FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

# Identification of autophagy signatures in breast cancer using The Cancer Genome Atlas data

*Master's Thesis*

**MSc Bioinformatics**

**Marina Vabistsevits** <rzq995@alumni.ku.dk>

## Supervisors

Dr Elena Papaleo <elenap@cancer.dk>
Prof Albin Sandelin <albin@binf.ku.dk>
Dr Kristoffer Vitting-Seerup <kristoffer.vittingseerup@bio.ku.dk>

*7 August, 2017*

# Abstract

Breast cancer is the second leading cause of death among women in developed countries. It is an extremely heterogeneous disease with distinct subtypes and clinical implications. One of the mechanisms through which cancer development could be controlled is autophagy. Autophagy is a highly-conserved process of cellular self-degradation in response to stress or nutrient starvation. It can promote both cell-survival and cell-death, hence it has been linked to many human pathologies including cancer. In cancer, autophagy is believed to have a complex and context-dependent role. At initial stages of cancer, it helps to recycle damaged molecules thereby reducing cytotoxic damage and preventing tumorogenesis. Impaired autophagy contributes to cancer progression. At later stages of cancer, tumour cells learn to use autophagy for their own cytoprotection and thereby evade cancer therapies.

However, the current understanding of the exact link between autophagy regulation and breast cancer development is very superficial and there are many unanswered questions. Both breast cancer and autophagy research fields have seen a number of highlights and breakthroughs in the past decade, therefore now it is time to start bridging the knowledge gap between the two by identifying autophagy signatures in breast cancer.

In this project, data from one of the largest publicly available cancer resources, The Cancer Genome Atlas, was used to analyse gene expression changes between various breast cancer classification subgroups, including tumour morphology, cancer stage, and PAM50 molecular profile. The sample classification subgroups were extensively investigated with different exploratory analysis methods, which confirmed the tremendous heterogeneity of breast cancer. The exploratory analysis results were used to guide the differential expression analysis setup. Additionally, soft-clustering was performed to detect clusters of genes that have similar expression behaviour with respect to their changes along cancer stages. Following that, differentially expressed genes and the genes assigned to individual clusters were tested for autophagy enrichment.

The enrichment analysis has shown that autophagy genes are overrepresented among the genes that are downregulated in cancer versus normal, regardless of subtype or morphology. This signature was found both in differential expression analysis and soft-clustering results. A consensus set of genes between the results of two methods was identified, thereby providing a list of candidate autophagy genes that can be further explored for their role in breast cancer.

# Contents

# List of Abbreviations

| | |
|---|---|
| AJCC | The American Joint Committee on Cancer |
| ATG | Autophagy-related gene |
| CBL | Computational Biology Lab |
| CPM | Counts per million |
| DCC | Data Coordinating Centre |
| DCRC | Danish Cancer Research Centre |
| DE | Differential expression / differentially expressed |
| DEA | Differential expression analysis |
| DEGs | Differentially expressed genes |
| EDA | Exploratory data analysis |
| ER | Estrogen receptor |
| FPKM | Fragments per kilobase of transcript per million |
| GDC | Genomic Data Commons |
| HER2 | Human epidermal growth factor receptor 2 |
| ICD-O-3 | International Classification of Diseases for Oncology, 3rd edition |
| IDC | Invasive ductal carcinoma |
| IHC | Immunohistochemistry |
| ILC | Invasive lobular carcinoma |
| LIR | LC3-interacting region |
| NCI | National Cancer Institute |
| NHGRI | National Human Genome Research Institute |
| PAM50 | Prediction Analysis of Microarray, based on 50 gene set |
| PAS | Phagophore assembly site |
| PCA | Principal component analysis |
| PCs | Principal components |
| PE | Phosphatidylethanolamine |
| PI3P | Phosphatidylinositol 3-phosphate |
| PPI | Protein-protein interactions |
| PR | Progesterone receptor |
| qRT-PCR | Quantitative reverse transcriptase polymerase chain reaction |
| ROS | Reactive oxygen species |
| RPKM | Reads per kilobase of transcript per million |
| STRING | Search Tool for the Retrieval of Interacting Genes/Proteins (database) |
| TCGA | The Cancer Genome Atlas |
| TCGA-BRCA | The Cancer Genome Atlas breast cancer data |
| TNBC | Triple Negative Breast Cancer |
| TNM | Tumour, Nodes, Metastasis (system) |
| UICC | International Union Against Cancer |
| WHO | World Health Organisation |

# List of Figures

# List of Tables

Chapter 1

# Introduction

## 1.1 Breast Cancer

### 1.1.1 Disease Overview

Breast cancer is the leading cause of cancer-related deaths among women in developed countries [1]. It has been estimated that approximately 2.4 million females developed breast cancer in 2015, and it was the cause of death for more than 520,000 individuals [2]. Denmark has the second highest disease incidence rate per 100,000 individuals in the world [3].

Screening programs, education, and improved therapeutic strategies have decreased the mortality rates from this disease, but not at the desired magnitude. The most plausible explanation for this discordance is the lack of a complete picture of the biological heterogeneity of breast cancers [4]. Breast cancer is not a single disease, but is composed of multiple subtypes with distinct morphologies and clinical implications [5]. Growing evidence implies that carcinomas with different histopathological and biological features exhibit distinct behaviours that lead to different treatment responses and require tailored therapeutic approaches [6]. Therefore, accurate grouping of breast cancers into clinically relevant subtypes is of high importance for prognosis prediction and treatment decision making.

### 1.1.2 Management and Prognosis

Historically, prognostication in breast cancer has relied on the clinicopathological parameters such as patient age, tumour size, lymph node involvement, presence of metastasis, histological grade, and status of individual molecular markers such as hormone receptors and proliferation markers expression. All of these parameters are routinely used in clinics to stratify patients for prognostic predictions, assign treatments, and include patients into clinical trials [4].

The limitations of these markers parameters in predicting risk of recurrence has led to the use of mRNA- and DNA-based markers. The advent of high-throughput platforms for gene expression profiling has shown that tumour cell response to treatment is not determined by anatomical prognostic factors but rather intrinsic molecular characteristics [7]. The large-scale analysis of the genetic makeup of tumours has permitted understanding of the genomic and transcriptomic landscape of breast cancer. It has brought the concept of breast cancer heterogeneity to the forefront of cancer research, and the fact that distinct subtypes of breast cancer are completely different diseases that affect the same anatomical site [8]. This new strategy has changed how breast cancer patients are managed and treated, which provided an incremental increase in the reproducibility and accuracy of disease prognosis and therapy selection [9]. Importantly, though, the prognostic and predictive power of molecular profiling has been shown to be complementary to, rather than a replacement for, traditional clinicopathological parameters [8].

### 1.1.3 Classification Methods

Breast cancers can be classified by different schemata. Each of the classification assignments influence treatment response and prognosis. This section will present the main traditional and novel breast cancer classification conventions to introduce the overwhelming heterogeneity observed among the breast cancer patients.

#### 1.1.3.1 Cancer Staging by TNM

Cancer staging is a way of determining and describing cancer location and spread in the body. The underlying purpose of staging is to characterise the extent or severity of an individual's cancer, and to bring together cancers that have similar prognosis and treatment [10].

Breast cancer is staged using the TNM system developed by The American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC) [11]. The TNM staging system is the most common tool used by clinicians to converge the results from diagnostic tests and scans, and it involves two steps. Firstly, cancer is classified by several factors, **T** for the extent of the tumour, **N** for the extent of spread to the nodes, and **M** for the presence of metastasis. Then, these are grouped as TNM factors to find the overall cancer stage. Table 1.1 shows the overview of TNM combinations. There are 5 stages: stage 0 (zero), which is noninvasive *in situ* carcinoma, and stages I through IV (1 through 4), which are used for invasive breast cancer.

**Table 1.1:** Cancer staging/ prognostic groups TNM system reference table (from AJCC/UICC). The combination of T, N, and M is used to assign the overall stage (substage).

| Stage | | Tumour | Nodes | Metastasis |
|---|---|---|---|---|
| **0** | 0 | Tis | N0 | M0 |
| **1** | IA | T1 | N0 | M0 |
| | IB | T0 | N1mi | M0 |
| | | T1 | N1mi | M0 |
| **2** | IIA | T0 | N1 | M0 |
| | | T1 | N1 | M0 |
| | | T2 | N0 | M0 |
| | IIB | T2 | N1 | M0 |
| | | T3 | N0 | M0 |
| **3** | IIIA | T0 | N2 | M0 |
| | | T1 | N2 | M0 |
| | | T2 | N2 | M0 |
| | | T3 | N1 | M0 |
| | | T3 | N2 | M0 |
| | IIIB | T4 | N0 | M0 |
| | | T4 | N1 | M0 |
| | | T4 | N2 | M0 |
| | IIIC | Any T | N3 | M0 |
| **4** | IV | Any T | Any N | M1 |

M0 includes cM0(i+). N1mi indicates cancer in the axillary lymph nodes $> 0.2mm$ but $< 2mm$ (micro-metastasis). *Table is adapted from [11].*

**T** – The tumour values (`TX, T0, Tis, T1, T2, T3 or T4`) depend on the cancer at the primary site of origin in the breast. `TX` refers to an inability to assess that site; `T0` means no evidence of primary tumour; `Tis` refers to noninvasive *in situ* carcinoma or Paget's disease [11]. The numbered `T`'s refer to the size of the tumour, ranging from less that 1mm to larger than 50mm in the greatest dimension [10].

**N** – The lymph node values (`NX, N0, N1, N2 or N3`) depend on the number, size and location of breast cancer cell deposits in various regional lymph nodes, such as the armpit (axillary lymph nodes), the collar area (supraclavicular lymph nodes), and inside the chest (internal mammary lymph nodes) [12]. `N0` refers to no regional node metastases.

**M** – The two metastatic values (`M0 and M1`) refer respectively to no clinical or radiographic evidence of distant metastases, and the presence of breast cancer cells in locations other than the breast and regional lymph nodes, such as to bone, brain, lung, i.e. detectable distant metastases. Another recently introduced category between the two, `cM0(i+)`, refers to molecularly or microscopically detected tumour cells in circulating blood, bone marrow or non-regional nodal tissue, no larger than 0.2 mm, but without evidence or symptoms or signs of metastases [11]. This category, however, does not change the stage grouping.

Historically, the TNM anatomic stage groups have been associated with outcome measures, including overall survival and disease-free survival [11]. For groups of patients it provides an accurate prediction of outcome, but within stage groups at individual patient level, outcome predictions are more problematic, as they have different biologic subtypes of cancers that express different biomarkers. In this way, while TNM classification remains the basis for cancer staging, but other factors, such as receptor status, histology, and molecular subtype are now incorporated into parallel prognostic stage groups. Despite the predictive power of intrinsic breast cancer subtypes (e.g. PAM50 classifier, discussed in Section 1.1.3.4), the anatomic TNM classification provides a common language for communicating disease burden.

#### 1.1.3.2 Tumour Morphology

Breast cancers are heterogeneous tumours that show a wide variation not only with regard to their clinical presentation and behaviour, but also morphological spectrum. The majority (95%) of breast tumours are adenocarcinomas – cancers that arise from the epithelial lining of the breast components [13], such as ducts and lobules.

The main division between mammary carcinomas is whether it is *in situ* or invasive (infiltrating) by its nature, meaning whether it is limited to the epithelial component or it has invaded the surrounding connective tissue [14]. *In situ* carcinomas have the potential to become invasive cancer, unless adequately and timely treated, while invasive carcinomas are capable of spreading to other sites of the body, such as lymph nodes or other organs, in the form of metastases [13].

Invasive and *in situ* carcinomas are further classified as ductal and lobular based on the site from which the tumour originated, thereby forming two major groups – ductal carcinomas and lobular carcinomas. Approximately 80% of breast carcinomas are Invasive Ductal Carcinoma (IDC), followed by Invasive Lobular Carcinomas (ILC) which account for approximately 10-15% of cases [14]. Apart from the these two, at least 18 different histological breast cancer types (morphological/ pathological entities) are described by the World Health Organization (WHO) [15], [16]. The remaining cases of invasive carcinoma are comprised of other special types of breast cancer that are characterised by unique pathological discoveries [13]. These special types include mucinous, metaplastic, medullary, micropapillary, papillary, tubular and others [14]. It is important to distinguish between these various subtypes, because they can have different prognoses and treatment implications.

Figure 1.1 shows the four morphologies that are of the most relevance to this project: IDC (8500/3), ILC (8520/3), metaplastic carcinoma (8575/3), and mucinous carcinoma (8480/3) [17]. The codes in brackets are ICD-O-3 codes (International Classification of Diseases for Oncology, $3^{rd}$ edition) from WHO.

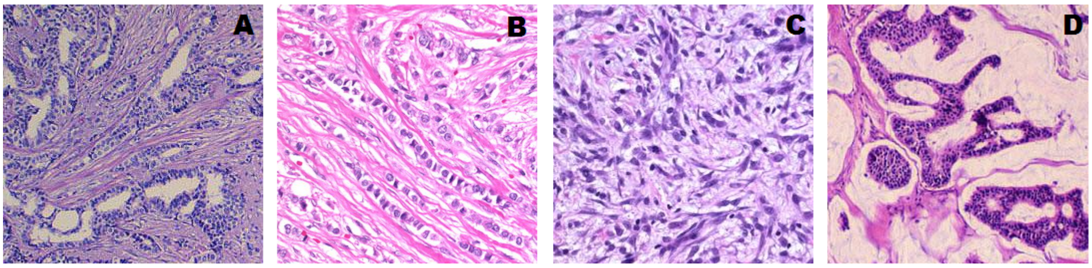

**Figure 1.1:** Breast carcinoma invasive morphologies/histologies. (A) Ductal, (B) Lobular, (C) Metaplastic, (D) Mucinous. Images taken from [18], [19]

IDC and ILC have distinct pathological features. Specifically, lobular carcinoma small cells are arranged individually, in single sheet pattern, and they have different molecular and genetic aberrations that distinguish them from ductal carcinomas [20]. The lobular phenotype is determined by dyregulation of cell-cell adhesion, primarily driven by lack of *E*-cadherin protein expression, which is often used as staining marker to tell it apart from the ductal morphology [19], [21]. Ductal carcinoma has no specific histological characteristics other than invasion through the basement membrane of a breast duct [14].

Metaplastic and mucinous carcinomas are very rare types of breast carcinomas that account for $> 1\%$ and 2% of all cases [13]. Metaplastic breast cancer is a histologically distinct type due to its characteristic outlook of complex admixture of differentiated cells [13]. It is made up of abnormally looking ductal-origin cells which are thought to have undergone a change in form (*metaplasia*) to become completely different cells that look like soft and connective tissue in the breast. Metaplastic breast cancers are also known to behave more aggressively than other kinds of breast cancers [22]. It has been show that $> 90\%$ of these cancers lack expression of ER/PR and HER2 (i.e. triple negative), and display a basal-like molecular profile [23] (more details in the following sections).

Mucinous carcinoma is less aggressive than more typical kinds of invasive cancer. The histological hallmark of this carcinoma is the excess of extracellular mucin, which surrounds the cancer cells and becomes a part of the tumour [24]. Mucinous tumours are usually ER/PR positive and HER2 negative, and consistently display a luminal phenotype [23].

#### 1.1.3.3 Receptor Status

The identification of breast cancer receptor status is routinely used for prognostic and predictive purposes [25]. A prognostic factor aims to give an indication of patient's overall clinical outcome (i.e. risk of recurrence and mortality), while a predictive factor is any measurement associated with response to a given therapy [26].

The most common method of testing for receptor status is immunohistochemistry (IHC), which stains the cells based on the presence of estrogen receptors (ER), progesterone receptors (PR) and human epidermal growth factor receptor 2 (HER2) [25]. Receptor status is a critical assessment for all breast cancers as it determines the suitability of using targeted adjuvant treatments such as tamoxifen and trastuzumab, which are now one of the most effective treatments of breast cancer [27].

Approximately 70% of invasive breast cancers are ER and/or PR positive. Estrogen and/or progesterone receptor positive cancer cells depend on estrogen or related hormones for their growth, therefore this kind of cancer can be treated with endocrine therapy drugs to reduce either the effect of estrogen (e.g. tamoxifen) or the actual levels of estrogen itself [28]. In this way, hormone therapy blocks the tumour from using estrogen and/or progesterone thereby slowing or stopping its growth.

15-20% of invasive breast cancers are characterised by overexpression of HER2 [27], which is a good example of both a prognostic and predictive biomarker. HER2 expression is associated with a worse prognosis and higher risk of recurrence, however, HER2+ patients can immensely benefit from the highly effective therapeutic option such as the monoclonal antibody therapy (trastuzamab) [29].

Cancers that do not express any of the three receptors (ER, PR, HER2) are referred to as triple-negative breast cancers (TNBC) [30]. Triple-negative breast cancers comprise a very heterogeneous group of cancers with a variety of prognoses, but most often thay are associated with a more aggressive outlook. These cancers are the most challenging type, because they do not respond to endocrine therapy or other available targeted agents [31].

#### 1.1.3.4 PAM50 Molecular Profile

With the wider adaptation of high-throughput gene expression profiling, it has been shown that tumour cell response to treatment is determined by 'molecular profiles' rather than physiological tumour characteristics and receptor status [32].

The original studies by Sørlie *et al.* [33] classified breast cancer tumours into five intrinsic subtypes with distinct clinical outcomes based on their 'molecular portrait': Luminal A, Luminal B, Basal-like, HER2-enriched, and Normal-like. The classification was guided by the differences underlying the gene expression patterns that reflect the fundamental differences of the tumours at the molecular level [34]. The observed five subtypes map quite well to the previously defined IHC receptor subtypes (Table 1.2), and have been repeated by several other studies with varying number of genes included in the subtypes' signature [5].

In 2009, Parker *et al.* [35] reported a clinically applicable 50-gene classifier, PAM50, containing mostly hormone receptor and proliferation-related genes. By comparing global gene expression data from microarray and qRT-PCR, a minimised set of 50 genes was identified that could reliably classify each tumour into one of the intrinsic subtypes with 93% accuracy. Over the past 7 years, the PAM50 intrinsic subtypes have shown to provide significant prognostic and predictive information beyond standard clinicopathological parameters [4], [36]. The PAM50 assay is now clinically implemented worldwide using the *nCounter* platform [4].

**Table 1.2:** Summary of the breast cancer molecular subtypes. PAM50 subtypes map to IHC status subtypes. Ki67 proliferative marker is used as a distinction between Luminal A and B. Luminal A and Normal-like share the IHC subtype. *Table adapted from [5].*

| **PAM50 subtype** | **IHC receptor status** | **Prognosis** |
|---|---|---|
| Luminal A | [ER+PR+] HER2– Ki67– | *Good* |
| Luminal B | [ER+PR+] HER2– Ki67+ | *Intermediate* |
| | [ER+PR+] HER2+ Ki67+ | *Poor* |
| HER2-enriched | [ER–PR–] HER2+ | *Poor* |
| Basal-like | [ER–PR–] HER2–, basal marker | *Poor* |
| Normal-like | [ER+PR+] HER2– Ki67– | *Intermediate* |

**Luminal tumours**

Luminal A and B subtypes are distinguished by the expression of two main biological processes: proliferation/cell cycle-related pathways and luminal/hormone-regulated pathways [4], and have expression patterns reminiscent of the luminal component of the breast [37]. Luminal tumours are the most common subtypes among breast cancer ( 60%) with Luminal A being the majority [5]. Luminal A is characterised by expression of ER-related genes and low expression of proliferative genes [38]. The IHC profile for Luminal A includes positive expression of ER, PR, cytokeratin CK8/18, and absence of HER2 expression and low Ki67 (proliferation marker) [4]. Luminal B tumours have higher expression of proliferation related genes, and lower expression of luminal-related genes or proteins such as PR and FOXA1, but not ER [39], which is found similarly expressed between two luminal subtypes and can only help distinguish luminal from non-luminal disease [4].

In general, the luminal subtypes carry a good prognosis, with Luminal B tumours having a significantly worse scenario than the Luminal A [34]. Treatment response differs between the two, but generally they respond well to hormone therapy and poorly to conventional chemotherapy [40]. Luminal A tumours could be adequately treated with just endocrine therapy, while Luminal B tumours which are more proliferative will benefit more from the combined therapeutic strategy of chemotherapy and hormonal treatment [41].

**HER2-enriched tumours**

The identification of HER2-enriched subtype among the molecular profiles found in breast cancer was reassuring because it confirmed the clinical impression that the tumours with HER2 overexpression are systematically different from other breast cancers [40]. The HER2-enriched subtype is characterised by the high expression of HER2-related and proliferation-related genes, intermediate expression of luminal-related genes, and low expression of basal-related genes and proteins [4]. The best mapping IHC subtype to HER2-enriched tumours is HER2-overexpressed (ER–/PR–/HER+), but it is not exclusive to them, i.e. it can also be associated with other subtypes [5]. Additionally, although the majority (68%) of HER2-enriched tumours have HER2 amplification, there are also cases of HER2-enriched subtype with HER2 negative receptor status [4].

HER2-enriched tumours carry poor prognosis that is derived from a higher risk of early relapse [42], but they can benefit greatly targeted therapeutic agents, such as anti-HER2 monoclonal antibody trastuzumab.

### Basal-like tumours

The Basal-like subtype name comes from the observation that the expression pattern of this subtype resembles that of the basal epithelial cells in other parts of the body and normal breast myoepithelial cells [37], [40]. The characteristics include lack of expression of ER-related (luminal) genes, low expression of HER2, and strong expression of basal markers (such as cytokeratins 5, 6, 17) [43]. Some of the important hallmarks of basal phenotype is low expression of BRCA genes [44] and aggressive features such as TP53 mutations [33].

Among all the intrinsic subtypes, Basal-like is the most distinct [45]. The unsupervised results in 2013 study by Prat *et al.* [46] revealed that a subgroup of breast cancers, Basal-like by PAM50, should be considered a molecular entity by itself just like ovarian or colorectal cancer, and that $> 70\%$ of Basal-like breast cancers were more similar to squamous cell lung cancer than to Luminal A or B disease [46].

Basal-like tumours account for 60-90% of triple negative breast cancers [47], and in the past the terms TNBC and Basal-like were used interchangeably [4]. However, within TNBC, all intrinsic molecular subtypes can be identified.

Triple negative Basal-like tumours are of particular interest because of their aggressive clinical course and currently lack any form of standard targeted systemic therapy. These tumours are associated with a lower disease-specific survival and a higher risk of local and regional relapse [31]. The size of basal tumours is, in general, larger than the other subtypes [48], and they also tend to show rapid growth [49]. The metastasis pattern also separates basal tumours from the other breast cancers, with its tendency towards internal organs (excluding bone) and less likely to involve lymph nodes [49].

The poor prognosis of Basal-like subtype has been confirmed by multiple independent data [40]. However, it is not clear if this prognosis is due to poor therapy options or inherent aggressiveness. Given the triple negative receptor status, basal tumours are not amenable to conventional targeted breast cancer therapies such as endocrine therapy or trastuzumab, leaving chemotherapy the only option [5], [40].

### Normal-like tumours

Normal-like intrinsic subtype is the smallest breast cancer group that accounts for less than 10% of the cases [5]. Normal-like subtype shares its IHC receptor status with Luminal A ([ER+ PR+] HER2− Ki67−), but differs on expression pattern. Also, as suggested by the name, Normal-like cancers are characterised by a normal breast tissue profiling [37]. Still, while Normal-like breast cancer carries a good prognosis, it is slightly worse than Luminal A cancer's prognosis.

## 1.2 The Cancer Genome Atlas

### 1.2.1 Purpose and Organisation

The Cancer Genome Atlas (TCGA) network [50], a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), and a part of NCI Genomic Data Commons (GDC) portal from 2016 [51], [52], maintains a public database of clinical and molecular data over 33 different tumour types with hundreds of cases per type, making it the most comprehensive repository of human cancer data [53]. Over the past decade TCGA research network has generated and maintained 2.5 petabytes of data describing tumour and matched normal tissues from more than 11,000 patients that is publicly available and has been used widely by the research community [53]. The data have contributed to more than a thousand studies of cancer by independent researchers and to the TCGA research network publications [54].

