This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Genome-wide expression profiling of human circulating monocytes and macrophages identifies diagnostic and prognostic signatures for cancer outcomes

Stamatina Fragkogianni

A Thesis submitted in fulfilment of the requirements of the degree of Doctor of Philosophy to the University of Edinburgh

April 2017
Abstract

**Background:** Breast and endometrial cancers are the most common gynaecological cancers in women in the UK. Early detection of tumours is crucial for improving patient survival rates. In breast cancer, mammography is the most reliable screening method for asymptomatic patients; however, its sensitivity is limited by breast density. Currently, there are no early screening assays for endometrial cancer. Thus, there is an urgent need to identify clinical biomarkers for improved non-invasive diagnosis of breast and endometrial cancer. Macrophages are abundant in the tumour microenvironment and their density has been associated with poor prognosis in breast cancer and decreased survival in endometrial cancer. Monocytes are precursors of macrophages and recent studies have shown an association with pro-tumoral functions. The aim of this study has been to examine the transcriptional profiles of human circulating monocytes and tumour associated macrophages (TAMs) in order to investigate their biological relevance and potential as biomarkers for cancer detection and prognosis.

**Methods:** RNA-sequencing was performed on purified monocytes (22 healthy individuals, 21 breast cancer, 16 endometrial cancer samples), as well as purified normal macrophages, TAMs from breast tissue (4 breast cancer, 4 healthy breast) and endometrium tissue (5 endometrial cancer, 9 healthy endometrium).

**Results:** A shift in the transcriptional profile of monocytes in cancer compared to controls was observed. Given these cancer-associated alterations circulating monocytes from cancer patients were called “Tumour Educated Monocytes” (TEMo). A TEMo-derived 13-gene signature was extracted that detected cancer, yielding an accuracy of 94%, a positive predictive value (PPV) of 92% and a negative predictive value (NPV) of 97%. External validation confirmed the ability of the signature to accurately identify cancer patients with perfect accuracy. Transcriptome profiling of TAMs revealed a significantly altered gene expression profile when compared to normal tissue resident macrophages. Furthermore, comparison of TAMs between breast and endometrial cancer also revealed differences suggesting that different tumour microenvironments induce different gene expression profiles in TAMs. Functional analysis of significant genes in breast
cancer revealed similar biological pathways to those of murine studies suggesting that TAMs in humans and mice may have similar functions. A gene list of transmembrane receptors has been extracted by comparing breast cancer TAMs with publicly available datasets that could serve as markers for their identification. Finally, exploratory analysis identified a subset of 49 genes associated with recurrence-free and overall survival in publicly available datasets.

**Conclusion:** To my knowledge this is the first genome-wide profiling study of human circulating monocytes and TAMs in breast and endometrial cancer. It provides evidence that monocytes and TAMs can alter their expression profile in the presence of cancer and, using bioinformatics tools I was able to identify biomarkers for diagnosis and prognosis of breast and endometrial cancer.
Lay summary

Early detection of tumours is an important way of improving patient survival rates. To date, in breast cancer mammographic screening is the most reliable screening method in asymptomatic patients; however, its sensitivity is limited by breast density. Previous reports have shown that extraction of a set of genes for cancer monitoring using whole blood is feasible. However, whole blood consists of a mixed populations of cells with different function in the presence of cancer, therefore a set of genes specific to tumour-associated immune cells could be more informative.

Monocytes are circulating blood cells with a half-life of about one to three days. Macrophages in tumours are almost entirely derived from circulating monocytes and are recruited into tumour sites. In studies using mouse models of cancer, tumour-associated macrophages (TAMs) have been associated with functions that promote malignancy and cancer progression. Their density has been associated with poor prognosis in human breast cancer and decreased survival in endometrial cancer. However, their functions in cancer are still unclear. The aim of this study has been to examine how circulating monocytes and TAMs respond in the presence of malignancy in order to develop a non-invasive blood test for the detection of cancer and identify markers for disease progression.

Cancer patients and healthy individuals (22 healthy samples, 21 breast cancer and 16 endometrial cancer samples) were recruited and using statistical modelling I showed that circulating monocytes from cancer patients as very different to those of healthy individuals. Given these alterations produced by cancer circulating monocytes have been now termed “Tumour Educated Monocytes” (TEMo). A set of 13 genes (molecular signature) was identified using predictive modelling that is able to identify the presence of cancer with accuracy higher than of current mammography. Further validation on an independent cohort of patients confirmed the robustness of the molecular signature. By comparing to a set of patients from infectious diseases but not cancer this study further showed that the molecular signature is specific to cancer. Furthermore, 9 patients that have undergone whole tissue biopsy (4 breast
cancer and 5 endometrial cancer patients) as well as 13 healthy individuals were recruited. Using statistical modelling, this study showed that TAMs are significantly different when compared to macrophages of healthy donors. Furthermore, comparison of TAMs between breast and endometrial cancer also revealed differences suggesting that tumours in different sites cause different profiles of TAMs. A set of genes specific to breast cancer TAMs was extracted by comparing our data set with publicly available datasets that could serve as markers for TAMs identification. Finally, a 73-gene TAM-derived signature was identified that was predictive for overall survival and recurrence-free survival.

This is to my knowledge the first study of human circulating monocytes and macrophages in breast and endometrial cancer. It shows evidence that monocytes and TAMs can alter their profiles in the presence of cancer and using mathematical modelling is it possible to identify biomarkers for diagnosis and prognosis of breast and endometrial cancer.
Declaration

The work in this thesis is entirely my own work unless otherwise stated, where credit to the contributor is given. No part of the work has been submitted for any other degree.

Stamatina Fragkogianni, April 2017
Acknowledgements

First and foremost, I would like to thank my supervisors Professor Jeffrey Pollard, Dr. Sander Granneman and Dr. Andy Sims for their excellent support and guidance throughout all these years. I feel exceptionally lucky and grateful having had the opportunity to work with such brilliant scientists.

I have also met and made friends with a number of exceptional people that helped me progress with my project but also taught me so much. I would like to thank Dr. Luca Cassetta for his patience and help throughout these years. His crazy ideas always mean more work but have also greatly helped this thesis. Special thanks goes to Dr. Agi Swierczak, she has been a great friend and a great colleague all these years. I am so grateful to her for proofreading and correcting this thesis so many times. I would also like to thank Dr. Gaël Sugano for bearing with my, most of the time, naïve biological questions. I might have missed the bus several times because of our hour-long discussions, but it was definitely worth it! Lastly, I would also like to thank all other members of the lab, past and present.

I would also like to thank my friends, Anastasia, Jana, Luca, Alison, and Diane for their support in the past years but also in the past few months. They helped put a smile on my face when I was grumpy and/or tired, and I’m really grateful to them. Special thanks goes to Sara for her enormous and tireless support even when I have been the most difficult person. I don’t think I would have been able to get through this without you. Lastly, I would like to thank my two thesis/library mates, Eleni and Yolanda, for making the past 3 months a lot more bearable. Looking forward to our trip to Morocco!

Finally, and most importantly, I would like to thank my parents, my grandparents and my sister for their love and total and unconditional support in everything I have chosen to do.
Table of contents

Abstract .................................................................................................................................................i
Lay summary ..........................................................................................................................................iii
Declaration ............................................................................................................................................v
Acknowledgements .............................................................................................................................vi
Table of contents .................................................................................................................................vii
List of figures ........................................................................................................................................xiii
List of tables ..........................................................................................................................................xv
Abbreviations .........................................................................................................................................xvi

1 General Introduction .........................................................................................................................1
  1.1 Breast cancer .................................................................................................................................1
    1.1.1 Incidence and statistics of survival .................................................................1
    1.1.2 Histological classification ...............................................................................2
    1.1.3 Staging .......................................................................................................................3
    1.1.4 Molecular portraits of breast cancer ...............................................................3
    1.1.5 Screening and detection ..................................................................................6
    1.1.6 Therapeutic strategies ......................................................................................7
  1.2 Endometrial cancer ......................................................................................................................9
    1.2.1 Incidence and statistics of survival .................................................................9
    1.2.2 Histological and molecular classification ..................................................9
    1.2.3 Staging ..................................................................................................................10
    1.2.4 Screening and detection ................................................................................10
    1.2.5 Therapeutic strategies ...................................................................................11
  1.3 Tumour microenvironment (TME) ..........................................................................................12
    1.3.1 Cells types within the tumour microenvironment ...........................................12
  1.4 Macrophages ...............................................................................................................................14
1.4.1 Tissue-resident macrophage origins and functions.................................14
1.4.2 Macrophage polarization ........................................................................16
1.4.3 Tumour-associated macrophages (TAMs)..............................................16
  1.4.3.1 TAM recruitment ...........................................................................17
  1.4.3.2 TAM functions in tumour development and progression ..................18
  1.4.3.3 Heterogeneity of TAM subpopulations .........................................18
  1.4.3.4 Macrophages as therapeutic targets .............................................19
1.4.3.5 TAM recruitment ...........................................................................17
1.4.3.6 TAM functions in tumour development and progression ..................18
1.4.3.7 Heterogeneity of TAM subpopulations .........................................18
1.4.3.8 Macrophages as therapeutic targets .............................................19
1.5 Monocytes ...............................................................................................20
  1.5.1 Monocyte subpopulations ..................................................................20
  1.5.2 Monocytes in cancer .........................................................................21
1.6 High-throughput gene expression profiling technologies .......................23
  1.6.1 RNA-sequencing technology ...............................................................23
    1.6.1.1 RNA-seq data analysis and considerations ..................................24
1.7 Gene signatures for diagnosis and prognosis of cancer .............................27
  1.7.1 Gene expression-based diagnostic assays ...........................................28
  1.7.2 First generation gene expression prognostic assays ............................29
  1.7.3 Stroma and immune-related prognostic signatures .............................32
  1.7.4 Considerations when developing diagnostic and prognostic signatures 33
    1.7.4.1 Sample size and patient heterogeneity ........................................33
    1.7.4.2 Model and feature selection .........................................................34
    1.7.4.3 Independent validation ................................................................34
    1.7.4.4 Reproducibility of analysis ..........................................................35
    1.7.4.5 Clinical application of predictive/prognostic signature ..................36
1.8 Aims and Objectives .................................................................................37
  1.8.1 Hypotheses .......................................................................................37
  1.8.2 Study Aims .......................................................................................38
2 Human circulating monocytes as biomarkers for early detection of breast
  and endometrial cancer .................................................................................39
  2.1 Acknowledgments and contributions ......................................................39
  2.2 Introduction ............................................................................................39
  2.3 Aims .......................................................................................................40
  2.4 Materials and methods ..........................................................................41
2.4.1 Patient samples ................................................................. 41
2.4.2 Sample processing .............................................................. 41
2.4.3 Sequence alignment and Quantification ................................. 42
2.4.4 Statistical analysis for differentially expressed genes ................. 42
2.4.5 Functional enrichment analysis and Pathway Analysis .................. 43
2.4.6 Gene selection ........................................................................ 43
2.4.7 Machine learning methods .................................................... 44
2.4.8 Cross validation and model evaluation ................................. 45
2.4.9 Comparison with publically available datasets ....................... 46

2.5 Results .................................................................................. 49

2.5.1 Monocytes exhibit significantly altered gene expression profiles in breast and endometrial cancer patients compared to those of healthy controls .......... 49
2.5.2 Distinct gene expression profiles of Tumour-Educated Monocytes (TEMo) and normal monocytes in cancer .......................................................... 56
2.5.3 TEMos as biomarkers for detection of cancer ............................ 60

2.5.3.1 Identification of a TEMo-derived signature for detection of cancer 60
2.5.3.2 Confirmation of gene signature in an independent validation set ..... 64
2.5.3.3 TEMo signature detects breast cancer at an early stage .......... 66
2.5.3.4 TEMo signature is specific to cancer ........................................ 68
2.5.3.5 TEMo as a pan-cancer diagnostic signature ............................. 69
2.5.3.6 TEMo signature is specific to monocytes in comparison with PBMCs

2.6 Discussion ............................................................................. 73

2.7 Conclusions ............................................................................ 77

3 Transcriptional profiling of TAMs highlights differences in breast and endometrial tumour microenvironments ................................................................. 78

3.1 Acknowledgments and contributions ........................................ 78
3.2 Introduction ........................................................................... 78
3.3 Aims .................................................................................... 79
3.4 Materials and methods .............................................................. 81

3.4.1 Patients ............................................................................... 81
3.4.2 Sample processing and RNA-sequencing ................................. 81
3.4.3  Sequencing alignment and Quantification ........................................ 82
3.4.4  Statistical analysis for differentially expressed genes ...................... 82
3.4.5  Functional enrichment and pathway analysis .................................. 83
3.4.6  Comparison with publically available datasets ............................... 83
3.4.7  Identification and validation of a prognostic TAM-derived signature ... 85

3.5  Results .................................................................................................. 86

3.5.1  TAMs from breast and endometrial cancer show different transcriptome
       profiles compared to normal resident tissue macrophages and monocytic
       progenitors .......................................................................................... 86
   3.5.1.1  Transcriptional profiles of TAMs are significantly different
            compared to normal resident macrophages in breast cancer .......... 86
   3.5.1.2  Transcriptional profiles of TAMs are significantly different
            compared to normal resident macrophages in endometrial cancer .... 89
   3.5.1.3  Gene expression profiles of TAMs have limited overlap with TEMo
            91
3.5.2  Macrophage transcriptome profiles show tissue specificity ............. 93
   3.5.2.1  Resident macrophages and TAMs from breast and endometrial tissue
            are transcriptionally different .......................................................... 93
3.5.3  Human breast and mouse mammary cancer TAMs share common
       biological profiles ................................................................................ 94
3.5.4  Mixed polarization of M1- and M2-like macrophages in human TAMs
       and resident macrophages .................................................................. 96
3.5.5  Exploratory analysis for identification of TAM specific markers ........ 97
   3.5.5.1  Identification of candidate gene markers for TAMs in breast cancer
            97
   3.5.5.2  Exploratory analysis of TAMs as prognostic markers for breast
            cancer 100

3.6  Discussion ............................................................................................ 105

3.7  Conclusions .......................................................................................... 111

4  General discussion and future work ......................................................... 112

4.1  TEMos as biomarkers for detection of breast and endometrial human
    cancers ..................................................................................................... 112
4.1.1 TEMos show distinct transcriptional profiles between breast and endometrial cancer patients compared to healthy individuals ........................................ 112
4.1.2 TEMo as a source of biomarkers for detection of breast and endometrial cancer 114

4.2 Genome-wide transcriptomic analysis of TAMs and resident macrophages in breast and endometrial cancer .................................................. 116
  4.2.1 TAMs and resident macrophages from breast and endometrial cancer comprise distinct populations ................................................................. 116
  4.2.2 TAMs and resident macrophages are transcriptional distinct depending on microenvironment ................................................................. 118
  4.2.3 Identification of candidate markers for TAMs identification .......... 118
  4.2.4 TAMs as biomarkers for prognosis of breast cancer ..................... 119
  4.2.5 There is little overlap between the transcriptional profiles of TAMs and their precursors monocytes ......................................................... 121

4.3 Future directions .................................................................................. 122
  4.3.1 Validation of the TEMo-derived diagnostic signature on larger cohorts and other settings ................................................................. 122
    4.3.1.1 Larger cohort and molecular subtypes .................................. 122
    4.3.1.2 Other cancer types ............................................................. 123
    4.3.1.3 TEMo signature for patients during/after treatment ........... 123
  4.3.2 Clinical application of the TEMo-derived diagnostic signature .......... 124
    4.3.2.1 Gene expression assays based on targeted genes ................. 124
  4.3.3 TAM-specific markers for therapeutic targeting ............................. 124
  4.3.4 Single-cell RNA sequencing of different regions of the tumour for identification of TAMs subpopulations ........................................... 125
  4.3.5 Prognostic significance of TAM markers ...................................... 125

4.4 Conclusion ............................................................................................. 126

5 References ................................................................................................. 127

6 Appendix .................................................................................................... 152
  6.1 R scripts .................................................................................................. 152
  6.2 Table of the 13-genes included in the TEMo-signature .................... 159
6.3 List of 73 candidate gene markers for identification of TAMs ..........160
6.4 Details of monocyte samples sequencing and alignment ..................163
6.5 Details of macrophage samples sequencing and alignment ...............167
List of figures

Figure 1. Female breast cancer incidence and mortality in the UK between 1972 and 2008 .......................................................... 1
Figure 2. Molecular classification of breast cancer and clinical association ....... 5
Figure 3. The primary tumour microenvironment .................................. 14
Figure 4. Overview of RNA-seq bioinformatics workflow .......................... 25
Figure 5. Analysis workflow of RNA-seq samples ..................................... 49
Figure 6. Distribution of CPM log2 normalized gene counts ...................... 50
Figure 7. Transcriptome analysis of purified monocytes showed differences in breast and endometrial cancer compared to healthy patients .....................51
Figure 10. Differential expression analysis of purified monocytes showed little differences between breast cancer and endometrial patients .......................55
Figure 11. Venn diagram of commonly up-regulated and down-regulated transcripts between monocytes from breast cancer and endometrial cancer compared to healthy patients, and monocytes from breast cancer compared to endometrial cancer ........................................................................56
Figure 13. Functional enrichment analysis of significantly regulated transcripts in TEMo .....................................................................59
Figure 14. Workflow for the identification and validation of the TEMo-derived signature .....................................................................61
Figure 15. Performance of RF-X^2 classifier in internal and independent validation ........................................................................ 63
Figure 16. Performance of random signatures and random classifiers ............ 66
Figure 17. TEMo can detect the presence of early stage breast cancer ............ 67
Figure 18. TEMo signature shows potential for pan-cancer diagnosis ............. 70
Figure 19. TEMo signature is specific to monocytes ................................... 72
Figure 20. Transcriptome analysis of purified TAMs from breast cancer compared to normal resident macrophages shows distinct gene expression profiles ........................................................................ 87
Figure 21. Functional enrichment analysis of significantly regulated transcripts in breast TAMs ...................................................................89
Figure 22. Transcriptome analysis of purified TAMs from endometrial cancer compared to resident macrophages shows distinct gene expression profiles......90
Figure 23. Venn diagram of transcripts commonly up-regulated and down-regulated in TAMs from breast cancer and endometrial cancer compared to healthy patients, and TEMo from breast cancer and endometrial cancer compared to healthy patients ........................................................................................................92
Figure 24. Gene expression analysis of purified TAMs from breast cancer compared to TAMs from endometrial cancer, and resident macrophages form healthy breast tissue compared to resident macrophages from healthy endometrial tissue reveals tissue specificity of macrophages .................................................93
Figure 25. Comparison of human breast cancer TAMs and mouse mammary tumour TAMs .........................................................................................................................................95
Figure 26. Comparison of transcripts expressed in TAMs and resident macrophages from breast and endometrial tissue against the M1-like and M2-like genes from the Martinez et al. study (Martinez et al. 2006) .........................................................97
Figure 27. Identification of TAM-specific markers ........................................................................................................................99
Figure 28. TAM-derived signature (49 genes) identified clinically relevant groups in the NKI dataset .............................................................................................................................................101
Figure 29. TAM-derived signature (49 genes) identifies clinically relevant groups in the MAINZ dataset ..............................................................................................................................................104
List of tables

Table 1. Clinical characteristics of samples used in the study ..................................48
Table 2. Performance of evaluated machine learning algorithms using both $X^2$
and RFE feature selection ........................................................................................................62
Table 3. Table of performance of the RF-$X^2$ classifier on a negative control
dataset .......................................................................................................................................69
Table 4. Clinical information of samples included in this study .........................................85
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCO</td>
<td>American Society of Clinical Oncology</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the Curve</td>
</tr>
<tr>
<td>AUROC</td>
<td>Area Under the Receiver Operating Characteristic</td>
</tr>
<tr>
<td>BCI</td>
<td>Breast Cancer Index</td>
</tr>
<tr>
<td>BCS</td>
<td>Breast Conserving Surgery</td>
</tr>
<tr>
<td>CAGE</td>
<td>Cap Analysis of Gene Expression</td>
</tr>
<tr>
<td>CTCs</td>
<td>Circulating Cancer Cells</td>
</tr>
<tr>
<td>ctDNA</td>
<td>Cell Free DNA</td>
</tr>
<tr>
<td>DCIS</td>
<td>Ductal Carcinoma In Situ</td>
</tr>
<tr>
<td>ECM</td>
<td>Extracellular Matrix</td>
</tr>
<tr>
<td>ER</td>
<td>Estrogen Receptor</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescent-Activated Cell Sorting</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FFPE</td>
<td>Formalin-Fixed Paraffin-Embedded</td>
</tr>
<tr>
<td>FIGO</td>
<td>Federation of Gynaecology and Obstetrics</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragment Per Kilobase Of Exon Per Million</td>
</tr>
<tr>
<td>FS</td>
<td>Feature Selection</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>GGI</td>
<td>Genomic Grade Index</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>HER2</td>
<td>Human Epidermal Growth Factor Receptor 2</td>
</tr>
<tr>
<td>HTPF</td>
<td>Human Tissue Procurement Facility</td>
</tr>
<tr>
<td>IDC</td>
<td>Invasive Ductal Carcinoma</td>
</tr>
</tbody>
</table>
IHC   Immunohistochemistry
IPA   Ingenuity Pathway Analysis
Log₂FC Binary logarithm of Fold Change
LOOCV Leave One Out Cross Validation
MAMs Metastasis-Associated Macrophages
MDS   Multi-Dimensional Scaling
MDSCs Myeloid Derived Suppressor Cells
ML    Machine Learning
MRI   Magnetic Resonance Imaging
MSCs  Mesenchymal Stem Cells
MSI   Microsatellite Instable
NK cells Natural Killer cells
NPV   Negative Predictive Value
OOB   Out-Of-Bag
PBMCs Peripheral Blood Mononuclear Cells
PCR   Polymerase Chain Reaction
PPV   Positive Predictive Value
PR    Progesterone Receptor
PyMT  Polyoma Middle T oncoprotein
qPCR  Quantitative Polymerase Chain Reaction
RF    Random Forest
RF-Χ² Random Forest Chi-Square
RFE   Recursive Feature Elimination
RIN   RNA Integrity Number
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>ROR</td>
<td>Risk Of Recurrence</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads Per Kilobase Of transcript per Million</td>
</tr>
<tr>
<td>RS</td>
<td>Recurrence score</td>
</tr>
<tr>
<td>SAGE</td>
<td>Serial Analysis of Gene Expression</td>
</tr>
<tr>
<td>SVA</td>
<td>Surrogate Variable Analysis</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TAM</td>
<td>Tumour-Associated Macrophage</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome ‘atlas</td>
</tr>
<tr>
<td>TEMo</td>
<td>Tumour-Educated Monocyte</td>
</tr>
<tr>
<td>TILs</td>
<td>Tumour Infiltrating Lymphocytes</td>
</tr>
<tr>
<td>TME</td>
<td>Tumour Microenvironment</td>
</tr>
<tr>
<td>TN</td>
<td>Triple-Negative</td>
</tr>
</tbody>
</table>
1 General Introduction

1.1 Breast cancer

Breast cancer is the most commonly diagnosed female cancer (Siegel et al. 2016) representing approximately 12% of all new cancer cases worldwide (McGuire et al. 2015). Approximately 30% of women diagnosed with early stage tumours will develop metastatic disease, which is currently considered incurable (O'Shaughnessy 2005). Hence, an advancement of current diagnostic and prognostic methods will greatly improve patient stratification for individualised clinical decisions, leading to better treatments and outcomes.

1.1.1 Incidence and statistics of survival

Breast cancer is the most common reproductive tract cancer in the UK with approximately 53,000 new cases reported each year (Cancer Research UK 2016).

![Graph showing breast cancer incidence and mortality in the UK between 1972 and 2008. Incidence rates have increased by 85% while mortality rates have decreased by 33% since the 1970s (Source: Office of National Statistics).]

Figure 1. Female breast cancer incidence and mortality in the UK between 1972 and 2008. Incidence rates of breast cancer have increased by 85% while mortality rates have decreased by 33% since the 1970s (Source: Office of National Statistics).
Although the incidence rates of female breast cancer have increased by 85% since 1970, death rates have decreased by 33% in the same time period (Office for National Statistics) (Figure 1). The reduction in mortality is likely due to early detection and prevention methods, as well as the use of systemic adjuvant treatment (EBCTCG 2005).

1.1.2 Histological classification
Breast cancer is a heterogeneous disease composed of distinct clinical and genetic features (Stingl & Caldas 2007). Histologically, breast cancer can be classified into \textit{in situ} carcinoma or invasive (infiltrating) carcinoma. \textit{In situ} carcinomas are located within the basement membrane and can be further sub-classified in lobular or ductal; with ductal carcinomas representing the most common type of non-invasive carcinoma.

Ductal carcinoma in situ (DCIS) accounts for 90% of total \textit{in situ} cases (Li et al. 2003), and can be further sub-classified according to the architectural features of the tumour including, a) Comedo, b) Cribiform, c) Micropapillary, d) Papillary, and e) Solid. It is thought that 20-50% of DCIS cases will progress to invasive breast cancer, however it is still unclear when and how this will occur (Cowell et al. 2013). The progression to invasive carcinomas is highly variable, and can take years or might not occur within the lifetime of the patient, leading to unnecessary clinical interventions.

Invasive carcinomas are defined by the invasion of malignant cells through the basement membrane. As with \textit{in situ} breast carcinomas, invasive breast carcinomas are a heterogeneous group of tumours that can be classified in 7 major histological subtypes, a) Infiltrating Ductal, b) Invasive Lobular, c) Ductal/Lobular, f) Mucinous, e) Tubular, f) Medullary and g) Papillary carcinomas. Of those, invasive ductal carcinoma (IDC) ‘no special type’ is by far the most common type of invasive carcinomas, accounting for 70-80% of all diagnosed invasive breast tumours (Li et al. 2005). It is characterised by the invasion of cancer cells through the basement membrane of the duct and the replacement of the surrounding normal tissue. IDC can be further classified as grade I (well differentiated), grade II (moderately
differentiated) and grade III (poorly differentiated) based on the Nottingham combined histologic grade (Elston-Ellis modification of Scarff-Bloom-Richardson grading system) to determine the aggressiveness of the tumour (Bloom & Richardson 1957; Elston & Ellis 1991). In lower grade tumours cells resemble normal-like cells and appear to grow slowly resulting in better patient outcome. In contrast, in higher-grade tumours, cells grow rapidly resulting in worse patient outcomes. Histological grade has been identified as a strong independent predictor of patient outcome in invasive breast carcinomas (Rakha et al. 2008).

1.1.3 Staging
Tumour staging of invasive breast carcinomas is performed using the TNM staging system and/or the overall stage grouping system. In the TNM system, T corresponds to tumour size, N corresponds to invasion of lymph nodes and M corresponds to the presence of metastasis. An alternative staging system includes 4 stages of breast cancer from 0 (early form, non-invasive) to IV (metastatic) (Singletary & Greene 2003).

1.1.4 Molecular portraits of breast cancer
Immunohistochemically (IHC), breast cancers are characterised by the expression of oestrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2/ERBB2). These markers define 3 basic therapeutic groups, the ER-positive group, the HER2 enriched (or ERBB2) group and the triple-negative group (TN) group. In an effort to better characterise breast cancer heterogeneity more markers of proliferation and invasiveness have been determined such as a) TP53, a protein involved in tumour suppression (Allred et al. 1993), b) BCL2 (B-cell lymphoma), a protein involved in cell death regulation (Q. Yang et al. 2009; Min et al. 2016), c) claudin, a protein involved in epithelium development (Lanigan et al. 2009; Lu et al. 2013), d) Ki67, a tumour proliferation marker (Nishimura et al. 2010) and e) TOP2A (topoisomerase II alpha), a protein involved in chromosome segregation (Şahin et al. 2016; Rody, Karn, et al. 2009).
While histological characterization of tumours has been the standard method for patient stratification, the advent of high-throughput technologies has enabled the profiling of thousands of genes simultaneously (section 1.6.1). In a seminal study, Perou and Sørlie examined 8,103 human genes from 42 individuals using cDNA microarrays. They identified an ‘intrinsic’ gene list (436 genes), and using hierarchical clustering identified four molecular subtypes, namely luminal, normal breast-like, HER2 and basal-like, mainly distinguished by the presence or absence of ER expression (Perou et al. 2000). In a following study on a larger cohort they showed that the luminal subtype can be further subdivided into luminal-A and luminal-B and identified distinct clinical outcomes associated with each different subtype (Figure 2) (Sørlie et al. 2001). These findings were further validated in independent cohorts by the same (Sørlie et al. 2003; Hu et al. 2006) and different groups (Sotiriou et al. 2003). Consistently, all microarray studies identified profound differences in outcomes and gene expression between ER-positive and ER-negative tumours.

Luminal-A cancers are generally characterized by high expression of ER and PR, low expression of proliferation-related genes, and absence of the HER2 oncogene. Additionally, they consist of low-grade tumours and are associated with very good overall prognosis. In contrast, luminal-B tumours demonstrate higher expression of proliferation related genes, are higher grade and are associated with significantly worse clinical outcomes (Perou et al. 2000; Sørlie et al. 2001; Sørlie et al. 2003; Parker et al. 2009). The significance of normal-like breast tumours is still largely undefined and it has been suggested that normal-like subtype might represent a contamination artifact (Parker et al. 2009). HER2-enriched cancers overexpress the ERBB2 gene resulting in aggressive, intermediate- to high-grade tumours associated with poor clinical outcomes (Sørlie et al. 2001). Basal-like carcinomas are usually high grade and overexpress proliferation- and carcinogenesis-related genes (Cancer & Atlas 2012). They are more common in young women of African and Hispanic origin, and have been linked to BRAC1-mutated breast cancers (Prat et al. 2014).
Figure 2. Molecular classification of breast cancer and clinical association. A) A series of gene expression profiling studies have highlighted that breast cancer is a heterogeneous disease, with distinct molecular subtypes. In a series of studies, Perou and colleagues identified 5 different subtypes of breast cancer (Perou et al. 2000). B) In subsequent studies the five molecular subtypes were associated with distinct clinical outcomes (Sørlie et al. 2003; Sørlie et al. 2001). The prognostic outcomes are shown as overall survival. ERBB2-positive tumours and basal subtypes have the worst outcome, whereas luminal tumours have better outcomes. Luminal-A tumours have generally the most favourable outcome. Illustration adapted from (Vargo-Gogola & Rosen 2007).

Additional studies have reported at least three more molecular subtypes, however their clinical significance is still to be determined: a) claudin-low subtype (Herschkowitz et al. 2007) , b) ‘molecular apocrine’ subtype (Farmer et al. 2005), and c) ‘interferon’ subtype (Hu et al. 2006). Furthermore, the mutational landscape of breast cancer is also now considered for patient classification (Nik-Zainal et al. 2016; Pereira et al. 2016). Recent studies through integration of genomic and transcriptomic datasets of 1,000 breast tumours have identified 10 subtypes (IC10).
based on molecular drivers that are significantly associated with relapse-free survival and chemosensitivity (Curtis et al. 2012; Ali et al. 2014).

Gene expression profiling of DCIS samples indicated that the molecular subtypes can be identified from an early pre-invasive disease stage (Clark et al. 2011). Additionally, gene expression comparison of primary breasts tumours to matched metastatic samples revealed that gene expression patterns of the tumours are conserved during tumour progression suggesting the presence of ‘intrinsic’ molecular subtypes at all stages of the disease (Weigelt et al. 2003).

Overall these studies highlight the heterogeneity of breast cancer and the potential for personalised medicine. However, despite systematic efforts for demonstrating the clinical significance of breast cancer subtypes their use in the clinic is still limited.

1.1.5 Screening and detection
The main screening methods for breast cancer in the UK include, self-examination, mammography, ultrasound, magnetic resonance imaging (MRI) and biopsy. Since the introduction of screening programs there has been an intense debate about its benefits and harms. The main benefit of screening is the early detection of breast carcinomas, and as a direct result, improved clinical prognosis. In fact the Independent UK Panel on Breast Cancer Screening has reported a 20% reduction in mortality within a 20-year screening period as 90% of women diagnosed with early stage breast cancer will survive for at least five years (Siegel et al. 2016). In contrast, the most direct harm of screening programs is over-diagnosis. Over-diagnosis is defined as the detection of cancers that would have never progressed within the lifetime of the patient and would otherwise have remained undetected. In most cases, over-diagnosis leads to overtreatment and unnecessary interventions (Marmot et al. 2013).

