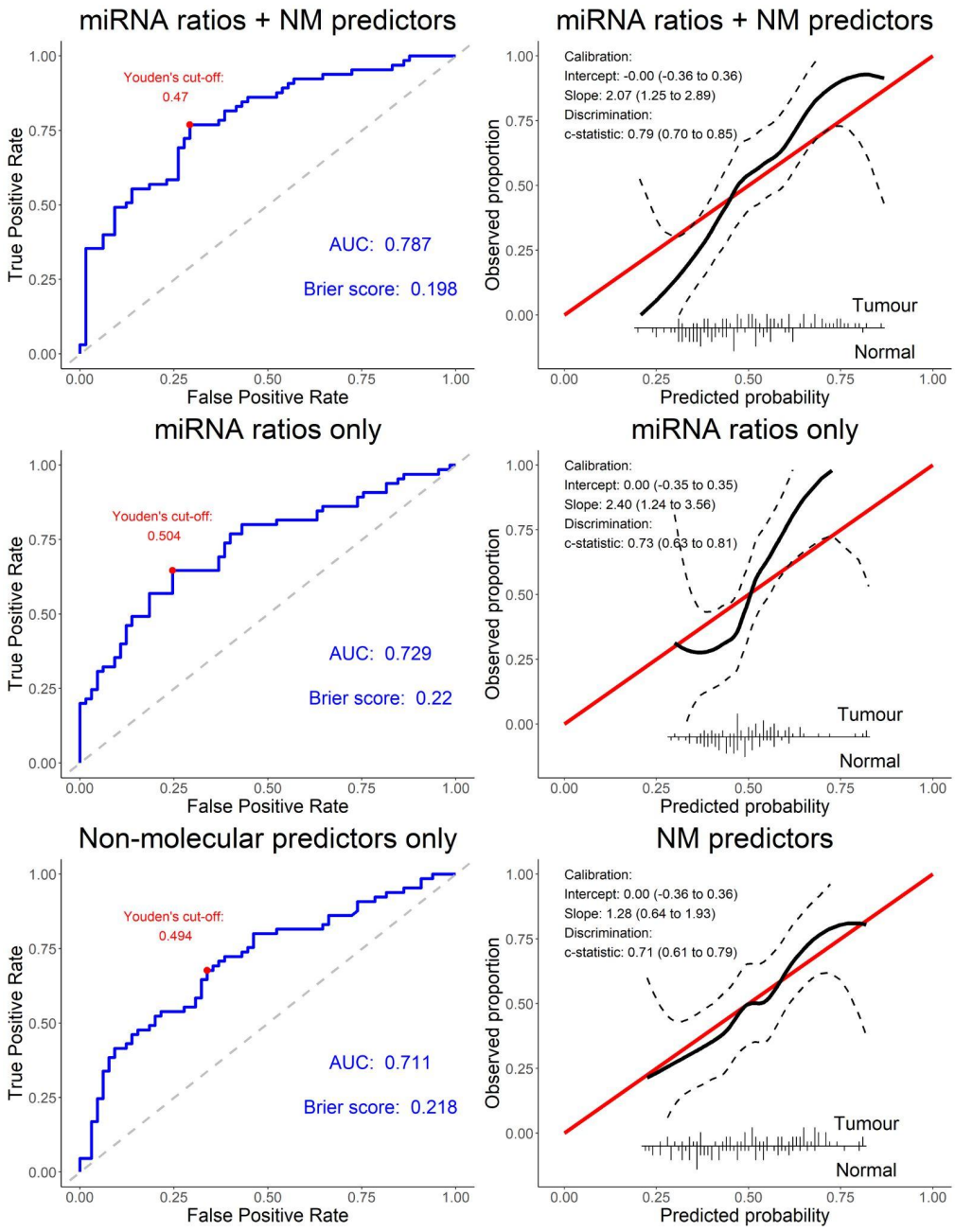


Figure 59. ROC AUC and calibration plots for the three LASSO logistic regression models are shown. Within the ROC AUCs the Youden's cut-off, AUC and Brier score are reported, while within the calibration plots, the intercept and slope of the calibration curve are reported.

The best performing model was the model on miRNA ratios together with non-molecular variables. Overall, the models created on RT-qPCR data were not optimally calibrated, with all predictions being slightly underestimated. Notably, the intercept of the calibration plot was usually on the optimal 0 point, whereas the slope was suboptimal. Finally, based on the paired DeLong test, the model on miRNA ratios together with non-molecular variables had a significantly better

AUC than non-molecular variables alone (**Table 19**), demonstrating the biomarker potential of the found miRNA ratios.

Table 19. DeLong test comparing the AUCs of the three LASSO logistic regression models.

Comparison	Z	p-value
miRNA only vs miRNA + NM	-1.44	0.150
miRNA only vs NM only	0.29	0.774
miRNA + NM vs NM only	2.80	0.005

Five of the seven miRNA ratios in the final model had significant associations with clinicopathological characteristics based on the RT-qPCR data (**Figure 60**). Namely, miR-93-5p_miR-19b-3p was lower in ER+ compared to ER− invasive BC patients ($p = 0.037$). miR-26b-5p_miR-142-5p was lower in Ki-67+ compared to Ki-67− invasive BCs ($p = 0.048$). Interestingly, miR-21-5p_miR-23a-3p was higher in ER+ than in ER− invasive BC patients ($p = 0.030$), in PgR+ versus PgR− ($p = 0.036$) as well as in Ki-67− in contrast to Ki-67+ BC invasive patients ($p = 0.033$). Lastly, let-7a-5p_miR-22-3p was lower in ductal compared to other BC histotypes ($p = 0.050$).

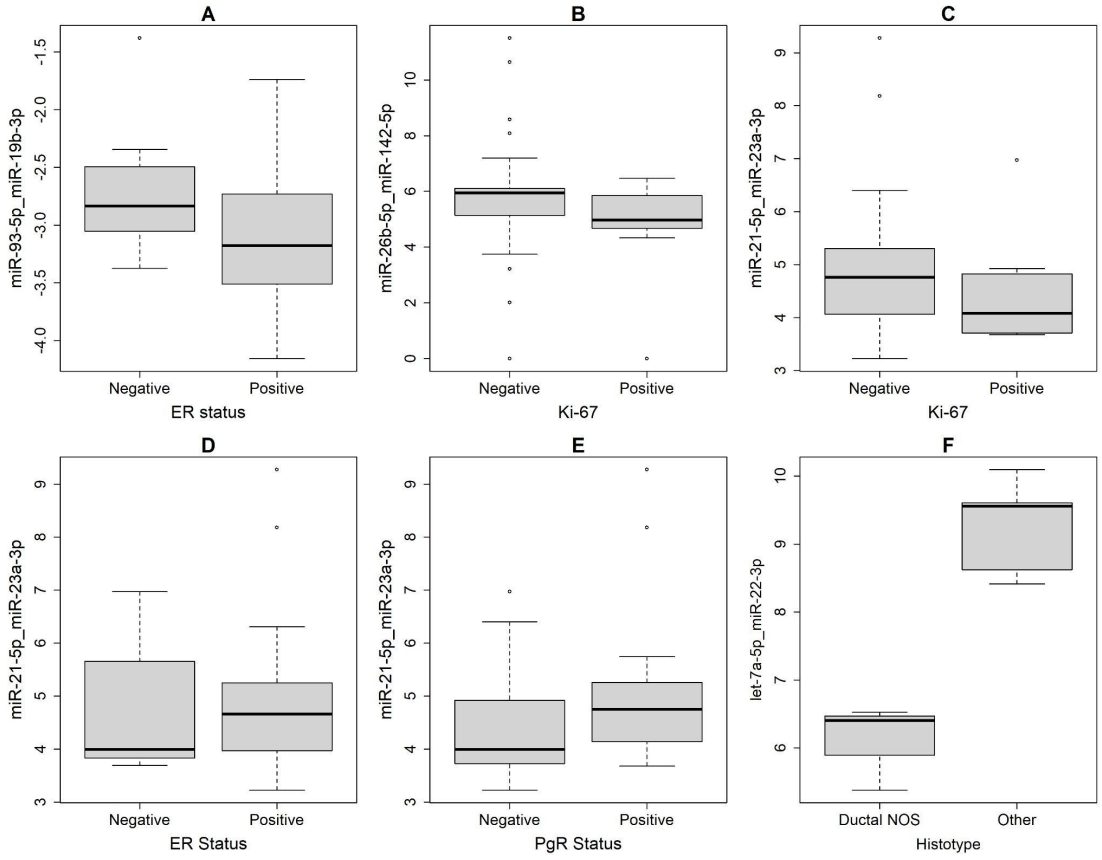


Figure 60. Expression values for ratios associated with clinicopathological BC cases characteristics. Panels A-E refer to invasive tumours whereas panel F to in situ ones.

We then evaluated the ratios selected by the hierarchical shrinkage models, which could be assembled using the miRNAs assayed by RT-qPCR (one of the ratios was the one found in the reported frequentist approach: let-7a-5p_miR-106b-5p). These were the ratios selected from strategy 1 and two of the ratios selected from strategy 2. Out of the selected three, only one ratio could be assembled based on a hierarchical shrinkage model on all 4656 ratios. This was the ratio let-7f-5p_miR-22-3p and was included in the strategy 1 selection. We performed a univariate logistic regression on each of the five ratios and the results are shown in **Table 20**.

Table 20. Univariate logistic regression results of the five ratios selected by hierarchical shrinkage models on discovery cohort RT-qPCR data.

miRNA ratio	Cases		Controls		Univariate LR		
	Median	IQ* range	Median	IQ range	OR	95% CI	P
let-7a-5p_miR-106b-5p	-0.13	[-0.67, 0.65]	-0.36	[-0.36, 0.76]	0.93	[0.70, 1.20]	0.561
miR-15b-5p_miR-122-5p	0.41	[-0.53, 2.15]	-0.66	[-0.66, 1.42]	1.06	[0.87, 1.28]	0.586
miR-26b-5p_miR-19b-3p	-2.17	[-2.50, -1.80]	-2.45	[-2.45, -1.50]	0.71	[0.48, 0.97]	0.053
miR-122-5p_miR-21-5p	-2.93	[-4.09, -1.60]	-3.50	[-3.50, -1.80]	0.95	[0.76, 1.18]	0.627
let-7f-5p_miR-22-3p	4.95	[1.68, 7.06]	2.20	[2.20, 7.76]	0.91	[0.81, 1.01]	0.071

*Interquartile

None of the statistical tests yielded significant results; however, two ratios were close to having a significant odds ratio. These ratios were miR-26b-5p_miR-19b-3p and let-7f-5p_miR-22-3p with OR of 0.71 [0.48 to 0.97] and 0.91 [0.81 to 1], respectively. Next, the ridge logistic regression was created on all five ratios combined from the two lists of selected ratios. An ROC AUC of 0.607 was obtained, which is inferior to the AUC obtained from the miRNA ratios yielded by the frequentist approach. It is important to stress that ratios included in this approach were not all selected using the same hierarchical model and that, overall, there were fewer predictors than in the frequentist approach. Thus, there is a lower probability of overfitting and less optimistic results, as well as a higher chance that miRNA ratios are correlated, reducing their overall discriminating ability. The ROC curve and the calibration plots of the model can be seen in **Figure 61**.

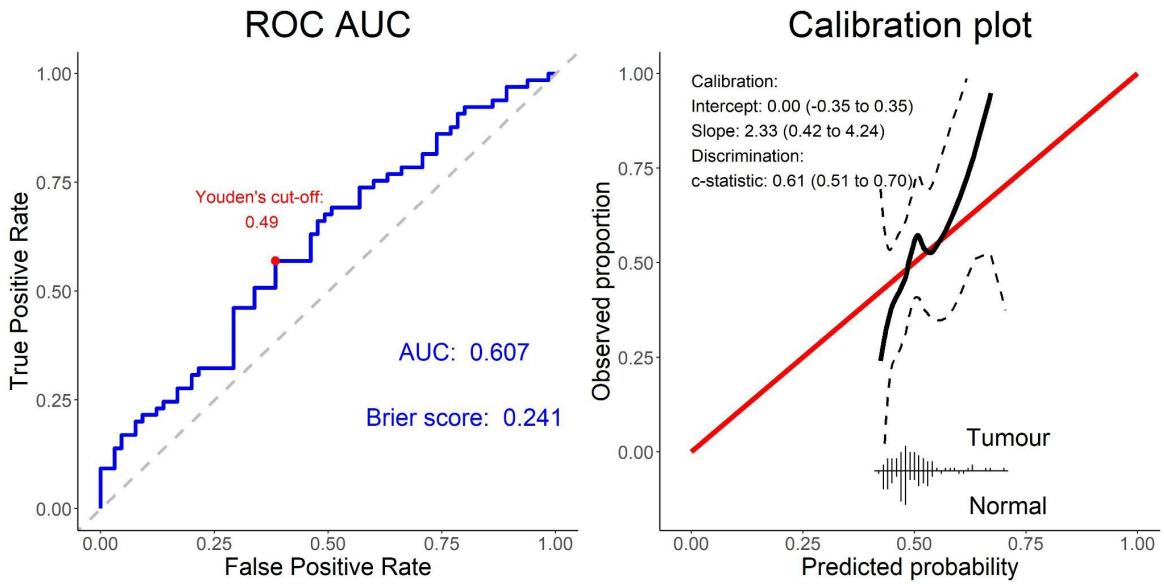


Figure 61. ROC AUC and calibration plot of the five combined ratios obtained using hierarchical shrinkage modelling. Within the ROC AUCs the Youden's cut-off, AUC and Brier score are reported while within the calibration plot, the intercept and slope of the calibration curve are reported.

Target enrichment and network analysis

The functional target enrichment analysis was performed on the miRNAs comprising the 7-ratio signature. Due to the software limitation of the possible number of miRNAs in a single functional enrichment analysis, we excluded let-7b-5p as it has a very similar mature sequence and function to let-7a-5p, which was included in the analysis. Based on the Wikipathways database, functional enrichment on the ten experimentally validated miRNA targets revealed their general involvement in cancer and breast cancer pathways, PI3K/AKT signalling pathway as well as the ATM-dependent DNA damage response. Additionally, they were involved in AR signalling and EGF/EGFR signalling pathways (**Figure 62**).

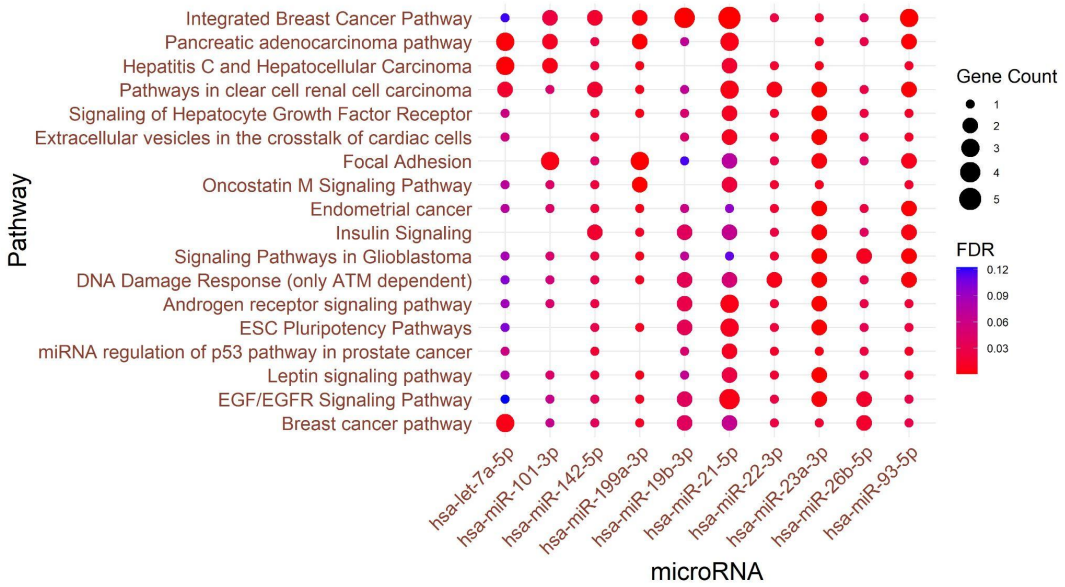


Figure 62. Wikipathways database enrichment results for the experimentally validated targets of the ten miRNAs making up the 7-miRNA ratio signature.

The KEGG database showed generally concordant results to the Wikipathways database (**Figure 63**). Interestingly, enrichment in cellular senescence was also observed.

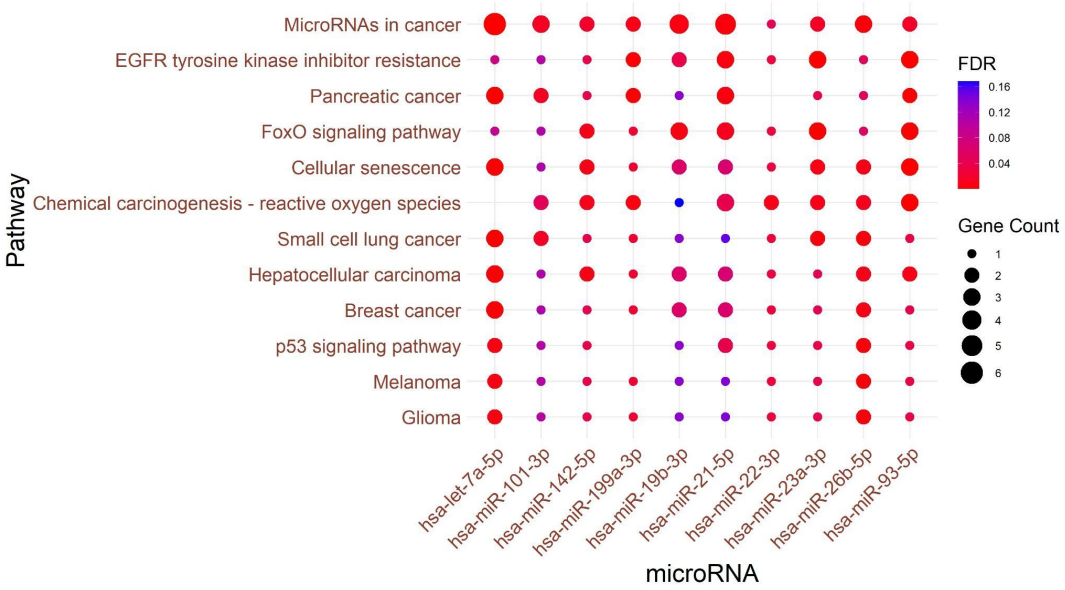


Figure 63. KEGG database enrichment results for the experimentally validated targets of the ten miRNAs making up the 7-miRNA ratio signature.

In the Reactome pathway database results (**Figure 64**), some of the notable pathways which were overrepresented in the majority of the experimentally validated targets of the ten miRNAs were cellular response to stress, Interleukin−4 and 13 signalling, PTEN regulation and deubiquitination.

Validated targets were also enriched in many other cancer diseases, indicating that these miRNAs could be pan-cancer biomarkers.

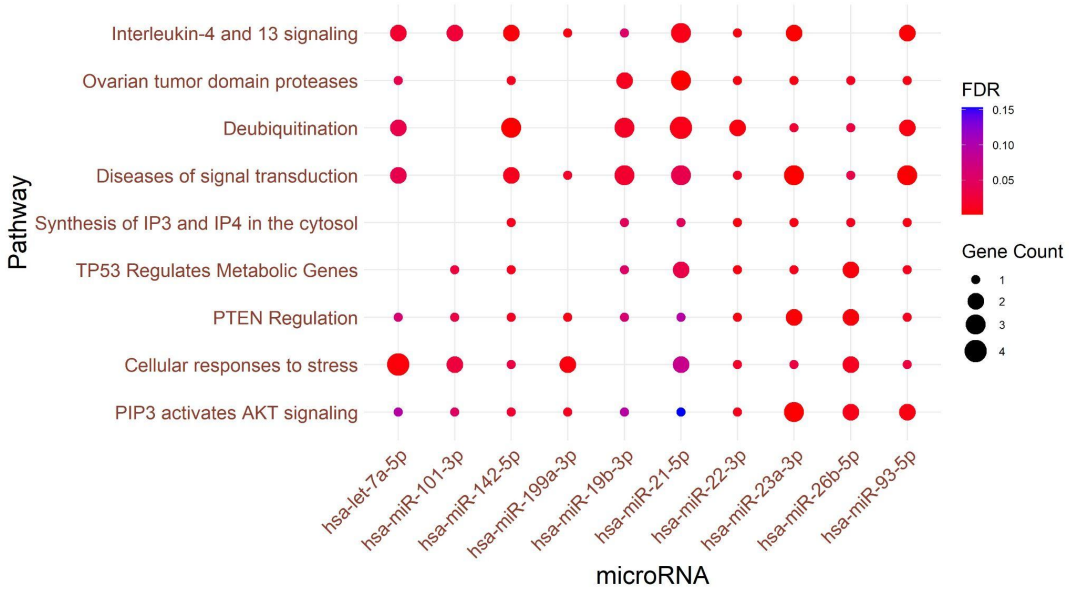


Figure 64. Reactome database enrichment results for the experimentally validated targets of the ten miRNAs making up the 7-miRNA ratio signature.

Finally, based on the Mienturnet software, messenger RNAs of 12 genes were commonly targeted by at least 5 of the 10 analysed miRNAs. The most targeted genes were the tumour suppressor phosphatase and tensin homolog (*PTEN*) (7 miRNAs) and Nuclear FMR1 Interacting Protein 2 (*NUFIP2*) (6 miRNAs).

We performed a network analysis using the MetaCore software on the 11 miRNAs comprising the seven miRNA ratios. The output of the software were small subnetworks containing relevant genes and miRNAs (those of interest and others if relevant to the pathway/network). The pathway maps of the involved genes in the networks, with an FDR lower than 0.05, among other processes, were the following: regulation of microRNAs in colorectal cancer, anti-inflammatory and cardioprotective adiponectin signalling as well as TGF-beta signalling via microRNAs in BC (**Figure 65**). A total of 30 transcription factors were found that interact with the ten miRNAs or their targeted genes (**Table 21**). The transcription factors with the highest number of interactions were RelA, EGR1, HIF1A, and p53.

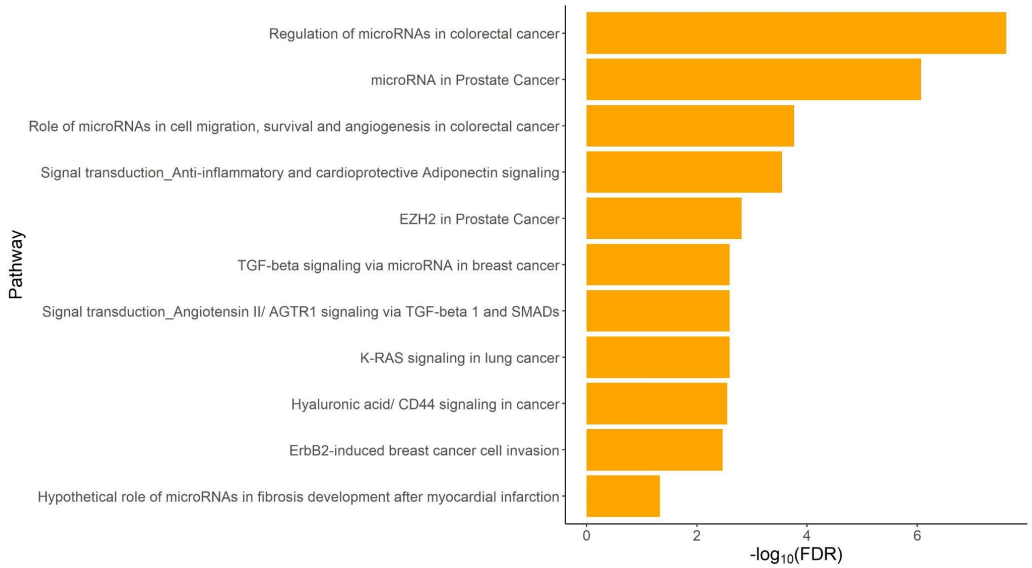


Figure 65. Pathway maps result of the 11 miRNAs making up the 7-miRNA ratio signature.

Table 21. Transcription factors with the highest number of interactions with the 11 miRNAs.

TF	Actual	R	Expected	Ratio	p-value	z-score
RelA	8	2373	1.83	4.37	< 0.001	4.68
EGR1	7	681	0.53	13.31	< 0.001	9.00
HIF1A	6	1225	0.95	6.34	< 0.001	5.27
p53	6	1935	1.50	4.02	0.003	3.77
EPAS1	5	291	0.23	22.25	< 0.001	10.11
c-Fos	5	496	0.38	13.05	< 0.001	7.50
c-Jun	5	1029	0.80	6.29	0.001	4.77
STAT3	5	1453	1.12	4.46	0.005	3.72
E2F3	4	240	0.19	21.58	< 0.001	8.89
SMAD4	4	446	0.35	11.61	< 0.001	6.26
STAT1	4	618	0.48	8.38	0.001	5.14
NRSF	4	622	0.48	8.33	0.001	5.12
KLF4	4	15040	11.62	0.34	0.003	-2.74
MYOD	3	226	0.18	17.19	< 0.001	6.78
TWIST1	3	382	0.30	10.17	0.003	5.00
NF-kB1	3	507	0.39	7.66	0.007	4.19
C/EBPalpha	3	609	0.47	6.38	0.011	3.72
GATA-1	3	12317	9.51	0.32	0.007	-2.48
NF-AT4	2	67	0.05	38.65	0.001	8.57
CREM	2	82	0.06	31.58	0.002	7.71
GFI-1	2	102	0.08	25.39	0.003	6.86
ETS2	2	153	0.12	16.92	0.006	5.49
Max	2	171	0.13	15.14	0.008	5.15
MITF	2	215	0.17	12.04	0.012	4.51
GATA-2	2	10400	8.03	0.25	0.007	-2.43
NANOG	2	12923	9.98	0.20	< 0.001	-2.99
MTA1	1	10	0.01	129.50	0.008	11.30
GCR	1	7912	6.11	0.16	0.010	-2.28

TF	Actual	R	Expected	Ratio	p-value	z-score
ETS1	1	8154	6.30	0.16	0.008	-2.33
HNF3-alpha	1	9363	7.23	0.14	0.003	-2.60

Actual: number of network objects in the activated dataset(s) which interact with the chosen object
R: number of network objects in the complete database or background list which interact with the chosen object
Expected: mean value for hypergeometric distribution (n*R/N); N in this case represents total number of gene-based objects in the complete database or background list (45315)
Ratio: connectivity ratio (Actual/Expected)

Seven networks were created from the 11 miRNAs with a z-score larger than 60 (**Table 22**).

Table 22. Results of the network analysis on 11 miRNAs making up the 7-miRNA ratio signature. Shown are the miRNAs included in each subnetwork, the associated GO processes as well as the network statistics.

Network	p-value	Z	g-score
miR-21-5p, miR-26b-5p, miR-23a-3p, microRNA 21, miR-21-3p	< 0.001	202.34	202.34
miR-93-5p, microRNA let-7a-1, miR-142-5p, microRNA let-7b, miR-23a-3p	< 0.001	184.89	184.89
miR-23a-3p, miR-93-5p, microRNA 21, miR-26b-5p, miR-let-7b-5p	< 0.001	158.91	158.91
miR-22-3p, miR-23a-3p, miR-142-5p, miR-93-5p, miR-let-7b-5p	< 0.001	140.06	140.06
miR-let-7a-2-3p, miR-93-5p, miR-21-3p, microRNA 21, miR-142-5p	< 0.001	78.59	78.59
microRNA let-7°-1, microRNA 23°, microRNA 19b-1, miR-23°-5p, STAT1	< 0.001	68.02	68.02
microRNA 199°-1, miR-22-5p, microRNA 22, SP1, Mn(‘3+) + Apotransferrin = Mn(III)-Apotransferrin	< 0.001	48.68	48.68

The g-score is a statistic modifying the Z-score based on the number of linear canonical pathway units within the network.

In network 1 (**Figure 66**), *PTEN* was the central gene and was inhibited by several miRNAs found among the seven miRNA ratios. Additionally, a gene which interacts with *PTEN*, neuron-restrictive silencer factor (*NRSF*), more commonly known as RE1 Silencing Transcription Factor (*REST*), was found to inhibit miR-199a and miR-93. In the second network, Sirtuin 1 (*SIRT1*) was the central gene which was inhibited by miR-23a-3p, miR-142, miR-22 and miR-93. In addition to *SIRT1*, miR-93 also inhibited the Estrogen Receptor 1 (*ESR1*) gene. *SIRT1* might be relevant to epigenetic gene silencing and promotes the formation of breast cancer through modulating Akt activity. Cyclin Dependent Kinase 4 (*CDK4*) was found to be inhibited by two miRNAs and a miRNA which also inhibited *SIRT1*, while *ERK2* was found to activate miR-101 and was functionally associated with let-7a and miR-26b.

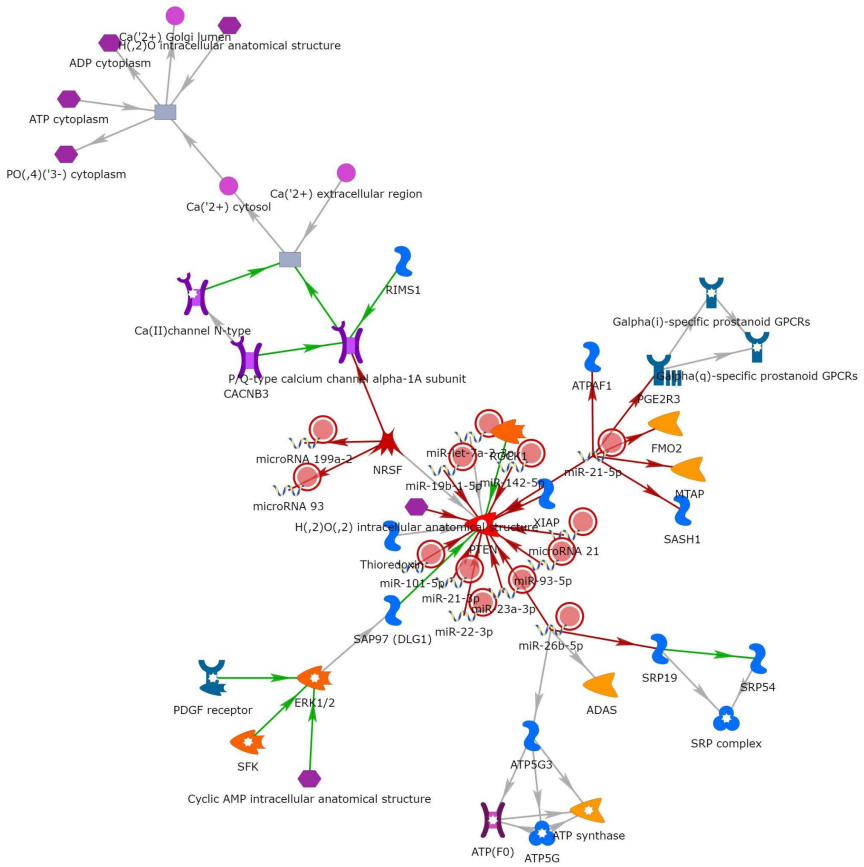


Figure 66. Graphical representation of network 1 from **Table 22**. The miRNAs with a red circle next to them were the input miRNAs.

Network 3 can be divided into three hubs: the protein kinase *JAK1*, transcription factor *BLIMP1* and receptor ligand CTGF (**Figure 67**). JAK1 was inhibited by miR-93, which also inhibited Nuclear Receptor Coactivator 3 (*NCOA3*). Additionally, *JAK1* was inhibited by miR-23a, which inhibited the *BLIMP1* transcription factor. *BLIMP1* was also inhibited by four other miRNAs of interest (miR-21, let-7b, let-7a and miR-22. The connection of the second and third hubs was reflected in the activation of both *BLIMP1* and CTGF by the transcription factor *SP1*. The CTGF receptor ligand was found to be inhibited by miR-26b, miR-19, miR-21 and miR-19b.

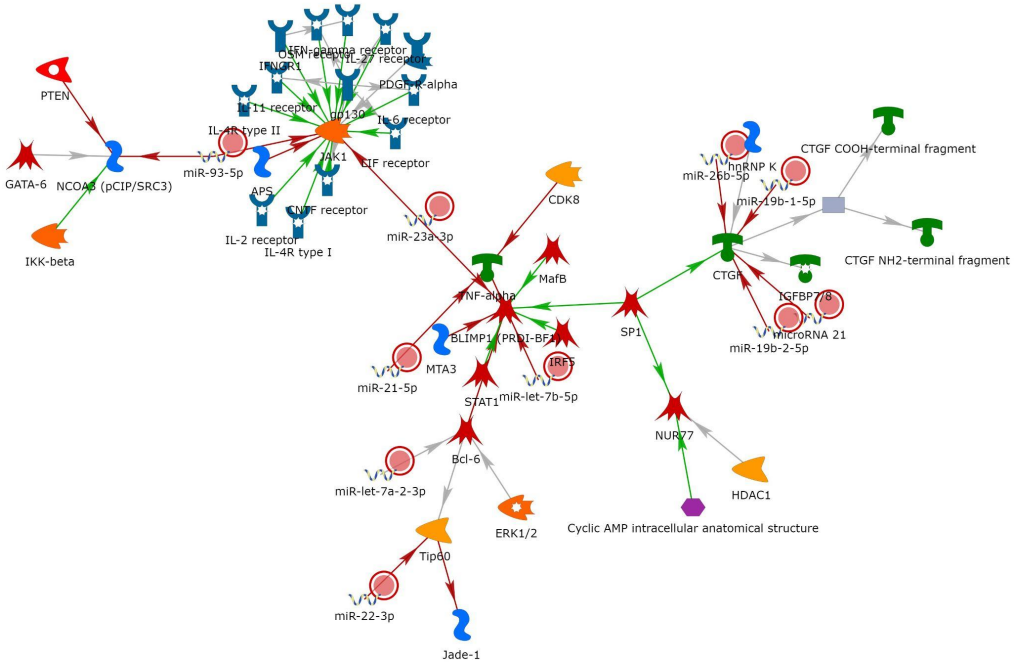


Figure 67. Graphical representation of network 3 from **Table 22**. The miRNAs with a red circle next to them were the input miRNAs.

Model application in the validation cohort

The miRNAs making up the seven miRNA ratios selected by the penalised LASSO logistic regression in the discovery cohort were assessed in the previously described validation cohort with 127 controls and 32 cases. The key differences between the discovery and validation cohorts are the inclusion of controls which went for a second-level investigation after the mammography, and the inclusion of BC cases which were diagnosed several months after blood sampling. Importantly, as both of them are found among the seven miRNA ratios and their Cts were highly correlated ($\rho = 0.96$), let-7b-5p was replaced by let-7a-5p in the one ratio it was a part of. Another reason for this decision was that the mature sequences of the two miRNAs are very similar, with only two nucleotides being different, and, in both miRNAs, the differing bases were purines. This enabled an overall much more cost-efficient RT-qPCR run. The flowchart of the biomarker analysis in the validation cohort can be seen in **Figure 68**.

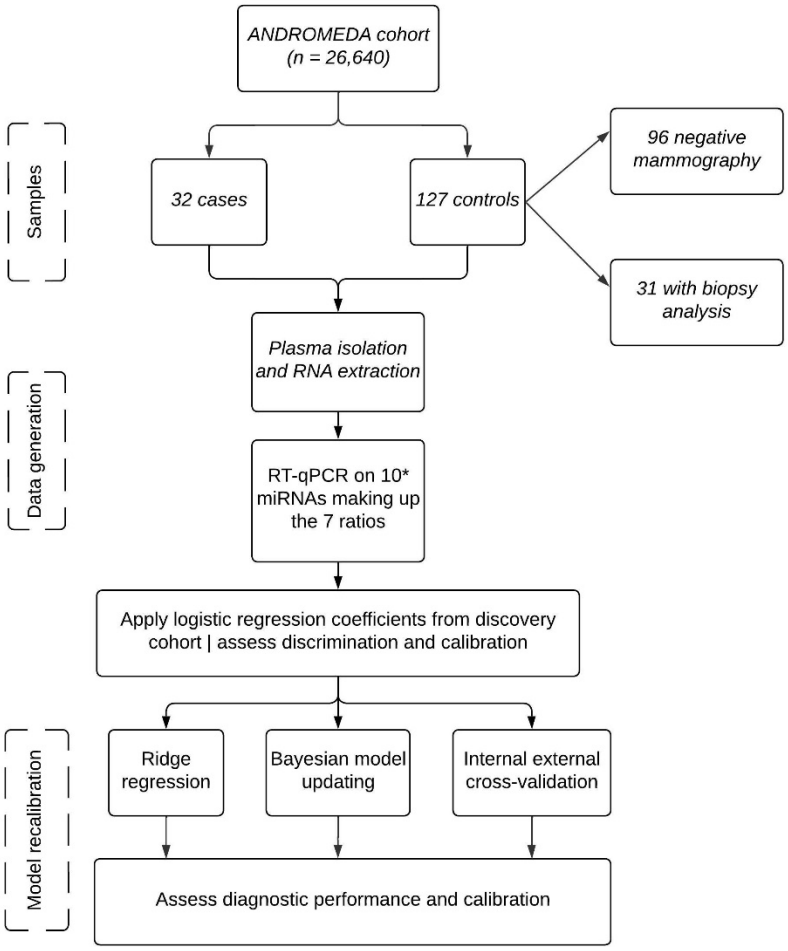


Figure 68. Validation cohort pipeline in which we assessed the discrimination power and predicted risk calibration of the miRNA ratio signature. *let-7b-5p was replaced by let-7a-5p in the ratios due to the high correlation of Cts between the two miRNAs within the discovery cohort.

The mean Ct and SD for each of the ten miRNAs across the RT-qPCR duplicates, stratified by BC status, can be seen in **Figure 69**. Almost all miRNAs had high-quality Ct values from RT-qPCR, as they were within the expected Ct value range and had generally small SD across replicates. Like in the discovery cohort, miR-22-3p was again the only exception as its SD was higher, and it had several samples with the Ct approaching 40. Interestingly, although not normalised, visually the Ct values tend to be lower in cases when compared to controls, indicating a tendency of higher expression of most analysed miRNAs in cases.

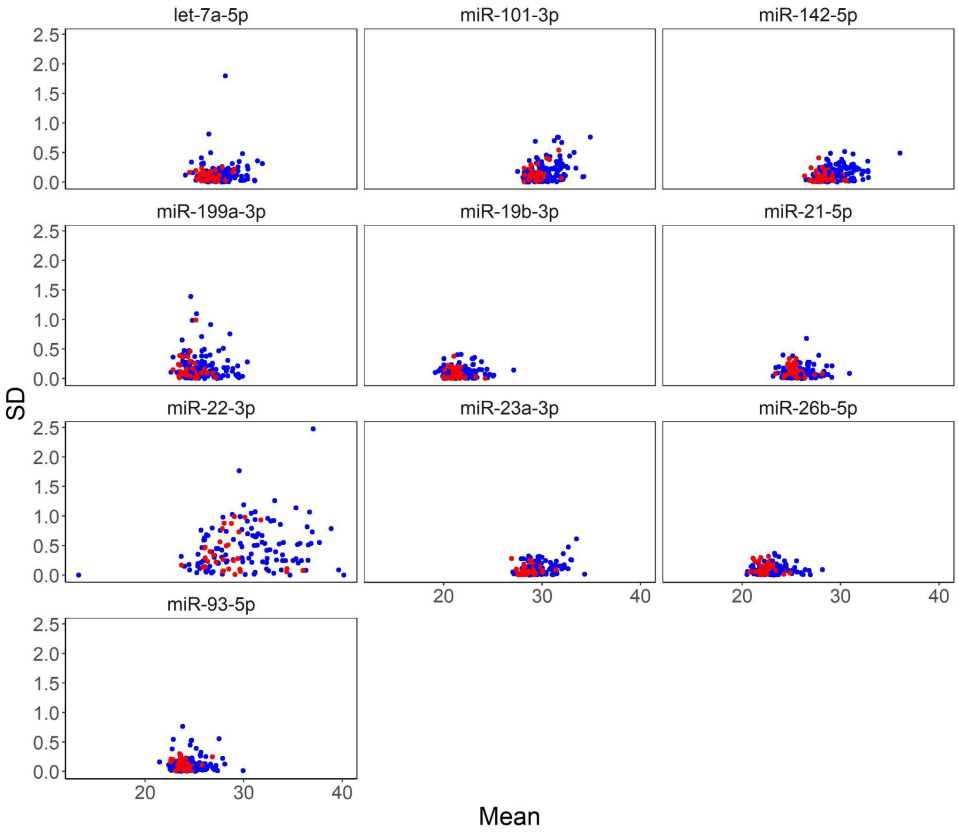


Figure 69. Mean and SD of Cts for each of the ten miRNAs assayed by RT-qPCR in the validation cohort. The red points represent BC cases, while blue points represent controls.

The mean, SD and coefficient of variation were calculated for each of the seven miRNA ratios analysed in the validation cohort. The density plots of mean and CV can be seen in **Figure 70** (CV was plotted instead of SD due to a more stable distribution of CV in this particular data).

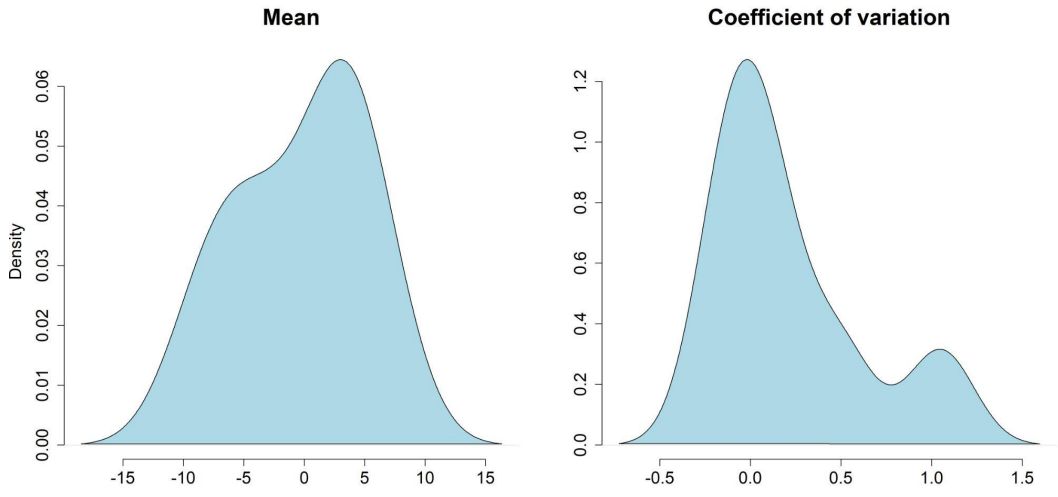


Figure 70. Density plots of mean and CV of the seven miRNA ratios.

Importantly, the value ranges of the individual miRNA ratios, as well as the other mentioned statistics (i.e., overall mean and SD) were comparable between the RT-qPCR data in the discovery and validation cohort. Additionally, four of the seven miRNA ratios were differentially expressed ($p < 0.05$) between BC cases and controls (**Figure 71**).

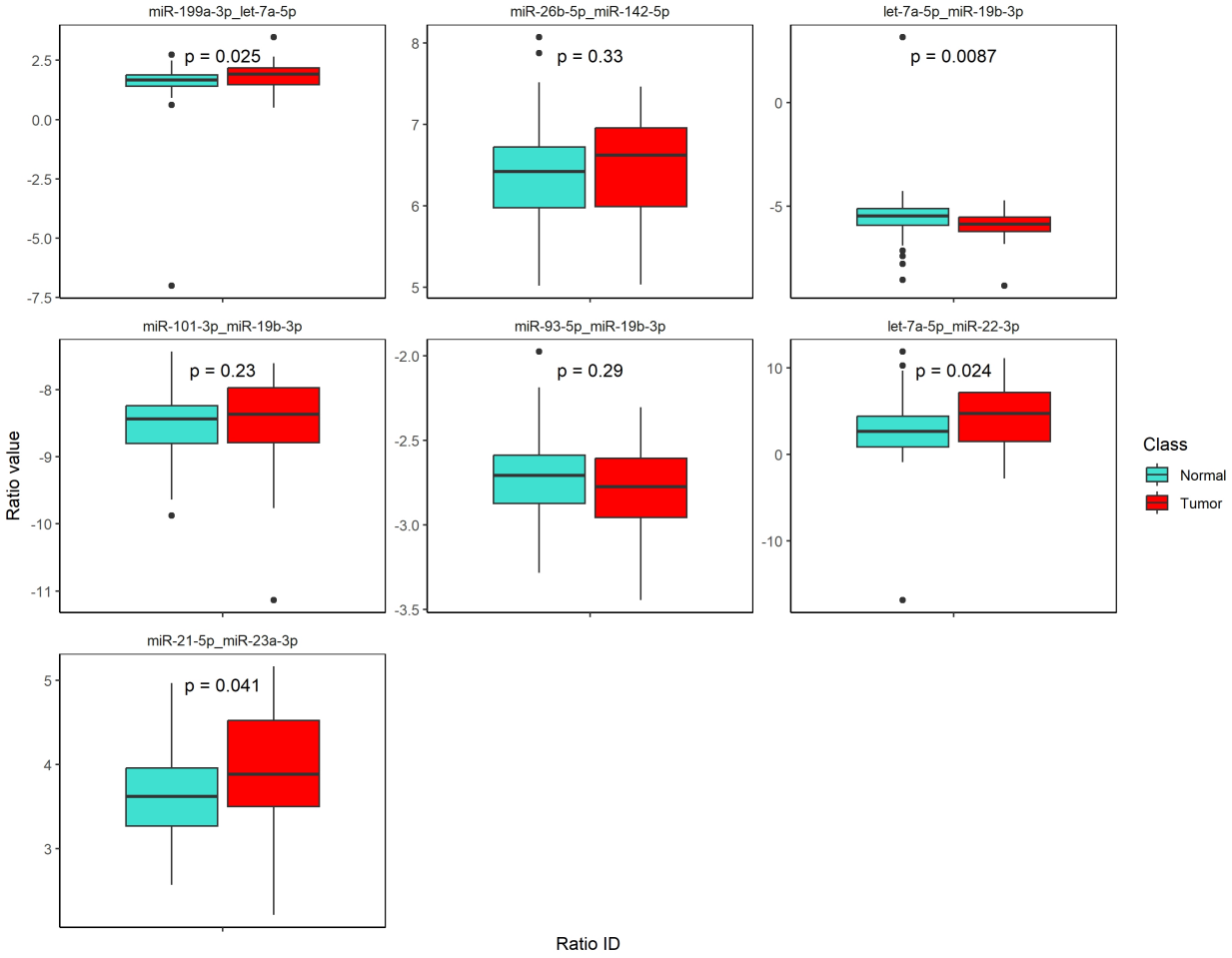


Figure 71. Boxplot of seven miRNA ratios computed in the validation cohort stratified by BC status.