The structure of TCGA is well organised and involves several cooperating centres responsible for collection and sample processing, followed by high-throughput sequencing and bioinformatics data analyses [53], [55]. The generated data is made available to the research community through public free-access databases such as NCI GDC data portal, GDC Legacy Archive and the Broad Institute's Firehose [56].
To provide comprehensive analysis of cancer genome profiles, the TCGA research network works with many centres utilising different high-throughput platforms to provide global information of cancer genomics. Some of the applied methods include: RNA-sequencing, miRNA-sequencing, exon sequencing, SNP genotyping, DNA methylation profiling, Reverse-phase protein array (RPPA) [53]. The multidimensional analyses performed on these distinct platforms provide scientists with a better understanding of cancer biology leading to improved cancer classification as well as development of new diagnostic methods and therapeutic approaches.

### 1.2.2 Data Access

Two versions of TCGA data are available: harmonised and legacy. The harmonised data is accessible via the NCI GDC data portal [51], and it represents a subset of the full TGCA data that has been harmonised against GRCh38 (hg38) genome using GDC Bioinformatics Pipelines. The GDC Legacy Archive provides access to an unmodified copy of data that was previously stored in the TCGA Data Portal hosted by the TCGA Data Coordinating Center (DCC), and which uses as references GRCh37 (hg19) and GRCh36 (hg18) [56].

The legacy data is provided as different levels that are defined in terms of a specific combination of both processing level (raw, normalised, integrated) and access level (controlled or open access), while the GDC open access data does not require authorisation to access the high level genomic data that is not individually identifiable [51], [56].

Finally, the data provided by GDC data portal and GDC Legacy Archive can be accessed using R/Bioconductor package TCGAbiolinks [57]. The Bioconductor project ensures high-quality, well-documented and interoperable software and the possibility of integration with hundreds of available packages within R, and is a highly valuable bioinformatics resource [58].
TCGAbiolinks aids in querying, downloading, pre-processing, and analysing TCGA within a single package, allowing user to have a better control over the data and ensuring results reproduciblity [57]. The full clinical report and molecular data (quantified by a variety of methods mentioned above) are prepared to be downloaded as a 'SummarizedExperiment' object [59], which allows easy integration with other data types and statistical methods that are common in the Bioconductor repository. In line with that, the TCGAbiolinks package has a variety of incorporated methodologies for processing and filtering the data.

### 1.2.3 Breast Cancer Dataset

TGCA breast cancer dataset (TCGA-BRCA) is the largest by the number of patients cancer type dataset available in TCGA. One of the most complete breast cancer characterisation studies that has ever been performed is the 2012 TCGA Research Network study [45] that succeeded to identify comprehensive molecular portraits of human breast tumours.

In this study, around 500 primary breast cancers were extensively profiled at the DNA (i.e., methylation, chromosomal copy number changes, and somatic and germ line mutations), RNA (i.e., miRNA and mRNA expressions), and protein (i.e., protein and phosphorprotein expression) levels using the most recent technologies [4], [45].

By classifying tumours using each individual platform and comparing results at different levels through combining the data together in a cluster of clusters, they concluded that diverse genetic and epigenetic alterations converge phenotypically into four major breast tumour subgroups [45]. But also importantly, these four entities were found to be recapitulated very well by the four main intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, and Basal-like) as previously defined by mRNA expression only in the Parker's study [35]. The Normal-like subtype had limited amount of samples, and therefore was not rigorously explored. Overall, these results suggested that intrinsic subtyping captures a great amount of biological diversity that occurs in breast cancer.

Another large-scale TCGA-BRCA study was conducted by Ciriello *et al.* in 2015 [21] on profiling 817 breast cancer samples. The study had a much larger proportion of lobular carcinoma tumours than the original TCGA work, where those were underrepresented. The study shed new light on the genetic bases of lobular morphology and provided more insights into the intrinsic subtypes and their distribution across different morphologies.

## 1.3 Autophagy

### 1.3.1 Process Overview

Autophagy is a highly conserved self-degradative process that has a key role in cellular stress adaptation and survival. Cell homeostasis is achieved by balancing biosynthesis and turnover, and autophagy is the main catabolic process responsible for degrading and recycling cytoplasmic organelles and protein aggregates [60]. The formation of a specialised cargo vesicle for capturing those macromolecules, called autophagosome, is the hallmark of *macroautophagy*, which is the best characterised autophagy pathway and the main focus of this project.

Turnover of cytoplasmic content is an essential cellular house-keeping process, so autophagy happens at any given moment in every cell at low, basal levels [61]. Autophagy is adaptive and tightly-regulated, and it can be rapidly induced by stimulation from various stresses such as nutrient or growth factor deprivation, hypoxia, DNA damage, damaged organelles, or intracellular pathogens [62]. It can be selective or non-selective, depending on the stimulus and cellular context, leading to different outcomes [60]. Non-selective autophagy is primarily a cytoprotective response to the aforementioned stresses. Besides autophagy, the cellular response to stress involves numerous other pathways including those that regulate nutrient uptake, metabolism, cell cycle and growth control. Therefore, it is not surprising that there is a close integration between signals that regulate those processes and autophagy, and also the fact that autophagy is involved in a variety of different physiological and pathological processes, including inflammation, development, energy homeostasis, cancer, cell survival and cell death [61].

Upon starvation or stress conditions, autophagy is upregulated. Damaged organelles and cytoplasm are sequestered by an expanding phagophore, leading to the formation of double-membrane autophagosome [60], as shown in Figure 1.2. The autophagosome subsequently fuses with a lysosome, followed by degradation of the sequestered cargo by resident hydrolases [63]. Degradation allows cells to eliminate damaged or harmful components through catabolism and recycling. Then, the breakdown products are released back in cytosol through permeases, where they can be reused as building blocks to generate energy to maintain cell viability under unfavourable conditions [64].

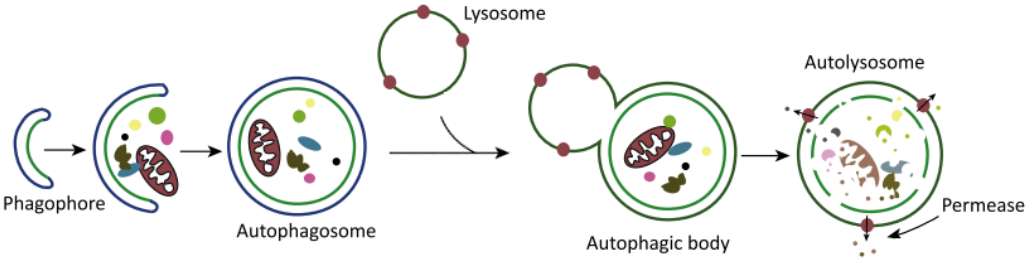

**Figure 1.2:** Overview of autophagosome formation, with primary membrane structures involved in autophagy depicted. The phagophore is the initial sequestering compartment that expands into a double-membrane autophagosome. Autophagosome fuses with lysosome to initiate degradation of the captured cargo (organelles and damaged proteins). The breakdown products are released to cytosol via permeases for reuse. *Image adapted from [60].*

The molecular cascade that regulates and executes autophagy has been the subject of a number of recent, comprehensive reviews [60], [65], [66], [67], but the full mechanism remains not fully understood. In 2016, Nobel Prize in Physiology or Medicine was given to Prof Yoshinori Ohsumi [68], a renowned scientist in the autophagy research field, for his success in elucidating the sophisticated machinery of the autophagy pathway. The work of Ohsumi and the colleagues has led to identification several dozens of autophagy-related genes, allowing for targeted research aimed at understanding of the autophagy mechanisms [60].

In the following sections, the overview of autophagosome biogenesis and a brief summary of the main signalling cascades involved in the regulation of autophagy are presented.

### 1.3.2 The Core Pathway of Autophagy

Autophagy is regulated by a set of genes called AuTophaGy-related genes (ATGs). Among them, the group of genes that is essential for autophagosome formation is referred to as the 'core' molecular machinery [69]. These core ATG proteins form three functional complexes:

- the ULK1/2–Atg13–FIP200–Atg101 complex;
- the Beclin-1–class III phosphatidylinositol 3-kinase (PtdIns3K/PI3K) complex;
- two ubiquitin-like protein conjugation systems (Atg12 and LC3);

and various proteins that mediate fusion between autophagosomes and lysosomes [60], [70].

The autophagy core process is divided into mechanistically distinct steps [71], which include induction, elongation and expansion of the phagophore, cargo recognition and selection, vesicle formation, autophagosome maturation via docking and fusion with lysosome/endosome, and breakdown of cargo followed by release of degradation products to cytosol.

The autophagosome formation takes place at the phagophore assembly site (PAS) [72], to which most of the core ATG proteins are recruited. The main sign of autophagy initiation is the ULK complex formation. ULK1/2 forms a large complex with the Atg13 gene product and the scaffold protein FIP200, Figure 1.3 (abbreviations in the legend). This takes place downstream of the mTOR complex (mammalian target of rapamycin), which is a critical regulator of autophagy induction, whose role is to inhibit autophagy under normal conditions [73].

When mTOR is activated (via Akt and MAPK signalling) it suppresses autophagy, but negative regulation of mTOR (AMPK and p53 signaling) promotes it [70]. Under nutrient-rich conditions, mTOR is associated with the ULK1/2 and phosphorylates them and Atg13; upon starvation, mTOR disassociates from the ULK complex, which then phosphorylates FIP200 and in this way initiates autophagosome formation [74].

Next, in the nucleation stage, the activated ULK complex targets a class III PI3K complex — consisting of Beclin 1, p150, and Atg14 — to promote the local production of a pool of phosphatidylinositol 3-phosphate (PI3P), which is needed for recruitment of protein complexes and lipids to expand autophagosome membrane [67], [75]. Ambra1 mediates ULK dimerisation [76], and its interaction with Beclin 1 is regulated by both ULK1 and Bcl-2.

Finally, in the autophagosome expansion stage (Figure 1.3, bottom left), the Atg12–Atg5–Atg16 complex is recruited to the autophagosome membrane where it facilitates the lipidation of LC3 (also known as MAP1LC3) with phosphatidylethanolamine (PE). The complex formation takes place as follows: Atg12 is conjugated to Atg5 in a ubiquitin-like reaction that requires Atg7 and Atg10. The Atg12-Atg5 conjugate then interacts noncovalently with Atg16 to form a large complex [71]. LC3 is the modified target of the second conjugation pathway. LC3 is cleaved at its C-terminus by Atg4 protease to generate the cytosolic LC3-I. LC3-I is conjugated to PE also in a ubiquitin-like reaction that requires Atg7 and Atg3 [71]. The lipidated form of LC3, known as LC3-II, is selectively attached to the forming autophagosome membrane with the help of p62/SQSTM1. SQSTM1 is a scaffolding protein that is expressed and turned-over by autophagy-induced selective degradation [75]. All in all, the lipidation of LC3 is widely used to monitor autophagy induction.

Selective autophagy, which is a targeted engulfment of specific cargos marked with degradation signals, is becoming increasingly appreciated [67]. The most studied example is *mitophagy* (Figure 1.3, top right) – a process specifically responsible for the removal of dysfunctional mitochondria from the cell. Upon mitochondrial damage, the protein PINK, which is continually degraded by PARL under normal conditions, is stabilised and recruits the E3 ligase Parkin (also known as PARK2) to initiate mitophagy [62]. Polyubiquitination of mitochondrial membrane proteins by Parkin results in the recruitment of *autophagy adaptor proteins* SQSTM1/p62, NBR1, and Ambra1. These are the proteins that mediate the targeting of autophagosomes to cargo (in this example, mitochondria) [67]. The targeting happens via binding to LC3 via their LC3-interacting region (LIR) and ubiquitin. In addition, BNIP3 and BNIP3L/NIX, which also contain LIRs, directly recruit autophagic machinery by a ubiquitin-independent mechanism to induce autophagosome formation in certain cell types [62].

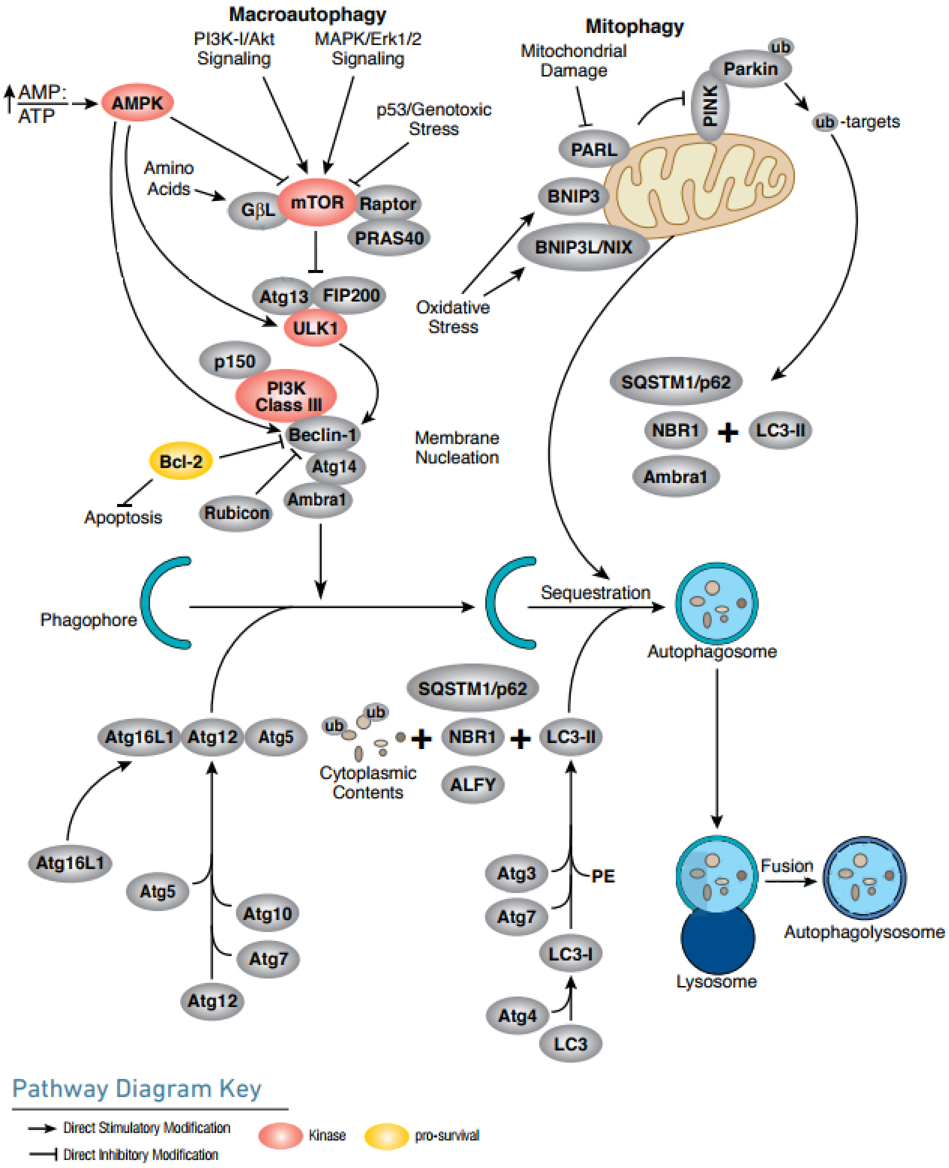

**Figure 1.3:** The core pathway of autophagy regulation. Autophagy is regulated by a complex signaling network of various stimulatory (arrowheads) and inhibitory (bars) inputs. Details in the main text. *Abbreviations:* **PE** – phosphatidylethanolamine, **MAP1LC3/LC3** – microtubule-associated protein 1 light chain 3, **Ambra1** – activating molecule in Beclin 1-regulated autophagy 1, **Bcl-2** – B cell lymphoma 2, **FIP200** – FAK family kinase interacting protein of 200 kDa, **mTOR** - mammalian target of rapamycin, **SQSTM1/p62** – sequestosome 1, **ULK** – unc-51-like kinase, **NBR1** – next to BRCA1 gene 1 protein, **AMPK** – AMP-activated protein kinase, **PINK** – PTEN induced putative kinase, **BNIP3L/NIX** – Bcl-2 adenovirus E1a nineteen kDa interacting protein 3, **PI3K** – phosphatidylinositol 3-kinase. *Image taken from [77].*

### 1.3.3 Signalling Pathways Regulating Autophagy

As established in the previous section, autophagy initiation is ultimately controlled by mTOR. Its inactivation in response to nutrient depletion results in the activation of the ULK complex and thus the induction of autophagy. The mTOR signalling pathway is critical because of its ability to integrate the information from nutrient, metabolic, and hormonal signals [74]. However, hypoxic stimuli can induce autophagy through an mTOR-independent pathway.
This section briefly summarises the mTOR-dependent and independent pathways of autophagy initiation.

Figure 1.4 presents the signalling regulation of mammalian autophagy. The two main signalling pathways, the class I PtdIns3K signalling and amino acid-dependent signalling, are integrated and maintained via mTOR.

Insulin and IGF1 (insulin-like growth factor 1) have been shown to inhibit autophagy via mTOR acivation. Activation of the insulin/growth factor receptor (yellow in Figure 1.4) and its adaptors, IRS1/2, stimulate the PtdIns3K complex and small GTPase RAS, leading to activation of the PtdIns3K–PKB–TOR pathway [70]. PKD1 activates proto-oncogene PKB (Akt), which then phosphorylates and inhibits the TSC1–TSC2 complex, leading to the stabilisation of Rheb, which in turn activates mTOR, leading to autophagy inhibition [78], [79].

Amino acids inhibit the Raf-1–MEK1/2–ERK1/2 signalling cascade, causing autophagy inhibition (i.e. nutrient-rich environment). Low cellular energy or stress causes AMPK to be phosphorylated by LKB1, activativating it. AMPK phosphorylates and activates TSC1–TSC2, leading to inactivation of mTOR and hence, autophagy induction [62]. p70S6K kinase is a substrate of mTOR that may negatively feed back on TOR activity, ensuring basal levels of autophagy that are important for homeostasis [79].
Finally, JNK1 and DAPK phosphorylate and disrupt the association of anti-apoptotic proteins, Bcl-2 and Bcl-XL with Beclin 1, which leads to the activation of the Beclin 1-associated class III PtdIns3K complex and mTOR-independent (and Beclin-1-dependent) stimulation of autophagy [80].

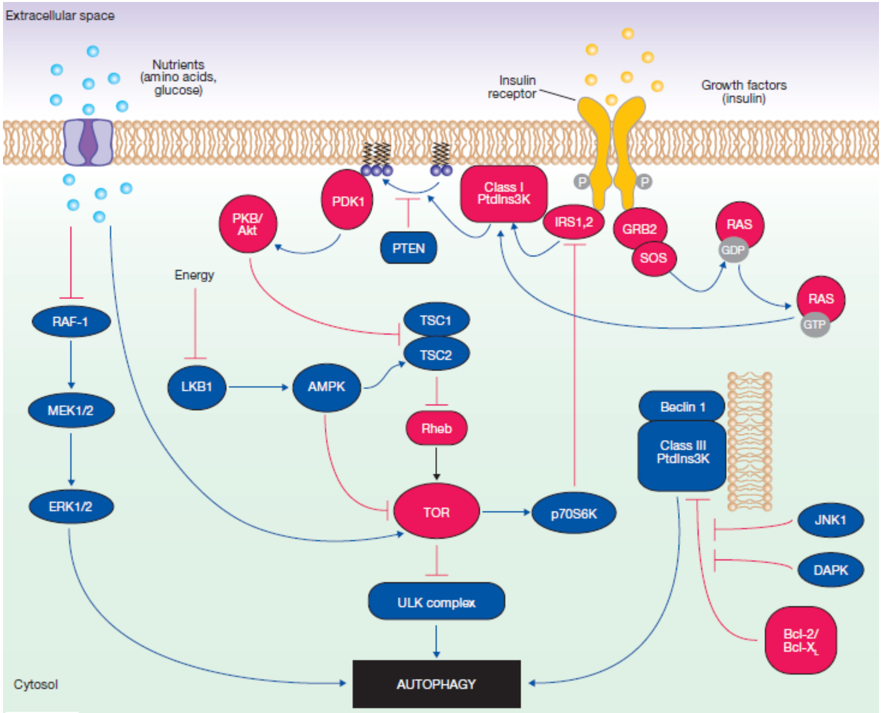

**Figure 1.4:** Signalling pathways regulating autophagy. Autophagy is induced by deprivation of nutrients (amino acids), hormones, energy. In the figure, the components coloured blue represent the factors that stimulate autophagy, whereas the red ones correspond to inhibitory factors. The signalling network is regulated by stimulatory (blue arrows) and inhibitory (red bars) inputs. Details in main text. *Abbreviations:* **PKB** – protein kinase B, **PDK1** – 3-phosphoinositide-dependent protein kinase 1, **JNK1** – c-Jun amino-terminal kinase 1, **DAPK** – death-associated protein kinase, **AMPK** – AMP-activated protein kinase, **TCS1/2** –tuberous sclerosis complex 1/2, **IRS1/2** – insulin receptor substrate, **LKB1** – liver kinase B1, **PtdIns3K/PI3K** – phosphatidylinositol 3-kinase complex. *Image taken from [79].*

### 1.3.4 Autophagy in Cancer

Tight control and fine-tuning of molecular signals is crucial for fidelity of functional autophagy, as deregulation at any of the above mentioned steps could result in too much or too little autophagy. Dysfunctional autophagy has been linked to a variety of diseases, including cancer [75].

Impaired autophagy compromises the ability of a cell to survive stressful conditions, which results in cell death. Chronic cell death leads to a prolonged inflammatory response which may be oncogenic. In fact, cancer-related inflammation is one of the established cancer hallmarks [81]. Moreover, autophagy also limits genotoxic damage by reducing the formation of reactive oxygen species (ROS) and clearing damaged mitochondria [82]. When autophagy is impaired, damaged mitochondria accumulate in the cell, increasing ROS production, and leading to protein, organelle, and DNA damage. This promotes tumourigenesis, supporting the idea that inadequate autophagy is a contributor to cancer development [83].

The role of autophagy in cancer is complicated and context-dependent, and presumed to differ between stages of cancer progression [84]. At the initial stages of cancer progression, autophagy is believed to have a tumour suppression function, as already described above. It facilitates turnover of damaged proteins and organelles, thereby limiting inflammation, tissue damage, and genome instability, which are the known promoters of cancer initiation [85]. In this way, it has been suggested that autophagy induction is beneficial for cancer prevention.
On the other hand, at later stages of cancer progression, autophagy gains a tumour-promoting function, as cancer cells begin to use it for their own cytoprotection [84]. Induction of autophagy allows cancer cells to survive under conditions of metabolic and genotoxic stress, such as hypoxia and acidity in tumour microenvironments [86]. Furthermore, autophagy also provides cancer cells with resistance to chemotherapy and radiotherapy, by allowing them to survive the hostile conditions created by those treatments [87].
Recent work has also begun to explore function of autophagy in cancer metastasis [87], [88]. Upregulation of autophagy at the later stages of cancer progression is observed outside of the primary tumour, in fact, Mowers *et al.* (2017) reported that a role of autophagy at nearly every phase of metastatic cascade has been identified, including in such essential processes as cell motility and invasion, cancer cell differentiation, resistance to anoikis, and tumour dormancy [88].

Because of the fundamental importance of autophagy in the development and progression of cancer and its capacity to affect treatment response, there has been an explosion of research on the molecular regulation and signalling pathways that control autophagy [87].

The high levels of autophagy are observed in tumour cells following essentially every anti-cancer treatment [89], which has motivated a great interest in combining autophagy inhibition with other established therapies in breast cancer to collectively eliminate cancer cells and improve clinical outcome [90]. For instance, recent functional studies indicated that autophagy inhibition, which was achieved either by pharmacological means or RNAi-mediated silencing of ATGs, sensitises hormone receptor (ER) positive breast cancer cells to tamoxifen, thereby promoting cytotoxicity and preventing the development of anti-estrogen resistance [90], [91]. In this way, autophagy inhibition may be useful as a combination strategy in this subset of breast cancer patients. Similarly, in various breast cancer cell culture models, autophagy inhibition appears to decrease the resistance of HER2+ positive breast cancer cells to the anti-HER2 monoclonal antibody trastuzimab [92].