Mammographic screening represents the most reliable and least invasive method for early detection of breast cancer to date. Its sensitivity in women over 40 years old is 78%, however this percentage drops to 48% in younger women or women with
higher breast density (Kolb et al. 2002). As with other screening methods, the benefits and harms of mammography have been debated (Løberg et al. 2015). One of the direct benefits of mammography is the early detection of tumours leading to rapid treatment and decreased risk of metastasis. The main disadvantages of mammography are over-diagnosis and false positive results. An exact rate of over-diagnosed patients is not available but observational studies and statistical modelling studies indicate it may be up to 54% (Kalager et al. 2012) (de Koning et al. 2005). It has been estimated that the risk of a false positive result in the age group from 50 to 69 years is 20%, with a 3% risk of undergoing unnecessary biopsy (Hofvind et al. 2012). In an attempt to improve diagnosis of carcinomas, digital mammography is increasingly used in the clinic (Weigel et al. 2013). However, given the limitations of mammographic screening there is clear need for a more reliable and accurate method for the detection of breast cancer.

1.1.6 Therapeutic strategies
Management of breast cancer currently involves five established therapeutic strategies: surgery, radiation therapy, chemotherapy, hormone therapy and targeted-HER2 therapy. Additionally, immunotherapies have been recently introduced. Factors such as the stage of cancer, tumour grade, patient’s age and overall health also affect the decision of which patients receive which combination of treatments.

Surgery represents the most effective method for treatment of breast cancer. It consists of either breast conserving surgery (BCS) or mastectomy (Poortmans 2007). BCS, also known as lumpectomy, is recommended in women with early stage breast cancer and involves the removal of the tumour as well as the surrounding normal tissue. Mastectomy is a radical operation that involves the removal of the whole breast or both breasts. Studies have shown that patients who undergo BCS have equivalent overall survival compared to patients who undergo total mastectomy alone or mastectomy followed by radiation (Agarwal et al. 2014; Litière et al. 2012).

Chemotherapy involves the administration of combinations of cytotoxic agents (alkylating agents, anthracyclines, antimetabolites and taxanes) to kill rapidly proliferating cancer cells. It can be administered preoperatively (neoadjuvant
therapy), usually in patients that presented with locally advanced disease, or after surgery (adjuvant therapy).

Hormone therapy, also known as endocrine therapy, involves blocking or lowering the level of estrogen and progesterone. Endocrine therapy is administered to women diagnosed with ER- and/or PR-positive early and advanced stage disease in both the neoadjuvant and adjuvant setting, and it has been very effective and a major reason for increased survival (Johnston 2010).

Anti-HER2 targeted therapy refers to the use of humanised monoclonal antibodies for directed targeting of the HER2 receptor (i.e. trastuzumab) (Slamon et al. 2001) or both the HER1 and HER2 receptors (i.e. lapatinib) (Geyer et al. 2009). HER2 targeted therapy alone or in conjunction with chemotherapy has drastically improved overall survival of HER2-enriched cancers (25% to 35% of total breast cancers) (Slamon et al. 2001).

Although, great efforts have been made in developing tailored treatments for breast cancer patients, a proportion of breast cancer patients will not respond or will develop resistance (Dixon 2014). It has been demonstrated that breast cancer is a systemic disease and the cross talk between cancer cells and immune cells in the surrounding tumour microenvironment (TME) plays an important role in the development of malignancy and tumour progression. Therefore, development of therapies that target the immune system and its components might improve clinical response and increase patient survival (Hanahan & Weinberg 2011; Joyce & Pollard 2009). Thus, the most recent efforts to improve treatment of cancer include immunotherapy. Immunotherapies are relatively new but have shown promising results pre-clinically. A number of clinical trials are underway for immunotherapies in breast cancer a) NCT02843126 (Intervention: Trastuzumab, NK immunotherapy), b) NCT00301730 (Intervention: aldesleukin, Tumour infiltrating lymphocytes, trastuzumab, paclitaxel), c) NCT00879489 (Intervention: Dendritic cell vaccination), NCT00367250 (Intervention: Cetuximab, Trastuzumab), d) NCT02792114 (Intervention: Cyclophosphamide, Mesothelin-targeted T-cells), among others, (ClinicalTrials.gov).
1.2 Endometrial cancer

Endometrial cancer is one of the most common female reproductive cancer and the sixth most common cancer in women overall (Siegel et al. 2016). The tumour originates from the endometrium. The disease is usually symptomatic and is often discovered at earlier stages. However, there is currently no method for systematic screening of endometrial cancer.

1.2.1 Incidence and statistics of survival

Endometrial cancer is the fourth most common female cancer in the UK with approximately 9000 new cases reported in 2013 (Cancer Research UK 2016). As reported by the Cancer Research UK, there were approximately 2,000 deaths in 2014; indicating a 7% decrease in mortality rates since the 1970s (Cancer Research UK 2016). Diagnosis of endometrial cancer at an early stage results in increased survival with a 5-year survival rate of 95%.

1.2.2 Histological and molecular classification

Endometrial cancer can be broadly classified into estrogen-dependent endometrioid carcinoma (Type I) and estrogen-independent non-endometrioid carcinoma (Type II) (Bokhman 1983; Colombo et al. 2013). Type I tumours are more common in premenopausal women, are associated with obesity and endometrial hyperplasia and show good overall prognosis. Type I tumours are the most common subtype accounting for 80% of all diagnosed cases. Type II tumours are more aggressive and associated with high risk of metastasis and poor prognosis (Bokhman 1983). However, this binary classification has been criticised for not capturing the heterogeneity within the tumour (Murali et al. 2014). Histologically, endometrial cancer can be classified into 4 distinct groups, endometrioid carcinoma, serous carcinoma, carcino-sarcoma, and clear-cell carcinoma (Silverberg & Mutter 2003). Generally, type I tumours correlate well with endometrioid histology and most type II cancers with serous carcinomas (Murali et al. 2014). More molecular subtypes
have been proposed either by IHC (Łapińska-Szumczyk et al. 2015; Łapińska Szumczyk et al. 2014) or gene profiling (Gao et al. 2013) however their clinical relevance is still largely undefined.

The molecular classification of endometrial cancer is still on-going. The most comprehensive study to date was conducted by The Cancer Genome Atlas (TCGA) (TCGA 2013). They assessed tumour samples and corresponding germline DNA from 373 patients with endometrioid and serous endometrial carcinomas using next generation sequencing at the genomic and proteomic level identifying four molecular subgroups; POLE (ultramutated), MSI (microsatellite-instable [hypermutated]), copy-number-low (endometrioid), and copy-number-high (serous-like). Of those, the newly identified POLE subgroup was characterised by high mutation frequency and very good prognosis. On the contrary, the serous-like group was characterised by low mutation frequency and poor prognosis. This new genomic classification of endometrial tumours allows for better stratification of patients for treatment decisions and targeted therapies. In fact, Murali et al. have proposed the integration of molecular and clinico-pathological parameters in order to improve prognostic accuracy (Murali et al. 2014).

1.2.3 Staging
Staging of endometrial carcinoma is determined according to the International Federation of Gynaecology and Obstetrics (FIGO) updated guidelines as published in 2009 (Pecorelli 2009). The FIGO staging comprises a numbering system of 4 stages ranging from Stage I (tumour is confined in the corpus uteri) to Stage IV (distant metastasis present). Alternatively, the American Joint Committee on Cancer TNM staging system is used, which is based on the extent of the tumour (T), spread to lymph nodes (N), and spread to distant sites (M) (Greene 2002).

1.2.4 Screening and detection
To date, there is no standard diagnostic test for identification of early stage asymptomatic patients. The majority of early stage endometrial cancer patients are diagnosed with symptoms of abnormal uterine bleeding, as it is present in 90% of the
cases (Colombo et al. 2013). Initial examination involves pelvic screening in order to rule out other causes of uterine bleeding, such as endometriosis. The standard method of diagnosis in the UK is endometrial biopsy in conjunction with transvaginal ultrasound. Endometrial tissue samples are obtained through hysterectomy or dilatation and curettage (Saso et al. 2011). Histology represents the most reliable method for a definite diagnosis. Histology is heavily dependent on tissue quality; scant biopsy material can lead to difficulties in interpretation of tumour types (Wei et al. 2013). Additionally, a large portion of endometrial carcinomas have similar morphological features, which makes it difficult to classify by histology alone (Wei et al. 2013). Therefore, the development of a non-invasive diagnostic method for endometrial carcinomas would be beneficial for the patients.

1.2.5 Therapeutic strategies

The main therapeutic strategies for treatment of endometrial cancer include, surgery, radiation therapy, hormonal therapy and chemotherapy (pre- or post-surgical). Decisions for specific patients’ treatment are made based on tumour stage and grade (Cancer Research UK, 2016). Surgery is the standard treatment for endometrial cancer and consists of hysterectomy, often in combination with removal of ovaries and fallopian tubes, and removal of lymph nodes. In cases of young women that wish to have children hormonal therapy might be considered (i.e. progestin, Tamoxifen, Letrozole). Radiation and chemotherapy are usually administered after surgery to patients with positive nodes, and/or late stage tumours (stage 3 or 4) with cervical involvement (Lachance et al. 2008; Meyer et al. 2015). As with breast cancers, there has been a recent interest in immunotherapies for endometrial cancer, although on a smaller scale. However, there are currently two immunotherapies in clinical trial for the treatment of endometrial cancer, amongst other cancers (ClinicalTrials.gov identifiers: NCT01174121, NCT00423254). As discussed previously in this chapter, given the importance of the cross talk between cancer cells and immune cells in the surrounding tumour microenvironment, the development of therapies that target the immune system may improve response to therapies and increase patient survival in endometrial cancer as well.
1.3 Tumour microenvironment (TME)

Tumours are complex tissues consisting of different cell types that engage in heterotypic interactions with one another (Bissell & Radisky 2001; Hanahan & Weinberg 2011). Cancer research has been focused primarily on the tumour cells but it has been demonstrated that mutations alone cannot account for the heterogeneity of tumours. In a seminal review, Hanahan and Weinberg proposed six essential alterations (“hallmarks of cancer”) of normal cells that govern the development of malignancy, namely self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan & Weinberg 2000). However, tissue invasion and metastasis are the major causes of mortality in human cancers and represent the most important difference between benign and malignant tumours (Lazebnik 2010). In a more recent review, the authors suggested tumour-promoting inflammation as an ‘enabling characteristic’ to the acquisition of the six hallmarks of cancer (Hanahan & Weinberg 2011).

Tumour progression is affected by extracellular signals such as cytokines, chemokines, extracellular matrix (ECM) molecules and adhesion molecules (Joyce & Pollard 2009). Expression of these factors induces biological processes such as angiogenesis, inflammation and recruitment of circulating cells in the tumour site. Thus, cancer cells in the primary tumour are surrounded by a microenvironment consisting of malignant, stromal and inflammatory cells (Figure 3). Increasing evidence suggests that the TME and the comprising cells types play an important role in tumour progression (Joyce & Pollard 2009; Hanahan & Coussens 2012; Quail & Joyce 2013).

1.3.1 Cells types within the tumour microenvironment

The TME in addition to cancer cells consists of various components that play different roles during malignancy. It is comprised of extracellular matrix, cells of mesenchymal origin such as, fibroblasts, myofibroblasts, mesenchymal stem cells (MSCs), adipocytes and endothelial cells, and cells of hematopoietic origin that can
be subdivided into cells of lymphoid and myeloid lineage. Cells of the lymphoid lineage include T cells, B cells and natural killer (NK) cells, while cells of the myeloid lineage include macrophages, neutrophils and myeloid-derived suppressor cells (MDSCs) (Pattabiraman & Weinberg 2014). Each of these cell types can have distinct pro-tumoral or anti-tumoral roles during tumour progression.

Several studies have shown a correlation between immune cell infiltration and patient prognosis (Bingle et al 2002b; Joyce & Pollard 2009). Reportedly, leukocyte infiltrates can comprise up to 50% of the tumour mass (Pollard 2004). Clinical and epidemiological studies have implicated macrophages with cancer initiation and promotion (Balkwill et al. 2005), and correlated their density to poor prognosis (Bingle et al 2002b). However, this is variable between, not only individual tumours, but also different cancers. Tumour-associated macrophages, or TAMs, have been shown to play a number of key roles in breast tumour progression (Pollard 2004; Joyce & Pollard 2009), with many therapies now focusing on inhibition of their functions (DeNardo et al. 2011). Depending on the cancer, TAMs can be indicators of either a positive or negative prognosis in human disease (see section 1.4.3). High TAM infiltration in stomach cancer results in a favourable prognosis (Ohno et al. 2003), whereas there is conflicting evidence for their role in prostate, lung and brain tumours (Bingle et al 2002a).
Figure 3. The primary tumour microenvironment. The TME consists of many different components with distinct roles during development and progression of malignancy. It comprises of extracellular matrix, cells of mesenchymal origin, and cells of hematopoietic origin. Illustration modified from (Joyce & Pollard 2009).

### 1.4 Macrophages

#### 1.4.1 Tissue-resident macrophage origins and functions

Macrophages are phagocytic cells of the myeloid lineage that regulate innate immune responses. Tissue-resident macrophages can be found throughout the body such as, Kupffer cells located in the liver, alveolar macrophages located in the lung, marginal zone macrophages located in the spleen and microglia located in the brain. Historically, tissue-resident macrophages are thought to be derived from circulating monocytes, thus having a hematopoietic progenitor in the bone marrow (van Furth & Cohn 1968). However, recent evidence in mouse models suggests that this is not the case. It has been shown that tissue-resident populations maintain their population through local proliferation with very little contribution from circulating monocytes.
(Jenkins et al. 2011; Schulz et al. 2012) and additionally, recent evidence using lineage tracing techniques has also shown that the majority of tissue resident populations are primarily derived from the yolk sac. In contrast with Langerhans cells that seem to have a mixed origin from yolk sac and fetal liver (Ginhoux et al. 2010; Hoeffel et al. 2012). Although, dermal and intestinal macrophages have been demonstrated to be continuously replaced by circulating monocytes (Zigmond et al. 2012; Bain et al. 2013). Taken all together, these data suggest three different embryological origins of macrophages in the mouse. On the contrary, few studies have examined the ontogeny of human tissue resident macrophages in humans due to the difficulties in performing genetic lineage in humans. During HSC transplantation, dermal macrophages displayed increased survival and slower replacement compared to dermal dendritic cells (DC), reinforcing the idea of self-maintenance in humans. Moreover, patients lacking blood monocytes due to a GATA2 mutation displayed normal numbers of Langerhans cells and macrophages in the skin and lung supporting the idea that these populations are independent of monocytes and DCs (Guilliams et al. 2014; Epelman et al. 2014). Although, there is some evidence that human macrophages might have similarities to mouse macrophages further work is needed.

Macrophages are among the first cells to arrive at sites of infections and through the secretion of growth factors, angiogenic factors and proteases, carry out pro-inflammatory functions and also promote tissue repair. They also produce cytokines and chemokines that regulate the recruitment and functions of other immune cells. They can be antigen-presenting cells and are involved in killing of pathogens through the production of reactive oxygen and nitrogen radicals. In addition to their inflammatory functions, macrophages play also important roles in tissue morphogenesis during development (Wynn et al. 2013). Mouse models with a deficiency for macrophages result in developmental defects such as osteopetrosis, dermal dysplasia and abnormal morphogenesis of the mammary gland (Pollard 2009).
1.4.2 Macrophage polarization

Macrophages are plastic cells and adopt different transcriptional profiles and functions based on signals from the local environmental (Lavin et al. 2014). The diversity of phenotypes and functions of macrophages has led to a number of classification attempts. A classical binary classification referring to inflammatory states in vitro consists of ‘classically activated’ and ‘alternatively activated’ macrophages. These two activation states are characterised by responses to interferon-γ (IFN-γ), Tumour-necrosis factor-α (TNF-α) and Toll-like receptor-4 (TLR4) (classically activated), and interleukin-4 (IL-4) and IL-13 (alternatively activated) (Gordon 2003). This classification was later extended to Th1/M1 (classically activated) and Th2/M2 (alternatively) in in vivo models (Mantovani et al. 2002; Mills 2012a). M1 activated macrophages are described as proinflammatory and cytotoxic and M2 activated macrophages are described are anti-inflammatory with tissue trophic and repair functions (Mantovani et al. 2002). However, this dualistic classification has been criticized as being overly simplistic and not capturing the complex environmental signals and distinct fates of different macrophage populations (Pollard 2009; Ostuni et al. 2015). In fact a comprehensive transcriptomic analysis of tissue–resident macrophages from several mouse tissue organs conducted from the Immunological Genome Project demonstrated discrete gene expression signatures between the different macrophage populations with underlying themes (Gautier et al. 2012).

1.4.3 Tumour-associated macrophages (TAMs)

In contrast to tissue-resident macrophages, it has been demonstrated that most TAMs in mouse models are derived from a classical “inflammatory” monocytic population (Ly6Ch) in lung (Qian et al. 2011), grafted tumours (Movahedi et al. 2010), and mouse mammary tumours (Franklin et al. 2014). For a long time it was thought that high number of macrophages within or near the tumour indicated an immune response against tumour development. However, now it is appreciated that macrophages at least in animal models, play an important role in tumour promotion and metastasis (Joyce & Pollard 2009). Support for these roles in humans comes
from clinical data showing that increased macrophage density has been associated with poor survival in many cancers (Bingle et al. 2002a) including thyroid (Ryder et al. 2008), lung (Chen et al. 2005), endometrial (Kelly et al. 2015) and hepatocellular carcinoma (Zhu et al. 2008). Based on a meta-analysis high TAM density was associated with worse overall survival in patients with gastric cancers, urogenital cancers and head and neck cancers (Zhang et al. 2012). However, high macrophage density found to be correlated with favorable patient outcomes in pancreatic cancers (Kim et al. 2008) and colorectal cancers (Zhang et al. 2012). In mouse models of breast cancer, transcriptional analysis of isolated TAMs showed an association with poor overall survival (Ojalvo et al. 2009).

1.4.3.1 TAM recruitment

Many different regulatory factors for TAM recruitment exist, however the most studied in cancer are colony stimulating factor 1 (CSF1) and CCL2. CSF1 represents the major regulator of macrophages; its overexpression has been associated with poor prognosis in many cancer types, some of which include breast cancer, endometrial cancer and clear-cell renal carcinoma (Smith et al. 1995; Lin et al. 2002; Mroczko et al. 2007; Yang et al. 2015). In the Polyoma Middle T oncoprotein (PyMT) mouse model, depletion of CSF1 resulted in decreased TAM density, delayed tumour progression and reduced metastasis (Lin et al. 2001). Beck et al identified a CSF1 response signature composed of genes over-expressed in tumours that are mostly populated by CSF1R expressing macrophages (Beck et al. 2009). The signature did not show a significant association with poor survival in 4 out the 5 publically available breast cancer datasets, but it was associated with decreased survival between grade 1 and grade 2 tumours (Beck et al. 2009). Interestingly, the signature was associated with a good prognosis in ER-negative tumours and tumours with TP53 mutations (Beck et al. 2009). Additionally, CCL2 plays a critical role in recruitment of TAMs in neoplastic sites, and treatment with anti-CCL2 antibodies resulted in reduced metastasis in mouse mammary tumour models (Qian et al. 2011; Kitamura et al. 2015).
1.4.3.2 TAM functions in tumour development and progression

TAMs carry out a number of functions that play an important role for tumour survival, growth and progression. TAMs promote tumour cell survival and proliferation through secretion of various growth factors, such as platelet-derived growth factor (PDGF), basic fibroblast growth factor (bFGF), TGFα and β and EGF (Pollard 2004). TAMs also support angiogenesis, a key process required for tumour growth (Qian & Pollard 2010). In the PyMT mouse expression of VEGF over loss of CSF1 resulted in increased macrophage density, sustained angiogenesis and increased tumour growth (Lin et al. 2007). TAMs can also promote tumour cell migration and invasion (Wyckoff et al. 2007; Chen et al. 2011), and tumour cell seeding of metastatic sites (Qian et al. 2009; Noy & Pollard 2014). All together, these findings give evidence of the major role of TAMs in tumour development and promotion of malignancy. TAMs have also been described to possess immunosuppressive functions (Mantovani et al. 2002) through their secretion of anti-inflammatory cytokines such as IL-10 that are able to suppress immune responses against the tumour (DeNardo et al. 2011). Indeed, blockade of IL-10 receptor resulted in increased primary tumour response to chemotherapy (Ruffell et al. 2014). Although, IL10 did not directly suppress CD8-positive T cell activation, in vitro studies have shown a direct suppression of CD8-positive T cell activation by TAMs (DeNardo et al. 2011; Movahedi et al. 2010; Doedens et al. 2010).

1.4.3.3 Heterogeneity of TAM subpopulations

Although TAMs appear to share a common monocytic progenitor they don’t represent a phenotypically homogenous population (Qian & Pollard 2010; Lahmar et al. 2016). Similarly to resident macrophages, classification attempts have suggested that TAMs within/near the tumour represent a population that is more similar to M2-rather than M1-like phenotype (Mantovani et al. 2002). In reality, studies have shown that distinct TAM subsets (MHC II hi, MHC II low) exist and express genes associated with both M2 and M1 polarization. In addition, these distinct TAM subsets are localized in either hypoxic (MHC II low) or well-oxygenated regions (MHC II hi) (Movahedi et al. 2010; Laoui et al. 2014; Ruffell et al. 2014). Distinct
The phenotypes of TAMs were found in different regions of the tumour with distinct behavioral and molecular characteristics (Wyckoff et al. 2007; Egeblad et al. 2008). Gene expression profiling identified an ‘invasive TAM’ population associated with promotion of tumour cell invasion that showed a distinct gene expression pattern from the general TAM population (Ojalvo et al. 2010). These results together support the notion that distinct subpopulations of TAMs are associated with distinct functions (Lewis & Pollard 2006). In addition to TAMs, metastatic-associated macrophages (MAMs) have been shown to play roles in tumour cell extravasation, survival and tumour growth in breast cancer lung metastasis models (Qian et al. 2009; Qian et al. 2011). The TME can affect the functions and phenotypes of macrophages, and MAMs represent yet another phenotypically distinct cancer-associated macrophage population.

1.4.3.4 Macrophages as therapeutic targets

Given the myriad of roles macrophage play in tumour progression there is growing interest in the therapeutic targeting of these cells. The main therapeutic approaches include a) inhibition of the CSF1/CSF1R signalling pathway, b) blocking of CCL2 monocyte recruitment (Qian et al. 2011) and c) reprogramming of TAMs towards anti-tumoral responses (Ruffell et al. 2014). The most encouraging therapeutic approach relies on the inhibition of the CSF1/CSF1R signalling pathway; indeed blocking the CSF1R resulted in decreased tumour growth and stimulation of CD8+ cytotoxic T cells in mouse models of breast and cervical cancer (Lin et al. 2001; DeNardo et al. 2011; Strachan et al. 2013). In a mouse model of glioblastoma, inhibition of the CSF1 receptor resulted in increased survival and regression of established tumours (Pyonteck et al. 2013). Strikingly, the use of a monoclonal antibody (RG7155) against the CSF1 receptor decreased the number of TAMs and increased the CD8+/CD4+ T cell ratio in animal models and patients with diffuse-type giant cell tumours, the latter resulting in therapeutic benefit (Ries et al. 2014).

A key limitation of current targeting of TAMs is that there are no TAM-specific markers, hence current therapeutic strategies also deplete non-tumour macrophage
populations leading to side effects. These include deficiencies in the immune system, elevated liver enzymes, dose limiting toxicology and oedema. Therefore, identifying TAM-specific markers may aid in reducing these side effects and enhance the efficacy of TAM-based therapies. Perhaps most importantly, most of our knowledge regarding the functions and phenotypes of macrophages comes from mouse models and may not be translatable into humans. Thus, it is important to investigate the transcriptional profiles of TAMs in human cancers in more depth, in order to determine if they play the similar roles in human cancer to the roles they play in promoting tumour progression in mouse cancer models. Hence, it is the purpose of this thesis to define human TAMs.

1.5 Monocytes

Monocytes are a heterogeneous cell population that represents 5-10% of peripheral blood mononuclear cells (PBMCs). They are bone marrow-derived cells and circulate in the blood for one to three days. They represent immune effector cells that through chemokine and adhesion receptors migrate to tissues and give rise to macrophages and dendritic cells (DCs) (Geissmann et al. 2010).

1.5.1 Monocyte subpopulations

Circulating blood monocytes are classified in three distinct populations: classical “inflammatory” (Human: CCR2HighCD14++CD16−, mouse: CCR2HighLy6CHigh), intermediate monocytes (CD14++CD16⁺), and nonclassical “patrolling” monocytes (Human: CX3CR1HighCD14⁺CD16+++, mouse: CX3CR1HighLy6CLow) (Geissmann et al. 2003; Ziegler-Heitbrock et al. 2010). Transcriptional analysis of monocyte subsets between human and mice showed some discrepancies, however their differentiation and immune functions was comparable (Ingersoll et al. 2010). Monocyte heterogeneity has been demonstrated in both human and mouse monocytes by gene expression profiling of the CD16-positive and CD16-negative subpopulations revealing distinct transcriptional profiles and biological functions (Frankenberger et al. 2012; Ancuta et al. 2009). Generally, it was thought that classical monocytes are the first to arrive at sites of inflammation and mediate inflammatory responses,
whereas the non-classical monocytes were recruited later and promoted tissue repair (Ginhoux & Jung 2014); however it is not as clear-cut. Ly6C\textsuperscript{low}/CD14\textsuperscript{+}CD16\textsuperscript{+}
monocytes, also termed patrolling monocytes, have been shown to patrol blood vessels walls and rapidly invade tissue during inflammation and differentiate to macrophages (Auffray et al. 2007; Cros et al. 2010). Since then it has been established that human patrolling monocytes patrol blood vessels and exert effector functions in the inflammatory responses against viruses, resembling mouse patrolling monocytes (Cros et al. 2010). Currently, the full spectrum of functions of monocytes, and in particular the specific functions of the monocyte subpopulations are still unclear.

1.5.2 Monocytes in cancer
To date, the mechanisms by which monocytes are involved in tumour progression are not known. It has been shown that classical monocytes are recruited from the circulation to primary tumour sites where they differentiate into pro-tumoral macrophages (Qian et al. 2011; Movahedi et al. 2010). In the PyMT mammary tumour model, it was found that non-classical monocytes were recruited to metastatic lungs through the CCL2-CCR2 signalling pathway, whereby they extravasated and differentiated into macrophages, and promoted metastatic growth (Qian et al. 2011). Sanford et al. demonstrated an association of inflammatory monocytes with survival in pancreatic cancer and an important role of the CCL2/CCR2 axis in the recruitment of inflammatory monocytes to the premetastatic liver (Sanford et al. 2013). Hanna et al. proposed a cancer surveillance role of patrolling monocytes that suppressed metastasis to lung via scavenging of tumour materials, as well as the recruitment and activation of NK cells (Hanna et al. 2015). Therefore, further investigation is required into how monocytes, and their sub-populations, are involved in cancer metastasis.

In human cancers elevated absolute monocyte numbers have been associated with poor survival studies in breast cancer (Wen et al. 2015), colorectal cancer (Sasaki et al. 2007), hepatocellular carcinoma (Sasaki et al. 2006), melanoma (H. Schmidt et al. 2005), and head and neck cancer (Chen et al. 2009) suggesting a possible role of
monocytes in the development of neoplasia. Feng et al. found an increase in the “patrolling” monocyte subpopulation in breast cancer patients and a significant negative association with tumour size and stage (Feng et al. 2011). Using microarray gene profiling, Chittezhath et al. demonstrated that isolated monocytes from renal carcinoma patients are transcriptionally distinct from those of healthy individuals and exhibited a proinflammatory and pro-tumoral profile (Chittezhath et al. 2014). More recently, Hamm et al. reported distinct gene expression profiles of monocytes in colorectal cancer compared to healthy individuals. Additionally, they identified a 23-gene signature able to diagnose colorectal cancer with high specificity and sensitivity both in an internal, and an independent cohort. The reported signature was demonstrated to be specific to colorectal cancer (compared to gastric and stomach datasets) and specific to monocytes (compared to PBMC datasets). Additionally, they presented preliminary results for the potential of the signature to detect disease relapse (Hamm et al. 2015).

These studies confirm the plasticity of circulating monocytes and their potential as diagnostic and prognostic markers. However, to date, little is known about the gene expression profiles and functions of circulating monocytes in human cancer. Seeing as monocytes are TAM precursors, and have been shown to promote metastasis in pre-clinical studies, it is possible that they also actively promote metastasis in human cancers. Therefore it is important to investigate their transcriptomes using sequencing technologies that are not restricted to known probes and evaluate not only their diagnostic potential but also the mechanisms by which they can promote tumour progression. An interesting question worth investigating is related to monocyte heterogeneity; are monocytes similar or distinct between healthy individuals and cancer patients, and are they distinct between different tumour types? This question is addressed in this thesis.
1.6 *High-throughput gene expression profiling technologies*

Cells respond to environmental stimuli by altering the expression of genes. Activated genes are transcribed to messenger RNAs. The quantification of the amount of transcripts produced for a gene is defined as gene expression (Mariadason et al. 2003). Experimental techniques for the quantification of gene expression include northern blotting (Alwine et al. 1977) and quantitative polymerase chain reaction (qPCR) (Gibson et al. 1996). However, these techniques have limitations in terms of the number of genes that can be tested simultaneously (van Hal et al. 2000). The development of hybridization-based methods such as DNA microarrays allows for the profiling of thousands of genes simultaneously and provides a reliable, fast and cost effective method for profiling gene expression patterns (Lipshutz et al. 1995; Schena et al. 1995). Microarray gene expression profiling of tumour samples has provided insights into important molecular processes, and enabled the identification of ‘molecular signatures’ for improved cancer diagnosis as described above. In breast cancer, gene expression profiling has had many applications including the identification of distinct molecular subtypes of breast cancer as described in section 1.1.4; the refinement of the prognostic stratification of patients based on the tumour heterogeneity; and the development of targeted therapeutic treatments (Bao & Davidson 2008).

In contrast to microarrays, tag-based approaches such as serial analysis of gene expression (SAGE) (Velculescu et al. 1995) and cap analysis of gene expression (CAGE) (Shiraki et al. 2003) measure the absolute abundance of transcripts. Although SAGE sequencing has been utilized for the transcriptomic analysis of various diseases, it is quite laborious and therefore the uptake of this method has been more limited (Harbers & Carninci 2005; Costa et al. 2010).

1.6.1 RNA-sequencing technology

RNA-seq is the first sequencing-based method that allows high-throughput quantification of the transcriptome (Mortazavi et al. 2008; Wang et al. 2009a). Compared to microarrays, RNA-seq is not limited to previous knowledge of the genomic sequence and therefore can be used for identification of new splice isoforms.
RNA-seq data are highly reproducible with a higher dynamic range, allowing for the detection of differentially expressed genes with higher fold changes (Malone & Oliver 2011; S. Zhao et al. 2014). However, dynamic range of RNA-seq experiments is very much dependent on the depth of sequencing and the total number of mapped reads. Due to the general high cost of sequencing laboratories choose to sequence with low depth, i.e. < 10M reads, therefore resulting in dynamic ranges comparable to those of microarrays. Additionally, the number of total reads is proportional to the expression levels of transcripts. Consequently, longer transcripts will have more reads mapping compared to shorter transcripts with a similar expression, therefore giving more power to detect differential expression for longer genes (Oshlack & Wakefield 2009). Overall, RNA-seq offers advantages comparing to microarrays, however it posses additional limitations that need to be considered during study design.

1.6.1.1 RNA-seq data analysis and considerations

RNA-seq experiments consist of a library preparation step and a subsequent data analysis step (Figure 4). The first step of any RNA-seq experiment is RNA isolation. RNA (total or fragmented) is converted into a library of cDNA fragments using first-strand and second-strand synthesis. After cDNA conversion, adaptors are attached to one or both ends and an optional step of amplification and size selection follows. The library is assessed for its quality and sequenced (Wang et al. 2009a; Zhao et al. 2016).
Figure 4. Overview of RNA-seq bioinformatics workflow. The first step of any RNA-seq experiment is RNA isolation and library preparation. The library is assessed for its quality and sequenced. Mapped reads are then used for quantification at the gene/transcript/exon level. Reads that fall into known genomic regions are counted. Summarised data are first normalised and then used for differential expression analysis. Differentially expressed genes between different phenotypes are used for pathway and enrichment analysis. Illustration adapted from (Han et al. 2015).

The result of a sequencing run consists of millions of raw reads that will be used downstream in the bioinformatics data analysis. The first and most important step of data analysis is quality control. Poor quality reads can arise from problems in library preparation, polymerase chain reaction (PCR) artefacts or possible contaminations.
Therefore is very important to filter data by removing low quality reads or trimming adaptors and overrepresented sequences.

After going through quality assessment, reads can be used for mapping to the reference genome. If a reference genome is not available, de novo assembly of the genome is performed. A variety of different tools have been developed for the alignment of raw reads to the reference genome and have been extensively evaluated (Engström et al. 2013). A problem that can arise during alignment is that of multiple reads, which occur when a read is aligned to multiple locations in the genome. A common solution is to set user-defined cut-offs for the number of possible multiple reads (Zhao et al. 2016).