We then created a heatmap and correlation plot of the mentioned ratios to determine how the miRNA ratio expression clusters and if any of the ratios are correlated (Pearson correlation). Eleven pairs of ratios were significantly correlated (**Figure 72**). It is important to stress that the sample clustering in the heatmap was influenced by the different proportions of cases and controls within the validation cohort.

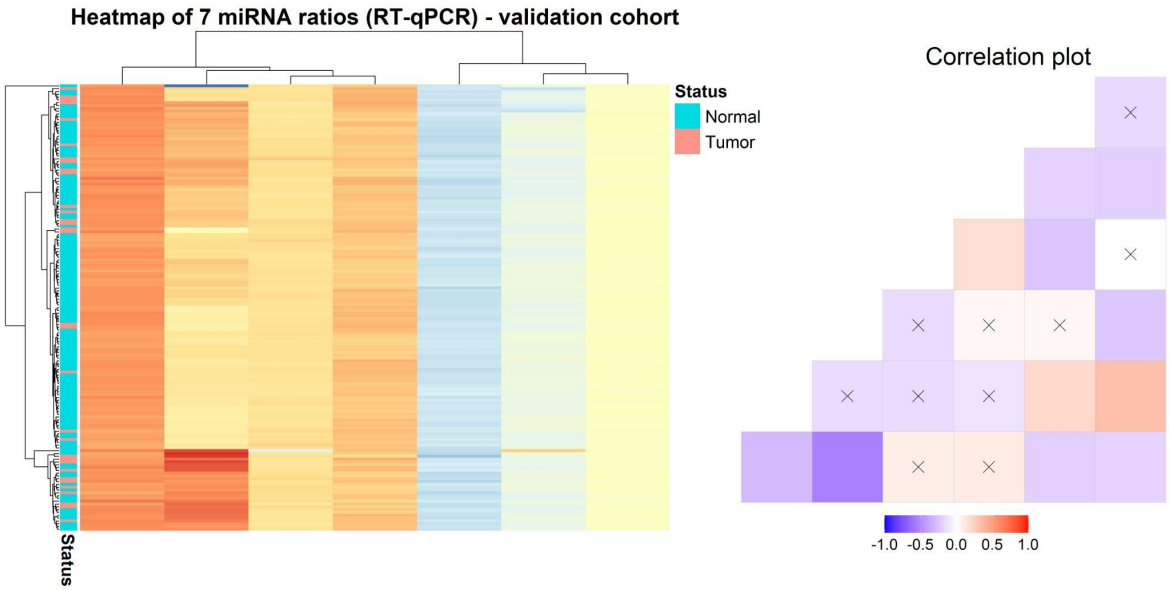


Figure 72. Heatmap of the seven miRNA ratios analysed in the validation cohort (left) and their correlation plot (right) where the squares without x refer to significantly positively (red) or negatively (blue) correlated pairs.

We performed univariate logistic regression on the individual seven ratios, and four of them were significantly associated with BC ($p < 0.05$) (*Supplementary Table 11* – Appendix B). From the significantly associated ratios, let-7a-5p_miR-22-3p had a discordant OR compared to the discovery cohort; in the validation cohort the OR was above 1, while in the discovery, it was < 1. All non-significantly associated ratios had discordant ORs between the discovery and validation set.

The multivariate model from the discovery cohort included seven miRNA ratios and three non-molecular variables (breast density, interaction of centred BMI and menopausal status and WCRF lifestyle score). The coefficients of the variables obtained from the discovery cohort were applied to the validation cohort using the sigmoid function for logistic regression. After applying the coefficients, we obtained 0.71 [0.61 to 0.80] ROC AUC. The relatively poor performance could be attributed to the differences between the discovery and validation cohorts with respect to time to diagnosis in cases or differences between controls as a subset of controls in the validation cohort underwent biopsy due to suspicion of a positive diagnosis. However, even after applying the coefficients to subgroups of cases (depending on their time of diagnosis) or subgroups of controls, we still obtained suboptimal prediction results without any significant improvement compared to the model on all samples (**Table 23**).

Table 23. ROC AUCs and their confidence intervals on models including all predictors, only miRNA ratios and only non-molecular predictors. Results for various subgroups are also shown.

	All predictors		miRNA ratios		Non-molecular predictors	
Sample subgroup	**AUC**	**95% CI**	**AUC**	**95% CI**	**AUC**	**95% CI**
All samples	0.71	[0.61, 0.80]	0.51	[0.39, 0.62]	0.74	[0.64, 0.82]
Without controls with additional biopsy	0.73	[0.62, 0.81]	0.52	[0.40, 0.65]	0.75	[0.64, 0.83]
Without controls with negative mammography result	0.66	[0.52, 0.78]	0.45	[0.31, 0.60]	0.72	[0.58, 0.83]
Without cases diagnosed more than 2 years after blood sampling	0.71	[0.58, 0.81]	0.55	[0.38, 0.71]	0.70	[0.56, 0.81]
Without cases diagnosed less than 2 years after blood sampling	0.72	[0.56, 0.83]	0.47	[0.31, 0.63]	0.77	[0.63, 0.87]

Notably, the time passed after blood sampling until diagnosis was not associated with the predicted risk (**Figure 73**).

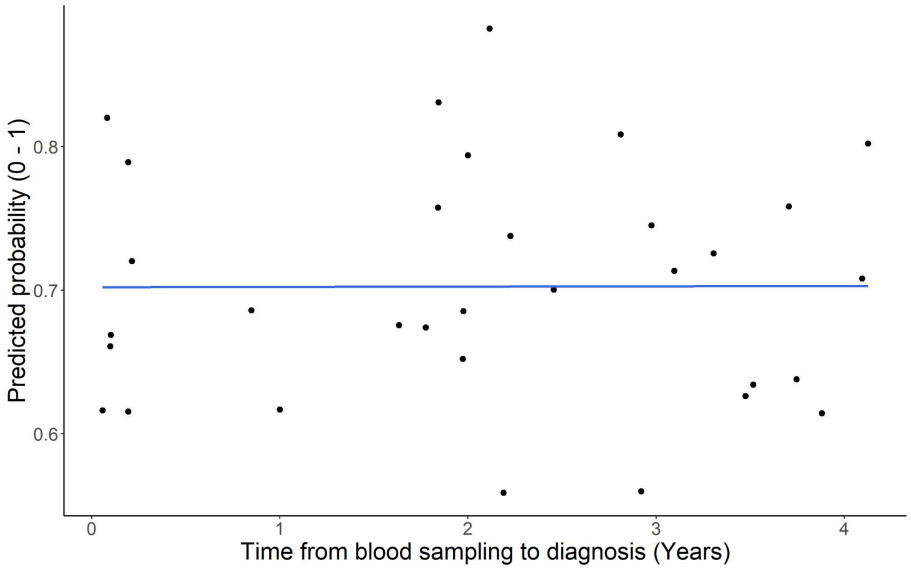


Figure 73. Scatter plot of time from blood sampling to diagnosis and predicted probability after applying the coefficients to the validation cohort (based on the seven miRNA ratios and non-molecular variables).

Finally, we ordered the 159 samples based on predicted risk and investigated the distribution of true positives, false positives, true negatives and false negatives at Youden's optimal cut-off (0.614) (**Figure 74**). A substantial miscalibration of the predicted probabilities and a large proportion of false positive classifications was observed.

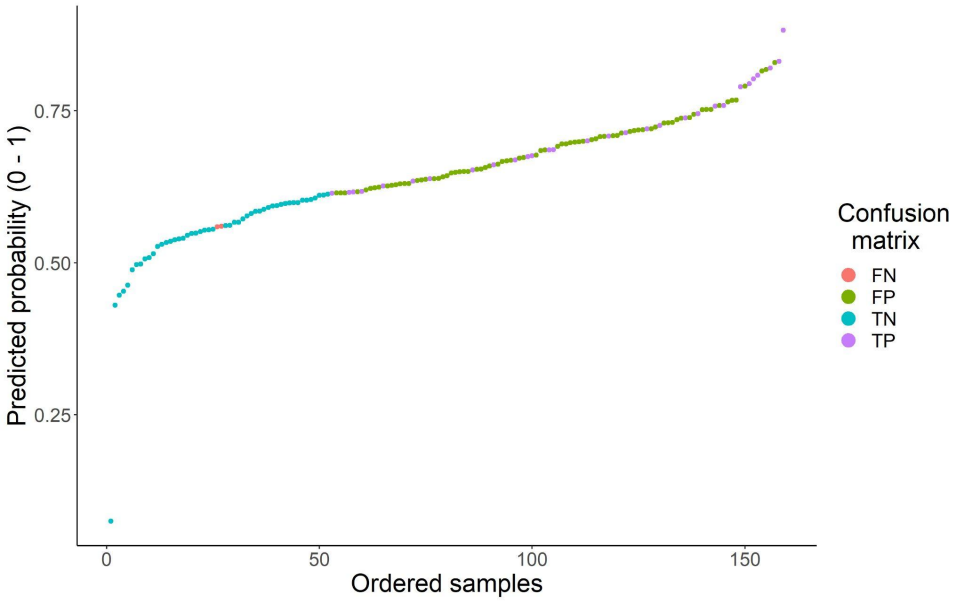


Figure 74. Validation cohort samples ordered by the predicted probability of being BC positive (based on the model combining miRNA ratios and non-molecular predictors). The samples were then classified into predicted case or control based on Youden's cut-off and the resulting prediction was labelled as TP, FP, TN or FN.

We then tested whether the distributions and variances of the predictors in the mentioned model differ significantly between the two healthy control subgroups. Out of the 12 tested predictors (we also included centred BMI and menopausal status as they made up the interaction term in the model), three miRNA ratios had significantly different variances (miR-199a-3p_let-7a-5p, let-7b-5p_miR-19b-3p, let-7a-5p_miR-22-3p) at the p-value < 0.01 cut-off. No other significant differences based on the three tests were found between the subgroups (*Supplementary Table 12* – Appendix B).

When applying the coefficients from the miRNA ratio only model on the validation cohort data, we obtained an ROC AUC of 0.51 [0.39 to 0.62], while for the non-molecular variables model, we obtained an ROC AUC of 0.74 [0.64 to 0.82], indicating that the non-molecular variables are more homogenous between the discovery and validation cohort (**Table 23**). For the models on miRNA ratios alone and on non-molecular variables alone, we also tested whether the time from blood collection until diagnosis was associated with the predicted risk (*Supplementary Figure 1* – Appendix B) and, as in the model with all predictors, no significant association was found. The distribution of true positives, false positives, true negatives and false negatives at Youden's optimal cut-off (0.656 for the model on miRNA ratios and 0.464 for the model on non-molecular predictors) for the two models were also investigated (*Supplementary Figure 2* – Appendix B).

An important aspect of every model is the calibration of its prediction estimates. Therefore, we generated a logistic calibration curve and investigated its intercept and slope. We only assessed

the calibration of the model with all samples and not the subgroups, as there was no evidence that any subgroups performed better than the complete cohort. The model on miRNA ratios and non-molecular variables applied to the validation cohort data was not calibrated, as seen in **Figure 75**. Consequently, there was a substantial overestimation of risk within the predictions (calibration plot metrics[1]: intercept was -2.45 [-2.84 to -2.06] and the slope was 1.31 [0.33 to 2.29]). The predicted risks of miRNA ratios only and non-molecular predictors only were also miscalibrated.

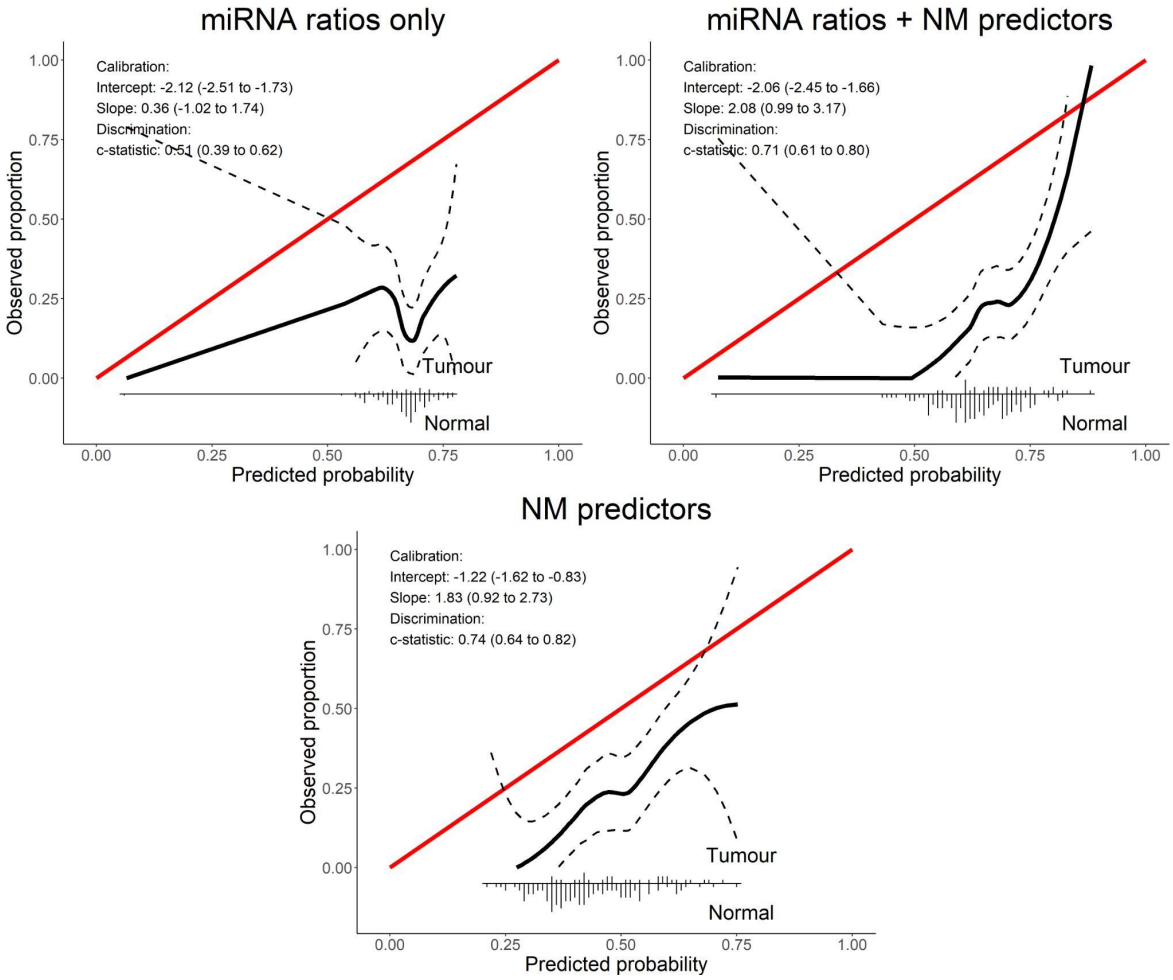


Figure 75. Calibration curve plots of the predicted probabilities of the three models (miRNA ratios together with non-molecular variables, miRNA ratios alone and non-molecular variables alone) applied to the validation cohort. The intercept and slope of the calibration curves are shown.

Since clear miscalibration was observed in all applied models, the next step was to recalibrate the predicted probabilities. To do so, we employed the closed testing method, which evaluated different recalibration approaches while striving to maintain the type I error: calibration of the intercept, calibration of intercept and the overall slope, or to have a complete model revision where

[1] All subsequent brackets of such format where we report the intercept and slope, refer to the respective calibration plots.

the coefficients and the intercept are re-evaluated. Based on the closed testing approach, we performed a complete model revision (re-estimated the coefficients and intercept) for all the mentioned models on miRNA ratios and non-molecular variables as well as miRNA ratios alone.

In order to perform model revision but avoid overfitting as much as possible, we used penalised ridge regression modelling. On the model on miRNA ratios together with non-molecular variables we obtained an ROC AUC of 0.90 [0.83 to 0.94] and a more calibrated model (intercept: 0.00 [-0.45 to 0.45] and slope: 1.43 [0.92 to 1.94]).

After performing ridge regression on the seven miRNA ratios alone, we obtained a more calibrated model (intercept: 0.00 [-0.41 to 0.41] and slope: 1.49 [0.88 to 2.11]) with an ROC AUC of 0.81 [0.72 to 0.87]. Finally, the ridge regression on non-molecular variables (WCRF lifestyle score, breast density and interaction term between centred BMI and menopause status) showed relatively good performance (ROC AUC = 0.78) and model calibration (intercept: 0.00 [-0.41 to 0.41] and slope: 1.23 [0.70 to 1.76]).

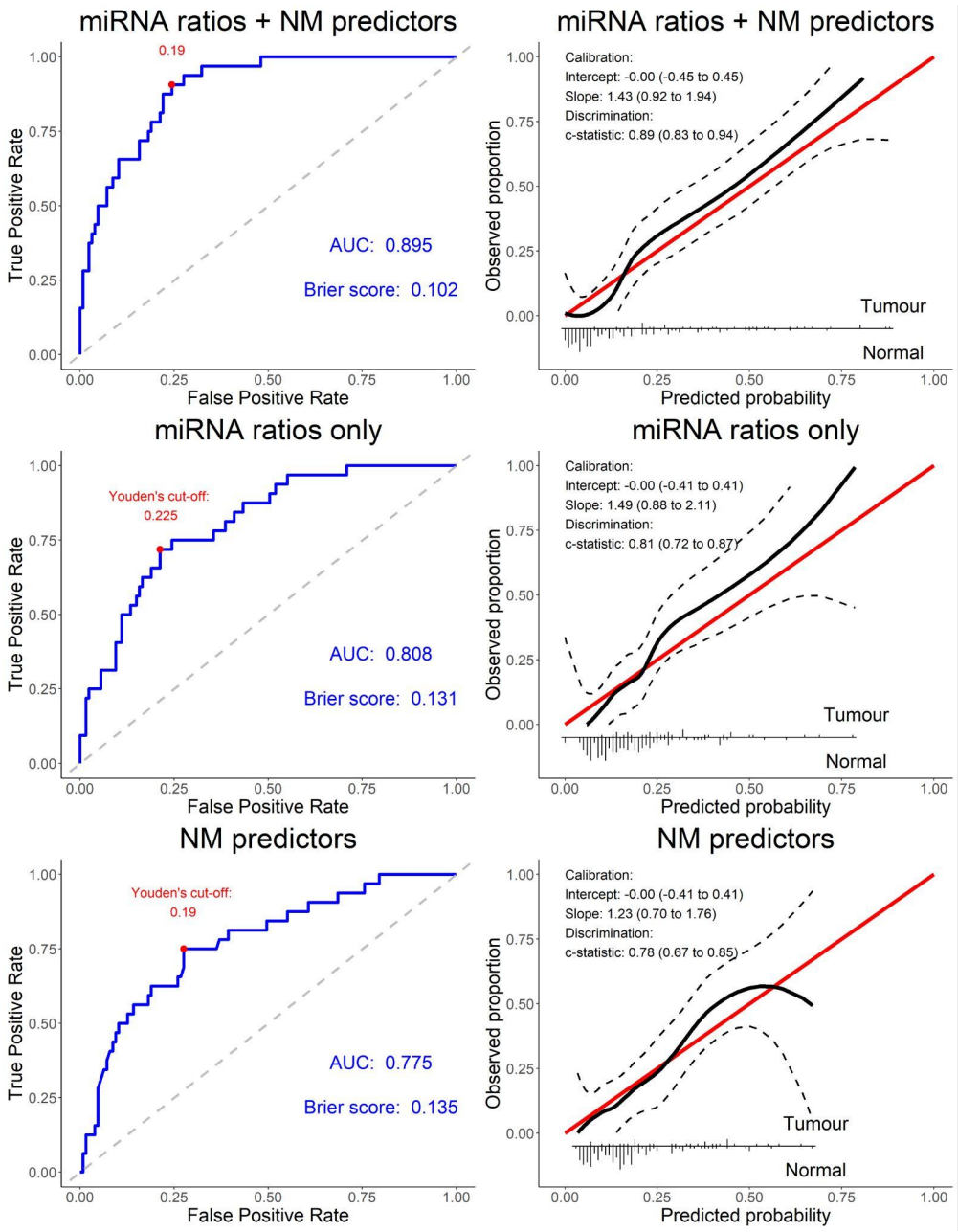


Figure 76. ROC AUC and calibration plots of the ridge regression models (model recalibration). Within the ROC AUCs, the Youden's cut-off, AUC and Brier score are reported, while within the calibration plots, the intercept and slope are reported.

After model revision, based on the DeLong test, the model with miRNA ratios and non-molecular variables performed significantly better than the other two models (**Table 24**). The new coefficients and intercepts of the three models can be seen in **Table 25**. Importantly, all revised coefficients in the validation cohort are concordant with the miRNA ratios in the discovery cohort except for let-7a-5p_miR-22-3p and miR-101-3p_miR-19b-3p.

Table 24. DeLong test on AUCs of the three recalibrated models in the validation cohort.

Comparison	Z	p-value
miRNA only vs miRNA + NM	-2.54	0.011
miRNA only vs NM only	0.49	0.625
miRNA + NM vs NM only	3.06	0.002

Table 25. Recalibrated coefficients in the validation cohort of the predictors included in the three models.

	All	miRNA only	NM only
Intercept	-11.637	-6.988	-3.140
miR-199a-3p_let-7a-5p	1.527	1.178	-
miR-26b-5p_miR-142-5p	-0.015	-0.029	-
miR-101-3p_miR-19b-3p	0.210	0.292	-
miR-93-5p_miR-19b-3p	0.191	-0.133	-
miR-21-5p_miR-23a-3p	1.157	0.912	-
let-7b-5p_miR-19b-3p	-0.409	-0.304	-
let-7a-5p_miR-22-3p	0.235	0.203	-
breast density	1.290	-	1.101
BMI*Menopause	0.677	-	0.566
WCRF lifestyle score	-0.043	-	-0.100

After model revision, the predicted probabilities were more evenly distributed in all three models, although still not fully optimal (in terms of underestimating the overall risk). This could be attributed to the sample size and lower event rate (**Figure 77**).

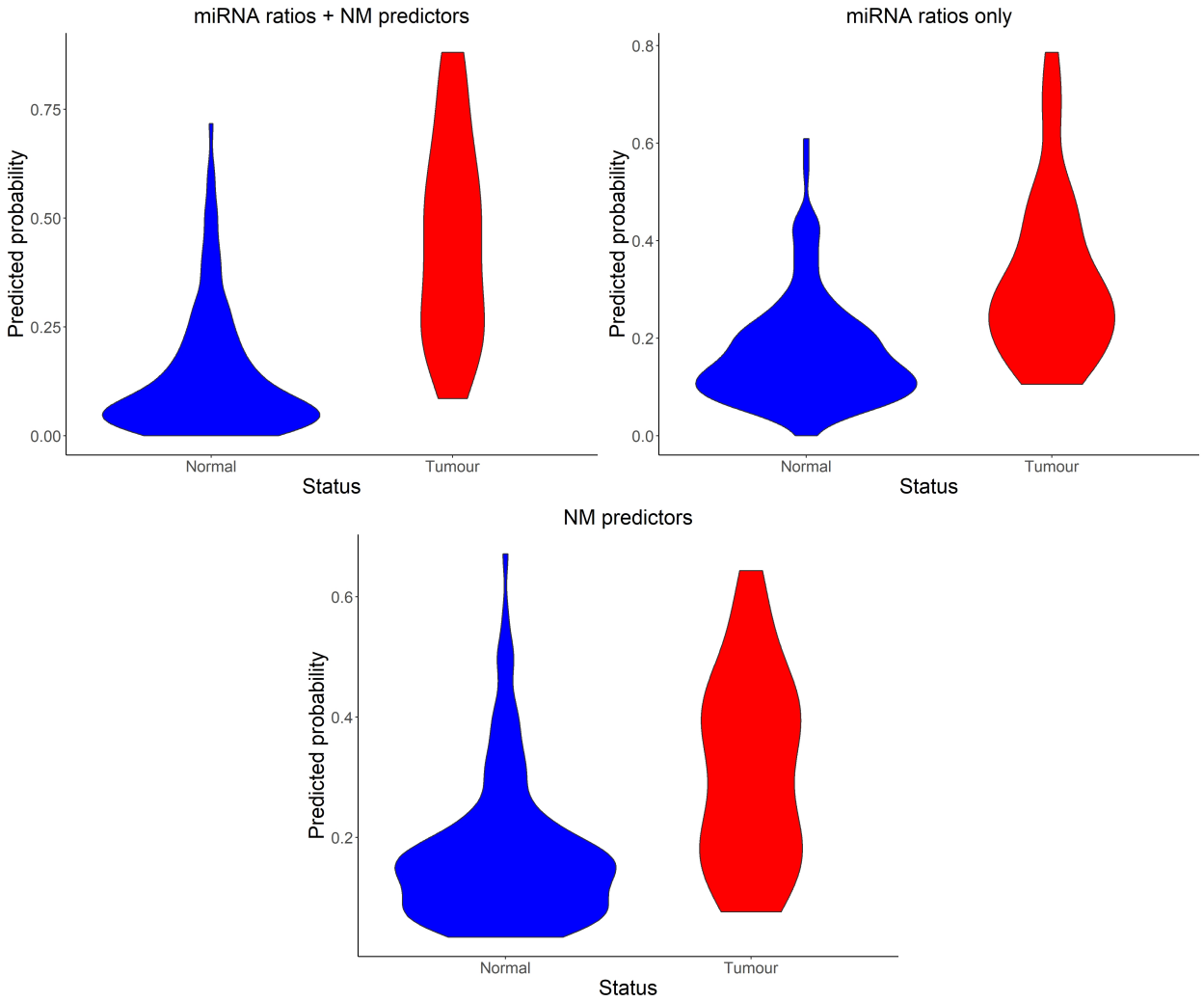


Figure 77. Violin plots of the calibrated predicted probabilities based on the three models.

Additionally, after calibrating the models, the time after blood sampling until diagnosis remained without significant association with the predicted risk (*Supplementary Figure 3* – Appendix B). The confusion matrix at Youden's optimal cut-off was labelled over ordered samples based on predicted probability (*Supplementary Figure 4* – Appendix B).

To additionally account for overfitting and overoptimism, we performed an ordinary bootstrap (n = 2000) on all the ridge regression models. The ROC AUC distributions of the models can be seen below. The 95% CI of the AUC based on the bootstrap is quite satisfactory for the model on miRNA ratios and non-molecular variables (0.785 to 0.903), indicating that these variables, when calibrated, could be useful biomarkers for early BC detection. Bootstrap results of this model as well as the miRNA ratio-only and non-molecular variables-only models can be seen in **Figure 78**.

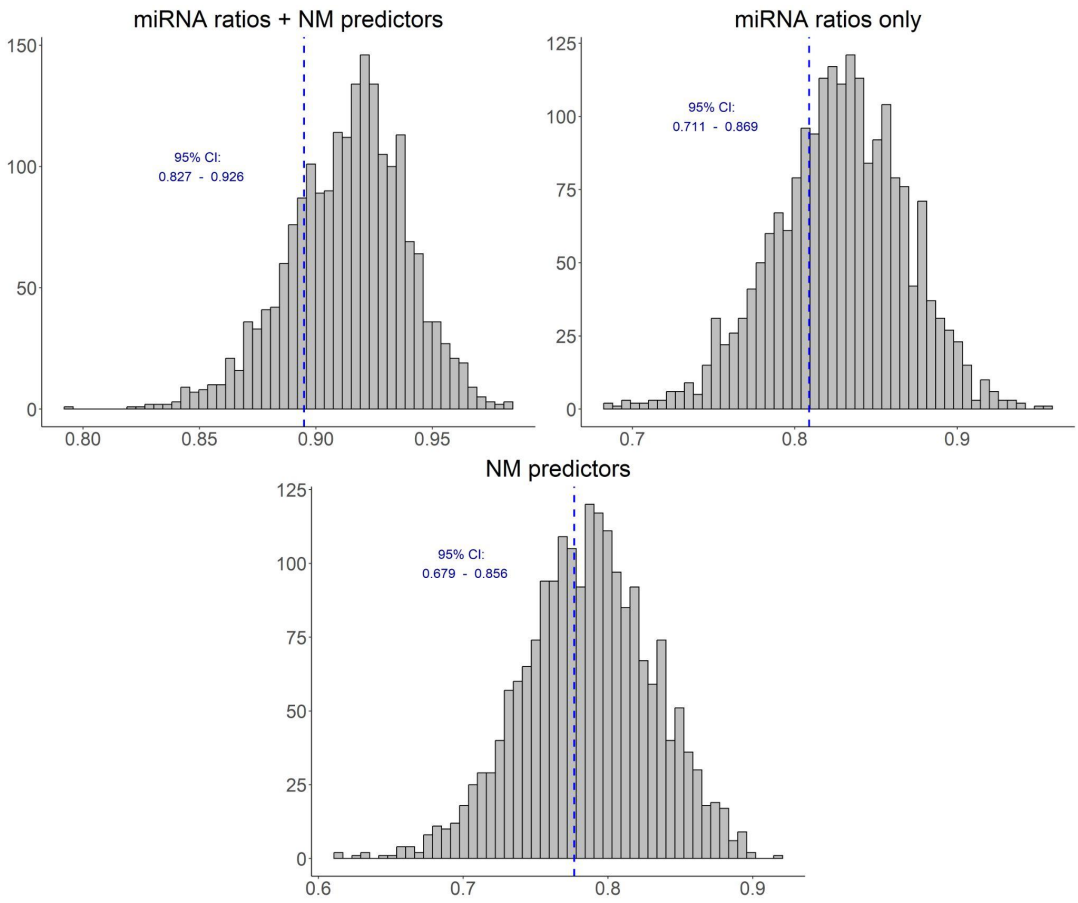


Figure 78. Histogram of bootstrapped ROC AUCs based on the ridge regression on the three models in the validation cohort.

To test the robustness of the frequentist ridge regression estimates and compare then to other methods, we also performed model updating using the Bayesian approach. We performed the Bayesian updating by setting the coefficients reported from the discovery cohort as prior means, while the standard deviation was set as the constant ln(4)/2 [280]. We obtained an ROC AUC of 0.87 [0.80 to 0.92], with only a slight miscalibration of the model (intercept: -0.21 [-0.65 to 0.22] and slope: 1.47 [0.93 to 2.01]). The Bayesian model updating was also performed on miRNA ratio-only and non-molecular variables-only models. The former model had an ROC AUC of 0.74 and was relatively well calibrated (intercept: -0.21 [-0.65 to 0.22] and slope: 1.47 [0.93 to 2.01]), while the latter showed similar results (ROC AUC = 0.78). Detailed plots on the calibration of Bayesian models can be seen in **Figure 79**.

In summary, the Bayesian model updating is comparable to the ridge regression reported previously, based on the discriminatory statistic and calibration. However, the miRNA ratio-only model performed better with the ridge regression.

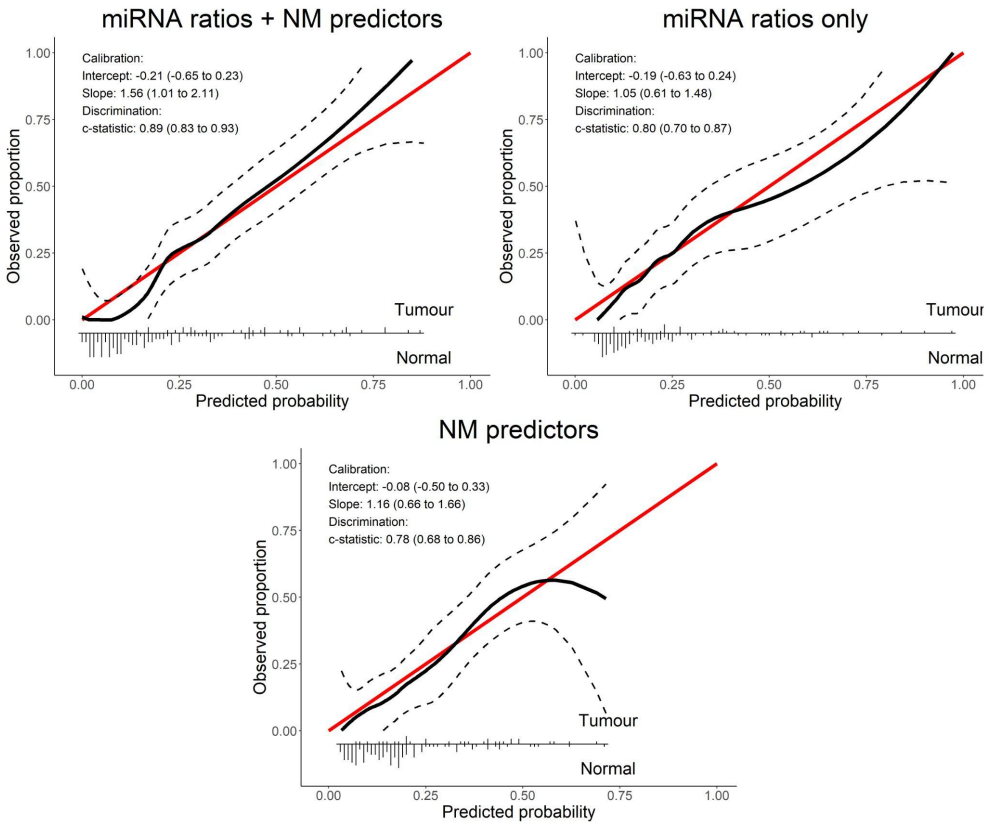


Figure 79. Calibration plots of the models calibrated using Bayesian model updating. The intercept and slope of the calibration plots are also shown.

The individual patient data (IPD) across different cohorts with the same predictors can be merged to create a new model while accounting for the cohorts. This can be done by utilising the IECV method. The IECV develops a model based on data from K-1 studies and tests it on the remaining study. Hence, in our case, we merged the discovery and validation cohort and, based on the heterogeneity estimate (of the Brier score), evaluated whether the mentioned models should be constructed in separate or combined cohorts. Additionally, since IECV can also be used to expand an intercept-only model by iteratively adding predictors common between cohorts, we used it to identify the set of predictors with the highest generalisability between our discovery and validation cohort. Importantly, since we only merged two cohorts, the IECV method has limited reliability.

Due to the relatively large heterogeneity ($tau^2 = 0.052$), combining the cohorts on all the predictors to create one prediction model would not create a more informative model than the ones obtained from individual cohorts. Nevertheless, the IECV can improve the generalizability and reduce the heterogeneity by finding the most generalisable predictors. After performing this on the two merged cohorts, a model with relatively low heterogeneity ($tau^2 = 0.002$), which included four predictors was created (miR-26b-5p_miR-142-5p, miR-21-5p_miR-23a-3p, interaction term of centred BMI and menopausal status, breast density). This model had a reduced discriminatory ability with an ROC AUC of 0.80 [0.74 to 0.84] but much higher generalisability compared to the

model utilising all predictors (the generalisability corresponds to the meta-analysed estimate of Brier score and was 0.43 in model with all predictors and 0.2 in the model with selected predictors).

The IECV on miRNA only selected two miRNA ratios (miR-199a-3p_let-7a-5p and miR-21-5p_miR-23a-3p), again with low heterogeneity (tau^2 = 0.002), reduced discriminatory ability (ROC AUC of 0.70 [0.64 to 0.77]) and higher generalisability compared to when using all miRNA ratios. This higher generalisability reflects the OR concordance of the two selected miRNA ratios between the two cohorts. Furthermore, the IECV was done on the non-molecular variables only, and the model on all predictors had a slightly lower heterogeneity (tau^2 = 0.001) than the model with filtered predictors (tau^2 = 0.002). This could be explained by the fact that non-molecular variables were very similar between the two cohorts. The ROC AUC of the model on all three non-molecular predictors was 0.74 [0.68 to 0.80]. Overall, the AUCs using the IECV generalisable predictors were lower but had superior calibration metrics compared to other recalibration methods (i.e., ridge regression or Bayesian model updating), as seen in **Figure 80**.

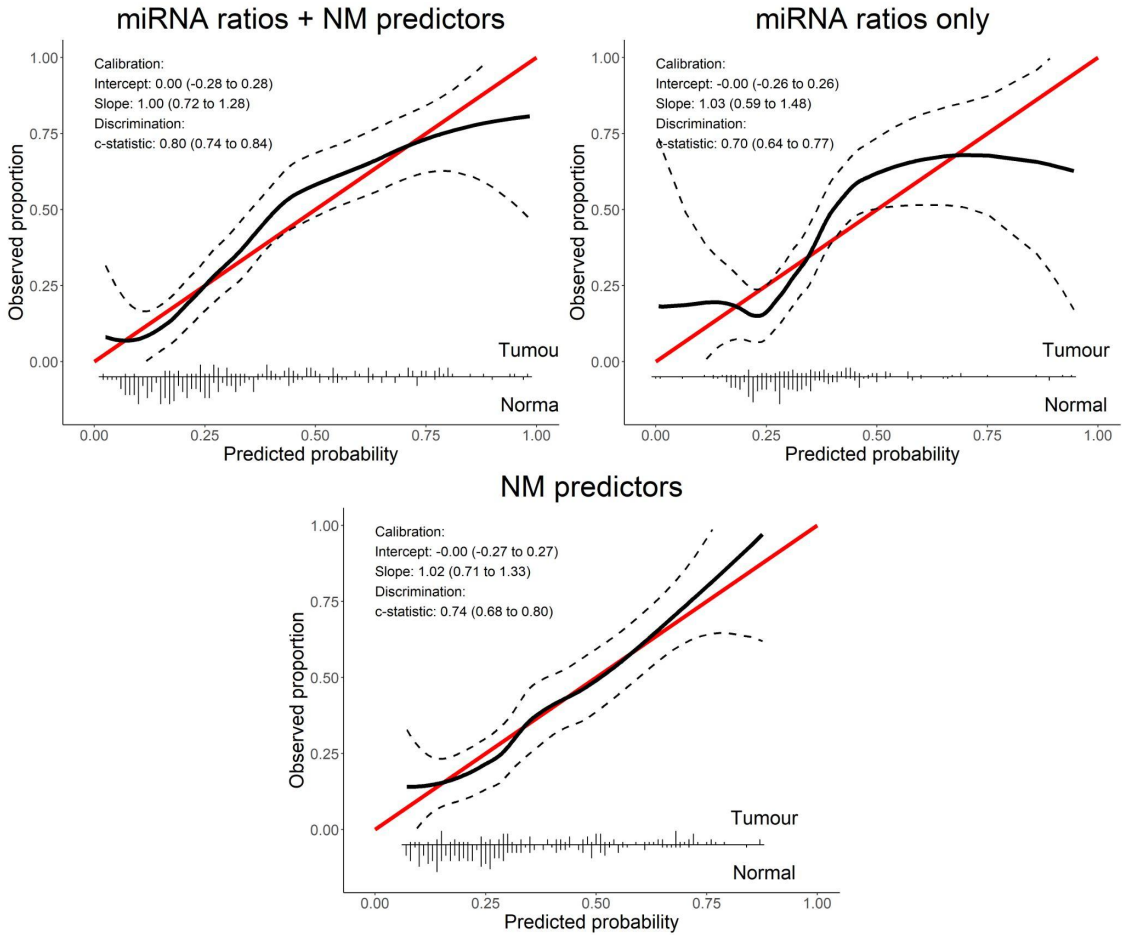


Figure 80. Calibration plot of the generalisable predictors of the three IECV models on the combined data from the discovery and validation cohort. The intercept and slope of the calibration plots are also shown.

Lastly, we tested the association of the miRNA ratios analysed in the validation cohort with clinicopathological parameters. Two miRNA ratios were associated with the ER status. Namely, miR-199a-3p_let-7a-5p was higher in ER+ compared to ER- cases (p-value: 0.049), while the opposite was found for miR-26b-5p_miR-142-5p (p-value: 0.027). Additionally, miR-93-5p_miR-19b-3p was lower in PgR− compared to PgR+ cases (p = 0.036), and let-7b-5p_miR-19b-3p was found to be associated with Tabar's classification of breast density (p = 0.025). The expressions of these miRNA ratios stratified by the clinicopathological variables they are associated with can be seen in **Figure 81**.

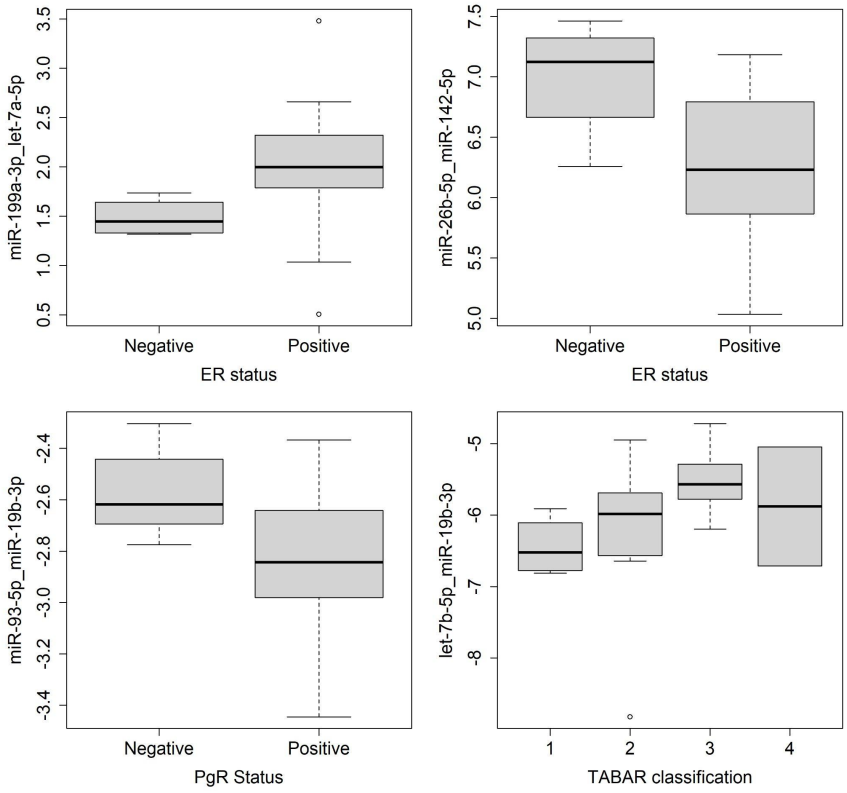


Figure 81. Expression values for miRNA ratios associated with clinicopathological BC cases characteristics in the validation cohort.

Subgroup and sensitivity analyses

As the sample size of our cohort was too small to perform biomarker discovery on subgroups, such as identifying miRNA ratios among different molecular subtypes of BC, we performed a set of analyses on the RT-qPCR data and non-molecular variables, which aimed to identify whether the variances and distributions of the predictors were the same between subgroups. Additionally, when the sample size allowed (i.e., when there were only two subgroups), we applied the previously reported model on the subgroups. These analyses were performed on both the discovery and validation cohorts. The set of subgroups we analysed were:

1) Family history status
2) Invasive vs in situ tumours
3) PRS stratification
4) Intrinsic molecular subtype
5) Lateral location of tumour

In the discovery cohort, no significant differences in variance or distribution were found between the 20 miRNA ratios and five non-molecular variables when comparing individuals with BC family history and those without. In the validation cohort, the same results were found except for miR-199a-3p_let-7a-5p, which had a significantly different variance between the two subgroups (F-statistic: 4.7, adjusted $p = 0.0002$). When comparing the predictors between in situ and invasive patients in the discovery cohort[2], three miRNA ratios were found to have significantly different variance (miR-101-3p_miR-19b-3p, miR-26b-5p_miR-142-5p and miR-20a-5p_miR-19b-3p). The PRS score based on the 77 SNPs was available only in the discovery cohort, and for the subgroup analysis, the individuals were stratified based on $PRS > 1$ and $PRS \leq 1$. There were no significantly different predictors based on any of the mentioned statistics, and the PRS data was not available in the validation cohort. Importantly, none of the predictors were significantly different between the intrinsic molecular tumour subtypes in the discovery cohort and only menopausal status was found to be different between some of the molecular subtypes in the validation cohort (Kruskal-Wallis test). These findings indicate that the identified miRNA ratios could be generalisable across all BC molecular subtypes. Finally, no differences in predictors were observed when comparing left to right breast tumour location in both the discovery and validation sets. We tried performing a LASSO logistic regression on the discovery cohort or applying the coefficients obtained on the total sample in the validation cohort subgroups, but due to a low event number, the results were not very conclusive or reliable.

We performed sensitivity analyses by rerunning the LASSO penalised logistic regression model on the discovery cohort (RT-qPCR miRNA data + non-molecular variables) and excluding various predictors to assess the performances.

To assess their utility without Tabar's breast density classification, the two models with non-molecular variables (miRNA ratios + non-molecular variables and non-molecular variables alone) were generated without the breast density variable. Without breast density, the model with miRNA ratios and non-molecular variables performed better than miRNA ratios alone, although not significantly ($z = 0.987$, p-value = 0.324). Just like in the models that included breast density, the model on miRNA ratios and non-molecular variables had a significantly better ROC AUC compared to non-molecular variables alone ($z = 2.862$, p-value = 0.004). In the model with miRNA ratios and non-molecular variables, the same 7 out of 20 miRNA ratios were selected, together with the interaction of BMI and menopause and WCRF lifestyle score. An ROC AUC of 0.77

[2] The validation cohort had only one in situ case so this comparison could not be made.

[0.68 to 0.84] was obtained, which is quite comparable to the model with breast density (**Figure 82**).

Next, the BMI, menopause status and their interaction were excluded, and a model was generated. The model selected eight miRNA ratios, breast density and WCRF lifestyle score with an ROC AUC of 0.79 [0.71 to 0.86]. The additional miRNA ratio selected was miR-335-5p_let-7f-5p with a relatively low coefficient (-0.022), which did appear multiple times when running the penalised LASSO logistic regression 100 times on different seeds. The model without WCRF lifestyle score selected the same miRNAs as the model without BMI and menopause status and their interaction. The non-molecular variables were breast density and the interaction between BMI and menopausal status. An ROC AUC of 0.78 [0.70 to 0.85] was obtained. Importantly, the predicted probabilities of all the models reported in this section were similarly calibrated (all had slight underestimations of risk).