Such studies undoubtedly hold promise, but given the tumour suppressing functions of autophagy, caution should be taken in interpreting the effects of these inhibitors and rapidly translating autophagy inhibitors as an all-purpose treatment for breast cancer. Since autophagy is important for normal tissues, the critical task is make autophagy inactivation sufficiently selective to impair cancer growth while sparing normal tissues from the deleterious consequences [93].

Thus, identification of the autophagy marker genes and signalling pathways that support cellular survival under oncogenic or microenvironmetal stress is of great importance. This will help to determine the best means to therapeutically inhibit autophagy for cancer treatmant, and to determine which patient subgroups would most benefit from this approach [93].

## 1.4 Aims and Objectives

The aim of this project is to identify autophagy signatures in breast cancer gene expression data collected by the TCGA Research Network. Currently, the research community is still in a very preliminary phase of understanding the intertwined relationship of autophagy and cancer. Hence, the project research objectives are:

- TCGA-BRCA dataset will be thoroughly explored in order to get a good understanding of its structure and the variety of available samples. The known breast cancer sample classification methods will be applied to the data and the resulting groups and their relationships will be evaluated.
- The breast cancer classification methods will be used as a guide for selecting groups of samples to perform differential expression analysis on. Differential expression analysis will help to identify subsets of deregulated genes to test for presence of autophagy signatures in them. In this way, groups of breast cancer samples with apparent autophagy signature will be identified.
- The data will also be clustered according to the patterns of gene expression change across cancer stages. This will display the major gene expression trends in cancer subtypes and will help to identify expression patterns which are enriched in co-expressed autophagy genes.
- Finally, a set of autophagy genes with a potential signature in breast cancer will be settled and explored for presence of biological interpretation. The candidate genes identified by this project will be made available to the collaborators in the Unit of Cell Stress and Survival at DCRC for experimental validation in cancer cell lines.

# Chapter 2

# Methods

## 2.1 Data

The full TCGA-BRCA RNA-sequencing dataset was downloaded from the NCI's GDC data portal [51] using R/Bioconductor package TCGAbiolinks v2.8.13 [57]. The overview of available breast cancer RNA-seq samples collected by the TCGA Research Network is displayed in Table 2.1.

**Table 2.1:** Overview of RNA-Seq samples in the TCGA-BRCA dataset. *Top row:* all available samples, *Bottom row:* samples used in the project.

| | **Samples** | **Tumour** | **Normal** | **Metastasis** |
|---|---|---|---|---|
| **Full dataset** | 1212 | 1093<br>(F: 1081, M: 12) | 112<br>(F: 112, M :0) | 7<br>(F: 7, M: 0) |
| **Final dataset** | 969 | F:857 | F:112 | |

The full dataset was reduced to include samples that matched three criteria: **i)** samples from female patients only (to reduce biological variation coming from gender) **ii)** samples that have been manually curated for their classifications **iii)** samples that are provided with sufficient metadata to benefit exploratory analysis. Only primary tumour and normal samples were included in the analysis.

### 2.1.1 Samples Annotation

Morphological and stage annotation of samples included in the analysis was curated by the members of CBL group and the collaborators from other groups at DCRC. Table 2.2 shows the number of samples that were represented in the morphological groups (right) and stages (left) in the final dataset.

**Table 2.2:** The number of samples in each morphology type and stage in the final dataset. 'Stage X' is unknown/unidentifiable stage. 'Other morphologies' comprises types that are represented by only a few samples.

| **Morphology** | *ICD-O-3 code* | |
|---|---|---|
| Invasive lobular carcinoma | 8520/3 | 143 |
| Invasive ductal carcinoma | 8500/3 | 644 |
| Ductal and lobular *in situ* carcinoma | 8522/3 | 24 |
| Metaplastic carcinoma | 8575/3 | 7 |
| Mucinous carcinoma | 8480/3 | 12 |
| Other morphologies | | 27 |

ICD-O-3, International Classification of Diseases for Oncology 3rd edition

| **Stage** | |
|---|---|
| stage 1 | 148 |
| stage 2 | 481 |
| stage 3 | 195 |
| stage 4 | 12 |
| stage X | 21 |

All samples in the final dataset were annotated with clinical data, which included PAM50 molecular subtype, patient age subgroup, race/ethnicity, menopause status, tumour size/grade, nodal involvement, metastasis status, year sample was taken, and tissue source site. The annotations extracted with TCGAbiolinks were integrated with further information from the work by Rahman *et al.* [94] on reprocessing TCGA data. The original set of PAM50 subtype annotations available with TCGA data that was obtained by a large TCGA-BRCA study in 2012 [45], was further complemented with the additional subtype labels for the previously unclassified samples from a recent TCGA-BRCA study by Ciriello *et al.* in 2015 [21]. A small minority of samples that had discrepancies in labels between the two studies were omitted in this project. Table 2.3 shows the sample counts for each PAM50 subtype.

**Table 2.3:** The number of samples in each PAM50 molecular subtype in the final dataset.

**PAM50**

| | |
|---|---|
| Luminal A | 420 |
| Luminal B | 183 |
| Basal-like | 153 |
| HER2-enriched | 75 |
| Normal-like | 26 |

### 2.1.2 Genes Annotation

A curated collection of autophagy-related gene lists was provided by experts in the field at DCRC (Cell Death and Metabolism and Cell Stress and Survival Units). Table 2.4 shows the functional groups that autophagy-related genes are managed by. The autophagy core genes and transcription factors are of the most interest in terms of molecular signatures.

**Table 2.4:** Functional groups of autophagy-related genes. The numbers are reported for the dataset after preprocessing with TCGAbiolinks. Some genes are present in more that one group.

| **Functional Group** | **Number of genes** |
|---|---|
| Autophagy core | 156 |
| Transcription factors | 101 |
| Lipid-related processes | 33 |
| Phosphatidylinositol | 40 |
| Endo- and exosomes | 132 |
| Transport | 216 |
| RABs and effectors | 131 |
| Docking and fusion | 14 |
| Mitophagy | 65 |
| Receptors and ligands | 66 |
| mTOR induction | 138 |
| Lysogenesis induction | 62 |
| Lysosome | 218 |
| *Total* | *1112* |

### 2.1.3 Extraction and Preparation

The TCGA-BRCA RNA-seq data was generated using *Illumina HiSeq 2000 RNA Sequencing Version 2 analysis* platform and quantified at University of North Carolina (UNC) Center for Bioinformatics for the TCGA project [95]. The quantification pipeline included using Mapsplice v12.07 [96] for mapping the data to reference genome (GRCh37/hg19), RSEM v1.1.13 [97] for transcript quantification [95].

Legacy TCGA-BRCA gene expression dataset was downloaded from the GDC portal and prepared using the TCGAbiolinks preprocessing pipeline. The pipeline contains integrated functions from the EDASeq package [98] for within-lane normalisation procedures to adjust for GC-content and gene length effects on read counts, as well as between-lane normalisation method to adjust for distributional differences between lanes (e.g. sequencing depth), such as quantile normalisation [56], [57]. In this project, the dataset was normalised for GC-content and filtered with with quantile cut-off of 0.10. The pipeline transforms the data into a '*SummarizedExperiment*' [59] object (counts table), with genes and samples as rows and columns, respectively.

After data pre-processing and selecting the samples with sufficient clinical information, the final dataset included 857 tumour and 112 normal samples (969 in total), and the gene expression matrix was reduced to 17372 genes. Figure 2.1 summarises the data flow from the original dataset thorough to the downstream analysis. The next section will describe the further processing step applied to the data prior to analysis.

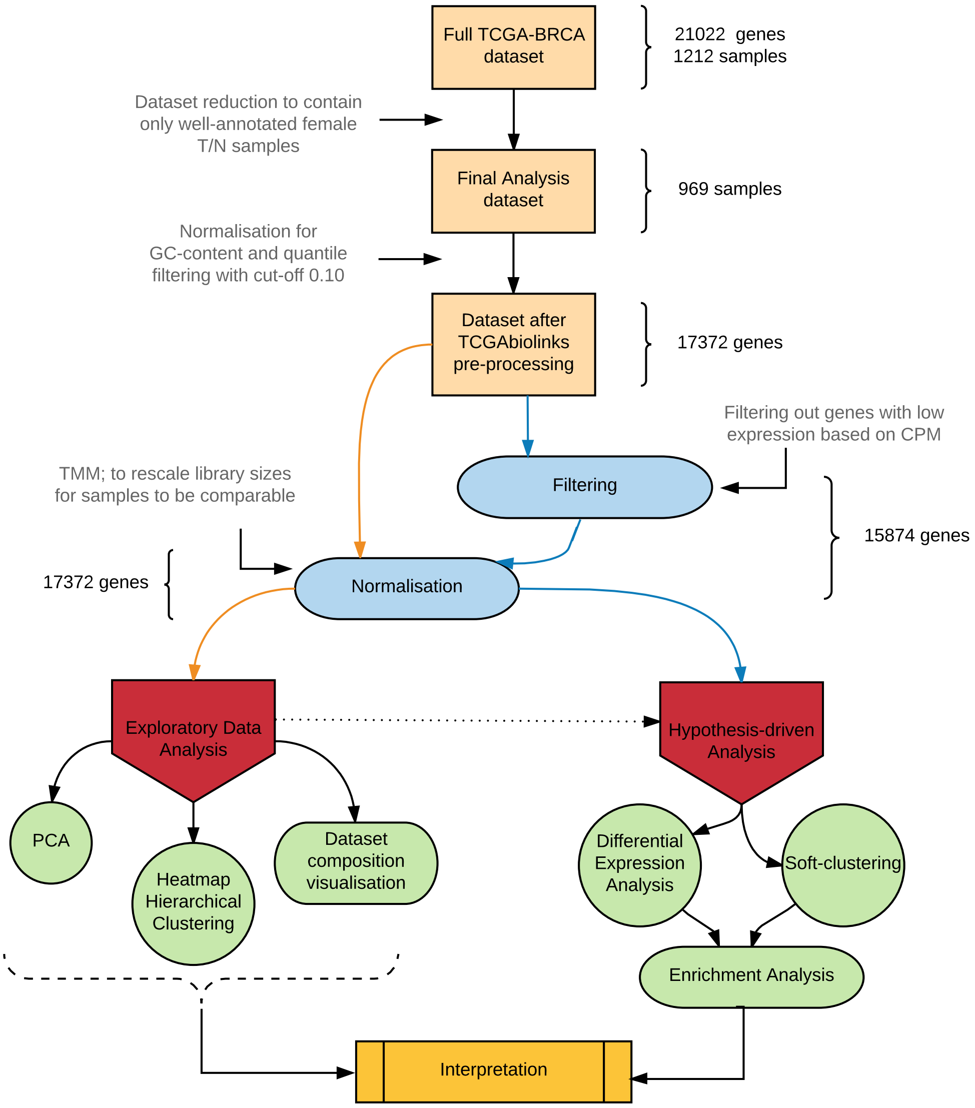
Full TCGA-BRCA dataset
21022 genes
1212 samples
Dataset reduction to contain only well-annotated female T/N samples
Final Analysis dataset
969 samples
Normalisation for GC-content and quantile filtering with cut-off 0.10
Dataset after TCGAbiolinks pre-processing
17372 genes
Filtering out genes with low expression based on CPM
Filtering
15874 genes
TMM; to rescale library sizes for samples to be comparable
17372 genes
Normalisation
Exploratory Data Analysis
Hypothesis-driven Analysis
PCA
Heatmap Hierarchical Clustering
Dataset composition visualisation
Differential Expression Analysis
Soft-clustering
Enrichment Analysis
Interpretation

**Figure 2.1:** Project workflow overview. Several pre-processing and filtering steps reduce the number of samples and genes in the original dataset prior to exploratory and hypothesis-driven data analyses. The data is normalised so that samples can be more appropriately compared. The results of exploratory analysis guide the hypothesis-driven analysis. The observations from both are used for biological interpretation.

### 2.1.4 Filtering and Normalisation

Throughout different parts of the analysis workflow the counts data is stored in a simple list-based data object *DGEList* from R package `edgeR` [99], which also contains sample annotation and library size information.

For differential expression and related analyses, gene expression is not considered as raw counts, because library sizes (total number of mapped reads) differ and the counts data is incomparable. Therefore, it is a common practice to transform raw counts onto a scale that accounts for such library size differences [100]. Popular transformations include (log2-)counts per million (log-CPM/CPM), and reads/fragments per kilobase of transcript per million (RPKM/FPKM). CPM is a simple measure of read abundance that can be compared across libraries of different sizes, i.e. it is relative change in expression [101]. Standardising further by gene length gives RPKM, which is an absolute expression. Data in the form of CPM, log-CPM, and RPKM will be used for different parts of analysis in this project.

Prior to the analysis it is important to filter out genes with very low counts across all libraries, as they provide more noise than evidence, e.g. for differential expression. Genes that are not expressed at a biologically meaningful level in any condition should be discarded to reduce the subset of genes to those that are of interest, and to minimise the number of tests carried out downstream. In this project, filtering was done based on CPM to account for library sizes. The selected cut-off for keeping the genes was guided by the mean-variance plot produced by `voom` function (discussed in Section 2.3.1.2). The genes with at least 2 CPM expression in at least 19 samples passed the threshold and were used in the analysis.

After filtering, the samples have to be normalised. Normalisation adjusts global properties of measurements for individual samples so that they can be more appropriately compared. In this way, it is ensured that the expression distribution of samples is similar across the entire experiment [102]. In this project, normalisation by the method of trimmed mean of M-values (TMM) [103] was performed using the `calcNormFactors` function in `edgeR` to rescale the library sizes to take into account differences in RNA-composition. As a result, the effective library size replaces the original library size in all downstream analyses.

However, it is important to note that normalisation does not remove batch effects, which can affect subsets of genes and/or samples in different and unpredictable ways. The batch effects will be taken into account at later steps.

## 2.2 Exploratory analysis methods

Exploratory data analysis (EDA) is an essential step in working with large publicly available datasets, such as the TCGA-BRCA in this project. Application of exploratory analysis to transcriptomics data can be a means of visualising the global structure of the data, and also serve three major roles:

1. Discover patterns and spot outliers/abnormalities
2. Frame the hypothesis
3. Check assumptions

After normalisation described in the previous section, the data is ready to be analysed. Rigorous exploration of the metadata available for samples will be used to maximise the insight into the dataset and extract important features.

Principal component analysis and clustering (as part of a heatmap and not) are the most commonly used exploratory tools. The following sections will present a general introduction to PCA and clustering, and provide simple examples of use. The underlying statistics and algorithms available for the calculations involved in PCA and clustering dendogram generation are fundamentally the same for the available packages in R/Bioconductor.

### 2.2.1 Principal Component Analysis

Principal component analysis (PCA) is technique that is often used to emphasise variation and bring out strong patterns in a dataset. This method linearly transforms a multivariate dataset into a set of uncorrelated variables, principal components (PCs), ordered by the amount of variance explained [104]. In this way, the first few PCs explain the largest amount of the variation in the data. PCA results can be visualised in a 2D scatter plot, where $x$ and $y$ axes are the selected PCs. The samples are projected onto the 2D plane such that they spread out in the two directions that capture the most of the variance across samples [105]. In a PCA 2D scatter plot, each data point represents a sample, which allows visualisation of sample clustering and dataset structure. The relationship between two samples is reflected by the distance between corresponding dots in the plot. Therefore, the more similar gene expression profiles are, the closer the data points are.

Figure 2.2 shows an example of separation of transcriptional profiles of cancer (pink) and normal (blue) samples. The primary source of variation (PC1) accounts for 11% of the total variation in the data. The second principal component (PC2) accounts for 8.6% of the variation. Here, and throughout the project, PCA is performed using the R function `prcomp()`, the plots are generated with R package `ggplot2` [106].

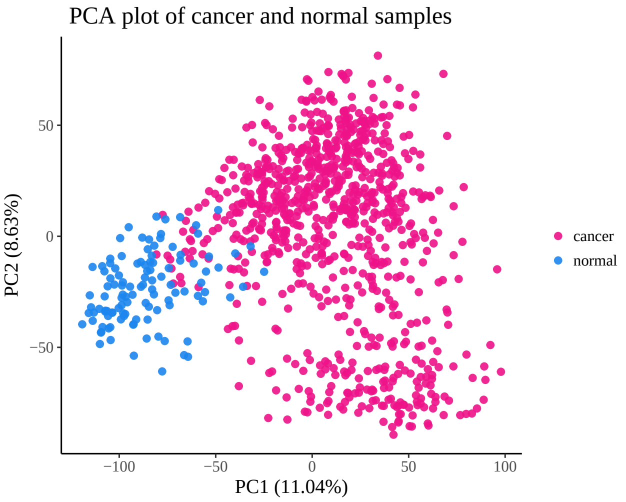

**Figure 2.2:** An example of PCA 2D scatter plot, showing the variance along PC1 and PC2 for cancer and normal samples.

Another way of exploring dataset variation characterised by PCA is to visualise variation captured by each principal component in a series of one-dimensional (1D) box plots. Figure 2.3 shows the variation seen in each PC (1-9) for the cancer/normal dataset that was shown previously as a scatter plot of first two PCs (Figure 2.2).
In contrast to 2D scatter plots, 1D PCA plots are able to show variation along more than two PCs at a time. For each PC, the variation of each condition group, here - cancer and normal, is represented by a box plot. The PCs where condition boxes have the smallest overlap (e.g. PC1) will shows the clearest separation when plotted in 2D. The PCA results were handled with `dplyr` R package [107] to create 1D box plots.

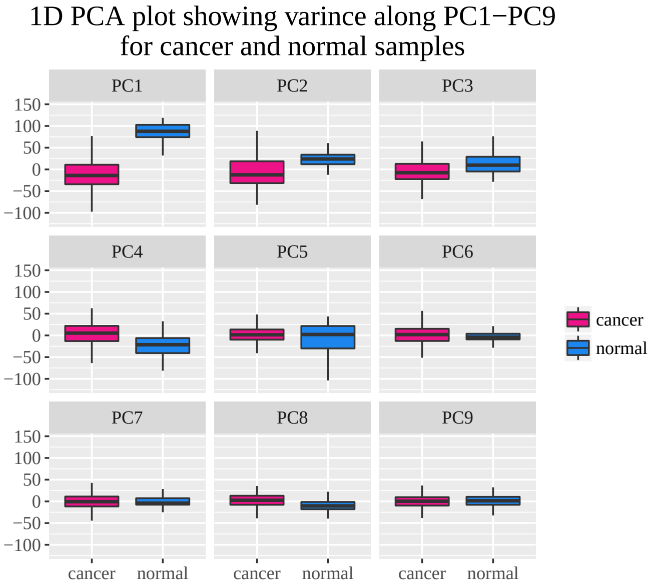

**Figure 2.3:** An example of one-dimensional PCA plots for the cancer/normal dataset.

The 1D PCA method is extremely useful for checking for presence of potential batch effects or signatures left by other cofactors in any of the PCs. Batch effects are a common and powerful source of variation in high-throughput biology. They are artifacts not related to the biological variation of scientific interests, and may substantially affect the downstream analysis results if not dealt with properly [108]. For example, batch effects may occur if an experiment was run on two different days, or two different lots of reagents or instruments were used. Principal components can capture both biological and technical variability and therefore can be used to quantify the effects of such artifacts on the data [109].

### 2.2.2 Clustering and Heatmap representation

In clustering, or unsupervised classification, the aim is to identify subsets (clusters) of data points based on the similarity between single objects. Similar objects should be assigned to the same cluster, while objects which are not similar to each other, should be assigned to different clusters. This can help to reveal the data structure and give first insights into the data, which is especially useful if prior knowledge is little or non-existent. Clustering can, therefore, be seen as an exploratory data analysis tool.

The purpose of clustering transcriptomics data is to statistically group samples according to their gene expression, in order to reduce complexity and dimensionality of the data, predict function or identify shared regulatory mechanisms [110]. Clustering can be performed as a part of heatmap. Heatmap is a data matrix visualising values in the cells by the use of a colour gradient. This gives a good overview of the largest and smallest values in the matrix. Rows (genes) and/or columns (samples) of the matrix are clustered to facilitate interpretation of sets of rows or columns rather than individual ones [110].

One of the popular methods is hierarchical clustering, which involves re-ordering of samples based on their distance in high dimensional space. The cluster is constructed based on the determination of two parameters — the distance metric and the linkage criterion. Objects close to each other in the hierarchy, measured by tracing the branch heights, are also close by some measure of distance — for example, individuals with similar expression profiles will be close together in terms of branch lengths.

The two most common distance measures used for clustering are Euclidean and Manhattan distances. Typically the results of the two are quite similar, and most studies default to using the Euclidean measure.

The Euclidean distance involves computing the square root of square differences between two coordinates. In this way, the shortest path diagonally is calculated:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

where the first point is $(x_1, y_1)$ and the second point is $(x_2, y_2)$.

An example of unsupervised clustering of samples with heatmap is shown in Figure 2.4. The data is clustered by columns (samples) and rows (genes), with dendograms showing how clusters are formed. The colour bar above the heatmap shows cancer/normal samples as pink/blue vertical lines, and clustering forms blocks of colour. The heatmap colours represent gene expression intensity according to the scale (high expression - dark blue, low expression - light yellow). This example shows how cancer and normal samples form two major clusters in the dendogram, which is noticeable in the heatmap colouring as well. Some of the samples are not within the expected clusters, i.e. are outliers, but that is in agreement with the observed slight overlap between cancer and normal samples clusters on the PCA plot (Figure 2.2).

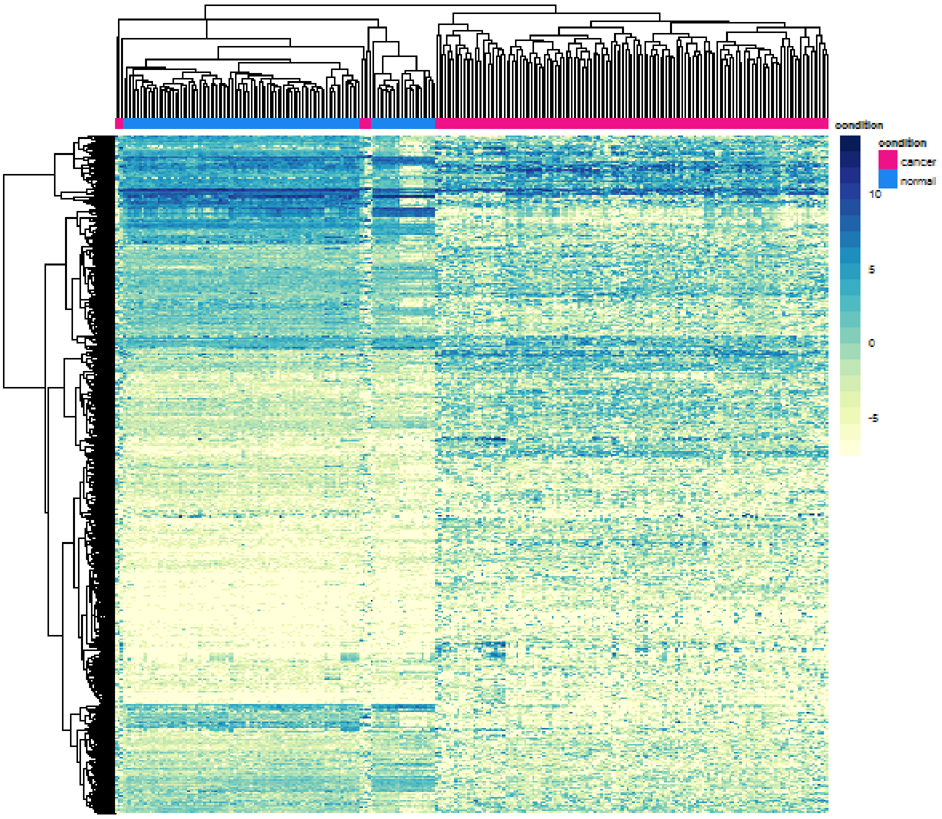

**Figure 2.4:** An example of clustering of the cancer/normal samples with heatmap representation. The data was clustered with Euclidean distance and average linkage.