Mapped reads are then used for quantification at the gene/transcript/exon level. The quantification step involves counting of reads that fall into a known genomic region. The number of reads in each genomic region represents the measured abundance. In this step the selection of tools for inferring gene abundance and gene annotation are very important. Different tools and different parameters generate different read numbers that affect downstream analysis. A recent study showed that out of the 21,958 genes common between RefGene, Ensembl and UCSC annotation only 16.3% obtained identical quantification results (Zhao & Zhang 2015).

Summarised data are first normalised and then used for differential expression analysis. Data normalization is a crucial step that directly affects the accuracy of gene expression and the results of the differential expression analysis. Sequencing depth or library size (total number of mapped reads) varies between samples, which makes the estimated abundance of genes not directly comparable. An equally important factor is gene length as it can introduce bias that influences the number of differentially expressed genes (Han et al. 2015; Zhao et al. 2016).

After the pre-processing steps, a fundamental research aim is to identify differences in genes between sample groups. A plethora of different tools based on different statistical methods has been developed for differential expression analysis. Even though several comparative studies have evaluated the performance of different methods, a general consensus regarding optimal study design and software for RNA-seq data analysis is missing (Rapaport et al. 2013; Soneson & Delorenzi 2013;
Zhang et al. 2014; Seyednasrollah et al. 2015). Other important factors to consider are the number of biological replicates and the sample sizes. Small sample sizes are a common phenomenon as RNA-seq can be costly. Soneson and Delorenzi evaluated 11 methods for differential expression analysis and concluded that small sample sizes affect all compared methods and results should be interpreted with caution (Soneson & Delorenzi 2013). The number of replicates to be included in any RNA-seq experiment depends both on technical variability during RNA-seq procedures but also on biological variability. Conesa et al. recommended that a minimum of three replicates per condition is needed to capture the biological variability within samples with sufficient statistical power (Conesa et al. 2016). Other factors that need to be considered before differential expression analysis are batch effects and other technical artefacts that can be introduced during sample preparation or sequencing. It is important to use tools to control for batch effects as they can lead to bias and variability within the final results.

The output of differential expression analysis is a list of differentially expressed genes that serve as a starting point to gaining biological insight into experimental systems. Downstream analysis of differentially expressed genes includes gene ontology and pathway analysis (Garber et al. 2011; Zhao et al. 2016; Han et al. 2015).

1.7 Gene signatures for diagnosis and prognosis of cancer

One of the main goals of cancer research has been the identification of relevant biomarkers (gene signatures) for early detection, prognosis and targeted therapy. With the advent of high-throughput technologies large amounts of cancer genomic data have been made available to the research community. Machine learning (ML) methods are increasingly applied in cancer research in an effort towards personalized clinical care.

Machine learning is a branch of artificial intelligence that is used to identify patterns within complex datasets. It uses prior knowledge to predict patterns and trends in new datasets. ML algorithms can be broadly divided in supervised and unsupervised.
Supervised learning algorithms are using training data with known labels/classes to ‘learn’ how to determine or map the input to the desired output. One instance of supervised learning that is widely used in cancer research is classification, where data are categorized in distinct classes. In contrast, unsupervised learning algorithms do not use labels/classes and the goal this time is to identify hidden patterns or groups. A commonly used unsupervised learning method is clustering, where groups/clusters are identified from unlabeled data.

Much work has been done on the identification of multigene signatures for molecular classification and prognosis of clinical outcomes (Sims 2009; Weigelt et al. 2010). These studies have highlighted the heterogeneity of breast and endometrial cancer and the fact that distinct subtypes can have distinct clinical outcomes. It is evident that the usage of ML classification methods significantly improves the accuracy of cancer prediction outcomes (by 15%-20%) (Cruz & Wishart 2006).

### 1.7.1 Gene expression-based diagnostic assays

Early detection of breast and endometrial cancer has resulted in a significant decrease of mortality rates. However, there is substantial room for improvement within the current early detection methods for breast cancer, and there is yet no early detection method for endometrial cancer. Thus, there is an urgent need for development of non-invasive screening methods for asymptomatic women that do not rely on assessing malignant tissue.

One alternative to tissue biopsies is assessing blood cells. PBMCs comprise of different types of cells including monocytes, T-cells, B-cells and NK cells. Gene expression profiling of PBMCs can provide an easy minimally-invasive source for early detection and monitoring of malignancy. Nichita et al. identified a 4-gene signature that could distinguish colorectal cancer patients from healthy individuals (Nichita et al. 2014a). Similar studies in breast cancer have identified signatures for potential early identification of breast cancer (Sharma et al. 2005; Aarøe et al. 2010a; LaBreche et al. 2011; Dumeaux et al. 2014). Sharma et al. was the first to report the effect of breast cancer on PBMCs and identified a 37-gene signature for early
detection of breast cancer (Sharma et al. 2005). Aarøe et al. described a classifier based on 738 probes for the stratification of breast cancer patients from healthy individuals (Aarøe et al. 2010a). LaBreche et al. developed a classifier derived from a mouse mammary tumour model that could distinguish PBMC samples from human breast cancer and normal samples (LaBreche et al. 2011). PBMCs have been described as promising early detection biomarkers in several other cancer types including pancreatic cancer (Baine et al. 2011), lung cancer (Showe et al. 2009), and renal cell carcinoma (Twine et al. 2003). Despite these findings, the clinical application of PBMCs is still unclear.

Nevertheless, PBMCs represent a mixed population of immune cells with a spectrum of pro-tumoral and anti-tumoral functions during malignancy. Therefore, changes in gene expression can be lost when looking at these cell populations as a whole and understanding the specific functions and responses of individual immune cells types within this mixed population it is likely to offer better diagnostic strategies for cancer. In line with this, it was demonstrated that mRNA of tumour-educated platelets could be used for stratification of cancer patients from healthy individuals with high accuracy (Best et al. 2015). Additionally, transcriptional profiling of circulating monocytes in colorectal cancer identified signatures for early diagnosis of colorectal cancer (Hamm et al. 2015).

Recently, blood-based ‘liquid biopsies’ from circulating tumour cells (CTCs) and cell-free tumour DNA (ctDNA) have been proposed. Although, these approaches are very promising, they currently face challenges regarding sensitivity and specificity (Alix-Panabières & Pantel 2016). However, combination of gene expression and emerging liquid biopsy diagnostic assays could improve current diagnostic assays.

### 1.7.2 First generation gene expression prognostic assays

Several prognostic assays have been identified for guided treatment of breast cancer patients, however the majority of these prognostic signatures have been useful primarily in predicting prognosis in patients with ER-positive tumours but their efficacy is limited in ER-negative tumours (Fan et al. 2009; Prat et al. 2012; Wirapati et al. 2008; Iwamoto et al. 2011). For breast cancer three of the most advanced
prognostic signatures for available for clinical use are, MammaPrint®, OncotypeDx® and Prosigna. OncotypeDx® and Prosigna have been recently recommended for guiding adjuvant therapy decisions by the American Society of Clinical Oncology (ASCO) (Harris et al. 2016).

The 70-gene MammaPrint® is a microarray-based gene expression prognostic assay tailored for breast cancer patients with stage I/II, node negative disease that stratifies patients into high- and low-risk prognostic groups independently of clinico-pathological factors (Van ’t Veer et al. 2002). MammaPrint® showed potential in selecting patients for adjuvant chemotherapy, identifying a subset of high-risk patients (based on clinico-pathological factors) that were reclassified as having low genomic risk and could be spared chemotherapy treatment (Cardoso et al. 2016).

The OncotypeDx® uses ‘formalin fixed paraffin embedded’ (FFPE) tissue with qPCR and is the most widely used prognostic assay for women with hormone receptor positive, node negative breast cancer. It is based on the expression of 16 cancer related genes and 5 housekeeping genes. Patients are stratified in three risk groups based on a recurrence score (RS) ranging from 0 to 100; low-risk (RS <18), intermediate-risk (18< RS <30) and high-risk (RS >=31) (Paik et al. 2004; Sparano & Paik 2008). Recently, a prospective validation study showed that ER-positive patients with high RS showed benefit to chemotherapy treatment compared to low RS patients (Sparano et al. 2015). However, the optimal therapeutic strategies for intermediate-risk patients are still unclear.

The PAM50/Prosigna is a gene expression assay based on 50 genes that were initially used for the identification of breast cancer subtypes. The test computes a risk of recurrence (ROR) score for estimation of distant recurrence-free survival for stage I/II women with ER-positive cancer treated with adjuvant endocrine therapy (Parker et al. 2009).

Several other prognostic assays have been developed for breast cancer such as MapQuant Dx, Breast Cancer Index (BCI) and EndoPredict. The Genomic Grade Index (GGI) MapQuant Dx is a microarray gene expression prognostic assay based on a 97-gene signature. The test is used on ER-positive cancer and reclassifies intermediate-grade tumours into high- or low-grade with associated clinical
prognosis (Sotiriou et al. 2006). EndoPredict has been validated for the stratification of ER-positive, HER2-negative cancers. The test classifies patients treated with adjuvant therapy in low- or high-risk recurrence groups based on the expression of 8 genes (plus 3 housekeeping) in order to calculate an EP score. EP score in combination with nodal status and tumour size estimates a comprehensive risk score referred to as EPclin. It should be noted that, EndoPredict and BCI tests have been recently recommended by ASCO to guide decisions regarding adjuvant therapy treatments (Harris et al. 2016). No such tests exist for endometrial cancer, although Levan et al. through hierarchical clustering, identified a subgroup of endometrioid tumours associated with decreased survival (Levan et al. 2010), indicating the potential for a gene-signature-based test for endometrial cancer.

Numerous other prognostic gene signatures have been described beyond those discussed herein. Despite the fact that there is minimal overlap of individual genes between prognostic signatures, the discriminatory value of each test is comparable especially in ER-positive tumours (Prat et al. 2012; Zhao et al. 2014; Desmedt et al. 2008). This has been attributed to the fact that the prognostic power of these tests is based on proliferation-related genes (Desmedt et al. 2008; Wirapati et al. 2008). However, retrospective studies comparing different prognostic tests on the same patient sample population have demonstrated significant discordance in terms of assigned risk category as well as of individual patients prognosis (Varga et al. 2013; Vargas et al. 2012; Kelly et al. 2012; Dowsett et al. 2013; Iwamoto et al. 2011; Bartlett et al. 2016).

Hence in breast cancer, the current prognostic assays provide additional benefit to guided treatment decisions than conventional clinico-pathological guidelines for ER-positive tumours, however their application on ER-negative cancers is limited. Also, current prognostic signatures for breast cancer rely on genes enriched for pathways related to cancer cells with little regard to TME. Whereas in endometrial cancer, no prognostic assays currently exist. Thus there is a need for improved prognostic assays for breast cancer and the development of prognostic assays for endometrial
cancer. Furthermore, a need exists for the development of detection assays for both cancers.

1.7.3 Stroma and immune-related prognostic signatures

The contribution of the TME and its cellular compartment in tumour progression and development has been recognised. Most prognostic assays described so far are derived from whole tissue consisting of tumour cells and surrounding stroma. Samples with low tumour epithelial cell content are usually excluded for further analysis. In this way, important components of the TME such as tumour-associated stromal cells and infiltrating immune cells are ignored (Manjili et al. 2012).

To determine the prognostic value of stromal cells, Finak et al. used laser capture microdissection to isolate normal and matched tumour stroma from breast cancer individuals. Using microarrays and ML, they identified a 26-gene signature that could classify patients into poor and good outcome groups. Their signature showed greater prognostic power than predictors derived from whole tissue datasets; and especially in the HER2-positive subgroup. Additionally, the good outcome group was enriched in immune related genes including T cells and NK cell markers. In contrast, the poor outcome group was enriched in genes related to angiogenesis and hypoxia (Finak et al. 2008). Desmedt et al. identified a 12-gene signature through comparison of tumour-associated CD10+ cells with CD10+ cells in normal breast tissue. Their signature was particularly accurate in differentiating between DCIS and invasive tumours in the HER2+ subgroup (Desmedt et al. 2012). Gene expression profiling of fine-needle (stroma-poor) and core-needle (stroma-rich) biopsies in breast cancer individuals identified a B cell/plasma gene set associated with good prognosis in ER-positive and ER-negative high proliferating tumours (Bianchini et al. 2010). Another, stroma-derived signature has been described to predict poor response to adjuvant chemotherapy (Farmer et al. 2009).

There is increasing evidence showing an important role of the immune system in early prognosis of breast cancer. Several immune-based signatures have been proposed for prognosis of ER-positive and ER-negative tumours (Desmedt et al. 2008; Schmidt et al. 2008; Teschendorff et al. 2007; Rody, Holtrich, et al. 2009).
Teschendorff et al. identified a subgroup of ER-negative tumours where over-expression of immune-related genes was associated with good prognosis (Teschendorff et al. 2007). Recently, it was demonstrated that immune-related genes could predict clinical benefit of adjuvant trastuzumab (Perez et al. 2015). However, the value of these immune-related signatures needs to be further validated (De Mattos-Arruda et al. 2014).

Tumour-infiltrating lymphocytes (TILs) are now emerging as markers for good prognosis in ER-negative breast tumours (Alexe et al. 2007; Calabrò et al. 2009; Manjili et al. 2012). Indeed, in a meta-analysis of 1,044 breast cancer samples, Calabrò et al. demonstrated that high expression of TIL markers in ER-positive cancers is associated with shorter survival, whereas high expression of TIL markers in ER-negative tumours is associated with favorable prognosis (Calabrò et al. 2009). More recent studies have shown a strong association of increased number of TILs with longer disease-free survival and favorable prognosis in TN breast cancers (Loi et al. 2014; Pruneri et al. 2016; Wang et al. 2016; García-Teijido et al. 2016). Given these findings, and the previously described pro-tumoral roles of TAMs it is likely that assessment of transcriptional changes in these cells may yield improved prognostic tests.

Overall these studies highlight the importance of the TME and the role of infiltrating immune cells in prognosis of breast cancer, and especially on ER-negative tumours. It is clear that gene expression profiling of distinct cell types within the TME that exert tumour-promoting functions such as TAMs could offer powerful markers for prognosis and targeted therapy of breast cancer and endometrial cancers.

1.7.4 Considerations when developing diagnostic and prognostic signatures

1.7.4.1 Sample size and patient heterogeneity

A big body of research has focused on the identification of signatures for prognosis and diagnosis of cancer. However, a very common issue of many of these studies is the lack of attention to sample size. A basic requirement for any ML algorithm is
dataset big enough to be partitioned into training and testing subsets. In most cases, training too many times in too few samples results in overtraining that leads to an overestimated performance of the prognostic/diagnostic model (Cruz & Wishart 2006). Another frequent problem is this of imbalanced datasets, i.e. more cancers samples than controls, again this leads to distorted results (Cruz & Wishart 2006; Sims 2009). Gene signatures derived from single time points such as pre-treatment or after neoadjuvant therapy might be affected by patient heterogeneity (tumour size, age, tumour grade, lymph-node status). In order to minimise this bias, one approach is to utilise “matched samples” from the same patient of tumour and normal tissue. A different approach would be to increase the sample size in order to make the patient population as similar as possible and increase the likelihood that the identified variation is only due to real variation between groups (Sims 2009).

1.7.4.2 Model and feature selection
One of the most important decisions when building prognostic and diagnostic assays is model selection. There is a plethora of methods, with different assumptions, data requirements and strengths and weaknesses. Therefore, it is very important to compare different ML algorithms on a given training dataset in order to choose the right tool for the right application (Cruz & Wishart 2006). Along with data quality and sample size, feature selection plays an important role in downstream interpretation (Song et al. 2015). Selecting gene subsets that are reproducible will lead to robust and accurate models (Cruz & Wishart 2006). Current prognostic signatures have been criticized for heavily relying on proliferation-related genes (Wirapati et al. 2008). Additionally, Venet et al. demonstrated that most random gene expression signatures are associated with clinical outcome in breast cancer (Venet et al. 2011). For that reason, rigorous validation is needed when selecting prognostic/diagnostic gene signatures.

1.7.4.3 Independent validation
A major limitation of many studies is the lack of external independent validation in order to verify that the findings have a real association with clinical outcomes.
Validation on an external cohort that has not been used during training should be used to validate the robustness of the signature. Negative validation is needed to determine that the findings are associated with a true prognosis/prediction. External validation is often done on publically available datasets that may or may not have the same study design, inclusion criteria or the purpose of study might be different. This needs to be taken into consideration as it can introduce further bias during interpretation of the results. Further validation of the signature should be done by alternative methods of measuring gene expression such as qPCR. It needs to be noted, that for novel studies there might be a lack of available public datasets to validate findings.

1.7.4.4 Reproducibility of analysis

The prognostic/diagnostic capacity of many identified signatures is often not reproducible partly due to sample size and data quality but also due to technology bias (Sims 2009; Boutros 2015). Studies have shown that gene expression of tissue samples is well correlated between RNA-seq and microarrays (Mooney et al. 2013; Wang et al. 2014). Fumagalli et al assessed the reliability of evaluating existing breast cancer prognostic signatures derived from microarrays with RNA-seq technology; they demonstrated a high correlation of gene expression between the two technologies (Fumagalli et al. 2014). Similarly, Zhang et al showed that the technology platform does not affect the clinical endpoint, and that both technologies have similar performance (Zhang et al. 2015). It needs to be noted that in the study by Fumagalli et al, RNA-seq normalized reads were log-transformed to match the log-scale of microarrays, indicating that pre-processing might play an important role in dataset agreement. Indeed, Thompson et al highlighted the importance of pre-processing for cross-platform comparisons suggesting a new normalization technique for cross-platform machine learning applications (Thompson et al. 2016). Moreover, studies have shown that even the smallest changes in the analysis pipeline can lead to big differences in the final biomarker set (Fox et al. 2014; Starmans et al. 2012). Therefore, a careful standardized pre-processing of the dataset is important for reproducible results, especially for cross-platform comparisons.
Another consideration regarding biomarker reproducibility is the complexity and heterogeneity of tumours; tumours consist of various cell types in different frequencies and with different functions. Existing prognostic markers are derived from studies focusing on inter-tumoral heterogeneity and are not capturing the intra-tumoral heterogeneity that exists within tumours (Boutros 2015).

1.7.4.5 Clinical application of predictive/prognostic signature

Gene expression profiling technologies have altered the way we view breast and endometrial cancer and have vastly improved our understanding of their heterogeneity (Sotiriou & Pusztai 2009; Sims 2009). However, there are a number of considerations to be addressed before they are routinely used in the clinic for disease diagnosis and prognosis. Firstly, it is important to determine if the signature outperforms existing methods or a “golden standard” by evaluating the sensitivity and specificity of the proposed signature. The new signature should perform equally well or better (reproducibility) than current standards (Byron et al. 2016). Ideally, this needs to be shown in a prospective validation study on a large cohort of patients. Additional though must be given to the chosen technology; RNA-seq and microarray technologies are more relevant for investigation purposes rather than clinical practice. Therefore, signatures might be converted to lower cost technologies such as qPCR or targeted RNA-sequencing. Studies have shown that gene expression data from RNA-seq and qPCR studies have excellent correlation (Mortazavi et al. 2008; Wang et al. 2009a). However, this brings attention to the number of genes included in a signature set. Large number of genes might prohibit the use of low cost technologies such as qPCR. Furthermore, most pathological samples are collected as FFPE tissues. FFPE samples are easier to collect and store; however this method results in partially degraded RNA. This is a clear challenge for the clinical application of FFPE samples to the clinic; although, efforts to develop RNA-seq protocols for FFPE samples have shown promising results (Zhao et al. 2014; Cieslik et al. 2015; Byron et al. 2016). To address reproducibility issues, it is recommended that a strict common protocol from sample collection to bioinformatics analysis should be implemented.
1.8 Aims and Objectives

Beast and endometrial cancers are the most common female cancers in the UK. In breast cancer, mammography is the most accurate diagnostic method to date; however its sensitivity is limited in women with higher breast density. In addition to this, routine screening methods for detection of endometrial cancer are lacking. Therefore, there is an urgent clinical need for an accurate non-invasive screening method for both breast and endometrial cancer.

In parallel, the TME and the tumour stroma are increasingly recognized as major players in tumour development and progression. Additionally, the tumour-infiltrating cellular compartment represents promising biomarkers for prognosis and diagnosis of cancer. In mouse models, TAMs have been associated with poor prognosis in both breast and endometrial cancers; however their roles in human cancers are still unclear. Moreover, their precursors, circulating monocytes, have been described to change their transcriptional profiles in the presence of cancer, such as renal carcinoma, and can act as biomarkers for diagnosis and in colorectal cancer.

1.8.1 Hypotheses

1. Circulating monocytes and TAMs will respond to the presence of tumours in regards to their transcriptional profiles.

2. Comparison of the gene expression profiles of circulating monocytes from cancer patients to those of healthy individuals will identify novel biomarkers for the detection of cancer

3. Human TAMs will display altered transcriptomes compared to resident macrophages allowing for identification of TAM-specific markers.

4. Transcriptional profiles of human TAMs in breast cancer will identify biomarkers able to distinguish patients based on clinical outcome.
1.8.2 Study Aims

1. Circulating monocytes
   a) To compare the transcriptional profiles of circulating monocytes coming from breast and endometrial cancer patients to healthy individuals as well as between cancer types.
   b) To identify a diagnostic signature and develop a classification model for stratification of cancer patients.
   c) To validate the robustness of the diagnostic signature on independent patient cohorts as well as datasets of monocytes from different cancer types.

2. Tumour-associated macrophages (TAMs)
   a) To analyze the transcriptional profiles of TAMs compared to resident macrophages in breast and endometrial cancer/tissue.
   b) To compare the gene expression profiles of TAMs between human and mouse models and to identify a set of candidate markers for identification of TAMs.
   c) To investigate the potential of TAMs as biomarkers for prognosis of breast cancer.
2 Human circulating monocytes as biomarkers for early detection of breast and endometrial cancer

2.1 Acknowledgments and contributions

I would like to acknowledge the contributions made to this chapter by Dr. Luca Cassetta and Hui Zhang, who carried out the isolation of monocytes and RNA extraction for sequencing as described in the materials and methods (section 2.4.1 and 2.4.2). I would also like to thank Dr. Luca Cassetta for selecting the genes in Figure 12C.

2.2 Introduction

Breast and endometrial cancer are two of the most commonly diagnosed cancers in women in the UK. There is increasing evidence that early detection of tumours improves survival rates. Mammography is the most accurate and non-invasive method to date for early detection of breast cancer, however its efficacy drops in women with higher breast density, resulting in overdiagnosis and potentially unnecessary interventions (Kolb et al. 2002). Currently there is no available screening method for early detection of endometrial cancer. Therefore, there is an urgent need for a non-invasive method for early detection of breast and endometrial cancer.

The advent of high-throughput microarray technologies has resulted in an increased interest in identification of multigene signatures for diagnosis and prognosis of cancer. Peripheral blood mononuclear cells (PBMCs) have been shown to be a promising source of diagnostic biomarkers in several cancer types, including breast cancer (Sharma et al. 2005; Aarøe et al. 2010a; LaBreche et al. 2011; Dumeaux et al. 2014). However, PBMCs represent a heterogeneous cell population, presenting with pro-tumoral and anti-tumoral functions. Hence, it might be more informative to study individual cell types. Monocytes are circulating blood cells that migrate to tissues and give rise to macrophages. Recent studies have revealed a pro-tumoral profile for
these cells in renal carcinoma and colorectal cancer (Chittezhath et al. 2014; Hamm et al. 2015).

Consequently, this study investigated the transcriptomes of circulating monocytes in breast and endometrial cancer in order to develop a blood-based, non-invasive diagnostic test for the detection of cancer. This test would act as a clinical tool for efficient screening of patients at risk.

2.3 Aims

- **Exploratory Analysis:** Determine the differences in the gene expression of circulating monocytes between breast and endometrial cancer patients, and healthy volunteers.

- **Identification of a diagnostic signature:** Development of a classifier for stratification of cancer patients.

- **Internal, external and negative validation:** Analyse additional cohorts of patients to evaluate the accuracy and robustness of the proposed classifier.
2.4 Materials and methods

2.4.1 Patient samples
Mononuclear cells were isolated from peripheral blood obtained from healthy female individuals through the New York Blood Centre or Cambridge Bioscience. Peripheral blood samples (20ml) from breast cancer patients and endometrial cancer were obtained from the Montefiore Medical Center, Bronx, NY, USA and the NHS, Edinburgh, Scotland. All the blood samples used in the study were collected into Venous Blood Collection Tubes containing EDTA and stored immediately at 4°C after collection.

Exclusion criteria of patients or controls at baseline included systemic metastatic disease, any inflammatory disorder, active infection or immunocompromised status not related to cancer. None of the patients recruited received chemotherapy or radiotherapy before collection, which was performed before surgery.

Pathologically, breast cancer patients had ductal carcinoma in situ (DCIS) or invasive breast cancers and endometrial cancer patients had Type I and Type II cancers.

2.4.2 Sample processing
Monocytes were isolated from healthy and cancer individuals using Fluorescent-activated cell sorting (FACS) with the following markers (CD45+/CD3-CD19-CD56-/HLA-DR*MHC II+/CD14+CD16*). Immediately after FACS sorting (carried out by Dr. Luca Cassetta), all samples were centrifuged at 450 RCF for 10 min at 4°C. Cell pellets were re-suspended in 350 ul of RLT lysis buffer and RNA extracted with RNAeasy Microkit (Qiagen) according to manufacturer’s instructions. RNA quantity was determined by QUBIT (Invitrogen); total RNA integrity was assessed by Agilent Bioanalyzer and the RNA Integrity Number (RIN) was calculated; samples that had a RIN > 7 were selected for RNA amplification and sequencing. RNA was amplified with Ovation RNAseq Amplification kit v2 (Nugen) according to manufacturer’s instructions; amplified RNA was sent to Albert Einstein Genomic Facility or BGI (Philadelphia). Standard Illumina un-stranded poly-A enriched libraries were prepared, fragmented and then multiplexed generating 100bp paired-end reads per
sample (HiSeq 2000 and 2500, Illumina). On average 30M reads were sequenced. Details of the number of reads generated and aligned to the reference genome, as well as information about the sequencing facility can be found in Section 6.4.

2.4.3 Sequence alignment and Quantification
FastQ files of 2x100bp paired-end reads were quality controlled using FASTQC (S. Kim et al. 2013). Samples were examined using the FASTQC interface for per base sequence quality as well as per sequence quality control. All samples but one were considered as good quality on the basis of a Phred score >=20 and were used for downstream processing. Quality controlled reads were then aligned to the GENCODE Human reference genome Release 19 (GRCh37.p13) using STAR aligner (Dobin et al. 2012) (version 2.3) for Linux Ubuntu, with default options as shown below and ran over 8 threads per sample (--GenomeDir GenomeDir/ --readsFilesIn /monocytes –runThreadN 8 –outFileNamePrefix MON). STAR is an ultra-fast spliced aligner tailored for RNA-seq data and was selected because is spliced aware and shows comparable performance to other aligners (Engström et al. 2013). For improved accuracy, the exon junction coordinates from the reference annotation were used. Furthermore, quantification of genes was performed using the count function of HTSeq (Anders et al. 2014). Reads with an alignment quality of less than 10 were discarded and reads were counted at the gene level.

2.4.4 Statistical analysis for differentially expressed genes
All statistical calculations have been performed in R programming language (version 3.2.3) (Chai & Wang 2015). The strategy is as follows: a) Filtered genes with count per million (CPM) reads > 1 in at least N samples (N number of the fewest replicates of a phenotype) were retained, b) Normalization; gene expression levels were normalized using reads per kilobase per million (RPKM) values and log₂ transformed using the cmp function from the EdgeR package in R (version 3.12) (Robinson et al. 2010) c) Batch effect correction; samples were corrected for batch effects using Combat function of the Surrogate Variable Analysis (SVA) package (version 3.18) (Cini et al. 2015). d) Differential expression analysis; Limma statistical software
was used to identify significantly differentially expressed transcripts with controlled False Positive Rate at 5% (FDR <0.05). Up-regulated genes were selected at a minimum log₂ fold change of 1.5 and down-regulated genes at a minimum log₂ fold change of -1.5. R package gplots (version 2.17) (Warnes et al. 2014) was used to plot heatmaps. Pearson correlation and complete linkage were used for hierarchical clustering.

The combination of RPKM and combat was used on the basis of the standard procedure for correction of batch effects as shown in the tool manual. Additionally, a variety of different tools were evaluated for differential expression analysis such as edgeR and DESEQ2, however these tools require raw reads as an input and pre-processing of counts for batch effect correction didn’t allow this.

### 2.4.5 Functional enrichment analysis and Pathway Analysis

DAVID database (Database for Annotation, Visualization and Integrated Discovery) (Huang et al. 2007) was used for Gene Ontology (GO) analysis using the list of previously identified significantly differentially expressed transcripts. Important biological processes were selected based on an FDR <= 0.05. Ingenuity Pathway Analysis (IPA) software was utilised for canonical pathway analysis.

### 2.4.6 Gene selection

Feature selection (FS) is one of the most important pre-processing steps in data mining. FS is used for the reduction of the feature space whilst maintaining maximum performance by eliminating redundant features. Feature selection can be broadly divided into (i) Filter methods, generic and independent of the classification algorithms (ii) Wrapper methods, consider the classification algorithm as a black box and use it in order to select the best subset. This approach is computationally intensive and the subset of features is very dependent on the classification algorithm, and (iii) Embedded methods, combine the feature selection step with the construction of the classifier. I used the following:

- Chi-square (x² statistic of independence) (Huan Liu & Setiono 1995). Chi-square is a filter based method; Chi-square score is used to assess the
independence of each gene in respect to a class. The gene expression of each
gene is first discretized into several intervals using an entropy-based
discretization approach. Genes with higher chi-square values are considered
more important. Chi.square function was utilised as implemented in the
FSselector package in R (Romanski 2009). The function returns the weight of
each attribute based on a chi-square test. The top 20 ranked features were
selected for classification and further assessment of classification
performance.

- **Recursive feature elimination (RFE):** RFE is a wrapper method that depends
  on a classification model in order to select the most informative features. The
  method starts with all the features; ranks them based on a classification
  model; at each iteration the classification model is re-fitted and eliminates
  features with low importance. The subset of genes that maximizes the area
  under the ROC curve is then selected. The rfe function with random forest
  (RF-RFE) and support vector machines (SVM-RFE) (GUYON et al. 2002)
  were used for classification as implemented in the caret (Kuhn 2015) package
  in R.

### 2.4.7 Machine learning methods

To analyse the data two supervised machine learning algorithms were assessed, (i)
Random forest (RF) and (ii) Support vector machines with a radial basis kernel
(SVM.radial) (Vapnik & Kotz 1982). To avoid over-fitting, optimization of the SVM
parameters was performed using an “internal” cross validation.

**Random forest, RF**

Random forest (RF) is an ensemble learning method used for classification. Random
forest constructs many decision trees (forest) during training and with the notion that
an ensemble predicts a class. Assuming N number of samples, the algorithm creates a
subset of the data using replacement of the N samples, and at each node, m number
of genes are selected from all the genes. The variable m that gives the best split is
selected and used for a binary split. This procedure is repeated for each node until the
tree is grown to terminal nodes k. The out-of-bag (OOB) samples are used to
evaluate the tree. The implementation of the caret package in \textit{R} was used \citep{Kuhn2015}. The number of constructed trees is denoted by the variable \textit{ntree} and the number of predictors at each node by the variable \textit{m}. Variable \textit{ntree} was kept constant at 1000 trees and variable \textit{mtry} was kept constant at $\sqrt{p}$, where \( p \) is the total number of genes.

\textbf{Support vector machines, SVMs}

SVM is a binary classification method that draws a hyperplane that best separates the classes by maximizing the margin between the classes. The training data were centred and scaled before the training of the model. SVMs were trained using the training set based on the implementation in the \textit{caret} package in \textit{R} \citep{Kuhn2015}. Optimization of the cost parameter \textit{C} used the following values 0.125, 0.25, 0.5, 1, 2, 4, 8 and the \textit{sigest} function for the \textit{kernlab} package in \textit{R} \citep{Karatzoglou2004} in order to calculate an estimation of the \textit{sigma} parameter. The optimization of the parameters on the SVM-RFE method was done using an “inner” 2-fold cross validation (due to high computational cost). For the $X^2$-SVM model 5 times 10-fold cross validation was used.