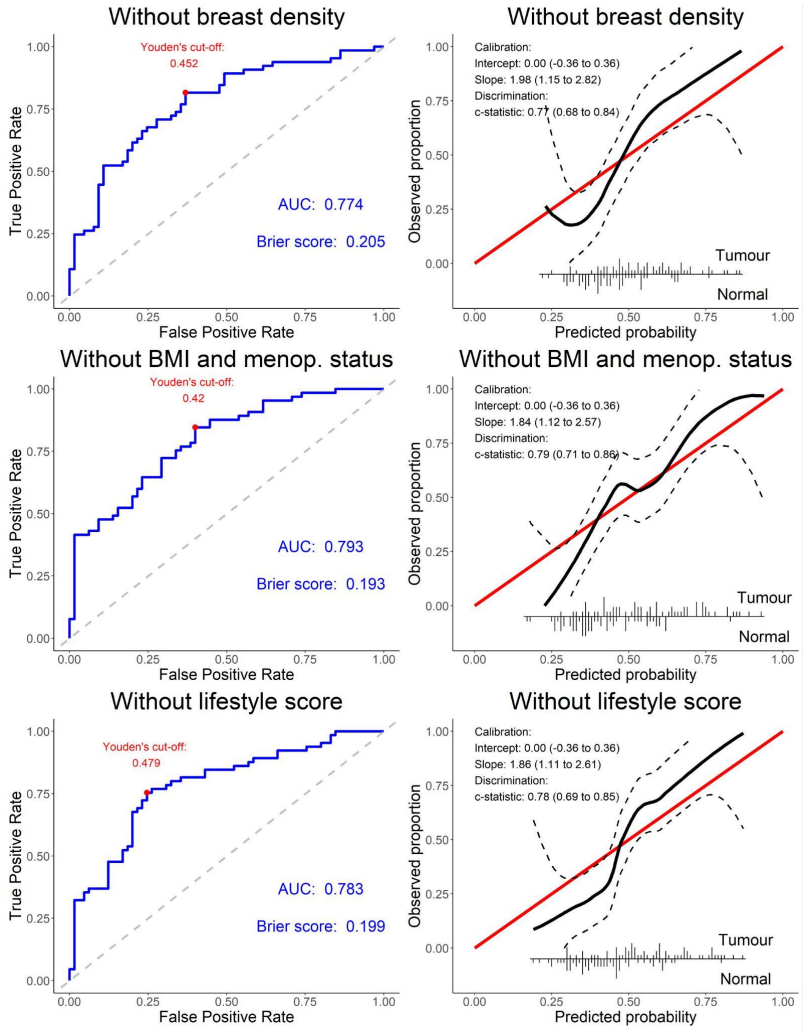


Figure 82. ROC AUC and calibration plots of the LASSO logistic regression models in the discovery cohort without specific predictors. Within the ROC AUCs, Youden's cut-off, AUC and Brier score are reported while within the calibration plots, the intercept and slope are reported.

To further test the discriminatory ability of the models without the breast density measurement, for which mammographic screening would be required, the model coefficients without the breast density predictor from the discovery cohort were applied to the validation cohort, and the results are shown in **Table 26**. Compared to the complete model, much poorer results without breast density were obtained. Furthermore, miscalibration of predicted probabilities was observed in the applied model without breast density.

After complete model revision, the performance and calibration improved significantly in the model with all predictors with an ROC AUC of 0.81. On the other hand, the model on non-molecular variables only needed the intercept to be recalibrated, and the resulting ROC AUC was 0.61 (**Figure 83**). The predicted probabilities of the recalibrated model on miRNA ratios and non-molecular variables without breast density had an optimal intercept but a slightly higher slope than optimal (1.49), while the predicted probabilities of the model based on non-molecular predictors had both the intercept and slope miscalibrated, indicating that the model on non-molecular variables without breast density was suboptimal.

Table 26. Model performance in validation cohort when applying the coefficients from the discovery cohort of models without the breast density predictor. The performances of the models within case and control subgroups were analysed as well.

	All predictors		**Non-molecular predictors**	
Sample subgroup	**AUC**	**95% CI**	**AUC**	**95% CI**
All samples	0.59	[0.48, 0.69]	0.61	[0.50, 0.71]
Without controls with additional biopsy	0.61	[0.49, 0.71]	0.62	[0.51, 0.72]
Without controls without additional biopsy	0.53	[0.39, 0.67]	0.57	[0.43, 0.71]
Without cases diagnosed more than 2 years after blood sampling	0.54	[0.38, 0.69]	0.54	[0.39, 0.68]
Without cases diagnosed less than 2 years after blood sampling	0.63	[0.49, 0.75]	0.68	[0.54, 0.78]

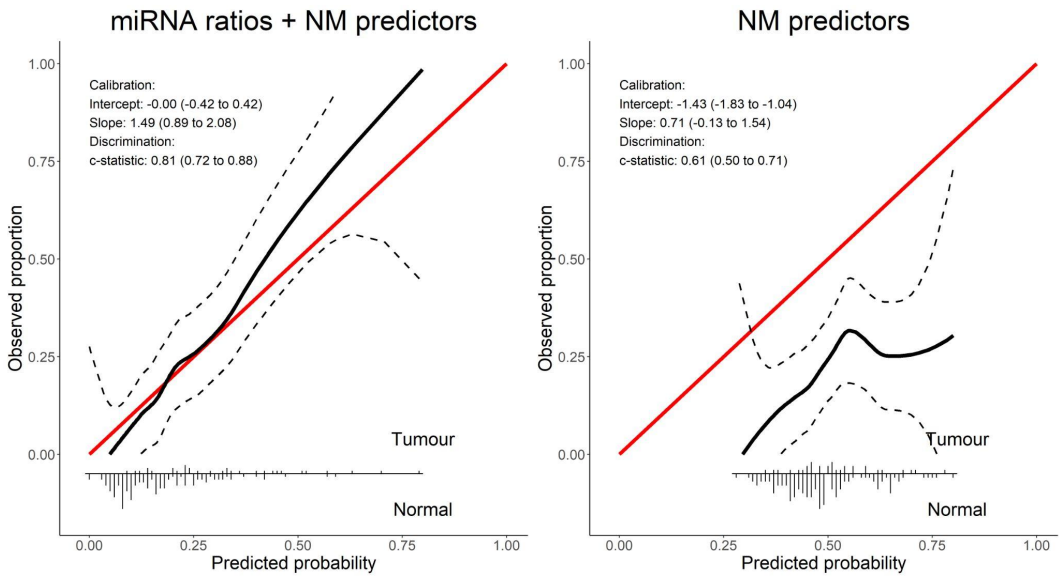


Figure 83. Calibration plots of the calibrated models on miRNA ratios and non-molecular predictors as well as non-molecular predictors only without breast density.

Candidate miRNA ratios in TCGA data

TCGA microRNA expression data (via small-RNA sequencing) was downloaded to evaluate the ability of the ratios to discriminate between healthy and tumour tissues. The dataset included 1,078 primary tumour samples and 104 adjacent healthy tissue, but we focused on the paired tumour and adjacent healthy tissue samples (103 pairs). The age at diagnosis for the analysed samples ranged from 30 to 90 years, with an average of 57.86 ± 14.7. We investigated which of the seven candidate miRNA ratios are differentially expressed between tumours and their adjacent normal tissues. From the seven miRNA ratios assessed in the validation cohort, six were differentially expressed ($p < 0.05$). The only non-differentially expressed ratio was let-7a-5p_miR-22-3p.

The mean fold change for each ratio was calculated by taking the mean of the fold changes across the paired samples (**Table 27**). Additionally, we performed a univariate conditional logistic regression on each ratio to account for the paired samples. Based on the conditional logistic regression, all ratios but let-7a-5p_miR-22-3p were significantly associated with BC ($p < 0.05$). The largest fold change and OR was observed for the ratio miR-21-5p_miR-23a-3p, which is expected as miR-21 is one of the most commonly dysregulated miRNAs in breast cancers but also across other types of cancer. Additionally, six of the seven ratios showed concordant OR when comparing the RT-qPCR miRNA ratio data from the discovery cohort to TCGA data (let-7a-5p_miR-22-3p was discordant). However, two ratios had an OR close to 1 (miR-21-5p_miR-23a-3p and miR-101-3p_miR-19b-3p).

Table 27. Univariate conditional logistic regression and paired Mann–Whitney U test on the seven candidate miRNA ratios in the tissue TCGA dataset.

miRNA ratio	OR	95% CI	P*	FC	log_2FC	P**
miR-199a-3p_let-7a-5p	2.23	[1.50, 3.32]	< 0.001	0.86	-0.22	1.14E-05
miR-26b-5p_miR-142-5p	0.39	[0.27, 0.57]	< 0.001	0.78	-0.35	5.38E-11
let-7b-5p_miR-19b-3p	0.73	[0.58, 0.92]	0.008	0.95	-0.08	4.48E-03
miR-101-3p_miR-19b-3p	0.71	[0.54, 0.94]	0.016	0.96	-0.06	1.95E-02
miR-93-5p_miR-19b-3p	2.35	[1.60, 3.44]	< 0.001	1.20	0.26	4.57E-07
let-7a-5p_miR-22-3p	1.17	[0.87, 1.56]	0.300	1.55	0.63	6.12E-01
miR-21-5p_miR-23a-3p	10.04	[3.44, 29.27]	< 0.001	1.72	0.78	5.82E-18

*P-value of the univariate logistic regression

**P-value of the Mann–Whitney U test

Puberty-associated CpG sites linked to BC

In this section, I analysed the CpG sites associated with pubertal timing or development from Sehovic et al. 2023 [234] in the context of BC, because early pubertal timing is a risk for BC in women. These CpG sites could be important biomarkers linking puberty and BC onset or risk. In addition, miRNAs significantly targeting the genes mapped to the CpGs of interest were identified.

From the mentioned study, based on the diseases and functions enrichment using the IPA tool, 2,711 CpG sites were enriched in BC processes as well as associated with puberty (standardised effect size > |0.13|). Furthermore, eight CpG sites associated with puberty, the same cut-off as the one above, were also found to be associated with BC risk based on peripheral blood samples. Six of the eight (**Table 28**) were found by a group from FIMM through twin discordance analyses on monozygotic twin pairs, indicating an environmental driver for the association between BC and methylation. All six were negatively associated with BC status. The two remaining CpGs were found to be associated with BC risk by a study on a prospectively sampled cohort (n = 162) of women where the CpG sites in peripheral blood were analysed. Both CpGs had a negative effect size.

I investigated the methylation of these CpG sites in blood and breast tissue, gene expression of the mapped genes as well as the miRNAs significantly targeting the underlying genes. The two sets of CpGs (based on IPA and those associated with risk) were analysed through two separate pipelines.

Table 28. List of CpG sites associated with puberty and BC risk which were investigated in this project.

CpG	Effect size or hazard ratio*	Source	Mapped Gene
cg00195561	0.16	Twin modelling	CHRM4
cg02079421	0.14	Twin modelling	PCNT
cg06579481	0.07	Twin modelling	
cg14018434	0.31	Twin modelling	SLC2A8
cg14038259	0.23	Twin modelling	RNF213
cg19212550	0.10	Twin modelling	DNMBP
cg00124920	-0.03	EPIC cohort	C1orf220
cg26772788	-0.02	EPIC cohort	

*Effect sizes from the conditional logistic regression performed on the EPIC cohort and hazard ratios from the paired Cox proportional hazard modelling on the Finnish twins.

Relevant miRNAs

To gain further epigenetic context for BC risk and puberty, I sought to understand which miRNAs play a role in regulating genes associated with puberty through DNA methylation sites. To do that, I performed a target enrichment analysis on the genes mapped to the 2,711 CpG sites associated with BC processes (from now on CpG set 1) as well as to the genes mapped to the eight CpG sites

associated with BC risk (from now on CpG set 2). Based on 1,990 identified genes (mapped to CpGs from set 1) by the Mienturnet software, 63 miRNAs were found to significantly target those genes (FDR < 0.05). On the other hand, the six genes mapped to CpG set 2 were significantly targeted by five miRNAs (FDR < 0.2). Due to the lower number of input genes, I increased the FDR cut-off. Four out of five miRNAs targeted the same two genes (*SLC2A8* and *C1orf220*) from the initial six in the input.

After obtaining the list of relevant miRNAs for both CpG set 1 and 2, I evaluated their expression in both breast tissue and blood. The blood data included the small-RNA sequencing on the 131 nested case–control cohort mentioned earlier, while the breast tissue data included the paired tumour and adjacent healthy tissue samples (n = 103 pairs from TCGA) on which small-RNA sequencing was performed. Forty miRNAs, targeting the genes mapped to CpGs in set 1, were found to be differentially expressed between BC tumour tissue and adjacent normal samples (**Figure 84**). While in plasma, four miRNAs (miR-21-5p, miR-22-3p, miR-19b-3p, miR-16-5p) which significantly targeted the genes mapped to CpGs in set 1 were differentially expressed. Three of the four microRNAs differentially expressed in plasma were also differentially expressed in tissue (all four but miR-19b-3p).

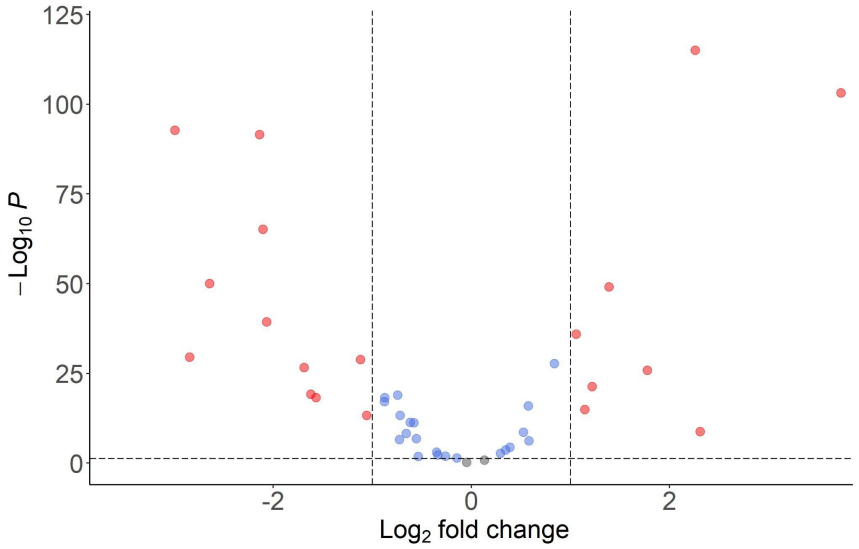


Figure 84. Volcano plot of paired class comparison of miRNAs in tumour and adjacent normal tissue (TCGA). The p-value shown in the plot is the Benjamini–Hochberg adjusted p-value. The red points indicate miRNAs above the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the blue points indicate miRNAs below the $\log_2$ fold deregulation cut-off and below the p-value cut-off, while the grey points indicate miRNAs that do not meet either of the criteria.

Among the five miRNAs targeting the genes mapped to CpGs in set 2, miR-26b-5p was differentially expressed in both tissue and plasma (**Figure 85**). The volcano plot was not shown due to the small number of input miRNAs.

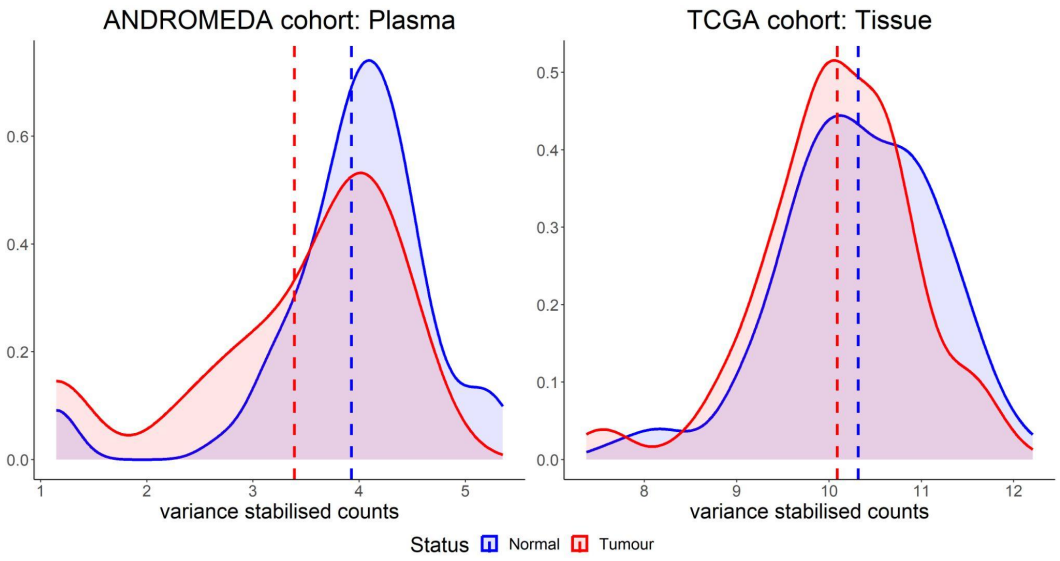


Figure 85. Density plots of miR-26b-5p stratified by BC status in plasma and tissue. The vertical dashed lines represent the mean expression.

Two genes, Cyclin Dependent Kinase 6 (*CDK6*) and Sp1 Transcription Factor (*SP1*), were commonly targeted by the three miRNAs differentially expressed in blood and tissue from set 1 analysis (**Figure 86**). On the other hand, the two genes targeted by miR-26b-5p, from set 2, were Dynamin Binding Protein (*DNMBP*) and *PCNT*.

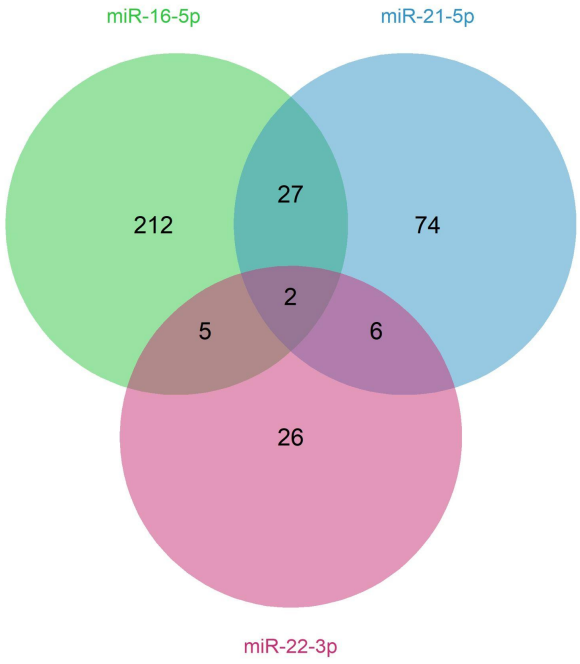


Figure 86. Venn diagram of the number of commonly targeted genes associated with puberty and BC by the miRNAs differentially expressed in both plasma and tissue.

Differentially expressed genes

Next, we looked at differential gene expression of the genes mapped to CpG sets 1 and 2. This was performed on RNA sequencing data on paired tumour and adjacent healthy breast tissues obtained from TCGA (n = 58 pairs). There were 50 differentially expressed genes between tumour and adjacent normal tissue (adjusted $p < 0.05$) in set 1 (**Figure 87**) and only one gene in set 2 (*DNMBP*).

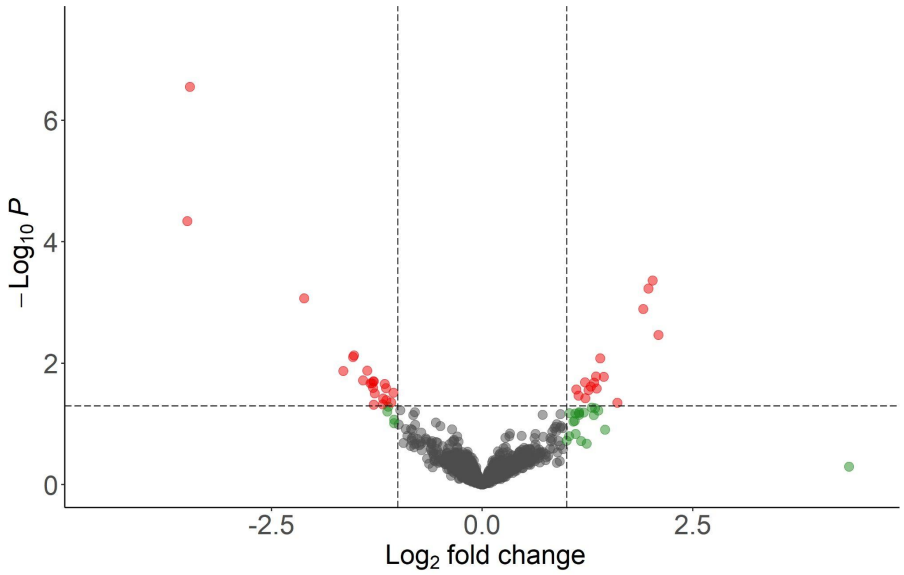


Figure 87. Volcano plot of paired differential expression analysis on TCGA tissue data of genes mapped to the CpGs associated with puberty and BC. The p-value shown in the plot is the Benjamini–Hochberg adjusted p-value. The red points indicate miRNAs above the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the blue points indicate miRNAs below the $\log_2$ fold deregulation cut-off and below the p-value cut-off, the green points are miRNAs above the $\log_2$ fold deregulation cut-off and above the p-value cut-off, while the grey points indicate miRNAs that do not meet either of the criteria.

The two genes commonly targeted by the three miRNAs mentioned above were not differentially expressed, while the gene targeted by miR-26b-5p, *DNMBP*, was differentially expressed. Thereafter, I explored Pearson's correlation in expression, based on variance-stabilised counts, between miRNAs that significantly target genes mapped to CpGs and differentially expressed genes they significantly target. Hence, for the CpG set 1, a matrix with 44 microRNAs, as the rest were filtered out based on count mean, and 50 differentially expressed genes was created. Overall, a high correlation between miRNAs and target genes was observed (**Figure 88**).

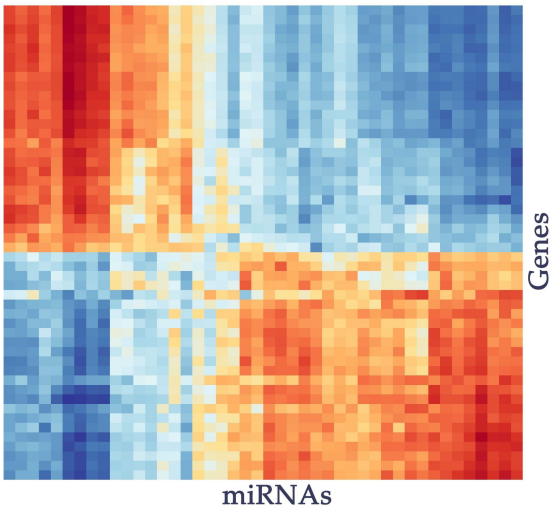


Figure 88. Correlation heatmap between miRNAs significantly targeting the genes mapped to the CpGs associated with puberty and BC and the gene expression of the targeted differentially expressed genes. The values range from -1 (blue) to 1 (red).

For the second set of CpGs, a matrix of one miRNA (miR-26b-5p), as the other miRNAs were filtered out due to low mean count, and six genes was created. Only one gene, *DNMBP*, had a correlation coefficient larger than |0.20| with the miR-26b-5p. In addition, from the 44 miRNAs on which we had correlation data, I selected the three that were differentially expressed in blood and tissue for further investigation. A slightly higher correlation across the 50 genes was observed for miR-21-5p compared to the other two miRNAs, indicating a more direct or prevalent role in regulating the mRNAs of the 50 genes (**Figure 89**).

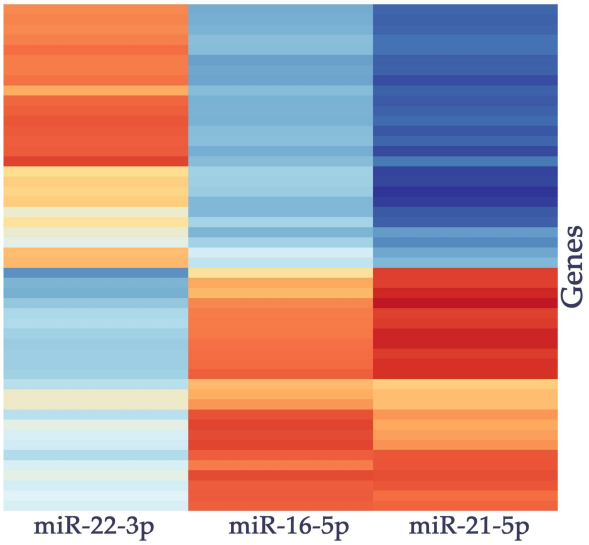


Figure 89. Correlation heatmap between three differentially expressed miRNAs in tissue and blood significantly targeting the genes mapped to the CpGs associated with puberty and BC and the gene expression of the targeted differentially expressed genes. The values range from -1 (blue) to 1 (red).

Additionally, I summed up the absolute correlation coefficients across the miRNAs for each gene, and a correlation sum cut-off > 1.2 was considered to select genes that were considered positively or negatively correlated with all three microRNAs. The reason 1.2 was chosen as the cut-off was that it was assumed to represent an average correlation coefficient of 0.4 with each miRNA. There were 22 genes that correlated with all three miRNAs.

Differentially methylated CpG sites

I studied the methylation differences of CpG set 1 and CpG set 2 between BC cases and controls in blood and breast tissue. To identify differentially methylated CpG sites in tissue, I used 57 pairs of tumour and adjacent healthy breast tissue from the TCGA methylation dataset. From CpG set 1, only one CpG was found to be differentially methylated (adjusted $p < 0.05$), cg23553576 to which the gene SYTL2 is mapped (p-adjusted: 0.005; log2FC: -0.323). None of the CpGs were differentially methylated in breast tissue from CpG set 2. Within the Italian prospective cohort (EPIC), which included BC cases and controls and had DNA methylation data on peripheral blood, 67 differentially methylated CpGs from set 1 were found (adjusted $p < 0.05$) (**Figure 90**).

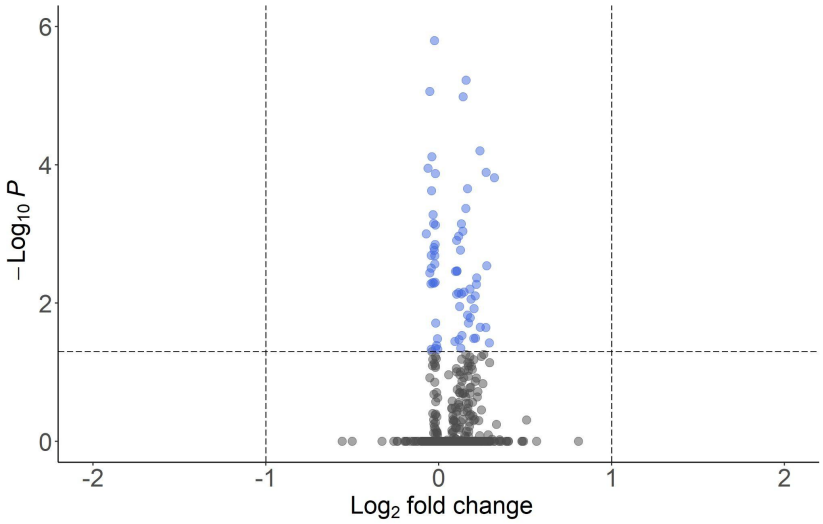


Figure 90. Volcano plot of the differential methylation analysis (in peripheral blood) of CpGs associated with puberty and BC (set 1). The p-value shown in the plot is the Benjamini–Hochberg adjusted p-value. The blue points indicate miRNAs below the $\log_2$ fold deregulation cut-off and below the p-value cut-off, while the grey points indicate miRNAs that do not meet either of the criteria.

Additionally, all eight CpGs from set 2 were significantly differentially methylated in blood between cases and controls. Notably, all eight CpGs had higher methylation in controls compared to cases (**Table 29**).

Table 29. Differential expression results of the CpGs from set 2 in peripheral blood. The log_2 fold change is also reported.

CpG	p-value	Mean Tumour	Mean Normal	log_2 FC
cg00124920	7.69E-10	0.74	0.77	-0.07
cg00195561	2.70E-07	0.86	0.88	-0.03
cg02079421	2.49E-05	0.88	0.89	-0.01
cg06579481	3.20E-04	0.76	0.80	-0.07
cg14018434	2.64E-02	0.94	0.94	-0.01
cg14038259	3.06E-04	0.88	0.88	-0.01
cg19212550	3.40E-05	0.70	0.74	-0.06
cg26772788	8.15E-10	0.76	0.79	-0.05

Finally, I evaluated the Pearson correlation of CpG set 1 and CpG set 2 with their mapped gene. For this, I used the TCGA samples which had both the methylation and gene expression data (n = 866). The correlation was performed between normalised Deseq counts (variance stabilising transformation) for gene expression and Beta values for methylation values. A total of 1,066 unique CpGs significantly correlated with their underlying gene in set 1 (adjusted $p < 0.05$). Conversely, in CpG set 2, out of the six CpGs with a mapped gene, four had a statistically significant correlation. The CpG differentially methylated in tissue was not significantly correlated with its mapped gene Synaptotagmin Like 2 (*SYTL2*) ($\rho = -0.15$, $p = 0.165$). The correlation results were merged with the annotated CpG data frame for future filtering.

Network analysis

In order to obtain a list of CpG sites on which to perform a network analysis, but which would also be studied more functionally and genomically, the parameters of the previously mentioned analyses were used to create a filtered list of CpGs:

1) CpGs associated with BC risk (n = 8)
2) CpGs mapped to the two common genes targeted by the three significantly differentially expressed miRNAs in blood and tissue (n = 4)
3) CpGs differentially methylated in tissue and CpGs differentially methylated in blood with a log_2 fold change > 0.25 or < -0.25 (n = 6)
4) CpGs mapped to differentially expressed genes with a log_2 FC > 2 or < -2 (n = 5)
5) CpGs mapped to genes which are correlated with all three miRNAs of interest mentioned earlier (n = 32)

Firstly, a network analysis using MetaCore was performed on the combined list of genes (number of unique genes = 42) mapped to CpGs/loci from the five subcategories. Based on the pathway map analysis in MetaCore, the genes involved in the networks were enriched in metaphase checkpoint, progesterone-mediated oocyte maturation and some pathways more directly related to breast cancer, such as PDGF signalling via PI3K/AKT and NFkB pathways (**Figure 91**).

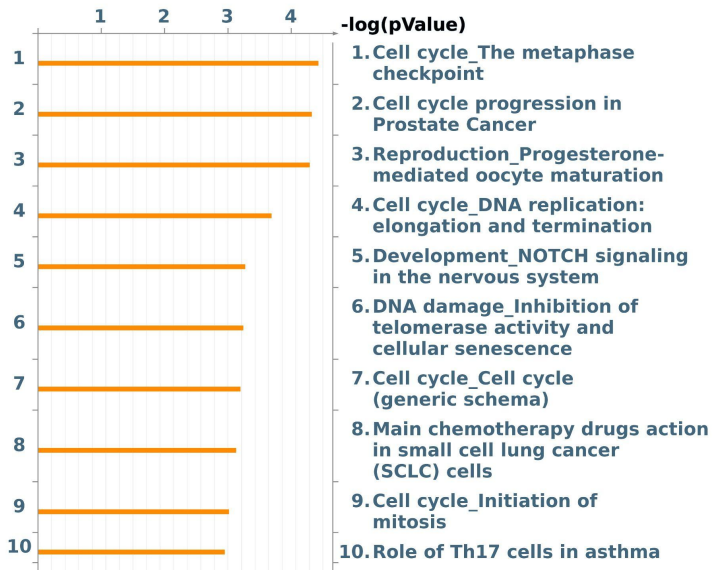


Figure 91. Pathway map results on the 42 unique genes mapped to CpGs associated with puberty and BC, which were the input for the network analysis.

Furthermore, a total of 166 unique transcription factors were found to interact with the 42 genes linked to puberty and BC. The transcription factors with the highest number of interactions were *LBP9*, *KLF4*, *GATA-1*, *FBI-1* (**Table 30**). Additionally, there were nine subnetworks, out of the 21 created by the software, based on the 42 unique genes, which had a z-score larger than 60 (**Table 31**).

Table 30. Transcription factors with the highest number of interactions with the 42 genes linked to puberty and BC.

Transcription factor	Actual	R	Expected	Ratio	Z
LBP9	33	12762	13.09	2.52	6.48
KLF4	33	15047	15.43	2.14	5.46
GATA-1	32	12405	12.72	2.52	6.33
FBI-1	31	11928	12.23	2.53	6.24
KLF17	29	11762	12.06	2.40	5.66
NANOG	29	12938	13.27	2.19	5.10
GATA-2	28	10404	10.67	2.63	6.04
c-Myc	27	12453	12.77	2.11	4.67
SOX2	26	7389	7.58	3.43	7.31
HNF3-alpha	25	9373	9.61	2.60	5.57
ETS1	24	8166	8.37	2.87	5.96
TAL1	23	9489	9.73	2.36	4.78
CTCF	22	6962	7.14	3.08	6.04
E2F1	21	5368	5.51	3.82	7.03
GCR	21	8316	8.53	2.46	4.72
CREB1	20	7400	7.59	2.64	4.92
SOX17	20	9346	9.58	2.09	3.77

Transcription factor	Actual	R	Expected	Ratio	Z
RXRA	20	9544	9.79	2.04	3.67

Note: all p-values were lower than 0.0001 and because the z-score is reported they were omitted from the table.

Actual: number of network objects in the activated dataset(s) which interact with the chosen object

R: number of network objects in the complete database or background list which interact with the chosen object

Expected: mean value for hypergeometric distribution (n*R/N); N in this case represents the total number of gene-based objects in the complete database or background list (45,315)

Ratio: connectivity ratio (Actual/Expected)

Table 31. Results of the network analysis on 42 genes linked to puberty and BC. The genes included in each subnetwork and the network statistics are shown.

Network	Seed nodes	p-value	Z	g-score
CDK1 (p34), HOXB13, CDCA1, OTR, GRO-2	11	< 0.001	152.69	152.69
TOP2 alpha, MMP-13, BUB1, SP1, UHRF1	7	< 0.001	95.17	95.17
CDK6, UHRF1, SP1, CDK1 (p34), PBK	6	< 0.001	81.57	81.57
SP1, C2orf48, ID4, BUB1, TPX2	5	< 0.001	72.45	72.45
MYH11, Desmuslin, ACM4, GRO-3, GRO-2	5	< 0.001	67.96	70.46
CGI-116, TNNT1, PPAPDC1A, CLCA4, DAND5	5	< 0.001	70.10	70.10
RRM2, MMP-13, E2F8, CDK1 (p34), HOXB13	5	< 0.001	68.65	68.65
PAQR4, DPP6, GABA-A, SLC2A8, DNMBP	5	< 0.001	68.65	68.65
DAZ2, DAZ	1	< 0.001	68.02	68.02
SLC2A8, 2-Deoxy-D-glucose	1	< 0.001	55.53	55.53
CNTN6, Connexin 26, Kinase MYT1, Beta-catenin,Actin cytoskeletal	3	< 0.001	43.95	43.95
E2F8, CDK6, CDK1, CDC18L, AKT1	3	< 0.001	42.03	42.03
ZNF687, CENP-F, UBE2C, LBP9, LBH	3	< 0.001	40.75	40.75
CDK1 (p34), Claspin, E2F1, MAOA	2	< 0.001	28.00	29.25
Kinase MYT1, CDK1 (p34), LBP9	2	< 0.001	27.70	28.95
Kendrin, ZNF687, LBP9, SLC41A1, C8orf37	2	< 0.001	27.14	27.14
UBE2C, LBP9, APOA4, ARP2, p53	1	0.003	17.51	18.76
E2F8, Rad51, RBBP4 (RbAp48), LBP9, AKT1	1	0.003	18.46	18.46
ZNF664, LBP9, RAMP1, UGT1A9	1	0.005	14.12	14.12
ANCO-2, LBP9, CRLR, Rab8B, F264	1	0.005	13.82	13.82

The g-score is a statistic modifying the Z-score based on the number of linear canonical pathway units within the network.

In network 1 (**Figure 92**), the key hub of interest is the one with the TF *YY1*, which interacts with several transcription factors, some of which are among the 42 genes (e.g., *HOXB13*) as well as the protein kinase *CDK1* which activates *PBK*.

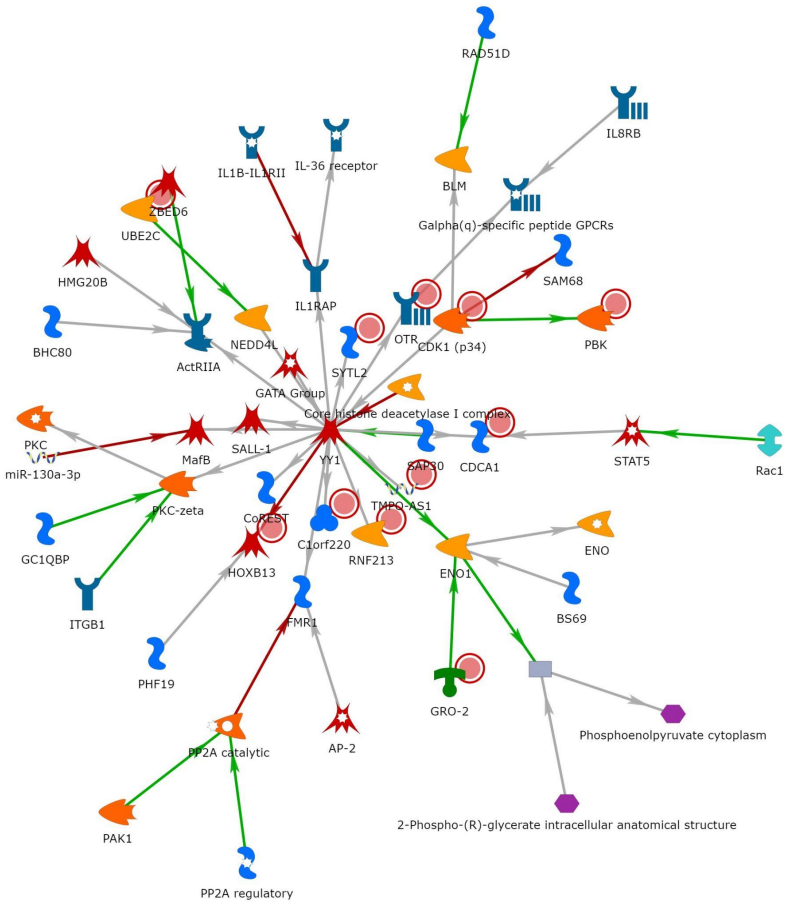


Figure 92. Graphical representation of network 1 from **Table 31**. The genes with a red circle next to them were the input genes.

In network 2 (**Figure 93**), there is one central hub around binding protein Vimentin, which are class-III intermediate filaments, and is activated by *SP1* and *TOP2A*. *TOP2A* is also a part of a small gene hub and is activated by *UHRF1* and has an unspecified interaction with *E2F8*. Two additional gene hubs around *E2F2* and *Sin3A* were created, however, none of the genes were a part of the 42 input genes.

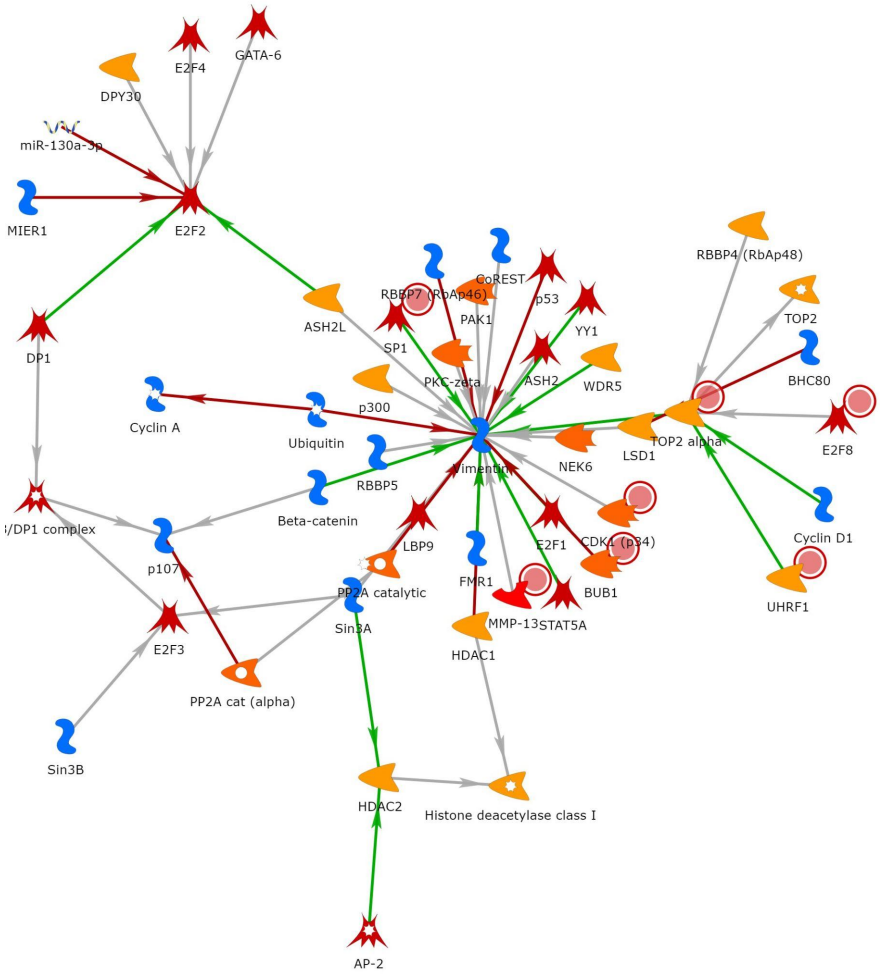


Figure 93. Graphical representation of network 2 from **Table 31**. The genes with a red circle next to them were the input genes.

In network 3, the main gene hub was created around the binding protein Cyclin D1. Cyclin D1 is activated by *E2F8* and *SP1* transcription factors, which were part of the input gene list. Additionally, *UHRF1* and *PBK* indirectly interact with Cyclin D1 through *STAT5* and *MAF* transcription factors, respectively (**Figure 94**).

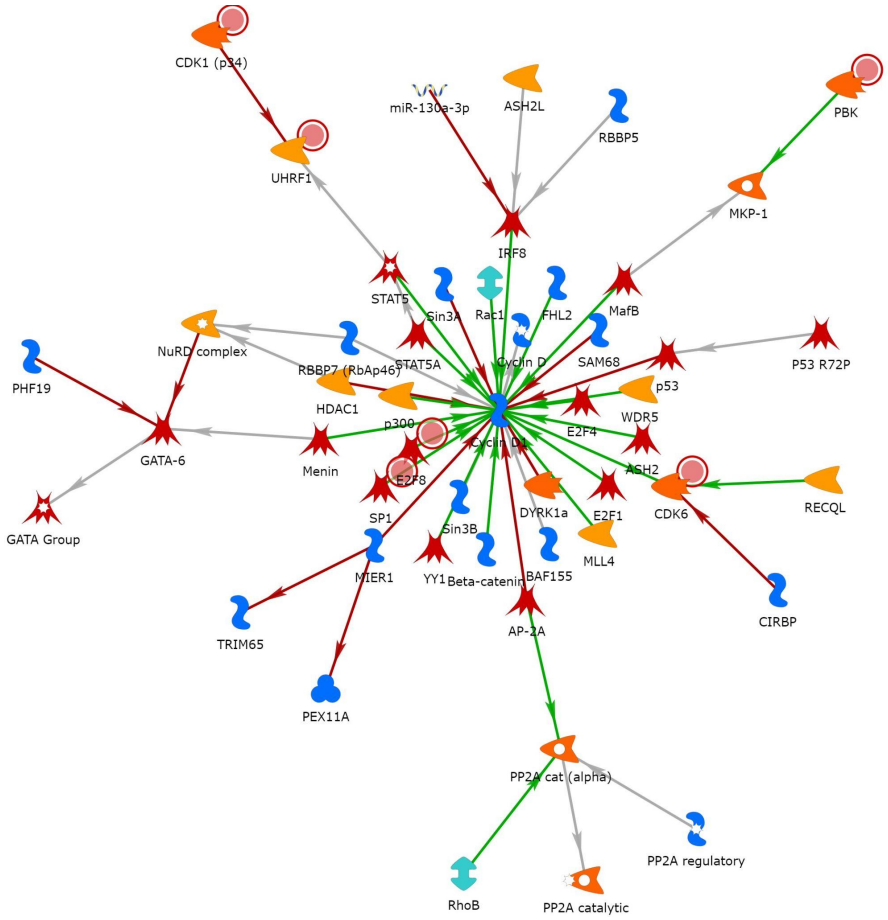


Figure 94. Graphical representation of network 3 from **Table 31**. The genes with a red circle next to them were the input genes.

In the direct interactions network, there were three main hubs around the transcription factor *SP1*, protein kinase *CDK1* and transcription factor *HOXB13* (**Figure 95**). Notably, *UHRF1* is inhibited by *CDK1* and activated by the transcription factor *SP1*. Additionally, it activates *TOP2* and inhibits Calponin-1 and *MYH11*. *CDK6* and *ID4* both activate *GRO-2*. *CDK6* also activates *GRO-3* and is functionally associated with *ANCO-2*. *PBK* is activated by *CDK1* and inhibits *PTEN*. The protein kinase *BUB1* is activated by *CDK1* and *HOXB13*, which also activates *CDCA1* and interacts with *UBE2C* and *TNNT1*.

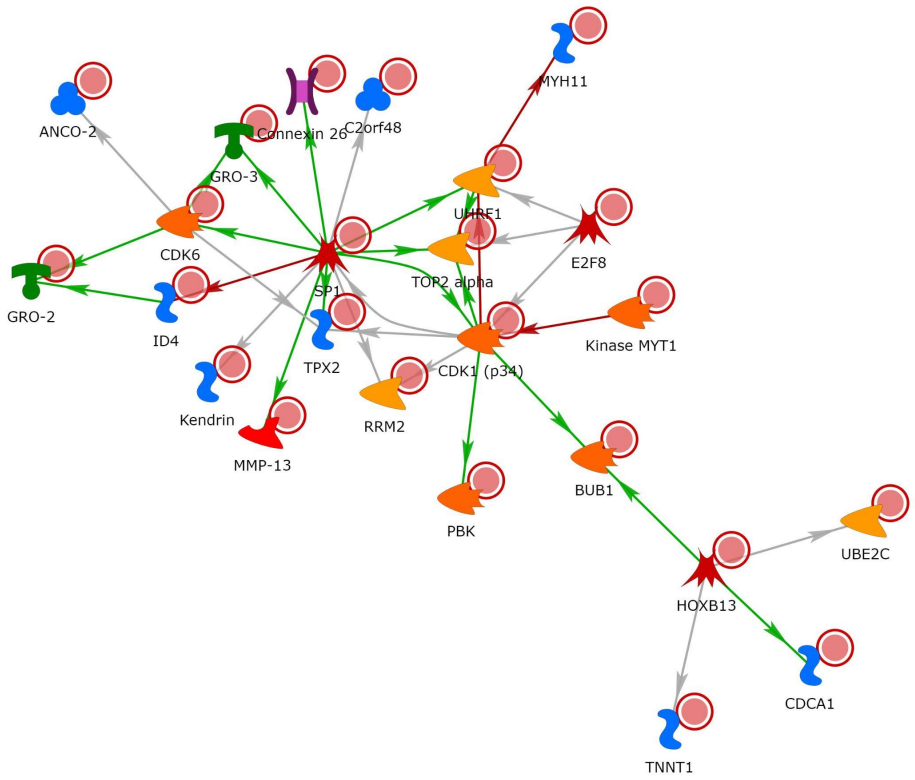


Figure 95. Graphical representation of the direct interaction network between the 42 unique input genes.