Prior to clustering the data is normalised and filtered for low expression as described in Section 2.1.4. The heatmaps were generated using Euclidean distance with average linkage (default) using the R package `pheatmap` [111]. Clustering without heatmap representation was done with the R function `hclust()` with default parameters.

## 2.3 Hypothesis-driven analysis methods

### 2.3.1 Differential Expression Testing

#### 2.3.1.1 Approach Motivation and Background

In RNA-seq, mapped reads are aggregated to counts at different levels, such as transcripts, exons, or genes. The counts for a given gene quantify the expression of that gene. The main goal is to find which genes have different levels of expression under different experimental conditions — this is known as differential expression. The detection of differentially expressed genes between two or more conditions is one of the most commonly asked questions in biological research, and RNA-sequencing is currently the primary technology used for gene expression profiling.

There are several tools within Bioconductor that have been developed for the differential analysis of count data, including `DESeq` [112], `DESeq2` [113], `edgeR` [99], `limma (voom)` [114].

There are two main goals behind using differential expression analysis tools:

1. Estimate the magnitude of differential expression between two or more conditions based on read counts from replicated samples, i.e. calculate the fold change of read counts, taking into account the differences in sequencing depth and variability.
2. Estimate the significance of the difference and correct for multiple testing.

#### 2.3.1.2 Limma-voom

*limma* is an R/Bioconductor package [114], [115] that provides an integrated set of tools for differential expression analysis. The *limma* pipeline includes linear modeling with extensive features for handling complex experimental designs with multiple treatment factors, and empirical Bayes statistical methods to borrow strength between genes to overcome the problem of small sample sizes [114].

The *limma* approach estimates the mean-variance relationship of gene expression data in log-CPM (log-counts normalised for sequence depth) [101]. The `voom()` method (an acronym for 'variance modeling at the observational level') incorporates the mean-variance trend into a precision weight for each individual normalised observation. The normalised log-counts and associated precision weights are then passed on to the *limma* empirical Bayes analysis pipeline [101].

Simulation studies show that the *limma-voom* performs as well or better than other count-based methods even when the data are generated according to the assumptions of the earlier methods [101]. A key advantage of the *limma* pipeline is that it provides accurate type I error rate control even when the number of samples in the analysis is small [116], which is a problem for other methods. Additionally, *voom* can robustly handle heterogeneous data with outliers and hypervariable genes [117], and also the analysis run-time is fast, which makes it the most suitable tool for differential expression testing in this project.

#### 2.3.1.3 Linear Models

The hallmark of the *limma* approach is the use of gene-wise linear models to analyse entire experiments as an integrated whole rather than simple comparisons between pairs of treatments [114], therefore allowing more flexible hypotheses. The effect of sharing information between samples allows one to model correlations that may exist between samples due to replication of samples or presence of covariates. This kind of information would not be accessible if the data was partitioned into subsets and analysed as a series of pairwise comparisons [105]. In this way, linear models permit very general analyses, in which it is possible to adjust for the effects of multiple experimental factors or batch effects [114].

In the linear model approach, two matrices have to be specified. The first is the *design matrix* which provides a representation of the experimental design, i.e. each column corresponds to a coefficient that describes the sample source. The second is the *contrast matrix* which specifies which comparisons have to be made between the samples. In this way, linear models describe how the treatment factors are assigned to the different samples, in matrix terms:

$$E(y_g) = X\beta_g$$

where $y_g$ is the vector of log-CPM values (expression data) for a gene $g$, and $X$ is the design matrix with the $x_i$ as rows, and $\beta_g$ is a vector of coefficients. The contrasts of interest are given by

$$\alpha_j = C^T\beta_g$$

where $C$ is the contrast matrix. The coefficients component of the fitted model produced by *limma* function `lmFit()` contains estimated values for the $\beta_g$. After applying `contrasts.fit()`, the coefficients component now contains estimated values for the $\alpha_j$ [118].

#### 2.3.1.4 Significance Testing

The number of differentially expressed (i.e. up- and downregulated) genes for each contrast (i.e. condition comparison) can be obtained after the linear model has been fit. Significance of differential expression is defined using an adjusted p-value cut-off and a log fold change (logFC) threshold. Fold change shows how much a gene's expression level has changed between two conditions. This value is reported on a logarithmic scale to base 2: for example, a $log2$ fold change of 1.5 means that the gene's expression is increased by a multiplicative factor of $2^{1.5} \approx 2.82$.

The results of fitting a linear model are summarised and subjected to hypothesis testing to see whether there is enough evidence to reject the null hypothesis that there is zero effect of the condition on the gene and that the observed difference between the conditions is due to experimental variability (i.e. variability expected to be between different samples in the same treatment group) [105].

The results of hypothesis testing is reported as a p-value. A p-value indicates the probability that a fold change as strong as the observed one, or even stronger, would be seen under the situation described by the null hypothesis [105]. For analyses comprising a large number of hypothesis tests, as it is in differential expression studies, a large number of inferences may occur by chance, leading to falsely significant results [119]. As the hypothesis is tested for every comparison, the p-values need to be corrected/adjusted for multiple testing. The most popular form of adjustment is 'FDR' which is Benjamini and Hochberg's method to control the false discovery rate [120]. The adjusted p-values provide a useful metric for assessing which genes to target for further analysis.

#### 2.3.1.5 Application to the dataset

Limma-voom was used to perform differential expression (DE) analysis on the final dataset of TCGA-BRCA RNA-seq samples. The differential expression was quantified between the subgroups of the three main classification methods/effects (PAM50, stage, morphology) and the normal samples. For each main effect, a separate design matrix was created and used to fit the model. All three effects were not included into one model together, as this would have resulted in non-estimable coefficients due to linear dependency of the effects. Each of the separate models included other cofactors and batch effect terms.

However, as the combination of subgroups of the main classification methods can have a unique effect in terms of differential expression, additional models with each existing pair of subgroups as factors were created. In this way it was possible to test DE not only between stages, but also between stages of a particular PAM50 subtype or morphology.

Differential expression results from the tested models used were filtered with significance threshold of logFC $> 1$ and adjusted p-value $< 0.05$. The genes declared as significantly differentially expressed were used in the enrichment analysis.

## 2.3.2 Gene Expression Clustering

### 2.3.2.1 Approach Motivation and Background

Clustering, or unsupervised classification, is a powerful tool in gene expression data analysis, as it can help to reveal the structured expression patterns hidden in high-dimensional gene expression datasets [121], as already discussed in Section 2.2.2 in exploratory context. Genes showing similar expression patterns (*co-expressed* genes) are often functionally related and controlled by the same regulatory mechanisms (*co-regulated* genes), resulting in expression clusters frequently being enriched by genes of certain functions. This makes clustering an attractive method for searching for autophagy signatures.

There are many clustering methods that can be applied to gene expression data, such as k-means clustering or the aforementioned hierarchical clustering. These common methods, however, produce hard partitions of the data, i.e. genes are assigned to exactly one cluster even if their expression profile is similar to several cluster patterns. For experiments where the change of expression over time is of interest, this might not be the best approach. It is known that regulation of genes is generally not in an 'on-off' fashion, but rather is a gradual change which allows a more refined control of the genes' functions [121]. Therefore, clustering should ideally take into account this complexity and produce more flexible clusters by allowing gene assignment to several clusters.

This type of clustering method is termed *soft-clustering*, as it does not create hard boundaries between the clusters.

### 2.3.2.2 Soft-clustering

The soft-clustering algorithm used in this project is implemented in R package `mfuzz` [121] using the fuzzy c-means algorithm (of the `e1071` package) based on the iterative optimization of an objective function to minimize the variation of objects within clusters [122].

Fuzzy clustering differentiates how closely a gene follows the dominant cluster patterns and assigns it degrees of membership to a cluster. The membership value $\mu_{ij}$ can vary between zero and one, and is an indication of how well gene $i$ is represented by cluster $j$. This approach strongly contrasts hard-clustering where membership is binary (i.e. 0 or 1), and enables fuzzy clustering to provide more information about the structure of gene expression data [121]. The formed clusters are visualised by the mfuzz plotting function, an example shown in Figure 2.5.

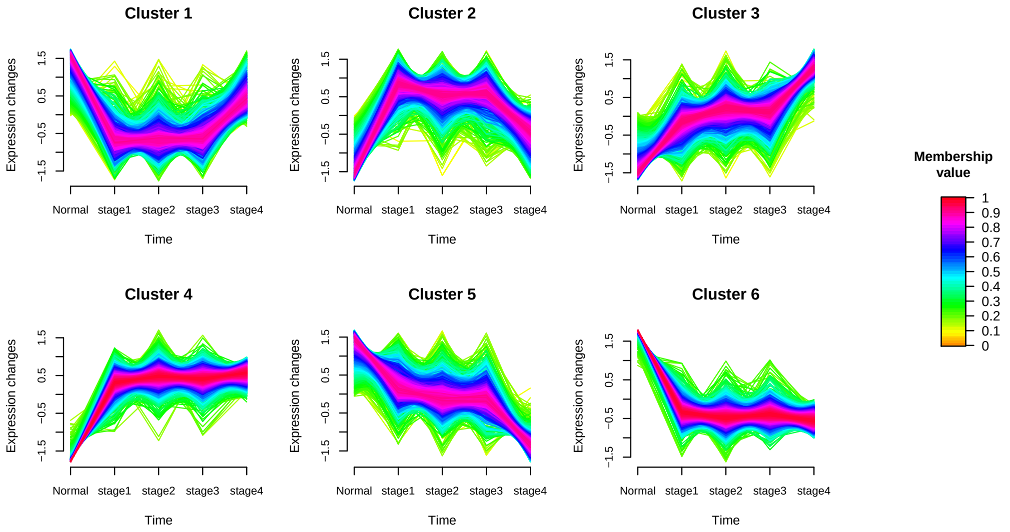

**Figure 2.5:** An example of soft-clustering results from gene expression data. Each cluster describes an expression pattern in the dataset (y-axis), made up by genes that follow its dominant pattern across the stages (x-axis). Green coloured lines correspond to genes with low membership value; red and purple coloured lines – high membership value.

The genes in each cluster are represented as lines colour-coded by their cluster membership value. Each cluster describes a particular expression pattern, for instance, Cluster 1 comprises genes that are upregulated at all stages of cancer compared to Normal, whereas Cluster 5 shows the opposite, i.e., downregulation trend. Other, more stage-specific patterns, such as in Clusters 2 and 6 with distinct expression is stage 4, can also be identified with this method.

The two parameters that are specified in fuzzy clustering are the number of clusters $c$ and the fuzziness parameter $m$. The fuzziness parameter value should be chosen to minimise clustering of random data and produce maximally stable clusters. Stable clusters are usually compact and are not affected by variation of $m$, whereas weak clusters disappear if $m$ is increased [121]. The $m$ parameter is often easiest to estimate by using the `mestimate` function specifically built for this purpose [123].

The method was originally introduced for microarray data, but it is acceptable to apply it to RNA-seq data in FPKM/RPKM format [124]. The normalised and filtered data can be directly used in the mfuzz standardisation function prior to clustering, which scales the expression change to zero mean and standard deviation of one (scale of the y-axis). The genes in the resulting clusters can be accessed and filtered by their membership score for inclusion in further analysis. The widely accepted membership score to use as a cut-off to select genes representing a particular cluster is 0.6 [125].

#### 2.3.2.3 Application to the dataset

Mfuzz was used to perform soft-clustering on the gene expression data with the input set of genes being the same as for differential experssion analysis, i.e. 15874 genes. The count data was TMM-normalised and transformed to log-RPKM format. Then, `removeBatchEffect()` function from `limma` package [114] was used create expression matrix where two batch effect sources, year and sample source site, have been taken into account prior to clustering the data.

Clustering was performed on all samples together and on individual PAM50 subtypes. For consistency, all analysis runs were done for six clusters ($c$) with fuzziness parameter ($m$) estimated by `mestimate` function. In each cluster, only genes with membership value of 0.6 or higher were used in the subsequent enrichment analysis.

## 2.3.3 Enrichment Analysis

### 2.3.3.1 Approach Motivation and Background

Gene expression analysis approaches produce lists of genes that are classified as 'interesting' in terms of their expression pattern (e.g. differentially expressed or clustered genes), but their immediate relevance to the study objective may be unclear. If 1,000 genes have changed in an experimental condition, it may be difficult to interpret the meaning of this change, and also one may be biased to only pay attention to the genes that are relevant to the study.

A solution is the use of enrichment/overrepresentation analysis, which aims to detect whether a group of objects has certain properties more (or less) frequent than can be expected by chance [126]. In biological sense, whether a candidate set of genes appears in the list of 'interesting' genes more frequently than genes picked at random would.

A gene set is a classified groups of genes into a biologically relevant group, which can be the members of the same biochemical pathway, be annotated with the same molecular function or expressed in the same cellular compartments. In this project, the gene set is the curated list of autophagy-related genes.

For enrichment analysis, four groups of genes are defined (Figure 2.6):

$m$ is the total number of genes  
$n$ is the number of genes in a set, e.g. autophagy genes  
$j$ is the number of 'interesting' genes, e.g. differentially expressed genes or genes in a cluster  
$k$ is the number of 'interesting' genes in the gene set

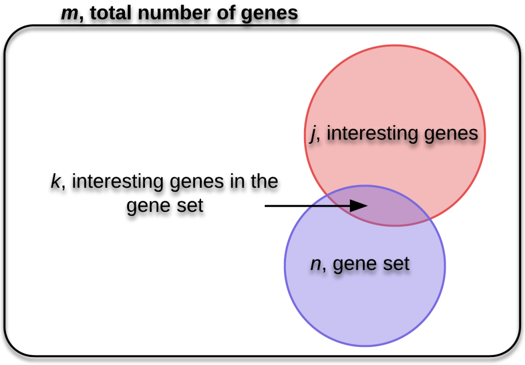
*m*, total number of genes
*j*, interesting genes
*k*, interesting genes in the gene set
*n*, gene set

**Figure 2.6:** The relationship between gene groups defined in Enrichment Analysis.

### 2.3.3.2 Fisher's Exact Test

A common approach for performing enrichment analysis is the Fisher's exact test. It is a statistical significance procedure for examining the association between two categorical variables [127].

To answer the question of whether a gene set is enriched in the list of interesting genes, it is common to start with a *2-by-2* contingency table. Table 2.5 shows the setup for a contingency table with the distribution of one variable in rows and another in columns. The character group labels ($m$,$n$,$k$,$j$) indicate how each value is calculated. The resulting matrix is used directly in the R function `fisher.test()`.

**Table 2.5:** The set up for enrichment analysis *2-by-2* contingency table. The gene set in this project is the autophagy-related genes, and the 'interesting' genes can be differentially expressed genes or genes represented by a cluster.

| | **Not interesting genes** | **Interesting genes** |
|---|---|---|
| **Other genes** | *m-n-j+k* | *j-k* |
| **Gene set genes** | *n-k* | *k* |

The null hypothesis of the Fisher's test assumes that the variables are independent and the sums of columns and rows are fixed [126]. Consequently, the values in contingency table are used to calculate the probability that this or any table with more extreme joint values (unobserved) would occur under the null hypothesis [128]. The calculated probabilities are expressed as p-values, and a small p-value indicates a discrepancy between the data and the null hypothesis of no association between variables.

In addition to p-value, `fisher.test()` also returns an *odds-ratio*, which is the ratio between proportions in and outside the sample. In other words, it is a measure of the magnitude of the difference. When testing for enrichment, the upper tail of the distribution is considered, which means that a one-sided test is performed. A two-sided test would imply also testing the the lower tail of the distribution, i.e. testing for depletion.

#### 2.3.3.3 Application to the dataset

Enrichment analysis was carried out on the results of differential expression testing to check whether autophagy-related genes are enriched among the DE genes in any of the model contrasts. The Fisher's test was performed on all DE genes, as well as on up- or downregulated separately. As three individual gene sets, all autophagy genes, only autophagy core genes, and only autophagy-related transcription factors were used.

The same enrichment analysis set-up was applied to the soft-clustering results. Genes in separate clusters with membership value of at least 0.6 were tested for enrichment of the gene sets.

The significance of autophagy enrichment was defined by odds-ratio $> 1$, and FDR-adjusted p-value $> 0.05$.

# Chapter 3

# Results

## 3.1 Exploratory Analysis

### 3.1.1 Sample Classifications

There are three main effects by which breast cancer samples can be stratified into groups: morphology, stage and PAM50 molecular profile. PCA separates samples into tumour/normal sample clusters fairly well (shown as an example in Methods, Figure 2.2). However, it is more interesting to apply PCA to explore variation between groups of samples classified by different methods. PCA was applied to 969 samples (857T, 112N), which were divided into subgroups by three systematic effect groups (i.e. PAM50, tumour morphology, cancer stage), the sizes of which are shown in Tables 2.2 and 2.3.

The PCA plot of data coloured by PAM50 molecular subtypes is shown in Figure 3.1. PC1 accounts for 11% of the total variation in the data, and is driven by the differences between normal and cancer samples. PC2 accounts for 8.6% of the variation, and characterises the variation among breast cancer subtypes. Basal-like samples form a separate cluster (orange), which emphasises its complete dissimilarity at molecular level. Luminal A and Luminal B form a partially overlapping cluster (yellow and red), which is expected from the known similarities of these subtypes. HER2-enriched cluster (blue) is understandably located between Luminal B and Basal-like as it partially shares receptor status subtype with the two. And lastly, seeing Normal-like subtype (pink) overlapping with normal samples (green) as well as with Luminal A subtype fits nicely with the fact that Normal-like subtype shares morphology with the former and IHC subtype with the latter.

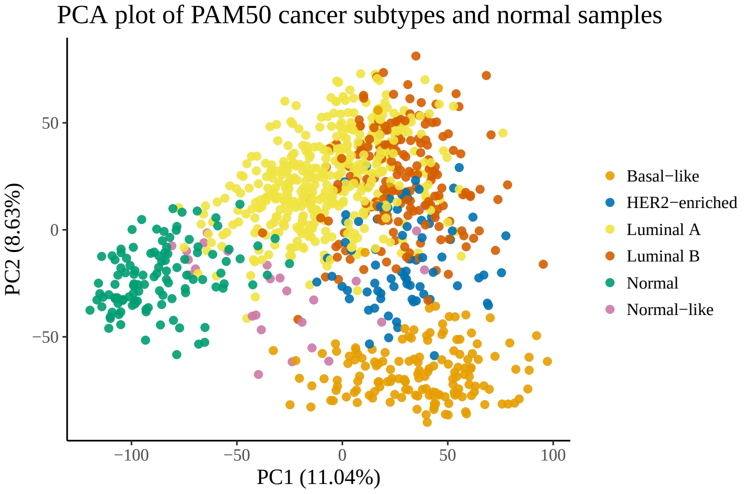

**Figure 3.1:** PCA plot showing the variance along PC1 and PC2 for PAM50 molecular subtypes of breast cancer and normal samples.

One dimensional PCA plot based on PAM50 subtypes is shown in Figure 3.2, where the variation along the first nine PCs is displayed as explained in Section 2.2.1. Looking at the results in this way highlights again that PC1 is driven by tumour/normal variation. Together with PC2 they also capture the differences among PAM50 subtypes. As the PCs numbers increase, variation captured by them decreases.

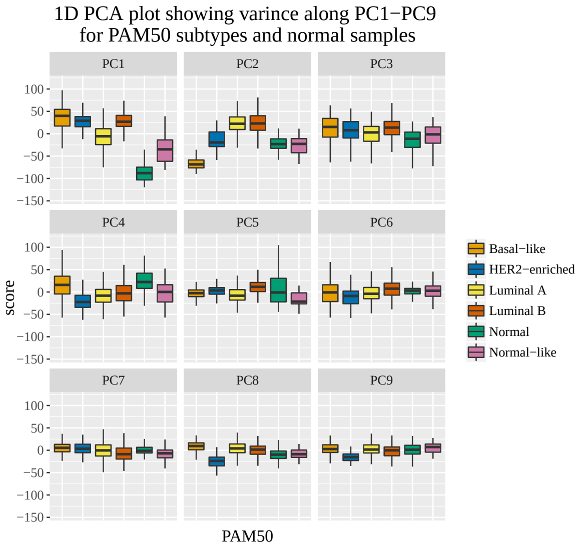

**Figure 3.2:** One-dimensional PCA plots showing variance along PC1-PC9 for PAM50 molecular subtypes of breast cancer and normal samples.

The PCA results for stages and morphology classifications are shown in Figures 3.3 and 3.4 (2D ad 1D). PCA was not able to capture as much variation between stages and morphology groups as between PAM50 subtypes. From one-dimensional plots for both stages and morphology it is evident that PC1 captures cancer/normal differences in both classification types. However, the rest of PCs do not capture enough variation for the groups to be clearly separated on 2D PCA plots.

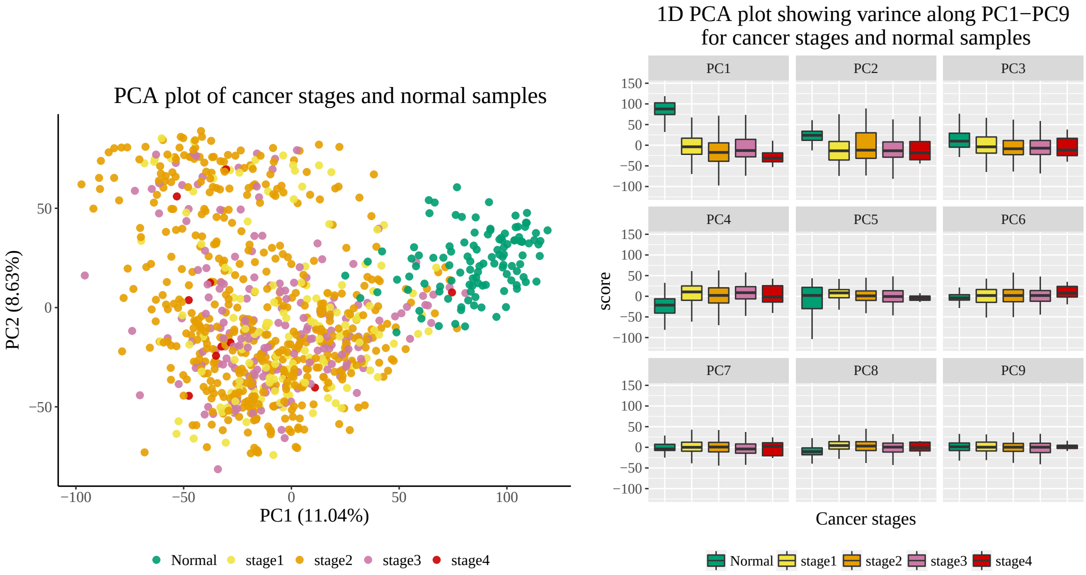

**Figure 3.3:** PCA results showing variance captured between cancer stages and normal samples; left - 2D, right 1D (PCs 1-9). Unknown stage samples are excluded.

Interestingly, PCA did not capture any noticeable differences even between distant stages (i.e. stage 1 and stage 4), where the differences in expression should be substantial and therefore reflected in the PCA results. This is also the case for two main distinct morphologies - Lobular and Ductal carcinomas, which overlap greatly in the plots.