\subsection*{2.4.8 Cross validation and model evaluation}

The models described above were trained on the training set, which consisted of 59 samples (22 healthy, 21 breast and 16 endometrial) and then tested on the independently collected testing set, which consisted of 19 samples (5 healthy, 13 breast and 1 endometrial) (Table 1). To evaluate the performance of the classifier on the training set 5 times repeated 10-fold cross-validation was used and “inner” 2-fold cross validation (for SVM-RFE) was used to optimise the parameters. The model was evaluated using the following metrics; accuracy, sensitivity, specificity and area under the receiver-operating characteristic (AUROC). In short, the training samples were randomly partitioned into \( k \) (\( k=10 \) ) subsamples. Out of the \( k \) sub-samples, one is kept for testing the classifier, and the remaining \( k-1 \) subsamples are used for training the classifier. Then the whole process is repeated \( k \) times, with each of the \( k \) subsamples used as a testing set once. After the whole process, the model with the highest value of AUROC curve was selected as optimal. Overall accuracy, sensitivity
and specificity were calculated using the cross-validated predictions. The receiver-operating characteristic (ROC) curves were drawn using the ROCR package in R (Sing et al. 2005). The optimal classifier was evaluated on an independent cohort of 19 samples (5 healthy and 14 cancer) that had not been used for training or gene selection. To determine the accuracy rates of the classifiers and gene signatures that can be obtained by chance, performance of randomly extracted gene signatures and permutations of expression values in each sample were calculated using the R package SigCheck (Stark & Norden). \textit{sigCheck} function was run to carried out the baseline analysis based on the following parameters (classes="treatment",scoreMethod="classifier",signature=subset,annotation="ensemble_id", classifierMethod = randomForestI). Next, \textit{sigCheckRandom} function was used to compare the performance of the primary signature to the performance of signatures composed of the same number of randomly-selected features based on 1000 iterations. The empirical p-value is calculated based on the percentile rank of the performance of the primary signature compared to a null distribution of the performance of the random signatures. Finally, \textit{sigCheckPermutated} function was used to compare the performance of the primary signature on the original data to its performance on permuted data by samples (expression values of all features permuted within each sample). The following parameters were used (toPermute="samples", check = check, iterations = 1000).

2.4.9 Comparison with publically available datasets

To validate the performance of the diagnostic signature, five publically available datasets were downloaded from the Gene Expression Omnibus (GEO) database.

1) **Negative control (RNAseq):** Lyme disease dataset (GSE63085) from (Bouquet et al. 2016). Samples that were diagnosed as Lyme disease patients (n=28) and healthy controls (n=13) were selected. FPKM (Fragments Per Kilobase per Million reads) normalized values were further normalized using voom for the statistical package Limma (Ritchie et al. 2015) in R.

2) **Renal carcinoma (Microarray):** (GSE38424) from (Chittezhath et al. 2014) consisting of 8 samples (4 normal and 4 cancer patients).
3) **Colorectal cancer (Microarray):** (GSE47756) from (Hamm et al. 2015) consisting of 93 samples (38 normal and 55 colorectal cancer patients).

4) **Breast cancer PBMCs (Microarray):** (GSE16443) from (Aarøe et al. 2010b) consisting of 130 samples (63 normal and 67 breast cancer samples).

5) **Breast cancer PBMCs (Microarray):** (GSE27562) from (LaBreche et al. 2011). Healthy samples and malignant samples that had not received any treatment resulting in 88 samples (31 normal and 57 breast cancer samples) were selected.

Microarray raw data were log$_2$ transformed, quantile normalized and filtered for probes without annotation using Limma (Ritchie et al. 2015). Duplicated probes were collapsed to the average expression of each gene.
Table 1. Clinical characteristics of samples used in the study. (F: female, M: male, TN: triple negative, ND: not done, N/A: not applicable). Healthy samples from the independent validation set were all collected from the blood bank. No further information was available for healthy samples for either cohort. Clinical information was not available for two patients from the breast cancer cohort (training). Note that HER2-positive patients were also ER-positive and PR-positive.

<table>
<thead>
<tr>
<th>Sample Group</th>
<th>Training</th>
<th>Independent validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Breast cancer N=21</td>
<td>Endometrial cancer N=16</td>
</tr>
<tr>
<td>Age (Median)</td>
<td>62</td>
<td>68</td>
</tr>
<tr>
<td>Gender (F/M)</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Breast cancer ER-positive</td>
<td>13/21</td>
<td>ND</td>
</tr>
<tr>
<td>Breast cancer PR-positive</td>
<td>13/21</td>
<td>ND</td>
</tr>
<tr>
<td>Breast cancer HER2-positive</td>
<td>1/21</td>
<td>ND</td>
</tr>
<tr>
<td>Breast cancer TN</td>
<td>6/21</td>
<td>ND</td>
</tr>
<tr>
<td>Grade (I/II/III)</td>
<td>N/A</td>
<td>6/0/10</td>
</tr>
<tr>
<td>Endometrial cancer Type (I or II)</td>
<td>N/A</td>
<td>Type I/II (8/8)</td>
</tr>
</tbody>
</table>
### 2.5 Results

#### 2.5.1 Monocytes exhibit significantly altered gene expression profiles in breast and endometrial cancer patients compared to those of healthy controls

Samples were collected and blood monocytes were isolated from healthy women (n=22) and untreated women with breast (n=21) and endometrial cancer (n=16) (Table 1). Monocyte RNA was isolated and evaluated for quality and quantity and used for mRNA amplification and sequencing (Material and Methods section 2.4). Figure 5 depicts the monocyte sequencing and analysis workflow, respectively.

**Figure 5. Analysis workflow of RNA-seq samples.** (A) Sample collection and processing workflow. 20 ml of blood was collected from healthy individuals and cancer patients. FACS was used to sort the monocytes. RNA was isolated from samples and sent for sequencing. (B) Bioinformatics analysis pipeline. Raw reads were quality controlled and aligned to the reference genome. Aligned reads were quantified at the gene level, normalized and assessed.
for differential expression between conditions. Differentially expressed transcripts were used for GO and pathway analysis.

Raw reads were aligned to the reference genome using a splice-aware aligner and quantified to the gene level. Samples were normalised and filtered for genes with low expression or no expression. Variability in sequencing depth as well as different gene length can result in skewed results, for that reason gene expression counts were normalised to account for this variability. Figure 6 shows the normalised gene counts of monocytes from breast and endometrial cancer as well as healthy donors; the second healthy sample shows high variability and it could be considered as an outlier, however it was kept for further analysis due to small sample size.

**Figure 6. Distribution of CPM log2 normalized gene counts.** Raw gene counts were quantified using HT-Seq. Gene counts were normalised for library size and log2 transformed. Healthy samples are shown in blue, breast cancer samples are shown in red and endometrial cancer samples shown in purple. Y-axis represents the log2 normalised counts per sample.

Unsupervised multidimensional scaling analysis (MDS) of 15,455 transcripts showed distinct clusters of breast and endometrial cancer samples compared to healthy donors, as shown in Figure 7A and Figure 7B, respectively. A linear model was fitted to identify differentially expressed transcripts between monocytes from breast cancer and endometrial samples compared to healthy individuals.
Comparison of the transcriptional profiles of monocytes between breast and healthy individuals identified 5,449 transcripts (out of 15,455) significantly differentially regulated between breast cancer and healthy samples with controlled false positive rate at 5% (FDR <= 0.05). Stringent criteria of log$_2$ fold change greater or less than 1.5/-1.5 resulted in 2,144 transcripts up- and down-regulated (1792 up, 352 down). Hierarchical clustering of the top 50 most significantly differentially expressed transcripts (Figure 8A) clearly separated the monocytes from breast cancer patients from those from healthy volunteers, although 5 samples from healthy donors clustered with those from breast cancer patients. The presence of outliers is likely due to high variation of the human genome but also due to high sensitivity of circulating monocytes. Additionally, it is very likely that some of these outliers are a result of low sequencing quality and low number of mapped reads to the reference genome due to contamination. Figure 6 depicts high variability in the distribution of normalised expression in some samples from healthy patients. Differentially expressed transcripts were subjected to IPA functional enrichment (Figure 8B). “Cancer” was reported as the most significant disease enriched in breast cancer monocytes, followed by “reproductive system disease”, “inflammatory response” and “cell to cell signalling”. Canonical pathway analysis showed predicted activation
of “pro-inflammatory” and “immune response” signalling pathways (Figure 8C). More specifically, “immune response IL-1” signalling pathway was predicted significantly activated through up-regulation of kinases such as mitogen-activated protein 6 (“MAP2K6”), “MAP2K10”, protein kinase cAMP-activated catalytic subunit beta (“PRKACB”) and “PRKAR1A”. “Toll-like receptor signalling (TLR)” was predicted significantly up-regulated through up-regulation of transmembrane receptors such as “TLR7”, “TLR5” and “LY86”.

Figure 8. Hierarchical clustering and enriched pathways of significant transcripts expressed in monocytes in breast cancer. (A) Gene expression heatmap top 50 most significant transcripts between monocytes coming from breast cancer (Red) and healthy samples (Blue); samples are clustered using complete linkage and Pearson correlation.
Samples are arranged horizontally; genes are arranged vertically; Red color in the heatmap indicates up-regulation, and green color indicates down-regulation based on the raw z-score (range [-2, 2]). (B) Top diseases and functions, (C) and top canonical pathways predicted to be involved in breast cancer, pathways, and transcripts ranked by the negative log of the P value of the enrichment score. The color scheme for panel C is based on Z scores, with activation in red, and undetermined directionality in gray. (PPRs: Pattern Recognition Receptors, TLR: Toll-Like Receptors, ATM: Ataxia Telangiectasia Mutated, TREM1: Triggering Receptor Expressed on Myeloid cells 1, ERK5: Extracellular signal Regulated Kinase 5, cAMP: Cyclic adenosine 3’5’-Monophosphate)

Comparison of transcriptional profiles of monocytes between endometrial cancer patients to those of healthy individuals identified 5,887 transcripts (out of 15,455) significantly differentially expressed with controlled false positive rate at 5% (FDR <= 0.05). Out of those, 2,722 transcripts were significantly up- and down-regulated using stringent criteria of log2 fold change greater or less than 1.5/1.5 (2,429 up, 223 down). Differential expression analysis of 13,874 transcripts between the different subtypes of endometrial cancer (TI-TII) did not show any significant differences at FDR <= 0.05. Like the breast cancer monocytes, hierarchical clustering on the top 50 most significant transcripts showed a clear separation between monocytes coming from endometrial cancer compared to those of healthy patients, and an up-regulation in gene expression compared to healthy controls; although 5 samples from healthy donors clustered with those from endometrial cancer patients (Figure 9A). Functional enrichment analysis of differentially expressed transcripts showed enriched functions associated with “cancer”, “inflammatory responses” and “cell signalling” (Figure 9B). Canonical pathway analysis revealed an activation of the “cAMP-mediated signalling” pathways as well as the “role of pattern recognition receptors (PPRs)” through the activation of a variety of pattern recognition receptors (“TLR7”, “TL7”, “TLR8”), G-proteins (“LTB4R”, “LPARI”, “GRM7”, “FR2”, “CXCR2”) and kinases (“PRKACB”, “PRKAR1A”, “PRKAR1B”, “PRKAR2B”). Interestingly enough, the “Granzyme B signalling” pathways that is associated with apoptosis and cell death was predicted to be significantly up-regulated through activation of “CASP3”, “APAF1” and “PARP1” (Figure 9C).
Figure 9. Hierarchical clustering and enriched pathways of significant transcripts expressed in monocytes in endometrial cancer. (A) Gene expression heatmap of top 50 most significant transcripts between monocytes coming from endometrial cancer (Purple) and healthy samples (Blue); samples are clustered using complete linkage and Pearson correlation. Samples are arranged horizontally; Genes are arranged vertically; Red color in the heatmap indicates up-regulation, and green color indicates down regulation based on the raw z-score (range [-2, 2]). (B) Top diseases and functions, (C) and top canonical pathways predicted to be involved in endometrial cancer, pathways, and transcripts ranked by the negative log of the $P$ value of the enrichment score. The color scheme for panel C is based on $Z$ scores, with activation in red, and undetermined directionality in grey.

Out of 15,455 expressed transcripts, only 59 were significantly differentially expressed between breast and endometrial monocytes with controlled false positive rate at 5% (FDR $\leq 0.05$), and out of those, only 35 transcripts were significantly
changed using stringent criteria of log₂ fold change greater or less than 1.5/-1.5 (0 up, 35 down) (Figure 10).

Figure 10. Differential expression analysis of purified monocytes showed little differences between breast cancer and endometrial patients. (A) MDS plot did not show distinct clusters of monocytes coming from breast (red) and endometrial cancer patients (purple). Volcano plot of 15,455 transcripts showing the gene expression profiles of monocytes from breast cancer compared to endometrial cancer samples. Log₂FC is plotted against the x-axis and the p-value for statistical difference on the y-axis. Points shown in red are transcripts that are not significantly expressed. Points shown in cyan below 0 are transcripts significantly decreased in monocytes from breast cancer compared to endometrial cancer. Points shown in cyan higher than 0 are transcripts significantly increased in monocytes from breast cancer compared to endometrial cancer.

A Venn diagram showed that 71% of differentially expressed transcripts from monocytes in breast cancer, and 56% of differentially expressed transcripts in endometrial cancer (compared to healthy controls) were commonly regulated between the two cancers (1,357 up, 169 down) (Figure 11). This may be indicative that the presence of cancer induces similar changes in circulating monocyte phenotypes in breast and endometrial cancer. Thus, I named the circulating monocytes from cancer patients as Tumour-Educated Monocytes (TEMo).
Figure 11. Venn diagram of commonly up-regulated and down-regulated transcripts between monocytes from breast cancer and endometrial cancer compared to healthy patients, and monocytes from breast cancer compared to endometrial cancer. Numbers in red indicate up-regulated genes, numbers in green down-regulated genes. There was a big overlap between breast and endometrial samples compared to control (FDR <0.05, log_2 fold change greater or less than 1.5/-1.5).

2.5.2 Distinct gene expression profiles of Tumour-Educated Monocytes (TEMo) and normal monocytes in cancer

In order to understand the role of TEMo (monocytes from cancer patients) in cancer progression I compared their gene expression profiles, regardless of cancer type, against monocytes from healthy donors (normal monocytes). Unsupervised MDS and hierarchical clustering clearly separated the gene expression profiles of monocytes from healthy controls compared to those of cancer patients, suggesting the populations to be transcriptionally distinct (Figure 12A and Figure 12B). Differential expression analysis identified 2,169 transcripts (out of 14,543) significantly altered...
in TEMo compared to monocytes coming from healthy patients (1,946 up and 223 down; log₂ fold change greater or less than 1.5/-1.5, FDR <= 0.05).

Figure 12. Transcriptome analysis of monocytes and TEMo. (A) MDS plot of 14,543 transcripts showed two distinct clusters of monocytes coming from cancer (green) and healthy patients (blue). (B) Gene expression heatmap of top 50 most significant genes between monocytes coming from cancer (green) and healthy samples (blue); samples are clustered using complete linkage and Pearson correlation. Samples are arranged horizontally; Genes are arranged vertically; Red color in the heatmap indicated up-regulation, and green color indicated down regulation based on the raw z-score (range [-3, 3]). (C) Bar plot of transcripts associated with macrophage biology coloured by type (red = Adhesion molecules, orange = Kinases, cyan = Transcription factors, blue = Soluble factors and green = Transmembrane receptors) found significantly expressed between cancer patients and
healthy donors (FDR <= 0.05). X-axis represents the logarithm of fold change (Log\(_2\)FC) of monocytes between healthy individuals and patients with cancer.

In order to better characterise the transcriptional profiles of TEMos, the expression of transmembrane receptors, soluble factors, transcription factors and enzymes were investigated (FDR < 0.05, log\(_2\)FC greater or less 1.5/-1.5) (Figure 12C). A series of pro-inflammatory chemokines, “CCR2”, “CCR5” and “CX3CR1”, important for monocyte recruitment (Weber et al. 2000) were found to be up-regulated. Additionally, up-regulated molecules included toll-like receptors, “TLR5” and “TLR7” that play a role in innate immune system regulation and “CD200R1”, important regulator of the adaptive immune system. C-type lectins “CLEC1A”, “CLEC1B” and “CLEC4A”, and the immunoglobulin-like lectins “SIGLEC9” and “SIGLEC16”, important regulators of the immune system were also found to be up-regulated (Qian et al. 2011; Zhang et al. 2000; Macauley et al. 2014). “FCGr3a (CD16)” an Fc receptor mainly expressed by non-classical monocytes (Ancuta et al. 2009) and “CD163L1” receptor expressed by tissue macrophages (González-Domínguez et al. 2015), both had elevated expression in TEMo compared to healthy patients. Interestingly, Angiopoietin (ANG) 1 and 2, known regulators of angiogenesis and inflammation, along with the Hepatocyte Growth Factor (HGF) were likewise significantly up-regulated in TEMo. ANG1 has been previously described to have a role in stimulation of pro-inflammatory responses in monocytes (Seok et al. 2013). On the contrary, another pro-angiogenic factor “VEGF-A” was predicted to be down-regulated in TEMo compared to normal monocytes. Interestingly, the death ligand “TNFSF10” (TRAIL) was also up-regulated in TEMo.
Figure 13. Functional enrichment analysis of significantly regulated transcripts in TEMo. (A) Top canonical pathways, (B), and top disease and functions predicted to be involved in breast cancer TEMos. Pathways and transcripts ranked by the negative log of the $P$ value of the enrichment score. The color scheme for panel B is based on $Z$ scores, with activation in red, and undetermined directionality in grey.

Differentially regulated transcripts in TEMo were analysed using IPA in order to identify enriched pathways. Functional enrichment analysis reported “cancer” as the most significant disease enriched in TEMo, followed by “reproductive system disease”, “cell to cell signalling”, “cellular movement and immunological disease” (Figure 13A). Canonical pathways analysis identified immune-related pathways enriched in the TEMos, including “Pattern recognition receptors”, “TREM1 signalling”, “Tissue Factors in cancer”, “Toll-like receptor signalling”, “IL-1 signalling” and “CD27 signalling in lymphocytes” (Figure 13B). Granzyme B signalling pathway was predicted as being up-regulated through up-regulation of “APAF1”, “CASP3”, “LMNB1”, and “PARP1”. This is quite surprising since Granzyme B is a serine protease produced by cytotoxic lymphocytes, NK cells and cytotoxic T cells. Therefore, this discrepancy might be due to potential contamination of samples that seem to be outliers in the analysis or an artefact of the IPA analysis software.
Taken together, the transcriptome data of TEMos from breast and endometrial cancers indicated a distinct gene expression profile from normal monocytes suggesting an altered function in the presence of malignancy. For that reason, I sought to investigate the validity of TEMos as biomarkers for early detection of cancer.

2.5.3 TEMos as biomarkers for detection of cancer

2.5.3.1 Identification of a TEMo-derived signature for detection of cancer

The analyses performed thus far indicated that a blood-based diagnostic signature might be contained within the TEMos transcriptome that could stratify cancer patients from healthy individuals. Clinical characteristics of the study population are summarised in Table 1. The study included 22 samples from healthy donors (normal monocytes) and 37 cancer samples from patients at the time of diagnosis (TEMo). I determined the diagnostic accuracy of TEMos in a training set (n=59, 22 healthy and 37 cancer) by evaluating two supervised machine leaning algorithms, namely random forest (RF) and support vector machines (SVM). Two feature selection methods were evaluated for the selection of informative features, namely recursive feature elimination (RFE) and chi-square statistic ($X^2$). Feature selection is an important pre-processing step aiming to identify a minimum number of non-redundant sets of genes able to discriminate between classes. The experimental design of the development of the classifier is shown in Figure 14.
Figure 14. Workflow for the identification and validation of the TEMo-derived signature. Internal validation and feature selection was conducted on 59 samples using 5 times 10-fold cross-validation. Different methods (RF and SVM) were used with different feature selection methods (RFE and $X^2$). The best performing method was selected as the optimized classifier. Independent validation of the classifier was conducted in a dataset of 19 samples from patients with breast and endometrial cancer. Negative validation was conducted on a dataset from PBMCs coming from patient with Lyme disease. The classifier was validated on two publicly available datasets from purified monocytes coming from renal carcinoma (n=8) and colorectal cancer (n=93).

Classifiers were optimised using 5 times 10-fold cross-validation on the training set and the best performing method was selected. Prior to optimization, $X^2$ was used to rank the list of expressed genes based on their $X^2$ score; out of those the 20 most important genes were selected as input genes for the classification methods. RFE feature selection ranked the features within each step of the cross validation resulting in a ranked list of the most important genes. The performances of the generated classification models were compared on the basis of: accuracy, positive predictive value (PPV), negative predictive value (NPV) and area under the curve (AUC).
Table 2. Performance of evaluated machine learning algorithms using both $X^2$ and RFE feature selection. The method selected (RF-$X^2$) is highlighted in bold.

Table 2 depicts the achieved classification performances of the machine learning algorithms on the internal validation set. Based on accuracy and AUC, random forest with chi-square (RF-$X^2$) feature selection was the method that performed best; out of the top 20 transcripts with the highest $X^2$ score, 13 transcripts showed the highest performance (Figure 15A, Section 6.2); yielding an accuracy of 94%, a PPV of 92%, a NPV of 97% and an AUC of 98% in the training set (Figure 15B and Figure 15C).

Overall, all methods evaluated showed high accuracy suggesting that monocytes are very different between cancer and healthy donors and could be used as biomarkers for patient stratification.
Figure 15. **Performance of RF-\(X^2\) classifier in internal and independent validation.** (A) Dot plot of the performance of top 20 genes as ranked by the \(X^2\) statistic. X-axis shows the number of genes used and y-axis their performance. 13 genes gave the best performance. (B) ROC curve showing the performance in the internal validation set (AUC = 98%) achieved from the 13 genes signature in blue. The performance of the independent validation set is shown in red (AUC = 100%) (C) Table of performance of the RF-\(X^2\) classifier internally using 5-times 10-fold cross validation on 59 samples as well as in the independent validation set.

To further test the uniqueness and specificity of the classifier and its ability to discriminate between cancer and healthy samples I extracted 1000 randomly selected gene signatures comprising 13 genes to build a RF classifier and tested its performance using a leave-one-out cross validation (LOOCV). The 13-gene signature showed the highest accuracy compared to the random signatures that yielded a mean accuracy of 80% (SD ±4%, \(p<0.001\)) ((Figure 16A). Furthermore,
random classifiers as determined by 1000 rounds of random permutations of the expression values for each sample, had no predictive power (mean accuracy: 61%, SD: 6%, p <0.001) (Figure 166B).

2.5.3.2 Confirmation of gene signature in an independent validation set

To further validate the robustness of the RF-X^2 classifier to distinguish between cancer and healthy individuals, I utilised an independently collected dataset of 19 samples, 5 normal monocytes from healthy patients and 14 TEMo from patients at the time of diagnosis (13 breast cancer, 1 endometrial cancer) Table 1 The independent validation dataset was not used during training or during feature selection. The optimised 13-gene RF-X^2 classifier yielded an accuracy of 100%, a sensitivity of 100%, a specificity of 100% and an area under the curve 100% Figure 15B and Figure 15C). Surprisingly, the performance of the classifier was higher than the internal validation set, however this is likely due to the nature of the independent validation samples. The internal validation set represents a population of healthy and benign patients, however in the independent cohort the healthy samples (n=5) are coming from patients with no sign of the disease, and the cancer samples (n=14) represent patients with invasive disease; these samples represent two very transcriptionally different cohorts, therefore is somewhat expected that the classifier would perform very well.
Check: Random Signatures

Percentile: 1.00 (Tests: 1000  p<0.001)
N = 1000   Bandwidth = 0.01118
Density
Signature [0.949]  Mode [0.627]

Check: Permuted Data: [samples]

Percentile: 1.00 (Tests: 1000  p<0.001)
N = 1000   Bandwidth = 0.01144
Density
Signature [0.949]  Mode [0.627]
Figure 166. Performance of random signatures and random classifiers (A) Background distribution based on the performance of the random signatures. Signatures were tested using a leave-one-out cross validation (LOOCV). Solid vertical red line represents the performance of the TEMo 13-gene signature; dotted vertical red line represents the performance of the mode classifier. (B) Background distribution based on classification performance of the 13-gene signature on permutated expression values for each sample.

2.5.3.3 TEMo signature detects breast cancer at an early stage

Within the 19 samples of the independent validation set, 4 patients have been diagnosed with DCIS. Several studies have carried out gene expression profiling of DCIS, however the diagnosis and monitoring of DCIS patients remains very challenging (Hannemann et al. 2006; Adeyinka et al. 2002; Seth et al. 2003). A careful look at the results of the classification revealed that all samples from patients with DCIS were correctly classified as having cancer suggesting that TEMos can identify the presence of pre-malignant disease even if the cancer cells are non-invasive and still confined within the basement membrane. In order to further investigate the transcriptional profiles of monocytes in DCIS patients, I used an MDS plot using only the 13-gene predictor on samples coming from invasive breast cancer, DCIS and healthy patients (n = 52, 27 healthy, 21 breast cancer and 4 DCIS) (Figure 17A).
Figure 17. TEMo can detect the presence of early stage breast cancer. (A) MDS plot on all breast cancer samples (internal and independent) using the 13-gene predictor highlighting the DCIS samples within the cancer class. (B) Venn diagram of commonly differentially expressed transcripts expressed in monocytes between invasive and DCIS samples compared to those of healthy patients (Log$_2$FC greater or less than 1.5/-1.5, FDR <= 0.05).

The MDS plot clearly separated the invasive breast cancer patients from healthy individuals, with all the DCIS samples included in the breast cancer cluster. Differential expression analysis between TEMos coming from invasive breast cancer compared to those of DCIS patients showed no significant differences (FDR < 0.05). I next looked at the commonly regulated transcripts (Log$_2$FC greater or less than 1.5/-1.5, FDR <= 0.05) between invasive and DCIS patients compared to healthy
patients. It is clear that TEMos from DCIS patients are very different from monocytes coming from healthy individuals. The Venn diagram in Figure 17B, showed that 1536 transcripts (1247 up, 289 down) were shared between the two groups; more specifically 76% of transcripts from invasive breast cancer and 56% of transcripts from DCIS were in common. Therefore, TEMos from DCIS and invasive breast cancer patients are very similar compared to healthy donors.
These data indicate that circulating TEMos are influenced by the formation and growth of tumours from an early stage.

2.5.3.4 TEMo signature is specific to cancer
To establish whether X²-RF classifier was specific to cancer, I would have ideally compared the signature against datasets of purified monocytes from normal and individuals with other, non-cancer afflictions. However, to the best of my knowledge, there is no such dataset available from purified monocytes in the literature; thus, I compared against a dataset (RNA-seq) from PBMCs isolated from individuals with Lyme disease (Bouquet et al. 2016), an infectious disease caused by bacteria of the Borrelia type. This dataset comprises 29 Lyme disease patients and 13 healthy individuals. Out of the 13 gene predictors selected, 12 were also represented in the Lyme dataset (excl. “RP11-469M7.1”). X²-RF classifier successfully classified all Lyme disease patients as normal, and all normal patients as normal thus indicating cancer specificity for the signature (Table 3).
Table 3. Table of performance of the RF-X$^2$ classifier on a negative control dataset. The dataset of Bouquet et al. from isolated PBMCs from Lyme disease was used as a negative control. The dataset comprises 29 Lyme disease patients and 13 healthy individuals (Bouquet et al. 2016).

2.5.3.5 TEMo as a pan-cancer diagnostic signature

Based on previous results presented in section 2.5.3.4 suggesting that the TEMo signature is specific to cancer, I sought to investigate its validity on other types of cancer. To that end, I compared against the only two available microarray datasets from isolated monocytes to-date, (1) renal carcinoma (Chittezhath et al. 2014) (4 healthy and 4 cancer) and (2) colorectal cancer (Hamm et al. 2015) (38 healthy and 53 cancer).

Out of the 13 gene predictors included in the TEMo signature, 10 were also represented in the renal carcinoma microarray (excl. “TMTC2”, “ZNF114”, “RP11-469M7.1”). Data were quantile normalised; applying the previously established RF-X$^2$ classifier it was impossible to separate the cancer patients from the healthy individuals using the random cut-off (50%). However, using ROC curves I identified an optimal cut-off of 0.824 for predicting healthy patients, yielding an accuracy of 87.5%, a sensitivity of 100%, a specificity of 75% and an AUC of 87.5% (Figure 18A and Figure 18B).
Figure 18. TEMo signature shows potential for pan-cancer diagnosis. (A) Table of performance of RF-$X^2$ classifier on renal carcinoma and colorectal cancer patients. (B) ROC curve analysis of the optimized classifier RF-$X^2$ on external datasets from purified monocytes from renal carcinoma and colorectal cancer patients.

For the colorectal cancer dataset, 12 out of the 13 gene predictors were represented in the microarray data (excl. “RP11-469M7.1”). Applying the RF-$X^2$ classifier using the random cut-off (50%) yielded an accuracy of 52%, a sensitivity of 34% and a
specificity of 78%. However, using ROC curves I identified an optimal cut-off of 0.534 for predicting healthy patients, yielding an accuracy of 63.4%, a sensitivity of 76.7%, a specificity of 43.3% and an AUC of 60% (Figure 18A and Figure 18B).

Overall, the TEMo signature showed a high predictive performance in the renal carcinoma dataset and a modest performance in colorectal cancer dataset. This indicates a potential of a pan-cancer diagnostic test based on circulating monocytes/TEMos.

2.5.3.6 TEMo signature is specific to monocytes in comparison with PBMCs

To assess if the TEMo-derived signature is specific to monocytes or whether it can classify patients based upon the whole PBMC population I compared it against two publically available datasets. Initially I compared the TEMo signature against the dataset from Aarøe et al. consisting of 130 samples (63 healthy individuals and 67 breast cancer patients) (Aarøe et al. 2010b). Only 5 out of the 13 gene predictors were present in the microarray (“PPIF”, “PTP4A1”, “CEPT1”, “IFI16”, “MCTP1”). It should be noted that re-analysis of the Aarøe dataset did not show any significantly differentially expressed genes in PBMCs between healthy and breast cancer individuals at FDR <= 0.05. Consistently with this, the optimised RF-\(X^2\) classifier failed to separate between healthy and cancer patients (Figure 19A). Additional analysis, using the 5 genes found on the Aarøe dataset on the TEMo samples resulted in an accuracy of 91%, sensitivity of 95 %, specificity 85%, PPV of 91% and NPV if 90%.
Figure 19. TEMo signature is specific to monocytes. (A) Table of performance of the RF-$X^2$ classifier on 2 datasets from isolated PBMCs from breast cancer (Aarøe et al. 2010a; LaBreche et al. 2011). (B) ROC curve analysis of the optimized classifier RF-$X^2$ on breast cancer PBMC datasets.

Next, I compared against the dataset from LaBreche et al. consisting of 162 samples (LaBreche et al. 2011); out of those healthy individuals and patients that had not received any treatment were selected, resulting in 88 samples (31 normal and 57 breast cancer samples). Twelve out of the 13 transcript predictors (excl. the long non-

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer – PBMCs</th>
<th>Breast cancer – PBMCs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Aarøe et al)</td>
<td>(LaBreche et al)</td>
</tr>
<tr>
<td></td>
<td>(63 normal, 67 cancer)</td>
<td>(31 normal, 57 cancer)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>49%</td>
<td>35%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Specificity</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Positive Predictive Value (PPV)</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Negative Predictive Value (NPV)</td>
<td>49%</td>
<td>35%</td>
</tr>
<tr>
<td>Area under the curve (AUC)</td>
<td>50%</td>
<td>45%</td>
</tr>
</tbody>
</table>

Random Forest & chi-square

Accuracy
Sensitivity
Specificity
Positive Predictive Value (PPV)
Negative Predictive Value (NPV)
Area under the curve (AUC)
coding RNA “RP11-469M7.1”) were represented in the microarray. Similarly to the Aarøe PBMC dataset the classifier had no predictive power (Figure 19A and B).

All together, in this section it was demonstrated that the performance of the TEMo signature is limited on PBMC samples, suggesting its specificity to monocytes.

### 2.6 Discussion

The development of a blood-based test for detection of cancer will potentially enable not only early detection, but also easy monitoring of the disease during therapy. This study presented compelling data to support the hypothesis that circulating monocytes respond to the presence of the tumour and could be used as a non-invasive diagnostic method. Using genome-wide transcriptional analysis, this study found that the monocyte transcriptomes are altered in both breast and endometrial cancer patients and show distinct profiles compared to monocytes from healthy individuals (Figure 7, Figure 8, and Figure 9). Comparison of monocytes coming from breast and endometrial patients revealed very similar transcriptional profiles (Figure 10, and Figure 11). Thus this study pooled the monocytes from breast and endometrial cancer together and named them Tumour-Educated Monocytes (TEMo). Differential expression analysis between TEMo and monocytes from healthy donors revealed that the two populations are distinct at the transcript level (Figure 12).