CpGs located on regulatory loci

The next step was to understand functionally and genomically the most important CpGs and their respective loci, genes or related miRNAs. Firstly, I examined the overlap of the super-enhancers in the human mammary epithelium (obtained from SEdb 2.0) with the genomic loci of the CpG sites found. A total of 190 unique super-enhancers were found across 247 CpGs from set 1, while only one super-enhancer was found in set 2 on CpG cg06579481. This CpG is not mapped to any gene and was not found to be differentially methylated in tissue but was differentially methylated in blood (**Table 29**). These super-enhancer mapping results were then annotated to the complete CpG data frame and were used as factors indicating a higher probability of regulatory function.

On the five filtered subsets of CpGs mentioned above, I then checked various genomic properties, described in detail in the methods section, as additional factors indicating CpG loci which could have regulatory functions or other functions affecting the underlying mapped gene. A total of 26 genomic loci with 21 unique genes, on which the CpG sites of interest were found, were classified as having a high probability of having a genomic regulatory function. The specific criteria met for the 26 CpG sites can be seen in **Table 32**. For some CpG sites, more than one criterion was met.

Table 32. CpG sites associated with puberty and BC that met one of the previously mentioned five criteria and have a high probability of being on a genomic regulatory site.

CpG	Gene	Chr.	Position*	Criteria met‡	cCRE signature§
cg00543329	CLSPN	1	36235839	4	promoter-like
cg01579019	PBK	8	27695516	4	promoter-like
cg01619846	UHRF1	19	4911475	2	
cg02079421	PCNT	21	47851545	3	
cg05951351	SYNM	15	99658349	1	
cg06258179	CDK6	7	92463261	4	promoter-like
cg06579481		7	104621597	1	
cg07085895	E2F8	11	19262124	2	
cg08892705	SYNM	15	99646202	1	
cg09593767	ZNF664	12	124457669	4	promoter-like
cg11823214	NUF2	1	163291487	4	promoter-like
cg11823214	NUF2	1	163291487	4	promoter-like
cg12695586	OXTR	3	8810077	2	
cg13510262	SP1	12	53774037	1	
cg15345369	RRM2	2	10262827	2 and 4	promoter-like
cg17305436	ID4	6	19837319	4	proximal enhancer-like
cg19048863	WASHC3	12	102455840	4	promoter-like
cg19347576	ANKRD12	18	9136711	1 and 4	promoter-like
cg19594360	ZNF687	1	151255303	4	proximal enhancer-like
cg20712426	CDK6	7	92464980	2 and 4	proximal enhancer-like
cg23097686	TMPO-AS1	12	98910128	4	promoter-like
cg19212550	DNMBP	10	101767908	2	
cg22288637	CDK6	7	92464428	2	
cg09863659	TPX2	20	30328246	2	
cg22041712	CENPF	1	214834360	2	
cg14038259	RNF213	17	78311557	2	

*hg19 genome position

‡Criteria on higher likelihood of being related to genomic regulatory site – refer to methods

§cCRE signature based on the aggregated cell types and human breast epithelium

The genomic location of 13 out of 26 CpGs had all four regulatory marker data available for breast epithelium in the ENCODE database. Therefore, for those CpG sites it was possible to infer the regulatory group of the candidate cCREs. In the breast epithelium cell type, ten of the cCREs were classified as having promoter-like signatures, while three were classified as having proximal enhancer-like signatures. All 13 cCREs had a concordant cCRE characterisation between aggregated cell types and breast epithelium. An extensive literature search in the context of BC and puberty was performed on the CpGs/genes classified as having a high probability of having a genomic regulatory function. Additionally, on the same set of genes, I also investigated the protein and RNA expression in breast tissue as well as the cell types expressing the RNA in breast or peripheral blood mononuclear cells (single-cell RNA sequencing data) using the Human Protein Atlas.

A total of 21 genes were thoroughly investigated in the literature using the PubMed database or Google Scholar search engine. The relevance of the genes in BC development and puberty was investigated, and the results are summarised in **Table 33**. Several genes were associated with puberty and BC, but five genes were found to have particularly relevant functions to BC risk and onset. Those five genes are *UHRF1*, *CDK6*, *PCNT*, *SP1* and Oxytocin Receptor (*OXTR*). *UHRF1* was found to be essential for germ cell development in males and females via regulation of several epigenetic pathways [297], and was associated with precocious puberty. In addition, several SNPs (rs12185519, rs12974635, rs16992771, rs2307209, rs2656924, rs3786941, rs4807665) on or close to the *UHRF1* gene were found to be associated with body height. When it comes to BC, *UHRF1* was found to modulate BC cell growth via oestrogen signalling and plays a role in the development of invasive ductal BC [298,299].

Twenty SNPs on *CDK6* were found to be associated with height. A study found two SNPs related to *CDK6* and *C6orf106* to be associated with pubertal growth spurt timing, and based on the Harmonizome knowledgebase, *CDK6* was associated with precocious puberty. When within the *D-CDK6* complex, *CDK6* was found to be important for tumour initiation. Additionally, it has been found to be relevant in mammary epithelial proliferation and BC initiation and maintenance.

Next, the *PCNT* gene has been linked to central precocious puberty [300], and mutations on this gene have been linked to Microcephalic Osteodysplastic Primordial Dwarfism, possibly linking precocious puberty and MOPDII [301]. A missense SNP on *PCNT* (rs7279204) was found to be associated with an increased risk of BC. Since eQTL analysis demonstrated that the SNP also correlated with other nearby genes, implying its potential role in regulating some cancer susceptibility genes, the authors argued it might play a role in regulating some cancer susceptibility genes [302]. Additionally, we found that *PCNT* gene expression correlated highly with the underlying puberty-associated CpG based on the TCGA samples.

The *SP1* gene also has several SNPs associated with height (rs10876469, rs11170394, rs12422555, rs191643352, rs1971762, rs2293059, rs574708537, rs7134628, rs7310771) and was found to be associated with precocious puberty. Moreover, it was found to play a role in folliculogenesis [303]. *SP1* regulates a number of cancer-related genes, and in the context of BC, it controls its proliferation via interaction with the insulin-like growth factors I receptor [304,305].

Finally, the *OXTR* gene is linked to precocious puberty and has SNPs that are associated with height. Furthermore, it can also be linked to puberty due to its associations with social affiliation, attachment, social support, trust, empathy, and other social or reproductive behaviours [306,307]. Although the exact effect does not seem to be linear, studies have linked *OXTR* to BC development and pathogenesis [308,309]. One study in mice found that the overexpression of *OXTR* induced cancer through prolactin or p-STAT5 pathways, which creates a microenvironment suitable for tumorigenesis [310].

Table 33. Genes mapped to the CpG sites which met the criteria mentioned in **Table 32**. Literature findings were summarised based on the gene's association with puberty and BC as well as the appropriate references.

Gene	Puberty			Breast cancer§	References
	Precocious puberty‡	Height (SNPs)	Other*		
ANKRD12	Yes	0	No	Yes	[311]
CDK6	Yes	19	Yes	Yes	[312–316]
CENPF	Yes	0	No	Yes	[317,318]
CLSPN	No	0	Yes	Yes	[319–321]
DNMBP	No	6	No	No	-
E2F8	Yes	0	No	Yes	[322,323]
ID4	No	12	Yes	Yes	[324–327]
NUF2	Yes	3	No	Yes	[328,329]
OXTR	Yes	3	Yes	Yes	[308–310]
PBK	Yes	6	Yes	NC	[330,331]
PCNT	Yes	0	NC	Yes	[300–302]
RNF213	No	0	No	NC	[332]
RRM2	Yes	4	NC	Yes	[333–335]
SP1	No	9	Yes	Yes	[303–305]
SYNM	No	3	Yes	Yes	[336–338]
TMPO-AS1	No	0	Yes	Yes	[339,340]
TPX2	Yes	1	No	Yes	[341,342]
UHRF1	Yes	7	Yes	Yes	[297–299,343,344]
WASHC3	No	11	No	No	-
ZNF664	No	0	Yes	NC	[345,346]
ZNF687	No	0	Yes	No	[347]

‡ Harmonizome gene knowledgebase was used to check whether a gene was associated with precocious puberty

*Studies which have shown relevance of the analysed genes in pubertal development

§Studies which have shown relevance of the analysed genes in BC onset and progression

NC: Not conclusive

The final step in linking puberty-associated genes with BC risk involved assessing the expression of these genes in breast and blood tissues, using single cell and mass data available on the Human Protein Atlas. The *UHRF1* gene is relatively low expressed in bulk breast tissue compared to other tissues such as the thymus or bone marrow (*Supplementary Figure 5* – Appendix C). Nevertheless, it does have its protein expressed in breast tissue, unlike many other tissues with similar gene expression levels. In breast tissue single-cell data, *UHRF1* was mainly expressed in breast glandular cells, while in PBMC, it was mainly expressed in dendritic cells.

Probably due to its involvement in cell-cycle progression, *CDK6* protein was expressed across all tissue and cell types. A similar result was found for *PCNT*, with the exception that it had overall low RNA expression across tissue types (except for skeletal muscle, tongue and heart muscle), indicating the presence of protein despite the low TPM value. More interesting were the single-

cell data results of transcription factor *SP1*, which showed high expression in breast glandular cells and adipocytes, while in PBMC it was primarily expressed in macrophages and monocytes (*Supplementary Figure 6* – Appendix C).

Interestingly, as seen in *Supplementary Figure 7* (Appendix C), the mRNA of the *OXTR* gene had the highest expression in breast tissue and in the single cell data it was only expressed in breast myoepithelial cells (breast tissue) and platelets (PBMC). This might indicate a more specific function of *OXTR* related to BC development and puberty.

Another notable gene was *ID4*, which had a relatively low TPM in bulk breast tissue but was found to have a high expression in breast glandular and myoepithelial cells as well as smooth muscle cells. It was not expressed in PBMC. Despite its low TPM in bulk breast tissue, *SYNM* protein was highly expressed in breast tissue. Upon investigation of single-cell data, we found that *SYNM* is mostly expressed in breast myoepithelial, endothelial, smooth muscle and breast glandular cells. In the PBMC single-cell data it was mostly expressed in platelets. With a relatively high TPM across all bulk tissues, including the breast, the gene *ZNF664* was found to be mainly expressed in breast glandular cells within the single-cell data. Not all figures were shown for the reported genes as they are readily available in the Human Protein Atlas.

Discussion

BC remains one of the highest incident cancers in women and its primary, and secondary prevention are essential to reducing mortality [348–351], improving the quality of life of women at risk [352,353] as well as reducing the health care costs associated with disease treatment [354]. Implementation of lifestyle awareness-raising campaigns or socially inclined activities could be a way of improving primary prevention for numerous diseases, including breast cancer. Early cancer detection through secondary prevention enables a much better prognosis due to the tumour having less time to evolve into different clonal expansions and to adapt to its environment or therapy [355,356]. The secondary prevention of BC is quite effective and includes breast self-examination and population screening [357]. Nevertheless, the tools used in the screening programs, such as mammography, which is the golden standard, do have room for improvement due to relatively high false positive reads or interval cancers that occur during routine BC screening [358]. Hence, early BC detection could be improved with accurate, non-invasive and cost-effective biomarkers that would complement or tailor the currently employed mammographic screening.

Numerous types of non-invasive biomarkers associated with BC risk or diagnosis were analysed in the past decade. Some examples of such biomarkers are polygenic risk scores [67] or methylation patterns of gene promoters [141,359], both assayed on germline DNA, or circulating DNAs [114,360], mRNAs [124] and non-coding RNAs such as miRNAs [224,232]. Circulating cell-free miRNAs are of special focus in this project as they are relatively stable in blood [361], and many of them were found to be candidates for early BC detection [224]. However, there are few commonly reported miRNAs or miRNA panels, possibly due to a lack of standardised experimental procedure and a scarcity of prospective studies [232]. In this project, we meta-analysed the most important studies concerning cell-free miRNAs for BC detection and developed a biomarker discovery pipeline within a prospective cohort study in a screening context to identify the most promising non-invasive biomarkers associated with BC.

Meta-analysis of cfc miRNAs

As cfc miRNAs are the most extensively studied type of non-coding RNAs in the diagnostic context and have been found to be promising biomarkers for the (early) detection of BC [224], through our meta-analysis, we evaluated the overall diagnostic performance capability of the thus-far reported circulating miRNA-based tools. We also investigated the lack of standardisation between the studies as well as other factors that might be causing discordant results and a lack of commonly appearing miRNAs that could be clinically viable diagnostic biomarkers. The observed pooled sensitivity (0.85) and specificity (0.83) obtained on all the reported models was quite satisfactory, especially since models with relatively poor performance were also included in the

pool. The obtained estimate of the pooled sensitivity is quite robust and reliable: after repeating the bivariate analysis without the influential models, a similar pooled sensitivity (0.84) and specificity (0.84) were obtained. However, it is important to note that a highly significant publication bias was observed based on Egger's test, which might also suggest the tendency of primary report authors to report the best-performing models instead of all a priori plausible models. In addition, studies tend to have slightly worse diagnostic performance, mainly reflected in specificity, when having a lower probability of bias or a lower probability of poor applicability (based on our tailored QUADAS-2 assessment). This indicates that studies without rigorous methodological practices and transparent reporting tend to overestimate the results, which should be considered when estimating the overall diagnostic ability of miRNAs.

Moreover, single or multiple miRNA panel and normaliser type were significant fixed effects in the bivariate model on all reported models. The subgroup analyses also confirmed the significance of the fixed effects model as we see a significantly better performance, especially in sensitivity, of multiple miRNA panels compared to single, as well as a superior performance of models utilising endogenous compared to exogenous normalisers. Considering that in the bivariate analysis there was a sample disparity between the models that used endogenous and exogenous normalisers, the issue was less severe in the univariate analysis based on the log-DOR. Nevertheless, in the univariate analysis, we also observed that models based on endogenous normalisers perform better than exogenous normalisers.

Various endogenous and exogenous normalising miRNAs or genes have been used in the meta-analysed studies and in studies working with circulating miRNAs in other fields. However, none were found to be an optimal solution for normalising RT-qPCR miRNA data [271]. This may be due to the absence of housekeeping circulating miRNAs and due to heterogeneity caused by differing batches or manufacturers of exogenous normalisers. Hence, the selection of the normalising molecule is one of the most important factors that contribute to the heterogeneity of results. One solution to the normaliser issue, which might produce more consistent results, as proposed by [271], is to compute ratios and compare them between cases and controls. Only one study [362] out of the 56 which we meta-analysed used the ratio-based values. Mimics of miRNAs and the mean threshold cycle of 50 miRNAs with the highest mean expression were two other types of normalisation methods found within three distinct meta-analysed studies [363–365]. However, we believe that the lack of experimental practicality and efficiency of the former and the lack of between-study comparability of the latter method may limit the use of such normalisation methods in a standardised way. Although not significant in the fixed effect model, a slight diagnostic performance difference between models with and without stages III and IV was observed. The same is true for models with and without stage IV. This indicates that the stage distribution could play a role in the between-study bias. Two other important factors could contribute to the increase of more consistent results: the use of validation cohorts and random selection of cases and controls with prospective sampling [366]. As shown in **Figure 11**, only about 40% of the studies used a validation cohort, while few studies performed or explicitly stated that they performed prospective

sampling without knowing the status of cases and controls. Independent internal and external cohorts are a fundamental requirement in the process of biomarker validation, while prospective random sampling and sampling of blood before biopsy would enable a non-biased and generalisable biomarker evaluation [366]. Blood sampling before biopsy would minimise the influence of biopsychological or physical effects that could also influence the level of circulating miRNAs [367]. Despite not being significant in this meta-analysis, differences in specimen type might influence the heterogeneity of the obtained results. Utilising plasma as specimen type runs the risk of having haemolysed samples which affects the miRNA content of the samples [368–370], as plasma contains cellular components that may contribute miRNAs from apoptotic or lysed cells (e.g., red blood cells, platelets). Therefore, studies using plasma as the specimen type need to check for haemolysed samples and exclude them [371] or to evaluate the influence of potential haemolysis on candidate miRNAs before their analysis in plasma samples [372]. On the other hand, during coagulation of serum samples, RNA molecules are released and may change the true profile of circulating miRNAs [370]. Hence, these issues are crucial in standardising the circulating miRNA detection procedure. Taken together, to obtain clinically viable diagnostic miRNAs that could be applied on the target population (women eligible for routine mammographic screening), a standardised laboratory protocol should be created. Moreover, future studies with random case–control selection from a prospective sample of women undergoing routine screening will allow for a standardised stage distribution and higher applicability of novel diagnostic biomarkers to the target population.

Among the meta-analysed models, there were slightly more models with a balanced case–control ratio than models with significantly more cases than controls. Models with significantly fewer cases were less common than the previous two groups. Sensitivity across the three groups seemed consistent, while the group with significantly fewer cases tends to have a larger FPR. Thus, the ratio of cases and controls has an effect on diagnostic accuracies, while the ratio of positive to negative predicted screens is influenced by or is a similarity of model preference for sensitivity or specificity. This was visualised in **Figure 21**, where we used the three cut-points for the case/control ratio, and **Figure 22**, where we used five cut-points. It is important to consider the effect of the case/control ratio when designing a diagnostic biomarker study, as it could have a major effect on the relationship between sensitivity and FPR. Hence, the reasoning behind a study's case/control ratio should be thought out in advance and reported to the readers. In our validation cohort, there were about four times fewer cases than controls because we wanted to study the performance of identified biomarkers in situations with more controls than cases, such as in the context of BC screening. When applying the model with miRNA ratios and non-molecular variables, using Youden's cut-off, we observed a relatively large FPR. There were no other deliberate statistical modelling or research design decisions that affected the preference for sensitivity and specificity. As will be discussed in detail below, the main objective of our project was to identify new biomarkers in a BC screening context. A model that can be applied more confidently on external cohorts and for which decision-making that affects the preference for sensitivity or specificity is more impactful needs to be developed on a much larger sample size.

For such models, the probability cut-off should be determined according to how the model would be applied (e.g., if it would be used to assist mammography, specificity would be preferable).

Either due to the model designs or authors' perceived costs of misdiagnosis, a slight preference for sensitivity or specificity was observed for some models. We tried to capture the trend of preference for specificity or sensitivity using the alpha at Q minimum and author's relative perceived cost (c_1) methods. On all reported models, a preference trend was seen using the alpha method, while for the c_1 method, the trend started being inconsistent around value 1. On the most important model per study, a trend was only visible using the c_1 method, which again broke down around the value of 1. Importantly, the underlying sensitivity and specificity depend on a plethora of factors, such as the biology behind the predictor, measurement tools, statistical modelling, population, etc. Therefore, the proportion of sensitivity to specificity is not fully robust and cannot be the only metric evaluated when assessing these methods. After further investigation of the models and studies labelled as having $c_1 > 1$, no association with QUADAS-2 reporting bias was found. Nevertheless, overfitting of some of the models could be causing a bias and noise within the alpha and c_1 preference estimates. Therefore, it is recommended to use these methods on out-of-sample performance results (i.e., ROC AUC). One reason why the c_1 method might be more optimal for the ideal scenario where we have out-of-sample results is that the alpha may vary for multiple analyses of the same dataset, whereas c_1 is based on the (implicit) "optimal" pair in the context of the study. In a nutshell, the recommendation for future preference evaluation methods would be to use c_1 only on models with out-of-sample performance results, for which it is deemed to be the most optimal for a study and for which the reasoning behind the design (regarding sample size, selection of predictors, etc.) has been reported, while the alpha metric could always accompany the c_1 metric as a check on correctness.

Prior to our meta-analysis, two meta-analyses on BC diagnostic circulating miRNAs were performed in 2014 [229,230]. The two studies meta-analysed a total of seventeen unique studies. Seven out of the previously meta-analysed 17 studies were included in this meta-analysis. This difference in included studies is reflected in the fact that we excluded studies with > 4.5% stage IV cases, while the previous studies did not. The reason for excluding these studies was the expectation of an overestimation of diagnostic performance in studies that include a higher proportion of stage IV cases than expected in BC community screening [236]. The pooled sensitivity and specificity obtained in our meta-analysis are in agreement with [229]. However, [230] have obtained a slightly lower pooled sensitivity and slightly higher specificity. This suggests that the overall diagnostic performance of circulating miRNAs on detection of BC has not significantly improved over the years. On the other hand, the pooled diagnostic performance obtained from the most important model of each study has shown an improvement in both sensitivity and specificity. Interestingly, the percentage of studies with high, low and unclear evaluations on the four key domains of QUADAS-2 were very similar between this study and [230]. As it is the most commonly analysed miRNA among the meta-analysed studies, we have evaluated the pooled sensitivity and specificity of miRNA-21-5p. A study in 2014 performed a meta-analysis on BC diagnostic serum

miRNA-21 [373]. Marginally lower pooled sensitivity and specificity of miRNA-21 were obtained in our meta-analysis compared to the estimates of [373].

The key strengths of our meta-analysis are the evaluation of all the reported models from each study (as opposed to singling out one model per study), exploration of the model or author preference for sensitivity or specificity and robust, comprehensive results obtained from bivariate analyses, complemented by univariate analyses when necessary. The main limitation is uncertainty due to unmodeled factors: laboratory and experimental differences, differences in stage composition of analysed cases within the studies, as well as different levels of statistical robustness of the models reported in primary studies. Another limitation is the relatively low number of databases assessed. Although we cannot exclude the possibility that we may have overlooked some studies in the research phase, based on the suggestions in the current literature regarding the choice of databases [374,375], we believe that the potential for systematic bias is low. Due to their complementarity, the databases chosen for this study have around 90% median recall rate when compared to the most elaborate approach with four databases (EMBASE, MedLine, Google Scholar and Web of Science) [376].

Some of the findings and conclusions from the meta-analysis regarding standardisation were used to improve the methods of our miRNA biomarker discovery study. For instance, we decided that utilising ratio-based computation could have the highest chance of removing laboratory bias and increasing study reproducibility. Considering that no significant difference in performance was observed between serum and plasma specimens, we decided to analyse the non-coding RNAs in plasma. Furthermore, we discussed how prospective sampling and sampling before biopsy or treatment are essential for reliable diagnostic models. Therefore, in our biomarker discovery study we adhered to these important details. Finally, we discussed and reported how our sample size might affect the model preferences for sensitivity or specificity.

Identifying novel biomarkers in a screening setting

There is an urgent need for non-invasive, easily reproducible and cost-effective biomarkers for BC detection. Moreover, to our knowledge, no study has focused on the potential role of circulating miRNAs in asymptomatic women undergoing general mammographic screening. Therefore, in the second part of the project we sought to identify novel and reliable non-invasive circulating biomarkers that could be used for early detection of BC in a screening setting. This was performed on a discovery cohort (70 cases and 70 controls) and a validation cohort (32 cases and 127 controls). The main differences between the cohorts are that in the validation cohort the cases had a longer average time between blood sampling and diagnosis (3 months in the discovery and 2.1 years in the validation cohort) and that the validation cohort included some controls that underwent biopsy due to a false-positive mammography result. Due to the much larger time window between

blood sampling and diagnosis in the validation cohort, the biomarkers analysed in the validation cohort can be considered somewhat predictive.

Different types of circulating biomarkers, as well as non-molecular variables (such as BMI, lifestyle score, etc.), were first assessed in the discovery cohort, and promising biomarkers were used to construct a diagnostic model that was applied in the validation cohort. Both the discovery and validation cohorts originate from a large BC screening study (ANDROMEDA), which screened 26,640 women of which 13,323 agreed to blood sample collection [377]. Importantly, all blood samples were collected immediately at the time of enrolment, thus before diagnosis and consequently before any treatment or intervention. Such an approach increased the chance of obtaining unbiased and reproducible results [378]. The biomarkers analysed in blood were methylation of *RARB*, *APC* and *BRCA1* promoters (using the MS-HRM method), PRS based on the 77 SNPs reported by Mavaddat et al. [109], all analysed on genomic DNA from the buffy coat, and cfc miRNAs analysed in the plasma.

Very few non-molecular variables were associated with BC in our discovery or validation cohort. For some variables, such as the recruiting hospital, an association was not expected to be observed. BMI, lifestyle according to WCRF guidelines and breast density were the significantly associated non-molecular variables in the discovery cohort. BMI has been confirmed to be a risk factor in numerous cancers, including postmenopausal BC [379]. Notably, BMI has a different effect on risk depending on the menopausal status, where a risk ratio (RR) of 0.94 [0.80 to 1.11] and 1.33 [1.20 to 1.48] was estimated for pre- and post-menopausal women, respectively [76]. Therefore, as will be discussed in more detail below, in spite of not being significantly associated with BC in our cohort, menopausal status was considered in the modelling phase, as well as its interaction with BMI. Furthermore, the WCRF/AICR lifestyle guidelines were found to be associated with the risk of various cancer types, including BC [380]. Some studies did not find it to be associated with BC [381], which could be related to the way of conducting the questionnaire or the underlying population. Our lifestyle score was calculated as in the Romaguera and colleagues [252] and was based on adherence to the WCRF recommendations from 2007. A newer guideline was published [382], and we have tested it on our cohort, but it performed slightly worse than the older version. Lastly, higher breast density is known to be a risk factor for BC [383].

In the validation cohort, the non-molecular variables associated with BC were previous benign biopsies, breastfeeding, waist circumference and breast density. The differing associations between the discovery and validation cohorts are probably an artefact of the relatively small size of the cohorts. The variables associated with BC in the validation cohort were previously described in the literature. For instance, women with a previous benign diagnosis had a 1.77 times increased risk compared to women without [384], while breastfeeding was found to reduce the risk of BC with a relative risk reduction of 4.3% for every 12 months of breastfeeding [74]. Through its effect on BMI, waist circumference has also been found to be associated with BC [385].

DNA methylation is an important mechanism when it comes to cancer risk and carcinogenesis in general as well as in BC [386]. Cancer cells can evolve to exploit it in numerous ways, such as hypermethylation at promoters of tumour suppressors and hypomethylation or methylation alterations at intergenic or intragenic regions [387]. Additionally, DNA methylation alterations in enhancers are also important for cancer, including BC [388]. Using the MS-HRM method we evaluated the methylation of *RARB*, *APC* and *BRCA1* promoters in the discovery cohort samples. MS-HRM is a fast and robust method for quantifying methylation at specific loci and has successfully been used with various specimen types [140]. Since not all samples could fit on one plate, they were split across two PCR plates, and the plate was included as a covariate in the analyses. The methylation estimates were very low across all samples, and no significant difference in methylation was observed between cases and controls. Moreover, due to the high frequency of zero methylation estimates, we performed a two-part analysis using the B^2 statistic on the zero data and the Wilcoxon rank sum exact test on the non-zero data [268]. Additionally, zero-inflated regression models were performed, but in neither test, for all three gene promoters, was methylation different between cases or controls or associated with cases or controls. The three gene promoters analysed have already been studied in the context of BC, both in blood and tissue. In a 2015 meta-analysis, *RARB* was found to be more methylated in cases than controls in both blood and tissue, with an OR of 7.27 [359]. Stratified by the source material, the OR in blood was 12.47 and 4.01 in tissue [359]. Nevertheless, some studies were unable to find a significant association between *RARB* methylation and BC [389]. In a meta-analysis from 2016, the *APC* gene promoter was also found to be associated with BC in both blood and tissue, with an OR of 8.92 [390]. Stratified by material, the OR was 9.93 and 9.44 in tissue and blood, respectively [390]. Finally, in another meta-analysis, *BRCA1* was also previously analysed in the same context of promoter methylation and an OR of 3.15 was reported [141]. Stratified by material, the OR in blood was 1.87 and 4.75 in tissue [141]. Furthermore, the three gene promoters were also associated with BC prognosis [391–393]. Unfortunately, we could not reproduce the mentioned results, possibly due to differences in cohort and study design (e.g., prospective or retrospective) and different approaches to obtaining methylation data.

We calculated the PRS on the samples from the training cohort based on the 77 SNPs from Mavaddat et al [109]. With an AUC of 0.5 and a non-significant OR of 0.98, PRS could not differentiate BC cases from controls in our cohort. The mentioned study reported a significant association of the PRS score based on 77 SNPs and BC with an OR of 1.55 [1.52 to 1.58]. The non-significant PRS result in our cohort could be due to population and sample size differences. The same group published another study in 2019 on 94,075 cases and 75,017 controls, which reported a PRS for predicting BC based on a novel signature of 313 SNPs. The study reported an OR of 1.65 [1.59 to 1.72], highlighting the potential of PRS for the detection or risk stratification of BC [67]. Nonetheless, its AUC is relatively low, indicating that PRS should ideally be used in combination with other biomarkers.

Several types of non-coding RNAs found in blood and tissue were previously found to be associated with BC risk and BC prognosis [172,394–396]. In this project we focused on cfc miRNAs in a BC screening setting. The initial high-throughput screening of the miRNAs was performed by small-RNA sequencing on extracted RNA from plasma. The miRNAs were first analysed individually, but the main biomarker pipeline involved ratio computation in NGS data, RT-qPCR assaying and testing the promising biomarkers in a validation cohort.

When analysed individually, based on small-RNA sequencing data, we found 27 miRNAs differentially expressed between BC cases and controls. The three miRNAs with the largest positive log fold change were miR-122, miR-3591 and miR-7, while the miRNAs with the largest negative log fold change were let-7f, let-7a and miR-26a. miR-122 plays a role in various cancers such as liver, gastric, breast and several others [397] and is a candidate circulating diagnostic and prognostic biomarker in numerous cancers [398–400], including BC [401]. In BC, depending on the cancer stage, miR-122 can play both tumour-suppressive [397,402] and pro-metastatic roles [403]. miR-3591 is involved in various signalling pathways associated with BC progression, such as IGF1-AKT or PI3K/AKT [404–406]. miR-7 plays a role in tumour suppression in BC [407,408] but was found to be overexpressed in the plasma of tumour samples in our study, implying the possible existence of mechanisms that export this miRNA outside the tumour cells and blood microenvironment as an adaptive mechanism. The same could also be true for miR-122. On the other hand, from the miRNAs overexpressed in healthy controls, miR-26a is believed to inhibit the proliferation and migration of BC by targeting the MCL-1 [409] or FAM98A [410] genes.

The candidate biomarkers differentiating between BC cases and healthy controls were designated to be validated using the RT-qPCR platform. As mentioned above, since there is no optimal normaliser for small non-coding RNAs when using RT-qPCR, we decided to calculate miRNA pairwise ratios from small-RNA sequencing data and perform a biomarker discovery analysis [271]. Candidate ratios obtained based on NGS data could then be compared directly with the ratios from the RT-qPCR platform, and the problem of normalisation and reproducibility of RT-qPCR data would be avoided.

A total of 20 miRNA ratios, consisting of 24 unique miRNAs, were obtained as potential biomarker candidates based on small-RNA sequencing data in the discovery cohort. The 24 miRNAs were then further tested with RT-qPCR on the same cohort. To assess the diagnostic ability of the candidate miRNA biomarkers and make a comparison to other non-molecular variables associated with BC in our cohort, three diagnostic models were built: a model based only on non-molecular variables, a model based only on miRNA ratios and a model with miRNA ratios and non-molecular variables combined. The multivariable model, which included three non-molecular variables, identified a signature of seven miRNA ratios consisting of 11 unique miRNAs. Four of the seven ratios were found to be associated with clinicopathological features of the cases, such as ER status or Ki-67, implying a possible direct function of the miRNAs comprising the ratios in cancer formation and progression. The target enrichment analysis of the

miRNAs that make up the seven ratios revealed that their target genes are involved in cancer pathways, including BC. Importantly, all 10 analysed miRNAs (one was excluded due to software limitations) were enriched in the PI3/AKT signalling pathway, which is relevant to tumour progression and endocrine resistance in BC [406]. The genes commonly targeted by the majority of the 10 miRNAs were *PTEN* and *NUFIP2*. *PTEN* is a known tumour suppressor blocking the PI3K signalling [411], while *NUFIP2* is an RNA-binding protein [412].

As mentioned in a previous section, performing variable selection on the $\log_2$-transformed miRNAs would have been better. Nevertheless, variable selection on the $\log_2$-transformed miRNAs showed similar results compared to the ratios of the raw miRNAs.

Five of the unique 11 miRNAs identified in the discovery cohort, from the model including miRNA ratios and non-molecular variables, were previously detected as potential diagnostic circulating biomarkers in other BC studies whose TNM stage distribution of cases also roughly matched the distribution of stages observed in BC screening programs [236]. These five miRNAs are: let-7a-5p [413], miR-19b-3p [363,414,415], let-7b-5p [401,414], miR-93-5p [414,416] and miR-21-5p [417]. With the exception of miR-19b-3p and miR-21-5p, the mentioned miRNAs are believed to be tumour suppressors or to have a protective role in BC tissue [418–420]. For instance, let-7a is believed to suppress BC cell migration by downregulating the CC chemokine receptor 7 [421]. Moreover, through IL-8 regulation, let-7b suppresses the cancer-promoting nature of BC-associated fibroblasts [418]. Additionally, circulating miR-21-5p was the most commonly found miRNA studied in the context of BC diagnosis. The discriminatory diagnostic capability of miRNA ratios (both alone and when combined with non-molecular variables) showed promising results in the discovery cohort (AUCs of 0.73 and 0.79, respectively) and was comparable to those obtained in previous studies [230,362,363,422,423]. For example, Fang et al. 2019, who also used a plasma-based miRNA ratio model (five ratios) with multi-platform validation on 131 samples, obtained a sensitivity and specificity of 71.7% and 78.2%, respectively [362]. The five ratios used by Fang et al. 2019 consisted of seven unique miRNAs, none of which correspond to the miRNAs in our final model. This could, in part, be due to the different reference populations or variations in experimental and analytical methods. For instance, in Fang et al. 2019, miRNA ratios were calculated using RT-qPCR data only. Another study performed in 2015 [363], using a profiling (n = 86) and validation cohort (n = 196), reported an 8-miRNA model (miR-16, let-7d, miR-103, miR-107, miR-148a, let-7i, miR-19b, miR-22-5p) with a 91% sensitivity and 49% specificity and an AUC of 0.81. One of the miRNAs in their model, miR-19b, was included in three ratios obtained in our final models.

Due to the small sample size, the models we created are only indicative of which biomarkers are promising, and their characteristics (e.g., logistic regression coefficients or calibration curve intercept and slope) should be optimised on a large cohort. Therefore, we did not create agnostic models that would perform variable selection on all available molecular and non-molecular biomarkers, but only on those which seemed to be associated with BC in our cohort.

The three models (miRNA ratios and non-molecular predictors, miRNA ratios only and non-molecular predictors only) developed in the discovery cohort were applied to our validation cohort of 32 cases and 127 controls. Performance was relatively poor after applying the model coefficients, where the model on miRNA ratios alone had an ROC AUC of 0.51 and the model on miRNA ratios and non-molecular predictors had an ROC AUC of 0.71. The poor discriminatory performance was probably due to the miscalibration of the models, as all three models were suboptimal when considering calibration-in-the-large [424]. The model based on non-molecular predictors had the best calibration and performance (ROC AUC: 0.74), indicating greater stability and generalisability of non-molecular predictors between the discovery and validation cohorts than miRNA ratios.

We performed model revision of all three models using ridge regression and applied the new coefficients to the validation cohort. As in the discovery cohort, the model with miRNA ratios and non-molecular predictors was the best-performing model, with a significantly better ROC AUC (0.89) than the model on non-molecular predictors alone, highlighting the discriminatory power added by including miRNA ratios. As an alternative to the frequentist ridge regression, in order to check the consistency of the results, the three models mentioned above were also updated using the Bayesian model updating approach. It was described and proposed as an alternative in several studies [281] and was recommended, especially in cases where the sample size is relatively low [279]. Importantly, the ROC AUC and predicted probabilities of the models were highly comparable to the ones obtained using the frequentist method. As mentioned earlier, due to the large time lag between diagnosis and blood sampling in the cases of the validation cohort, the predictive ability of biomarkers was somewhat tested. This could partly explain the initial miscalibration of the ratio predictors and could indicate a possible predictive potential of these biomarkers in addition to the discriminatory one at the time of diagnosis.

In another study, a panel of eight miRNAs validated and developed on a relatively sizeable prospective cohort achieved an ROC AUC of 0.915 [425], and one of these miRNAs is included in our ratio signature of seven miRNAs (miR-19b-3p). It was not possible to compare the model calibration between our and the mentioned study, as the study cited did not report the necessary data. However, considering that our study was based on a screening cohort and considering the limited sample size, further validation of these miRNA ratio sets on larger screening cohorts is desirable.

To evaluate whether a model could be constructed on the merged data of the discovery and validation cohort, as well as to identify the most generalisable predictors between the two cohorts, the IECV method was employed [282]. Due to the considerable heterogeneity between the cohorts when using all common predictors, modelling on merged cohorts is not optimal. Nevertheless, we exploited this method to identify the most generalisable predictors across the cohorts and obtained two miRNA ratios (miR-26b-5p_miR-142-5p and miR-21-5p_miR-23a-3p), the interaction between centred BMI and menopausal status and breast density. In the IECV on miRNA ratio

predictors, again there were two generalisable miRNA ratios, but instead of miR-26b-5p_miR-142-5p, miR-199a-3p_let-7a-5p was selected. These two miRNA ratios had the largest coefficients in the validation cohort after model updating, indicating their potential as diagnostic biomarkers.

The biomarkers identified in this project were based on cohorts that included various types of molecular and histological BC tumours, with the aim of identifying a biomarker signature that would be able to discriminate most, if not all, BC tumours from healthy controls. Unfortunately, due to the sample size limitations, it was not possible to accurately assess the biomarkers' performance and applicability among the different BC subtypes. In any case, for each of the candidate biomarkers (miRNA ratios or selected non-molecular predictors), we tested whether there were differences in distribution and variance between the BC subtypes or differences between other patient characteristics. For almost all the cohort characteristics, the selected predictors did not have significantly different distributions or variances. This suggests a possible generalisability of the diagnostic biomarkers we have identified to various subtypes of BC. It is paramount to further validate these biomarkers in a screening setting on a much larger sample size and investigate their actual applicability across various BC subtypes.

To evaluate the robustness of our model on miRNA ratios and non-molecular predictors, we performed various sensitivity analyses in the discovery cohort. We evaluated the performance of the model without some of the non-molecular predictors, and in all instances, the models were relatively stable, and their performance did not decline significantly. Additionally, the models with non-molecular predictors but without breast density were applied to the validation cohort, and a much worse performance was obtained than that of the models with breast density. The ROC AUC in miRNA ratios and non-molecular predictors was 0.59, while in non-molecular predictors only, it was 0.61. As in the models with breast density, the predicted probabilities were highly miscalibrated. After recalibration, the performance of miRNA ratios and non-molecular predictors improved (ROC AUC = 0.81). These results demonstrate the potential of the identified miRNA ratios to be used as a diagnostic or risk-stratifying tool before women undergo mammography screening and breast density is not yet known.

Finally, six of the seven candidate miRNA ratios were associated with BC when analysed on paired tumour and adjacent normal tissue samples from TCGA. Furthermore, these miRNA ratios were concordant based on the OR between the TCGA data and RT-qPCR data in the discovery cohort. Hence, these results provide additional evidence that the miRNAs that make up the miRNA ratios play a role in BC progression or onset. For a holistic understanding of the miRNA ratios as biomarkers, it is crucial to understand why a concordant result between blood and tissue was observed and whether it was a coincidence. One hypothesis is that cancers can affect the miRNA expression profile by releasing miRNAs into the bloodstream (for cell communication, response to various stimuli etc.) and, in that way, create relative miRNA proportions similar to those observed in the cancer cells [426,427]. Additionally, dead cancer cells could also be a source of the

miRNAs in the bloodstream [426,427]. Nevertheless, it is important to further investigate these hypotheses for a better understanding of the found cfc biomarkers.

As was observed in our meta-analysis, in most similar studies that have analysed cfc miRNAs in the context of early diagnosis of BC, the controls usually come from healthy donors recruited in a separate setting from the cases, which were generally diagnosed prior to blood sampling. Therefore, the main strength of our study is that all samples came from a similar screening setting and were taken prospectively, with the limitation of a relatively small sample size [428]. Moreover, most published studies on diagnostic cfc miRNAs are based on endogenous or exogenous miRNA normalisers [224]. As mentioned, an essential aspect of the standardisation of cfc miRNA analysis is the normalisation method [224,228], and utilising values based on the ratio of miRNAs is a good step to overcome the lack of optimal endogenous or exogenous normalisers [271]. In addition, taking blood prior to biopsy and before knowing BC status could offer a better chance of obtaining a non-confounded circulating miRNA profile.

DNA methylation sites associated with BC and puberty

Gene expression can be epigenetically regulated through various mechanisms, such as DNA methylation, histone modification, RNA-induced silencing complex based on miRNAs, etc [429–431]. Like other epigenetic mechanisms, DNA methylation is an important aspect of the formation and progression of most tumours, including BC. DNA methylation can also be associated with development [432], puberty [433,434] as well as lifestyle factors [435], and can be used as an ageing clock [436]. Therefore, with the group from the Finnish Institute of Molecular Medicine, I identified DNA methylation sites associated with puberty among young adult Finnish twins and assessed their potential function or involvement in associated pathologies. [234]. Because blood sampling occurred several years after the completion of puberty, the identified CpG sites should mainly be considered as biomarkers for pubertal development. Since early puberty is a risk factor for BC [287], I further investigated the puberty-associated CpG sites that were linked to BC processes (set 1) via IPA or linked to BC risk (set 2). I then identified miRNAs that significantly target genes mapped to the above-mentioned CpG sites and CpG sites that are located at potential genomic regulatory points, such as enhancers and super-enhancers.

Forty miRNAs identified as differentially expressed in BC tissue and four in blood significantly targeted the genes mapped to the CpG sites associated with puberty and BC. Three miRNAs (miR-21-5p, miR-22-3p and miR-16-5p) commonly targeted *CDK6* and *SP1*, mapped to CpG sites from set 1. Importantly, these three miRNAs were found in the 21 miRNA ratios identified as candidate biomarkers for BC detection in small-RNA sequencing. Two of the three miRNAs (miR-21-5p and miR-22-3p) also made up the ratios in the model built in the discovery cohort and tested in the validation cohort. miR-21 is consistently upregulated in BC cases relative to controls in both tissue and blood [437,438]. It is considered to be an oncomiR [439], and one of its described functions

involves regulating the expression of *STAT3* and being linked to cell proliferation, colony formation, migration and invasion [440]. On the other hand, miR-22 is a protective miRNA as it has been found to suppress tumorigenesis in BC [441] and some other cancers [442]. Additionally, miR-22 is involved in oestrogen signalling by inhibiting ERα expression [443], which makes it also relevant for puberty due to the importance of oestrogen signalling during puberty [109]. Furthermore, miR-26b-5p significantly targeted *DNMBP* and *PCNT*, mapped to CpGs from set 2. miR-26b-5p is also a protective miRNA as it inhibits proliferation in triple-negative BC [444] as well as some other cancers, such as thyroid cancer [445].

We found 50 differentially expressed genes between BC tumour and healthy adjacent tissue mapped to CpG sites from set 1 and one differentially expressed gene mapped to CpG sites from set 2. Additionally, one and 67 CpG sites from set 1 were differentially methylated in tissue and blood, respectively. Regarding CpG sites from set 2, none were differentially methylated in tissue, while all eight were differentially methylated in blood. Since a high overall correlation was observed between the genes mapped to CpGs of set 1 and the miRNAs that significantly target them, it is plausible that these miRNAs are important for pubertal development and the onset or progression of BC. Furthermore, 22 genes were selected based on their high correlation with the three miRNAs differentially expressed in blood and tissue. Methylation of 1,066 and 4 CpG sites correlated with the gene expression of their underlying gene in CpG set 1 and set 2, respectively. The identified differentially expressed genes and differentially methylated CpG sites could be biomarkers linking puberty and BC onset.

Network analysis provided further information on the genes associated with puberty and BC. From the input gene list, several smaller highly significant networks were generated, which usually have a main node around an external gene or protein that interacts with the input gene list. Some notable central hub genes are the transcription factor *YY1* and binding proteins Cyclin D1 and Vimentin. *YY1* is a transcription activator and repressor expressed in numerous cell types and has been associated with cell cycle progression [411]. It was found to be an oncogene in BC, and p27 was one of its targets [446]. Eleven genes, such as *CDK1* and *HOXB13*, associated with puberty and BC were found to interact directly or indirectly with *YY1*. The other central gene, Cyclin D1, is also related to the cell cycle and is activated by *E2F8*, *SP1* and *STAT5*, which also interacts with *UHRF1*. Lastly, Vimentin is a part of the intermediate filament protein family and was found to be overexpressed in various cancers [447]. In BC, it is believed to play a role in tumour migration and invasion [448,449]. The network analyses on genes associated with puberty and BC have shown some general cell-cycle and developmental gene interactions, indicating the possible relevance of these genes to BC risk. However, further experimental and bioinformatic analyses should be done to clarify whether and how these genes are related to both BC and puberty.

Through literature search (**Table 33**) and the Human Protein Atlas database, we investigated the functions and expression patterns of genes mapped to CpG sites associated with puberty and BC that are found on gene regulatory sites or exons. Therefore, these genes represent the list of genes

whose expression or function, without claiming causality, could be associated with DNA methylation underlying the identified CpG sites. In the literature, a considerable number of genes were linked to both puberty and BC (**Table 33**); in particular, four of them (*UHRF1*, *CDK6*, *SP1* and *OXTR*) appear to be directly involved in the onset and progression of BC and have also been found to be directly or indirectly associated with puberty. *UHRF1* protein regulates gene expression through binding to certain DNA sequences and recruiting histone deacetylases [450]. It has an important role in the cell cycle, as it regulates topoisomerase IIα and retinoblastoma gene expression, both crucial to the cell cycle [343]. It is also involved in DNA damage checkpoints. *UHRF1* is believed to regulate breast cancer cell growth through oestrogen signalling [298] and was found to play a role in the development of invasive ductal BC [299]. Based on the Human Protein Atlas, *UHRF1* mRNA has the highest expression in the thymus and has a low expression in the breast. The single-cell data has shown that, within the breast tissue, *UHRF1* is most expressed in breast glandular cells. Finally, considering its low mRNA expression in the breast, the *UHRF1* protein is relatively abundant. Within the *D-CDK6* complex, *CDK6* is ubiquitously important for tumour initiation [313]. Additionally, it is important for mammary epithelial proliferation and BC initiation and maintenance [314]. The protein encoded by *SP1* is a transcription factor that binds to GC-rich motifs of numerous promoters and regulates cellular processes such as differentiation, cell growth, immune response, etc [451]. It regulates a plethora of cancer-associated genes [305] and is believed to control the proliferation of BC via interactions with insulin-like growth factors-I receptor (IGF-I) [304]. *SP1* mRNA is expressed in all tissue types and has an expression of around 25 TPM in breast tissue. It is expressed in numerous cell types within the breast, including glandular cells, adipocytes and fibroblasts. In blood, it has the highest expression in macrophages and monocytes. Finally, *OXTR* could be involved in BC development and progression [308,309], where one of the suggested mechanisms is through the prolactin/p-STAT5 pathway [310]. Interestingly, *OXTR* mRNA was low in all tissues but BC, where it was around 55 TPM. In the breast tissue, *OXTR* is mainly expressed in breast myoepithelial cells, while in peripheral blood it is mainly expressed in platelets.