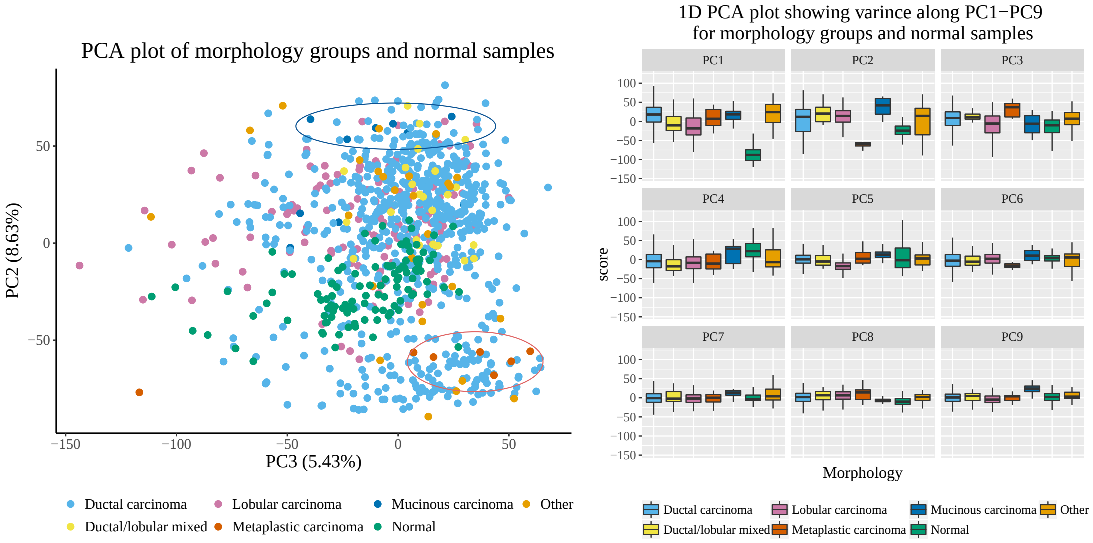

**Figure 3.4:** PCA results showing variance captured between morphology types and normal samples; left - 2D, right 1D (PCs 1-9).

A possible explanation for the observed PCA results is the difference in sizes of morphology and stages subgroups, as well as their composition in terms of PAM50 subtypes, which evidently characterises the main driving force behind differences between samples. To explore this idea, the sample count per subgroup was visualised as stacked barplots presented in Figure 3.5. Each bar shows the total number of samples of a chosen morphology (left plot) or stage (right plot). The difference in subgroups sizes in apparent.

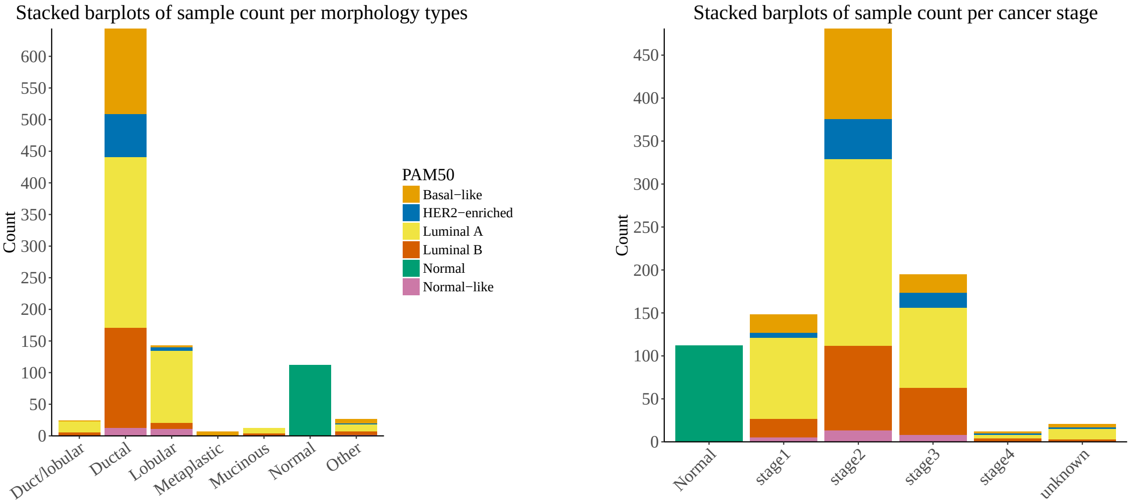

**Figure 3.5:** Stacked barplots of sample counts per subgroups in morphology types (left) and stages (right). The coloured proportions represent PAM50 subtypes samples within each subgroup.

Ductal carcinoma group makes up 75% of all samples (644/857), but is well-proportionate in terms of its PAM50 composition (i.e. similar proportions as the full dataset, with Luminal A dominating). The other morphologies, however, in addition to being smaller in sample count, are restricted to only a few PAM50 subtypes. For instance, Metaplastic carcinoma is made up completely of Basal-like samples, while Mucinous morphology is represented by only Luminal subtypes (the counts are shown in Table 3.1). This explains why these two mophologies are well-separated in PCA plot on PC2 (Figure 3.4, circled) – the difference between Luminal and Basal tissues is driving their variation. However, it is important to note, that there is some evidence in the literature, e.g. a study by Weigelt *et al.* [23], that Metaplastic and Mucinous morphologies actually do only exists as Basal and Luminal subtypes, respectively.

The sample count difference between stages is also noticeable, with stage 2 accounting for roughly 50% of the samples. The proportions of PAM50 subtypes within stages 1-3 appear to be well-balanced. However, this is not the case for stage 4, where in addition to alarmingly low sample count, Normal-like subtype is not represented. The counts are shown in Table 3.1.

**Table 3.1:** Sample counts in pairwise comparisons of three main breast cancer samples classifications. *Top table:* PAM50 and stages, *Middle table:* PAM50 and morphology, *Bottom table:* stages and morphology. Each table shows row and column sums for each subgroup. The total count of samples in the project is in the lower right corner: 857.

| | **Luminal A** | **Luminal B** | **Basal-like** | **HER2-enriched** | **Normal-like** | *rowsums* |
|---|---|---|---|---|---|---|
| **stage 1** | 94 | 22 | 21 | 6 | 5 | 148 |
| **stage 2** | 217 | 99 | 105 | 47 | 13 | 481 |
| **stage 3** | 93 | 55 | 21 | 18 | 8 | 195 |
| **stage 4** | 4 | 4 | 2 | 2 | x | 12 |
| **unknown** | 12 | 3 | 4 | 2 | x | 21 |
| *colsums* | 420 | 183 | 153 | 75 | 26 | *857* |

| | **Luminal A** | **Luminal B** | **Basal-like** | **HER2-enriched** | **Normal-like** | *rowsums* |
|---|---|---|---|---|---|---|
| **Lobular** | 114 | 10 | 3 | 5 | 11 | 143 |
| **Ductal** | 270 | 158 | 135 | 68 | 13 | 644 |
| **LobDuctal** | 17 | 6 | 1 | x | x | 24 |
| **Metaplastic** | x | x | 7 | x | x | 7 |
| **Mucinous** | 8 | 4 | x | x | x | 12 |
| **Other** | 11 | 5 | 7 | 2 | 2 | 27 |
| *colsums* | 420 | 183 | 153 | 75 | 26 | *857* |

| | **stage 1** | **stage 2** | **stage 3** | **stage 4** | **unknown** | *rowsums* |
|---|---|---|---|---|---|---|
| **Lobular** | 14 | 78 | 49 | x | 2 | 143 |
| **Ductal** | 118 | 368 | 129 | 11 | 18 | 644 |
| **LobDuctal** | 5 | 11 | 7 | x | 1 | 24 |
| **Metaplastic** | 2 | 4 | 1 | x | x | 7 |
| **Mucinous** | 3 | 5 | 4 | x | x | 12 |
| **Other** | 6 | 15 | 5 | 1 | x | 27 |
| *colsums* | 148 | 481 | 195 | 12 | 21 | *857* |

The alternative method of visualising dataset structure and exploring the relevance of different classifications is to perform sample clustering and represent it as a heatmap. Figure 3.6 shows the

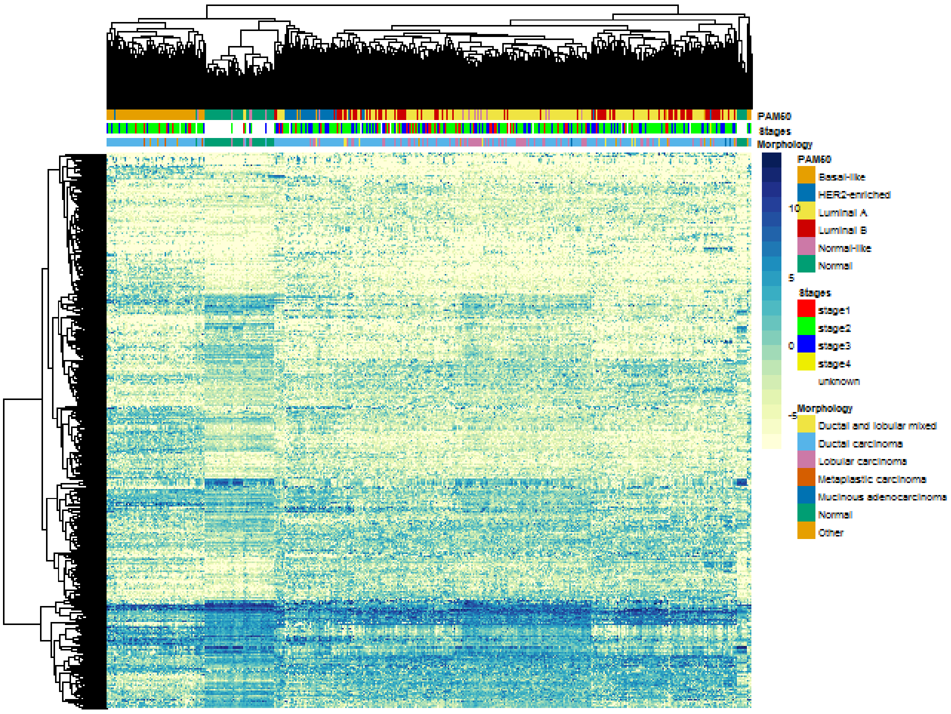

**Figure 3.6:** Heatmap representation of clustering top 1000 highest variance genes (calculated on cancer samples only). The data is clustered by columns (samples) and rows (genes), with dendograms showing how clusters are formed. The colour bars above the heatmap (PAM50, stages, morphology) show the subgroups to which each sample belongs, colour-coded according to the legend on the right. The colours of heatmap represent gene expression magnitude according to the scale (high expression - dark blue, low expression - light yellow). The data was clustered with Euclidean distance and average linkage.

clustering results based on top 1000 highest variance genes. Variance was computed for each row (gene) in cancer samples only, and the top 1000 genes were used in clustering to capture largest gene expression changes across the full dataset. The three main effect groups (PAM50, stages, morphology) are shown above the heatmap as colour bars, providing side-by-side comparisons.

First and foremost, the unsupervised clustering was able to form two main clusters: Luminal and non-Luminal samples. Cluster on the right contains the majority of Luminal A and Luminal B samples according to the top colour bar (PAM50). The left hand side cluster contains Basal-like, HER2-enriched, and normal samples. Normal samples have a distinct tanscriptomic profile which is seen from the heatmap. Normal-like subtype is scattered across the entire dendogram, with more pronounced appearance amongst Luminal A and normal samples, as expected. The Basal-like subtype cluster is well-defined, which highlights its uniqueness and distinction from the rest. The Luminal types are mixed as anticipated. Overall, clustering is in agreement with the PCA results for PAM50 subtypes.
Regarding the stages and morphology, it can be observed that clustering, much like PCA, was not able to find any distinct patterns among their subgroups. Again, the sizes of the subgroups may partially be responsible for that. It is a challenge to spot the minority-sized subgroups on the colour bar, but very clear to see how Ductal carcinoma dominates the dataset (light blue, bottom bar). The colour bar of stages also shows no distinct patterns.

As stages of cancer progression are of great interest in this project, an additional exploratory clustering analysis was done to investigate the relationship between cancer stages and the underlying PAM50 subtypes. The averages of PAM50 subtypes expression at each stage were clustered to produce dendogram in Figure 3.7.

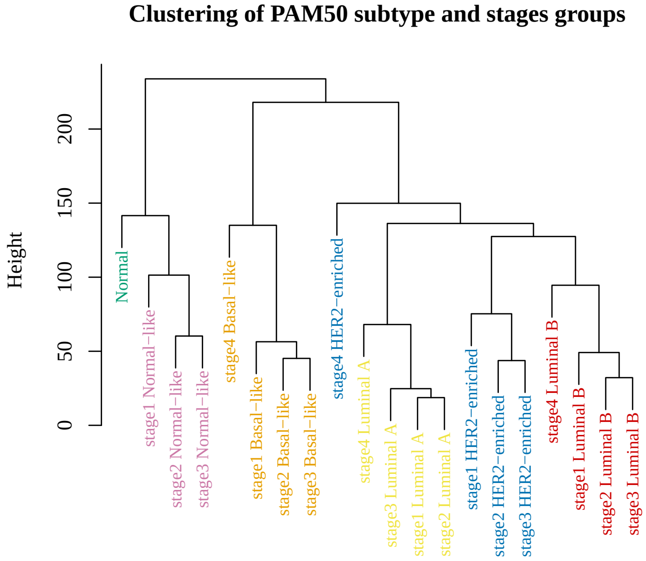

**Figure 3.7:** Dendogram showing clustering of PAM50 subtypes within stages (averages per stage per subtypes).

Remarkably, all subtypes form distinct branch clusters (expect HER2-enriched). Normal and Normal-like samples form one outgroup, and Basal-like subtype the other. Luminal A, Luminal B, and HER2-enriched are separated to well-defined branches. Importantly, each subtype cluster has stage 4 as an outgroup, with the exception of HER2-enriched, in which stage 4 is only made up of 2 samples, perhaps leading to an unstable average. Seeing stage 4 as an outgroup is an important observation, as it is the most different and severe stage of cancer progression.

Lastly, seeing the samples cluster primarily by PAM50 and only then by stage, is yet another evidence to the previously observed results in PCA and heatmap clustering. PAM50 classification describes the main driving force behind breast cancer samples differentiation the best.

### 3.1.2 Cofactors and batch effects

Morphology, cancer stages, and PAM50 molecular subtype are the anticipated sources of variation in the dataset, and namely, biological variation. However, in a large and heterogeneous dataset such as TCGA-BRCA there are several other potential sources of variation, which may be contributing to the differences in gene expression between samples. The abundance clinical and meta annotation available for the present samples allows these potential sources of technical variation to be explored.

Among the available annotations, the ones of the most interest are the patient age group, year sample was taken, and tissue source site. They were explored with PCA to visualise any outlying subgroups and check for the presence of unwanted batch effects among samples, such as differences caused by being processed in different years and across a variety of research labs.

One-dimensional PCA plots were used to visualise variation over the years and across the sample processing sites. As there are many (> 20) subgroups in both year and site data, 2D PCA scatter plots are not helpful for spotting outlying subgroups.

Figure 3.8 shows the first 6 PCs of from PCA results based on year data. Each box plot represents samples taken in a single year, with individual years shown as gitter-dots to visualise each year group approximate size and sample variation within it (according to y-axis). The box plots are shown for years in the sequential order (1988-2013).

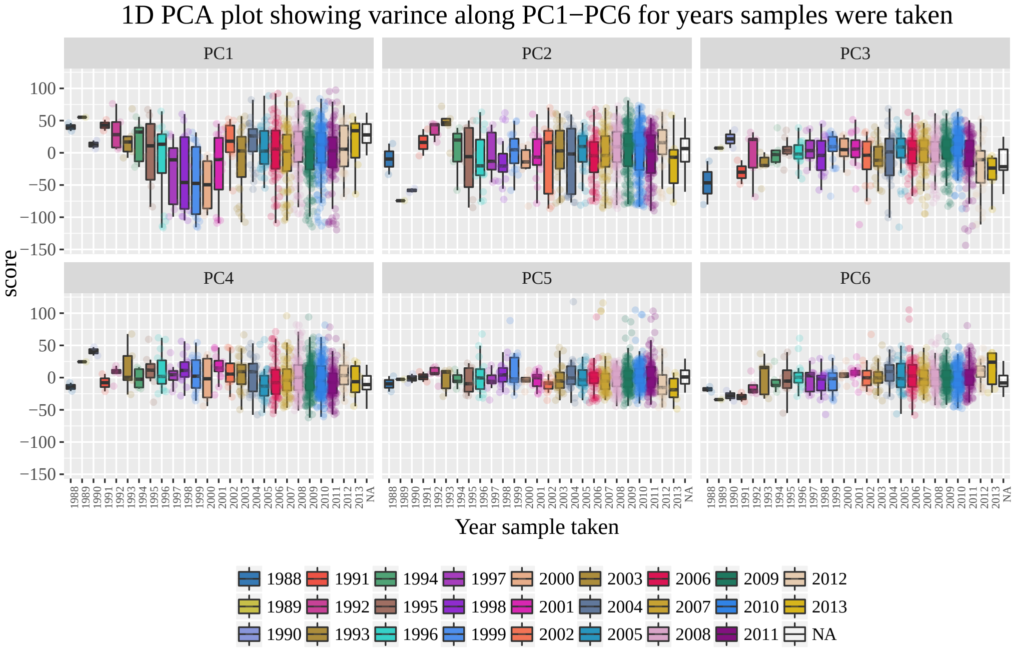

**Figure 3.8:** One-dimensional PCA plots showing variance along PC1-PC6 for all samples (cancer and normal) taken in years 1988-2013. The dots across each box plot represent samples and their variation on the y-axis scale, to give a better indication of year group size and sample variation within it.

Overall, it is evident that there is a lot of variation between the years. This is especially the case for the earlier years, however, it is partially due to the low number of samples taken in those years. The variation stabilises after year 2005, which can be explained by the fact that this is when the pilot of TCGA project started and the universal protocols were introduced [53]. Also from the large number of jitter dots over those years it can be seen that the majority of samples were collected after that time point. Another aspect worth noting is that samples in years 1997-2000 appear to be quite different to previous and following years (PC1). This potentially confounding effect was explored further.

As previously established, PAM50 subtypes is expected to be the main point of interest for further investigation, therefore, it is important to check the PAM50 composition of every year group to make sure that, for example, a particular subtype has *not* been collected in only one year, which would have confounded the dataset for further exploration. Figure 3.9 shows the stacked barplot of all sample counts organised by year. Each bar shows the total count of samples per year, and the proportions of samples representing each PAM50 subtype are coloured according to the legend.

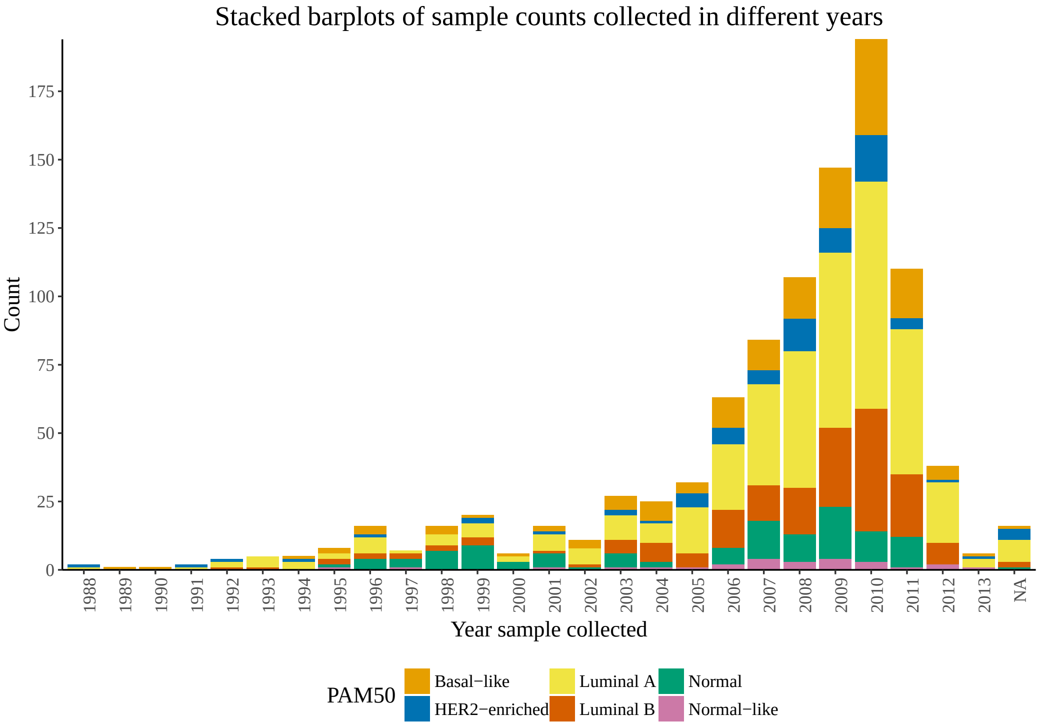

**Figure 3.9:** Stacked barplots of sample counts per year (1988-2013). The coloured proportions represent PAM50 subtypes samples within each year group.

As already noted from 1D PCA plot for the year data, the total number of samples taken in the last decade is marginally higher than in the earlier years. Overall, the distribution of PAM50 subtypes in each year is well-balanced, although some years inevitably have no normal samples and/or samples of a specific subtype. On the whole, it can be concluded that the dataset is not confounded by a particular year-subtype combination.

Moreover, visualising data in this way has shed light on the abnormality in years 1997-2000 seen in 1D PCA plot (Figure 3.8, PC1). It can be seen in Figure 3.9 that these years are represented by $\approx 50\%$ of normal samples, which is not the case for all other years. Having such high proportion of normal samples, makes their box plots to be shifted down compared to the rest, and hence appear different. Appendix A shows a 1D PCA plot for cancer samples only, confirming this explanation, i.e. years 1997-2000 do not stand out from the rest if only cancer samples are included in the analysis.

Similarly to year data, the sample source data was also explored with PCA and variation between different source sites assessed. A corresponding 1D PCA plot (PC 1-6) showing variation for samples grouped by sample source site (24 sites in total) is displayed in Figure 3.10.

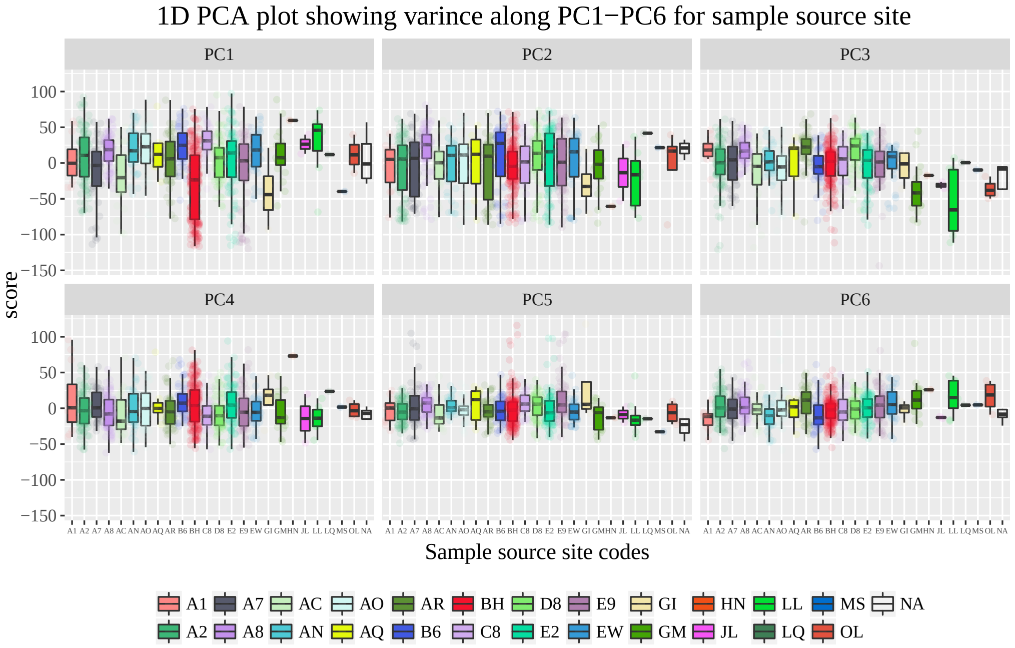

**Figure 3.10:** One-dimensional PCA plots showing variance along PC1-PC6 for all samples (cancer and normal) from different source sites (two-letter abbreviations in alphabetic order; NA - site unknown). The dots across each box plot represent samples and their variation on the y-axis scale, to give a better indication of year group size and sample variation within it.

The existence of variation between source sites can be seen unmistakably. As with year data, some source site groups contain very few samples, which makes the variation between them and the large groups more prominent. One source site that really stands out is BH (red), on PC1. A possible explanation to this observation, which would have made the dataset unusable, is the unexpected difference in the handling protocol at this site. Fortunately, this was not the case.

The source site data was visualised as stacked bar plots to inspect sample subtype variety coming from each site, Figure 3.11. As previously noted, some source sites have produced very few samples. The larger sample groups are relatively well-balanced in terms of their PAM50 subtype composition, with the exception of Normal-like, which comes only from 11 sites. This is, however, expected, as there are only 26 Normal-like samples in total.