The immune system recognises and responds to the presence of pathogens mainly through TLRs found on the cell surface of leukocytes such as monocytes (Medzhitov et al. 1997). Consistently, TEMo showed a significant up-regulation of the PRRs and TLR pathways through up-regulation of “TLR5”, “TLR7”, “TLR8”, and “TRL10” (Figure 13). Chittezhath et al. (Chittezhath et al. 2014) was the first to propose a tumour promoting phenotype of circulating monocytes in renal carcinoma through an IL1R-dependent mechanism. Indeed, by targeting IL1-IL1R in a human xenograft model they were able to prevent this pro-tumoral phenotype leading to a reduction in tumour growth. Consistently, IPA pathway analysis predicted an up-regulation of the IL1 pathway in TEMo through up-regulation of related kinases (“MAPK10”,

73
“MAP2K6”, “PRKAR1A”, “PRKACB”, “IKBKB” and “IRAK4”). Similarly to the Chittezhath study, the TEMo transcriptional profile was characterized by an up-regulation of pro-inflammatory chemokines (“CCRI”, “CCR2”, “CCR3”, “CCR5”) and transmembrane sialic acid receptors (“SIGLEC7”, “SIGLEC9”, “SIGLEC16”). These results give the first indication of an activated profile of TEMo in the presence of malignancy, however the mechanisms underlying their specific functions remain unclear. Furthermore, significant differences in the transcriptional profiles of monocytes from patients with early stage breast cancer (DCIS) to those with no signs of invasive disease indicate a potential for monocyte recognition of early stage tumours (Figure 17).

This work also reported a robust classification model based on a 13-gene TEMo derived signature that can accurately discriminate healthy individuals from cancer (breast and endometrial) patients. Previous work has described the use of gene expression profiles of PBMCs as an easily accessible source for detection of various cancer types, including breast cancer (Braakhuis et al. 2013; Nichita et al. 2014b; Aarøe et al. 2010a; Dumeaux et al. 2014; Sharma et al. 2005; Osman 2006; Twine et al. 2003). More recently, Hamm et al. (Hamm et al. 2015) proposed a 23 gene monocyte-derived signature for diagnosis and follow-up of colorectal cancer. The findings in this study support the use of a blood-based diagnostic method, however the use of TEMos might offer a superior method, with higher accuracy. In this work a set of 13 genes specifically expressed by TEMos coupled with a random forest classifier was able to detect the presence of cancer in an internal validation cohort with an accuracy (94%) and specificity (86%) that is higher than that of mammography (Figure 15). Independent validation of the model on a cohort of patients that was not used during training or feature selection showed 100% accuracy (Figure 15). This is a somewhat surprising result however this is possibly due to the sample population included in the independent validation set. In the internal validation set monocytes were derived from both healthy and benign patients as well as patients with invasive disease, however in the independent cohort monocytes were derived from patients with no sign of the disease (n=5), and from patients with invasive disease. Hence, the independent validation cohort included two very transcriptionally different cohorts. For that reason, additional independent validation
is needed on a larger patient population. Further validation of the specificity of the classifier on a non-cancer negative control dataset from Lyme disease patients showed that the TEMo-signature is specific to cancer (Table 3). Prediction on breast cancer datasets from PBMCs showed very limited performance suggesting specificity to TEMos (Figure 19). However, the observed low performance of the signature can be also due to cross-platform transfer and annotation limitations. For example, in the Aarøe study, only 5 out of the 13-gene predictors were found, which certainly limits the performance of the model. Although, a recent study by Fumagalli et al showed a good correlation of prognostic signatures derived from microarrays to RNA-seq platform (Fumagalli et al. 2014), pre-processing plays an important role in cross-platform machine-learning applications (Thompson et al. 2016). For that reason, the most appropriate comparison would be against purified monocytes or PBMCs RNA-seq datasets; which to the best of my knowledge are not available thus far. Furthermore, the signature showed high predictive performance in renal carcinoma patients and intermediate performance on colorectal cancer patients suggesting the potential of a pan-cancer diagnostic test based on circulating monocytes (Figure 18).

The proposed monocyte-derived signature comprises of 13 transcripts, some of which have been previously described to be associated with cancer, however their functions in relation to monocytes are still unclear (Appendix 5.1). “MCTP1” (multiple C2 and transmembrane domain containing 1) is located in the membrane of the cell and functions through the calcium mediated signalling. “PIBF1” (progesterone immunomodulatory binding factor 1) is located in the nucleus and plays a key role in pregnancy (Check et al. 1996). In breast cancer it has been reported that “PIBF1” is expressed in higher levels by tumour cells (Lachmann et al. 2004). “TMTC2” (transmembrane and tetratricopeptide repeat containing 2) is a protein-coding gene, located in the cytoplasm with functions involved in calcium ion homeostasis. “SLFI1” (SMC-5-SMC6 complex localisation factor 1) is located in the nucleus; and it regulates DNA damage and DNA repair functions. “CEPT1” (choline/ethanolamine phosphotranferase 1) is found in the cytoplasm and is associated with lipid metabolic processes. “ZNF114” (zinc finger protein 114) is involved in transcriptional regulation. “CRYBG3” (crystalline beta-gamma domain
containing 3) supports functions related to carbohydrate binding. “IFI16” (interferon-alpha inducible protein 16) is located in the nucleus and plays an important role in the activation of the adaptive immune system. Elevated expression of “IFI16” has been associated with increased cell growth in human melanoma and prostate cancer (Certa et al. 2001; Duan et al. 2011). “RP11-469M7.1” is a non-protein coding transcript located in chromosome 2; to the best of my knowledge there is no literature describing its functions. “PPIF” (peptidylprolyl cis trans isomerase) is located in the cytoplasm and regulates metabolism, apoptosis and cancer related inflammation (Yao et al. 2005). “SCD” (steroyl-coA desaturase) is an enzyme located in the cytoplasm and is involved in fatty acid metabolism. In breast and prostate cancer cells overexpression of “SCD” has been linked to increased proliferation (Peck et al. 2016; Kim et al. 2011). Blocked expression of “SCD” in lung adenocarcinoma cells resulted in reduced xenograft growth (Scaglia & Igal 2008; Hess et al. 2010)). Additionally, inhibition of “SCD” expression resulted in increased cell death of lung cancer cells (Hess et al. 2010). “PTP4A2” (protein tyrosine phosphatase type IVA 1) is an enzyme located in the cytoplasm that plays a key role in the regulation of cell proliferation and migration. “PTP4A2” (or PRL-1) has been described to be overexpressed in various cancers and to promote tumorogenesis and tumour metastasis. More specifically, in vitro blocking of “PRL-1” expression in lung cancer cells resulted in reduced cell migration (Achiwa & Lazo 2007; Stephens et al. 2005). “NRIP1” (nuclear receptor interacting protein 1) also known as RIP140 is located in the cell nucleus and is involved in the regulation of ER (Cavaillès et al. 1995) and in the development of the murine mammary gland (Nautiyal et al. 2013). Low expression of “NRIP1” has been correlated with poor overall prognosis in chronic lymphocytic leukemia (Herold et al. 2011) and in colon cancer (Lapierre et al. 2014). Nevertheless despite these data no common theme is apparent. This is probably due to the nature of the machine learning methods that seeks to find patterns and do not necessarily find highly over or under expressed transcript abundance.

The results presented herein introduce the idea of circulating monocytes/TEMos as a source for the diagnosis of cancer. As the first study of its kind, this study has limitations mainly imposed by the relatively small sample size. Thus, the aim of
future studies is to increase the number of patients as well as to include a more balanced representation of molecular subtypes. Including patients from distinct molecular subtypes would be important for investigating the prognostic relevance of a monocyte-derived signature given the differences between breast cancer sub-types. Further validation in a bigger independent cohort is needed, as well as validation in datasets from different types of infectious diseases. Comparisons with public datasets are currently limited due to the lack of available datasets from isolated monocytes. Despite the limitations, the results described here offer a novel approach for the development of methods for early non-invasive detection of breast and endometrial cancer.

2.7 Conclusions

In conclusion, this study presents the first clear evidence of a shift in the transcriptional profile of circulating monocytes/TEMos in the presence of breast and endometrial tumours. An altered gene expression profile of TEMos in early breast cancer development has been identified, suggesting that TEMos can act as important markers for the early detection of breast cancer. Moreover, it was shown that TEMos, at the transcript level, overexpress a variety of pro-inflammatory molecules as well as pattern recognition receptors. Additionally, a 13-gene TEMo-derived signature was identified and was able to accurately stratify cancer patients from healthy individuals with 94% accuracy internally, and 100% accuracy in an independent validation set. This approach showed higher accuracy than what is currently reported for mammography. Finally, the predictive model showed evidence of potential pan-cancer predictive capabilities. The work presented within this chapter strongly supports the clinical use of circulating monocytes/TEMos for the detection of cancer.
3 Transcriptional profiling of TAMs highlights differences in breast and endometrial tumour microenvironments

3.1 Acknowledgments and contributions

I would like to acknowledge the contributions made towards this chapter by Dr. Luca Cassetta and Hui Zhang, who carried out the isolation of macrophages and RNA extraction for sequencing as described in the material and methods (section 3.4.1 and 3.4.2). I would also like to acknowledge Dr. Luca Cassetta for selecting the transcripts in Figure 20C, Figure 22C and Figure 27C.

3.2 Introduction

Genomic features of tumours and their surrounding microenvironment are proving to be promising biomarkers for diagnosis and prognosis (DeNardo et al. 2011; Hanahan & Coussens 2012; Noy & Pollard 2014). Indeed, transcriptional studies have underlined the clinical relevance of the tumour stroma for prognosis of breast cancer (Finak et al. 2008; Farmer et al. 2009; Bianchini et al. 2010; Desmedt et al. 2012). A meta-analysis of gene expression signatures of immune-infiltrating cells across human cancers has demonstrated the heterogeneity of tumour-associated leucocytes within the tumour microenvironment (TME), but also their potential as prognostic and predictive biomarkers for cancer (Gentles et al. 2015).

Tumour-associated macrophages (TAMs) have been shown to promote biological processes required for tumour progression, including angiogenesis (Lin et al. 2006; De Palma & Naldini 2011; Yeo et al. 2014), tumour cell migration and invasion (Wyckoff et al. 2004; Wyckoff et al. 2007; Chen et al. 2011), and tumour cell seeding of metastatic sites (Qian et al. 2009). In multiple studies, increased TAM density has been associated with poor clinical outcomes in breast (Mahmoud et al. 2012; Medrek et al. 2012; Zhang et al. 2013) and endometrial cancer (Kübler et al. 2014). In breast cancer mouse models, gene expression profiling of purified TAMs showed an association with poor overall survival (Ojalvo et al. 2009). In contrast,
very little is known about the functions and roles of TAMs and normal resident macrophages in human cancers. Consequently, this study investigated the transcriptomes and the functions of TAMs in breast and endometrial cancers, as well as alterations in the gene expression profile of TAMs when compared to resident macrophages. Furthermore, this study compared the transcriptional profiles of TAMs with the transcriptional profiles of their precursors, monocytes, as well as with murine TAMs extracted from mouse mammary tumours. Finally, this study systematically examined the clinical relevance of genes expressed by TAMs as markers for identification of TAMs as well as for prognosis for breast cancer patients.

### 3.3 Aims

- **Exploratory Analysis:** Determine the differences in the transcriptomes of TAMs in breast and endometrial cancer compared to macrophages resident in normal tissue. Also, determine the differences in the transcriptomes of TAMs between different cancer types.

- **Comparison of mouse and human TAMs:** Comparison of pathways and transcriptional profiles of TAMs between species using human and mouse model datasets.

- **Meta-analysis of TAMs in tissue compared to Tumour-educated monocytes (TEMo) in blood:** Comparison of gene expression profiles of TAMs and TEMos. Identification of commonly expressed genes and investigation of their functions and biological pathways.

- **Identification of specific markers for TAMs:** Assessment of commonly regulated genes between TAMs (breast cancer TAMs vs normal breast, endometrial cancer TAMs vs normal endometrium), breast cancer stroma and mouse mammary tumour dataset, that could be used for identification of TAMs.

- **Exploratory analysis of gene expressed by TAMs in whole tissue datasets for predicting prognosis of breast cancer:** Cluster analysis of the TAM
markers on whole tissue datasets in order to investigate their clinical relevance.
3.4 Materials and methods

3.4.1 Patients
Normal breast tissue from mammoplasty reduction surgeries (25-50 grams) were obtained from the Human Tissue Procurement Facility (HTPF), Ohio State University, USA; normal/benign endometrial tissue (1-2 grams) was obtained after surgery for conditions unrelated to cancer from Montefiore Medical Center, NY, USA. Breast cancer tissue (0.1-1 grams) and endometrial cancer tissue (0.1-1 grams) was obtained from Montefiore Medical Center, NY, USA. Clinical study protocols were approved by the IRB of the Albert Einstein Medical College (NY, USA). Informed consent was obtained from all human subjects included in this study.

Exclusion criteria at baseline included systemic metastatic disease, any inflammatory disorders, active infection or immunocompromised status not related to cancer. All the patients recruited did not receive chemotherapy or radiotherapy treatment before sample collection.

Pathologically breast cancer patients exhibited invasive breast cancers. Endometrial cancer patients displayed Type I and Type II cancers.

3.4.2 Sample processing and RNA-sequencing
Isolation and Fluorescence-activated cell sorting (FACS) of human tissue macrophages was performed as described (Cassetta et al. 2016). Briefly, markers for FACS isolation of macrophages included CD45+/CD3−CD19−CD56−/CD11b+/CD14+/CD163+. Immediately after sorting, all samples were centrifuged at 450 RCF for 10 min at 4C. Cell pellets were re-suspended in 350 ul of RLT lysis buffer and RNA extracted with RNAeasy Microkit (Qiagen) according to manufacturer’s instructions. RNA quantity was determined by QUBIT (Invitrogen); total RNA integrity was assessed by Agilent Bioanalyzer and the RNA Integrity Number (RIN) was calculated; samples that had a RIN > 7 were selected for RNA amplification and sequencing. RNA was amplified with Ovation RNAseq Amplification kit v2 (Nugen) according to manufacturer’s instructions; amplified RNA was sent to BGI (Philadelphia). Standard Illumina un-stranded poly-A enriched libraries were prepared, fragmented and then multiplexed generating
100bp paired-end reads per sample (HiSeq 2000 and 2500, Illumina). On average 37M reads were sequenced. Details of the number of reads generated and aligned to the reference genome, as well as information about the sequencing facility can be found in Section 6.5.

3.4.3 Sequencing alignment and Quantification
FastQ files of 2x100bp paired-end reads from resident macrophages and TAM were quality controlled using FASTQC (Andrews 2012). Samples were examined using the FASTQC interface for per base sequence quality as well as per sequence quality control. All samples were considered as good quality on the basis of a Phred score >=20 and were used for downstream processing. Quality controlled reads were then aligned to the GENCODE Human reference genome Release 19 (GRCh37.p13) using STAR aligner (Dobin et al. 2012) (version 2.3) for Linux Ubuntu, with default options as shown below and ran over 8 threads per sample (--GenomeDir GenomeDir/ --readsFilesIn /monocytes –runThreadN 8 –outFileNamePrefix MAC). STAR is an ultra-fast spliced aligner tailored for RNA-seq data and was selected on the basis that it shows comparable performance to other aligners (Engström et al. 2013). For improved accuracy, the exon junction coordinates from the reference annotation were used. Furthermore, quantification of genes was performed using the count function of HTSeq (Anders et al. 2014). Reads with alignment quality less than 10 were discarded and reads were counted at the gene level.

3.4.4 Statistical analysis for differentially expressed genes
All statistical calculations were performed in R programming language (version 3.2.3). The strategy was as follows: a) Filtering: genes with count per million (CPM) reads > 1 in at least N samples (N number of the fewest replicates of a phenotype) were retained, b) Normalization: gene expression levels were normalized using reads per kilobase per million (RPKM) values and log2 transformed using the cmp function from the EdgeR package in R (version 3.12) (Robinson et al. 2010) c) Batch effect correction: samples were corrected for batch effects using Combat function of the Surrogate Variable Analysis (SVA) package (version 3.18) (Cini et al. 2015), d)
Differential expression analysis: Limma statistical software (version 3.26.7) (Ritchie et al. 2015) was used to identify significantly differentially expressed genes with controlled False Positive Rate at 5% (FDR <= 0.05). Up-regulated genes were selected with a minimum log₂ fold change of 1.5 and down-regulated genes with a minimum log₂ fold change of -1.5. R package gplots (version 2.17) (Warnes 2016) was used to plot heatmaps. Pearson correlation and complete linkage were used for hierarchical clustering.

3.4.5 Functional enrichment and pathway analysis
Gene ontology (GO) enrichment analysis was performed using ToppGene software suite on the list of differentially expressed transcripts (FDR < 0.05) (Chen, Bardes, et al. 2009). Ingenuity Pathway Analysis (IPA) software was utilised for canonical pathway analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

3.4.6 Comparison with publically available datasets
The following publically available microarray datasets were downloaded from the public repository Gene Expression Omnibus (GEO) (Barrett et al. 2013).

Breast cancer:
1) **GSE9014**: (Finak et al. 2008) a dataset of 59 samples, 53 breast cancer tumour stroma and 6 normal breast stroma including updated clinical information.
2) **GSE18295**: (Ojalvo et al. 2010) a mouse dataset from invasive TAMs and non-invasive TAMs including 10 samples (5 TAMs, 5 invasive TAMs).
4) **NKI Dataset**: a gene expression breast cancer dataset published by Van’t Veer et al. and van de Vijver et al (NKI) (Van’t Veer et al. 2002; van de Vijver et al. 2002). The NKI dataset includes 337 samples from breast cancer patients with complete clinical information. Treated patients were filtered out resulting in 207 breast cancer patients for further analysis. Downloaded though breastCancerNKI package in R (Schroeder et al. 2011).
5) **MAINZ dataset**: a gene expression dataset published by Schmidt et al. (Schmidt et al. 2008) including 200 node-negative breast cancer patients.
Dataset was downloaded though breastCancerNKI package in R (Schroeder et al. 2011).

For all datasets, quantile normalization was performed and multiple probes were collapsed to single gene using the average expression (avereps function of the Limma package in R (Ritchie et al. 2015)) and a linear model was fitted for the identification of differentially expressed transcripts. Transcripts with FDR $\leq 0.05$ and $\log_2$FC $> +1.0$ (up-regulated) and $\log_2$FC $< -1.0$ (down-regulated) were considered to be differentially expressed.

Gprofiler web tool was used for microarray annotation to mouse gene symbols. Mouse-to-human ortholog mapping was performed using OrthoRetriever (version 1.2). Venn diagrams were constructed based on the overlapping genes between the publically available datasets described above GSE18295 (Mouse mammary tumour), GSE9014 (human breast cancer stroma) and the human breast cancer TAMs dataset.

To investigate the macrophage polarization in the human breast TAMs dataset the list of M1- and M2-like genes was extracted from the study of Martinez et al. (Martinez et al. 2006). The genes extracted from the Martinez study were then overlapped against all the expressed transcripts in the breast and endometrial TAMs dataset. In brief, to identify the M1-like and M2-like genes within the TAM population, I compared the differentially expressed transcripts (breast TAMs vs resident breast macrophages and endometrial TAMs vs resident endometrial macrophages) ($\log_2$FC $> 0$) with the extracted list of M1- and M2-like genes from the Martinez study. Similarly, to identify the M1- and M2-like genes in the resident macrophage population, I compared the differentially expressed transcripts (resident breast macrophages vs breast TAMs, and resident endometrial macrophages vs endometrial TAMs) with the extracted list of M1 and M2-like genes. In this analysis the threshold is set to Log$_2$FC $> 0$ because the initial list of M1 and M2-like genes has been identified using microarrays. Therefore, selecting less stringent criteria will help identify more genes that are potentially expressed but due to technological bias will be missed.
3.4.7 Identification and validation of a prognostic TAM-derived signature

The TAM-derived signature was comprised of differentially expressed transcripts (FDR $\leq 0.05$ and log$_2$FC greater or less than 1.0/-1.0) overlapping the human breast TAM dataset, the human breast stroma dataset (Finak et al. 2008) and the mouse mammary tumour TAMs dataset (Ojalvo et al. 2010). Cluster analysis was performed in R as implemented in the gplots package (Warnes 2016) using complete linkage and Pearson correlation. Recurrence and overall survival analysis was performed using GraphPad Prism statistical software. Kaplan-Meier curves were compared using a log-rank test. Missing data or deaths due to causes other than breast cancer were censored.

<table>
<thead>
<tr>
<th>Sample Group</th>
<th>Breast</th>
<th>Endometrium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy Breast N=4</td>
<td>Breast cancer N=4</td>
</tr>
<tr>
<td>Age (Median)</td>
<td>57</td>
<td>63</td>
</tr>
<tr>
<td>Gender (F/M)</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>ER-positive</td>
<td>N/A</td>
<td>2/4</td>
</tr>
<tr>
<td>PR-positive</td>
<td>N/A</td>
<td>1/4</td>
</tr>
<tr>
<td>HER2-positive</td>
<td>N/A</td>
<td>0/4</td>
</tr>
<tr>
<td>TN</td>
<td>N/A</td>
<td>1/4</td>
</tr>
<tr>
<td>Grade (I/II/III)</td>
<td>N/A</td>
<td>ND</td>
</tr>
<tr>
<td>Type (I or II) or benign or MR</td>
<td>4 (MR)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 4. Clinical information of samples included in this study. (F: female, M: male, N/A: not applicable, ND: not done, TN: triple negative, MR: mammoplasty reduction).
3.5 Results

Tissue macrophages were collected and isolated from a total of 13 breast and endometrial cancer patients (4 breast and 9 endometrial) and a total of 9 healthy individuals (4 breast tissue and 5 endometrial tissue) (Material & Methods 3.4, Table 4). Macrophage RNA was isolated and evaluated for quality and quantity and used for mRNA amplification and sequencing. For the differential expression analysis the same workflow as for the TEMos was used as described in Chapter 2 (Figure 5B). Briefly, for all samples, raw reads were aligned to the reference genome using a splice-aware aligner and quantified at the gene level. Samples were normalised and filtered for low or no expression. Gene expression was normalised to account for variability in sequencing depth and different gene lengths.

3.5.1 TAMs from breast and endometrial cancer show different transcriptome profiles compared to normal resident tissue macrophages and monocytic progenitors

3.5.1.1 Transcriptional profiles of TAMs are significantly different compared to normal resident macrophages in breast cancer

To understand the effect of malignancy on the transcriptional profile of macrophages in breast cancer, I compared the gene expression profiles of resident normal breast tissue macrophages (n=4) to those of TAMs (n=4). Unsupervised multi-dimensional scaling analysis (MDS) of 16,339 transcripts showed distinct clusters of normal resident macrophages and TAMs in breast cancer, as shown in Figure 20A. Differential expression analysis identified 2,609 significantly differentially expressed transcripts with controlled false positive rate at 5% (FDR <= 0.05). Stringent criteria of log2 fold change greater or less than 1.5/-1.5 resulted in 1,995 transcripts significantly up- or down-regulated in TAMs compared to normal resident macrophages (1,089 up, 906 down). Hierarchical clustering of the top 50 most significantly regulated transcripts separated TAMs from normal resident macrophages further confirming an up-regulation of gene expression in cancer compared to healthy tissue (Figure 20B).
Figure 20. Transcriptome analysis of purified TAMs from breast cancer compared to normal resident macrophages shows distinct gene expression profiles. (A) MDS plot of 16,339 transcripts showed two distinct clusters of breast cancer TAMs (red) and resident macrophages (blue). (B) Gene expression heatmap of the top 50 most significant transcripts in breast cancer TAMs (red) and resident macrophages (blue); samples are clustered using complete linkage and Pearson correlation. Red color in the heatmap indicated up-regulation, and green color indicated down-regulation based on the raw z-score (range [-1.5, 1.5]). C) Bar plot of transcripts associated with macrophage biology coloured by type (red = Adhesion molecules, purple = Enzymes, blue = Soluble factors and green = Transmembrane receptors) found to be significantly differentially regulated in TAMs compared to resident macrophages (FDR <= 0.05, log₂FC greater or less than 1.5/−1.5). X-axis represents the logarithm of fold change (Log₂FC) of monocytes between healthy individuals and patients with cancer.
In order to better characterise the transcriptional profiles of TAMs, the expression of transmembrane receptors, soluble factors, transcription factors and enzymes was investigated (FDR <= 0.05, log₂FC greater or less 1.5/-1.5) (Figure 20C). Up-regulated transmembrane receptors were associated with cell activation and antigen presentation such as MHC class II molecules, Fc receptors “CD16”, “CD32” and “CD64” and T cell co-stimulatory molecules “CD80” and “CD83”. Additionally, “TREM”, “TREM1” and “TREM2”, members of the Ig receptor family, and pattern recognition receptors (PRRs) “TLR7” and “TLR3”, were found upregulated on TAMs compared to normal resident macrophages. Chemokine receptors “CCR2” and “CX3CR1” were also significantly upregulated in TAMs. Among the genes encoding soluble factors, proinflammatory cytokines, interleukin 1 alpha (IL1A) and interleukin 1 beta (IL1B), interleukin 18 (IL-18) and tumour necrosis factor (TNF) were upregulated. Furthermore, TAMs showed up-regulation of pro-inflammatory chemokines, chemokine ligand 2 (CCL2), “CCL3”, “CCL4”, “CCL5” and “CCL8” compared to normal resident macrophages. The matrix metalloproteinase “MMP9” was found up-regulated in TAMs. Adhesion molecules, intracellular adhesion molecule 1, (ICAM1), intracellular adhesion molecule 4 (ICAM4) and integrin alpha M (ITGAM) were also found to be up-regulated. Interestingly, gene expression of “CD163”, the established marker for macrophages, was not significantly different between the 2 groups.

Differentially expressed transcripts in TAMs were analysed using IPA in order to identify enriched pathways; functional enrichment analysis reported “cell movement” and “cancer” as the most significant functions/diseases enriched in TAMs, followed by “reproductive system disease”, “cell growth and proliferation”, “inflammatory response”, “lipid metabolism” and “metabolic disease” (Figure 21A). Canonical pathway analysis predicted activation of pathways related to inflammation. More specifically, “toll like receptors (TLR)” and “pattern recognition receptors (PRRs)” signalling pathways were predicted to be significantly activated through up-regulation of cell death cytokine “TNF”, inflammatory and pro-inflammatory cytokines “IL1A”, “IL1B”, and “IL18” and toll like receptors “TLR7” and “TLR3”. Additionally, “chemokine signalling” was predicted to be significantly activated through up-regulation of pro-inflammatory chemokine “CCL2”,
inflammatory chemokine receptor 5 (CCR5) and enzyme phospholipase β2 (PLCB2). In contrast, the “NF-κB signalling” pathway, responsible for regulating immune responses was predicted to be down-regulated due to down-regulation of epidermal growth factor (EGFR), interleukin 1 receptor 1 (IL1RI) and growth hormone receptor (GHR) (Figure 21B).

Figure 21. Functional enrichment analysis of significantly regulated transcripts in breast TAMs. (A) Top canonical pathways, (B), and top disease and functions predicted to be involved in breast cancer TAMs. Pathways and transcripts were ranked by the negative log of the $P$ value of the enrichment score. The color scheme for panel B is based on $z$ scores, with activation in red, and inhibition in green. (STAT3: Signal transducer and activator of transcription 3, NF-κB: nuclear factor kappa-light-chain-enhancer of activated B cells)

3.5.1.2 Transcriptional profiles of TAMs are significantly different compared to normal resident macrophages in endometrial cancer

Next, I compared the gene expression profiles of normal resident endometrial macrophages (n=5) to those of endometrial TAMs (n=9). Similarly to macrophages from breast tissue, unsupervised MDS analysis of 16,558 transcripts showed distinct clusters of normal resident endometrial macrophages and endometrial TAMs (Figure 22A). Differential expression analysis identified a lower number (674) of differentially expressed transcripts with controlled false discovery rate at 5% (FDR <= 0.05) compared to breast TAMs. Stringent criteria of log$_2$FC greater or less than
1.5/−1.5 resulted in 446 transcripts significantly up- and down-regulated in TAMs compared to normal resident macrophages (81 up, 365 down).

**Figure 22. Transcriptome analysis of purified TAMs from endometrial cancer compared to resident macrophages shows distinct gene expression profiles.** (A) MDS plot of 16,558 transcripts showed two distinct clusters of endometrial cancer TAMs (purple) and resident macrophages (blue). (B) Gene expression heatmap of top 50 most significant transcripts between endometrial cancer TAMs (purple) and resident macrophages (blue). Samples are clustered using complete linkage and Pearson correlation. Red color in the heatmap indicated up-regulation, and green color indicated down-regulation based on the raw z-score (range [−2, 2]). (C) Top diseases and functions predicted to be involved in endometrial cancer TAMs, pathways ranked by the negative log of the P value of the enrichment score. (D) Bar plot of transcripts associated with macrophage biology coloured by type (red = Adhesion molecules, purple = Enzymes, blue = Soluble factors and green = Transmembrane receptors) found to be significantly differentially regulated in TAMs.
compared to resident macrophages from endometrial tissue. X-axis represents the logarithm of fold change ($\log_2 FC$) of monocytes between healthy individuals and patients with cancer.

In contrast to breast cancer, the top 50 most significant transcripts separating endometrial TAMs from resident macrophages were down-regulated in TAMs compared to resident endometrial macrophages (Figure 22B). Enrichment analysis using IPA did not identify any significantly activated canonical pathways but functional analysis identified “cancer”, and “reproductive system disease” as the top enriched diseases. Other enriched functions included “inflammatory response”, “cell growth and proliferation” and “lipid metabolism” (Figure 22C). At the transcript level the only transmembrane receptors found to be up-regulated were the macrophage receptor with collagenous structure (MARCO), a scavenger receptor found on the cell surface of macrophages responsible for phagocytosis, and the C-type lectin domain family 5 member A (CLEC5A), a transmembrane protein (Figure 22D).

### 3.5.1.3 Gene expression profiles of TAMs have limited overlap with TEMo

In mouse models, circulating monocytes are considered to be the precursors of TAMs (Movahedi et al. 2010; Augier et al. 2010; Qian et al. 2011), however very little is known about TAM origins in humans. Thus, I decided to compare the transcriptional profiles of TAMs in breast and endometrial cancer to those of TEMo (monocytes from breast and endometrial cancer, as described in Chapter 2). Interestingly, transcripts that distinguish TAMs from resident macrophages were very different from those that distinguish TEMos from normal monocytes with only 10% of commonly differentially expressed transcripts found (Figure 23A). Between endometrial TAMs and TEMo only 14 transcripts (11 up, 3 down) were commonly dysregulated. These commonly dysregulated transcripts were predicted to be enriched in metabolic processes through up-regulation of “PFKFB4” and “HK2” (Figure 23B). Both groups up-regulated the G coupled receptor, formyl peptide receptor 2 (FPR2), previously described to promote anti-tumor activity in macrophages (Liu et al. 2013). Ninety-nine transcripts were commonly regulated (95 up, 4 down) between breast TAMs and TEMos. Functional enrichment analysis
identified functions related to “immune responses”, “defence responses” and “cytokine-mediated signalling” (Figure 23C). More specifically, the Fc receptor “CD64”, the chemokines receptors “CCR5” and “CX3CR1” and “TLR7”, were found commonly up-regulated while the “IL-1” receptor and “IL1R1” were down-regulated in both groups.

Figure 23. Venn diagram of transcripts commonly up-regulated and down-regulated in TAMs from breast cancer and endometrial cancer compared to healthy patients, and TEMo from breast cancer and endometrial cancer compared to healthy patients. Human endometrial TAMs and human breast TAMs share 42 transcripts (8 up, 33 down). Human endometrial TAMs and human TEMo shared 14 transcripts (11 up, 3 down). Human breast TAMs and human TEMos shared 99 transcripts (95 up, 4 down). Numbers in red indicate up-regulated genes (log₂FC > 1.5, FDR <= 0.05), numbers in green downmodulated genes (log₂FC < -1.5, FDR <= 0.05). (B) Enriched biological processes of 14 transcripts shared between endometrial TAMs and TEMo. (C) Enriched biological processes of 99 transcripts shared between breast TAMs and TEMo.
3.5.2 Macrophage transcriptome profiles show tissue specificity

3.5.2.1 Resident macrophages and TAMs from breast and endometrial tissue are transcriptionally different

In order to investigate the diversity of the gene expression profiles of macrophages, I compared the transcriptomes of TAMs between breast and endometrial cancer patients as well as the gene expression profiles of normal resident macrophages in between healthy breast and endometrial tissue.