The CpG sites, miRNAs and genes that have been identified in the context of puberty and breast cancer cannot be causally associated with the progression and onset of breast cancer. The main reason is that CpG sites have been identified in peripheral blood. Therefore, the potential effect of CpG sites and genes on the risk of BC has to be studied in the breast, endocrine system or other relevant tissues. Blood could still be used as a medium for the transport of miRNAs or as a medium where cancer cells export unwanted molecules. The identified CpG sites, miRNAs and genes are mainly biomarkers linking puberty and BC, and future investigations should clarify their function.

Conclusions

Breast cancer is the malignancy with the highest incidence and mortality among women. It is a disease with a significant healthcare burden, and due to its high incidence, the quality of life of many women is negatively affected. The problems of mortality, health burden and decreased quality of life related to BC can be mitigated through primary and secondary prevention. Breast self-examination and BC screening programs are the core secondary prevention methods. Nonetheless, mammography, which is the gold standard tool used in BC screening programs, has some disadvantages, such as a relatively higher false positive rate, especially in women with denser breasts, interval tumours, inflexible scheduling and radiation exposure.

To overcome the drawbacks of current BC screening, it is necessary to identify and implement non-invasive and cost-effective diagnostic biomarkers for early BC detection. In this project we first evaluated the diagnostic performance of circulating cell-free miRNAs in the thus far published studies, as they are one of the most studied types of non-invasive biomarkers. In our meta-analysis we presented reliable diagnostic performance estimates of cfc miRNAs and showed that they are promising biomarkers for (early) detection of BC. The subgroup analysis revealed that single miRNAs perform worse on average compared to multiple miRNA panels. In addition, differences in performance were also observed between models based on exogenous and endogenous RT-qPCR normalisers.

Using novel methods to evaluate model or author preference for sensitivity or specificity, we have determined that overall, the meta-analysed studies tend to prefer specificity. Additionally, the case–control ratio likely has an impact on diagnostic accuracy, while the preference for sensitivity or specificity influences the ratio of predicted positive to predicted negative screens. For this reason, we emphasised the importance of the authors disclosing their motivations for the research design and the possible implications they might have on the preference for sensitivity or specificity. Furthermore, we concluded that prospective random sampling of cases and controls, independent validation cohorts as well as standardisation of studies, especially on normalising methods, patient flow and specimen type, are paramount in achieving consistent and homogeneous results across studies. This would discover reliable miRNA candidate models for the diagnosis of BC that would have to be independently validated by several laboratories.

In the second part of the project, we identified non-invasive plasma biomarkers which could assist, together with non-molecular parameters, in early BC detection. This was done on discovery and validation case–control cohorts nested in a large screening cohort of BC, and we followed the conclusions from our meta-analysis to obtain reliable and reproducible circulating biomarkers. Seven miRNA ratios were identified as promising biomarkers for early BC detection that can be measured through a widespread and low-cost technique (RT-qPCR). The miRNAs showed a

certain degree of heterogeneity between the discovery and validation cohorts, but after recalibration, we demonstrated their potential as biomarkers for early diagnosis of BC. The miR-21-5p, also found to be the most promising miRNA in our meta-analysis, was found to make up one of the seven ratios, further demonstrating its potential as a BC diagnostic biomarker. The lifestyle score was among the non-molecular variables found to differentiate between cases and controls. Hence, in addition to optimising secondary prevention, it is vital to raise awareness and organise social initiatives that would improve people's lifestyles. This would contribute to the increase of overall quality of life and health among the population as well as decrease healthcare costs.

We identified the DNA methylation sites associated with puberty and BC, which are mapped to genes targeted significantly by two of the miRNAs among the seven-miRNA ratio signature. In addition, a subset of DNA methylation sites located in regulatory regions or exons was also identified, as they may be more likely to influence gene expression of the mapped genes. However, since they were identified in blood, the actual functional relevance of the identified CpG sites, genes and miRNAs for the onset and progression of BC has yet to be established.

Finally, considering the small sample size in this project, further evaluation and reconstruction of the model using the seven miRNA ratios on a much larger cohort is required. This would be a challenging task due to the extensive time required by the screening programme to obtain a sufficient number of events, as prospective samples are required. However, collaboration between hospitals and research centres could make this feasible.

Appendix

Appendix A

Additional file 1. QUADAS-2 tailored for diagnostic cfc miRNAs for early BC detection using RT-qPCR.

Title:

Phase 1: State the review question:

Patients: The study needs to report the type of sampling that was performed which optimally should be random or consecutive sampling of the patients/controls. In addition, information on whether the patients were matched to the controls or not needs to be disclosed. The institution responsible for giving out the ethical approval as well as the clinic or institution the samples were obtained from needs to be disclosed.

Index test(s): The index test(s) reported by the study must be circulating cell-free microRNAs. The expression of the index test(s) must be performed by RT-qPCR. In addition, a detailed protocol of RNA extraction, reverse transcription (and if performed, pre-amplification) must be described. Moreover, the RT-qPCR evaluation of the miRNAs requires a normalizing method, which needs to be disclosed by the study. The method of calculating and analysing the Delta Cts or other methods of quantifying the relative expression of miRNAs (i.e., miRNA ratios) need to be disclosed.

Reference standard and target condition: The primary target condition needs to include patients with malignant breast cancer. The reference standard is usually a histopathological analysis of the cancer tissue which is obtained by performing a biopsy. However, since the biopsy is invasive it is usually not performed on healthy controls. Nevertheless, a mammographic screening of the healthy individuals should be performed to have any kind of confirmation that the individual does not have breast cancer. The studies need to report whether the mammographic screening was performed on the healthy controls and which institution performed the histopathological analysis on included cases (or potential benign samples) and when with respect to the sampling (before or after). The histopathological analysis needs to confirm that the patient indeed has malignant breast cancer as well as the stage to which the cancer has progressed. This is relevant information to evaluating the index test(s) applicability to clinical setting.

Phase 2: Draw a flow diagram for the primary study

PDF of the hand-drawn flow diagram.

Phase 3: Risk of bias and applicability judgments

0. **DOMAIN 1: PATIENT SELECTION Risk of Bias**

Describe the methods of patient selection:

Signalling questions:

- Was random case/control selection from a prospective cohort performed? Yes/No/Unclear
- Did the study avoid inappropriate exclusions? Yes/No/Unclear

Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting)**:**

Is there concern that the included patients do not match the review question?
CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST(S)

0. **If more than one test was used, please complete for each test (in this case for each model). Risk of Bias**

Describe the index test and how it was conducted and interpreted:

Signalling questions:

- Was a validation cohort included in the study? Yes/No/Unclear
- If a threshold was used, was it pre-specified? Yes/No/Unclear
- Was a normalizing method utilized? Yes/No/Unclear
- Was the experiment methodologically sound? Yes/No/Unclear
- Was the performance of the index test(s) properly reported? Yes/No/Unclear

Could the conduct or interpretation of the index test have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability

Is there concern that the index test, its conduct, or interpretation differ from the review question? CONCERN: LOW/HIGH/UNCLEAR

0. **DOMAIN 3: REFERENCE STANDARD Risk of Bias**

Describe the reference standard and how it was conducted and interpreted:

Signalling questions:

- Is the reference standard likely to correctly classify the target condition? Yes/No/Unclear
- Were the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear
- Did all BC patients receive a reference standard? Yes/No/Unclear
- Did patients receive the same reference standard? Yes/No/Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias? RISK: LOW /HIGH/UNCLEAR

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question? CONCERN: LOW /HIGH/UNCLEAR

0. **DOMAIN 4: FLOW AND TIMING Risk of Bias**

Describe any patients who did not receive the index test(s) and/or reference standard or whowere excluded from the 2x2 table (refer to flow diagram):

Describe the time interval and any interventions between index test(s) and reference standard:

Signalling questions:

- Were patients who undergone treatment removed? Yes/No/Unclear
- Were all patients included in the analysis? Yes/No/Unclear
- Was the sampling performed before the biopsy? Yes/No/Unclear
- Was the sampling performed before the surgery? Yes/No/Unclear

Could the patient flow have introduced bias? RISK: LOW /HIGH/UNCLEAR

Supplementary Table 1. General information about the studies included in the meta-analysis.

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
Swellam et al. [452]	2019	Egypt	Serum	182 (39 + 47)	- miR-21	0.86
					- miR-126	1
					- miR-155	1
				111 (39 + 47)#	- miR-21	0.40/0.93
					- miR-126	0.76/1
					- miR-155	0.96/0.97
Zhang et al. [453]	2017	China	Whole Blood	28 (13)	- miR-30b-5p	0.93
					- miR-96-5p	0.77
					- miR-182-5p	0.76
					- miR-374b-5p	0.83
					- miR-942-5p	0.81
Mar-Aguilar et al. [454]	2013	Mexico	Serum	71 (10)	- miR-10b	0.95
					- miR-21	0.95
					- miR-125b	0.95
					- miR-145	0.98
					- miR-155	0.99
					- miR-191	0.79
					- miR-382	0.97
					- miR-145/miR-155/miR-382	0.99
Wu et al. [455]	2012	China	Serum	100 (50)	- miR-222-3p	0.67
Diansyah et al. [456]	2021	Indonesia	Plasma	42 (16)	- miR-21	0.92
Hosseini Mojahed et al. [457]	2020	Iran	Serum	72 (36)	- miR-155	0.89
Pena-Cano et al. [423]	2019	Mexico	Serum	100 (50)	- miR-195-5p	0.88
Kim et al. [458]	2020	South Korea	Plasma	60 (30)	- miR-202	0.95
Heydari et al. [422]	2018	Iran	Serum	80 (40)	- miR-140-3p	0.66
Motamedi et al. [437]	2019	Iran	Plasma	47 (24)	- miR-21	0.83
Swellam et al. [459]	2019	Egypt	Serum	150 (30 + 40)	- miR-17-5p	0.87
					- miR-155	0.99
					- miR-222-3p	0.86
				103 (30 + 40)#	- miR-17-5p	1/0.76
					- miR-155	0.94/0.94
					- miR-222-3p	1/0.79
Matamala et al. [460]	2015	Spain	Plasma	230 (116)	- miR-505-5p	0.72
					- miR-96-5p	0.72
					- miR-125b-5p	0.64
					- miR-21	0.61

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
Li et al. [401]	2019	China	Plasma	226 (113)	- let-7b-5p/miR-122-5p/ miR-146-5p/miR-210-3p/ miR-215-5p	0.97
Han et al. [461]	2017	China	Serum	120 (21)	- miR-21	0.79
				71 (21)	- miR-125b	0.56
				120 (21)	- miR-145	0.59
				70 (21)	- miR-155	0.75
				120 (21)	- miR-365	0.80
				70 (21)	- miR-21/miR-155	0.87
				70 (21)	- miR-21/miR-155/miR-365	0.92
				120 (21)	- miR-21/miR-365	0.87
Zhao et al. [462]	2010	USA	Plasma	30 (15)	- let-7c	0.78
				30 (15)	- miR-589	0.85
				20 (10)	- miR-425	0.83
				20 (10)	- let-7d	0.99
Pastor-Navarro et al. [463]	2020	Spain	Serum	90 (45)	- miR-21/miR-205	0.77
					- miR-21	0.77
					- miR-205	0.65
Si et al. [464]	2013	China	Serum	120 (20)	- miR-92a	0.92
					- miR-21	0.93
Freres et al. [363]	2015	Belgium	Plasma	196 (88)	- miR-16/let-7d/miR-103/miR-107/miR-148a/let-7i/miR-19b/miR-22*/	0.81
					- miR-16/let-7d/miR-103/ miR-181a/miR-107/miR-142-3p/miR-148a/let-7f-1/miR-199a-5p/miR-590-5p/miR-32	0.80
Schrauder et al. [465]	2012	Germany	Whole Blood	48 (24)	- miR-202	0.68
Ng et al. [466]	2013	China	Plasma	120 (50)	- miR-145/miR-451a	0.93
Li et al. [415]	2018	China	Plasma	292 (146)	- miR-106a-3p/miR-106a-5p/ miR-20b-5p/miR-92a-5p	0.83
			Serum	298 (148)	- miR-106a-5p/miR-19b-3p/ miR-20b-5p/miR-92a-3p	0.97
Shen et al. [467]	2014	USA	Serum	100 (50)	- miR-133a/miR-148b	0.86
Antolin et al. [468]	2015	Spain	Whole Blood	64 (20)	- miR-200c	0.85
				37 (20)	- miR-200c	0.82
Soleimanpour et al. [469]	2019	Iran	Plasma	60 (30)	- miR-21	0.99
					- miR-155	0.92
Nashtahossein i et al. [470]	2021	Iran	Serum	72 (38)	- miR-660-5p	0.77
				62 (38)#	- miR-660-5p	0.82
				72 (38)	- miR-210-3p	0.72
				62 (38)#	- miR-210-3p	0.65
Han et al. [471]	2020	China	Serum	182 (38)	- miR-1204	0.82
Chen et al. [472]	2016	USA	Plasma	102 (49)	- miR-21	0.61
					- miR-152	0.69

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
Yu et al. [417]	2018	China	Serum	160 (47)	- miR-21-5p/miR-21-3p/ miR-99a-5p	0.90
Zou et al. [414]	2021	China	Serum	246 (122)	- let-7b-5p/miR-106a-5p/ miR-16-5p/miR-19a-3p/miR-19b-3p/miR-20a-5p/miR-223-3p/miR-25-3p/miR-425-5p/miR-451a/miR-92a-3p/miR-93-5p	0.96
Fang et al. [362]	2019	China	Plasma	131 (38 + 40)	- miR-324-3p/miR-382-5p/ miR-21-3p/miR-324-3p/ miR-30a-5p/miR-30e-5p/ miR-221-3p/miR-324-3p	0.90
					- miR-324-3p/miR-382-5p/miR-21-3p/miR-324-3p/miR-30a-5p/miR-30e-5p/ miR-221-3p/miR-324-3p	0.82
An et al. [473]	2018	China	Serum	109 (24)	- miR-24	0.72
					- miR-103a	0.72
Hu et al. [474]	2012	China	Serum	152 (76)	- miR-16/miR-25/ miR-222/miR-324-5p	0.93
Zhang et al. [475]	2015	China	Serum	151 (93)	- miR-205	0.84
Eichelser et al. [416]	2013	Germany	Serum	160 (40)	- miR-34a	0.64
					- miR-93	0.70
					- miR-373	0.88
Wang et al. [476]	2018	China	Serum	102 (44)	- miR-130b-5p/miR-151a-5p/ miR-206/miR-222-3p	0.93
					- miR-130b-5p	0.73
					- miR-151a-5p	0.80
					- miR-206	0.86
					- miR-222-3p	0.89
Zhang et al. [477]	2017	China	Plasma	125 (50)	- miR-200c	0.56
					- miR-141	0.58
Feliciano et al. [478]	2020	Spain	Serum	80 (60)	- miR-125b/miR-29c/miR-16/ miR-1260/miR-451a	1/0.82
				188 (92)	- miR-125b/miR-29c/miR-16/ miR-1260/miR-451a	0.96/0.92
Ibrahim et al. [479]	2020	Egypt	Plasma	50 (20)	- miR-10b	0.73
					- miR-21-3p	0.78
					- miR-181a	0.70
					- miR-145	0.70
Swellam et al. [480]	2021	Egypt	Serum	94 (20 + 30)	- miR-27a	0.82/0.92
Jang et al. [481]	2021	South Korea	Plasma	136 (56)	- miR-1246	0.96
					- miR-206	0.94
					- miR-24	0.97
					- miR-373	0.94

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
					- miR-1246/miR-206	0.99
					- miR-1246/miR-206/miR-373	0.99
					- miR-1246/miR-206/ miR-24/miR-373	0.99
Guo et al. [482]	2020	China	Plasma	79 (40)	- miR-21	0.66
					- miR-1273g-3p	0.63
Huang et al. [413]	2018	China	Serum	235 (107)	- let-7a	0.68
					- miR-155	0.64
					- miR-574-5p	0.89
Ashirbkekov et al. [483]	2020	Kazakhstan	Plasma	68 (33)	- miR-16-5p	0.66
					- miR-210-3p	0.71
					- miR-222-3p	0.76
					- miR-29c-3p	0.74
					- miR-145-5p	0.93
					- miR-191-5p	0.90
					- miR-21	0.71
					- miR-145-5p/miR-191-5p	0.98
					- miR-145-5p/miR-21-5p	0.93
					- miR-191-5p/miR-21-5p	0.92
					- miR-145-5p/miR-191-5p/ miR-21-5p	0.98
Guo et al. [364]	2018	China	Serum	60 (30)	- miR-1915-3p	0.88
					- miR-455-3p	0.78
Cuk et al. [484]	2013	Germany	Plasma	180 (60)	- miR-127-3p	0.65
					- miR-148b	0.70
					- miR-376a	0.59
					- miR-376c	0.59
					- miR-409-3p	0.62
					- miR-652	0.75
					- miR-801	0.72
					- Panel of 7 miRs above	0.81
Raheem et al. [485]	2019	Iraq	Serum	60 (30)	- miR-34a	0.67
Zhu et al. [486]	2020	China	Serum	120 (60)	- miR-1908-3p	0.84
Ahmed Mohmmed et al. [487]	2021	Egypt	Serum	80 (30)	- miR-106a	0.95
Sadeghi et al. [488]	2021	Iran	Whole Blood	130 (60)	- miR-145	0.65/0.61
					- miR-106b-5p/miR-126-3p/ miR-140-3p/miR-193a-5p/ miR-10b-5p	0.79/0.86
Itani et al. [489]	2021	Lebanon	Plasma	73 (32)	- miR-21	0.76
					- miR-155	0.70
					- miR-23a	0.74
					- miR-130a	0.78
					- miR-145	0.81

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
					- miR-425-5p	0.83
					- miR-139-5p	0.83
					- miR-451	0.73
					- miR-145/miR-425-5p	0.83
					- miR-21/miR-23a	0.80
					- miR-21/miR-130a	0.82
					- miR-21/miR-23a/miR-130a	0.82
					-miR-145/miR-139-5p/mir-130a	0.96
					- miR-145/miR-139-5p/mir-130a /miR-425-5p	0.97
Mahmoud et al. [490]	2021	Egypt	Serum	95 (25)	- miR-185-5p	0.84
					- miR-301a-3p	0.90
Zou et al. [425]	2022	Multiple	Serum	374 (197)	- miR-133a-3p/miR-497-5p/mir-24-3p/miR-125b-5p/miR-377-3p/ miR-374c-5p/miR-324-5p/miR-19b-3p	0.92
				379 (199)	- miR-133a-3p/miR-497-5p/mir-24-3p/miR-125b-5p/miR-377-3p/ miR-374c-5p/miR-324-5p/miR-19b-3p	0.92
				325 (199)#	- miR-133a-3p/miR-497-5p/mir-24-3p/miR-125b-5p/miR-377-3p/ miR-374c-5p/miR-324-5p/miR-19b-3p	0.92
				210 (199)¶	- miR-133a-3p/miR-497-5p/mir-24-3p/miR-125b-5p/miR-377-3p/ miR-374c-5p/miR-324-5p/miR-19b-3p	0.95
Zou et al. [491]	2021	Singapore	Serum	369 (100+196)	- miR-451a/miR-195-5p/miR-126-5p/miR-423-3p/miR-192-5p/miR-17-5p	0.87
Li et al. [492]	2022	China	Serum	98 (49)	- miR-9-5p	0.85/0.94
					- miR-17-5p	0.71/0.65
					- miR-148a-3p	0.87/0.88
Shaker et al. [493]	2021	Egypt	Serum	450 (150+120)	- miR-29	0.92
					- miR-182	0.97
Uyisenga et al. [365]	2021	Rwanda	Plasma	45 (18)	- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-342-3p/miR-33a-5p/miR-130a-3p/let-7i-5p/miR-328-3p/miR-	0.87

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
					29b-3p/miR-146a-5p/miR-29a-3p/miR-126-3p	
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-5p/miR-130a-3p/miR-185-5p/let-7i-5p/miR-328-3p/miR-29b-3p/miR-146a-5p/miR-210-3p/miR-126-3p	0.87
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-5p/miR-130a-3p/let-7i-5p/miR-328-3p/miR-29b-3p/miR-210-3p/miR-126-3p	0.87
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-342-3p/miR-33a-5p/miR-130a-3p/let-7i-5p/miR-328-3p/miR-29b-3p/miR-146a-5p/miR-210-3p/miR-126-3p	0.87
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-5p/miR-130a-3p/miR-185-5p/let-7i-5p/miR-328-3p/miR-29b-3p/miR-146a-5p/miR-29a-3p/miR-126-3p	0.86
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-5p/miR-130a-3p/let-7i-5p/miR-328-3p/miR-29b-3p/miR-146a-5p/miR-210-3p/miR-126-3p	0.86
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-5p/miR-130a-3p/let-7i-5p/miR-29b-3p/miR-146a-5p/miR-210-3p/miR-126-3p	0.86
					- let-7a-5p/miR-150-5p/miR-940/miR-33a-5p/miR-130a-3p/miR-328-3p/miR-29a-3p/miR-126-3p	0.86
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-5p/miR-130a-3p/let-7i-5p/miR-328-3p/miR-29b-3p/miR-29a-3p/miR-126-3p	0.86
					- let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-33a-	0.86

Authors	Year	Country[§]	Source	Sample size (Healthy controls + benign)[†]	Index test (model)	Diagnostic Performance[‡]
					5p/let-7i-5p/miR-29b-3p/miR-146a-5p/miR-29a-3p/miR-126-3p - let-7a-5p/miR-150-5p/miR-940/miR-32-5p/miR-130a-3p/miR-185-5p/let-7i-5p/miR-29b-3p/miR-146a-5p/miR-126-3p	0.86
					- let-7a-5p/miR-150-5p/miR-940/miR-130a-3p/miR-328-3p/miR-29a-3p/miR-210-3p/miR-126-3p	0.86

§Country from which the cases and controls of the reported model were sampled.
†Sample size (cases, controls and benign) of the reported model.
‡For each reported model, its ROC AUC is shown. If not available, then the sensitivity and specificity pair are reported.
#Model with cases up to TNM stage II.
¶Model with TNM stage III and IV cases.

Supplementary Table 2. Summary of the bivariate analysis on all the reported models and its corresponding subgroup analyses. Subgroups marked with an asterix (*) do not have a large enough model sample size in order for the result to be reliable.

Subgroup		Fixed Effects		Random Effects					
				Model			Study		
		Estimates	CI	SD	Corr.	n	SD	Corr.	n
All models	Sen	0.85	[0.81, 0.88]	0.85	-0.17	146	0.70	0.06	46
	Spe	0.83	[0.79, 0.87]	0.60	-0.17	146	0.74	0.06	46
Plasma	Sen	0.83	[0.77, 0.87]	0.75	-0.09	64	0.49	-0.09	15
	Spe	0.85	[0.78, 0.91]	0.47	-0.09	64	0.86	-0.09	15
Serum	Sen	0.87	[0.81, 0.91]	0.94	-0.29	73	0.84	0.25	29
	Spe	0.83	[0.78, 0.87]	0.63	-0.29	73	0.71	0.25	29
Single miR panel	Sen	0.82	[0.77, 0.86]	0.80	-0.28	96	0.73	0.08	34
	Spe	0.83	[0.78, 0.87]	0.67	-0.28	96	0.75	0.08	34
Multiple miR panel	Sen	0.90	[0.86, 0.93]	0.55	0.09	50	0.65	0.11	20
	Spe	0.86	[0.80, 0.90]	0.35	0.09	50	0.77	0.11	20
Endogenous normaliser	Sen	0.82	[0.77, 0.86]	0.80	-0.28	96	0.73	0.08	34
	Spe	0.83	[0.78, 0.87]	0.67	-0.28	96	0.75	0.08	34
Exogenous normaliser*	Sen	0.82	[0.60, 0.93]	1.38	-0.63	9	0.51	-1	4
	Spe	0.76	[0.63, 0.86]	0.88	-0.63	9	0.13	-1	4
With stage III & IV cases	Sen	0.85	[0.80, 0.88]	0.85	-0.19	125	0.69	0.09	38
	Spe	0.84	[0.80, 0.88]	0.65	-0.19	125	0.75	0.09	38
Without stage III & IV cases	Sen	0.86	[0.77, 0.91]	0.95	0.04	21	0.55	-0.14	13
	Spe	0.82	[0.74, 0.88]	0.47	0.04	21	0.73	-0.14	13
With stage IV cases*	Sen	0.87	[0.61, 0.97]	0.98	0.22	17	1.35	-1	4
	Spe	0.86	[0.80, 0.90]	0.68	0.22	17	0.16	-1	4
Without stage IV cases	Sen	0.85	[0.81, 0.88]	0.83	-0.23	129	0.62	0.06	43
	Spe	0.84	[0.80, 0.88]	0.60	-0.23	129	0.76	0.06	43
miRNA-21-5p*	Sen	0.74	[0.64, 0.83]	0.52	1.00	10	0.46	-1	9
	Spe	0.81	[0.70, 0.89]	0.11	1.00	10	0.81	-1	9
QUADAS-2: > 3 "LOW"	Sen	0.82	[0.78, 0.86]	0.85	-0.11	109	0.51	-0.09	39
	Spe	0.82	[0.78, 0.86]	0.63	-0.11	109	0.76	-0.09	39
QUADAS-2: > 4 "LOW"	Sen	0.82	[0.77, 0.85]	0.83	-0.02	78	0.43	-0.08	29
	Spe	0.80	[0.74, 0.85]	0.56	-0.02	78	0.78	-0.08	29
QUADAS-2: > 5 "LOW"	Sen	0.79	[0.73, 0.84]	0.78	-0.13	41	0.43	-0.11	20
	Spe	0.77	[0.69, 0.83]	0.58	-0.13	41	0.69	-0.11	20

Supplementary Table 3. Summary of the bivariate analysis on the most important model of each study and its corresponding subgroup analyses. Subgroups marked with an asterix (*) do not have a large enough model sample size in order for the result to be reliable.

Subgroup		Fixed Effects		Random Effects Model		
		Estimates	CI	SD	Corr.	n
Most important models	Sen	0.88	[0.85, 0.91]	0.86	0.23	46
	Spe	0.88	[0.84, 0.91]	1.00	0.23	46
Plasma	Sen	0.89	[0.83, 0.93]	0.84	0.04	14
	Spe	0.90	[0.82, 0.95]	1.16	0.04	14
Serum	Sen	0.87	[0.83, 0.91]	0.91	0.31	29
	Spe	0.86	[0.81, 0.90]	0.96	0.31	29
Single miR panel	Sen	0.85	[0.80, 0.89]	0.76	0.14	26
	Spe	0.87	[0.80, 0.92]	1.12	0.14	26
Multiple miR panel	Sen	0.90	[0.86, 0.94]	0.89	0.28	20
	Spe	0.89	[0.83, 0.92]	0.87	0.28	20
Endogenous normaliser	Sen	0.85	[0.80, 0.89]	0.76	0.14	26
	Spe	0.87	[0.80, 0.92]	1.12	0.14	26
Exogenous normaliser*	Sen	0.93	[0.81, 0.97]	0.87	1	3
	Spe	0.75	[0.63, 0.84]	0.45	1	3
With stage III & IV cases	Sen	0.88	[0.84, 0.91]	0.90	0.27	37
	Spe	0.89	[0.85, 0.92]	1.02	0.27	37
Without stage III & IV cases	Sen	0.88	[0.81, 0.92]	0.65	0.12	9
	Spe	0.77	[0.67, 0.84]	0.63	0.12	9
With stage IV cases	Sen	0.87	[0.63, 0.97]	1.37	0.52	4
	Spe	0.89	[0.71, 0.96]	1.10	0.52	4
Without stage IV cases	Sen	0.88	[0.85, 0.91]	0.79	0.19	42
	Spe	0.88	[0.83, 0.91]	0.99	0.19	42
miRNA-21-5p	Sen	0.75	[0.66, 0.83]	0.58	-0.43	9
	Spe	0.81	[0.70, 0.89]	0.83	-0.43	9
QUADAS-2: > 3 "LOW"	Sen	0.86	[0.82, 0.89]	0.77	0.13	39
	Spe	0.86	[0.81, 0.90]	1.06	0.13	39
QUADAS-2: > 4 "LOW"	Sen	0.85	[0.80, 0.88]	0.77	0.18	29
	Spe	0.84	[0.77, 0.89]	1.04	0.18	29
QUADAS-2: > 5 "LOW"	Sen	0.84	[0.78, 0.89]	0.84	0.1	20
	Spe	0.80	[0.73, 0.86]	0.91	0.1	20

Appendix B

Supplementary Table 4. Demographic, family, reproductive and screening history, lifestyle, anthropometric measurements, education, breast density and PRS are reported for the discovery cohort. Results of univariate logistic regression for each variable are also reported.

	Cases (n = 65)		Controls (n = 66)		Cases vs controls	P
	Mean/N	SD/%	Mean/N	SD/%	OR [95% CI]	
Age at enrolment (years)						
Mean ± SD	59.15	6.00	57.82	5.92	1.04 [0.98, 1.10]	0.201
Centre						
Biella	16	24.62	20	30.3	1 (ref)	
Torino	49	75.38	46	69.7	1.33 [0.62, 2.91]	0.467
Previous negative second-level screening rounds						
0	59	90.77	62	93.94	1 (ref)	
≥ 1	6	9.23	4	6.06	1.58 [0.43, 6.43]	0.497
Previous benign biopsies						
0	49	77.78	57	86.36	1 (ref)	
≥ 1	14	22.22	9	13.64	1.81 [0.73, 4.69]	0.207
Missing	2					
Education						
Low	22	34.38	21	32.31	1 (ref)	
Medium	28	43.75	31	47.69	0.86 [0.39, 1.90]	0.712
High	14	21.88	13	20	1.03 [0.39, 2.71]	0.955
Missing	1		1			
Nr. of first-degree relatives with BC						
0	56	88.89	58	87.88	1 (ref)	
≥ 1	5	7.94	8	12.12	0.65 [0.19, 2.06]	0.469
Missing	2					
Age at menarche (years)						
≤ 11	19	29.69	22	33.33	1 (ref)	
12–13	33	51.56	33	50	1.16 [0.53, 2.54]	0.713
≥ 14	12	18.75	11	16.67	1.26 [0.45, 3.55]	0.654
Missing	1					
Age at first full pregnancy (years)						
Nulliparous	11	16.92	19	28.79	0.39 [0.13, 1.13]	0.087
≤ 19	1	1.54	3	4.55	0.22 [0.01, 2.02]	0.219
20–24	15	23.08	11	16.67	1 (ref)	
25–29	16	24.62	21	31.82	0.51 [0.18, 1.41]	0.198
≥ 30	22	33.85	12	18.18	1.39 [0.47, 4.13]	0.545
Contraceptive therapy						
No OR use < 1 year	31	48.44	29	45.31	1 (ref)	
1–4 years	5	7.81	10	15.62	0.47 [0.13, 1.48]	0.210
≥ 5 years	28	43.75	25	39.06	1.04 [0.50, 2.20]	0.902
Missing	1		2			
Breastfeeding						
Nulliparous OR no breastf. OR breastf. < 6 months	39	60.94	42	63.64	1 (ref)	
≥ 6 months	25	39.06	24	36.36	1.12 [0.55, 2.29]	0.751

	Cases (n = 65) Mean/N	SD/%	Controls (n = 66) Mean/N	SD/%	Cases vs controls OR [95% CI]	P
Missing	1		0			
Menopausal status						
Not in menopause	12	18.75	12	18.18	1 (ref)	
Menopause	52	81.25	54	81.82	0.74 [0.30, 1.78]	0.504
Missing	1		0			
HRT use						
Not in menopause	12	18.75	12	18.18	1 (ref)	
No HRT use OR HRT use < 1 year	45	70.31	43	65.15	1.04 [0.42, 2.60]	0.921
≥ 1 year	7	10.94	11	16.17	0.63 [0.18, 2.18]	0.475
Missing	1		0			
Measured BMI (kg/m^2)						
Mean ± SD	28.02	6.24	25.76	5.15	1.07 [1.01, 1.15]	0.029
Waist circumference (cm)						
Mean ± SD	92.69	17.55	88.00	11.83	1.02 [1.00, 1.05]	0.087
Missing	2		2			
Level of occupational physical activity at age 30–39 years						
Exclusively/mainly sitting	23	35.9	17	25.8	1 (ref)	
Standing or average	34	53.1	43	65.2	0.58 [0.27, 1.26]	0.172
Heavy or very heavy	7	10.9	6	9.1	0.86 [0.24, 3.12]	0.817
Missing	1		0			
Level of leisure time physical activity at 30–39 years						
< 2 h/week	34	53.1	35	53	1 (ref)	
≥ 2 h/week	30	46.9	31	47	1.00 [0.50, 1.99]	0.991
Missing	1		0			
Alcohol habit						
Never or ex drinker	19	29.69	16	24.24	1 (ref)	
Drinker (incl. occasionally)	45	70.31	50	75.76	0.76 [0.35, 1.65]	0.485
Missing	1		0			
Smoking habit						
Never smoker	26	41.94	31	47.69	1 (ref)	
Ex-smoker	25	40.32	19	29.23	1.57 [0.71, 3.50]	0.265
Occasionally/Smoker	11	17.74	15	23.08	0.87 [0.34, 2.22]	0.779
Missing	3		1			
BI-RADS breast density						
1	21	32.31	21	31.82	1 (ref)	
2	27	41.54	34	51.52	0.79 [0.36, 1.75]	0.566
3 or 4	17	26.15	11	16.67	1.55 [0.59, 4.15]	0.379
Tabar breast density						
1	10	15.38	22	33.33	1 (ref)	
2	23	35.38	25	37.88	2.02 [0.80, 5.32]	0.141
3	7	10.77	6	9.09	2.57 [0.69, 10.01]	0.162
4 or 5	25	38.46	13	19.7	4.23 [1.59, 11.97]	0.005
WCRF/AICR lifestyle score						
Mean ± SD	5.12	1.11	5.52	0.98	0.68 [0.47, 0.96]	0.034
PRS						
Mean ± SD	1.00	0.41	0.98	0.43	1.09 [0.47, 2.51]	0.842
Missing	0		4			

Supplementary Table 5. Histological and molecular subtype characteristics of invasive and in situ breast cancer cases of the discovery cohort.

Invasive (n = 57)			In situ (n = 8)		
	N	%		N	%
Histotype			**Histotype**		
Ductal NOS	30	58.82	Ductal NOS	3	37.5
Lobular	8	15.69	Solid	1	12.5
Tubular	4	7.84	Micropapillary	1	12.5
Other	9	17.65	Papillary	1	12.5
Missing	6		Other	2	25
Grade			**Grade**		
I	18	36.73	I	2	25
II	25	51.02	II	2	25
III	6	12.24	III	4	50
Missing	8		**Tumour size (mm)**		
pT			1–10	4	50
1a-1b-1mic	25	46.3	11–20	2	25
1c	24	44.44	21 +	2	25
2 +	5	9.26			
Missing	3				
Tumour size (mm)					
1–10	25	46.3			
11–20	24	44.44			
21 +	5	9.26			
Missing	3				
Stage					
IA	42	87.5			
IIA	3	6.25			
IIIC	2	4.17			
IV	1	2.08			
Missing	9				
Molecular subtypes					
ER					
Negative	8	15.09			
Positive (> 10%)	45	84.91			
Missing or undetermined	4				
PgR					
Negative	17	32.08			
Positive (> 10%)	36	69.92			
Missing or undetermined	4				
Her2					
Negative	45	86.54			
Positive	7	13.46			
Missing or undetermined	5				
Ki-67					
Negative	39	76.47			
Positive (> 20%)	12	23.53			
Missing or undetermined	6				
Intrinsic subtype					

	Invasive (n = 57)		In situ (n = 8)	
	N	%	N	%
Luminal A-like	27	52.94		
Luminal B-like (HER2 negative)	13	25.49		
Luminal B-like (HER2 positive)	5	9.8		
HER2 positive (non-luminal)	2	3.92		
Triple negative	4	7.84		
Missing	6			

Supplementary Table 6. Demographic, family, reproductive and screening history, lifestyle, anthropometric measurements, education and breast density are reported for the validation cohort. Results of univariate logistic regression for each variable are also reported.

	Cases (n = 32)		Controls (n = 127)		Cases vs. controls	
	Mean/N	SD/%	Mean/N	SD/%	OR [95% CI]	P
Age at enrolment (years)	64.7	6.33	63.11	5.89	1.04 [0.98, 1.12]	0.193
Previous benign biopsies						
0	24	75	118	92.91	1 (Ref)	
≥ 1	6	18.75	9	7.09	3.28 [1.02, 9.98]	0.038
Missing	2	6.25	0			
Education						
Low	13	40.63	55	45.08	1 (Ref)	
Medium	12	37.5	55	45.08	0.92 [0.38, 2.21]	0.857
High	2	6.25	12	9.84	0.71 [0.10, 3.02]	0.671
Missing	5	15.63	5			
Nr. of first-degree relatives with BC						
0	25	78.13	106	83.46	1 (Ref)	
≥ 1	7	21.88	21	16.54	1.41 [0.51, 3.57]	0.480
Missing	0	0	0			
Age at menarche (years)						
≤ 11	7	21.88	27	21.26	1 (Ref)	
12-13	16	50	66	51.97	1.02 [0.34, 3.19]	0.971
≥ 14	9	28.13	34	26.77	0.94 [0.36, 2.66]	0.895
Missing	0	0	0			
Age at first full pregnancy (years)						
Nulliparous	12	37.5	25	19.69	1.82 [0.69, 4.95]	0.229
≥ 19	2	6.25	12	9.45	0.63 [0.09, 2.85]	0.588
20-24	10	31.25	38	29.92	1 (Ref)	
25-29	3	9.38	29	22.83	0.39 [0.08, 1.42]	0.184
≥ 30	5	15.63	23	18.11	0.83 [0.23, 2.64]	0.753
Missing	0	0	0			
Contraceptive use						
None OR < 1 year	27	84.38	100	80.65	1 (Ref)	
1-4 years	0	0	0	0		
≥ 5 years	3	9.38	24	19.35	0.46 [0.10, 1.46]	0.236
Missing	2	6.25	3			
Breastfeeding						
Nulliparous OR no breastf. OR breastf. < 6 months	28	87.5	78	61.42	1 (Ref)	

	Cases (n = 32)		Controls (n = 127)		Cases vs. controls	
	Mean/N	SD/%	Mean/N	SD/%	OR [95% CI]	P
≥ 6 months	4	12.5	49	38.58	0.23 [0.06, 0.62]	0.009
Missing	0	0	0			
Menopausal status						
Not in menopause	8	25	22	17.32	1 (Ref)	
Menopause	24	75	105	82.68	0.63 [0.26, 1.66]	0.324
Missing	0	0	0			
HRT						
Not in menopause	8	25	22	17.32	1 (Ref)	
No HRT use OR HRT use < 1 year	19	59.38	93	73.23	0.56 [0.22, 1.51]	0.233
≥ 1 year	3	9.38	12	9.45	0.69 [0.13, 2.90]	0.625
Missing	2	6.25	0			
Measured BMI (kg/m2)	26.3	4.31	25.06	4.88	1.05 [0.97, 1.14]	0.194
Waist circumference (cm)	89.8	12.1	84.82	12.35	1.03 [1.00, 1.06]	0.047
Level of occupational physical activity at age 30–39 years						
Exclusively/mainly sitting	10	31.25	33	25.98	1 (Ref)	
Standing or average	13	40.63	57	44.88	0.75 [0.30, 1.94]	0.549
Heavy or very heavy	7	21.88	37	29.13	0.62 [0.21, 1.81]	0.390
Missing	2	6.25	0			
Level of leisure time physical activity at 30–39 years						
< 2 h/week	16	50	67	52.76	1 (Ref)	
≥ 2 h/week	14	43.75	60	47.24	0.98 [0.44, 2.17]	0.955
Missing	2	6.25	0			
Alcohol habit						
Never drinker or ex drinker	6	18.75	39	30.71	1 (Ref)	
Drinker, also occasionally	24	75	88	69.29	1.77 [0.71, 5.09]	0.248
Missing	2	6.25	0			
Smoking habit						
Never smoker	17	53.13	77	60.63	1 (Ref)	
Ex-smoker	6	18.75	26	20.47	1.05 [0.35, 2.82]	0.933
Occasionally/Smoker	9	28.13	24	18.90	1.7 [0.65, 4.25]	0.264
Missing	0	0	0			
TABAR breast density						
1	4	12.5	50	39.37	1 (Ref)	
2	14	43.75	63	49.61	2.78 [0.93, 10.27]	0.087
3	12	37.5	9	7.09	16.67 [4.72, 71.36]	<0.001
4 or 5	2	6.25	5	3.94	5 [0.59, 33.66]	0.102
Missing	0	0	0			
WCRF lifestyle score	5.08	1.34	5.37	1.19	0.83 [0.61, 1.13]	0.229

Supplementary Table 7. Histological and molecular subtype characteristics of invasive and in situ breast cancer cases of the validation cohort.