Another important observation is that the majority of normal samples come from one site, BH. This should not affect the analysis, as the main interest lies in the difference between cancer subgroups. However, seeing this explains the abnormal variation seen in 1D PCA plot for this site. As with year data, this one site is $\approx$ 50% composed of normal samples, making it appear very different from the rest, because PC1 is driven by cancer/normal distinction. Appendix A contains the plot resulting from using only cancer samples.

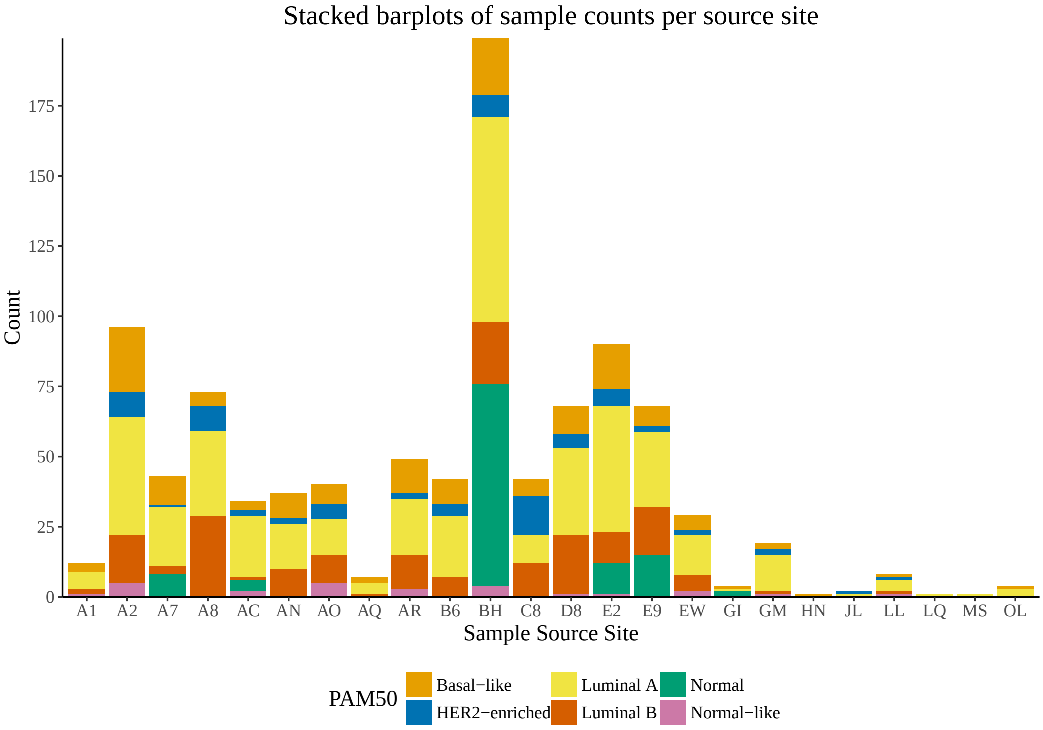

**Figure 3.11:** Stacked barplots showing count of each PAM50 subtype collected at each sample source site (shown in alphabetic order). The coloured proportions represent PAM50 subtypes samples within each year group.

### 3.1.3 Summary

Exploratory analysis was performed on the available samples from TCGA-BRCA dataset to explore variation described by three main sample classification methods (PAM50 molecular profile, tumour morphology, cancer stage) and to assess the variation coming from two potential sources of batch effect – sample source site and year sample taken.

PAM50 intrinsic subtypes describe the main differences between the samples, which is distinctly shown by PCA and clustering results. Morphology and stage sample classifications do not capture enough variation on their own. This could be due to very disproportionate subgroup sizes, as well as the fact that PAM50 composition of subgroups has a much stronger variation signal which leads to masking their own variation. Therefore, when using other classification methods, PAM50 subtypes have to be taken into account. For example, comparing two morphologies might not produce truthful results if their PAM50 subtype sample composition is the main driving effect of their differences. Hence, the heterogeneity of groups has to be taken into account, and the downstream analysis has to be done on combinations of subgroups.

The exploration of year and sample source data has shown the apparent technical variation coming from both. Therefore, in the further analyses, for example in differential expression testing, those two sources of batch effects have to be included in the model.

## 3.2 Identifying Autophagy Signatures

### 3.2.1 Differential Expression Analysis

Differential expression (DE) analysis was performed on the final dataset which included 857 tumour and 112 normal samples. After all filtering steps the dataset included 15784 genes, 1090 of which were autophagy-related genes.

Limma-voom method was used to quantify differential expression between the samples representing subgroups of the main breast cancer classification methods (PAM50 molecular subtypes, cancer stages, tumour morphology). The subgroups are described in Section 1.1.3 and the numbers of samples per subgroup in the final dataset are presented in Tables 2.2 and 2.3. The outline of differential analysis workflow is shown in Figure 3.12.

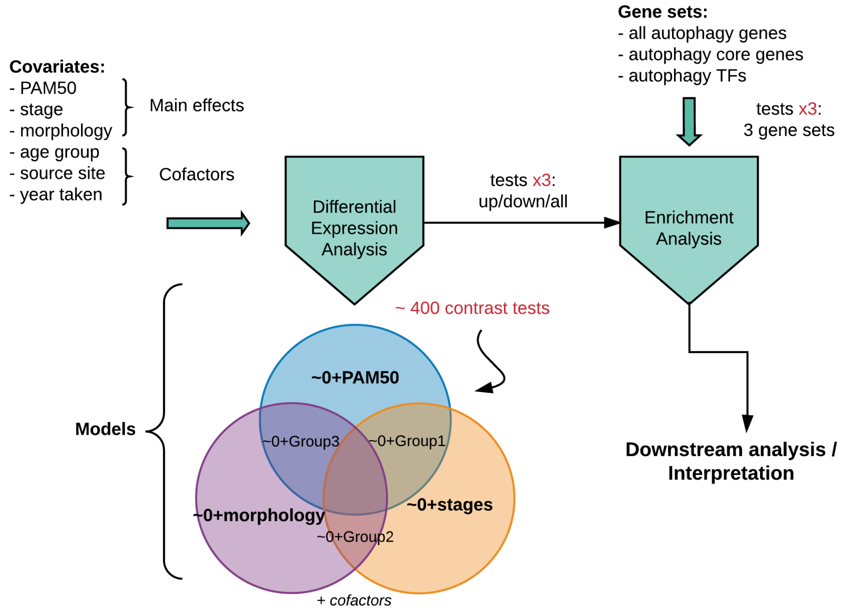

**Figure 3.12:** Differential expression analysis workflow for the TCGA-BRCA dataset. The main classification groups/effects (PAM50, stages, morphology) as well as the additional cofactors (age group, sample source site and year) were included in the models for differential expression testing. The Venn diagram shows the six tested models; the outer sections show the three main effect models, and the overlapping 'Groups' show the three models of combined main effects. In total, ≈ 400 contrasts across the six models were made to test for differential expression. For enrichment analysis, three gene sets were used: all autophagy genes (n=1090), core only (n=155), and TFs only (n=97). Enrichment was tested in all differentially expressed genes, and also in the groups of up- or downregulated genes individually. In total, 400x3x3 tests were made prior to downstream analysis.

#### 3.2.1.1 Main Effect Models

Each of the main sample classifications (PAM50, stages, morphology) were used as the main effect in individual models for DE testing. For each main effect, all combinations of subgroups and normal samples were included in the contrast matrix. In this way, for each PAM50 subtype (or each stage and morphology) differentially expressed genes were identified compared to the rest of subtypes and the normal samples. All models included the age group covariate as well as the two sources of batch effect: year and site. The raw numbers of differentially expressed genes in each contrast are available from Appendix B and are described below.

Differential expression testing based on **PAM50 subtypes** has detected a fairly large number of differentially expressed genes (DEGs) in each comparison (min: 5% - Luminal A vs Luminal B, max: 31% - Basal-like vs Normal). Understandably, contrasts of normal samples vs a subtype detected much more DEGs, a large number of which are downregulated in cancer. The proportions of differentially expressed genes between PAM50 subtypes ranged between 5% and 22% (Luminal A vs Luminal B and Luminal A vs Basal-like), whereas a subtype vs normal 12% (Normal-like) to 31% (Basal-like); details in Appendix B.

Differential expression testing based on **cancer stages** showed a considerable number of differentially expressed genes between stages and normal samples, ≈ 23% at each stage. Interestingly, the analysis has found only very few genes to be differentially expressed between individual stages.

Differential expression testing based on **tumour morphology** showed that the most DEGs-rich contrasts are again the ones involving normal samples. In the contrasts where morphology types were compared to normal samples the proportion of DEGs varied between 21% and 28%. The two main morphologies, Lobular and Ductal carcinomas, had only 6% (924) DEGs between them. When both were tested with the mixed Lobular+Ductal morphology, both had less than 20 DEGs, meaning that this morphology is too much of a blend between the two separate ones to detect sufficient differential expression. Metaplastic and Mucinous morphologies were found to have the most DEGs with Lobular morphology (15% and 13%, respectively), and less with Ductal (6% and 9%) and Lobular+Ductal (10% and 6%) morphologies. The largest number of DEGs between two morphologies was between Metaplastic and Mucinous - 19%.

Overall, large numbers of differentially expressed genes between PAM50 subtypes is an expected and reassuring result, as it is known from exploratory analysis that the differences between them is the strongest source of variation in the data. Not being able to see many differentially expressed genes between stages can be due to heterogeneity of the dataset, which is also probably contributing to differential expression in larger morphology types.

#### 3.2.1.2 Combined Effect Models

The reduce the effect of data heterogeneity on the differential expression analysis, additional models with combined group effects were tested. Three models were created: PAM50 + stage (Group1), morphology + stage (Group2), PAM50 + morphology (Group3), as shown in the Venn diagram in Figure 3.12. This analysis setup allowed testing for differential expression in more precisely defined groups, for example, Group1 model is able to quantify expression between stages of a specific subtype, i.e. between stages 1 and 2 of Luminal A. Again, all models included the age group, year, and source site as cofactors.

A general observation for using combined models is that a large number of separate PAM50 subtype-specific differentially expressed genes were found at individual stages and in specific morphologies. As the number of contrasts for each Group is around 100, the quantitative details will be omitted.

Differential expression analysis with **Group1 model** (PAM50 + stage) has shown that all subtypes at all stages have a large number of DEGs compared to normal samples. Between different stages of a single subtype, only Basal-like and Luminal B had a reasonable number of differentially expressed genes. Contrasts between the same stages of different PAM50 subtypes have shown that Luminal A and Basal-like have the largest number of differentially expressed genes consistently at all stages. Interestingly, the smallest numbers of differentially expressed genes between subtypes was at stage 4.

Differential expression analysis with **Group2 model** (stage + morphology) has shown that Ductal morphology has the largest number of DEGs at all stages. Ductal is the only morphology that contains stage 4 samples in this dataset, so the comparisons for the rest of morphologies were made based only on the first three stages. The stage-wise comparison of Lobular and Ductal carcinomas has shown that the number of DEGs between them increases with stage. This observation can be considered reliable, as both of them have a adequate number of samples per stage, and also both are represented by a good mix of PAM50 subtypes (Table 3.1).
This is, however, not the case for contrasts involving Mucinous and Metaplastic types. As mentioned before, their PAM50 composition is restricted to Luminal and Basal subtypes, respectively, which makes all comparisons made with them dubious, as the true driving force of this variation is unknown. The DE results between the two have shown a very large number of differentially expressed genes, particularly at stage 2. In this contrast the comparison that is being made is between only 4 and 5 samples, which are exclusively Basal-like and Luminal A. In an earlier comparison these two subtypes showed the highest number of DEGs on the all-samples level, and here, this compelling difference appears to be maintained at the level of a few samples too.

Differential expression analysis with **Group3 model** (PAM50 + morphology) is of interest mostly only for Lobular and Ductal morphologies, as only they contain enough samples to represent each PAM50 subtype (Table 3.1). Both have a large number of differentially expressed genes between all subtypes, with Ductal carcinoma having a distinctly large number for the contrast between Luminal A and Basal-like.

#### 3.2.1.3 Enrichment Analysis

Ultimately, the project interest lies not in the numbers of differentially expressed genes in each contrast, but in whether these subsets of DEGs are enriched for autophagy-related genes. As shown in the diagram in Figure 3.12, groups of up-, down-, and both direction regulated genes were tested for enrichment. Three separate gene sets were used in the Fisher's test - all autophagy genes, only autophagy core genes, and only autophagy-related transcription factors (TFs). The significance of autophagy enrichment was evaluated based on the cut-off for adjusted p-value ($< 0.05$) and the odds-ratios score ($> 1$). The results significant at non-adjusted p-value were also considered.

However, the enrichment results have largely been unfruitful, particularly on the differential expression results from the individual main effect models. The raw data is shown in Appendix X2 and described below.

From the individual main effect models (only PAM50, stage, or morphology) none of the subgroups of differentially expressed genes detected in contrasting tests have been assigned an odds-ratio $> 1$ and simultaneously been significant at adjusted p-value level. There were several contrasts enriched for TFs but only at non-adjusted p-value level. All of those contrasts were downregulation in a particular subset of cancer samples compared to normal:

- Luminal A, Luminal B, and Basal-like (both directions in Luminal B)
- Stage 2, stage 3, and stage 4
- Metaplastic, Mucinous, and Lobular+Ductal mixed

When using Group models, i.e. combining two effects, some contrasts were found enriched at acceptable levels.

In **Group1**, 'all autophagy' genes were enriched in the downregulated genes at stage 4 of Basal-like subtype compared to stage 4 of Luminal A and Luminal B, and also to normal samples. At non-adjusted p-value level, there were 13 contrasts enriched for TFs downregulated in a cancer type vs normal. Among these, there was stage 1 of every PAM50 subtype, all stages of Basal-like subtype (details in Appendix C).

In **Group2**, autophagy TFs were enriched in the downregulated genes in every morphology compared to normal, at varying stages. There was also a strong enrichment in TFs that are differentially expressed between Lobular and Ductal carcinomas at stage 1. At non-adjusted p-value level, there was a large number of contrasts enriched for autophagy TFs, mainly in the downregulated in cancer vs normal fashion. Among these, there are all stages of Ductal and Mucinous types, as well as stages 1 and 3 of Metaplastic. Interestingly, Lobular carcinoma did not appear in this context at all. Conversely, groups of upregulated genes in Lobular compared to Metaplastic (stage 3) and Mucinous (stage 2) were enriched for autophagy (Appendix C).

In **Group3**, 'all autophagy' genes were enriched in the downregulated genes in Basal-like samples compared to Luminal A and Normal-like samples (all of Lobular morphology) and normal samples. When looking at non-adjusted p-value, Luminal B can also be added to this pattern. As for non-adjusted p-value enrichment for TFs, there were multiple enriched contrasts, again including several instances of downregulation in cancer vs normal (particularly Basal-like and Luminal B), DE between morphologies within Basal-like subtype, and again similar observations for Lobular carcinoma as for 'all autophagy'.

Overall, it seems that using the combined main effects models permits to detect more autophagy enrichment in the differential expression results. As only specific combinations of samples are considered, the heterogeneity is decreased, allowing for a stronger differential expression signal, i.e. more genes are reported. However, it still cannot be claimed that enrichment results attained from the combined models provide strong evidence of a particular group of samples being enriched for autophagy. Hence, the main observations are:

- The main reoccurring trend among all tested contrasts of differentially expressed genes is autophagy enrichment among genes that are being downregulated in cancer.
- There is a consistent enrichment for autophagy TFs in genes downregulated in separate stages/morphologies (at adjusted and non-adjusted p-value level)
- There is no autophagy enrichment in genes that are differentially expressed between stages. This is the case for both the main effect model, and the models with stage combined with other classification subgroups.
- Among the frequently seen potentially enriched groups of samples are Basal-like subtype, Lobular morphology, stage 4. However, exploring these groups further cannot be justified, as there are very few samples in them. Table 3.1 shows that stage 4 samples are available only for Ductal morphology, and of samples that are Basal-like and have Lobular morphology there are only 3.

#### 3.2.1.4 Summary

Differential expression analysis of the TCGA-BRCA dataset was performed based on three sample classification methods – stage, morphology, PAM50 molecular profile. The analysis involved approximately 400 differential expression tests across six model. The results have shown that contrasts between samples classified by their molecular profile detect the largest amount of DEGs, both when using the model with a single main effect and models with combined effects. Conversely, very minimal (if any) differential expression was detected between cancer stages, both with single main effect model and models with combined effects. With morphology classification, the most DEGs-abundant contrast was between two smallest in size mophologies, Metaplastic and Mucinous, which are made up solely of two distinct PAM50 subtypes. Considerably less DEGs were found between two largest morphologies, Ductal and Lobular carcinomas. However, the number of differentially expressed genes between them has increased when they were compared at individual PAM50 subtype level.

Overall, enrichment analysis was unsuccessful in identifying autophagy enrichment in any of the groups of differentially expressed genes. There were several contrasts of samples changes between which included a significantly large number of autophagy genes and therefore were considered enriched. However, this enrichment was not consistent and usually appeared in the groups that were represented by too few samples to be reliable. The general trend for autophagy genes is downregulation in various subgroups of cancer samples compared to normal, and particularly this trend is true for autophagy-regulating transcription factors.

### 3.2.2 Soft-clustering

Soft-clustering was performed on all samples together to evaluate the expression patterns in the dataset as whole, and also on separate PAM50 subtypes individually to detect any molecular profile-specific patterns. Classification by morphological groups was decided not to be included in the analysis due to the underlying imbalance of PAM50 composition of each morphology, and also the fact that only one morphology contains samples of all stages (Ductal carcinoma).

#### 3.2.2.1 Full Dataset Clustering

Gene clustering results based on full dataset expression data are presented in Figure 3.13. The genes were sorted into six clusters according to expression changes between normal samples and cancer samples of different stages. Clusters 4 and 5 represent genes that are downregulated or upregulated, respectively, in cancer compared to normal, regardless of the stage. Clusters 1 and 2 represent patterns of more gradual down- and up- regulation in cancer, with a distinct rapid change in stage 4. Clusters 3 and 6 are assigned genes that have altered expression at initial stages of cancer, while towards the most serious stage, i.e. stage 4, the expression resembles normal expression. Here, only the genes with $> 0.6$ cluster membership score are presented, as they are the main contributors to a cluster's stability and dominant expression pattern. Only those genes are considered in downstream analysis as cluster representatives.

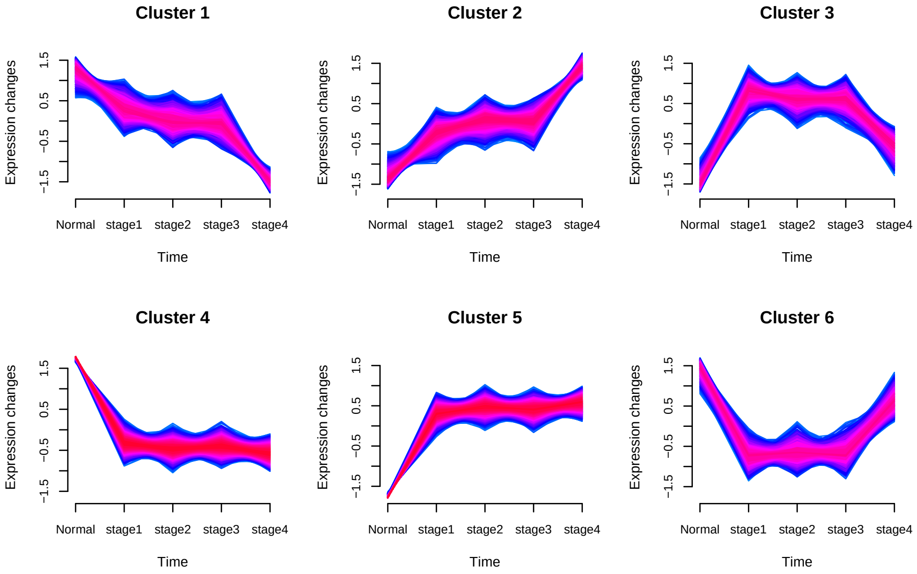

**Figure 3.13:** Soft-clustering results based on all samples data. Each cluster describes an expression change pattern (y-axis), made up by genes that follow its dominant pattern across the stages (x-axis). Genes only with membership score $> 0.6$ are shown. Number of genes per cluster: 1 - 792, 2 - 1005, 3 - 796, 4 - 3224, 5 - 2889, 6 - 743.

In this way, the obtained six gene groups were tested for autophagy enrichment. As previously with differential expression results, three gene sets were used: all autophagy genes, only autophagy core genes, and only autophagy-related transcription factors (TFs). The significance of autophagy enrichment was evaluated based on the cut-off for adjusted p-value ($< 0.05$) and the odds-ratios score ($> 1$).

Table 3.2 presents the enrichment results for soft-clustering done on all samples. For each gene set, the number of autophagy genes present in a cluster is reported, along with the odds-ratio score, p-value, and adjusted p-value. The cluster that is found to be enriched for autophagy is marked with bold (Cluster 4), at both all genes and TFs-only levels. Cluster 4 comprises the genes that are downregulated at all cancer stages compared to normal. 264 autophagy-related genes display this expression pattern, 39 of them are TFs.

**Table 3.2:** Enrichment for autophagy analysis results based on soft-clustering patterns of all samples. Three autophagy gene sets were tested: all genes, core autophagy genes, and autophagy TFs. The cluster that is significantly enriched is Cluster 4 (at all genes and TFs-only levels).

| | **n genes** | **odds-ratio** | **p-value** | **adj. p-value** |
|---|---|---|---|---|
| *all autophagy genes* | | | | |
| cluster1 | 47 | 0.84 | 0.31 | 0.63 |
| cluster2 | 69 | 0.99 | 1 | 1 |
| cluster3 | 51 | 0.92 | 0.61 | 0.92 |
| **cluster4** | 264 | 1.26 | 2e-03 | 7e-03 |
| cluster5 | 150 | 0.69 | 4e-05 | 2e-04 |
| cluster6 | 53 | 1.04 | 0.76 | 0.92 |
| *autophagy core genes* | | | | |
| cluster1 | 4 | 0.49 | 0.19 | 0.42 |
| cluster2 | 8 | 0.79 | 0.74 | 0.89 |
| cluster3 | 9 | 1.16 | 0.58 | 0.87 |
| cluster4 | 32 | 1.01 | 1 | 1 |
| cluster5 | 22 | 0.73 | 0.21 | 0.42 |
| cluster6 | 13 | 1.87 | 0.05 | 0.31 |
| *autophagy TFs* | | | | |
| cluster1 | 7 | 1.48 | 0.34 | 0.41 |
| cluster2 | 2 | 0.31 | 0.09 | 0.19 |
| cluster3 | 5 | 1.02 | 0.82 | 0.82 |
| **cluster4** | 39 | 2.62 | 1e-05 | 7e-05 |
| cluster5 | 10 | 0.51 | 0.05 | 0.14 |
| cluster6 | 2 | 0.42 | 0.33 | 0.41 |

#### 3.2.2.2 Subtype-specific Clustering

Clustering was performed on individual PAM50 subtypes samples to evaluate the presence of any characteristic expression patterns and test for subtype-specific autophagy enrichment. For individual subtypes, figures with the complete set of clusters and enrichment results tables can be found in Appendix D. Here, only selected cluster details are presented.

**Basal-like subtype**

The expression patterns of Basal-like samples are different from the full dataset patterns. There are no clusters with strict up- or downregulation trend in cancer, instead, there are patterns with additional up- or downregulation in stage 4 (Appendix D; Clusters 1 and 4). Additionally, there are two new patterns, in which the expression is the same in normal and stages 1-3, but drastically up- or downregulated in stage 4 (Appendix D; Clusters 2 and 6). Autophagy is enriched in Basal-like samples also in the pattern of downregulation in cancer (Cluster 4, individual cluster is shown in Figure 3.14), but the pattern has extra downregulation at stage 4. Cluster 4 contains 177 autophagy genes, and enrichment is defined by odds-ratio score 1.25 and adjusted p-value 0.03.