Figure 24. Gene expression analysis of purified TAMs from breast cancer compared to TAMs from endometrial cancer, and resident macrophages form healthy breast tissue compared to resident macrophages from healthy endometrial tissue reveals tissue specificity of macrophages. (A) MDS plot of 17,120 transcripts showed two distinct clusters of breast cancer TAMs (red) and endometrial TAMs (purple) (B) MDS plot of 16,688 transcripts showed two distinct clusters of resident macrophages from healthy breast tissue (green) and resident macrophages from healthy breast tissue (orange).

Unsupervised MDS analysis of 17,120 transcripts identified distinct clusters of TAMs from breast and endometrial cancer (Figure 24A). Differential expression analysis identified 537 transcripts with controlled false discovery rate at 5% (FDR <= 0.05). Stringent criteria of log$_2$FC greater or less than 1.5/-1.5 resulted in 473 transcripts significantly up- and down-regulated in breast TAMs compared to endometrial TAMs (461 up, 12 down). Additionally, shared transcripts between breast TAMs and endometrial TAMs when compared to resident macrophages showed very little overlap (8 up, 33 down) (Figure 23A).
Similarly to the TAMs, unsupervised MDS analysis of 16,688 transcripts revealed unique clustering of resident macrophages from breast and endometrial tissue (Figure 24B). Differential expression analysis identified 3,396 transcripts significantly regulated (FDR < 0.05). Filtering of significant transcripts with log₂FC greater or less than 1.5/−1.5 revealed 2,199 genes in breast resident macrophages compared to endometrial resident macrophages (1,137 up, 1,062 down).

Overall, these results highlight the heterogeneity of macrophages and indicate tissue specific phenotypes for TAMs and resident macrophages. Furthermore, these results suggest that different tumour microenvironments induce different gene expression profiles in TAMs.

3.5.3 Human breast and mouse mammary cancer TAMs share common biological profiles

The vast majority of literature describing the functions of TAMs is from experiments on mouse models. Consequently, in order to investigate similarities between mouse and human TAMs, I compared the human RNA-seq breast TAMs dataset, derived from patients with breast cancer, to a published study on invasive TAMs extracted from a mouse mammary cancer model (Ojalvo et al. 2010). Comparison of the differentially expressed transcripts in human breast TAMs to those of the mouse mammary tumour TAMs revealed that only 265 (~15%) transcripts were shared between the two groups, (161 up, 104 down) (Figure 25A). Comparative canonical pathway analysis revealed a number of key pathways being shared between the two datasets, however the “IL-8 signalling” and “NF-kB signalling” pathways were predicted to be inhibited in human TAMs, but activated in mouse mammary tumour TAMs (Figure 25B).

This is the first comparison of human breast cancer TAMs and mouse mammary tumour TAMs and it reveals a significant overlap of pathways despite the limited number of shared transcripts.
Figure 25. Comparison of human breast cancer TAMs and mouse mammary tumour TAMs. (A) Venn diagram of commonly up-regulated and down-regulated transcripts in TAMs from human breast cancer and mouse mammary tumours. Numbers in red indicate up-regulated genes (log₂FC > 1.0, FDR <= 0.05), numbers in green downmodulated genes (log₂FC < -1.0, FDR <= 0.05). (B) Bar plot of shared top canonical pathways between differentially expressed transcripts in human breast cancer TAMs and mouse mammary tumour TAMs. Pathways are ranked by the negative log of the P value of the enrichment score. The color scheme is based on z score, with activation in red, inhibition in green and non-predicted enrichment in gray. Solid colored bars for pathways enriched in the human TAMs and stripped bars for mouse mammary tumour TAMs.
3.5.4 Mixed polarization of M1- and M2-like macrophages in human TAMs and resident macrophages

To interrogate the polarization status of TAMs, I compared the human breast TAMs dataset with that published by Martinez et al. (Martinez et al. 2006) on monocyte-derived macrophages exposed to TH1 versus TH2-type stimuli. To that end, extracted transcripts described to be associated with an M1-like (n=54) phenotype and transcripts associated with an M2-like (n=43) phenotype were compared with the breast TAMs dataset. Out of the 16,339 transcripts expressed in the breast TAMs dataset, 7,091 were found up-regulated (log\(_2\)FC > 0). Out of the 54 previously described M1-like genes, 29 transcripts were identified with an M1-related phenotype (53% of total), and out of the 43 M2-like transcripts, 24 transcripts were identified with a M2-like phenotype (55% of total) in the breast TAMs dataset. Out of the 16,558 genes expressed in the endometrial TAMs dataset, 8,854 transcripts were up-regulated (log\(_2\)FC > 0). Out of the 54 M1-like transcripts, 19 transcripts were identified with a M1-related phenotype (35% of total), and out of the 43 M2-like transcripts, 11 transcripts were identified with a M2 phenotype (25% of total) in the endometrial TAMs dataset (Figure 26A).

Additionally, I investigated the polarisation status of the resident macrophages in breast and endometrial tissue. In the breast tissue, 16,339 transcripts expressed, 9,248 transcripts were up-regulated (log\(_2\)FC > 0). Out of the 54 M1-like transcripts, 10 transcripts were identified with a M1-like related phenotype (18% of total) and out of the 43 M2-like transcripts, 9 transcripts were identified with a M2-like phenotype (21% of total); In the endometrial tissue, out of the 16,558 genes expressed in the endometrial TAMs dataset, 7,704 transcripts were up-regulated (log\(_2\)FC > 0). Out of the 54 M1-like transcripts, 22 transcripts were identified with a M1-like related phenotype (41% of total) and out of the 43 M2-like transcripts, 23 transcripts were identified with a M2-like phenotype (53% of total) (Figure 26B).

These results indicate that human TAMs and resident macrophages represent mixed gene expression profiles that cannot be classified based on the previously described M1/M2-like phenotype.
3.5.5 Exploratory analysis for identification of TAM specific markers

3.5.5.1 Identification of candidate gene markers for TAMs in breast cancer

The most commonly used marker for identification of TAMs is CD163; however it is not TAM-specific (Shabo et al. 2008). In order to identify specific human TAMs markers, I compared the human breast TAMs dataset with a human breast stroma dataset from Finak et al. (Finak et al. 2008) (Human breast stroma) and a gene expression, Polyoma Middle T-positive mice (Mouse mammary tumour TAMs) (Ojalvo et al. 2010). The human breast stroma dataset consists of 59 samples (53
samples from breast cancer tumour stroma, 6 samples from normal breast stroma) and the mouse mammary tumour of 10 samples (5 invasive TAMs and 5 from general TAMs). Finak et al. showed that the poor outcome group of their study was enriched in genes related to macrophage functions. For that reason, the human breast cancer stroma dataset was selected in order to investigate the presence of commonly expressed genes with the human TAMs dataset as well as to investigate the impact of TAMs in patient survival. The mouse mammary tumour datasets was selected in order to be able to identify transcripts that are present in the mouse in order to facilitate the translation of finding into mouse models for validation purposes. Venn diagram (Figure 27A) identified 73 transcripts commonly up and down-regulated in all three dataset (51 up, 22 down, Section 6.3) at Log_{2}FC greater or less than 1.0/-1.0, FDR < 0.05. Additionally, 355 transcripts (180 up and 175 down) were exclusively shared between the human breast TAMs dataset and the human breast stroma dataset, and 192 (110 up and 82 down) exclusively shared between the human breast TAMs dataset and the mouse mammary tumour TAMs dataset. Gene ontology of the 73 TAM and breast stroma transcripts revealed significantly enriched biological process such as “Angiogenesis”, “Immune response” and “Cell activation”. Additionally, the 73 transcripts were associated with molecular functions such as “Cytokine/Chemokine binding”, “Transmembrane receptors” and “Growth factor signalling” (Figure 27B).

Out of the 73 transcripts, 9 transmembrane receptors were commonly up-regulated in all 3 datasets, including chemokine receptors “CCR5”, “CX3CR1”, sialic acid binding protein “SIGLEC1”, C-Type lectin “CLEC5A”, toll-like receptors “TLR7”, glycoproteins “CD83”, scavenger receptors (STAB1), interleukin receptor “IL1RN”, and members of the TNF receptors superfamily “TNFRSF9”. Additionally, 18 receptors were commonly up-regulated in the human breast TAMs dataset and the human breast stroma dataset, while 4 receptors were exclusively up-regulated only in the human breast TAMs dataset; namely, “SIGLEC10”, “FFAR3”, “EDAR” and “ADRA2b” (Figure 27C).
To my knowledge this is the first list of candidate transcripts, expressed by human breast TAMs, and could be used for TAM-specific identification. These markers are also up-regulated by mouse mammary tumour TAMs and therefore could facilitate experimental validation in mouse models.

**Figure 27. Identification of TAM-specific markers.** (A) Venn diagram between the human breast TAMs dataset, the human breast stroma dataset and the mouse mammary tumour TAMs dataset. 73 transcripts were shared between all datasets (FDR < 0.05, Log₂ FC greater or less than 1.0/-1.0). (B) Bar plot of biological processes and molecular functions enriched in the 73 transcripts found to be commonly regulated between all 3 datasets (FDR < 0.05). (C) Bar plot of significant transcripts (FDR < 0.05) shared between datasets. X-axis shows log₂ FC of transcripts in the human breast TAMs dataset. The transcripts in green
color indicate the 9 transmembrane receptors that represent candidates for TAMs identification.

3.5.5.2 Exploratory analysis of TAMs as prognostic markers for breast cancer

Given the contribution of macrophages in the progression of breast cancer (Pollard 2004; Qian & Pollard 2010) and their ability to discriminate breast cancer patients from healthy individuals as shown in section 3.5.1.1, I sought to investigate their potential as biomarkers for prognosis of breast cancer. Consequently, two datasets from whole tissue were utilised, (a) NKI dataset (Van ’t Veer et al. 2002; van de Vijver et al. 2002), and (b) MAINZ dataset (Schmidt et al. 2008). The NKI dataset has been the basis for the MammaPrint prognostic assay and consists of 337 samples from treated and untreated patients of mostly node-negative tumours. The MAINZ dataset has been one of the first datasets to highlight the importance of the immune system in cancer and its prognostic significance and consists of 200 node-negative breast cancer tumours, from patients that did not receive any treatment. These datasets were generated on different microarray platforms and include different probes therefore only 49 out of the 73 genes identified in section 3.5.5.1 (67%) were present in all 3 datasets. Thus, the 49 genes that were present in all 3 datasets were used downstream (Section 6.3).
Figure 28. TAM-derived signature (49 genes) identified clinically relevant groups in the NKI dataset. (A) Heatmap of expression of 49 genes in the NKI dataset highlighting different clinical groups. Red color in the heatmap indicated up-regulation, and green color indicated down-regulation based on the raw z-score (range [-5, 5]). (B) Pie chart of clinical information for the 3 clinically relevant groups; Group I (38% grade III tumours, 7.8% ER-negative, 33% distant metastasis), Group II (70% grade III and 75% ER-negative, 30% distant metastasis), and Group III (78% grade III tumour, 75% ER-negative, 65% distant survival).
metastasis). (C) Kaplan Meier curves of the 3 outcome groups for distant metastasis-free survival (P = 0.0013) with the group III showing the worst survival (HR= 2.7, P = 0.0003, 95% CI [1.9-8.66]), (D) Kaplan Meier curves of the 3 outcome groups for overall survival (P < 0.0001), with group III having the least favourable outcome (HR=4.167, P <0.0001, 95% CI [3.8-34.9]).

The NKI dataset was filtered for patients that did not receive any treatment, resulting in 207 patients for further analysis. Average follow up time was 2494.3 days (6.3 years) and 33% of the patients had a distant metastasis event (70 out 207 samples). Hierarchical clustering analysis (with k = 3) of the 49 genes on the NKI dataset identified three distinct groups of patients; Group I (n = 129 samples), Group II (n = 42 samples) and Group III (n = 36 samples) (Figure 28A). More specifically 78% of patients in the Group III presented with grade III tumours compared to 38% of patients in the Group I. Additionally, 75% of patients in the Group III presented with ER-negative cancers; and only 7.8% in Group I. Lastly, 65% of patients in Group III had a positive metastasis event compared to 33% in Group I. Interestingly, Group II included patients with high-grade invasive tumours (70% grade III and 75% ER-negative) but with low metastasis percentages (Figure 28B). In order to investigate the clinical relevance of these groups, survival analysis was conducted on distant metastasis-free survival and overall survival. The Kaplan-Meier curves showed a highly significant difference in distant metastasis-free survival between the three groups (P = 0.0013), with the Group III showing the worst survival (HR= 2.7, P = 0.0003, 95% CI [1.9-8.66] (Figure 28C). Additionally, Kaplan-Meier curves showed a highly significant difference in overall survival (P < 0.0001), with Group III having the least favourable outcome (HR=4.167, P <0.0001, 95% CI [3.8-34.9]) (Figure 28D).

Similarly, a second cluster analysis (k=3) was performed on the MAINZ dataset in order to identify 3 clinically relevant groups (Figure 29A). Average follow up time was 2816 days (or 7.71 years) and 23% of patients had a positive event of distant metastasis event (42 out of 200 samples). Group I included patients with lower grade tumours (7% grade III, 75% grade II) and a high percentage of ER-positive cancers
(93% ER-positive / 7% ER-negative) compared to the Group III (18% grade III, 82% ER-positive). Only 13% of patients presented with distant metastasis in the Group I compared to 41% of patients in the Group III. Surprisingly, the Group II included high-grade tumours (grade III 31%) and a lower percentage of ER-positive patients (65%) compared to the Group III; however the percentage of patients that presented with distant metastasis was lower (30%) than the Group III (40%) (Figure 29B). Kaplan-Meier curves showed a highly significant difference in distant metastasis-free survival ($P = 0.004$); Group III was associated with the worst distant metastasis-free survival ($HR = 3.6$, $P = 0.001$, 95% CI [2.1 - 20.0]) (Figure 29C).

All together this section presented an exploratory analysis of human TAMs as markers for prognosis in breast cancer. Although there is some evidence of a clinical relevance for human TAMs for distant metastasis and overall survival, this study represents a very small cohort and adequate analysis is needed for further validation.
Figure 29. TAM-derived signature (49 genes) identifies clinically relevant groups in the MAINZ dataset. (A) Heatmap of expression of 49 genes on the MAINZ dataset highlighting the different groups. Red color in the heatmap indicated up-regulation, and green color indicated down-regulation based on the raw z-score (range [-5, 5]). (B) Pie chart of clinical information for the 3 clinically relevant groups. Group I (7% grade III tumours, 7% ER-negative, 14% distant metastasis), Group II (31% grade III and 35% ER-negative, 30% distant metastasis), Group III (18% grade III tumour, 18% ER-negative, 41% distant metastasis). (C) Kaplan Meier curves of the 3 groups for distant metastasis-free survival (P = 0.004); Group III was associated with the worst distant metastasis-free survival (HR = 3.6, P = 0.001, 95% CI [2.1 – 20.0]).
3.6 Discussion

In experimental mouse models, it has been demonstrated that TAMs are a major component of the tumour microenvironment and, more importantly, they promote tumour development and progression (Pollard 2004; Joyce & Pollard 2009). These findings are contradictory to the classical immunological perspective of their phagocytic and antigen presenting functions. Therefore, studies on how TAMs promote tumour growth could offer new therapeutic opportunities. This is especially important in humans, because most of our knowledge comes from studies of mouse models. It has been suggested that macrophages can make up to 50% of the tumour mass (Solinas et al. 2009). In breast cancer, high TAM infiltration has been correlated with poor disease-free survival and found to be an independent prognostic predictor (Gwak et al. 2015; Zhang et al. 2012; Medrek et al. 2012; Zhang et al. 2013). In endometrial cancer increased TAM infiltration has been associated with advanced tumour grade, and lymph node metastasis. However, very little is known about the phenotypes and functions of TAMs in human cancers. Thus, this study examined the transcriptional profiles of human resident macrophages and TAMs in breast and endometrial cancer in order to dissect specific markers but also examine their prognostic value in breast cancer.

It is clear that high-throughput studies have improved our understanding of the heterogeneity of breast and endometrial cancers. This is also true for macrophage biology where a series of microarray studies have elucidated the transcriptomes of macrophages within the TME (Ojalvo et al. 2009; Ojalvo et al. 2010). However, the advent of RNA-sequencing technology offers a more sensitive method for measuring gene expression than microarrays that is not limited to known probes. Here, isolated macrophages from breast and endometrial cancer and healthy tissues were subjected to RNA-sequencing and subsequent analysis. Gene expression profiles of resident macrophages were found to be different from those of TAMs in breast and endometrial cancer. Unsupervised MDS analysis demonstrated that resident macrophages and TAMs clustered separately (Figure 20 and Figure 22). This indicates that TAMs and resident macrophages consist of distinct populations at the transcript level. Similarly, in mouse models of breast cancer Ojalvo et al.
demonstrated that TAMs from mammary mouse tumours are transcriptionally different from resident macrophages in the spleen (Ojalvo et al. 2009). Differentially regulated transcripts of human breast TAMs were enriched in pathways related to immune responses such as “TLR signaling”, “PPRs signaling” and “chemokine signaling” (Figure 21). TLRs have been shown to have a significant role in cancer progression (Hallam et al. 2009; Williams et al. 2016). Transcript abundance of MMP9, was found significantly up-regulated in breast TAMs compared to resident macrophages (Figure 20C). Matrix metalloproteinases play important roles during chronic inflammation and MMP9 is thought to be crucial for tumour progression and metastasis. In a recent study of ovarian cancer, it was demonstrated that TAMs promoted up-regulation of MMP9 and enhanced cancer cell invasion through the TLR signaling pathway (Ke et al. 2016).

Very little is known about the functions of TAMs in endometrial cancer. To the best of my knowledge there are no publically available transcriptomic datasets of isolated TAMs from endometrial carcinomas. In this study, unsupervised MDS analysis demonstrated that endometrial TAMs and resident macrophages clustered separately indicating two distinct populations at the transcript level (Figure 22A). Differentially regulated transcripts were enriched in functions related to “cancer”, “inflammatory responses” and “cellular growth and proliferation” (Figure 22C). Analysis of significant transcripts identified transmembrane receptors MARCO and CLEC5A as up-regulated. Expression of MARCO has been identified in a subset of TAMs in mouse models of breast and melanoma; and specific targeting of these TAMs using an anti-MARCO monoclonal antibody resulted in a pro-inflammatory phenotype and decreased tumour growth and metastasis (Georgoudaki et al. 2016). In this study, the number of differentially regulated transcripts in endometrial TAMs compared to resident endometrial macrophages was low. This is probably due the fact that resident endometrial macrophages were extracted from benign samples rather than healthy tissue. This is one of the limitations of this study as healthy endometrial tissue is very hard to collect. Although this data indicates that TAMs may also have a tumour-promoting phenotype in endometrial cancer it would be important to compare TAMs to resident macrophages coming from healthy endometrial tissue.
Macrophages are highly plastic cells that populate most organs of the human body. The diversity among resident macrophages from different organs has been previously described (Gautier et al. 2012; Lavin et al. 2014). In mouse models, comparison of gene expression of 7 different macrophage populations (brain, spleen, liver, lung, peritoneal cavity, intestine) revealed distinct gene expression profiles and that only a limited number of genes was shared between populations (Lavin et al. 2014). Consistent with these findings, a comparison of TAMs between breast and endometrial cancer also identified 473 transcripts differentially expressed between the two TAM populations, indicating TAM specificity in distinct TMEs (Figure 24A). In mouse models, TAMs have been shown to consist of distinct subsets within the same tumour and distinct TAM subsets have been identified in different tumour regions (Movahedi et al. 2010; Laoui et al. 2014; Ruffell et al. 2014). In addition to this, differential expression analysis revealed distinct global gene expression profiles of human resident macrophages between breast and endometrial tissue (Figure 24B). Approximately two thousand mRNA transcripts were differentially regulated by at least 1.5-fold (log₂-scale) between the two resident macrophage populations. These data suggest that distinct resident macrophage populations exist within different organs, expressing unique mRNA transcripts enabling them to carry out specific functions. Therefore, the findings in this study support the notion of tissue specificity and heterogeneity of macrophages in healthy tissue, but also importantly of TAMs in different tumours.

The pro-tumoral activity of macrophages has been previously linked to an M2-like polarization, however factors such as the plasticity of macrophages, the location within the tumour and the cancer types have indicated that both M1 and M2-like polarized states are associated with tumour progression (Movahedi et al. 2010; Franklin et al. 2014; Biswas & Mantovani 2010). This is also supported by evidence from other transcriptomic studies describing a spectrum of activation states that are not captured by the dualistic M1-M2 classification system (Xue et al. 2014; Gautier et al. 2012). Indeed, in this study transcriptional profiling of TAMs and resident macrophages failed to reveal a unique polarization state (Figure 26). Taking into consideration the inherent plasticity of macrophages is very likely that their phenotypes in vivo are more complicated than the rather simplistic M1-M2 polarized
states. Therefore, it is suggested that a functional classification rather than a polarization state would better describe macrophages and their subpopulations in human cancers. In line with this, *in vivo* studies of TAMs in a xenograft model of human renal carcinoma demonstrated pro-tumoral functions irrespective of the M1-M2 phenotype (Chittezhath et al. 2014). Hence, although TAMs display tissue specificity it is likely that they also display a variety of phenotypes within the same tumour.

It is thought that TAMs arise from circulating monocytes which migrate to primary and metastatic sites and differentiate into TAMs (Augier et al. 2010; Lin & Pollard 2004; Qian et al. 2011; Franklin et al. 2014). A comparison between differentially expressed transcripts between TEMos (isolated circulating monocytes from breast and endometrial cancer patients) and TAMs from endometrial and breast cancer patients revealed that few transcripts are shared between the two populations, suggesting that when circulating monocytes migrate to the tissue and differentiate they change their transcriptional profile.

The majority of literature describing the function of TAMs comes from experiments carried out in mouse models. Therefore, there is value in examining cross-species agreement of transcripts and enriched pathways. In this study, differentially regulated transcripts between breast TAMs and resident macrophages were overlaid with differentially regulated transcripts from a subpopulation of invasive mouse TAMs. Although, only 15% of total transcripts were shared between species, top canonical pathways such as “TLR signalling”, “Leukocyte extravasation” and “TREM1 signalling” were consistently predicted to be up-regulated in both mouse and human (Figure 25). A disagreement was observed in the proinflammatory NF-κB signalling pathway as well as in the IL-8 signalling pathway. Given the heterogeneity of TAMs it is possible that the TAMs in this study represent a distinct population that is different to the mouse invasive TAM population. It would be worthwhile to compare TAMs from invasive areas in human tumours to the invasive mouse TAMs. However, given the difficulties in isolating human TAMs in sufficient numbers this may not be possible. Furthermore, there remains a need for better markers for identification and isolation of TAMs, and their subpopulations.
Clinical identification of macrophages is usually performed using immunohistochemical markers; CD68 is a pan-macrophage marker that has been widely used for the identification of TAMs. Indeed, high numbers of CD68-positive TAMs have been associated with high-vasculosity, and decreased disease-free survival in breast cancer (Leek et al. 1996; Mahmoud et al. 2012); however, recent studies have suggested that CD68 is not a macrophage-specific marker, at least for breast cancer (Ruffell et al. 2012), and it is also expressed by other cell types such as fibroblasts (Gottfried et al. 2008). Another frequently used marker for macrophage identification is CD163. CD163-positive TAMs have been associated with less favorable clinical outcomes compared to CD68-positive TAMs (Medrek et al. 2012). However, studies have demonstrated that breast cancer cells also express CD163; therefore it is also not restricted to macrophages (Shabo et al. 2008). Given the heterogeneity of TAM subpopulations (Qian & Pollard 2010), another major limitation of these markers is that they cannot distinguish between pro-tumoral or anti-tumoral macrophages (Tang 2013). Therefore there is an urgent need to identify specific markers for identification of TAMs. Thus, this study overlaid 3 datasets coming from human breast tumour stroma, mouse mammary tumours and human breast TAMs in order to identify a set of relevant genes that could be used as potential markers for TAMs. The mouse mammary tumour dataset was included in order to facilitate evaluation of these markers in mouse models. The analysis identified a set of 73 genes that was shared between all three datasets (Figure 27). Gene ontology analysis of the biological functions revealed that 73 genes were associated with “angiogenesis”, “defense and immune responses” as well as “cell activation” Figure 27B. Out of those, a set of 9 transmembrane receptors that showed up-regulated abundance in breast TAMs were selected. This list represents candidate markers for identification of TAMs.

Given the need for better prognostic markers of breast cancer many studies have now focused on identifying prognostic gene signatures. However, most of these multigene signatures for prognosis of breast cancer have been derived from bulk tissue datasets (Sotiriou & Pusztai 2009). There is now strong evidence that the tumour stroma of breast cancer patients can be used for prediction of clinical outcome, and is also more accurate than a signature derived from whole tumour tissue (Finak et al. 2008;
Desmedt et al. 2012). Additionally, there is increasing evidence that immune-related signatures can also be used for prediction of prognosis in ER-negative tumours. For instance, tumour-infiltrating lymphocytes have been associated with good overall prognosis in ER-negative tumours (Calabrò et al. 2009; Loi et al. 2014; Pruneri et al. 2016). As a result, this study conducted an exploratory analysis examining the potential of a TAM-derived signature to predict patient outcomes in two publically available microarray datasets from whole breast cancer tissue (NKI: 207 samples & MAINZ: 200 samples). Due to missing probes in the microarrays, 49 out of the 73 genes identified in section 3.5.5.1 were used downstream for hierarchical clustering. The number of clusters was set to three (k=3) in order to identify clinically relevant outcome groups. In both datasets 3 groups of patients had significantly different clinical outcomes with one group of patients displaying significantly worse relapse-free survival and worse overall outcome. These findings indicate a clinical relevance for macrophages in human breast cancer, however this is an exploratory study and further analysis is required to validate these findings.

The analysis presented herein has several limitations: 1) The signature was derived from datasets comparing healthy tissue to breast cancer and therefore it is not the most appropriate training set for deriving a signature for stratification of patients in different prognostic groups. Furthermore, the dataset described in this chapter represents the only transcriptomic study to date from purified macrophages in humans consisting of a relative small sample size (4 healthy and 4 breast cancer individuals). A more diverse cohort with additional clinical information/features would be the most appropriate dataset to derive a TAM-specific prognostic gene signature. 2) The publically available datasets were retrospectively accrued, with no balance for molecular subtypes. Additionally, the 49-gene signature was extracted from ER-positive and ER-negative patients; therefore a cohort with a variety of molecular subtypes is required in order to make definite conclusions. 3) The sample size included in this study is relatively small. A limiting factor that didn’t allow for a larger sample size was the difficulty with extracting purified macrophages from biopsy samples, however future studies aim to increase the sample size. 4) A proportion of genes were not present in all datasets owing to different microarray platforms or changes in the annotation. Once the study is fully validated and a bigger
number of TAM samples are available future work will assess its validity in larger patient cohort such as TCGA (The Cancer Genome Atlas) or METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) is required. 5) The gene signature comprises a gene list of commonly differentially expressed genes between 3 datasets. It would be informative to evaluate the accuracy of prediction of the signature against permutated signatures. One such strategy would involve the random permutation of the expression values or the random sampling of genes to generate random signatures of equal size (49-genes) and utilise a machine learning algorithm (logistic regression, supervised classification) in order to compare the accuracy of the primary signature against the permutated or random signatures (Stark & Norden).

3.7 Conclusions

This is the first study to carry out genome-wide profiling of TAMs in human breast and endometrial cancer. The findings presented herein also support the idea of distinct populations of TAMs and resident macrophages both in breast and endometrial cancer and tissue, respectively, as well as subpopulation heterogeneity of TAMs within tumours. Furthermore, this study presents a list of up-regulated transcripts that may yield TAM-specific markers that will provide a valuable tool for further investigation into the functions and roles of these cells in breast and endometrial cancer, and potentially other cancers. Finally, this study presents preliminary results indicating that a set of 49 genes expressed by TAMs can identify groups associated with unfavorable outcomes in breast cancer.
4 General discussion and future work

My thesis analysed the transcriptional profiles of circulating monocytes and macrophages in order to identify biomarkers for diagnosis of breast and endometrial cancer as well as markers for identification of TAMs. This general discussion gives an overview of the broad findings of this study and their impact on diagnosis, prognosis and treatment of breast and endometrial cancer.

4.1 TEMos as biomarkers for detection of breast and endometrial human cancers

The findings in this thesis showed that circulating monocytes can respond to the presence of malignancy by altering their transcriptional profiles. Furthermore, a 13-gene signature was identified capable of discriminating between healthy individuals and cancer patients. As discussed in Chapter 1 (Section 1.5.2) very little is known about the functions of monocytes in human cancers. To the best of my knowledge, this is the first study to examine the gene expression profiles of circulating monocytes in breast and endometrial cancers and hence begin to address this deficiency in our understanding of human cancer-associated monocytes.

4.1.1 TEMos show distinct transcriptional profiles between breast and endometrial cancer patients compared to healthy individuals

In humans, monocytes have been classified in three distinct populations classical “inflammatory” (Human: CCR2^High^CD14^{++}CD16^{−}) intermediate monocytes (CD14^{++}CD16^{+}), and nonclassical “patrolling” monocytes (Human: CX3CR1^High^CD14^{+}CD16^{++}) (Geissmann et al. 2003; Ziegler-Heitbrock et al. 2010). In breast cancer, Feng et al. described an expansion in the patrolling monocyte subpopulation of patients, and a significant negative association with tumour size and stage (Feng et al. 2011). Although this study didn’t specifically investigate specific monocyte subpopulations, differential expression analysis revealed a mix of transcripts encoding for markers associated with each of the classical and
nonclassical subpopulation as shown in Figure 12C. More specifically, CD16 and CX3CR1 transmembrane receptors that are usually expressed by the “patrolling” subpopulation were found up-related in TEMo compared to normal monocytes. Additionally, CCR2 receptor associated with the “inflammatory” subpopulation was also found up-regulated. All together, these results suggest a mix of monocyte subpopulations presented in this dataset. Given that this work examined a mixed population rather than purifying monocytes for a specific subpopulation this results are not surprising; it would be very informative to examine the contribution of each population during malignancy in order to further understand the function of these monocytes during malignancy.

Epidemiological studies have shown an association between elevated monocyte count and poor prognosis in various cancers (Sasaki et al. 2007; Sasaki et al. 2006; H. Schmidt et al. 2005; Chen et al. 2009) including breast cancer (Wen et al. 2015). Recently, microarray gene expression profiling in renal carcinoma and colorectal cancer identified distinct monocyte profiles between cancer patients and healthy individuals (Chittezhath et al. 2014; Hamm et al. 2015). Consistently, in this study transcriptional profiling of human circulating monocytes using RNA-sequencing demonstrated distinct transcriptional profiles between monocytes in breast and endometrial cancers compared to healthy individuals (Figure 7, Figure 8, Figure 9). Interestingly, monocytes (TEMo) from breast and endometrial cancers showed very similar profiles suggesting a common mechanism for recognition of the presence of malignancy regardless of cancer type (Figure 10, Figure 11).

It appears that monocytes are altered and/or educated by the presence of malignancy resulting in a change in their transcriptional profile (Figure 12); however the roles and functions of monocytes in human cancers are still largely unknown. Chittezhath et al., proposed a pro-tumoral profile of circulating monocytes through up-regulation of pro-inflammatory cytokines and pro-tumoral genes. This pro-tumoral phenotype was supported by an IL1R-depedent mechanism (Chittezhath et al. 2014). In a similar way, TEMo exhibited an up-regulation in a number of proinflammatory cytokines and a predicted activation of the IL1 pathway (Figure 13B). Therefore it is possible that a similar pro-tumoral mechanism is present in TEMo from breast and
endometrial cancers. Additionally, in this study the TLR signalling pathway was predicted as one of the top up-regulated pathways in TEMos (Figure 13B). It can be postulated that TEMo, via cell surface TLR receptors and pattern recognition receptors, recognise signals coming from cancer cells which in turn mediates their recruitment to affected tissues. However, further studies are required to validate whether these pro-tumoral and recruitment mechanisms exist in cancer-associated monocytes.

Different from this study where freshly isolated monocytes were used for sequencing, in the Chittezhath study isolated monocytes were cultured before microarray profiling. It is likely that culturing of monocytes might induce transcriptional changes not related to the disease itself. Hence, this could lead to inconsistencies between the two studies, and also future studies, due to differences in experimental pre-processing. Therefore, a common protocol for isolation of monocytes is required to make consistent comparisons and identify common mechanisms, if any, of monocytes between different cancer types.