Invasive (n = 31)			In situ (n = 1)		
	N	**(%)**		**N**	**(%)**
Histotype			**Histotype**		
Ductal NOS	21	75	Ductal NOS	0	0
Lobular	4	14.29	Solid	1	100
Tubular	0	0	Micropapillary	0	0
Other	3	10.71	Papillary	0	0
Missing	3		Other	0	0
Grade			**Grade**		
I	3	10.34	I	0	0
II	16	55.17	II	1	100
III	10	34.48	III	0	0
Missing	2		**Tumour size (mm)**		
pT			1–10	0	0
1a-1b-1mic	7	25.00	11–20	1	100
1c	12	42.86	21 +	0	0
2 +	9	32.14			
Missing	3				
Tumour size (mm)					
1–10	7	25.93			
11–20	12	44.44			
21 +	8	29.63			
Missing	4				
Stage					
IA	14	50			
IIA	6	21.43			
IIB	4	14.29			
IIIA	3	10.71			
IIIC	1	3.57			
IV	0	0			
Missing	3				
Molecular subtypes					
ER					
Negative	4	14.81			
Positive (> 10%)	23	85.19			
Missing or undetermined	4				
PgR					
Negative	6	22.22			
Positive (> 10%)	21	77.78			
Missing or undetermined	4				
Her2					
Negative	23	88.46			
Positive	3	11.54			
Missing or undetermined	5				
Ki-67					
Negative	5	17.86			
Positive (> 20%)	23	82.14			

	Invasive (n = 31)		In situ (n = 1)	
	N	(%)	N	(%)
Missing or undetermined	3			
Intrinsic subtype				
Luminal A-like	4	17.39		
Luminal B-like (HER2 negative)	15	65.22		
Luminal B-like (HER2 positive)	2	8.7		
HER2 positive (non-luminal)	0	0		
Triple negative	2	8.7		
Missing	8			

Supplementary Table 8. Differentially expressed miRNAs between BC cases and controls in plasma based on small-RNA sequencing (n = 131).

miRNA	Base mean	log_2 FC	SE (log_2 FC)	p-adjusted
hsa-let-7f-1_hsa-let-7f-5p	25.57	-0.86	0.14	4.07E-08
hsa-let-7f-2_hsa-let-7f-5p	27.43	-0.87	0.14	6.73E-08
hsa-let-7a-2_hsa-let-7a-5p	24.50	-0.80	0.13	8.49E-08
hsa-let-7a-3_hsa-let-7a-5p	24.98	-0.68	0.13	3.69E-06
hsa-mir-22_hsa-miR-22-3p	67.71	0.55	0.11	1.38E-05
hsa-mir-423_hsa-miR-423-5p	47.52	0.48	0.10	1.80E-05
hsa-mir-3184_hsa-miR-3184-3p	47.52	0.48	0.10	1.80E-05
hsa-let-7a-1_hsa-let-7a-5p	25.81	-0.64	0.13	1.81E-05
hsa-let-7g_hsa-let-7g-5p	60.93	-0.47	0.11	2.18E-04
hsa-mir-3591_hsa-miR-3591-3p	200.89	0.77	0.19	3.78E-04
hsa-mir-122_hsa-miR-122-5p	200.92	0.77	0.19	3.78E-04
hsa-mir-21_hsa-miR-21-5p	162.05	-0.42	0.11	5.34E-04
hsa-let-7b_hsa-let-7b-5p	80.50	-0.35	0.09	7.15E-04
hsa-mir-320a_hsa-miR-320a	65.74	0.50	0.13	7.53E-04
hsa-mir-20a_hsa-miR-20a-5p	115.48	-0.36	0.09	8.17E-04
hsa-mir-7-1_hsa-miR-7-5p	31.87	0.72	0.19	1.04E-03
hsa-mir-26a-1_hsa-miR-26a-5p	28.60	-0.60	0.16	1.25E-03
hsa-mir-7-3_hsa-miR-7-5p	32.76	0.71	0.20	1.74E-03
hsa-mir-221_hsa-miR-221-3p	44.42	0.40	0.11	2.21E-03
hsa-mir-26b_hsa-miR-26b-5p	14.92	-0.46	0.13	3.40E-03
hsa-let-7d_hsa-let-7d-5p	20.67	-0.44	0.13	4.01E-03
hsa-mir-7-2_hsa-miR-7-5p	36.10	0.69	0.21	4.38E-03
hsa-mir-3529_hsa-miR-3529-3p	36.10	0.69	0.21	4.38E-03
hsa-mir-339_hsa-miR-339-5p	14.86	0.52	0.16	5.27E-03
hsa-mir-26a-2_hsa-miR-26a-5p	32.18	-0.49	0.15	5.53E-03
hsa-mir-146a_hsa-miR-146a-5p	36.46	0.36	0.11	5.78E-03
hsa-mir-126_hsa-miR-126-5p	148.36	-0.34	0.11	6.50E-03

Supplementary Table 9. Results of univariable logistic regression and AUCs performed on the 21 miRNA ratios based on discovery cohort NGS data.

miRNA ratio	Cases		Controls		Univariate LR			
	Median	IQ range	Median	IQ range	OR	95% CI	P	AUC
let-7a-5p-2_miR-106b-5p	-1.59	[-2.32, -1.26]	-0.86	[-1.22, -0.38]	0.39	[0.24, 0.59]	< 0.001	0.77
let-7a-5p-2_miR-22-3p	-2.00	[-2.92, -1.29]	-0.76	[-1.36, -0.03]	0.38	[0.24, 0.54]	< 0.001	0.81
let-7a-5p-2_miR-320a	-2.00	[-2.89, -1.04]	-0.61	[-1.02, -0.08]	0.45	[0.31, 0.62]	< 0.001	0.80
let-7b-5p_miR-19b-3p-1	-2.22	[-2.92, -1.77]	-1.70	[-2.03, -1.38]	0.33	[0.18, 0.55]	< 0.001	0.74
let-7f-5p-1_miR-103-1	-1.59	[-2.12, -1.11]	-0.81	[-1.33, -0.50]	0.40	[0.24, 0.61]	< 0.001	0.76
let-7f-5p-1_miR-19b-3p-1	-4.30	[-4.89, -3.76]	-3.20	[-3.78, -2.75]	0.41	[0.26, 0.59]	< 0.001	0.79
let-7f-5p-1_miR-103-2	-1.53	[-2.32, -1.08]	-0.76	[-1.19, -0.42]	0.38	[0.23, 0.58]	< 0.001	0.76
let-7f-5p-2_miR-146a-5p	-1.00	[-2.00, -0.07]	0.15	[-0.35, 0.74]	0.40	[0.27, 0.57]	< 0.001	0.78
miR-101-3p-2_miR-19b-3p-1	-2.39	[-2.76, -2.19]	-2.11	[-2.46, -1.82]	0.25	[0.11, 0.51]	< 0.001	0.68
miR-15a-5p_miR-16-5p-2	-1.33	[-1.87, -0.54]	-0.72	[-1.19, -0.20]	0.55	[0.36, 0.81]	0.004	0.66
miR-15b-5p_miR-16-5p-1	-2.22	[-3.43, -1.37]	-0.95	[-2.10, -0.35]	0.65	[0.50, 0.81]	< 0.001	0.71
miR-199a-3p-2_let-7a-5p-2	-0.42	[-1.34, 0.00]	-1.59	[-2.31, -0.79]	2.50	[1.72, 3.81]	< 0.001	0.75
miR-199a-3p-2_let-7f-5p-2	-0.60	[-1.52, 0.00]	-1.81	[-2.48, -1.03]	2.48	[1.72, 3.72]	< 0.001	0.76
miR-20a-5p_miR-19b-3p-1	-1.71	[-2.40, -1.45]	-1.15	[-1.42, -0.92]	0.21	[0.10, 0.40]	< 0.001	0.77
miR-21-5p_miR-23a-3p	-0.57	[-0.87, -0.24]	-0.14	[-0.65, 0.15]	0.33	[0.16, 0.63]	0.001	0.68
miR-22-3p_miR-19b-3p-2	-2.09	[-2.59, -1.66]	-2.58	[-3.01, -2.13]	2.69	[1.57, 4.88]	< 0.001	0.67
miR-26b-5p_miR-142-5p	-0.55	[-1.00, 0]	0.02	[-0.59, 0.46]	0.55	[0.35, 0.81]	0.004	0.68
miR-27a-3p_miR-122-5p	-2.54	[-3.48, -1.45]	-1.34	[-2.21, -0.73]	0.59	[0.43, 0.77]	< 0.001	0.71
miR-30d-5p_miR-20a-5p	-0.45	[-1.00, -0.09]	-0.88	[-1.32, -0.50]	2.37	[1.32, 4.46]	0.005	0.68
miR-335-5p_let-7f-5p-2	-1.00	[-1.97, 0.00]	-2.32	[-2.85, -1.41]	2.31	[1.66, 3.36]	< 0.001	0.76
miR-93-5p_miR-19b-3p-1	-3.84	[-4.15, -3.04]	-3.16	[-3.51, -2.84]	0.39	[0.22, 0.64]	< 0.001	0.69

IQ: Interquartile

Supplementary Table 10. Results of univariable logistic regression and AUCs performed on the 20 miRNA ratios based on discovery cohort RT-qPCR data.

miRNA ratio	Cases		Controls		Univariate LR			
	Median	IQ range	Median	IQ range	OR	95% CI	P	AUC
let-7a-5p_miR-106b-5p	-0.17	[-0.68, 0.38]	0.07	[-0.36, 0.70]	0.78	[0.50, 1.20]	0.270	0.59
let-7a-5p_miR-22-3p	5.48	[3.54, 7.81]	6.94	[4.08, 9.89]	0.85	[0.73, 0.98]	0.026	0.63
let-7a-5p_miR-320a	-4.30	[-4.94, -3.49]	-4.08	[-4.63, -3.16]	0.88	[0.65, 1.17]	0.399	0.57
let-7b-5p_miR-19b-3p	-3.19	[-3.61, -2.69]	-2.83	[-3.38, -2.54]	0.74	[0.47, 1.13]	0.176	0.59
let-7f-5p_miR-103a-3p	-1.23	[-1.88, -0.79]	-1.19	[-1.66, -0.82]	0.86	[0.61, 1.20]	0.390	0.53
let-7f-5p_miR-146a-5p	-7.82	[-8.60, -7.42]	-7.57	[-8.41, -6.71]	0.83	[0.62, 1.10]	0.210	0.59
let-7f-5p_miR-19b-3p	-7.79	[-8.36, -7.08]	-7.49	[-8.16, -6.66]	0.87	[0.64, 1.16]	0.355	0.57
miR-101-3p_miR-19b-3p	-8.09	[-8.63, -7.65]	-8.17	[-8.64, -7.69]	0.99	[0.70, 1.41]	0.969	0.49
miR-15a-5p_miR-16-5p	-19.10	[-20.43, -16.51]	-17.41	[-18.50, -16.96]	0.82	[0.65, 1.02]	0.087	0.63
miR-15b-5p_miR-16-5p	-10.13	[-10.92, -8.80]	-10.12	[-10.88, -9.32]	1.12	[0.92, 1.38]	0.276	0.52
miR-199a-3p_let-7a-5p	1.66	[1.08, 2.21]	1.24	[0.92, 1.90]	1.64	[1.05, 2.63]	0.033	0.61
miR-199a-3p_let-7f-5p	3.65	[2.97, 4.15]	3.2	[2.74, 3.93]	1.41	[0.97, 2.08]	0.077	0.58
miR-20a-5p_miR-19b-3p	-0.50	[-0.69, -0.14]	-0.48	[-0.76, -0.21]	0.78	[0.45, 1.28]	0.351	0.48
miR-21-5p_miR-23a-3p	4.34	[3.93, 4.91]	4.30	[3.80, 4.95]	1.02	[0.66, 1.58]	0.917	0.52
miR-22-3p_miR-19b-3p	-11.04	[-13.85, -9.74]	-12.53	[-15.49, -10.03]	1.12	[0.98, 1.30]	0.113	0.6
miR-26b_miR-142-5p	5.84	[5.08, 6.08]	6.05	[5.61, 6.44]	0.48	[0.28, 0.77]	0.005	0.65
miR-27a-3p_miR-122-5p	0.78	[-0.75, 2.06	0.65	[-0.80, 1.48]	1.12	[0.93, 1.36]	0.241	0.54
miR-30d-5p_miR-20a-5p	-4.27	[-4.79, -4.04]	-4.29	[-4.62, -4.03]	0.97	[0.63, 1.50]	0.905	0.5
miR-335-5p_let-7f-5p	1.44	[0.78, 2.27]	1.03	[0.46, 2.03]	1.18	[0.86, 1.64]	0.319	0.58
miR-93-5p_miR-19b-3p	-3.07	[-3.45, -2.79]	-3.29	[-3.52, -3.04]	2.05	[1.00, 4.50]	0.059	0.61

IQ: Interquartile

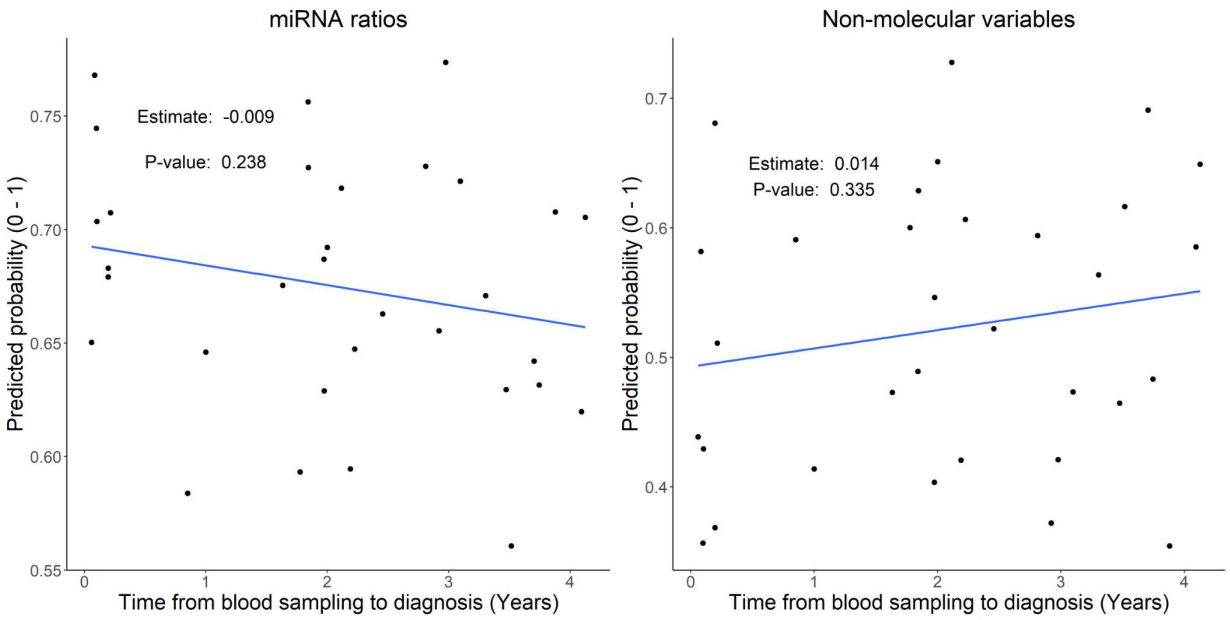
Supplementary Table 11. Univariate logistic regression results on the seven miRNA ratios analysed in the validation cohort. Interquartile ranges and medians of the ratios stratified by BC status were also reported.

	Cases		Controls		Univariate LR		
miRNA ratio	**Median**	**IQ range**	**Median**	**IQ range**	**OR**	**95% CI**	**P**
let-7a-5p_miR-22-3p	1.91	[1.47, 2.18]	1.68	[1.40, 1.89]	1.17	[1.03, 1.33]	0.016
let-7a-5p_miR-19b-3p	6.62	[5.99, 6.96]	6.42	[5.97, 6.72]	0.52	[0.30, 0.87]	0.016
miR-199a-3p_let-7a-5p	-5.86	[-6.22, -5.52]	-5.48	[-5.91, -5.11]	2.96	[1.23, 7.57]	0.019
miR-21-5p_miR-23a-5p	-8.37	[-8.79, -8.00]	-8.44	[-8.80, -8.20]	1.98	[1.03, 3.92]	0.044
miR-93-5p_miR-19b-3p	-2.78	[-2.96, -2.61]	-2.71	[-2.87, -2.59]	0.39	[0.07, 2.19]	0.286
miR-26b_miR-142-5p	4.75	[1.49, 7.18]	2.68	[0.86, 4.43]	1.26	[0.64, 2.53]	0.504
miR-101-3p_miR-19b-3p	3.89	[3.50, 4.52]	3.62	[3.27, 3.96]	1.18	[0.56, 2.66]	0.671

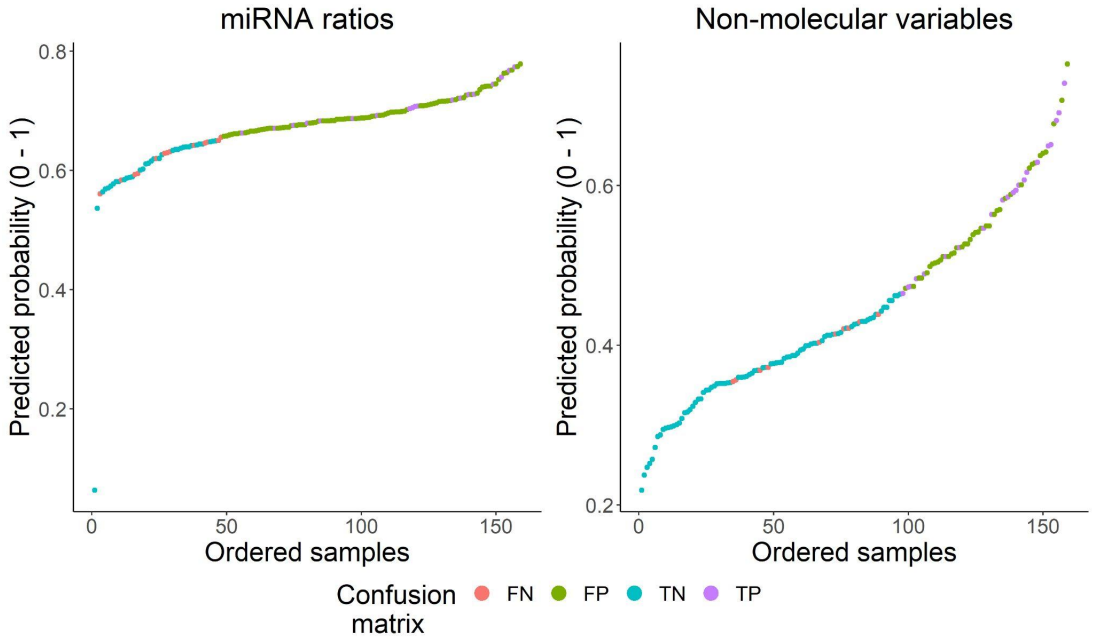
IQ: Interquartile

Supplementary Table 12. Testing the distribution and variance differences on the 12 predictors analysed in the validation cohort between controls which underwent a biopsy due to a suspicious mammography result and controls with a negative mammography result.

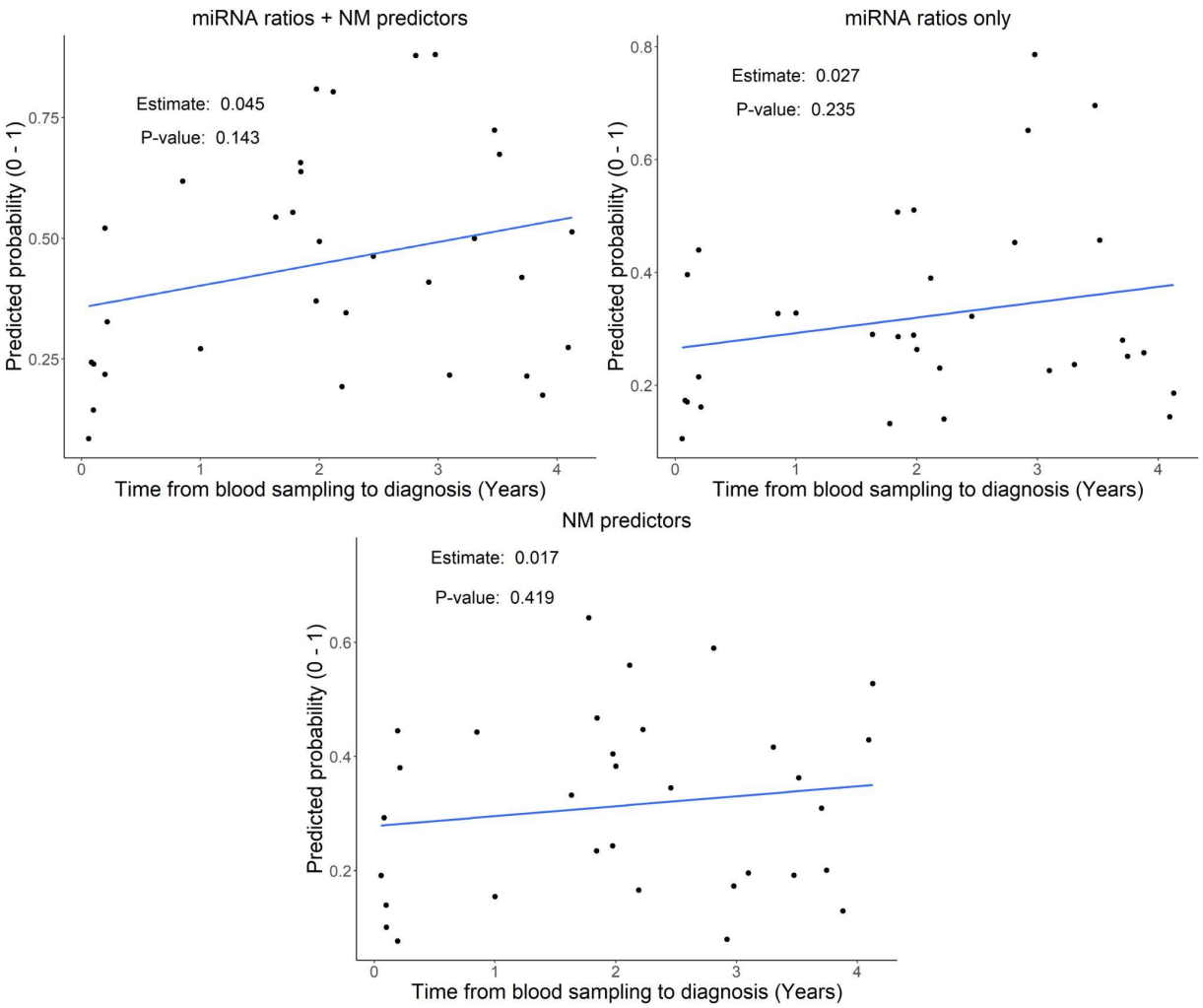
Predictor	**F-statistic**	**F (p-value)**	**W-statistic**	**W (p-value)**	**D-test**	**D (p-value)**
BMI*MS	1.78	0.074	1476	0.948	0.16	0.477
Breast density (TABAR)	1.20	0.577	1380	0.504	0.09	0.507
Centred BMI	1.64	0.124	1416	0.686	0.18	0.377
let-7a-5p_miR-22-3p	0.39	0.001	1689	0.260	0.17	0.469
let-7a-5p_miR-19b-3p	0.21	0.000	1347	0.430	0.16	0.549
Menopausal status	0.74	0.271	1592	0.378	0.07	0.416
miR-101-3p_miR-19b-3p	1.22	0.551	1541	0.768	0.14	0.714
miR-199a-3p_let-7a-5p	0.05	0.000	1061	0.017	0.30	0.025
miR-21-5p_miR-23a-5p	0.99	0.923	1581	0.604	0.14	0.718
miR-26b_miR-142-5p	0.53	0.020	1782	0.099	0.25	0.087
miR-93-5p_miR-19b-3p	0.79	0.384	1701	0.233	0.19	0.310
WCRF lifestyle score	0.61	0.071	1732	0.171	0.19	0.203



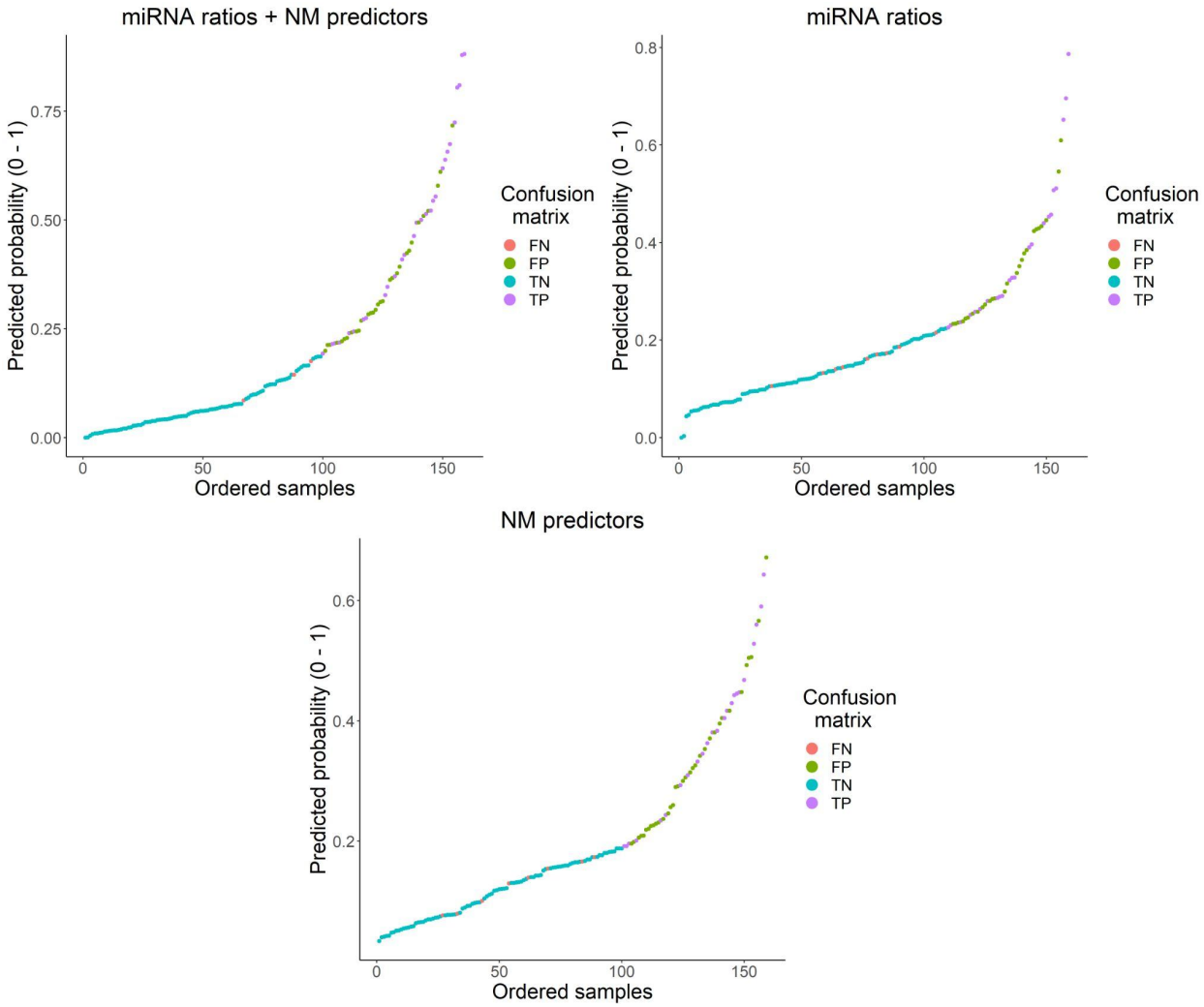
Supplementary Figure 1. Scatter plot of time from blood sampling to diagnosis and predicted probability after applying the coefficients to the validation cohort (miRNA ratios and non-molecular predictors assessed separately).



Supplementary Figure 2. Validation cohort samples ordered by the predicted probability of being BC positive (based on miRNA ratios and non-molecular predictors separately). The samples were then classified into predicted case or control based on the Youden's cut-off and the resulting prediction was labelled as TP, FP, TN or FN.

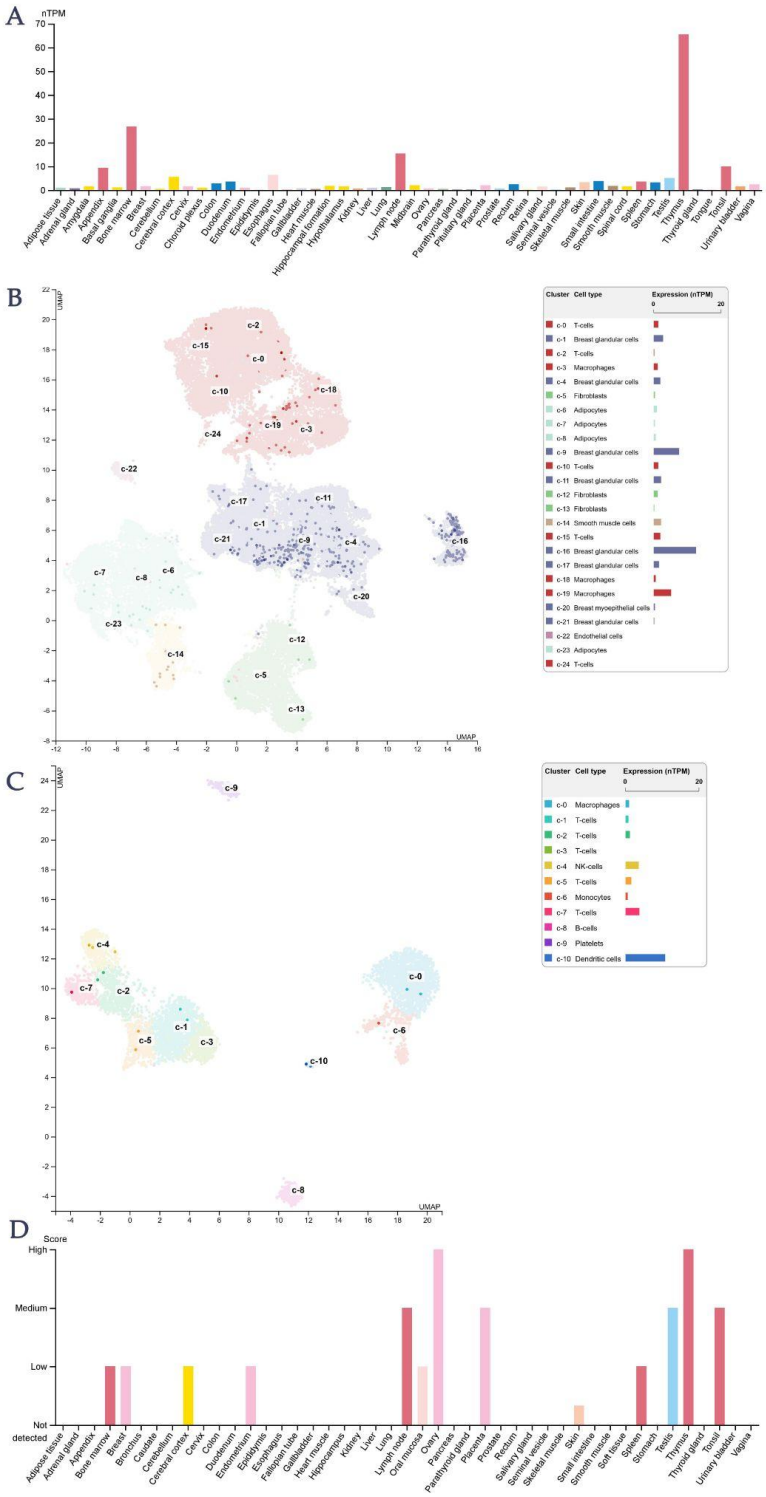


Supplementary Figure 3. Scatter plot of time from blood sampling to diagnosis and predicted probability after recalibrating the coefficients of the three models in the validation cohort.

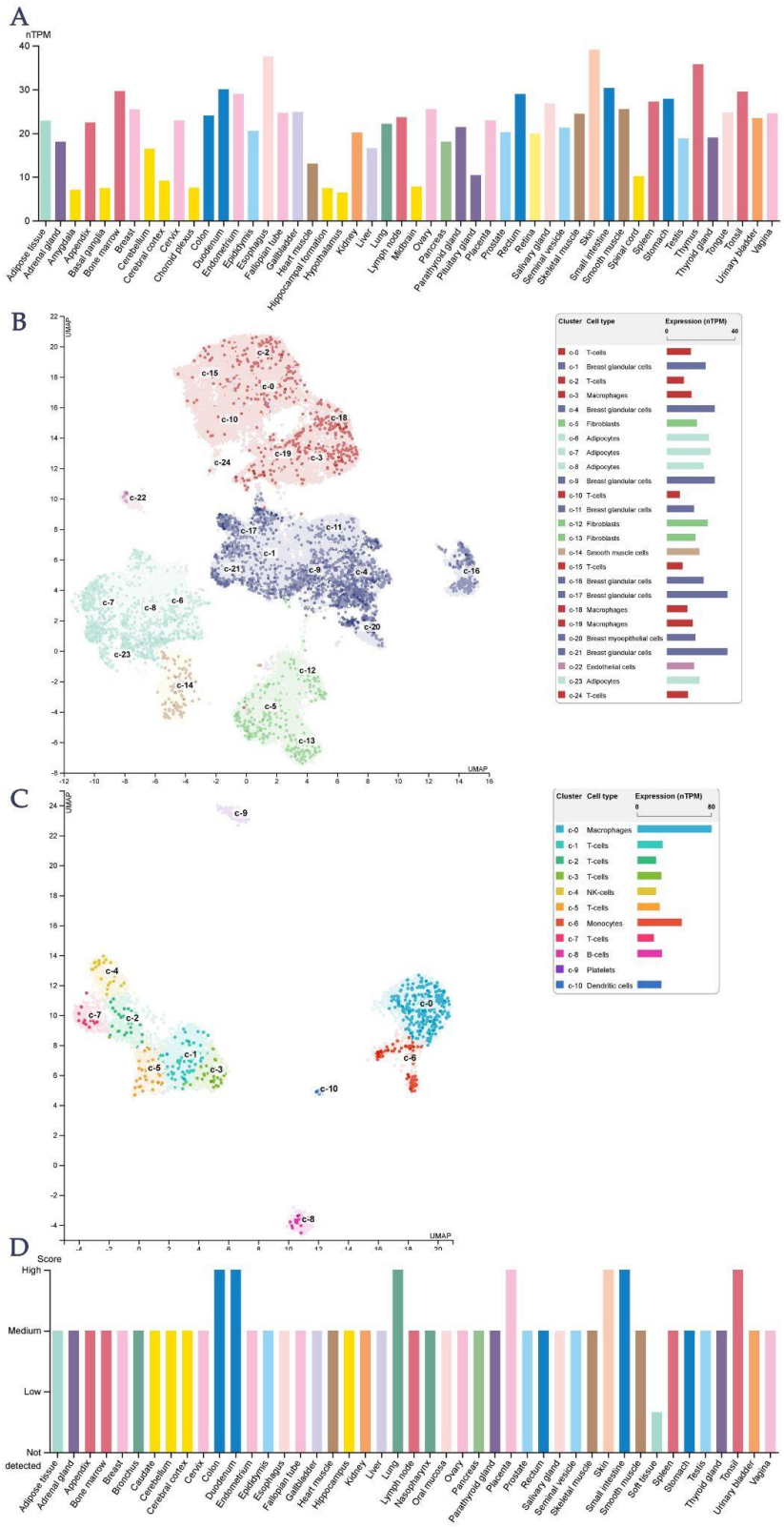


Supplementary Figure 4. Validation cohort samples ordered by the calibrated predicted probabilities of being BC positive (on all three models). The samples were then classified into predicted case or control based on the Youden's cut-off and the resulting prediction was labelled as TP, FP, TN or FN.

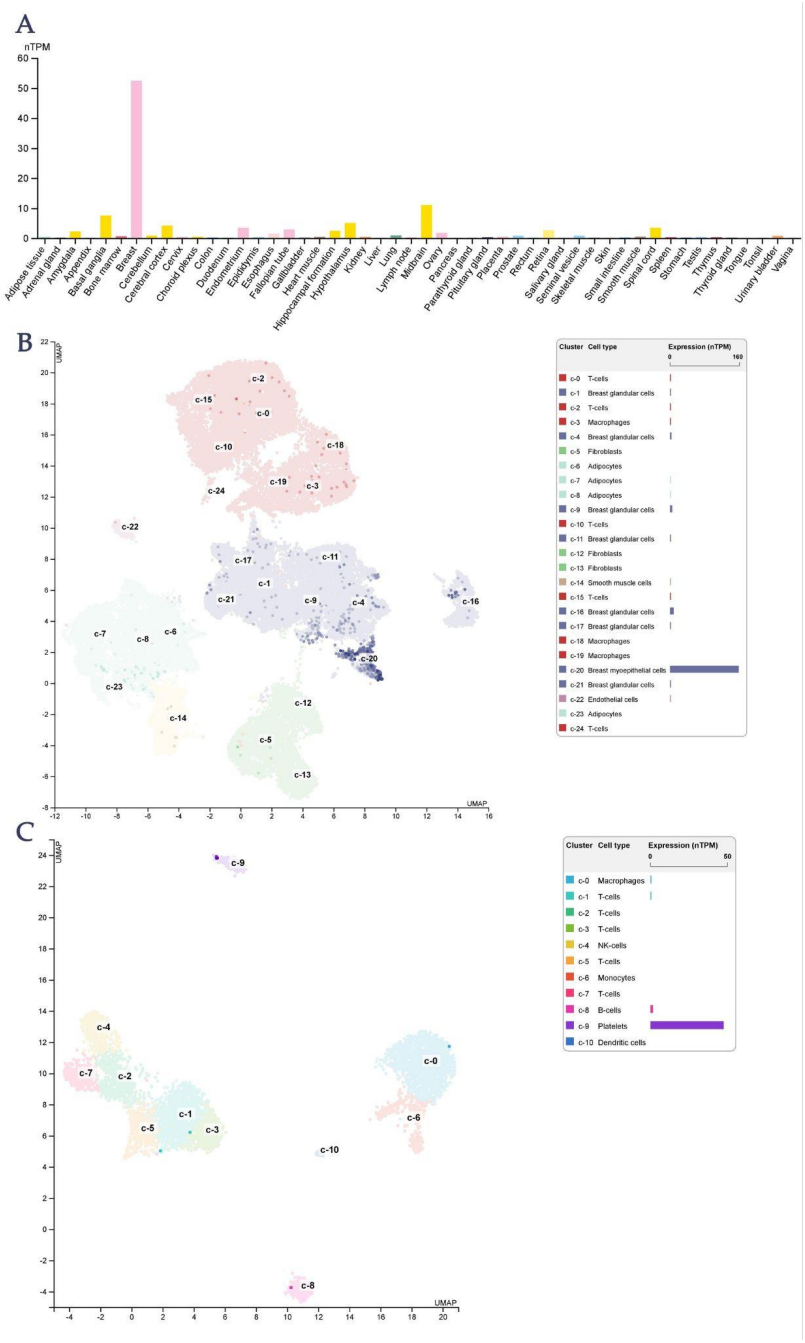
Appendix C



Supplementary Figure 5. *UHRF1* data obtained from the Human Protein Atlas. Shown are the bulk tissue (A) gene expression, breast (B) and PBMC (C) single cell gene expression as well as the tissue protein expression (D).



Supplementary Figure 6. *SP1* data obtained from the Human Protein Atlas. Shown are the bulk tissue (A) gene expression, breast (B) and PBMC (C) single cell gene expression as well as the tissue protein expression (D).



Supplementary Figure 7. *OXTR* data obtained from the Human Protein Atlas. Shown are the bulk tissue (A) gene expression, breast (B) and PBMC (C) single cell gene expression as well as the tissue protein expression (D).

References

1. Pezzella F, editor. Oxford textbook of cancer biology. First edition. Oxford: Oxford University Press; 2019. 483 p. (Oxford textbooks in oncology).
2. Hill W, Lim EL, Weeden CE, Lee C, Augustine M, Chen K, et al. Lung adenocarcinoma promotion by air pollutants. Nature. 2023 Apr 6;616(7955):159–67.
3. Sonnenschein C, Soto AM. Carcinogenesis explained within the context of a theory of organisms. Progress in Biophysics and Molecular Biology. 2016 Oct;122(1):70–6.
4. Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002 Dec;420(6917):860–7.
5. Greten FR, Grivennikov SI. Inflammation and Cancer: Triggers, Mechanisms, and Consequences. Immunity. 2019 Jul;51(1):27–41.
6. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discovery. 2022 Jan 1;12(1):31–46.
7. Sherr CJ, McCormick F. The RB and p53 pathways in cancer. Cancer Cell. 2002 Aug;2(2):103–12.
8. Mendiratta G, Ke E, Aziz M, Liarakos D, Tong M, Stites EC. Cancer gene mutation frequencies for the U.S. population. Nat Commun. 2021 Oct 13;12(1):5961.
9. Khezri MR, Jafari R, Yousefi K, Zolbanin NM. The PI3K/AKT signaling pathway in cancer: Molecular mechanisms and possible therapeutic interventions. Experimental and Molecular Pathology. 2022 Aug;127:104787.
10. Owen KL, Brockwell NK, Parker BS. JAK-STAT Signaling: A Double-Edged Sword of Immune Regulation and Cancer Progression. Cancers. 2019 Dec 12;11(12):2002.
11. Quintás-Cardama A, Verstovsek S. Molecular Pathways: JAK/STAT Pathway: Mutations, Inhibitors, and Resistance. Clinical Cancer Research. 2013 Apr 15;19(8):1933–40.
12. Meulmeester E, Ten Dijke P. The dynamic roles of TGF-β in cancer. The Journal of Pathology. 2011 Jan;223(2):206–19.
13. Guertin DA, Sabatini DM. Defining the Role of mTOR in Cancer. Cancer Cell. 2007 Jul;12(1):9–22.
14. Dreesen O, Brivanlou AH. Signaling Pathways in Cancer and Embryonic Stem Cells. Stem Cell Rev. 2007 Jan;3(1):7–17.
15. Sung H, Ferlay J, Siegel RL, 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 A Cancer J Clin. 2021 May;71(3):209–49.
16. Danckert B, Ferlay J, Engholm G, Hansen H, Johannesen T, Khan S, et al. NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 8.2 [Internet]. Association of the Nordic Cancer Registries. Danish Cancer Society. 2019 [cited 2021 Feb 21]. Available from: http://www.ancr.nu
17. Huang J, Chan PS, Lok V, Chen X, Ding H, Jin Y, et al. Global incidence and mortality of breast cancer: a trend analysis. Aging (Albany NY). 2021 Feb 11;13(4):5748–803.
18. Nelson DR, Brown J, Morikawa A, Method M. Breast cancer-specific mortality in early breast cancer as defined by high-risk clinical and pathologic characteristics. Wang Y, editor. PLoS ONE. 2022 Feb 25;17(2):e0264637.
19. Tsang JYS, Tse GM. Molecular Classification of Breast Cancer. Advances in Anatomic Pathology. 2020 Jan;27(1):27–35.

20. Lahkani S, Ellis I, Schnitt S, Tan P, van de Vijver M. WHO classification of tumours of the breast. 4th ed. Lyon: International agency for research on cancer; 2012. (World health organization classification of tumours).
21. Veta M, Pluim JPW, Van Diest PJ, Viergever MA. Breast Cancer Histopathology Image Analysis: A Review. IEEE Trans Biomed Eng. 2014 May;61(5):1400–11.
22. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010 Aug;12(4):207.
23. Sun X, Kaufman PD. Ki-67: more than a proliferation marker. Chromosoma. 2018 Jun;127(2):175–86.
24. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Annals of Oncology. 2013 Sep;24(9):2206–23.
25. Kwan ML, Kushi LH, Weltzien E, Maring B, Kutner SE, Fulton RS, et al. Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors. Breast Cancer Res. 2009 Jun;11(3):R31.
26. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. JCO. 2009 Mar 10;27(8):1160–7.
27. METABRIC Group, Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Jun;486(7403):346–52.
28. Amin MB, American Joint Committee on Cancer, American Cancer Society, editors. AJCC cancer staging manual. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP ; editors, Stephen B. Edge, MD, FACS [and 16 others] ; Donna M. Gress, RHIT, CTR-Technical editor ; Laura R. Meyer, CAPM-Managing editor. Chicago IL: American Joint Committee on Cancer, Springer; 2017. 1024 p.
29. Kalli S, Semine A, Cohen S, Naber SP, Makim SS, Bahl M. American Joint Committee on Cancer's Staging System for Breast Cancer, Eighth Edition: What the Radiologist Needs to Know. RadioGraphics. 2018 Nov;38(7):1921–33.
30. Giaquinto AN, Sung H, Miller KD, Kramer JL, Newman LA, Minihan A, et al. Breast Cancer Statistics, 2022. CA A Cancer J Clinicians. 2022 Nov;72(6):524–41.
31. Falck AK, Fernö M, Bendahl PO, Rydén L. St Gallen molecular subtypes in primary breast cancer and matched lymph node metastases - aspects on distribution and prognosis for patients with luminal A tumours: results from a prospective randomised trial. BMC Cancer. 2013 Dec;13(1):558.
32. Sant M, Allemani C, Capocaccia R, Hakulinen T, Aareleid T, Coebergh JW, et al. Stage at diagnosis is a key explanation of differences in breast cancer survival across Europe. Intl Journal of Cancer. 2003 Sep;106(3):416–22.
33. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MMA. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. BMJ. 2015 Oct 6;h4901.
34. Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, et al. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes & Diseases. 2018 Jun;5(2):77–106.
35. Said TK, Conneely OM, Medina D, O'Malley BW, Lydon JP. Progesterone, in Addition to Estrogen, Induces Cyclin D1 Expression in the Murine Mammary Epithelial Cell, in Vivo*. Endocrinology. 1997 Sep 1;138(9):3933–9.

36. Cicatiello L, Addeo R, Sasso A, Altucci L, Petrizzi VB, Borgo R, et al. Estrogens and Progesterone Promote Persistent *CCND1* Gene Activation during G 1 by Inducing Transcriptional Derepression via c- *Jun* /c- *Fos* /Estrogen Receptor (Progesterone Receptor) Complex Assembly to a Distal Regulatory Element and Recruitment of Cyclin D1 to Its Own Gene Promoter. Molecular and Cellular Biology. 2004 Aug 1;24(16):7260–74.
37. Zwijsen RML, Wientjens E, Klompmaker R, Van Der Sman J, Bernards R, Michalides RJAM. CDK-Independent Activation of Estrogen Receptor by Cyclin D1. Cell. 1997 Feb;88(3):405–15.
38. Burgess AW. EGFR family: Structure physiology signalling and therapeutic targets †. Growth Factors. 2008 Jan;26(5):263–74.
39. Mayer IA, Arteaga CL. The PI3K/AKT Pathway as a Target for Cancer Treatment. Annu Rev Med. 2016 Jan 14;67(1):11–28.
40. Koren S, Reavie L, Couto JP, De Silva D, Stadler MB, Roloff T, et al. PIK3CAH1047R induces multipotency and multi-lineage mammary tumours. Nature. 2015 Sep;525(7567):114–8.
41. Kakarala M, Wicha MS. Implications of the Cancer Stem-Cell Hypothesis for Breast Cancer Prevention and Therapy. JCO. 2008 Jun 10;26(17):2813–20.
42. Ayyanan A, Civenni G, Ciarloni L, Morel C, Mueller N, Lefort K, et al. Increased Wnt signaling triggers oncogenic conversion of human breast epithelial cells by a Notch-dependent mechanism. Proc Natl Acad Sci USA. 2006 Mar 7;103(10):3799–804.
43. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019 Jan 22;321(3):288.
44. Han HS, Vikas P, Costa RLB, Jahan N, Taye A, Stringer-Reasor EM. Early-Stage Triple-Negative Breast Cancer Journey: Beginning, End, and Everything in Between. American Society of Clinical Oncology Educational Book. 2023 Jun;(43):e390464.
45. Blumen H, Fitch K, Polkus V. Comparison of Treatment Costs for Breast Cancer, by Tumor Stage and Type of Service. Am Health Drug Benefits. 2016 Feb;9(1):23–32.
46. Salonen P, Kellokumpu-Lehtinen P, Tarkka M, Koivisto A, Kaunonen M. Changes in quality of life in patients with breast cancer. Journal of Clinical Nursing. 2011 Jan;20(1–2):255–66.
47. Reich M, Lesur A, Perdrizet-Chevallier C. Depression, quality of life and breast cancer: a review of the literature. Breast Cancer Res Treat. 2008 Jul;110(1):9–17.
48. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. BCTT. 2019 Apr;Volume 11:151–64.
49. Kamińska M, Ciszewski T, Łopacka-Szatan K, Miotła P, Starosławska E. Breast cancer risk factors. pm. 2015;3:196–202.
50. Ban KA, Godellas CV. Epidemiology of Breast Cancer. Surgical Oncology Clinics of North America. 2014 Jul;23(3):409–22.
51. Kim Y, Yoo KY, Goodman MT. Differences in Incidence, Mortality and Survival of Breast Cancer by Regions and Countries in Asia and Contributing Factors. Asian Pacific Journal of Cancer Prevention. 2015 Apr 14;16(7):2857–70.
52. Hsieh C, Trichopoulos D, Katsouyanni K, Yuasa S. Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: Associations and interactions in an international case-control study. Intl Journal of Cancer. 1990 Nov 15;46(5):796–800.
53. Thakur P, Seam RK, Gupta MK, Gupta M, Sharma M, Fotedar V. Breast cancer risk factor evaluation in a Western Himalayan state: A case–control study and comparison with the Western World. South Asian J Cancer. 2017 Jul;06(03):106–9.
54. Balekouzou A, Yin P, Pamatika CM, Bekolo CE, Nambei SW, Djeintote M, et al. Reproductive risk factors associated with breast cancer in women in Bangui: a case–control study. BMC Women's Health. 2017 Dec;17(1):14.