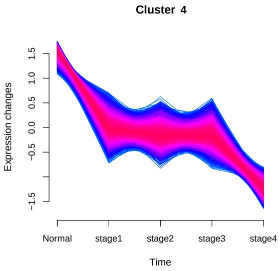

**Figure 3.14:** Basal-like downregulation cluster

**Luminal A subtype**

The expression in Luminal A samples is similar to the all-samples pattern. There are clusters with exclusive up- and downregulation in all cancer stages (Appendix D; Clusters 3 and 4). There are also clusters with the patterns observed in Basal-like – relatively unchanged expression until stage 3, and then rapid raise/drop in stage 4. In Luminal A, however, downregulation cluster (Figure 3.15) is not found to be enriched for autophagy at 'all autophagy genes' level, as it is for all and Basal-like-only samples, even though it is still the most gene-abundant cluster like in the other subtypes (Appendix D). Here, autophagy enrichment is detected only at TFs level: 36 genes, odds-ratio 2.92, adjusted p-value 1.1e-05.

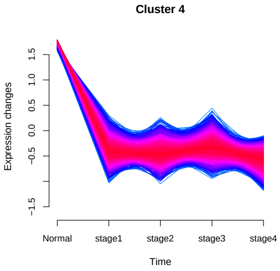

**Figure 3.15:** Luminal A downregulation cluster

**Luminal B subtype**

The samples of Luminal B subtype have gene expression patterns the most similar to full dataset than all other subtypes. Again, the downregulation cluster (Figure 3.16) contains the largest number of autophagy genes, but the enrichment is only significant in TFs gene set (23 genes, odds-ratio 2.15, adjusted p-value 0.01).

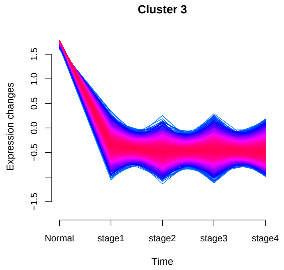

**Figure 3.16:** Luminal B downregulation cluster

**HER2-enriched subtype**

Gene clustering based on HER2-enriched samples generated overall less stable clusters. In Appendix D it can be seen that the cluster cores contain less samples with high membership score (i.e. appear less 'red'). Expression patterns are similar to the ones observed previously. Again, the downregulation cluster (Figure 3.17) contains the largest number of autophagy genes, but the enrichment is only significant in TFs gene set (21 genes, odds-ratio 2.13, adjusted p-value 0.02).

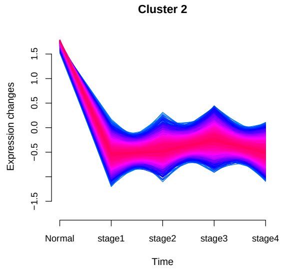

**Figure 3.17:** HER2-enriched downregulation cluster

**Normal-like subtype**

Unlike other subtypes, for Normal-like subtype, there are only samples of stages 1-3 available. Clustering of these samples has produced a variety of new cluster patterns not seen for other subtypes in addition to the reoccurring up- and downregulation trends. The number of genes in each cluster at $>$ 0.6 level is noticeably lower. As a consequence, there is no autophagy enrichment in any of the clusters for any of the gene sets. Figure 3.18 shows the downregulation cluster based on Normal-like samples.

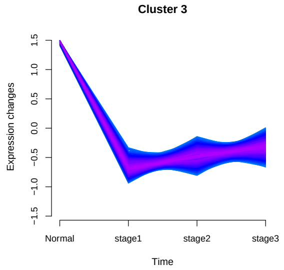

**Figure 3.18:** Normal-like downregulation cluster

#### 3.2.2.3 Summary

Overall, it seems that the downregulation trend characterises the autophagy genes behaviour the most. Both in all samples and in individual subtypes it is the largest cluster with the highest number of autophagy-related genes, and it is seen enriched at several instances.

The downregulation pattern is enriched for 'all autophagy' gene set in all-samples group and in Basal-like subtype. Then, for 'autophagy TFs' gene set it is enriched in all samples, Luminal A, Luminal B, and HER2-enriched subtypes. The exact pattern of downregulation cluster varies slightly between subtypes. In Luminal B the trend is ideally straight; in Luminal A and HER2-enriched the cluster has a very slight dip at stage 4; and in Basal-like it is changes between stage 3 and 4 as much as between normal and stage 1. Normal-like subtype has the weakest clusters with the least number of genes being $>$ 0.6 members, which is perhaps a result of having less samples in total (Table 3.1) – this may also be affecting the HER2-enriched group. In addition to that, not having stage 4 samples in Normal-like subtype, i.e. having a different number of time points, has brought out different expression patterns in this subtype.

## 3.3 Comparative Analysis and Interpretation

The results of differential expression analysis and soft-clustering overall agree on the fact that autophagy signature in the given dataset is observed among the genes downregulated in various breast cancer subtypes when compared to normal. Therefore, it is important to investigate the extent of this agreement both between the two methods, and also among subtypes and stages in separate results.

### 3.3.1 Identifying Candidate Genes

Gene expression clustering of individual PAM50 subtype samples has produced six clusters for each subtype, among which the downregulation trend clusters were the most autophagy genes-abundant (Figures 3.14-3.18). Therefore, it was interesting to investigate whether the autophagy genes with this expression pattern are the same across different subtypes. Thus, the downregulation cluster autophagy genes with > 0.6 membership score were extracted and overlapped. The resulting overlap is shown in Figure 3.19.

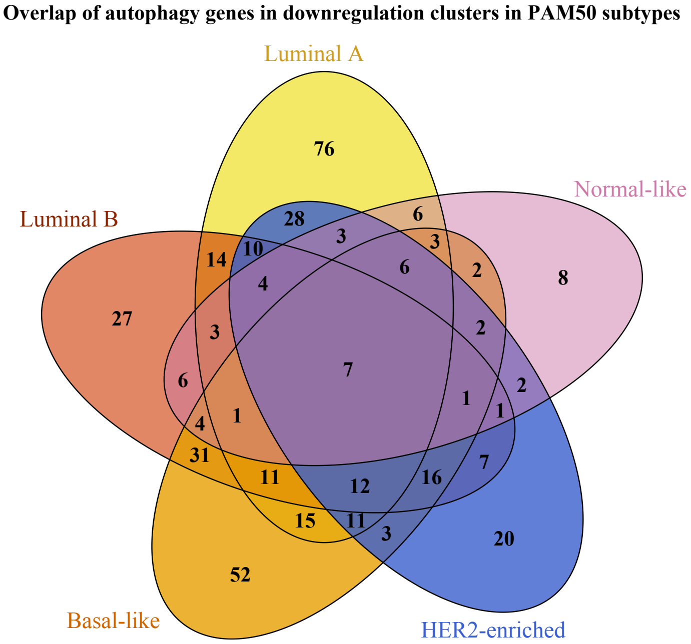

**Figure 3.19:** Venn diagram showing overlap between PAM50 subtype-specific autophagy genes assigned to downregulation clusters in soft-jclustering of individual subtypes. The diagram was made with `VennDiagram` R package [129].

Remarkably, each subtype has a fairly large number of genes unique to its downregulation cluster, e.g. Luminal A – 76 genes. There is a very low number of genes that are downregulated in all subtypes – 7 genes; or 19 (7+12) genes, if not taking into account Normal-like clustering data, as it may be affected by only having three stages. Luminal A and B share 62 genes, Luminal A and Basal-like - 67 genes , and Luminal A and Basal-like – 83, but between three of them, only 11 genes. This suggests that, overall, different subsets of autophagy genes are downregulated in different PAM50 subtypes.

In differential expression analysis, one of the tested models (Group1: PAM50+stage) allowed the comparison of each individual stage of every subtype to normal samples. In this way, downregulated autophagy genes at each stage in every subtype were identified. Table 3.3 shows the numbers of differentially expressed autophagy genes in each subgroup combination (subtype+stage) vs normal.

**Table 3.3:** Count of differentially expressed autophagy genes detected in contrasts of Group1 model. Gene counts for each stage of every PAM50 subtype in contrast to normal are shown.

| *DE contrast (vs normal)* | upregulated genes | downregulated genes |
|---|---|---|
| Luminal A, stage 1 | 65 | 122 |
| Luminal A, stage 2 | 64 | 127 |
| Luminal A, stage 3 | 56 | 116 |
| Luminal A, stage 4 | 30 | 59 |
| Luminal B, stage 1 | 65 | 147 |
| Luminal B, stage 2 | 77 | 156 |
| Luminal B, stage 3 | 86 | 154 |
| Luminal B, stage 4 | 74 | 136 |
| Basal-like, stage 1 | 81 | 152 |
| Basal-like, stage 2 | 83 | 182 |
| Basal-like, stage 3 | 72 | 173 |
| Basal-like, stage 4 | 59 | 141 |
| HER2-enriched, stage 1 | 75 | 155 |
| HER2-enriched, stage 2 | 74 | 182 |
| HER2-enriched, stage 3 | 82 | 173 |
| HER2-enriched, stage 4 | 43 | 141 |
| Normal-like, stage 1 | 28 | 49 |
| Normal-like, stage 2 | 37 | 48 |
| Normal-like, stage 3 | 19 | 22 |

By overlapping the these genes, it is possible to get an insight into how many genes are downregulated in all stages of one subtype (i.e. complete downregulation), and also how many are specific to a particular stage and how that compares across different subtypes.

Figure 3.20 shows such overlap for Basal-like subtype. There are 83 autophagy genes that are downregulated in all stages. 54 genes are downregulated in stages 1-3 but not in stage 4. In stage 4 there is quite a high number of genes (38) downregulated uniquely there.
Figure 3.21 shows the stages overlap for Luminal A. Again, as for Basal-like, the two largest overlap group are between all stages (54 genes) and between stages 1-3 only (57 genes). Interestingly, in stage 4 of Luminal A there are very few uniquely downregulated genes.

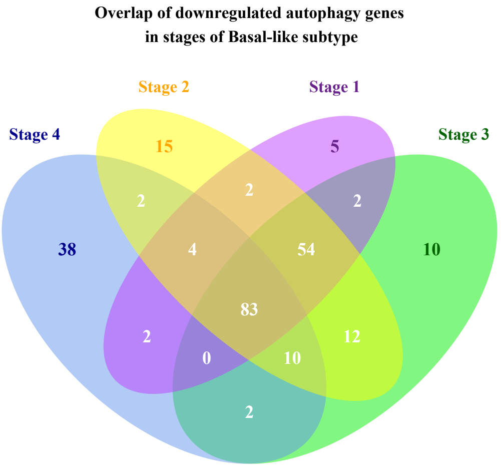

**Figure 3.20:** Overlap between downregulated autophagy genes in stages of Basal-like subtype

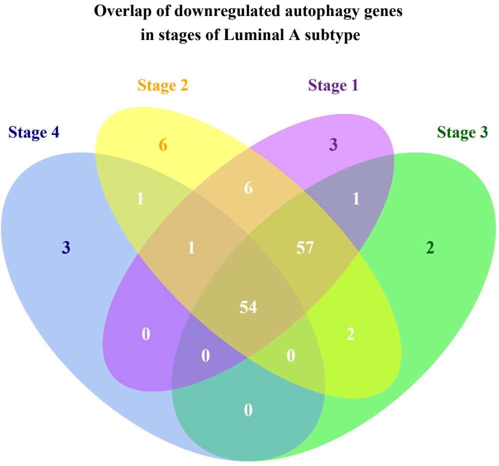

**Figure 3.21:** Overlap between downregulated autophagy genes in stages of Luminal A subtype

The analogous Venn diagrams for Luminal B, HER2-enriched, and Normal-like subtypes are available from Appendix E.

In Luminal B there are 87 genes shared by all stages, 40 genes shared by stages 1-3, and 31 genes unique to stage 4, showing a similar scenario to Basal-like subtype.
In HER2-enriched, the largest overlap group is stages 1-3 (79 genes), whereas all stages share only 32 genes.
Normal-like subtype is a slightly different case, as in addition to its inherent similarity to normal samples, there are no stage 4 samples, which cumulatively resulted in fewer DEGs in total. In Normal-like subtype, there are only 12 genes shared by the available stages (1-3).

Following the above comparisons, it can be seen that a large proportion of downregulated genes are shared by all stages.
This is the expression pattern that is found to be most enriched in clustering analysis too. Hence, the next step is to make a comparison between autophagy genes that

- are downregulated at all stages according to differentially expression analysis (‘*DEA*’), and
- are found in the most autophagy-enriched downregulation cluster (‘*cluster*’)

for individual subtypes.

The PAM50 subtypes have to be considered individually, as Figure 3.19 has shown that there is no solid set of genes that are consistently downregulated genes across all subtypes.

The comparison results are shown in Figure 3.22. The two sets of genes that are being compared for each subtype are labelled as *‘DEA’* and *‘cluster’* for presentation simplicity. The overlap between two sets represents autophagy genes that are downregulated at all stages according to both differential expression testing and soft-clustering analysis. It can be seen that the two methods overall agree, i.e. the majority of *‘DEA’* set genes are present within the larger *‘cluster’* gene set.

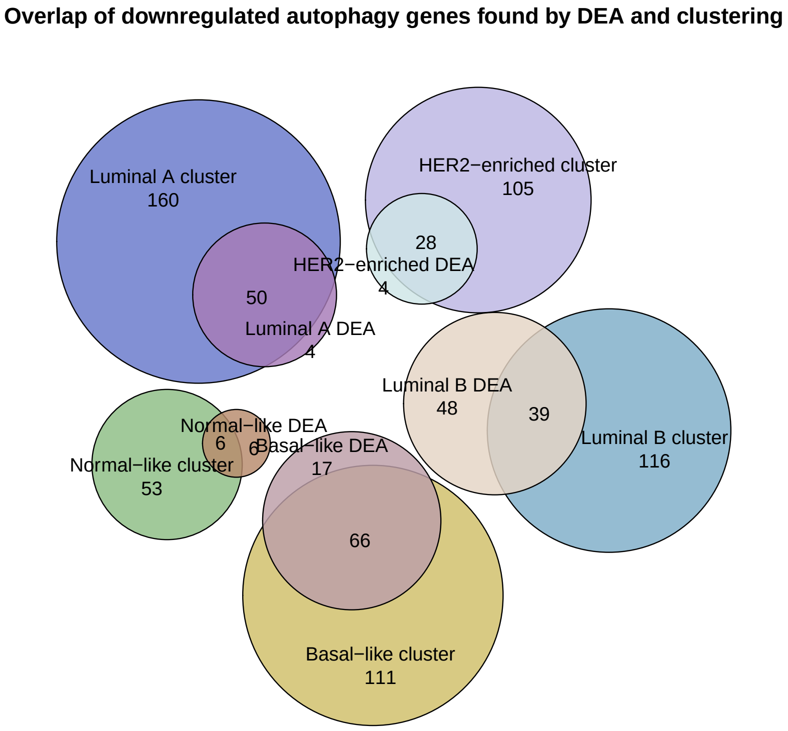

**Figure 3.22:** Subtype-specific overlaps between two autophagy gene sets: **DEA** – genes that are downregulated at all stages according to differentially expression analysis, **cluster** – genes found in the downregulation cluster, both for a specific subtype. The diagram was made with `eulerr` R package [130].

In this way, five subtype-specific lists of candidate autophagy genes were obtained (intersections in Figure 3.22). The following section focuses on exploring functional groups of these genes, while making comparisons between the subtypes, and evaluating the findings with a perspective on biological interpretation.

### 3.3.2 Exploring Candidate Genes

The full list of functional groups of autophagy genes used in the analysis is presented in Table 2.4 in the methods section.
The diagram in Figure 3.23 below provides an overview of the functions of the final sets of subtype-specific candidate genes, i.e. autophagy-related genes downregulated in cancer according to the agreement between two methods.

In the diagram, the genes belonging to each PAM50 subtype are shown as columns of connected boxes. For each subtype, the genes are grouped and coloured-coded by their functional category (shown on the left) into boxes, which are also aligned horizontally to guide visual perception. In total, there are 108 genes that occur in at least one subtype. Many genes are found in several subtypes, and also may appear in several functional categories within one subtype.

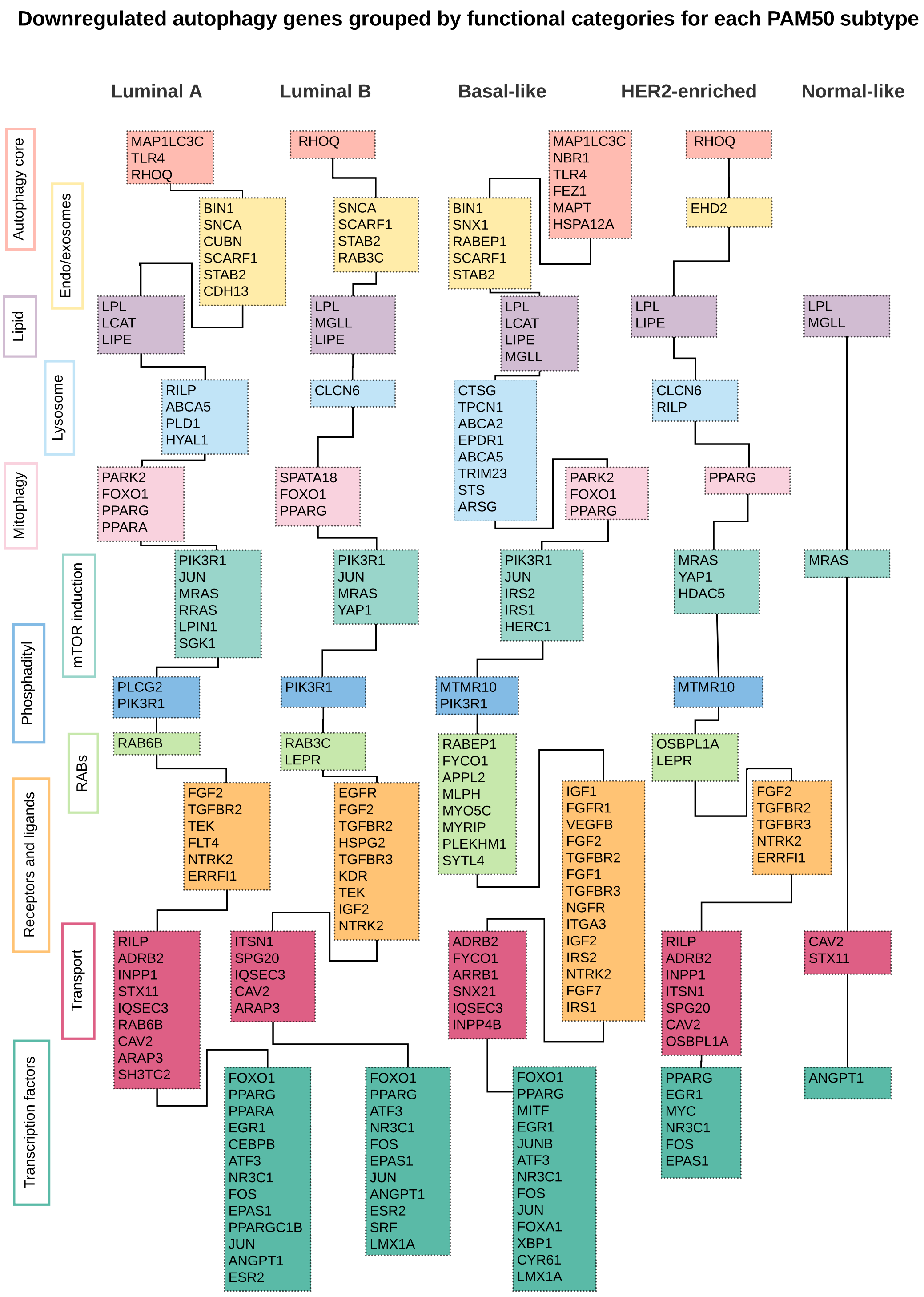

**Figure 3.23:** Diagram of autophagy genes downregulated in each PAM50 subtype grouped by functional categories indicating their role in the autophagy mechanism. Functional groups are colour-coded (group names are shown on the left).

The most gene-abundant groups are transcription factors, receptors/ligands, transport, and mTOR induction-related genes. However, among the smaller groups there are also several potentially interesting candidate genes. Below are some examples:

- **MAP1LC3C (LC3)**. This gene is found downregulated at all stages of Luminal A and Basal-like subtypes. LC3 is an integral gene in the autophagy core pathway. LC3 protein is a part of the autophagosome membrane, and it is often used as marker of autophagy induction (described in Section 1.3.2). Here, MAP1LC3C is one of the three human homologues of rat LC3 [131]. The three homologues were isolated and studied and were found to exhibit distinct expression patterns in different human tissues [131].
- **NBR1** also belongs to the autophagy core functional group and is found downregulated in Basal-like subtype. NBR1 is one of the adapter proteins that LC3 recruits to the autophagosomes to mediate selective autophagy of different cellular cargos through its LC3-interacting region (LIR). Misfolded and aggregated proteins are marked by ubiquitin, which is recognized by ubiquitin-binding domains of NBR1, which later targets the recognised substrates for autophagy [90]. Therefore, disruption of NBR1 function may contribute to defective protein/organelle turnover.
- **PIK3R1**. The PIK3R1 gene product is one on the subunits of the class I PIK3 complex that has a major role in mTOR-dependent autophagy regulation. This gene has been characterised by 2012 TCGA study [45] as one of newly discovered contributors to breast cancer development. It is downregulated in Luminal A, Luminal B, and Basal-like subtypes.

Although the current list of genes is quite focused on the expression behaviour of interest, it is still quite extensive and, therefore, has to be narrowed down to find hits for further investigation. One approach is to check if any of these genes are known tumour suppressor genes in breast cancer. Downregulation of such genes would be an interesting direction to explore experimentally.

A list of known tumour suppressor genes in breast cancer was extracted from the *TSGene database* [132]. The database provides a manually curated list of 589 human tumour suppressor genes downregulated in tumour versus normal in breast invasive carcinoma samples from TCGA. This list was compared to the current set of candidate genes and 18 matches were found. These genes are: CDH13, YAP1, EGR1, EPAS1, ESR2, FOXO1, BIN1, IGF1, ATF3, NGFR, PARK2, PLD1, ERRFI1, PPARA, PPARG, TGFBR2, TGFBR3, INPP4B.

The roles of identified genes were explored with a variety of published and online resources, and a number of them were found to have a specific role in the signalling pathways related to autophagy, hence giving a solid direction for further investigation. Below are the most interesting examples:

- **FOXO1.** The FOXO family proteins are involved in control of cell growth and differentiation and were one of the first transcriptional regulators to be linked to autophagy [133]. FOXOs are regulated by phosphorylation and, when activated, they translocate to the nucleus to induce the expression of a number of autophagy-related genes, including ATG12, BECN1, BNIP3, MAP1LC3, ULK1, and PIK3C3 [60]. Phosphorylation of FOXO1 and FOXO3 by Akt/PKB prevents them from translocating to the nucleus and activating genes that promote cell death, so mTOR promotes cell survival through Akt-mediated inhibition of FOXO1 and FOXO3 [66]. In summary, members of the FOXO family can act as autophagy inducers and repressors depending on their cellular localisation [134].
- **PARK2/parkin** is an E3 ligase that is important for mitochondrial homeostasis [62]. As described in Section 1.3.2, once PARK2 is activated by PINK1 it ubiquitinates proteins on the surface of damaged mitochondria, thereby targeting them for mitophagy [60]. Conversely, PARK2 has a negative role in non-selective autophagy. PARK2 stabilizes Bcl-2 via mono-ubiquitination, which strengthens the interaction between Bcl-2 and Beclin 1, thus suppressing autophagy [135].
- **IGF1**, insulin-like growth factor 1, is responsible for the activation of IRS1/2 (insulin receptor substrates 1/2), which mediate the signalling cascades through class I PIK3 that activate PKB, the best-characterised positive modifier of mTOR activity, resulting in negative regulation of autophagy induction, as depicted in Figure 1.4. Therefore, autophagy may be activated by IGF-1 inhibition. This aspect was explored and reviewed in [136]. Of note, the IRS1/2 are also downregulated in Basal-like subtype along with IGF1.