4.1.2 TEMo as a source of biomarkers for detection of breast and endometrial cancer

Peripheral blood mononuclear cells (PBMCs) have been used as markers for detection of various cancers, including breast cancer (Aarøe et al. 2010a; LaBreche et al. 2011; Dumeaux et al. 2014). However, the translational capacity of these signatures is not yet known. This study identified a 13-gene signature capable of distinguishing cancer patients from healthy individuals yielding an accuracy of 94% in an internal validation cohort. Consistently, in an independent validation cohort the signature yielded accuracy of 100% (Figure 15). This shows a consistency of the TEMo signature to correctly classify cancer patients from healthy individuals. However, such high accuracy in the independent validation set is likely due to the small sample size, therefore prospective studies will have to further validate its robustness in a population derived both from Edinburgh and other sources. Despite these limitations, there is strong evidence that this signature from TEMo-derived transcripts has higher accuracy than what is currently reported for mammographic
screening. In addition to that, the TEMo-derived signature showed higher or similar performance compared to signatures derived from PBMCs in breast cancer patients where the reported accuracy ranged between ~73% to ~92% (Aarøe et al. 2010a; LaBreche et al. 2011; Dumeaux et al. 2014) (Sharma et al. 2005). Hamm et al. identified a 23-gene signature derived from monocytes capable of discriminating between colorectal cancer patients and healthy individuals (Hamm et al. 2015), confirming the potential of circulating monocytes as markers for cancer diagnosis. Differently from the Hamm study, the TEMo-signature was also validated against a non-cancer dataset demonstrating its specificity to cancer (Table 3). It would be of interest to compare against various other datasets from purified circulating monocytes coming from various infectious diseases to further confirm its specificity, however none is available thus far. Furthermore, the TEMo-signature showed modest discriminating power on renal carcinoma and colorectal cancer datasets (Figure 18). However, this might be the result of the difference in the technology used to generate the datasets; it has been demonstrated that cross-platform normalization, especially for machine learning applications plays an important role (Thompson et al. 2016). Additionally, even the smallest differences in the analysis pipeline could lead to big differences in the final biomarker set (Fox et al. 2014; Starmans et al. 2012). Although, this modest performance could also indicate that TEMo-signature, thus far, is only able to discriminate between breast and endometrial cancer patients compared to healthy donors. Hence, it would be of interest to develop a common pipeline for the detection of signatures and use this pipeline to generate datasets from different types of cancers (as they become available) in order to investigate further the pan-cancer diagnostic ability of the TEMo signature.

Thus far, the totality of studies on PBMCs and circulating monocytes for identification of diagnostic signatures has been conducted using microarrays. The approach presented in Chapter 2 is the first study based on genome-wide transcriptome profiling of monocytes and thus is not limited to specific probes, as it is the case in microarrays studies. Additionally, it has been shown that RNA-seq has a wider dynamic range, identifying genes that might be missed by microarrays (Wang et al. 2009a). Therefore, I believe that this study is a useful source of information for monocytes in both cancer and healthy individuals and could be used
in subsequent studies for identification of mechanistic pathways as well as biomarkers. Furthermore, the TEMo signature described within this study has been identified using robust machine leaning methods and validated in an independent cancer population as well as negative validation datasets to confirm its robustness and specificity to cancer. Overall, this is the first study to report evidence that monocytes can provide a highly accurate, non-invasive blood-based method for the detection of breast and endometrial cancer.

4.2 Genome-wide transcriptomic analysis of TAMs and resident macrophages in breast and endometrial cancer

4.2.1 TAMs and resident macrophages from breast and endometrial cancer comprise distinct populations

Transcriptome analysis of freshly isolated TAMs compared to resident macrophages revealed two distinct macrophage populations in both breast and endometrial cancer/tissue (Figure 20). In mouse models, it has been shown that macrophages can alter their transcriptional profiles in response to the presence of malignancy. Specifically, TAMs have been described to comprise a distinct population from macrophages extracted from the spleen (Ojalvo et al. 2009). The data in Chapter 3 showed that TAMs were associated with pro-tumoral phenotypes as predicted by pathway enrichment analysis and through up-regulation of chemokines and pro-tumoral genes (Figure 21). A pro-tumoral profile of TAMs has also been recently demonstrated in human cell lines of ovarian cancer (Ke et al. 2016) and transcriptome analysis also identified a pro-tumoral profile of human TAMs in renal carcinoma (Chittezhath et al. 2014). The pro-tumoral profile of macrophages has been previously described to resemble an M2-like polarization. However, several studies have associated both M1- and M2-like polarised states with tumour progression (Movahedi et al. 2010; Franklin et al. 2014; Chittezhath et al. 2014). This study identified a mixed rather than a unique polarization in both TAMs and resident macrophages supporting the idea that TAMs have a mixed M1/M2-like phenotype with a variety of functions within the tumour (Figure 26).
Although it is not clear what is causing this shift in the phenotype of TAMs, it is known that during cancer development and progression, the tumour microenvironment in addition to cancer cells contains immune and stroma cells. TAMs and T-cells are the most frequently found immune cells in the tumour microenvironment. These different cell types communicate with each other either by direct contact or by producing cytokine and chemokine signals or by acting in autocrine and paracrine manners to affect tumour development. Different cytokine and chemokine signals can either promote or prevent tumour development and progression (Lin & Karin 2007). For instance, cytokines through activation of important signaling pathways such as NF-κb, AP-1, and STAT but also through activation of transcription factors such as SMAD can control the immune and inflammatory microenvironment resulting in anti-tumoural (IL-12, TRAIL) or pro-tumoural responses (IL-6, IL17, IL23) (Grivennikov et al. 2010). Wyckoff et al. showed that TAMs in mammary tumours interact with cancer cells through a paracrine loop (Wyckoff et al. 2004). TAMs express CSF1-R, which in turn binds to CSF1 secreted by cancer cells. TAMs produce epidermal growth factor (EGF) that activates the EGF receptor (EGFR) in cancer cells. This paracrine loop enables the co-migration of these cells, allowing their invasion in healthy tissue and intravasation. Sousa et al. showed that the mesenchymal-like cell line MDA-MB231 that secretes high levels of CSF-1 polarized macrophages towards a tumour promoting and immunosuppressive phenotype. Another tumour-derived signal that influences TAMs polarization is hypoxia. It has been shown that accumulation of TAMs in hypoxic regions of the tumour was associated with a pro-tumoral TAM phenotype (Laoui et al. 2014). There are many more different tumour-derived signals that can influence the phenotype of TAMs and other immune cells; however a comprehensive characterisation is lacking, especially at the human setting.

One of the main limitations of this analysis is related to its small sample size, consisting of 8 samples from breast tissue (4 healthy tissue and 4 breast cancer tissue) and 14 samples of endometrial tissue (5 healthy tissue and 9 endometrial cancer tissue). This is mainly due to the difficulty of isolating macrophages from the
tissue as well as the extraction of RNA from these cells. Nevertheless, despite the small sample size this study was able to dissect transcripts that discriminate between the TAMs in healthy and cancer tissues improving our understanding of macrophage biology in human cancers.

4.2.2 TAMs and resident macrophages are transcriptional distinct depending on microenvironment

In this work, comparison of differentially expressed genes of TAMs between breast and endometrial cancer revealed very different expression profiles between the two cancer types (Figure 24A). Similarly, gene expression profiles of resident macrophages were very different between breast and endometrial tissues (Figure 24B). These findings are indicative of unique phenotypes of TAMs and resident macrophages in different cancers/tissues. In breast cancer mouse models, transcriptional analysis of TAMs identified a unique invasive TAMs population (Ojalvo et al. 2010). Furthermore, functionally distinct TAMs subpopulations have been described within different regions of the tumour (Movahedi et al. 2010; Laoui et al. 2014). Hence, it is possible that like mouse models, different TAMs subpopulations exist in human tumours. Thus, it would be of interested to investigate such subpopulations and determine pro-tumoral from anti-tumoral subsets. Different approaches can be utilised to dissect TAM heterogeneity within tumours. This study used a combination of FACS sorting (Cassetta et al. 2016) and genome-wide RNA-seq analysis. Alternative approaches include single-cell RNA-seq for dissecting the intra-population heterogeneity of macrophages (Liu & Trapnell 2016). Additionally, multicolour immunofluorescence (IF) can be used to evaluate coexpression of multiple TAM markers and map the location of distinct TAM subpopulation within the same tumour (Almendro et al. 2013).

4.2.3 Identification of candidate markers for TAMs identification

The analysis in Chapter 3 identified a list of 73 genes expressed in human breast TAMs, mouse mammary tumours and human stroma (Figure 27A). Out of those 9 up-regulated transmembrane receptors were identified and could act as identification
markers for TAMs (Figure 27C). Several markers are currently used for the identification of macrophages, among them CD68 and CD163 are the most common. Both CD163- and CD68-positive TAMs have been associated with poor survival in cancer (Leek et al. 1996; Mahmoud et al. 2012; Medrek et al. 2012), however recent studies have shown that these markers are not macrophage specific (Ruffell et al. 2012; Gottfried et al. 2008; Shabo et al. 2008). Therefore, this study provides a useful list of receptors up-regulated in TAMs that could be used for specific targeting of TAMs.

As previously discussed TAMs are known to play an important role in tumour progression in mouse models of cancer via processes such as angiogenesis and matrix remodelling. Further to this, an immunosuppressive phenotype of TAMs has also been described (Mantovani et al. 2002). Recent mouse models of spontaneous breast cancer have expanded on this and have shown that TAMs can indirectly suppress the functions of CD8-positive T cells. Inhibition of TAM recruitment by IL-10R and/or CSF1R blocking has resulted in improved chemosensitivity, leading to decreased primary and metastatic tumour growth (DeNardo et al. 2011; Ruffell et al. 2014; Franklin et al. 2014). Hence, many therapeutic studies are now focusing on inhibiting TAMs and their functions in cancer. However, due to the lack of specific markers for TAMs, side effects are an issue. Therefore identification of markers specific to TAMs, such as the ones listed in this study, would allow for depletion or reprogramming of TAMs specifically, thus potentially reducing and/or eliminating current macrophage-based treatment side effects and hence possibly increasing the anti-tumour efficacy of such therapies. Moreover, given the phenotypic heterogeneity of macrophages more specific targeting of TAMs may further enhance anti-tumor T cell responses. In addition to these, these receptors were also found to be up-regulated in TAMs from mouse mammary tumours, thus allowing for their direct experimental validation as therapeutic targets in mouse models.

4.2.4 TAMs as biomarkers for prognosis of breast cancer

Gene expression profiling of bulk tissue has led to the identification of multigene signatures for prognosis of breast and endometrial cancer (Perou et al. 2000; Sørlie et
al. 2003; Parker et al. 2009; Paik et al. 2004). Such clinical signatures are currently used in the clinic to aid patient clinical decisions. However, these signatures have been criticised as to heavily rely on proliferation-related genes and being mostly effective in ER-positive node-negative tumours. Interestingly, one out of the 21 genes included in the Oncotype Dx genomic test is “CD68”. “CD68” is a pan-macrophage marker, and staining for CD68-positive TAMs in breast cancer has been shown to be an independent predictor for patient survival (Zhang et al. 2013). Additionally, the MammaPrint 70-gene signature includes “MMP9”, a matrix metallopeptidase with important roles in chronic inflammation and ECM degradation. In mouse models it was shown to be primarily produced by TAMs (Riabov et al. 2014); in ovarian cancer elevated expression of “MMP9” by TAMs resulted in enhanced cancer cell invasion (Ke et al. 2016). Therefore, it is very possible that a TAM-derived signature might have predictive power in whole tissue datasets. In parallel, transcriptional studies have presented evidence for the clinical relevance of the tumour stroma for prognosis of breast cancer (Finak et al. 2008; Farmer et al. 2009; Bianchini et al. 2010; Desmedt et al. 2012). TAMs are abundant within the tumour microenvironment (TME) and their infiltration has been correlated with poor overall survival in breast cancer, as well as advanced tumour grade and lymph node metastasis in endometrial cancer (Gwak et al. 2015; Zhang et al. 2012; Medrek et al. 2012; Zhang et al. 2013; Kübler et al. 2014). The work presented in this thesis represents a pilot study investigating the performance of a 49-gene signature, expressed by human breast TAMs, human breast tumour stroma and mouse mammary tumours, for clustering patients with different clinical outcomes. The 49-gene signature was able to identify groups associated with clinical relevance in two publically available breast cancer datasets derived from whole tissue tumours (Figure 28 and Figure 29). A comparison of the 49-gene signature with current clinical signatures such as Oncotype Dx, MammaPrint and EndoPredict didn’t show any overlap. However, comparison of the 49-gene signature against the Genomic Grade Index (GGI) MapQuant Dx (97-gene signature) identified “CX3CR1” being shared. “CX3CR1” has a role in modulating inflammatory responses. Upregulation of “CX3CR1” in TAMs has been correlated with poor prognosis in human colon carcinomas. Overall, it is expected that the 49 TAM-derived gene signature has
limited overlap with other signatures firstly because it has been extracted on a different patient cohort but also because current clinical signatures were extracted from whole tissue datasets and not from purified TAMs. Limited overlap between prognostic/diagnostic signatures is not surprising, as most current clinical signatures even if they have been derived from whole tissue datasets include different genes.

This study has several limitations, the sample size of the human breast TAMs dataset is relatively small therefore it is not possible to split it into training and testing subsets for validation. Additionally, the initial purpose of this dataset was to identify differences between resident and tumour macrophages therefore the signature mainly highlights differences between normal and cancer phenotypes and less so between patients that might progress to metastatic disease or not. Therefore, future studies will aim to increase the sample size but also follow-up of recruited patients for revalidation of the prognostic signature. Furthermore, this study performed class discovery using a fixed number of clusters (k=3); class discovery using hierarchical clustering has been criticised as to having little inter-observer reproducibility in molecular classification of breast cancer (Mackay et al. 2011). However, as discussed earlier due to small sample size it was not possible to perform more reliable classification methods, however the aim of future studies is to validate the reproducibility of the signature and evaluate a variety of other methods such as supervised classification and regression methods. Finally, the study utilised two publically available datasets for the evaluation of the prognostic relevance of the signature. Therefore, after acquisition of more samples and further validation of the 49-gene signature and it would be useful to compare against more samples but also more recent datasets i.e. METABRIC or TCGA.

Despite these limitations, this study provides preliminary evidence for the potential use of human TAMs as prognostic markers for breast cancer.

4.2.5 There is little overlap between the transcriptional profiles of TAMs and their precursors monocytes
To date, the mechanisms by which monocytes are involved in tumour progression are not known. In mouse models, it has been demonstrated that circulating monocytes
are recruited from the circulation to primary tumour sites where they differentiate into pro-tumoral macrophages (Qian et al. 2011; Movahedi et al. 2010). One of the main findings presented in Chapter 2 is the similarity of the transcriptional profiles of human monocytes during malignancy. In contrast, in Chapter 3, it was shown that gene expression profiles of breast TAMs and endometrial TAMs in humans are very different. This is suggestive of a profound shift in the profiles of circulating monocytes after differentiation to macrophages. Indeed, comparison of differentially expressed transcripts between TEMos and TAMs from breast and endometrial cancer revealed only a minority of transcripts being shared between the two populations (Figure 23). However this can only be speculative and further experiments are required in order to investigate this.

4.3 Future directions

4.3.1 Validation of the TEMo-derived diagnostic signature on larger cohorts and other settings

4.3.1.1 Larger cohort and molecular subtypes

The 13-gene TEMo-derived diagnostic signature has been shown to accurately distinguish breast and endometrial cancer patients from healthy individuals in a cohort of a total of 78 samples (Internal validation: 22 healthy, 21 breast and 16 endometrial cancer, Independent validation: 5 healthy, 13 breast and 1 endometrial cancer). Future work aims to increase the number of patients recruited in this study in order to further validate the robustness of the signature.

Previous gene profiling of whole tissue in breast and endometrial cancer has demonstrated the importance of inter-tumour heterogeneity. Several studies in breast and endometrial cancer have shown that different molecular subtypes are associated with distinct clinical outcomes (Perou et al. 2000; Sørlie et al. 2003; Sørlie et al. 2001; Gao et al. 2013; TCGA 2013). Thus far, this study was unable to discriminate between ER-positive and triple-negative patients, however, more samples are required to fully assess if the TEMo signature can discriminate between molecular subtypes. Thus, future work aims to collect sufficient numbers of distinct molecular
subtypes to elucidate if the TEMo-derived signature can discriminate between distinct molecular subtypes.

Finally, this study presented evidence that TEMo recognise the presence of malignancy at an early stage (DCIS) (Figure 17). Hence, future work will also focus on recruitment of a larger cohort of patients with DCIS disease. A proportion of DCIS patients will not progress to invasive disease and at the moment our ability to identify them is limited. This study was unable to find differences in the transcriptomes of TEMo between DCIS and invasive patients, although this might be due to the small sample size of DCIS patients (n=4). To that end, further studies aim to elucidate if TEMos are able to discriminate between patients that will progress to advanced disease and those who will not.

4.3.1.2 Other cancer types

The 13-gene TEMo-derived diagnostic signature has shown a potential use as a pan-cancer diagnostic blood test. The TEMo signature showed modest, but promising, results in discriminating renal carcinoma and colorectal carcinoma from healthy individuals. Future work aims to validate the TEMo signature on other cancers; including hormonal-related cancers such as ovarian cancer but also prostate and hepatocellular carcinoma.

4.3.1.3 TEMo signature for patients during/after treatment

In this study only untreated patients at the time of diagnosis were considered. Interestingly enough, Hamm et al. published preliminary results in a small patient cohort of the potential of monocytes as a prognostic follow-up test for detection of relapse (Hamm et al. 2015). It is therefore likely that a prognostic signature exists in the transcriptomes of monocytes for prognosis of cancer. Hence, future prospective studies aim to investigate the significance of a signature derived from TEMos to detect relapse and monitor metastatic events. It is likely that the current TEMo signature will not be able to discriminate between patients likely to relapse or respond to treatment because it hasn’t been trained to discriminate between these
settings, therefore a new signature will have to be extracted and trained on patients at different stages of treatment.

4.3.2 Clinical application of the TEMo-derived diagnostic signature
Based on the expression of only 13 genes the model has great potential for translation into a useful clinical or commercially available test. Future work will focus on the development of a protocol for the optimal clinical application of the signature for robust, accurate and reproducible diagnosis.

4.3.2.1 Gene expression assays based on targeted genes
An RNA-seq based clinical assay is not a suitable option for clinical application due to its unnecessary coverage and high cost. Commercially available prognostic assays such as Oncotype Dx and MapQuant Dx are based on low cost qPCR for measuring gene expression of genes of interest. Therefore a similar setup could be used for the 13-gene signature. Initially, gene expression levels of the 13 genes would have to be re-measured on a training set using qPCR technology; a classifier should be trained and applied to samples from newly recruited patients. Similarly to Oncotype Dx, it is likely that a set of housekeeping genes will have to be included.

Alternative approaches also include targeted RNA-sequencing (RNA captureSeq) for measuring abundance of transcripts of interest (Mercer et al. 2012). This approach is more cost effective than genome-wide RNA-seq and it is suitable for use as a diagnostic clinical test.

4.3.3 TAM-specific markers for therapeutic targeting
This study presented a list of up-regulated transmembrane receptors that could be used as markers for the identification of human breast TAMs. Future work will initially aim to validate the reproducibility of expression of these markers at the transcript level by qPCR. Furthermore, future work will utilise immunofluorescence (IF) in order to validate the expression of these markers at the protein level on a newly recruited patient cohort. Additionally, future work will use multicolour IF to
investigate colocalization of candidate markers with currently used TAM cell surface markers such as CD163.

4.3.4 Single-cell RNA sequencing of different regions of the tumour for identification of TAMs subpopulations

There is evidence in the literature for distinct subpopulations of TAMs within different tumour regions that are associated with distinct functions (Movahedi et al. 2010; Laoui et al. 2014). Additionally, this study presented evidence of distinct TAM populations between different cancers as well as a mixed polarization state rather than a unique M2-like polarization. Therefore it is possible that several TAM subpopulations with distinct functions and phenotypes exist in human tumours. In order to investigate this, future work will perform single-cell RNA-seq for identification of these distinct populations within the same tumour. However, to date it has not been capable to conduct single-cell RNA-seq on purified macrophages/TAMs. An alternative approach is single-cell qPCR, although this approach is limited to a specific number of genes.

4.3.5 Prognostic significance of TAM markers

This work identified a 49-gene signature able to group patients based on clinical outcome. Although, the data is preliminary results, it gives evidence for a prognostic relevance of human TAMs in breast cancer. Future work aims to increase the number of recruited patients with clinical follow-up information in order to provide a more appropriate training set for validation of the signature.

The prognostic value of TAMs in endometrial cancer is not yet clear, although Kübler et al. showed an association between TAM infiltration and advanced tumour grade and lymph-node metastasis (Kübler et al. 2014). This study did not investigate the prognostic potential of TAMs in endometrial cancer due to limitations similar to what was described for breast cancer TAMs (discussed earlier in section 4.2.4), however future studies aim to investigate the prognostic potential of TAMs in endometrial cancer. Given the differences in the transcriptomes of TAMs between breast and endometrial cancers is it likely that a signature derived from breast TAMs
will not have prognostic power in endometrial cancer samples. Therefore, future work will aim to increase the sample size of the endometrial cancer tissue cohort and collect more molecular subtypes in order to develop a robust prognostic model.

**4.4 Conclusion**

This thesis analysed the transcriptomic profiles of human circulating monocytes and TAMs in breast and endometrial cancers. A profound shift in the gene expression profiles of circulating monocytes in the presence of breast and endometrial cancer was reported. These findings led to the identification of a novel 13-gene signature suitable for diagnosis of breast and endometrial cancer. The robustness of the signature was validated internally using cross-validation and externally on an independently collected cohort yielding accuracy of 94% and of 100%, respectively. Importantly, the 13-gene signature has been shown to be specific to cancer compared to a non-cancer dataset from Lyme disease and to have higher performance than what is currently reported for mammography. Transcriptomic analysis showed that transcriptome profiles of TAMs differed from those of resident macrophages in breast and endometrial cancer/tissue, as well as between breast and endometrial cancer suggesting cancer specificity for TAMs and revealing potential TAM-specific markers. Finally, exploratory analysis identified a subset of 49 genes expressed by breast TAMs associated with recurrence and overall survival in publicly available datasets of breast cancer.
References


Ancuta, P. et al., 2009. Transcriptional profiling reveals developmental relationship and distinct biological functions of CD16+ and CD16- monocyte subsets. *BMC*
Anders, S., Pyl, P.T. & Huber, W., 2014. HTSeq A Python framework to work with high-throughput sequencing data,


Cardoso, F. et al., 2016. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *dx.doi.org*, 375(8), pp.717–729.


González-Domínguez, É. et al., 2015. CD163L1 and CLEC5A discriminate subsets of human resident and inflammatory macrophages in vivo. *Journal of Leukocyte


Hamm, A. et al., 2015. Tumour-educated circulating monocytes are powerful candidate biomarkers for diagnosis and disease follow-up of colorectal cancer. *Gut*, pp.1–12.

Han, Y. et al., 2015. Advanced Applications of RNA Sequencing and Challenges. *Bioinformatics and Biology Insights*, pp.29–18.


Karatzoglou, A. et al., 2004. kernlab - An S4 package for kernel methods in R.


Levan, K. et al., 2010. Identification of a Gene Expression Signature for Survival


Medrek, C. et al., 2012. The presence of tumor associated macrophages in tumor stroma as a prognostic marker for breast cancer patients. *BMC Cancer*, 12(1),


Nautiyal, J. et al., 2013. The transcriptional co-factor RIP140 regulates mammary gland development by promoting the generation of key mitogenic signals.


Pereira, B. et al., 2016. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7, p.11479.


Slamon, D.J. et al., 2001. Use of Chemotherapy plus a Monoclonal Antibody against
HER2 for Metastatic Breast Cancer That Overexpresses HER2. *dx.doi.org*, 344(11), pp.783–792.


Stark, R. & Norden, J., *Checking gene expression signatures against random and known signatures with SigCheck 2nd ed.*,.


Strachan, D.C. et al., 2013. CSF1R inhibition delays cervical and mammary tumor growth in murine models by attenuating the turnover of tumor-associated
macrophages and enhancing infiltration by CD8+ T cells. OncoImmunology, 2(12), p.e26968.


Vapnik, V.N. & Kotz, S., 1982. Estimation of dependences based on empirical data,

Varga, Z. et al., 2013. Comparison of EndoPredict and Oncotype DX Test Results in


Weigelt, B. et al., 2003. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proceedings of the National Academy of*
Sciences, 100(26), pp.15901–15905.


Zhao, S. et al., 2016. Bioinformatics for RNA-Seq Data Analysis.


Experimental Medicine, 8(11), pp.21981–21990.
6 Appendix

6.1 R scripts

Code for differential expression analysis between conditions. The same script can be modified to work for comparison of phenotypes for monocytes and macrophages.

```r
library(edgeR)
library(sva)
library(gplots)
library(ggplot2)
library(biomaRt)

#########################################################################
#Load data
#########################################################################

#import sample files from the path where files are saved
my_path = "/some_path"
sampleFiles = list.files(path = my_path , pattern=".count", full.names=FALSE)

#conditions: number of samples for condition normal (n1) and condition cancer (n2)
sampleConditions = c(rep("normal",n1), rep("cancer",n1))
d = readDGE(sampleFiles, NULL, columns=c(1,2), sampleConditions)

# Filtering and removing of low or not expressed transcripts # min: minimum number of #samples
keep = rowSums(cpm(d)>1) >= min d = d[keep,] table(keep)

#recalculate the library size after filtering
d$samples$lib.size = colSums(d$counts)

#########################################################################
Batch Correction if necessary
#########################################################################

#load file with batch information
batch_info_file = read.csv(file = "/some_path/batch_info")
b = as.factor(batch_info_file)
```

152
pheno = data.frame(colnames(d), sampleConditions, batch)

#normalise counts
edata = cpm(d, log=TRUE, prior.count=5)

# Run combat
modcombat = model.matrix(~1, data=pheno)
combat_edata = ComBat(dat = edata, batch = batch, mod = modcombat, prior.plots = FALSE, par.prior = TRUE)

#########################################################################
#MDS plot
########################################################################
plotMDS(combat_edata, main = "MDS Plot", pch = c(16,17)[as.numeric(pheno$sampleConditions), col=c("darkgreen","blue")][as.numeric(pheno$sample Conditions)], cex=1.3, xlab = "Dimension 1", ylab="Dimension 2", cex.lab=1.3, cex.axis = 2, cex.main = 2, ylim=c(-2,3), xlim=c(-2,4))

#########################################################################
#Differential expression analysis
########################################################################
#Translate ensemble IDs to gene symbol names using the bioMart database
ensembl = useMart(biomart = "ENSEMBL_MART_ENSEMBL", dataset="hsapiens_gene_ensembl", host = "grch37.ensembl.org") all_genes= getBM(filters = "ensembl_gene_id", attributes= c("external_gene_name","ensembl_gene_id"), values = rownames(combat_edata), mart = ensembl)
test_vlookup=merge(as.data.frame(combat_edata), all_genes, by.x="row.names", by.y="ensembl_gene_id")
rownames(combat_edata) = test_vlookup$external_gene_name

# Run the differential expression analysis
# Call the function run_DE that fits the linear model
# Set thresholds for selected differentially expressed transcripts
combat_fit = run_DE(combat_edata, sampleConditions)
tt = topTable(combat_fit, adjust.method = "BH", coef=1, number=999999) <= -1.5),
sigGenes= tt[tt$adj.P.Val <= 0.05,]
up.regulated.genes = sigGenes[sigGenes$logFC >= 1.5,]
down.regulated.genes = sigGenes[sigGenes$logFC <= -1.5, ]

# write a file with the significantly differentially expressed genes
write.table(sigGenes, file = "~/Desktop/TEMo_sigGenes.csv", sep=",")

###Volcano plot
volcanoplot(tt)

########################################################################
# Heatmap of top 50 significant genes
########################################################################

dist2 <- function(x, ...) as.dist(1-cor(t(x), method="pearson")) data = combat_edata
factors = sampleConditions
color.map <- function(factors) { if (factors="cancer") "#006400" else if(factors="normal") "#0000FF" else "#82CAFA") patientcolors <- unlist(lapply(factors, color.map))
hmcol <- colorRampPalette(c("green","black","red"))
test = sigGenes[1:50,]$ID

# get the row names of the most highly significant genes
df = data[rownames(data) %in% test,]

# get the count data of those 50 genes from the combat data and plot them
heatmap.2(df,col = hmcol, breaks = myBreaks ,scale="row", margin=c(4, 12),key=TRUE, symkey=TRUE, density.info="none", trace="none", dendrogram = "column", na.rm=TRUE, cexRow = 1.1, hclustfun = hclust2, distfun = dist2, ColSideColors=patientcolors)

# MY FUNCTIONS

# function that fits the linear model for differential expression
run.DE = function(data, sampleConditions){
factors=factor(sampleConditions)
factors = relevel(factors, ref="normal")
design = model.matrix(~0 + factors)
colnames(design) = levels(factors)
combat_fit = lmFit(data, design)
cont.matrix = makeContrasts(CvsN = cancer-normal, levels=design)
combat_fit2 = contrasts.fit(combat_fit, cont.matrix)
combat_fit2 = eBayes(combat_fit2)
return(combat_fit2)}

Code for feature selection and model fitting of the X²-RF model.

library(caret)
library(FSelector)
library(pROC)

# Load internal and external datasets
# Load the expression matrix of the TEMo
load("combat_edata")

# Load expression matrix of the independent validation set
load("testData")

# Format the data matrix of the internal validation with the class of each sample
data = combat_edata
class = rep(c("Normal", "Cancer"), c(22, 37))
data = t(data)
data = data.frame(data, class = class)
dataclass = as.factor(data$class)

# Format the data matrix of the independent validation with the class of each sample
testData_class = as.factor(c(rep(c("Normal", "Cancer"), c(5, 14))))
testData = data.frame(testData, class = testData_class)
# Use chi-square statistic function for feature selection and calculation of the chi-square score for each gene
weights <- chi.squared(class ~ ., data = data[, ncol(data)-1])

# Use the 13 genes with the highest chi-square score
i = 13
subset <- cutoff.k(weights, i)
new_data = data[, subset]
# Train the RF classifier using the caret function using cross validation on the 13 genes
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5, summaryFunction = twoClassSummary, classProbs = TRUE, verboseIter = FALSE, savePredictions = TRUE) set.seed(12)
rfFit_final <- train(x = new_data, y = data$class, method = "rf", metric = "ROC", ntree = 1000, tuneGrid = expand.grid(mtry=floor(sqrt(i))), trControl = ctrl)
predictors = predictors(rfFit_final)

# ROC plot for assessing the performance of the classifier internally
y=plot.roc(x = rfFit_final$pred$obs, predictor = rfFit_final$pred$Cancer, cex.axis = 2, legacy.axes = TRUE, col="blue")

# Confusion matrix for getting the performance metrics
confusionMatrix(data = rfFit_final$pred$pred, reference = rfFit_final$pred$obs)

# Prediction on the independent validation cohort
testData = testData[,subset]
testData = data.frame(testData, class = testData_class) prediction_b5_rf <- predict(rfFit_final, testData[,ncol(testData)]) prediction.prob.b5_rf <- predict(rfFit_final, testData[,ncol(testData)], type = "prob")

# Confusion matrix for assessing the performance on the independent cohort
confusionMatrix(data = prediction_b5_rf, reference = testData$class) rfFitROC_test <- roc(response = testData$class, predictor = prediction.prob.b5_rf$Cancer) plot(rfFitROC_test, legacy.axes = TRUE)

###############################################################################
# Prediction on Chittezhath dataset
###############################################################################
# Load the matrix from the Chittezhath study
load("~/my_path/biswas_annonated_normalised_data.RData")
testDataBiswas = t(biswas_normalised_data)

# Find how many of the 13-gene predictors are included in the Chittezhath study and Train a new
classifier based on the 10 genes found in the Chittezhath study
testDataBiswas = testDataBiswas[, which(colnames(testDataBiswas) %in% predictors)]

ew_data = data[, colnames(testDataBiswas)]
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5, summaryFunction =
twoClassSummary, classProbs = TRUE, verboseIter = FALSE, savePredictions = TRUE)

set.seed(12)
rfFit_final.biswas <- train(x = new_data, y = data$class, method = "rf", metric = "ROC", ntree = 1000,
tuneGrid = expand.grid(mtry = floor(sqrt(10))), trControl = ctrl)
rfFit_final.biswas
print(rfFit_final.biswas)

#Prediction on the Chittezhath study
prediction_b5_rf.biswas <- predict(rfFit_final.biswas, testDataBiswas) prediction.prob.b5_rf.biswas
<- predict(rfFit_final.biswas, testDataBiswas, type = "prob")

#Confusion matrix and ROC curves for the model
confusionMatrix(prediction_b5_rf.biswas, biswas.class)
rfFitROC_biswas <- roc(response = biswas.class, predictor = prediction.prob.b5_rf.biswas$Normal)
plot(rfFitROC_biswas, legacy.axes = TRUE, col = "green")