55. Hunter DJ, Colditz GA, Hankinson SE, Malspeis S, Spiegelman D, Chen W, et al. Oral Contraceptive Use and Breast Cancer: A Prospective Study of Young Women. Cancer Epidemiology, Biomarkers & Prevention. 2010 Oct 1;19(10):2496–502.
56. Narod SA. Hormone replacement therapy and the risk of breast cancer. Nat Rev Clin Oncol. 2011 Nov;8(11):669–76.
57. Collaborative Group on Hormonal Factors in Breast Cancer. Type and timing of menopausal hormone therapy and breast cancer risk: individual participant meta-analysis of the worldwide epidemiological evidence. Lancet. 2019 Sep 28;394(10204):1159–68.
58. Zolfaroli I, Tarín JJ, Cano A. Hormonal contraceptives and breast cancer: Clinical data. European Journal of Obstetrics & Gynecology and Reproductive Biology. 2018 Nov;230:212–6.
59. Shamseddin M, De Martino F, Constantin C, Scabia V, Lancelot A, Laszlo C, et al. Contraceptive progestins with androgenic properties stimulate breast epithelial cell proliferation. EMBO Mol Med. 2021 Jul 7;13(7):e14314.
60. Cobain EF, Milliron KJ, Merajver SD. Updates on breast cancer genetics: Clinical implications of detecting syndromes of inherited increased susceptibility to breast cancer. Seminars in Oncology. 2016 Oct;43(5):528–35.
61. Niederhuber JE, Armitage JO, Doroshow JH, Kastan MB, Tepper JE. Abeloff's clinical oncology. Sixth edition. Abeloff MD, editor. Philadelphia, PA: Elsevier; 2020.
62. Thompson D. Cancer Incidence in BRCA1 Mutation Carriers. CancerSpectrum Knowledge Environment. 2002 Sep 18;94(18):1358–65.
63. Semmler L, Reiter-Brennan C, Klein A. BRCA1 and Breast Cancer: a Review of the Underlying Mechanisms Resulting in the Tissue-Specific Tumorigenesis in Mutation Carriers. J Breast Cancer. 2019;22(1):1.
64. Metcalfe KA, Finch A, Poll A, Horsman D, Kim-Sing C, Scott J, et al. Breast cancer risks in women with a family history of breast or ovarian cancer who have tested negative for a BRCA1 or BRCA2 mutation. Br J Cancer. 2009 Jan;100(2):421–5.
65. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Annals of Oncology. 2015 Jul;26(7):1291–9.
66. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. The American Journal of Human Genetics. 2019 Jan;104(1):21–34.
67. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. The American Journal of Human Genetics. 2019 Jan;104(1):21–34.
68. Mars N, Widén E, Kerminen S, Meretoja T, Pirinen M, Della Briotta Parolo P, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. Nat Commun. 2020 Dec 14;11(1):6383.
69. Mars N, Lindbohm JV, Della Briotta Parolo P, Widén E, Kaprio J, Palotie A, et al. Systematic comparison of family history and polygenic risk across 24 common diseases. The American Journal of Human Genetics. 2022 Dec;109(12):2152–62.
70. Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). Radiologic Clinics of North America. 2002 May;40(3):409–30.
71. Gram IT, Funkhouser E, Tabár L. The Tabár classification of mammographic parenchymal patterns. European Journal of Radiology. 1997 Feb;24(2):131–6.
72. Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, et al. Benign Breast Disease and the Risk of Breast Cancer. N Engl J Med. 2005 Jul 21;353(3):229–37.

73. Arthur R, Wang Y, Ye K, Glass AG, Ginsberg M, Loudig O, et al. Association between lifestyle, menstrual/reproductive history, and histological factors and risk of breast cancer in women biopsied for benign breast disease. Breast Cancer Res Treat. 2017 Oct;165(3):623–31.
74. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. The Lancet. 2002 Jul;360(9328):187–95.
75. Argolo DF, Hudis CA, Iyengar NM. The Impact of Obesity on Breast Cancer. Curr Oncol Rep. 2018 Jun;20(6):47.
76. Chen Y, Liu L, Zhou Q, Imam MU, Cai J, Wang Y, et al. Body mass index had different effects on premenopausal and postmenopausal breast cancer risks: a dose-response meta-analysis with 3,318,796 subjects from 31 cohort studies. BMC Public Health. 2017 Dec;17(1):936.
77. Rose DP, Vona-Davis L. Interaction between menopausal status and obesity in affecting breast cancer risk. Maturitas. 2010 May;66(1):33–8.
78. Yager JD, Davidson NE. Estrogen Carcinogenesis in Breast Cancer. N Engl J Med. 2006 Jan 19;354(3):270–82.
79. Romieu I, Scoccianti C, Chajès V, De Batlle J, Biessy C, Dossus L, et al. Alcohol intake and breast cancer in the E uropean prospective investigation into cancer and nutrition. Intl Journal of Cancer. 2015 Oct 15;137(8):1921–30.
80. Luo J, Margolis KL, Wactawski-Wende J, Horn K, Messina C, Stefanick ML, et al. Association of active and passive smoking with risk of breast cancer among postmenopausal women: a prospective cohort study. BMJ. 2011 Mar 1;342(mar01 1):d1016–d1016.
81. Taylor EF, Burley VJ, Greenwood DC, Cade JE. Meat consumption and risk of breast cancer in the UK Women's Cohort Study. Br J Cancer. 2007 Apr 10;96(7):1139–46.
82. Sieri S, Krogh V, Ferrari P, Berrino F, Pala V, Thiébaut AC, et al. Dietary fat and breast cancer risk in the European Prospective Investigation into Cancer and Nutrition. The American Journal of Clinical Nutrition. 2008 Nov;88(5):1304–12.
83. McTiernan A, Kooperberg C, White E, Wilcox S, Coates R, Adams-Campbell LL, et al. Recreational Physical Activity and the Risk of Breast Cancer in Postmenopausal Women: The Women's Health Initiative Cohort Study. JAMA. 2003 Sep 10;290(10):1331.
84. John EM, Phipps AI, Knight JA, Milne RL, Dite GS, Hopper JL, et al. Medical radiation exposure and breast cancer risk: Findings from the Breast Cancer Family Registry. Intl Journal of Cancer. 2007 Jul 15;121(2):386–94.
85. Andersen ZJ, Stafoggia M, Weinmayr G, Pedersen M, Galassi C, Jørgensen JT, et al. Long-Term Exposure to Ambient Air Pollution and Incidence of Postmenopausal Breast Cancer in 15 European Cohorts within the ESCAPE Project. Environ Health Perspect. 2017 Oct 3;125(10):107005.
86. Schernhammer E, Bogl L, Hublin C, Strohmaier S, Zebrowska M, Erber A, et al. The association between night shift work and breast cancer risk in the Finnish twins cohort. Eur J Epidemiol. 2023 May;38(5):533–43.
87. Gonzalez TL, Rae JM, Colacino JA. Implication of environmental estrogens on breast cancer treatment and progression. Toxicology. 2019 Jun;421:41–8.
88. Eve L, Fervers B, Le Romancer M, Etienne-Selloum N. Exposure to Endocrine Disrupting Chemicals and Risk of Breast Cancer. IJMS. 2020 Nov 30;21(23):9139.
89. Oyelowo T. Estrogen Concepts. In: Mosby's Guide to Women's Health [Internet]. Elsevier; 2007 [cited 2023 Dec 16]. p. 8–10. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780323046015500034

90. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic Accuracy of Digital versus Film Mammography: Exploratory Analysis of Selected Population Subgroups in DMIST. Radiology. 2008 Feb;246(2):376–83.
91. Heywang-Köbrunner SH, Hacker A, Sedlacek S. Advantages and Disadvantages of Mammography Screening. Breast Care (Basel). 2011;6(3):199–207.
92. Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2 [Internet]. Washington, D.C.: National Academies Press; 2006 [cited 2024 Feb 1]. Available from: http://www.nap.edu/catalog/11340
93. Yaffe MJ, Mainprize JG. Risk of Radiation-induced Breast Cancer from Mammographic Screening. Radiology. 2011 Jan;258(1):98–105.
94. Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative Probability of False-Positive Recall or Biopsy Recommendation After 10 Years of Screening Mammography: A Cohort Study. Ann Intern Med. 2011 Oct 18;155(8):481.
95. Grimm LJ, Avery CS, Hendrick E, Baker JA. Benefits and Risks of Mammography Screening in Women Ages 40 to 49 Years. J Prim Care Community Health. 2022 Jan;13:215013272110583.
96. Von Euler-Chelpin M, Lillholm M, Vejborg I, Nielsen M, Lynge E. Sensitivity of screening mammography by density and texture: a cohort study from a population-based screening program in Denmark. Breast Cancer Res. 2019 Dec;21(1):111.
97. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic Density and the Risk and Detection of Breast Cancer. N Engl J Med. 2007 Jan 18;356(3):227–36.
98. Mandelson MT. Breast Density as a Predictor of Mammographic Detection: Comparison of Interval- and Screen-Detected Cancers. Journal of the National Cancer Institute. 2000 Jul 5;92(13):1081–7.
99. Daly MB, Pilarski R, Axilbund JE, Berry M, Buys SS, Crawford B, et al. Genetic/Familial High-Risk Assessment: Breast and Ovarian, Version 2.2015. J Natl Compr Canc Netw. 2016 Feb;14(2):153–62.
100. Venturini E, Losio C, Panizza P, Rodighiero MG, Fedele I, Tacchini S, et al. Tailored Breast Cancer Screening Program with Microdose Mammography, US, and MR Imaging: Short-term Results of a Pilot Study in 40–49-Year-Old Women. Radiology. 2013 Aug;268(2):347–55.
101. Allweis TM, Hermann N, Berenstein-Molho R, Guindy M. Personalized Screening for Breast Cancer: Rationale, Present Practices, and Future Directions. Ann Surg Oncol. 2021 Aug;28(8):4306–17.
102. Pashayan N, Morris S, Gilbert FJ, Pharoah PDP. Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model. JAMA Oncol. 2018 Nov 1;4(11):1504.
103. Schünemann HJ, Lerda D, Quinn C, Follmann M, Alonso-Coello P, Rossi PG, et al. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. Ann Intern Med. 2020 Jan 7;172(1):46.
104. Oeffinger KC, Fontham ETH, Etzioni R, Herzig A, Michaelson JS, Shih YCT, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA. 2015 Oct 20;314(15):1599.
105. Catana A, Apostu AP, Antemie RG. Multi gene panel testing for hereditary breast cancer - is it ready to be used? Medicine and Pharmacy Reports [Internet]. 2019 Jul 4 [cited 2023 Nov 21]; Available from: https://medpharmareports.com/index.php/mpr/article/view/1083
106. Fountzilas C, Kaklamani VG. Multi-gene Panel Testing in Breast Cancer Management. In: Gradishar WJ, editor. Optimizing Breast Cancer Management [Internet]. Cham: Springer International Publishing; 2018 [cited 2023 Nov 21]. p. 121–40. (Cancer Treatment and Research; vol. 173). Available from: http://link.springer.com/10.1007/978-3-319-70197-4_8

107. D'Argenio V, Esposito MV, Telese A, Precone V, Starnone F, Nunziato M, et al. The molecular analysis of BRCA1 and BRCA2: Next-generation sequencing supersedes conventional approaches. Clinica Chimica Acta. 2015 Jun;446:221–5.
108. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020 Sep 1;15(9):2759–72.
109. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. JNCI: Journal of the National Cancer Institute [Internet]. 2015 May [cited 2023 Jul 27];107(5). Available from: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/djv036
110. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. JNCI: Journal of the National Cancer Institute [Internet]. 2015 May [cited 2021 Jan 20];107(5). Available from: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/djv036
111. Aucamp J, Bronkhorst AJ, Badenhorst CPS, Pretorius PJ. The diverse origins of circulating cell-free DNA in the human body: a critical re-evaluation of the literature. Biological Reviews. 2018 Aug;93(3):1649–83.
112. Thierry AR, El Messaoudi S, Gahan PB, Anker P, Stroun M. Origins, structures, and functions of circulating DNA in oncology. Cancer Metastasis Rev. 2016 Sep;35(3):347–76.
113. Sant M, Bernat-Peguera A, Felip E, Margelí M. Role of ctDNA in Breast Cancer. Cancers. 2022 Jan 9;14(2):310.
114. Davidson BA, Croessmann S, Park BH. The breast is yet to come: current and future utility of circulating tumour DNA in breast cancer. Br J Cancer. 2021 Sep 14;125(6):780–8.
115. Yu D, Tong Y, Guo X, Feng L, Jiang Z, Ying S, et al. Diagnostic Value of Concentration of Circulating Cell-Free DNA in Breast Cancer: A Meta-Analysis. Front Oncol. 2019 Mar 1;9:95.
116. Liyanage UK, Moore TT, Joo HG, Tanaka Y, Herrmann V, Doherty G, et al. Prevalence of Regulatory T Cells Is Increased in Peripheral Blood and Tumor Microenvironment of Patients with Pancreas or Breast Adenocarcinoma. The Journal of Immunology. 2002 Sep 1;169(5):2756–61.
117. Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. Journal of Laboratory and Clinical Medicine. 2006 Mar;147(3):126–32.
118. Aarøe J, Lindahl T, Dumeaux V, Sæbø S, Tobin D, Hagen N, et al. Gene expression profiling of peripheral blood cells for early detection of breast cancer. Breast Cancer Res. 2010 Feb;12(1):R7.
119. Hou H, Lyu Y, Jiang J, Wang M, Zhang R, Liew CC, et al. Peripheral blood transcriptome identifies high-risk benign and malignant breast lesions. Deb S, editor. PLoS ONE. 2020 Jun 4;15(6):e0233713.
120. Lund E, Dumeaux V, Braaten T, Hjartaker A, Engeset D, Skeie G, et al. Cohort Profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. International Journal of Epidemiology. 2008 Feb 1;37(1):36–41.
121. Lund E, Holden L, Bøvelstad H, Plancade S, Mode N, Günther CC, et al. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the NOWAC postgenome cohort as a proof of principle. BMC Med Res Methodol. 2016 Dec;16(1):28.
122. Holden M, Holden L, Olsen KS, Lund E. Local in Time Statistics for detecting weak gene expression signals in blood – illustrated for prediction of metastases in breast cancer in the NOWAC Post-genome Cohort. AGG. 2017 Jul;Volume 7:11–28.
123. Syantra Inc. Syantra DX, A Blood Test for Breast Cancer [Internet]. 2021 [cited 2023 Nov 24]. Available from: https://www.syantra.com
124. Čelešnik H, Potočnik U. Blood-Based mRNA Tests as Emerging Diagnostic Tools for Personalised Medicine in Breast Cancer. Cancers. 2023 Feb 8;15(4):1087.

125. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010 Jan 1;31(1):27–36.
126. Yamada Y, Haga H, Yamada Y. Concise Review: Dedifferentiation Meets Cancer Development: Proof of Concept for Epigenetic Cancer. Stem Cells Translational Medicine. 2014 Oct 1;3(10):1182–7.
127. Kamińska K, Nalejska E, Kubiak M, Wojtysiak J, Żołna Ł, Kowalewski J, et al. Prognostic and Predictive Epigenetic Biomarkers in Oncology. Mol Diagn Ther. 2019 Feb;23(1):83–95.
128. Costa-Pinheiro P, Montezuma D, Henrique R, Jerónimo C. Diagnostic and prognostic epigenetic biomarkers in cancer. Epigenomics. 2015 Sep;7(6):1003–15.
129. Akinyemiju T. Epigenetic Biomarkers in Cancer Epidemiology. In: Epigenetic Mechanisms in Cancer [Internet]. Elsevier; 2018 [cited 2023 Dec 6]. p. 223–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128095522000097
130. Vietri M, D'elia G, Benincasa G, Ferraro G, Caliendo G, Nicoletti G, et al. DNA methylation and breast cancer: A way forward (Review). Int J Oncol. 2021 Oct 29;59(5):98.
131. Terry MB, McDonald JA, Wu HC, Eng S, Santella RM. Epigenetic Biomarkers of Breast Cancer Risk: Across the Breast Cancer Prevention Continuum. In: Stearns V, editor. Novel Biomarkers in the Continuum of Breast Cancer [Internet]. Cham: Springer International Publishing; 2016 [cited 2023 Dec 6]. p. 33–68. (Advances in Experimental Medicine and Biology; vol. 882). Available from: http://link.springer.com/10.1007/978-3-319-22909-6_2
132. Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet. 2018 Feb;19(2):81–92.
133. Robert MF, Morin S, Beaulieu N, Gauthier F, Chute IC, Barsalou A, et al. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nat Genet. 2003 Jan;33(1):61–5.
134. Bashtrykov P, Jeltsch A. DNMT1-associated DNA methylation changes in cancer. Cell Cycle. 2015 Jan 2;14(1):5–5.
135. Al-Kharashi LA, Al-Mohanna FH, Tulbah A, Aboussekhra A. The DNA methyl-transferase protein DNMT1 enhances tumor-promoting properties of breast stromal fibroblasts. Oncotarget. 2018 Jan 5;9(2):2329–43.
136. Angeloni A, Bogdanovic O. Enhancer DNA methylation: implications for gene regulation. Blewitt M, editor. Essays in Biochemistry. 2019 Dec 20;63(6):707–15.
137. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics & Chromatin. 2018 Dec;11(1):37.
138. Li Y, Tollefsbol TO. DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. In: Tollefsbol TO, editor. Epigenetics Protocols [Internet]. Totowa, NJ: Humana Press; 2011 [cited 2023 Dec 6]. p. 11–21. (Methods in Molecular Biology; vol. 791). Available from: http://link.springer.com/10.1007/978-1-61779-316-5_2
139. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct;98(4):288–95.
140. Hussmann D, Hansen LL. Methylation-Sensitive High Resolution Melting (MS-HRM). In: Tost J, editor. DNA Methylation Protocols [Internet]. New York, NY: Springer New York; 2018 [cited 2023 Jul 27]. p. 551–71. (Methods in Molecular Biology; vol. 1708). Available from: http://link.springer.com/10.1007/978-1-4939-7481-8_28
141. Zhang L, Long X. Association of BRCA1 promoter methylation with sporadic breast cancers: Evidence from 40 studies. Sci Rep. 2015 Dec 8;5(1):17869.

142. Bodelon C, Ambatipudi S, Dugué PA, Johansson A, Sampson JN, Hicks B, et al. Blood DNA methylation and breast cancer risk: a meta-analysis of four prospective cohort studies. Breast Cancer Res. 2019 Dec;21(1):62.
143. Van Veldhoven K, Polidoro S, Baglietto L, Severi G, Sacerdote C, Panico S, et al. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clin Epigenet. 2015 Dec;7(1):67.
144. Tang Q, Cheng J, Cao X, Surowy H, Burwinkel B. Blood-based DNA methylation as biomarker for breast cancer: a systematic review. Clin Epigenet. 2016 Dec;8(1):115.
145. Wang T, Li P, Qi Q, Zhang S, Xie Y, Wang J, et al. A multiplex blood-based assay targeting DNA methylation in PBMCs enables early detection of breast cancer. Nat Commun. 2023 Aug 7;14(1):4724.
146. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Reports. 2001 Nov;2(11):986–91.
147. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009 Mar;10(3):155–9.
148. Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023 Jun;24(6):430–47.
149. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021 Feb;22(2):96–118.
150. Flynn RA, Chang HY. Long Noncoding RNAs in Cell-Fate Programming and Reprogramming. Cell Stem Cell. 2014 Jun;14(6):752–61.
151. Chen L, Zhu QH, Kaufmann K. Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. Planta. 2020 Nov;252(5):92.
152. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, et al. A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response. Cell. 2010 Aug;142(3):409–19.
153. Fanucchi S, Fok ET, Dalla E, Shibayama Y, Börner K, Chang EY, et al. Immune genes are primed for robust transcription by proximal long noncoding RNAs located in nuclear compartments. Nat Genet. 2019 Jan;51(1):138–50.
154. Ruan X, Li P, Ma Y, Jiang C fei, Chen Y, Shi Y, et al. Identification of human long noncoding RNAs associated with nonalcoholic fatty liver disease and metabolic homeostasis. Journal of Clinical Investigation. 2021 Jan 4;131(1):e136336.
155. Hennessy EJ, Van Solingen C, Scacalossi KR, Ouimet M, Afonso MS, Prins J, et al. The long noncoding RNA CHROME regulates cholesterol homeostasis in primates. Nat Metab. 2018 Dec 3;1(1):98–110.
156. Du Q, Hoover AR, Dozmorov I, Raj P, Khan S, Molina E, et al. MIR205HG Is a Long Noncoding RNA that Regulates Growth Hormone and Prolactin Production in the Anterior Pituitary. Developmental Cell. 2019 May;49(4):618-631.e5.
157. Zheng X, Han H, Liu G, Ma Y, Pan R, Sang L, et al. LncRNA wires up Hippo and Hedgehog signaling to reprogramme glucose metabolism. The EMBO Journal. 2017 Nov 15;36(22):3325–35.
158. Zhang P, Cao L, Fan P, Mei Y, Wu M. LncRNA-MIF, a c-Myc-activated long non-coding RNA, suppresses glycolysis by promoting Fbxw7-mediated c-Myc degradation. EMBO Reports. 2016 Aug;17(8):1204–20.
159. McClintock MA, Dix CI, Johnson CM, McLaughlin SH, Maizels RJ, Hoang HT, et al. RNA-directed activation of cytoplasmic dynein-1 in reconstituted transport RNPs. eLife. 2018 Jun 26;7:e36312.

160. Lin A, Hu Q, Li C, Xing Z, Ma G, Wang C, et al. The LINK-A lncRNA interacts with PtdIns(3,4,5)P3 to hyperactivate AKT and confer resistance to AKT inhibitors. Nat Cell Biol. 2017 Mar;19(3):238–51.
161. Sang L jie, Ju H qiang, Liu G ping, Tian T, Ma G lin, Lu Y xin, et al. LncRNA CamK-A Regulates Ca2+-Signaling-Mediated Tumor Microenvironment Remodeling. Molecular Cell. 2018 Oct;72(1):71-83.e7.
162. Sánchez Y, Segura V, Marín-Béjar O, Athie A, Marchese FP, González J, et al. Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. Nat Commun. 2014 Dec 19;5(1):5812.
163. Hart JR, Roberts TC, Weinberg MS, Morris KV, Vogt PK. MYC regulates the non-coding transcriptome. Oncotarget. 2014 Dec 30;5(24):12543–54.
164. Kim T, Jeon YJ, Cui R, Lee JH, Peng Y, Kim SH, et al. Role of MYC-Regulated Long Noncoding RNAs in Cell Cycle Regulation and Tumorigenesis. JNCI: Journal of the National Cancer Institute [Internet]. 2015 Apr [cited 2023 Dec 6];107(4). Available from: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/dju505
165. Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun. 2014 Nov 21;5(1):5383.
166. Peng W xin, Huang J guo, Yang L, Gong A hua, Mo YY. Linc-RoR promotes MAPK/ERK signaling and confers estrogen-independent growth of breast cancer. Mol Cancer. 2017 Dec;16(1):161.
167. Hou P, Zhao Y, Li Z, Yao R, Ma M, Gao Y, et al. LincRNA-ROR induces epithelial-to-mesenchymal transition and contributes to breast cancer tumorigenesis and metastasis. Cell Death Dis. 2014 Jun 12;5(6):e1287–e1287.
168. Xu N, Chen F, Wang F, Lu X, Wang X, Lv M, et al. Clinical significance of high expression of circulating serum lncRNA RP11-445H22.4 in breast cancer patients: a Chinese population-based study. Tumor Biol. 2015 Oct;36(10):7659–65.
169. Vickers KC, Roteta LA, Hucheson-Dilks H, Han L, Guo Y. Mining diverse small RNA species in the deep transcriptome. Trends in Biochemical Sciences. 2015 Jan;40(1):4–7.
170. Jorjani H, Kehr S, Jedlinski DJ, Gumienny R, Hertel J, Stadler PF, et al. An updated human snoRNAome. Nucleic Acids Res. 2016 Jun 20;44(11):5068–82.
171. Bachellerie JP, Cavaillé J, Hüttenhofer A. The expanding snoRNA world. Biochimie. 2002 Aug;84(8):775–90.
172. Romano G, Veneziano D, Acunzo M, Croce CM. Small non-coding RNA and cancer. Carcinogenesis. 2017 May;38(5):485–91.
173. Liang J, Wen J, Huang Z, Chen X ping, Zhang B xiang, Chu L. Small Nucleolar RNAs: Insight Into Their Function in Cancer. Front Oncol. 2019 Jul 9;9:587.
174. Fang X, Yang D, Luo H, Wu S, Dong W, Xiao J, et al. SNORD126 promotes HCC and CRC cell growth by activating the PI3K–AKT pathway through FGFR2. J Mol Cell Biol. 2016 Dec 2;jmcb;mjw048v2.
175. Siprashvili Z, Webster DE, Johnston D, Shenoy RM, Ungewickell AJ, Bhaduri A, et al. The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer. Nat Genet. 2016 Jan;48(1):53–8.
176. Chen W, Moore MJ. Spliceosomes. Current Biology. 2015 Mar;25(5):R181–3.
177. Dvinge H, Guenthoer J, Porter PL, Bradley RK. RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. Genome Res. 2019 Oct;29(10):1591–604.

178. Cheng Z, Sun Y, Niu X, Shang Y, Ruan J, Chen Z, et al. Gene expression profiling reveals U1 snRNA regulates cancer gene expression. Oncotarget. 2017 Dec 22;8(68):112867–74.
179. Liu Y, Dou M, Song X, Dong Y, Liu S, Liu H, et al. The emerging role of the piRNA/piwi complex in cancer. Mol Cancer. 2019 Dec;18(1):123.
180. Moyano M, Stefani G. piRNA involvement in genome stability and human cancer. J Hematol Oncol. 2015 Dec;8(1):38.
181. Chalbatani GM, Dana H, Memari F, Gharagozlou E, Ashjaei S, Kheirandish P, et al. Biological function and molecular mechanism of piRNA in cancer. Practical Laboratory Medicine. 2019 Jan;13:e00113.
182. Tan L, Mai D, Zhang B, Jiang X, Zhang J, Bai R, et al. PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. Mol Cancer. 2019 Dec;18(1):9.
183. Khan S, Ayub H, Khan T, Wahid F. MicroRNA biogenesis, gene silencing mechanisms and role in breast, ovarian and prostate cancer. Biochimie. 2019 Dec;167:12–24.
184. Cammaerts S, Strazisar M, De Rijk P, Del Favero J. Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. Front Genet [Internet]. 2015 May 21 [cited 2023 Dec 6];6. Available from: http://www.frontiersin.org/RNA/10.3389/fgene.2015.00186/abstract
185. Olena AF, Patton JG. Genomic organization of microRNAs. Journal Cellular Physiology. 2010 Mar;222(3):540–5.
186. Axtell MJ, Westholm JO, Lai EC. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. Genome Biol. 2011;12(4):221.
187. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014 Aug;15(8):509–24.
188. Wahid F, Shehzad A, Khan T, Kim YY. MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2010 Nov;1803(11):1231–43.
189. Park JE, Heo I, Tian Y, Simanshu DK, Chang H, Jee D, et al. Dicer recognizes the 5′ end of RNA for efficient and accurate processing. Nature. 2011 Jul;475(7355):201–5.
190. Medley JC, Panzade G, Zinovyeva AY. microRNA strand selection: Unwinding the rules. WIREs RNA. 2021 May;12(3):e1627.
191. Song X, Cheng L, Zhou T, Guo X, Zhang X, Chen Y ping P, et al. Predicting miRNA-mediated gene silencing mode based on miRNA-target duplex features. Computers in Biology and Medicine. 2012 Jan;42(1):1–7.
192. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids—the mix of hormones and biomarkers. Nat Rev Clin Oncol. 2011 Aug;8(8):467–77.
193. Chen C, Tan R, Wong L, Fekete R, Halsey J. Quantitation of MicroRNAs by Real-Time RT-qPCR. In: Park DJ, editor. PCR Protocols [Internet]. Totowa, NJ: Humana Press; 2011 [cited 2023 Dec 7]. p. 113–34. (Methods in Molecular Biology; vol. 687). Available from: https://link.springer.com/10.1007/978-1-60761-944-4_8
194. Liu CG, Calin GA, Volinia S, Croce CM. MicroRNA expression profiling using microarrays. Nat Protoc. 2008 Apr;3(4):563–78.
195. Benesova S, Kubista M, Valihrach L. Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. Diagnostics. 2021 May 27;11(6):964.
196. Ahn WS, Kim YW, Liu JL, Kim H, Kim EY, Jeon D, et al. Differential microRNA expression signatures and cell type-specific association with Taxol resistance in ovarian cancer cells. DDDT. 2014 Feb;293.

197. Mulrane L, McGee SF, Gallagher WM, O'Connor DP. miRNA Dysregulation in Breast Cancer. Cancer Research. 2013 Nov 15;73(22):6554–62.
198. Srivastava K, Srivastava A. Comprehensive Review of Genetic Association Studies and Meta-Analyses on miRNA Polymorphisms and Cancer Risk. De Windt LJ, editor. PLoS ONE. 2012 Nov 30;7(11):e50966.
199. Reddy KB. MicroRNA (miRNA) in cancer. Cancer Cell Int. 2015 Dec;15(1):38.
200. Peng Y, Croce CM. The role of MicroRNAs in human cancer. Sig Transduct Target Ther. 2016 Jan 28;1(1):15004.
201. Iorio MV, Croce CM. microRNA involvement in human cancer. Carcinogenesis. 2012 Jun 1;33(6):1126–33.
202. Chang TC, Yu D, Lee YS, Wentzel EA, Arking DE, West KM, et al. Widespread microRNA repression by Myc contributes to tumorigenesis. Nat Genet. 2008 Jan;40(1):43–50.
203. Humphries B, Yang C. The microRNA-200 family: small molecules with novel roles in cancer development, progression and therapy. Oncotarget. 2015 Mar 30;6(9):6472–98.
204. Feng X, Wang Z, Fillmore R, Xi Y. MiR-200, a new star miRNA in human cancer. Cancer Letters. 2014 Mar;344(2):166–73.
205. Jalava SE, Urbanucci A, Latonen L, Waltering KK, Sahu B, Jänne OA, et al. Androgen-regulated miR-32 targets BTG2 and is overexpressed in castration-resistant prostate cancer. Oncogene. 2012 Oct 11;31(41):4460–71.
206. Wang J, Chen J, Sen S. MicroRNA as Biomarkers and Diagnostics. Journal Cellular Physiology. 2016 Jan;231(1):25–30.
207. Ghamlouche F, Yehya A, Zeid Y, Fakhereddine H, Fawaz J, Liu YN, et al. MicroRNAs as clinical tools for diagnosis, prognosis, and therapy in prostate cancer. Translational Oncology. 2023 Feb;28:101613.
208. Mello-Grand M, Gregnanin I, Sacchetto L, Ostano P, Zitella A, Bottoni G, et al. Circulating microRNAs combined with PSA for accurate and non-invasive prostate cancer detection. Carcinogenesis. 2019 Apr 29;40(2):246–53.
209. Fehlmann T, Kahraman M, Ludwig N, Backes C, Galata V, Keller V, et al. Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients. JAMA Oncol. 2020 May 1;6(5):714.
210. Inagaki M, Uchiyama M, Yoshikawa-Kawabe K, Ito M, Murakami H, Gunji M, et al. Comprehensive circulating microRNA profile as a supersensitive biomarker for early-stage lung cancer screening. J Cancer Res Clin Oncol. 2023 Sep;149(11):8297–305.
211. Wang H, Peng R, Wang J, Qin Z, Xue L. Circulating microRNAs as potential cancer biomarkers: the advantage and disadvantage. Clin Epigenet. 2018 Dec;10(1):59.
212. Wang J, Zhang KY, Liu SM, Sen S. Tumor-Associated Circulating MicroRNAs as Biomarkers of Cancer. Molecules. 2014 Feb 10;19(2):1912–38.
213. Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, Miyaji MM, et al. Blood Cell Origin of Circulating MicroRNAs: A Cautionary Note for Cancer Biomarker Studies. Cancer Prevention Research. 2012 Mar 1;5(3):492–7.
214. Van Schooneveld E, Wildiers H, Vergote I, Vermeulen PB, Dirix LY, Van Laere SJ. Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. Breast Cancer Res. 2015 Dec;17(1):21.
215. Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature. 2007 Oct 11;449(7163):682–8.

216. Yang H, Zhou J, Mi J, Ma K, Fan Y, Ning J, et al. HOXD10 acts as a tumor-suppressive factor via inhibition of the RHOC/AKT/MAPK pathway in human cholangiocellular carcinoma. Oncology Reports. 2015 Oct;34(4):1681–91.
217. Starr R, Willson TA, Viney EM, Murray LJL, Rayner JR, Jenkins BJ, et al. A family of cytokine-inducible inhibitors of signalling. Nature. 1997 Jun;387(6636):917–21.
218. Endo TA, Masuhara M, Yokouchi M, Suzuki R, Sakamoto H, Mitsui K, et al. A new protein containing an SH2 domain that inhibits JAK kinases. Nature. 1997 Jun;387(6636):921–4.
219. Jiang S, Zhang HW, Lu MH, He XH, Li Y, Gu H, et al. MicroRNA-155 Functions as an OncomiR in Breast Cancer by Targeting the *Suppressor of Cytokine Signaling 1* Gene. Cancer Research. 2010 Apr 15;70(8):3119–27.
220. Wang T, Niu G, Kortylewski M, Burdelya L, Shain K, Zhang S, et al. Regulation of the innate and adaptive immune responses by Stat-3 signaling in tumor cells. Nat Med. 2004 Jan 1;10(1):48–54.
221. Feng YH, Tsao CJ. Emerging role of microRNA-21 in cancer. Biomedical Reports. 2016 Oct;5(4):395–402.
222. Hopkins BD, Hodakoski C, Barrows D, Mense SM, Parsons RE. PTEN function: the long and the short of it. Trends in Biochemical Sciences. 2014 Apr;39(4):183–90.
223. Chan KK, Matchett KB, Coulter JA, Yuen HF, McCrudden CM, Zhang SD, et al. Erythropoietin drives breast cancer progression by activation of its receptor EPOR. Oncotarget. 2017 Jun 13;8(24):38251–63.
224. Aggarwal V, Priyanka K, Tuli HS. Emergence of Circulating MicroRNAs in Breast Cancer as Diagnostic and Therapeutic Efficacy Biomarkers. Mol Diagn Ther. 2020 Apr;24(2):153–73.
225. Boeri M, Verri C, Conte D, Roz L, Modena P, Facchinetti F, et al. MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci USA. 2011 Mar;108(9):3713–8.
226. Hong F, Li N, Feng Z, Zheng Y, Zhu C, Zhang F. Exosomal microRNAs as novel diagnostic biomarkers in breast cancer: A systematic evaluation and meta-analysis. Asian Journal of Surgery. 2023 Nov;46(11):4727–36.
227. Schwarzenbach H. Methods for quantification and characterization of microRNAs in cell-free plasma/serum, normal exosomes and tumor-derived exosomes. Transl Cancer Res. 2018 Mar;7(S2):S253–63.
228. Lee I, Baxter D, Lee MY, Scherler K, Wang K. The Importance of Standardization on Analyzing Circulating RNA. Mol Diagn Ther. 2017 Jun;21(3):259–68.
229. Cui Z, Lin D, Song W, Chen M, Li D. Diagnostic value of circulating microRNAs as biomarkers for breast cancer: a meta-analysis study. Tumour Biol. 2015 Feb;36(2):829–39.
230. Liu L, Wang S, Cao X, Liu J. Analysis of circulating microRNA biomarkers for breast cancer detection: a meta-analysis. Tumor Biol. 2014 Dec;35(12):12245–53.
231. Mayeux R. Biomarkers: potential uses and limitations. NeuroRx. 2004 Apr;1(2):182–8.
232. Sehovic E, Urru S, Chiorino G, Doebler P. Meta-analysis of diagnostic cell-free circulating microRNAs for breast cancer detection. BMC Cancer. 2022 Dec;22(1):634.
233. Chiorino G, Petracci E, Sehovic E, Gregnanin I, Camussi E, Mello-Grand M, et al. Plasma microRNA ratios associated with breast cancer detection in a nested case–control study from a mammography screening cohort. Sci Rep. 2023 Jul 25;13(1):12040.
234. Sehovic E, Zellers SM, Youssef MK, Heikkinen A, Kaprio J, Ollikainen M. DNA methylation sites in early adulthood characterised by pubertal timing and development: a twin study. Clin Epigenet. 2023 Nov 10;15(1):181.

235. McGrath TA, Alabousi M, Skidmore B, Korevaar DA, Bossuyt PMM, Moher D, et al. Recommendations for reporting of systematic reviews and meta-analyses of diagnostic test accuracy: a systematic review. Syst Rev. 2017 Dec;6(1):194.
236. Toss A, Isca C, Venturelli M, Nasso C, Ficarra G, Bellelli V, et al. Two-month stop in mammographic screening significantly impacts on breast cancer stage at diagnosis and upfront treatment in the COVID era. ESMO Open. 2021 Apr;6(2):100055.
237. Poisot T. The digitize Package: Extracting Numerical Data from Scatterplots. The R Journal. 2011;3(1):25–6.
238. Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011 Oct 18;155(8):529.
239. Ying GS, Maguire MG, Glynn RJ, Rosner B. Calculating Sensitivity, Specificity, and Predictive Values for Correlated Eye Data. Invest Ophthalmol Vis Sci. 2020 Sep 16;61(11):29.
240. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case–control studies. Statist Med. 2007 May 10;26(10):2170–83.
241. Altman DG, editor. Statistics with confidence: confidence intervals and statistical guidelines ; [includes disk]. 2. ed., [Nachdr.]. London: BMJ Books; 2011. 240 p.
242. Doebler P. mada: Meta-Analysis of Diagnostic Accuracy [Internet]. 2020 [cited 2021 Sep 29]. Available from: https://CRAN.R-project.org/package=mada
243. Dean CB, Nielsen JD. Generalized linear mixed models: a review and some extensions. Lifetime Data Anal. 2007 Dec;13(4):497–512.
244. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005 Oct;58(10):982–90.
245. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015 Oct 7;67:1–48.
246. Vogelgesang F, Schlattmann P, Dewey M. The Evaluation of Bivariate Mixed Models in Meta-analyses of Diagnostic Accuracy Studies with SAS, Stata and R. Methods Inf Med. 2018 May;57(3):111–9.
247. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software. 2010 Aug 5;36:1–48.
248. Nieuwenhuis R, Grotenhuis M te, Pelzer B. influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. The R Journal. 2012;4(2):38–47.
249. Doebler P, Holling H. Meta-analysis of Diagnostic Accuracy and ROC Curves with Covariate Adjusted Semiparametric Mixtures. Psychometrika. 2015 Dec;80(4):1084–104.
250. Holling H, Böhning W, Böhning D. Meta-analysis of diagnostic studies based upon SROC-curves: a mixed model approach using the Lehmann family. Statistical Modelling. 2012 Aug;12(4):347–75.
251. Doebler P, Holling H. Meta-analysis of Diagnostic Accuracy and ROC Curves with Covariate Adjusted Semiparametric Mixtures. Psychometrika. 2015 Dec;80(4):1084–104.
252. Romaguera D, Vergnaud AC, Peeters PH, Van Gils CH, Chan DS, Ferrari P, et al. Is concordance with World Cancer Research Fund/American Institute for Cancer Research guidelines for cancer prevention related to subsequent risk of cancer? Results from the EPIC study. The American Journal of Clinical Nutrition. 2012 Jul;96(1):150–63.
253. Karavasiloglou N, Hüsing A, Masala G, Van Gils CH, Turzanski Fortner R, Chang-Claude J, et al. Adherence to the World Cancer Research Fund/American Institute for Cancer Research cancer prevention recommendations and risk of in situ breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. BMC Med. 2019 Dec;17(1):221.

254. Clinton SK, Giovannucci EL, Hursting SD. The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet, Nutrition, Physical Activity, and Cancer: Impact and Future Directions. The Journal of Nutrition. 2020 Apr;150(4):663–71.
255. Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Soft [Internet]. 2011 [cited 2023 Nov 11];45(3). Available from: http://www.jstatsoft.org/v45/i03/
256. Appierto V, Callari M, Cavadini E, Morelli D, Daidone MG, Tiberio P. A lipemia-independent NanoDrop ® -based score to identify hemolysis in plasma and serum samples. Bioanalysis. 2014 May;6(9):1215–26.
257. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.
258. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923–30.
259. Düren Y, Lederer J, Qin LX. Depth normalization of small RNA sequencing: using data and biology to select a suitable method. Nucleic Acids Research. 2022 Jun 10;50(10):e56–e56.
260. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Research. 2019 Jan 8;47(D1):D155–62.
261. Urru S. Breast Cancer Prevention: Statistical Analysis Of Some Risk Factors [Master's thesis]. [Turin]: University of Turin; 2020.
262. Le A, Szaumkessel M, Tan T, Thiery JP, Thompson E, Dobrovic A. DNA Methylation Profiling of Breast Cancer Cell Lines along the Epithelial Mesenchymal Spectrum—Implications for the Choice of Circulating Tumour DNA Methylation Markers. IJMS. 2018 Aug 28;19(9):2553.
263. Ritz C, Spiess AN. *qpcR*: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinformatics. 2008 Jul 1;24(13):1549–51.
264. Tse MY, Ashbury JE, Zwingerman N, King WD, Taylor SA, Pang SC. A refined, rapid and reproducible high resolution melt (HRM)-based method suitable for quantification of global LINE-1 repetitive element methylation. BMC Res Notes. 2011 Dec;4(1):565.
265. Rödiger S, Böhm A, Schimke I. Surface Melting Curve Analysis with R. The R Journal. 2013;5(2):37.
266. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions [Internet]. 2022 [cited 2022 Apr 25]. Available from: https://cran.r-project.org/web/packages/boot/index.html
267. Lachenbruch PA. Analysis of data with excess zeros. Stat Methods Med Res. 2002 Aug;11(4):297–302.
268. Neuhäuser M, Boes T, Jöckel KH. Two-part permutation tests for DNA methylation and microarray data. BMC Bioinformatics. 2005;6(1):35.
269. Rigby RA, Stasinopoulos DM. Generalized Additive Models for Location, Scale and Shape. Journal of the Royal Statistical Society Series C: Applied Statistics. 2005 Jun 1;54(3):507–54.
270. Henningsen A. censReg: Censored Regression (Tobit) Models [Internet]. 2017 [cited 2022 May 25]. Available from: http://CRAN.R-Project.org/package=censReg.
271. Deng Y, Zhu Y, Wang H, Khadka VS, Hu L, Ai J, et al. Ratio-Based Method To Identify True Biomarkers by Normalizing Circulating ncRNA Sequencing and Quantitative PCR Data. Anal Chem. 2019 May 21;91(10):6746–53.
272. Tay JK, Narasimhan B, Hastie T. Elastic Net Regularization Paths for All Generalized Linear Models. J Stat Soft [Internet]. 2023 [cited 2023 Nov 11];106(1). Available from: https://www.jstatsoft.org/v106/i01/
273. Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958 Dec;45(3/4):562.

274. Colombo M, McKeigue P. hsstan: Hierarchical Shrinkage Stan Models for Biomarker Selection [Internet]. 2021 [cited 2022 Jul 12]. Available from: https://CRAN.R-project.org/package=hsstan
275. Campo BDC. Towards reliable predictive analytics: a generalized calibration framework [Internet]. arXiv; 2023 [cited 2024 Feb 10]. Available from: http://arxiv.org/abs/2309.08559
276. Campo BDC, Nieboer D, Van Calster B, Steyerberg E, Vergouwe Y. CalibrationCurves: Calibration Performance [Internet]. 2023 [cited 2024 Feb 10]. Available from: https://cran.r-project.org/web/packages/CalibrationCurves/index.html
277. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016 Jun;74:167–76.
278. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TPA, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models: Method selection to update a prediction model. Statist Med. 2017 Dec 10;36(28):4529–39.
279. Siregar S, Nieboer D, Versteegh MIM, Steyerberg EW, Takkenberg JJM. Methods for updating a risk prediction model for cardiac surgery: a statistical primer. Interactive CardioVascular and Thoracic Surgery. 2019 Mar 1;28(3):333–8.
280. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed., thoroughly rev. and updated. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008. 758 p.
281. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022 Dec 12;22(1):316.
282. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. Statistics in Medicine. 2021 Jul 10;40(15):3533–59.
283. Ruidong Li HQ. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, mRNA, and miRNA data in GDC [Internet]. Bioconductor; 2018 [cited 2023 Nov 11]. Available from: https://bioconductor.org/packages/GDCRNATools
284. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000. 350 p. (Statistics for biology and health).
285. Therneau TM, Thomas L, Elizabeth A, Cynthia C. survival: Survival Analysis [Internet]. 2023 [cited 2024 Feb 11]. Available from: https://cran.r-project.org/web/packages/survival/index.html
286. Licursi V, Conte F, Fiscon G, Paci P. MIENTURNET: an interactive web tool for microRNA-target enrichment and network-based analysis. BMC Bioinformatics. 2019 Dec;20(1):545.
287. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. The Lancet Oncology. 2012 Nov;13(11):1141–51.
288. Bode HF, He L, Hjelmborg J, Kaprio J, Ollikainen M. Pre-diagnosis blood DNA methylation profiling of twin pairs discordant for breast cancer points to the importance of environmental risk [Internet]. Genetic and Genomic Medicine; 2023 Aug [cited 2024 Jan 26]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.08.15.23293985
289. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research. 2016 May 5;44(8):e71–e71.
290. Pott S, Lieb JD. What are super-enhancers? Nat Genet. 2015 Jan;47(1):8–12.
291. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell. 2013 Apr;153(2):307–19.

292. Wang Y, Song C, Zhao J, Zhang Y, Zhao X, Feng C, et al. SEdb 2.0: a comprehensive super-enhancer database of human and mouse. Nucleic Acids Research. 2023 Jan 6;51(D1):D280–90.
293. The ENCODE Project Consortium, Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul 30;583(7818):699–710.
294. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. 2016;2016:baw100.
295. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.
296. Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al. A single–cell type transcriptomics map of human tissues. Sci Adv. 2021 Jul 30;7(31):eabh2169.
297. Wu Y, Duan P, Wen Y, Zhang J, Wang X, Dong J, et al. UHRF1 establishes crosstalk between somatic and germ cells in male reproduction. Cell Death Dis. 2022 Apr 19;13(4):377.
298. Luo G, Li Q, Yu M, Wang T, Zang Y, Liu Z, et al. UHRF1 modulates breast cancer cell growth via estrogen signaling. Med Oncol. 2022 Aug;39(8):111.
299. Yang Y, Liu G, Qin L, Ye L, Zhu F, Ying Y. Overexpression of UHRF1 and its potential role in the development of invasive ductal breast cancer validated by integrative bioinformatics and immunohistochemistry analyses. Transl Cancer Res. 2019 Aug;8(4):1086–96.
300. Moise-Silverman J, Silverman LA. A review of the genetics and epigenetics of central precocious puberty. Front Endocrinol. 2022 Dec 2;13:1029137.
301. Ma Y, Xu Z, Zhao J, Shen H. Novel compound heterozygous mutations of *PCNT* gene in MOPD type II with central precocious puberty. Gynecological Endocrinology. 2021 Feb 1;37(2):190–2.
302. Zhou J, Chen C, Zhao X, Jiang T, Jiang Y, Dai J, et al. Coding variants in the PCNT and CEP295 genes contribute to breast cancer risk in Chinese women. Pathology - Research and Practice. 2021 Sep;225:153581.
303. Cai H, Liu B, Wang H, Sun G, Feng L, Chen Z, et al. SP1 governs primordial folliculogenesis by regulating pregranulosa cell development in mice. Journal of Molecular Cell Biology. 2020 Apr 24;12(3):230–44.
304. Maor S, Mayer D, Yarden RI, Lee AV, Sarfstein R, Werner H, et al. Estrogen receptor regulates insulin-like growth factor-I receptor gene expression in breast tumor cells: involvement of transcription factor Sp1. Journal of Endocrinology. 2006 Dec;191(3):605–12.
305. Gao Y, Gan K, Liu K, Xu B, Chen M. SP1 Expression and the Clinicopathological Features of Tumors: A Meta-Analysis and Bioinformatics Analysis. Pathol Oncol Res. 2021 Jan 28;27:581998.
306. Maud C, Ryan J, McIntosh JE, Olsson CA. The role of oxytocin receptor gene (OXTR) DNA methylation (DNAm) in human social and emotional functioning: a systematic narrative review. BMC Psychiatry. 2018 Dec;18(1):154.
307. Creswell KG, Wright AGC, Troxel WM, Ferrell RE, Flory JD, Manuck SB. OXTR polymorphism predicts social relationships through its effects on social temperament. Social Cognitive and Affective Neuroscience. 2015 Jun 1;10(6):869–76.
308. Liu H, Gruber CW, Alewood PF, Möller A, Muttenthaler M. The oxytocin receptor signalling system and breast cancer: a critical review. Oncogene. 2020 Sep 10;39(37):5917–32.
309. Ariana M, Pornour M, Mehr SS, Vaseghi H, Ganji SM, Alivand MR, et al. Preventive effects of oxytocin and oxytocin receptor in breast cancer pathogenesis. Personalized Medicine. 2019 Jan;16(1):25–34.

310. Li D, San M, Zhang J, Yang A, Xie W, Chen Y, et al. Oxytocin receptor induces mammary tumorigenesis through prolactin/p-STAT5 pathway. Cell Death Dis. 2021 Jun 7;12(6):588.
311. Karedath T, Ahmed I, Al Ameri W, Al-Dasim FM, Andrews SS, Samuel S, et al. Silencing of ANKRD12 circRNA induces molecular and functional changes associated with invasive phenotypes. BMC Cancer. 2019 Dec;19(1):565.
312. Sovio U, Bennett AJ, Millwood IY, Molitor J, O'Reilly PF, Timpson NJ, et al. Genetic Determinants of Height Growth Assessed Longitudinally from Infancy to Adulthood in the Northern Finland Birth Cohort 1966. Gibson G, editor. PLoS Genet. 2009 Mar 6;5(3):e1000409.
313. Choi YJ, Li X, Hydbring P, Sanda T, Stefano J, Christie AL, et al. The Requirement for Cyclin D Function in Tumor Maintenance. Cancer Cell. 2012 Oct;22(4):438–51.
314. Watt AC, Goel S. Cellular mechanisms underlying response and resistance to CDK4/6 inhibitors in the treatment of hormone receptor-positive breast cancer. Breast Cancer Res. 2022 Dec;24(1):17.
315. Nebenfuehr S, Kollmann K, Sexl V. The role of CDK6 in cancer. Intl Journal of Cancer. 2020 Dec;147(11):2988–95.
316. Ghafouri-Fard S, Khoshbakht T, Hussen BM, Dong P, Gassler N, Taheri M, et al. A review on the role of cyclin dependent kinases in cancers. Cancer Cell Int. 2022 Oct 20;22(1):325.
317. Yadav DK, Sharma A, Dube P, Shaikh S, Vaghasia H, Rawal RM. Identification of crucial hub genes and potential molecular mechanisms in breast cancer by integrated bioinformatics analysis and experimental validation. Computers in Biology and Medicine. 2022 Oct;149:106036.
318. Chen Q, Xu H, Zhu J, Feng K, Hu C. LncRNA MCM3AP-AS1 promotes breast cancer progression via modulating miR-28-5p/CENPF axis. Biomedicine & Pharmacotherapy. 2020 Aug;128:110289.
319. Erkko H, Pylkäs K, Karppinen SM, Winqvist R. Germline alterations in the CLSPN gene in breast cancer families. Cancer Letters. 2008 Mar;261(1):93–7.
320. Trott JF, Schennink A, Horigan KC, Lemay DG, Cohen JR, Famula TR, et al. Unique Transcriptomic Changes Underlie Hormonal Interactions During Mammary Histomorphogenesis in Female Pigs. Endocrinology. 2022 Mar 1;163(3):bqab256.
321. Azenha D, Hernandez-Perez S, Martin Y, Viegas MS, Martins A, Lopes MC, et al. Implications of CLSPN Variants in Cellular Function and Susceptibility to Cancer. Cancers (Basel). 2020 Aug 24;12(9):2396.
322. Iino K, Mitobe Y, Ikeda K, Takayama K, Suzuki T, Kawabata H, et al. RNA-binding protein NONO promotes breast cancer proliferation by post-transcriptional regulation of *SKP2* and *E2F8*. Cancer Science. 2020 Jan;111(1):148–59.
323. Ye L, Guo L, He Z, Wang X, Lin C, Zhang X, et al. Upregulation of E2F8 promotes cell proliferation and tumorigenicity in breast cancer by modulating G1/S phase transition. Oncotarget. 2016 Apr 26;7(17):23757–71.
324. Dong J, Huang S, Caikovski M, Ji S, McGrath A, Custorio MG, et al. ID4 regulates mammary gland development by suppressing p38MAPK activity. Development. 2011 Dec 1;138(23):5247–56.
325. Holliday H, Roden D, Junankar S, Wu SZ, Baker LA, Krisp C, et al. Inhibitor of Differentiation 4 (ID4) represses mammary myoepithelial differentiation via inhibition of HEB. iScience. 2021 Feb;24(2):102072.
326. Junankar S, Baker LA, Roden DL, Nair R, Elsworth B, Gallego-Ortega D, et al. ID4 controls mammary stem cells and marks breast cancers with a stem cell-like phenotype. Nat Commun. 2015 Mar 27;6(1):6548.
327. Nasif D, Campoy E, Laurito S, Branham R, Urrutia G, Roqué M, et al. Epigenetic regulation of ID4 in breast cancer: tumor suppressor or oncogene? Clin Epigenetics. 2018 Aug 23;10(1):111.

328. Zhai X, Yang Z, Liu X, Dong Z, Zhou D. Identification of NUF2 and FAM83D as potential biomarkers in triple-negative breast cancer. PeerJ. 2020 Sep 21;8:e9975.
329. Lv S, Xu W, Zhang Y, Zhang J, Dong X. NUF2 as an anticancer therapeutic target and prognostic factor in breast cancer. Int J Oncol. 2020 Oct 26;57(6):1358–67.
330. Wen H, Chen Z, Li M, Huang Q, Deng Y, Zheng J, et al. An Integrative Pan-Cancer Analysis of PBK in Human Tumors. Front Mol Biosci. 2021 Nov 10;8:755911.
331. Qiao L, Ba J, Xie J, Zhu R, Wan Y, Zhang M, et al. Overexpression of PBK/TOPK relates to poor prognosis of patients with breast cancer: a retrospective analysis. World J Surg Onc. 2022 Sep 28;20(1):316.
332. Pollaci G, Gorla G, Potenza A, Carrozzini T, Canavero I, Bersano A, et al. Novel Multifaceted Roles for RNF213 Protein. IJMS. 2022 Apr 19;23(9):4492.
333. Kim SH, On J won, Pyo H, Ko KS, Won JC, Yang J, et al. Percentage fractions of urinary di(2-ethylhexyl) phthalate metabolites: Association with obesity and insulin resistance in Korean girls. Baixeras E, editor. PLoS ONE. 2018 Nov 27;13(11):e0208081.
334. Abdel-Rahman MA, Mahfouz M, Habashy HO. RRM2 expression in different molecular subtypes of breast cancer and its prognostic significance. Diagn Pathol. 2022 Dec;17(1):1.
335. Wilson EA, Sultana N, Shah KN, Elford HL, Faridi JS. Molecular Targeting of RRM2, NF-κB, and Mutant TP53 for the Treatment of Triple-Negative Breast Cancer. Molecular Cancer Therapeutics. 2021 Apr 1;20(4):655–64.
336. Zlotina A, Kiselev A, Sergushichev A, Parmon E, Kostareva A. Rare Case of Ulnar-Mammary-Like Syndrome With Left Ventricular Tachycardia and Lack of TBX3 Mutation. Front Genet. 2018 Jun 15;9:209.
337. Maggi L, Mavroidis M, Psarras S, Capetanaki Y, Lattanzi G. Skeletal and Cardiac Muscle Disorders Caused by Mutations in Genes Encoding Intermediate Filament Proteins. Int J Mol Sci. 2021 Apr 20;22(8):4256.
338. Noetzel E, Rose M, Sevinc E, Hilgers RD, Hartmann A, Naami A, et al. Intermediate filament dynamics and breast cancer: Aberrant promoter methylation of the Synemin gene is associated with early tumor relapse. Oncogene. 2010 Aug 26;29(34):4814–25.
339. Ning X, Zhao J, He F, Yuan Y, Li B, Ruan J. Long non-coding RNA TMPO-AS1 facilitates chemoresistance and invasion in breast cancer by modulating the miR-1179/TRIM37 axis. Oncol Lett. 2021 Apr 28;22(1):500.
340. Zhu D, Lv W, Zhou X, He Y, Yao H, Yu Y, et al. Long non-coding RNA TMPO-AS1 promotes tumor progression via sponging miR-140-5p in breast cancer. Exp Ther Med. 2020 Nov 5;21(1):1–1.
341. Yang Y, Li DP, Shen N, Yu XC, Li JB, Song Q, et al. TPX2 promotes migration and invasion of human breast cancer cells. Asian Pacific Journal of Tropical Medicine. 2015 Dec;8(12):1064–70.
342. Wang T, Zhang F, Zhang P. Role of the TPX2/NCOA5 axis in regulating proliferation, migration, invasion and angiogenesis of breast cancer cells. Exp Ther Med. 2023 May 9;25(6):304.
343. Ashraf W, Ibrahim A, Alhosin M, Zaayter L, Ouararhni K, Papin C, et al. The epigenetic integrator UHRF1: on the road to become a universal biomarker for cancer. Oncotarget. 2017 Aug 1;8(31):51946–62.
344. Bostick M, Kim JK, Estève PO, Clark A, Pradhan S, Jacobsen SE. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science. 2007 Sep 21;317(5845):1760–4.
345. Zhang J, Zhou YJ, Yu ZH, Chen AX, Yu Y, Wang X, et al. Identification of core genes and clinical roles in pregnancy-associated breast cancer based on integrated analysis of different microarray profile datasets. Bioscience Reports. 2019 Jun 28;39(6):BSR20190019.

346. Middelberg RPS, Heath AC, Madden PAF, Montgomery GW, Martin NG, Whitfield JB. Evidence of Differential Allelic Effects between Adolescents and Adults for Plasma High-Density Lipoprotein. Stoll M, editor. PLoS ONE. 2012 Apr 18;7(4):e35605.
347. Kao J, Salari K, Bocanegra M, Choi YL, Girard L, Gandhi J, et al. Molecular Profiling of Breast Cancer Cell Lines Defines Relevant Tumor Models and Provides a Resource for Cancer Gene Discovery. Blagosklonny MV, editor. PLoS ONE. 2009 Jul 3;4(7):e6146.
348. Birnbaum JK, Duggan C, Anderson BO, Etzioni R. Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study. The Lancet Global Health. 2018 Aug;6(8):e885–93.
349. Vineis P, Wild CP. Global cancer patterns: causes and prevention. The Lancet. 2014 Feb;383(9916):549–57.
350. Osborne M, Boyle P, Lipkin M. Cancer prevention. The Lancet. 1997 May;349:S27–30.
351. Adami HO, Day NE, Trichopoulos D, Willett WC. Primary and secondary prevention in the reduction of cancer morbidity and mortality. European Journal of Cancer. 2001 Sep;37:118–27.
352. Moshina N, Falk RS, Botteri E, Larsen M, Akslen LA, Cairns JA, et al. Quality of life among women with symptomatic, screen-detected, and interval breast cancer, and for women without breast cancer: a retrospective cross-sectional study from Norway. Qual Life Res. 2022 Apr;31(4):1057–68.
353. Miller SM, Bowen DJ, Lyle J, Clark M, Mohr D, Wardle J, et al. Primary prevention, aging, and cancer: Overview and future perspectives. Cancer. 2008 Dec 3;113(S12):3484–92.
354. McGarvey N, Gitlin M, Fadli E, Chung KC. Increased healthcare costs by later stage cancer diagnosis. BMC Health Serv Res. 2022 Sep 13;22(1):1155.
355. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2017 Apr;1867(2):151–61.
356. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 2017 Apr;18(4):213–29.
357. Conte L, De Nunzio G, Lupo R, Mieli M, Lezzi A, Vitale E, et al. Breast Cancer Prevention: The Key Role of Population Screening, Breast Self-Examination (BSE) and Technological Tools. Survey of Italian Women. J Canc Educ. 2023 Oct;38(5):1728–42.
358. Pashayan N, Duffy SW, Chowdhury S, Dent T, Burton H, Neal DE, et al. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. Br J Cancer. 2011 May;104(10):1656–63.
359. Fang C, Jian ZY, Shen XF, Wei XM, Yu GZ, Zeng XT. Promoter Methylation of the Retinoic Acid Receptor Beta2 (RARβ2) Is Associated with Increased Risk of Breast Cancer: A PRISMA Compliant Meta-Analysis. Coleman WB, editor. PLoS ONE. 2015 Oct 9;10(10):e0140329.
360. Gao Q, Zeng Q, Wang Z, Li C, Xu Y, Cui P, et al. Circulating cell-free DNA for cancer early detection. The Innovation. 2022 Jul;3(4):100259.
361. Umu SU, Langseth H, Bucher-Johannessen C, Fromm B, Keller A, Meese E, et al. A comprehensive profile of circulating RNAs in human serum. RNA Biology. 2018 Feb 1;15(2):242–50.
362. Fang R, Zhu Y, Hu L, Khadka VS, Ai J, Zou H, et al. Plasma MicroRNA Pair Panels as Novel Biomarkers for Detection of Early Stage Breast Cancer. Front Physiol. 2019 Jan 8;9:1879.
363. Frères P, Wenric S, Boukerroucha M, Fasquelle C, Thiry J, Bovy N, et al. Circulating microRNA-based screening tool for breast cancer. Oncotarget. 2015 Dec 29;7(5):5416–28.
364. Guo J, Liu C, Wang W, Liu Y, He H, Chen C, et al. Identification of serum miR-1915-3p and miR-455-3p as biomarkers for breast cancer. Coleman WB, editor. PLoS ONE. 2018 Jul 26;13(7):e0200716.

365. Uyisenga JP, Debit A, Poulet C, Frères P, Poncin A, Thiry J, et al. Differences in plasma microRNA content impair microRNA-based signature for breast cancer diagnosis in cohorts recruited from heterogeneous environmental sites. Sci Rep. 2021 Jun 3;11(1):11698.
366. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. JNCI: Journal of the National Cancer Institute. 2008 Oct 15;100(20):1432–8.
367. Tiberio P, Callari M, Angeloni V, Daidone MG, Appierto V. Challenges in Using Circulating miRNAs as Cancer Biomarkers. BioMed Research International. 2015;2015:1–10.
368. Kirschner MB, Edelman JJB, Kao SCH, Vallely MP, Van Zandwijk N, Reid G. The Impact of Hemolysis on Cell-Free microRNA Biomarkers. Front Genet [Internet]. 2013 [cited 2023 Nov 6];4. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2013.00094/abstract
369. Kirschner MB, Kao SC, Edelman JJ, Armstrong NJ, Vallely MP, Van Zandwijk N, et al. Haemolysis during Sample Preparation Alters microRNA Content of Plasma. Pfeffer S, editor. PLoS ONE. 2011 Sep 1;6(9):e24145.
370. Felekkis K, Papaneophytou C. Challenges in Using Circulating Micro-RNAs as Biomarkers for Cardiovascular Diseases. IJMS. 2020 Jan 15;21(2):561.
371. Yamada A, Cox MA, Gaffney KA, Moreland A, Boland CR, Goel A. Technical Factors Involved in the Measurement of Circulating MicroRNA Biomarkers for the Detection of Colorectal Neoplasia. Calin G, editor. PLoS ONE. 2014 Nov 18;9(11):e112481.
372. Pizzamiglio S, Zanutto S, Ciniselli CM, Belfiore A, Bottelli S, Gariboldi M, et al. A methodological procedure for evaluating the impact of hemolysis on circulating microRNAs. Oncology Letters. 2017 Jan;13(1):315–20.
373. Li S, Yang X, Yang J, Zhen J, Zhang D. Serum microRNA-21 as a potential diagnostic biomarker for breast cancer: a systematic review and meta-analysis. Clin Exp Med. 2016 Feb;16(1):29–35.
374. Hansen C, Steinmetz H, Block J. How to conduct a meta-analysis in eight steps: a practical guide. Manag Rev Q. 2022 Feb;72(1):1–19.
375. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research Synthesis Methods. 2020 Mar;11(2):181–217.
376. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017 Dec;6(1):245.
377. the Andromeda working group, Giordano L, Gallo F, Petracci E, Chiorino G, Segnan N. The ANDROMEDA prospective cohort study: predictive value of combined criteria to tailor breast cancer screening and new opportunities from circulating markers: study protocol. BMC Cancer. 2017 Dec;17(1):785.
378. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of Biomarker Development for Early Detection of Cancer. JNCI Journal of the National Cancer Institute. 2001 Jul 18;93(14):1054–61.
379. Pati S, Irfan W, Jameel A, Ahmed S, Shahid RK. Obesity and Cancer: A Current Overview of Epidemiology, Pathogenesis, Outcomes, and Management. Cancers (Basel). 2023 Jan 12;15(2):485.
380. Malcomson FC, Wiggins C, Parra-Soto S, Ho FK, Celis-Morales C, Sharp L, et al. Adherence to the 2018 World Cancer Research Fund/American Institute for Cancer Research Cancer Prevention Recommendations and cancer risk: A systematic review and meta-analysis. Cancer. 2023 Sep;129(17):2655–70.
381. Fanidi A, Ferrari P, Biessy C, Ortega C, Angeles-Llerenas A, Torres-Mejia G, et al. Adherence to the World Cancer Research Fund/American Institute for Cancer Research cancer prevention

recommendations and breast cancer risk in the Cancer de Màma (CAMA) study. Public Health Nutr. 2015 Dec;18(18):3337–48.

382. World Cancer Research Fund/American Institute for Cancer Research. Diet, nutrition, physical activity and cancer: a global perspective. Continuous update project expert report. 2018.
383. Bodewes FTH, Van Asselt AA, Dorrius MD, Greuter MJW, De Bock GH. Mammographic breast density and the risk of breast cancer: A systematic review and meta-analysis. The Breast. 2022 Dec;66:62–8.
384. Román M, Louro J, Posso M, Vidal C, Bargalló X, Vázquez I, et al. Long-Term Risk of Breast Cancer after Diagnosis of Benign Breast Disease by Screening Mammography. Int J Environ Res Public Health. 2022 Feb 24;19(5):2625.
385. Gaudet MM, Carter BD, Patel AV, Teras LR, Jacobs EJ, Gapstur SM. Waist circumference, body mass index, and postmenopausal breast cancer incidence in the Cancer Prevention Study-II Nutrition Cohort. Cancer Causes Control. 2014 Jun;25(6):737–45.
386. Lakshminarasimhan R, Liang G. The Role of DNA Methylation in Cancer. In: Jeltsch A, Jurkowska RZ, editors. DNA Methyltransferases - Role and Function [Internet]. Cham: Springer International Publishing; 2016 [cited 2023 Nov 3]. p. 151–72. (Advances in Experimental Medicine and Biology; vol. 945). Available from: http://link.springer.com/10.1007/978-3-319-43624-1_7
387. Dhar GA, Saha S, Mitra P, Nag Chaudhuri R. DNA methylation and regulation of gene expression: Guardian of our health. Nucleus (Calcutta). 2021;64(3):259–70.
388. Ankill J, Aure MR, Bjørklund S, Langberg S, Oslo Breast Cancer Consortium (OSBREAC), Bathen TF, et al. Epigenetic alterations at distal enhancers are linked to proliferation in human breast cancer. NAR Cancer. 2022 Jan 13;4(1):zcac008.
389. Danos P, Giannoni-Luza S, Murillo Carrasco AG, Acosta O, Guevara-Fujita ML, Cotrina Concha JM, et al. Promoter hypermethylation of *RARB* and *GSTP1* genes in plasma cell-free DNA as breast cancer biomarkers in Peruvian women. Molec Gen & Gen Med. 2023 Aug 7;e2260.
390. Zhou D, Tang W, Wang W, Pan X, An HX, Zhang Y. Association between aberrant APC promoter methylation and breast cancer pathogenesis: a meta-analysis of 35 observational studies. PeerJ. 2016;4:e2203.
391. Hsu NC, Huang YF, Yokoyama KK, Chu PY, Chen FM, Hou MF. Methylation of BRCA1 Promoter Region Is Associated with Unfavorable Prognosis in Women with Early-Stage Breast Cancer. Aboussekhra A, editor. PLoS ONE. 2013 Feb 6;8(2):e56256.
392. Jin Z, Tamura G, Tsuchiya T, Sakata K, Kashiwaba M, Osakabe M, et al. Adenomatous polyposis coli (APC) gene promoter hypermethylation in primary breast cancers. Br J Cancer. 2001 Jul 6;85(1):69–73.
393. De Ruijter TC, Van Der Heide F, Smits KM, Aarts MJ, Van Engeland M, Heijnen VCG. Prognostic DNA methylation markers for hormone receptor breast cancer: a systematic review. Breast Cancer Res. 2020 Dec;22(1):13.
394. Aggarwal V, Priyanka K, Tuli HS. Emergence of Circulating MicroRNAs in Breast Cancer as Diagnostic and Therapeutic Efficacy Biomarkers. Mol Diagn Ther. 2020 Apr;24(2):153–73.
395. Klinge C. Non-Coding RNAs in Breast Cancer: Intracellular and Intercellular Communication. ncRNA. 2018 Dec 12;4(4):40.
396. Dvorská D, Braný D, Ňachajová M, Halašová E, Danková Z. Breast Cancer and the Other Non-Coding RNAs. IJMS. 2021 Mar 23;22(6):3280.
397. Faramin Lashkarian M, Hashemipour N, Niaraki N, Soghala S, Moradi A, Sarhangi S, et al. MicroRNA-122 in human cancers: from mechanistic to clinical perspectives. Cancer Cell Int. 2023 Feb 20;23(1):29.

398. Maierthaler M, Benner A, Hoffmeister M, Surowy H, Jansen L, Knebel P, et al. Plasma miR-122 and miR-200 family are prognostic markers in colorectal cancer. Intl Journal of Cancer. 2017 Jan;140(1):176–87.
399. Chen Q, Ge X, Zhang Y, Xia H, Yuan D, Tang Q, et al. Plasma miR-122 and miR-192 as potential novel biomarkers for the early detection of distant metastasis of gastric cancer. Oncology Reports. 2014 Apr;31(4):1863–70.
400. Laterza OF, Scott MG, Garrett-Engele PW, Korenblat KM, Lockwood CM. Circulating miR-122 as a potential biomarker of liver disease. Biomarkers in Medicine. 2013 Apr;7(2):205–10.
401. Li M, Zou X, Xia T, Wang T, Liu P, Zhou X, et al. A five-miRNA panel in plasma was identified for breast cancer diagnosis. Cancer Medicine. 2019 Nov;8(16):7006–17.
402. Wang B, Wang H, Yang Z. MiR-122 Inhibits Cell Proliferation and Tumorigenesis of Breast Cancer by Targeting IGF1R. Mukhopadhyay P, editor. PLoS ONE. 2012 Oct 8;7(10):e47053.
403. Fong MY, Zhou W, Liu L, Alontaga AY, Chandra M, Ashby J, et al. Breast-cancer-secreted miR-122 reprograms glucose metabolism in premetastatic niche to promote metastasis. Nat Cell Biol. 2015 Feb;17(2):183–94.
404. Erturk E, Enes Onur O, Akgun O, Tuna G, Yildiz Y, Ari F. Mitochondrial miRNAs (MitomiRs): Their potential roles in breast and other cancers. Mitochondrion. 2022 Sep;66:74–81.
405. Christopoulos PF, Msaouel P, Koutsilieris M. The role of the insulin-like growth factor-1 system in breast cancer. Mol Cancer. 2015;14(1):43.
406. Paplomata E, O'Regan R. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. Ther Adv Med Oncol. 2014 Jul;6(4):154–66.
407. Li M, Pan M, Wang J, You C, Zhao F, Zheng D, et al. miR-7 Reduces Breast Cancer Stem Cell Metastasis via Inhibiting RELA to Decrease ESAM Expression. Molecular Therapy - Oncolytics. 2020 Sep;18:70–82.
408. Shi Y, Luo X, Li P, Tan J, Wang X, Xiang T, et al. miR-7-5p suppresses cell proliferation and induces apoptosis of breast cancer cells mainly by targeting REGγ. Cancer Letters. 2015 Mar;358(1):27–36.
409. Gao J, Li L, Wu M, Liu M, Xie X, Guo J, et al. MiR-26a Inhibits Proliferation and Migration of Breast Cancer through Repression of MCL-1. Cheng JQ, editor. PLoS ONE. 2013 Jun 4;8(6):e65138.
410. Liu T, Wang Z, Dong M, Wei J, Pan Y. MicroRNA-26a inhibits cell proliferation and invasion by targeting FAM98A in breast cancer. Oncol Lett. 2021 Mar 10;21(5):367.
411. Miller TW, Pérez-Torres M, Narasanna A, Guix M, Stål O, Pérez-Tenorio G, et al. Loss of *Phosphatase and Tensin Homologue Deleted on Chromosome 10* Engages ErbB3 and Insulin-Like Growth Factor-I Receptor Signaling to Promote Antiestrogen Resistance in Breast Cancer. Cancer Research. 2009 May 15;69(10):4192–201.
412. Bardoni B. 82-FIP, a novel FMRP (Fragile X Mental Retardation Protein) interacting protein, shows a cell cycle-dependent intracellular localization. Human Molecular Genetics. 2003 Jul 15;12(14):1689–98.
413. Huang S kai, Luo Q, Peng H, Li J, Zhao M, Wang J, et al. A Panel of Serum Noncoding RNAs for the Diagnosis and Monitoring of Response to Therapy in Patients with Breast Cancer. Med Sci Monit. 2018 Apr 23;24:2476–88.
414. Zou X, Xia T, Li M, Wang T, Liu P, Zhou X, et al. MicroRNA profiling in serum: Potential signatures for breast cancer diagnosis. CBM. 2021 Feb 9;30(1):41–53.
415. Li M, Zhou Y, Xia T, Zhou X, Huang Z, Zhang H, et al. Circulating microRNAs from the miR-106a–363 cluster on chromosome X as novel diagnostic biomarkers for breast cancer. Breast Cancer Res Treat. 2018 Jul;170(2):257–70.

416. Eichelser C, Flesch-Janys D, Chang-Claude J, Pantel K, Schwarzenbach H. Deregulated Serum Concentrations of Circulating Cell–Free MicroRNAs miR-17, miR-34a, miR-155, and miR-373 in Human Breast Cancer Development and Progression. Clinical Chemistry. 2013 Oct 1;59(10):1489–96.
417. Yu X, Liang J, Xu J, Li X, Xing S, Li H, et al. Identification and Validation of Circulating MicroRNA Signatures for Breast Cancer Early Detection Based on Large Scale Tissue-Derived Data. J Breast Cancer. 2018;21(4):363.
418. Al-Harbi B, Hendrayani SF, Silva G, Aboussekhra A. Let-7b inhibits cancer-promoting effects of breast cancer-associated fibroblasts through IL-8 repression. Oncotarget. 2018 Apr 3;9(25):17825–38.
419. Xiang Y, Liao XH, Yu CX, Yao A, Qin H, Li JP, et al. MiR-93-5p inhibits the EMT of breast cancer cells via targeting MKL-1 and STAT3. Experimental Cell Research. 2017 Aug;357(1):135–44.
420. Liu K, Zhang C, Li T, Ding Y, Tu T, Zhou F, et al. Let-7a inhibits growth and migration of breast cancer cells by targeting HMGA1. International Journal of Oncology. 2015 Jun;46(6):2526–34.
421. Kim SJ, Shin JY, Lee KD, Bae YK, Sung KW, Nam SJ, et al. MicroRNA let-7a suppresses breast cancer cell migration and invasion through downregulation of C-C chemokine receptor type 7. Breast Cancer Res. 2012 Feb;14(1):R14.
422. Heydari N, Nikbakhsh N, Sadeghi F, Farnoush N, Khafri S, Bastami M, et al. Overexpression of serum MicroRNA-140-3p in premenopausal women with newly diagnosed breast cancer. Gene. 2018 May;655:25–9.
423. Peña-Cano MI, Saucedo R, Morales-Avila E, Valencia J, Zavala-Moha JA, López A. Deregulated microRNAs and Adiponectin in Postmenopausal Women with Breast Cancer. Gynecol Obstet Invest. 2019;84(4):369–77.
424. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal. 2014 Aug 1;35(29):1925–31.
425. Zou R, Loke SY, Tang YC, Too HP, Zhou L, Lee ASG, et al. Development and validation of a circulating microRNA panel for the early detection of breast cancer. Br J Cancer. 2022 Feb 1;126(3):472–81.
426. Zen K, Zhang C. Circulating MicroRNAs: a novel class of biomarkers to diagnose and monitor human cancers. Medicinal Research Reviews. 2012 Mar;32(2):326–48.
427. Li L, Sun Y, Feng M, Wang L, Liu J. Clinical significance of blood-based miRNAs as biomarkers of non-small cell lung cancer (Review). Oncol Lett [Internet]. 2018 Apr 12 [cited 2024 Feb 13]; Available from: http://www.spandidos-publications.com/10.3892/ol.2018.8469
428. Van Calster B, Van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res. 2020 Nov;29(11):3166–78.
429. Pratt AJ, MacRae IJ. The RNA-induced Silencing Complex: A Versatile Gene-silencing Machine. Journal of Biological Chemistry. 2009 Jul;284(27):17897–901.
430. Miller JL, Grant PA. The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans. In: Kundu TK, editor. Epigenetics: Development and Disease [Internet]. Dordrecht: Springer Netherlands; 2013 [cited 2023 Nov 9]. p. 289–317. (Subcellular Biochemistry; vol. 61). Available from: http://link.springer.com/10.1007/978-94-007-4525-4_13
431. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacol. 2013 Jan;38(1):23–38.
432. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019 Oct;20(10):590–607.

433. Almstrup K, Lindhardt Johansen M, Busch AS, Hagen CP, Nielsen JE, Petersen JH, et al. Pubertal development in healthy children is mirrored by DNA methylation patterns in peripheral blood. Sci Rep. 2016 Jun 28;6(1):28657.
434. Thompson EE, Nicodemus-Johnson J, Kim KW, Gern JE, Jackson DJ, Lemanske RF, et al. Global DNA methylation changes spanning puberty are near predicted estrogen-responsive genes and enriched for genes involved in endocrine and immune processes. Clin Epigenet. 2018 Dec;10(1):62.
435. Lim U, Song MA. Dietary and Lifestyle Factors of DNA Methylation. In: Dumitrescu RG, Verma M, editors. Cancer Epigenetics [Internet]. Totowa, NJ: Humana Press; 2012 [cited 2023 Nov 9]. p. 359–76. (Methods in Molecular Biology; vol. 863). Available from: http://link.springer.com/10.1007/978-1-61779-612-8_23
436. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. Genome Biol. 2019 Dec;20(1):249.
437. Motamedi M, Hashemzadeh Chaleshtori M, Ghasemi S, Mokarian F. Plasma Level Of miR-21 And miR-451 In Primary And Recurrent Breast Cancer Patients. BCTT. 2019 Oct;Volume 11:293–301.
438. Mar-Aguilar F, Luna-Aguirre CM, Moreno-Rocha JC, Araiza-Chávez J, Trevino V, Rodríguez-Padilla C, et al. Differential expression of miR-21, miR-125b and miR-191 in breast cancer tissue. Asia-Pac J Clncl Oncology. 2013 Mar;9(1):53–9.
439. Petrović N. miR-21 Might be Involved in Breast Cancer Promotion and Invasion Rather than in Initial Events of Breast Cancer Development. Mol Diagn Ther. 2016 Apr;20(2):97–110.
440. Zhang C, Liu K, Li T, Fang J, Ding Y, Sun L, et al. miR-21: A gene of dual regulation in breast cancer. International Journal of Oncology. 2016 Jan;48(1):161–72.
441. Zhang X, Li Y, Wang D, Wei X. miR-22 suppresses tumorigenesis and improves radiosensitivity of breast cancer cells by targeting Sirt1. Biol Res. 2017 Dec;50(1):27.
442. Ling B, Wang GX, Long G, Qiu JH, Hu ZL. Tumor suppressor miR-22 suppresses lung cancer cell progression through post-transcriptional regulation of ErbB3. J Cancer Res Clin Oncol. 2012 Aug;138(8):1355–61.
443. Pandey DP, Picard D. miR-22 Inhibits Estrogen Signaling by Directly Targeting the Estrogen Receptor α mRNA. Molecular and Cellular Biology. 2009 Jul 1;29(13):3783–90.
444. Ma S, Wei H, Wang C, Han J, Chen X, Li Y. MiR-26b-5p inhibits cell proliferation and EMT by targeting MYCBP in triple-negative breast cancer. Cell Mol Biol Lett. 2021 Dec;26(1):52.
445. Zhou A, Chen G, Cheng X, Zhang C, Xu H, Qi M, et al. Inhibitory effects of miR-26b-5p on thyroid cancer. Mol Med Report [Internet]. 2019 May 31 [cited 2023 Nov 9]; Available from: http://www.spandidos-publications.com/10.3892/mmr.2019.10315
446. Wan M, Huang W, Kute TE, Miller LD, Zhang Q, Hatcher H, et al. Yin Yang 1 Plays an Essential Role in Breast Cancer and Negatively Regulates p27. The American Journal of Pathology. 2012 May;180(5):2120–33.
447. Satelli A, Li S. Vimentin in cancer and its potential as a molecular target for cancer therapy. Cell Mol Life Sci. 2011 Sep;68(18):3033–46.
448. Vora HH, Patel NA, Rajvik KN, Mehta SV, Brahmbhatt BV, Shah MJ, et al. Cytokeratin and Vimentin Expression in Breast Cancer. Int J Biol Markers. 2009 Jan;24(1):38–46.
449. Wang X, Ji S, Ma Y, Xing X, Zhou Y, Xu X, et al. Vimentin plays an important role in the promotion of breast cancer cell migration and invasion by leucine aminopeptidase 3. Cytotechnology. 2020 Oct;72(5):639–47.
450. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. UHRF1 ubiquitin like with PHD and ring finger domains 1 [Homo sapiens (human)] - Gene - NCBI [Internet]. 2004 [cited 2023 Nov 10]. Available from: https://www.ncbi.nlm.nih.gov/gene/29128

451. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Gene: SP1 [Internet]. 2004 [cited 2023 Nov 10]. Available from: https://www.ncbi.nlm.nih.gov/gene/6667
452. Swellam M, Ramadan A, El-Hussieny EA, Bakr NM, Hassan NM, Sobeih ME, et al. Clinical significance of blood-based miRNAs as diagnostic and prognostic nucleic acid markers in breast cancer: Comparative to conventional tumor markers. J Cell Biochem. 2019 Aug;120(8):12321–30.
453. Zhang K, Wang YW, Wang YY, Song Y, Zhu J, Si PC, et al. Identification of microRNA biomarkers in the blood of breast cancer patients based on microRNA profiling. Gene. 2017 Jul;619:10–20.
454. Mar-Aguilar F, Mendoza-Ramírez JA, Malagón-Santiago I, Espino-Silva PK, Santuario-Facio SK, Ruiz-Flores P, et al. Serum Circulating microRNA Profiling for Identification of Potential Breast Cancer Biomarkers. Disease Markers. 2013;34(3):163–9.
455. Wu Q, Wang C, Lu Z, Guo L, Ge Q. Analysis of serum genome-wide microRNAs for breast cancer detection. Clinica Chimica Acta. 2012 Jul;413(13–14):1058–65.
456. Diansyah MN, Prayogo AA, Sedana MP, Savitri M, Romadhon PZ, Amrita PNA, et al. Early detection breast cancer: role of circulating plasma miRNA-21 expression as a potential screening biomarker. Turk J Med Sci. 2021;51(2):562–9.
457. Hosseini Mojahed F, Aalami AH, Pouresmaeil V, Amirabadi A, Qasemi Rad M, Sahebkar A. Clinical Evaluation of the Diagnostic Role of MicroRNA-155 in Breast Cancer. International Journal of Genomics. 2020 Sep 8;2020:1–13.
458. Kim J, Park S, Hwang D, Kim SI, Lee H. Diagnostic Value of Circulating miR-202 in Early-Stage Breast Cancer in South Korea. Medicina. 2020 Jul 9;56(7):340.
459. Swellam M, Zahran RFK, Abo El-Sadat Taha H, El-Khazragy N, Abdel-Malak C. Role of some circulating MiRNAs on breast cancer diagnosis. Archives of Physiology and Biochemistry. 2019 Oct 20;125(5):456–64.
460. Matamala N, Vargas MT, González-Cámpora R, Miñambres R, Arias JI, Menéndez P, et al. Tumor MicroRNA Expression Profiling Identifies Circulating MicroRNAs for Early Breast Cancer Detection. Clinical Chemistry. 2015 Aug 1;61(8):1098–106.
461. Han JG, Jiang YD, Zhang CH, Yang YM, Pang D, Song YN, et al. A novel panel of serum miR-21/miR-155/miR-365 as a potential diagnostic biomarker for breast cancer. Ann Surg Treat Res. 2017;92(2):55.
462. Zhao H, Shen J, Medico L, Wang D, Ambrosone CB, Liu S. A Pilot Study of Circulating miRNAs as Potential Biomarkers of Early Stage Breast Cancer. Creighton C, editor. PLoS ONE. 2010 Oct 29;5(10):e13735.
463. Pastor-Navarro B, García-Flores M, Fernández-Serra A, Blanch-Tormo S, Martínez de Juan F, Martínez-Lapiedra C, et al. A Tetra-Panel of Serum Circulating miRNAs for the Diagnosis of the Four Most Prevalent Tumor Types. IJMS. 2020 Apr 16;21(8):2783.
464. Si H, Sun X, Chen Y, Cao Y, Chen S, Wang H, et al. Circulating microRNA-92a and microRNA-21 as novel minimally invasive biomarkers for primary breast cancer. J Cancer Res Clin Oncol. 2013 Feb;139(2):223–9.
465. Schrauder MG, Strick R, Schulz-Wendtland R, Strissel PL, Kahmann L, Loehberg CR, et al. Circulating Micro-RNAs as Potential Blood-Based Markers for Early Stage Breast Cancer Detection. Hoheisel JD, editor. PLoS ONE. 2012 Jan 5;7(1):e29770.
466. Ng EKO, Li R, Shin VY, Jin HC, Leung CPH, Ma ESK, et al. Circulating microRNAs as Specific Biomarkers for Breast Cancer Detection. Srivastava RK, editor. PLoS ONE. 2013 Jan 3;8(1):e53141.
467. Shen J, Hu Q, Schrauder M, Yan L, Wang D, Medico L, et al. Circulating miR-148b and miR-133a as biomarkers for breast cancer detection. Oncotarget. 2014 Jul 30;5(14):5284–94.

468. Antolín S, Calvo L, Blanco-Calvo M, Santiago MP, Lorenzo-Patiño MJ, Haz-Conde M, et al. Circulating miR-200c and miR-141 and outcomes in patients with breast cancer. BMC Cancer. 2015 Dec;15(1):297.
469. Soleimanpour E, Babaei E, Hosseinpour-Feizi MA, Montazeri V. Circulating miR-21 and miR-155 as potential noninvasive biomarkers in Iranian Azeri patients with breast carcinoma. J Can Res Ther. 2019;15(5):1092.
470. Nashtahosseini Z, Reza Aghamaali M, Sadeghi, Heydari N, Parsian. Circulating status of microRNAs 660-5p and 210-3p in breast cancer patients. The Journal of Gene Medicine. 2021;23(4):e3320.
471. Han S, Li P, Wang D, Yan H. Dysregulation of serum miR-1204 and its potential as a biomarker for the diagnosis and prognosis of breast cancer. Rev Assoc Med Bras. 2020 Jun;66(6):732–6.
472. Chen H, Liu H, Zou H, Chen R, Dou Y, Sheng S, et al. Evaluation of Plasma miR-21 and miR-152 as Diagnostic Biomarkers for Common Types of Human Cancers. J Cancer. 2016;7(5):490–9.
473. An X, Quan H, Lv J, Meng L, Wang C, Yu Z, et al. Serum microRNA as potential biomarker to detect breast atypical hyperplasia and early-stage breast cancer. Future Oncology. 2018 Dec;14(30):3145–61.
474. Hu Z, Dong J, Wang LE, Ma H, Liu J, Zhao Y, et al. Serum microRNA profiling and breast cancer risk: the use of miR-484/191 as endogenous controls. Carcinogenesis. 2012 Apr;33(4):828–34.
475. Zhang H, Li B, Zhao H, Chang J. The expression and clinical significance of serum miR-205 for breast cancer and its role in detection of human cancers. International journal of clinical and experimental medicine. 2015;8(2):3034.
476. Wang Y, Yin W, Lin Y, Yin K, Zhou L, Du Y, et al. Downregulated circulating microRNAs after surgery: potential noninvasive biomarkers for diagnosis and prognosis of early breast cancer. Cell Death Discov. 2018 Dec;4(1):87.
477. Zhang G, Zhang W, Li B, Stringer-Reasor E, Chu C, Sun L, et al. MicroRNA-200c and microRNA-141 are regulated by a FOXP3-KAT2B axis and associated with tumor metastasis in breast cancer. Breast Cancer Res. 2017 Dec;19(1):73.
478. Feliciano A, González L, Garcia-Mayea Y, Mir C, Artola M, Barragán N, et al. Five microRNAs in Serum Are Able to Differentiate Breast Cancer Patients From Healthy Individuals. Front Oncol. 2020 Nov 3;10:586268.
479. Ibrahim AM, Said MM, Hilal AM, Medhat AM, Elsalam IMA. Candidate circulating microRNAs as potential diagnostic and predictive biomarkers for the monitoring of locally advanced breast cancer patients. Tumor Biology. 2020;42(10):1010428320963811.
480. Swellam M, Zahran RFK, Ghonem SA, Abdel-Malak C. Serum MiRNA-27a as potential diagnostic nucleic marker for breast cancer. Archives of Physiology and Biochemistry. 2021 Jan 2;127(1):90–6.
481. Jang J, Kim Y, Kang K, Kim K, Park Y, Kim C. Multiple microRNAs as biomarkers for early breast cancer diagnosis. Mol Clin Oncol. 2020 Dec 17;14(2):31.
482. Guo H, Zeng X, Li H, Guo Y, Wang T, Guo H, et al. Plasma miR-1273g-3p acts as a potential biomarker for early Breast Ductal Cancer diagnosis. An Acad Bras Ciênc. 2020;92(1):e20181203.
483. Ashirbekov Y, Abaildayev A, Omarbayeva N, Botbayev D, Belkozhayev A, Askandirova A, et al. Combination of circulating miR-145-5p/miR-191-5p as biomarker for breast cancer detection. PeerJ. 2020 Dec 16;8:e10494.
484. Cuk K, Zucknick M, Madhavan D, Schott S, Golatta M, Heil J, et al. Plasma MicroRNA Panel for Minimally Invasive Detection of Breast Cancer. Miller TW, editor. PLoS ONE. 2013 Oct 23;8(10):e76729.
485. Raheem AR, Abdul-Rasheed OF, Al-Naqqash MA. The diagnostic power of circulating micro ribonucleic acid 34a in combination with cancer antigen 15-3 as a potential biomarker of breast cancer. SMJ. 2019 Dec;40(12):1218–26.

486. Zhu Y, Wang Q, Xia Y, Xiong X, Weng S, Ni H, et al. Evaluation of MiR-1908-3p as a novel serum biomarker for breast cancer and analysis its oncogenic function and target genes. BMC Cancer. 2020 Dec;20(1):644.
487. Ahmed Mohmmed E, Shousha W, El-Saiid A, Ramadan S. A Clinical Evaluation of Circulating MiR-106a and Raf-1 as Breast Cancer Diagnostic and Prognostic Markers. Asian Pac J Cancer Prev. 2021 Nov 1;22(11):3513–20.
488. Sadeghi H, Kamal A, Ahmadi M, Najafi H, Sharifi Zarchi A, Haddad P, et al. A novel panel of blood-based microRNAs capable of discrimination between benign breast disease and breast cancer at early stages. RNA Biology. 2021 Nov 12;18(sup2):747–56.
489. Itani MM, Nassar FJ, Tfayli AH, Talhouk RS, Chamandi GK, Itani ARS, et al. A Signature of Four Circulating microRNAs as Potential Biomarkers for Diagnosing Early-Stage Breast Cancer. IJMS. 2021 Jun 6;22(11):6121.
490. Mahmoud MM, Sanad EF, Elshimy RAA, Hamdy NM. Competitive Endogenous Role of the LINC00511/miR-185-3p Axis and miR-301a-3p From Liquid Biopsy as Molecular Markers for Breast Cancer Diagnosis. Front Oncol. 2021 Oct 20;11:749753.
491. Zou R, Loke SY, Tan VKM, Quek ST, Jagmohan P, Tang YC, et al. Development of a microRNA Panel for Classification of Abnormal Mammograms for Breast Cancer. Cancers. 2021 Apr 28;13(9):2130.
492. Li X, Tang X, Li K, Lu L. Evaluation of Serum MicroRNAs (miR-9-5p, miR-17-5p, and miR-148a-3p) as Potential Biomarkers of Breast Cancer. Bertero L, editor. BioMed Research International. 2022 Jan 24;2022:1–8.
493. Shaker O, Ayeldeen G, Abdelhamid A. The Impact of Single Nucleotide Polymorphism in the Long Non-coding MEG3 Gene on MicroRNA-182 and MicroRNA-29 Expression Levels in the Development of Breast Cancer in Egyptian Women. Front Genet. 2021 Aug 4;12:683809.

Curriculum Vitae

Emir Šehović was born on 31st August 1996. He obtained his Bachelor's degree in Genetics and Bioengineering at the International Burch University in Sarajevo in 2017. In his bachelor's thesis he covered the topic of Y-DNA haplogroups within the Balkan populations on which he also published a peer-reviewed article in Anthropological Review in 2018. He obtained his Master's degree in Genetics and Bioengineering in 2019 at the same University and for his thesis project worked on salivary microRNAs and autism spectrum disorder diagnosis on which he also published an article in PlosOne. In July 2020, he became an early stage researcher, under Marie Skłodowska-Curie Actions (MSCA) Innovative Training Network, within the CancerPrev consortium and enrolled as a PhD student at the University of Turin.