#########################################################################
#Prediction on Negative control (Lyme disease study)
#########################################################################

# Load the matrix from the Lyme disease study and identify how many genes predictors are present in the study
load("~/my_path/lyme_expression_data.RData")
signature = data.frame(names = predictors)
sig_names = merge(signature, all_genes, by.x="names", by.y="ensembl_gene_id")
lymeData = normalised_exps[which(rownames(normalised_exps) %in% sig_names$external_gene_name),] dim(lymeData)
lymeData = as.data.frame(t(lymeData))
lymeClass = as.factor(c(rep("Cancer", 28), rep("Normal", 13))) new_data = data[, colnames(lymeData)]

##Train the model on the datasets with the predictors found in common
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5, summaryFunction = twoClassSummary, classProbs = TRUE, verboseIter = FALSE, savePredictions = TRUE)
set.seed(12)
rfFit_final.lyme <- train(x = new_data, y = data$class, method = "rf", metric = "ROC", ntree = 1000, tuneGrid = expand.grid(mtry = floor(sqrt(8))), trControl = ctrl)

### Confusion matrix and ROC curves for the model
prediction_lyme <- predict(rfFit_final.lyme, lymeData) prediction.prob.lyme <- predict(rfFit_final.lyme, lymeData, type = "prob")
confusionMatrix(data = prediction_lyme, reference = lymeClass)
result.roc.model1 <- roc(response = lymeClass, predictor = prediction.prob.lyme$Cancer, levels = rev(levels(lymeClass)))

###############################################################################
#Prediction on the colorectal cancer dataset
###############################################################################
# Load the matrix from the Lyme disease study and identify how many genes predictors are present in the study
load(file="~/my_path/colorectal_cancer_normalised_data.RData")
predictors = predictors(rfFit_final) sum(rownames(gut_normalised_data) %in% predictors)
gut.dataset = gut_normalised_data[which(rownames(gut_normalised_data) %in% predictors),]

## Train the model on the datasets with the predictors found in common
new_data = data[, colnames(gut.dataset)]
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5, summaryFunction = twoClassSummary, classProbs = TRUE, verboseIter = FALSE, savePredictions = TRUE)
set.seed(12)
rfFit_final.gut <- train(x = new_data, y = data$class, method = "rf", metric = "ROC", ntree = 1000, tuneGrid = expand.grid(mtry = floor(sqrt(12))), trControl = ctrl)
### Table of the 13-genes included in the TEMo-signature

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Chromosome</th>
<th>Log$_2$FC</th>
<th>q-value (FDR)</th>
<th>Type</th>
<th>Gene ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTP1</td>
<td>5q15</td>
<td>1.57</td>
<td>4.44e-07</td>
<td>Calcium Binding Protein</td>
<td>calcium ion binding; calcium-mediated signaling;</td>
</tr>
<tr>
<td>PIBF1</td>
<td>13q22.1</td>
<td>3.10</td>
<td>1.76e-08</td>
<td>-</td>
<td>interleukin-4 receptor binding; immune system process; negative regulation of</td>
</tr>
<tr>
<td>TMTC2</td>
<td>12q21.31</td>
<td>2.80</td>
<td>1.50e-07</td>
<td>Endoplasmic Reticulum Protein</td>
<td>IL-12; negative regulation of NK-cell;</td>
</tr>
<tr>
<td>SLF1/ ANKRD32</td>
<td>5q15</td>
<td>2.99</td>
<td>1.80e-08</td>
<td>Transcription regulator</td>
<td>DNA repair; positive regulation of protein complex assembly; response to DNA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>damage stimulus; CDP-choline pathway; lipid metabolic process; phospholipid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>biosynthetic process; phospholipid metabolic process;</td>
</tr>
<tr>
<td>CEPT1</td>
<td>1p13.3</td>
<td>1.73</td>
<td>2.51e-07</td>
<td>Enzyme</td>
<td>regulation of transcription, DNA-dependent; transcription, DNA-dependent</td>
</tr>
<tr>
<td>ZNF114</td>
<td>19q.3.33</td>
<td>3.06</td>
<td>1.15e-08</td>
<td>Zinc finger protein</td>
<td>carbohydrate binding</td>
</tr>
<tr>
<td>CRYBG3</td>
<td>3q11.2</td>
<td>2.37</td>
<td>8.25e-06</td>
<td>-</td>
<td>activation of innate immune response; autophagy; B cell receptor signaling</td>
</tr>
<tr>
<td>IFI16</td>
<td>1q22</td>
<td>1.85</td>
<td>6.05e-05</td>
<td>Transcription regulator</td>
<td>pathway; cellular response to interferon-beta; regulation of transcription;</td>
</tr>
<tr>
<td>PPIF</td>
<td>10q22.3</td>
<td>-3.04</td>
<td>3.54e-07</td>
<td>Enzyme</td>
<td>apoptotic mitochondrial changes; apoptotic process; necroptosis; programmed</td>
</tr>
<tr>
<td>SCD</td>
<td>10q24.31</td>
<td>-3.50</td>
<td>8.82e-10</td>
<td>Enzyme</td>
<td>cell death; fatty acid biosynthetic process; cellular lipid metabolic process;</td>
</tr>
<tr>
<td>RP11-469M7.1</td>
<td>2q</td>
<td>-2.55</td>
<td>1.87-06</td>
<td>-</td>
<td>cell cycle; diphosphorylation; multicellular organismal development; androgen</td>
</tr>
<tr>
<td>PTP4A1</td>
<td>6q12</td>
<td>-1.38</td>
<td>1.31e-03</td>
<td>Phosphatase</td>
<td>receptor signaling pathway; lipid storage; ovarian follicle rupture;</td>
</tr>
<tr>
<td>NRIP1</td>
<td>21q11.2</td>
<td>-1.36</td>
<td>2.28e-05</td>
<td>Transcription regulator</td>
<td></td>
</tr>
</tbody>
</table>

159
## 6.3 List of 73 candidate gene markers for identification of TAMs

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Ensembl ID</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAB1</td>
<td>ENSG00000010327</td>
<td>stabilin 1</td>
</tr>
<tr>
<td>POU2F2</td>
<td>ENSG00000028277</td>
<td>POU class 2 homeobox 2</td>
</tr>
<tr>
<td>CAPG</td>
<td>ENSG00000042493</td>
<td>capping protein (actin filament), gelsolin-like</td>
</tr>
<tr>
<td>CP</td>
<td>ENSG00000047457</td>
<td>ceruloplasmin (ferroxidase)</td>
</tr>
<tr>
<td>TNFRSF9</td>
<td>ENSG00000049249</td>
<td>tumour necrosis factor receptor superfamily, member 9</td>
</tr>
<tr>
<td>TRAF3IP2</td>
<td>ENSG00000056972</td>
<td>TRAF3 interacting protein 2</td>
</tr>
<tr>
<td>NCKAP1</td>
<td>ENSG00000061676</td>
<td>NCK-associated protein 1</td>
</tr>
<tr>
<td>TGFBR3</td>
<td>ENSG00000069702</td>
<td>transforming growth factor, beta receptor III</td>
</tr>
<tr>
<td>RAB27A</td>
<td>ENSG00000069974</td>
<td>RAB27A, member RAS oncogene family</td>
</tr>
<tr>
<td>SIGLEC1</td>
<td>ENSG00000088827</td>
<td>sialic acid binding Ig-like lectin 1, sialoadhesin</td>
</tr>
<tr>
<td>LAMB1</td>
<td>ENSG00000091136</td>
<td>laminin, beta 1</td>
</tr>
<tr>
<td>TREM2</td>
<td>ENSG00000095970</td>
<td>triggering receptor expressed on myeloid cells 2</td>
</tr>
<tr>
<td>CTSZ</td>
<td>ENSG00000101160</td>
<td>cathepsin Z</td>
</tr>
<tr>
<td>IL21R</td>
<td>ENSG00000103522</td>
<td>interleukin 21 receptor</td>
</tr>
<tr>
<td>LIPA</td>
<td>ENSG00000107798</td>
<td>lipase A, lysosomal acid, cholesterol esterase</td>
</tr>
<tr>
<td>SLC15A3</td>
<td>ENSG00000110446</td>
<td>solute carrier family 15 (oligopeptide transporter), member 3</td>
</tr>
<tr>
<td>TRPV4</td>
<td>ENSG00000111199</td>
<td>transient receptor potential cation channel, subfamily V, member 4</td>
</tr>
<tr>
<td>DSE</td>
<td>ENSG00000111817</td>
<td>dermatan sulfate epimerase</td>
</tr>
<tr>
<td>HDDC2</td>
<td>ENSG00000111906</td>
<td>HD domain containing 2</td>
</tr>
<tr>
<td>CD83</td>
<td>ENSG00000112149</td>
<td>CD83 molecule</td>
</tr>
<tr>
<td>BAG2</td>
<td>ENSG00000112208</td>
<td>BCL2-associated athanogene 2</td>
</tr>
<tr>
<td>Symbol</td>
<td>Gene Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LY86</td>
<td>ENSG00000112799</td>
<td>lymphocyte antigen 86</td>
</tr>
<tr>
<td>ACVR2B</td>
<td>ENSG00000114739</td>
<td>activin A receptor, type IIB</td>
</tr>
<tr>
<td>PLEK</td>
<td>ENSG00000115956</td>
<td>plekstrin</td>
</tr>
<tr>
<td>CNN3</td>
<td>ENSG00000117519</td>
<td>calponin 3, acidic</td>
</tr>
<tr>
<td>CTGF</td>
<td>ENSG00000118523</td>
<td>connective tissue growth factor</td>
</tr>
<tr>
<td>IFIT2</td>
<td>ENSG00000119922</td>
<td>interferon-induced protein with tetratricopeptide repeats 2</td>
</tr>
<tr>
<td>CLU</td>
<td>ENSG00000120885</td>
<td>clusterin</td>
</tr>
<tr>
<td>ADCY7</td>
<td>ENSG00000121281</td>
<td>adenylate cyclase 7</td>
</tr>
<tr>
<td>PLAU</td>
<td>ENSG00000122861</td>
<td>plasminogen activator, urokinase</td>
</tr>
<tr>
<td>MMP19</td>
<td>ENSG00000123342</td>
<td>matrix metallopeptidase 19</td>
</tr>
<tr>
<td>F13A1</td>
<td>ENSG00000124491</td>
<td>coagulation factor XIII, A1 polypeptide</td>
</tr>
<tr>
<td>SGCE</td>
<td>ENSG00000127990</td>
<td>sarcoglycan, epsilon</td>
</tr>
<tr>
<td>EMILIN2</td>
<td>ENSG00000132205</td>
<td>elastin microfibril interfacfer 2</td>
</tr>
<tr>
<td>PNISR</td>
<td>ENSG00000132424</td>
<td>PNN-interacting serine/arginine-rich protein</td>
</tr>
<tr>
<td>RRAS2</td>
<td>ENSG00000133818</td>
<td>related RAS viral (r-ras) oncogene homolog 2</td>
</tr>
<tr>
<td>IL1RN</td>
<td>ENSG00000136689</td>
<td>interleukin 1 receptor antagonist</td>
</tr>
<tr>
<td>SLCO2B1</td>
<td>ENSG00000137491</td>
<td>solute carrier organic anion transporter family, member 2B1</td>
</tr>
<tr>
<td>PRCP</td>
<td>ENSG00000137509</td>
<td>prolylcarboxypeptidase (angiotensinase C)</td>
</tr>
<tr>
<td>YAP1</td>
<td>ENSG00000137693</td>
<td>Yes-associated protein 1</td>
</tr>
<tr>
<td>CH25H</td>
<td>ENSG00000138135</td>
<td>cholesterol 25-hydroxylase</td>
</tr>
<tr>
<td>CIR1</td>
<td>ENSG00000138433</td>
<td>corepressor interacting with RBPJ, 1</td>
</tr>
<tr>
<td>CXCL9</td>
<td>ENSG00000138755</td>
<td>chemokine (C-X-C motif) ligand 9</td>
</tr>
<tr>
<td>CSGALNACT1</td>
<td>ENSG00000147408</td>
<td>chondroitin sulfate N-acetylgalactosaminytransferase 1</td>
</tr>
<tr>
<td>Gene Symbol</td>
<td>Ensembl ID</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>RGS10</td>
<td>ENSG00000148908</td>
<td>regulator of G-protein signaling 10</td>
</tr>
<tr>
<td>ADAMTS1</td>
<td>ENSG00000154734</td>
<td>ADAM metallopeptidase with thrombospondin type 1 motif, 1</td>
</tr>
<tr>
<td>VSIG4</td>
<td>ENSG00000155659</td>
<td>V-set and immunoglobulin domain containing 4</td>
</tr>
<tr>
<td>ITGB2</td>
<td>ENSG00000160255</td>
<td>integrin, beta 2 (complement component 3 receptor 3 and 4 subunit)</td>
</tr>
<tr>
<td>CCR5</td>
<td>ENSG00000160791</td>
<td>chemokine (C-C motif) receptor 5 (gene/pseudogene)</td>
</tr>
<tr>
<td>HPGDS</td>
<td>ENSG00000163106</td>
<td>hematopoietic prostaglandin D synthase</td>
</tr>
<tr>
<td>AIM2</td>
<td>ENSG00000163568</td>
<td>absent in melanoma 2</td>
</tr>
<tr>
<td>DCSTAMP</td>
<td>ENSG00000164935</td>
<td>dendrocyte expressed seven transmembrane protein</td>
</tr>
<tr>
<td>GNGT2</td>
<td>ENSG00000167083</td>
<td>guanine nucleotide binding protein (G protein), gamma transducing activity polypeptide 2</td>
</tr>
<tr>
<td>CX3CR1</td>
<td>ENSG00000168329</td>
<td>chemokine (C-X3-C motif) receptor 1</td>
</tr>
<tr>
<td>PTAFR</td>
<td>ENSG00000169403</td>
<td>platelet-activating factor receptor</td>
</tr>
<tr>
<td>ITGAM</td>
<td>ENSG00000169896</td>
<td>integrin, alpha M (complement component 3 receptor 3 subunit)</td>
</tr>
<tr>
<td>TM4SF1</td>
<td>ENSG00000169908</td>
<td>transmembrane 4 L six family member 1</td>
</tr>
<tr>
<td>PTGER4</td>
<td>ENSG00000171522</td>
<td>prostaglandin E receptor 4 (subtype EP4)</td>
</tr>
<tr>
<td>GATM</td>
<td>ENSG00000171766</td>
<td>glycine amidinotransferase (L-arginine:glycine amidinotransferase)</td>
</tr>
<tr>
<td>HPSE</td>
<td>ENSG00000173083</td>
<td>heparanase</td>
</tr>
<tr>
<td>CYSLTR1</td>
<td>ENSG00000173198</td>
<td>cysteiny1 leukotriene receptor 1</td>
</tr>
<tr>
<td>C1QB</td>
<td>ENSG00000173369</td>
<td>complement component 1, q subcomponent, B chain</td>
</tr>
<tr>
<td>C1QA</td>
<td>ENSG00000173372</td>
<td>complement component 1, q subcomponent, A chain</td>
</tr>
<tr>
<td>PIK3R6</td>
<td>ENSG00000174083</td>
<td>phosphoinositol-3-kinase, regulatory subunit 6</td>
</tr>
<tr>
<td>ADCY6</td>
<td>ENSG00000174233</td>
<td>adenylate cyclase 6</td>
</tr>
<tr>
<td>SLC8A1</td>
<td>ENSG00000183023</td>
<td>solute carrier family 8 (sodium/calcium exchanger), member 1</td>
</tr>
<tr>
<td>CAMK1D</td>
<td>ENSG00000183049</td>
<td>calcium/calmodulin-dependent protein kinase ID</td>
</tr>
<tr>
<td>Genes</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>ADAP2</td>
<td>ArfGAP with dual PH domains 2</td>
<td></td>
</tr>
<tr>
<td>TNFAIP2</td>
<td>tumour necrosis factor, alpha-induced protein 2</td>
<td></td>
</tr>
<tr>
<td>SPRY4</td>
<td>sprouty homolog 4 (Drosophila)</td>
<td></td>
</tr>
<tr>
<td>PTPN1</td>
<td>protein tyrosine phosphatase, non-receptor type 1</td>
<td></td>
</tr>
<tr>
<td>TLR7</td>
<td>toll-like receptor 7</td>
<td></td>
</tr>
<tr>
<td>CLEC5A</td>
<td>C-type lectin domain family 5, member A</td>
<td></td>
</tr>
</tbody>
</table>

Genes highlighted in bold consist the 49-genes present in all external microarray datasets and used for clustering analysis in section 3.5.5.2.

### 6.4 Details of monocyte samples sequencing and alignment

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample Name</th>
<th>Set</th>
<th>Sequencing facility</th>
<th># of Input reads</th>
<th># Uniquely mapped reads</th>
<th>% of uniquely mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>mon1</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>36219431</td>
<td>30257606</td>
<td>83.54%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon2</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>11327878</td>
<td>6509103</td>
<td>57.46%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon3</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>32983605</td>
<td>27336799</td>
<td>82.88%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon4.ACOMJAACX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>11774134</td>
<td>8263868</td>
<td>70.19%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon4.AD13VMACX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>17628026</td>
<td>11606582</td>
<td>65.84%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon5</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>42930698</td>
<td>32876858</td>
<td>76.58%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon6.AD13UEACXX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>7843386</td>
<td>2782897</td>
<td>35.48%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon6.AD13VMACX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>33711693</td>
<td>8720532</td>
<td>25.87%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon7</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>50262984</td>
<td>39484469</td>
<td>78.56%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon8</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>24351380</td>
<td>13249844</td>
<td>56.26%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon9.ACOMJAACX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>26205647</td>
<td>19250858</td>
<td>73.46%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon9.AD13VMACX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>21479956</td>
<td>15112912</td>
<td>70.36%</td>
</tr>
<tr>
<td>Class</td>
<td>Mon</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>Score 1</td>
<td>Score 2</td>
<td>Accuracy</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>----------</td>
<td>-----------------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon 10.ACOMJAAC XX</td>
<td>Training</td>
<td></td>
<td>23998650</td>
<td>17348992</td>
<td>72.29%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon 10.AD13VMAC XX</td>
<td>Training</td>
<td></td>
<td>24560720</td>
<td>17126223</td>
<td>69.73%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 11</td>
<td>Training</td>
<td></td>
<td>40812157</td>
<td>32024985</td>
<td>78.47%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 12</td>
<td>Training</td>
<td></td>
<td>36391454</td>
<td>28376116</td>
<td>77.97%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 13</td>
<td>Training</td>
<td></td>
<td>36802620</td>
<td>25540364</td>
<td>69.40%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 14</td>
<td>Training</td>
<td></td>
<td>31033592</td>
<td>23337340</td>
<td>75.20%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 15</td>
<td>Training</td>
<td></td>
<td>37449978</td>
<td>28690561</td>
<td>76.61%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 16</td>
<td>Training</td>
<td></td>
<td>31033592</td>
<td>23337340</td>
<td>75.20%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 17</td>
<td>Training</td>
<td></td>
<td>18198967</td>
<td>10996495</td>
<td>60.42%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 18</td>
<td>Training</td>
<td></td>
<td>27443132</td>
<td>18870783</td>
<td>70.26%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 19.ACOMJAAC XX</td>
<td>Training</td>
<td></td>
<td>18198967</td>
<td>10996495</td>
<td>60.42%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 19.AD13VMAC XX</td>
<td>Training</td>
<td></td>
<td>25203984</td>
<td>14322042</td>
<td>56.82%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 20.ACOMJAAC XX</td>
<td>Training</td>
<td></td>
<td>46635327</td>
<td>34741820</td>
<td>74.50%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon 20.AD13VMAC XX</td>
<td>Training</td>
<td></td>
<td>13449350</td>
<td>9449035</td>
<td>70.26%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 21</td>
<td>Training</td>
<td></td>
<td>39664732</td>
<td>28002098</td>
<td>70.60%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 22.ACOMJAAC XX</td>
<td>Training</td>
<td></td>
<td>17380178</td>
<td>14670314</td>
<td>84.41%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 22.AD13VMAC XX</td>
<td>Training</td>
<td></td>
<td>15126408</td>
<td>12471497</td>
<td>82.45%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 23</td>
<td>Training</td>
<td></td>
<td>30102516</td>
<td>25070676</td>
<td>83.28%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 24</td>
<td>Training</td>
<td></td>
<td>24323927</td>
<td>19872126</td>
<td>81.70%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 25</td>
<td>Training</td>
<td></td>
<td>27106872</td>
<td>20099154</td>
<td>74.15%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 26</td>
<td>Training</td>
<td></td>
<td>44243583</td>
<td>33469068</td>
<td>75.65%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 27</td>
<td>Training</td>
<td></td>
<td>25034108</td>
<td>20436398</td>
<td>81.63%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 28.ACOMJAACXX</td>
<td>Training</td>
<td></td>
<td>37582839</td>
<td>29881715</td>
<td>79.51%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 28.AD13VMACXX</td>
<td>Training</td>
<td></td>
<td>18171931</td>
<td>13914117</td>
<td>76.57%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon 29.ACOMJAACXX</td>
<td>Training</td>
<td></td>
<td>17790781</td>
<td>10956294</td>
<td>61.58%</td>
</tr>
<tr>
<td>Tissue</td>
<td>Sample ID</td>
<td>Laboratory</td>
<td>Training</td>
<td>Drive Number</td>
<td>Validation</td>
<td>Accuracy</td>
</tr>
<tr>
<td>----------</td>
<td>----------------</td>
<td>------------</td>
<td>----------</td>
<td>--------------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon29.AD13VMACXX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>15344765</td>
<td>8628365</td>
<td>56.23%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon30.AC0MJAACXX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>13803802</td>
<td>9827281</td>
<td>71.19%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon30.AD13VMACXX</td>
<td>Training</td>
<td>Albert Einstein</td>
<td>7127518</td>
<td>4762494</td>
<td>66.82%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon56</td>
<td>Training</td>
<td>BGI</td>
<td>58289278</td>
<td>52288544</td>
<td>89.71%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon57</td>
<td>Training</td>
<td>BGI</td>
<td>45856974</td>
<td>39857932</td>
<td>86.92%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon58</td>
<td>Training</td>
<td>BGI</td>
<td>33850967</td>
<td>29726233</td>
<td>87.82%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon59</td>
<td>Training</td>
<td>BGI</td>
<td>38206882</td>
<td>34362905</td>
<td>89.94%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon60</td>
<td>Training</td>
<td>BGI</td>
<td>46461835</td>
<td>41605192</td>
<td>89.55%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon61</td>
<td>Training</td>
<td>BGI</td>
<td>31123425</td>
<td>27431895</td>
<td>88.14%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon62</td>
<td>Training</td>
<td>BGI</td>
<td>35681915</td>
<td>31923671</td>
<td>89.47%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon63</td>
<td>Training</td>
<td>BGI</td>
<td>25679235</td>
<td>22937354</td>
<td>89.32%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon64</td>
<td>Training</td>
<td>BGI</td>
<td>37274712</td>
<td>32439845</td>
<td>87.03%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon65</td>
<td>Training</td>
<td>BGI</td>
<td>48095894</td>
<td>42356873</td>
<td>88.07%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon66</td>
<td>Training</td>
<td>BGI</td>
<td>43395103</td>
<td>38320900</td>
<td>88.31%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon67</td>
<td>Training</td>
<td>BGI</td>
<td>51595517</td>
<td>44835607</td>
<td>86.90%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon68</td>
<td>Training</td>
<td>BGI</td>
<td>46602760</td>
<td>40908341</td>
<td>87.78%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon69</td>
<td>Training</td>
<td>BGI</td>
<td>33652476</td>
<td>29746395</td>
<td>88.39%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon70</td>
<td>Training</td>
<td>BGI</td>
<td>35138818</td>
<td>31323024</td>
<td>89.14%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon71</td>
<td>Training</td>
<td>BGI</td>
<td>46345295</td>
<td>41049937</td>
<td>88.57%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon72</td>
<td>Training</td>
<td>BGI</td>
<td>31571135</td>
<td>28192784</td>
<td>89.30%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon73</td>
<td>Training</td>
<td>BGI</td>
<td>28508845</td>
<td>25175582</td>
<td>88.31%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon74</td>
<td>Training</td>
<td>BGI</td>
<td>52298114</td>
<td>45135848</td>
<td>86.30%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon75</td>
<td>Training</td>
<td>BGI</td>
<td>40828374</td>
<td>35614404</td>
<td>87.23%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon76</td>
<td>Training</td>
<td>BGI</td>
<td>30262014</td>
<td>26252910</td>
<td>86.75%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon78</td>
<td>Training</td>
<td>BGI</td>
<td>58805262</td>
<td>51404866</td>
<td>87.42%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon79</td>
<td>Training</td>
<td>BGI</td>
<td>45693499</td>
<td>40195119</td>
<td>87.97%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon80</td>
<td>Training</td>
<td>BGI</td>
<td>38153401</td>
<td>33968337</td>
<td>89.03%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon81</td>
<td>Training</td>
<td>BGI</td>
<td>43161242</td>
<td>38994933</td>
<td>90.35%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon82</td>
<td>Training</td>
<td>BGI</td>
<td>47553368</td>
<td>43376901</td>
<td>91.22%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon83</td>
<td>Training</td>
<td>BGI</td>
<td>47814207</td>
<td>43542518</td>
<td>91.07%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon84</td>
<td>Training</td>
<td>BGI</td>
<td>41806828</td>
<td>38289462</td>
<td>91.59%</td>
</tr>
<tr>
<td>Tissue</td>
<td>Sample</td>
<td>Type</td>
<td>BGI</td>
<td>Size</td>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>----------</td>
<td>---------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>BRC</td>
<td>mon85</td>
<td>Training</td>
<td>BGI</td>
<td>39485993</td>
<td>35621006</td>
<td>90.21%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon87</td>
<td>Training</td>
<td>BGI</td>
<td>38898482</td>
<td>35595277</td>
<td>91.51%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon103</td>
<td>Test</td>
<td>BGI</td>
<td>25620468</td>
<td>22949441</td>
<td>89.57%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon104</td>
<td>Test</td>
<td>BGI</td>
<td>24561517</td>
<td>21802420</td>
<td>88.77%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon105</td>
<td>Test</td>
<td>BGI</td>
<td>25728799</td>
<td>22981214</td>
<td>89.32%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon106</td>
<td>Test</td>
<td>BGI</td>
<td>25736927</td>
<td>23010986</td>
<td>89.41%</td>
</tr>
<tr>
<td>Healthy</td>
<td>mon107</td>
<td>Test</td>
<td>BGI</td>
<td>25671989</td>
<td>22414496</td>
<td>87.31%</td>
</tr>
<tr>
<td>DCIS</td>
<td>mon98</td>
<td>Test</td>
<td>BGI</td>
<td>25591594</td>
<td>22613745</td>
<td>88.36%</td>
</tr>
<tr>
<td>DCIS</td>
<td>mon99</td>
<td>Test</td>
<td>BGI</td>
<td>25535291</td>
<td>22163159</td>
<td>86.79%</td>
</tr>
<tr>
<td>DCIS</td>
<td>mon100</td>
<td>Test</td>
<td>BGI</td>
<td>25551475</td>
<td>22319618</td>
<td>87.35%</td>
</tr>
<tr>
<td>DCIS</td>
<td>mon101-2</td>
<td>Test</td>
<td>BGI</td>
<td>25606027</td>
<td>22717203</td>
<td>88.72%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon93-2_L5</td>
<td>Test</td>
<td>BGI</td>
<td>13363012</td>
<td>12008587</td>
<td>89.86%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon93-2_L6</td>
<td>Test</td>
<td>BGI</td>
<td>12302623</td>
<td>11050617</td>
<td>89.82%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon94</td>
<td>Test</td>
<td>BGI</td>
<td>25376352</td>
<td>22540019</td>
<td>88.82%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon95</td>
<td>Test</td>
<td>BGI</td>
<td>25265732</td>
<td>22151264</td>
<td>87.67%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon96_L1</td>
<td>Test</td>
<td>BGI</td>
<td>14365624</td>
<td>12736470</td>
<td>88.66%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon96_L2</td>
<td>Test</td>
<td>BGI</td>
<td>12259781</td>
<td>10828547</td>
<td>88.33%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon97</td>
<td>Test</td>
<td>BGI</td>
<td>25519599</td>
<td>22590733</td>
<td>88.52%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon118</td>
<td>Test</td>
<td>BGI</td>
<td>25582722</td>
<td>22674029</td>
<td>88.63%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon119</td>
<td>Test</td>
<td>BGI</td>
<td>25596619</td>
<td>22714528</td>
<td>88.74%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon120</td>
<td>Test</td>
<td>BGI</td>
<td>25607234</td>
<td>22849575</td>
<td>89.23%</td>
</tr>
<tr>
<td>BRC</td>
<td>mon121</td>
<td>Test</td>
<td>BGI</td>
<td>25658925</td>
<td>23086265</td>
<td>89.97%</td>
</tr>
<tr>
<td>Endo Cancer</td>
<td>mon112</td>
<td>Test</td>
<td>BGI</td>
<td>25608802</td>
<td>22787617</td>
<td>88.98%</td>
</tr>
</tbody>
</table>
### 6.5 Details of macrophage samples sequencing and alignment

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample Name</th>
<th>Sequencing facility</th>
<th># of Input reads</th>
<th># of uniquely mapped reads</th>
<th>Uniquely mapped reads (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR-resident</td>
<td>MAC_3</td>
<td>BGI</td>
<td>37043054</td>
<td>32667201</td>
<td>88.19%</td>
</tr>
<tr>
<td>BR-resident</td>
<td>MAC_4</td>
<td>BGI</td>
<td>45173305</td>
<td>39639808</td>
<td>87.75%</td>
</tr>
<tr>
<td>BR-resident</td>
<td>MAC_5</td>
<td>BGI</td>
<td>41228220</td>
<td>34729203</td>
<td>84.24%</td>
</tr>
<tr>
<td>BR-TAMs</td>
<td>MAC_9</td>
<td>BGI</td>
<td>56067484</td>
<td>49162660</td>
<td>87.68%</td>
</tr>
<tr>
<td>BR-TAMs</td>
<td>MAC_12</td>
<td>BGI</td>
<td>38448007</td>
<td>33061966</td>
<td>85.99%</td>
</tr>
<tr>
<td>Endo-resident</td>
<td>MAC_14</td>
<td>BGI</td>
<td>30742690</td>
<td>26260224</td>
<td>85.42%</td>
</tr>
<tr>
<td>Endo-resident</td>
<td>MAC_16</td>
<td>BGI</td>
<td>43213483</td>
<td>38002703</td>
<td>87.94%</td>
</tr>
<tr>
<td>Endo-resident</td>
<td>MAC_17</td>
<td>BGI</td>
<td>38212518</td>
<td>33229031</td>
<td>86.96%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_19</td>
<td>BGI</td>
<td>35206493</td>
<td>30507265</td>
<td>86.65%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_21</td>
<td>BGI</td>
<td>49896107</td>
<td>42965646</td>
<td>86.11%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_22</td>
<td>BGI</td>
<td>39852217</td>
<td>33350750</td>
<td>83.69%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_23</td>
<td>BGI</td>
<td>45352036</td>
<td>37848749</td>
<td>83.46%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_24</td>
<td>BGI</td>
<td>59631838</td>
<td>52941205</td>
<td>88.78%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_25</td>
<td>BGI</td>
<td>56567465</td>
<td>48962930</td>
<td>86.56%</td>
</tr>
<tr>
<td>BR-TAMs</td>
<td>MAC_27</td>
<td>BGI</td>
<td>40859541</td>
<td>37533221</td>
<td>91.86%</td>
</tr>
<tr>
<td>BR-resident</td>
<td>MAC_28</td>
<td>BGI</td>
<td>25656732</td>
<td>23537557</td>
<td>91.74%</td>
</tr>
<tr>
<td>BR-TAMs</td>
<td>MAC_29</td>
<td>BGI</td>
<td>25470267</td>
<td>23382496</td>
<td>91.80%</td>
</tr>
<tr>
<td>Endo-resident</td>
<td>MAC_30</td>
<td>BGI</td>
<td>25660226</td>
<td>23329914</td>
<td>90.92%</td>
</tr>
<tr>
<td>Endo-resident</td>
<td>MAC_31</td>
<td>BGI</td>
<td>25630680</td>
<td>23329196</td>
<td>91.02%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_32</td>
<td>BGI</td>
<td>25488368</td>
<td>22977958</td>
<td>90.15%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_33</td>
<td>BGI</td>
<td>25506728</td>
<td>23039905</td>
<td>90.33%</td>
</tr>
<tr>
<td>Endo-TAMs</td>
<td>MAC_35</td>
<td>BGI</td>
<td>24522941</td>
<td>21662683</td>
<td>88.34%</td>
</tr>
</tbody>
</table>