# Chapter 4

# Discussion

Understanding of autophagy role in cancer is becoming an increasingly intriguing area of research as the two constituting fields – breast cancer and autophagy – have seen a lot of progress individually in the past decade. This project aimed to identify the expression signatures of autophagy-related genes in the breast cancer transcriptomics data available from TCGA.

Rigorous exploratory analysis was applied to the dataset to get an understanding of the main patterns characterising the available samples. Three main breast cancer classification methods that were used to stratify the patients into groups are cancer stage, tumour morphology, and PAM50 intrinsic molecular subtype. One of the main objectives of gene expression analysis was to identify a subset of samples described by those classification methods with a distinct autophagy-related signature. To do this, two complementary analysis methods were used – differential expression testing and soft-clustering. In differential expression analysis, a multitude of contrasting tests were performed across different combinations of subgroups defined by classification methods, and the identified sets of differentially expressed genes were tested for autophagy enrichment. Soft-clustering aimed to detect clusters of genes that have similar collective expression behaviour with respect to their changes along cancer stages. Then, enrichment analysis was used to identify expression patterns where autophagy genes are overrepresented. Finally, a comparative analysis of two methods' results produced a list of autophagy genes that can be used as a starting point for further exploration of autophagy role in cancer.

Exploratory analysis of TCGA-BRCA dataset was a crucial part of this project. Firstly, it has confirmed the view that is now widely accepted in the breast cancer research community on the extreme heterogeneity of this disease, and the fact that cancer patients can most accurately be classified into prognostic groups by their intrinsic molecular profile rather than by tumour morphology and stage. The major differences between PAM50 subtypes were highlighted by PCA, hierarchical clustering, as well as differential expression testing, where individual subtypes had considerably more DEGs between each other than tumour morphologies or stages. In fact, neither exploratory analysis nor differential expression testing have found any strong differences between morphological groups or stages. The PCA plots for both did not reveal any PCs that describe the proposed sample subgroups sufficiently well. DE analysis found exceptionally low number of DEGs between cancer stages (even stage 1 and stage 4), and also not a great amount of DEGs between two major established morphologies, Lobular and Ductal carcinomas. This result was surprising, but the proposed explanation was in the underlying imbalanced composition of the subgroups.

The imbalanced composition and sample size of the subgroups defined by classification methods became apparent in the exploratory analysis. The composition of subgroups in terms of the constituting PAM50 subtypes is important, as this classification method explains the main driving force that defines variation between samples. Therefore, there are two implications from this. First, if the groups are too mixed/heterogeneous (i.e. made up of various PAM50 subtypes), then it is difficult to capture differences between them, as the variation signal is too mixed and thereby weak. Second, if two groups that are being compared are exclusively made up of two different PAM50 subtypes, then the difference between them is likely to be driven by PAM50. A clear example of this was found when comparing Mucinous and Metaplastic morphologies, which are exclusively made up of Luminal and Basal samples, respectively. These two morphologies had the largest number of DEGs detected between them in comparison to the rest, in addition to being very clearly separated by PCA. However, the true source of their variation is difficult to define. To tackle the problem of heterogeneity created by the strong influence of PAM50 profiles on other classifications, DEA with combined models was performed. It allowed to characterise differential expression between individual stages or morphologies within each PAM50 subtype, thereby creating more opportunities to test for autophagy signatures.

However, the results of enrichment analysis on DEGs were ambiguous and largely unsuccessful in identifying autophagy signatures in any of the specific sample groups. The main observed reocurring trend was the enrichment of autophagy-related genes among the downregulated genes in various cancer subgroups (including individual morphologies and stages) versus normal. Interestingly, there were no autophagy enrichment signatures between different stages of cancer, which is something that is often brought up in the literature to describe the behaviour of autophagy processes in cancer. Similarly, no significant autophagy enrichment was identified in the genes differentially expressed between the PAM50 cancer subtypes. Therefore, it was concluded that the main observed and statistically quantified autophagy signature in this breast cancer dataset is the overall downregulation of autophagy-related genes in cancer versus normal, regardless of stage or subgroup. In line with that, soft-clustering analysis results have shown that the majority of autophagy genes are assigned to the cluster with expression pattern representing downregulation at all stages. This cluster pattern was found to be significantly enriched for autophagy, both at all genes or transcription factors-only level, and in separate PAM50 subtypes.

Overall, it is important to highlight the significance of finding the trend of downregulation in cancer to be the main autophagy signature in this dataset. In a broad sense, it makes sense to see a subset of autophagy-related genes to be downregulated in cancer. Deregulated autophagy means that the basal turnover of damaged macromolecules is impaired, which possesses an oncogenic impact. Additionally, seeing a group of autophagy genes downregulated consistently at all stages compared to normal, gives an indication of there being a part of the autophagy pathway that is inactivated is cancer in general.

However, the literature on autophagy role in cancer often brings up the ‘dual role’ behaviour, which involves upregulation of autophagy at different stages of cancer progression with opposite consequences (cytotoxicity and cytoprotection of cancer cells). The results of this project did not show any statistically significant indication of autophagy genes being upregulated at any specific stage or cancer in general. Although among the soft-clustering results there were several clusters of genes with potentially interesting expression patterns in relation to stage-specific upregulation, the enrichment analysis did not indicate that those clusters possess significant autophagy signatures.

The inability to identify any autophagy signatures that would be in agreement with the literature-proposed expression patterns or that are cancer subtype-specific (as opposed to just cancer vs normal) and are significant (or significant at non-adjusted p-value level, as p-value cut-offs are arbitrary), leads to the discussion of the potential limitations of this study. Firstly, it has to be reiterated that subgroup imbalance is very likely to be contributing to results both of exploratory and gene expression analyses. In addition to PAM50 composition influence discussed above, the number of samples in each group is also of major concern. For instance, as in the current dataset there are only 12 samples of stage 4 (total $n = 857$), and all of them are classified as one morphology, the conclusions made from analyses involving those samples have to regarded with caution. This is especially the case for when those samples are stratified by PAM50 subtype, as it results in only two samples representing the gene expression of Basal-like and HER2-enriched samples in differentially expression testing and soft clustering (see Table 3.1).

Hence, these dataset constraints may be the reason why no subtype-specific autophagy signatures were detected. Moreover, a possible interpretation of not seeing the literature-ascertained autophagy upregulation at later cancer stages that is related to metastasis initiation, is due to the fact that the primary tumour samples constituting this dataset are not a fit representation of cancer behaviour at later stages (even stage 4). Therefore, to see the true upregulation of autophagy as a result of cancer cells using it for their own benefit, the actual metastasis samples need to be analysed.

Another important aspect that was not taken into account in this project is the information on treatments that each patient has received. This information was not readily available at the time when it could have been utilised to benefit the project. However now, it is clear that having this information would have perhaps shed a light on the observed autophagy downregulation trend, or, rather the opposite – the lack of upregulation, as some treatments can have activating or inhibitory effect on autophagy.

In spite of the lack of subtype-specific signatures and overall dataset limitations, the comparative analysis of downregulated genes identified by two gene expression analysis methods was carried out to investigate the biological relevance of the putative autophagy signature. For each PAM50 subtype, a list of candidate genes that exhibit complete downregulation in cancer was produced. Among those genes, some interesting hits were identified, some of which are involved both in the key steps of autophagosome formation and the signalling pathways that conduct the best-characterised autophagy induction/inhibition.

## 4.1 Future directions

The main direction from the current phase of the project is to seek collaboration with the other units at DCRC who have the expert knowledge of autophagy and the related processes. The identified list of candidate genes will be made available to them, and with their help, a more informed biological interpretation of the potential impact of downregulation of the identified genes can be achieved. Following that, particularly interesting genes can be sent for experimental validation in MCF7 breast cancer cell lines available at DCRC.

This is, however, not the end of the investigation of autophagy signatures in the breast cancer data. TCGA is a rich source of data, but the aforementioned constraints make drawing statistically significant conclusions from the results quite difficult. Therefore, the analysis should be replicated on another, more balanced cohort of breast cancer patients.

With the current results at hand, there are several directions where future analyses could go. Firstly, it would be interesting to explore the results of the differential expression analysis and soft-clustering separately. The results could be used as a guide for network analysis aiming to explore the relationship between deregulated genes and how their protein products interact together or modify each other, which will be possible by using the available experimental information on protein-protein interaction (PPI) networks deposited in databases (e.g. STRING). Additionally, retrieving information on the transcription factors and miRNAs that are known to regulate the differentially expressed autophagy genes could be used to analyse correlations in their own and their targets' expression levels changes. It would be interesting to specifically identify those transcription modifiers that are altered in a certain breast cancer subtypes/stages/morphologies and for which the alteration is also reflected in the expression of the target autophagy genes.

## 4.2 Conclusion

The over-arching observation made in this project is the presence of autophagy-related signature among the genes that are downregulated in cancer versus normal. The widely accepted view on autophagy being upregulated at different stages of cancer progression was not confirmed by the results of data analysis in this study. The comparative analysis has helped to identify a set of autophagy-related genes that are downregulated at all stages in specific breast cancer subtypes according to both analysis methods. Obtaining this list of genes has opened a new avenue for research into the part of the autophagy mechanism that appears to be inactivated in cancer cells at all stages. A brief look at functions of the genes has shown that many of them are involved in the core steps of autophagy regulation, which is quite promising in itself.

To conclude, understanding of the convoluted relationship between autophagy and cancer is still at very preliminary phase. When greater understanding of the signalling that both regulates and executes autophagy in cancer is achieved, targeting specific pathway components as a part of cancer treatment would maximise the benefit to breast cancer patients.

# Acknowledgements

I would like to thank the Computational Biology Laboratory at the Danish Cancer Research Centre for allowing me to join their group to write my Master's thesis with them as a part of my degree at the University of Copenhagen.

I am very thankful to Dr Elena Papaleo, the group leader, for her professional guidance and valuable support throughout the project.

I would also like to express my gratitude to Dr Kristoffer Vitting-Seerup, who provided internal supervision of my project. His ability to explain difficult bioinformatics concepts and show how to correctly use the tools to answer biological questions has been greatly appreciated and has enabled me to complete high quality work.

I would also like to thank Prof Albin Sandelin for agreeing to supervise this project, and more importantly, for making the MSc courses on high-throughput data analysis that precede the thesis part of the degree extremely interesting, as this is what has motivated me to do my thesis in this research field.

Finally, I would like to acknowledge and thank the colleagues at DCRC (in the Cell Death & Metabolism and Cell Stress & Survival Units) for providing a list of expert-curated autophagy genes to use in this project.

# Bibliography

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012", *International journal of cancer*, vol. 136, no. 5, 2015.

[2] C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, D. J. Dicker, O. Chimed-Orchir, R. Dandona, L. Dandona, and T. Fleming, "Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015", *JAMA Oncology*, vol. 3, no. 4, p. 524, Apr. 2017.

[3] *Breast cancer statistics — World Cancer Research Fund International*, 2015. [Online]. Available: `http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics`.

[4] M. Vidal, L. Paré, and A. Prat, "Molecular Classification of Breast Cancer", in *Management of Breast Diseases*, Springer International Publishing, 2017, ch. Molecular, pp. 2003–219.

[5] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, and B. Shi, "Breast cancer intrinsic subtype classification, clinical use and future trends.", *American journal of cancer research*, vol. 5, no. 10, pp. 2929–43, 2015.

[6] F. M. Blows, K. E. Driver, M. K. Schmidt, A. Broeks, F. E. van Leeuwen, J. Wesseling, and M. C. Cheang, "Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: A collaborative analysis of data for 10,159 cases from 12 studies", *PLoS Medicine*, vol. 7, no. 5, 2010.

[7] T. Iwamoto and L. Pusztai, "Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data?", *Genome medicine*, vol. 2, no. 11, p. 81, Nov. 2010.

[8] B. Weigelt, F. L. Baehner, and J. S. Reis-Filho, "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: A retrospective of the last decade", *The Journal of pathology*, vol. 220, no. 2, pp. 263–280, 2010.

[9] L. Pusztai, K. Broglio, F. Andre, W. F. Symmans, K. R. Hess, and G. N. Hortobagyi, "Effect of molecular disease subsets on disease-free survival in randomized adjuvant chemotherapy trials for estrogen receptor–positive breast cancer", *Journal of Clinical Oncology*, vol. 26, no. 28, pp. 4679–4683, 2008, PMID: 18662965.

[10] *AJCC - What is Cancer Staging?*, 2017. [Online]. Available: `https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx`.

[11] A. E. Giuliano, J. L. Connolly, S. B. Edge, E. A. Mittendorf, H. S. Rugo, L. J. Solin, D. L. Weaver, D. J. Winchester, and G. N. Hortobagyi, "Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual", *CA: A Cancer Journal for Clinicians*, vol. 67, no. 4, pp. 290–303, Jul. 2017.

[12] J. C. Scatarige, I. Boxen, and R. L. Smathers, "Internal mammary lymphadenopathy: Imaging of a vital lymphatic pathway in breast cancer.", *Radiographics*, vol. 10, no. 5, pp. 857–870, 1990.

[13] J. Makki, "Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance", *Liberatas Academica. Clinical Medicine insights: Pathology*, no. 8, 2015.

[14] B. Weigelt, H. Horlings, B. Kreike, M. Hayes, M. Hauptmann, L. Wessels, D. de Jong, M. Van de Vijver, L. V. Veer, and J. Peterse, "Refinement of breast cancer classification by molecular characterization of histological special types", *The Journal of Pathology*, vol. 216, no. 2, pp. 141–150, Oct. 2008.

[15] R. Walker, “World health organization classification of tumours. pathology and genetics of tumours of the breast and female genital organs”, *Histopathology*, vol. 46, no. 2, pp. 229–229, 2005.

[16] *International Classification of Diseases for Oncology. World Health Organisation (WHO). Morphological codes, 3rd Ed.* 2011. [Online]. Available: http://codes.iarc.fr/codegroup/2.

[17] T. Gathani, D. Bull, J. Green, G. Reeves, and V. Beral, “Breast cancer histological classification: agreement between the Office for National Statistics and the National Health Service Breast Screening Programme”, *Breast Cancer Research*, vol. 7, no. 7, pp. 1090–1096, 2005.

[18] D. M. Ramnani, *Webpathology.com: A Collection of Surgical Pathology Images*, 2016. [Online]. Available: http://www.webpathology.com/category.asp?category=52.

[19] P. Abdelmessieh and J. Katz, *Breast Cancer Histology: Overview*, 2016. [Online]. Available: http://emedicine.medscape.com/article/1954658-overview#a1.

[20] B. Weigelt, F. C. Geyer, R. Natrajan, M. A. Lopez-Garcia, A. S. Ahmad, K. Savage, B. Kreike, and J. S. Reis-Filho, “The molecular underpinning of lobular histological growth pattern: A genome-wide transcriptomic analysis of invasive lobular carcinomas and grade- and molecular subtype-matched invasive ductal carcinomas of no special type”, *The Journal of pathology*, vol. 220, no. 1, pp. 45–57, 2010.

[21] G. Ciriello, M. L. Gatza, A. H. Beck, T. A. King, M. D. Wilkerson, S. K. Rhie, A. Pastore, H. Zhang, M. Mclellan, C. Yau, C. Kandoth, R. Bowlby, H. Shen, S. Hayat, R. Fieldhouse, S. C. Lester, G. M. K. Tse, R. E. Factor, L. C. Collins, K. H. Allison, Y.-Y. Chen, A. D. Cherniack, G. Robertson, C. Benz, C. Sander, P. W. Laird, K. A. Hoadley, and C. M. Perou, “Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer”, *CELL*, vol. 163, pp. 506–519, 2015.

[22] T. L. Schwartz, H. Mogal, C. Papageorgiou, J. Veerapong, and E. C. Hsueh, “Metaplastic breast cancer: Histologic characteristics, prognostic factors and systemic treatment strategies”, *Experimental hematology & oncology*, vol. 2, no. 1, p. 31, 2013.

[23] B. Weigelt, F. C. Geyer, and J. S. Reis-Filho, “Histological types of breast cancer: How special are they?”, *Molecular Oncology*, vol. 4, no. 3, pp. 192–208, Jun. 2010.

[24] A. Dumitru, A. Procop, A. Iliesiu, M. Tampa, L. Mitrache, M. Costache, M. Sajin, A. Lazaroiu, and M. Cirstoiu, “Mucinous breast cancer: A review study of 5 year experience from a hospital-based series of cases”, *Maedica*, vol. 10, no. 1, p. 14, 2015.

[25] D. C. Zaha, “Significance of immunohistochemistry in breast cancer.”, *World journal of clinical oncology*, vol. 5, no. 3, pp. 382–92, 2014.

[26] M. Cianfrocca and L. J. Goldstein, “Prognostic and predictive factors in early-stage breast cancer”, *The oncologist*, vol. 9, no. 6, pp. 606–616, 2004.

[27] E. Stickeler, “Prognostic and predictive markers for treatment decisions in early breast cancer”, *Breast Care*, vol. 6, no. 3, pp. 193–198, 2011.

[28] E. B. C. T. C. Group *et al.*, “Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials”, *The Lancet*, vol. 365, no. 9472, pp. 1687–1717, 2005.

[29] N. Iqbal and N. Iqbal, “Human epidermal growth factor receptor 2 (her2) in cancers: Overexpression and therapeutic implications”, *Molecular biology international*, vol. 2014, 2014.

[30] W. D. Foulkes, I. E. Smith, and J. S. Reis-Filho, “Triple-negative breast cancer”, *New England journal of medicine*, vol. 363, no. 20, pp. 1938–1948, 2010.

[31] C. A. Hudis and L. Gianni, “Triple-negative breast cancer: An unmet medical need”, *The oncologist*, vol. 16, no. Supplement 1, pp. 1–11, 2011.

[32] B. Weigelt, F. Baehner, and R.-F. Jorge, “The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade”, *The Journal of pathology*, vol. 220, no. September, pp. 114–125, 2010.

[33] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale, “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10 869–10 874, 2001.

[34] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-

Dale, and D. Botstein, “Repeated observation of breast tumor subtypes in independent gene expression data sets.”, *PNAS*, vol. 100, no. 14, pp. 8418–23, 2003.

[35] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard, “Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes”, *J Clin Oncol*, vol. 27, pp. 1160–1167,

[36] M. Gnant, M. Filipits, R. Greil, H. Stoeger, M. Rudas, Z. Bago-Horvath, B. Mlineritsch, W. Kwasny, M. Knauer, C. Singer, R. Jakesz, P. Dubsky, F. Fitzal, R. Bartsch, G. Steger, M. Balic, and S. Ressler, “Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone”,

[37] C. M. Perou, T. Sorlie, M. B. Eisen, M. Van De Rijn, *et al.*, “Molecular portraits of human breast tumours”, *Nature*, vol. 406, no. 6797, p. 747, 2000.

[38] P. Eroles, A. Bosch, J. A. Pérez-Fidalgo, and A. Lluch, “Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways”, *Cancer treatment reviews*, vol. 38, no. 6, pp. 698–707, 2012.

[39] A. Prat, M. C. U. Cheang, M. Martın, J. S. Parker, E. Carrasco, R. Caballero, S. Tyldesley, K. Gelmon, P. S. Bernard, T. O. Nielsen, *et al.*, “Prognostic significance of progesterone receptor–positive tumor cells within immunohistochemically defined luminal a breast cancer”, *Journal of clinical oncology*, vol. 31, no. 2, pp. 203–209, 2012.

[40] J. D. Brenton, L. A. Carey, A. A. Ahmed, and C. Caldas, “Molecular classification and molecular forecasting of breast cancer: Ready for clinical application?”, *Journal of clinical oncology*, vol. 23, no. 29, pp. 7350–7360, 2005.

[41] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, *et al.*, “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer”, *New England Journal of Medicine*, vol. 351, no. 27, pp. 2817–2826, 2004.

[42] L. A. Carey, E. C. Dees, L. Sawyer, L. Gatti, D. T. Moore, F. Collichio, D. W. Ollila, C. I. Sartor, M. L. Graham, and C. M. Perou, “The triple negative paradox: Primary tumor chemosensitivity of breast cancer subtypes”, *Clinical cancer research*, vol. 13, no. 8, pp. 2329–2334, 2007.

[43] C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu, “Breast cancer classification and prognosis based on gene expression profiles from a population-based study”, *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10 393–10 398, 2003.

[44] G. Callagy, E. Cattaneo, Y. Daigo, L. Happerfield, L. G. Bobrow, P. D. Pharoah, and C. Caldas, “Molecular classification of breast carcinomas using tissue microarrays”, *Diagnostic Molecular Pathology*, vol. 12, no. 1, pp. 27–34, 2003.

[45] TCGA Research Network, “Comprehensive molecular portraits of human breast tumours”, *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[46] A. Prat, B. Adamo, C. Fan, V. Peg, M. Vidal, P. Galván, A. Vivancos, P. Nuciforo, H. G. Palmer, S. Dawood, *et al.*, “Genomic analyses across six cancer types identify basal-like breast cancer as a unique molecular entity”, *Scientific reports*, vol. 3, p. 3544, 2013.

[47] C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. Van’t Veer, and C. M. Perou, “Concordance among gene-expression–based predictors for breast cancer”, *New England Journal of Medicine*, vol. 355, no. 6, pp. 560–569, 2006.

[48] E. Rakha, T. Putti, D. Abd El-Rehim, C. Paish, A. Green, D. Powe, A. Lee, J. Robertson, and I. Ellis, “Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation”, *The Journal of pathology*, vol. 208, no. 4, pp. 495–506, 2006.

[49] C. Ho-Yen, R. L. Bowen, and J. Jones, “Characterization of basal-like breast cancer: An update”, *Diagnostic Histopathology*, vol. 18, no. 3, pp. 104–111, 2012.

[50] *The Cancer Genome Atlas*. [Online]. Available: `https://cancergenome.nih.gov/`.

[51] *NCI Genomic Data Commons*. [Online]. Available: `https://gdc.cancer.gov/`.

[52] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, “Toward a shared vision for cancer genomic data”, *New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, 2016, PMID: 27653561.

[53] *Overview of The Cancer Genome Atlas (TCGA).* [Online]. Available: `https://cancergenome.nih.gov/abouttcga/overview`.

[54] Editorial., “The future of cancer genomics”, *Nature Medicine*, vol. 21, no. 2, pp. 99–99, Feb. 2015.

[55] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”, *Contemp Oncol (Pozn)*, vol. 19, no. 1A, pp. 68–77, 2015.

[56] T. C. Silva, A. Colaprico, C. Olsen, F. D’Angelo, G. Bontempi, M. Ceccarelli, and H. Noushmehr, “TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages”, *F1000Research*, vol. 5, p. 1542, 2016.

[57] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr, “TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data”, *Nucleic Acids Research*, vol. 44, no. 8, e71, 2016.

[58] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, “Bioconductor: Open software development for computational biology and bioinformatics”, *Genome biology*, vol. 5, no. 10, R80, 2004.

[59] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, “Orchestrating high-throughput genomic analysis with Bioconductor”, *Nature Methods*, vol. 12, no. 2, pp. 115–121, Jan. 2015.

[60] Y. Feng, Z. Yao, and D. J. Klionsky, “How to control self-digestion : and post-translational regulation of autophagy”, *Trends in Cell Biology*, vol. 25, no. 6, pp. 1–10, 2015.

[61] B. Levine and G. Kroemer, “Autophagy in the pathogenesis of disease”, *Cell*, vol. 132, no. 1, pp. 27–42, 2008.

[62] G. Kroemer, G. Mariño, and B. Levine, “Autophagy and the Integrated Stress Response”, *Molecular Cell*, vol. 40, no. 2, pp. 280–293, 2010.

[63] T. Yorimitsu and D. Klionsky, “Autophagy: molecular machinery for self-eating”, *Cell Death and Differentiation*, vol. 12, pp. 1542–1552, 2005.

[64] Y. Feng, D. He, Z. Yao, and D. J. Klionsky, “The machinery of macroautophagy”, *Nature Publishing Group*, vol. 24, no. 24, pp. 24–4124, 2013.

[65] A. Abada and Z. Elazar, “Getting ready for building: signaling and autophagosome biogenesis.”, *EMBO reports*, vol. 15, no. 8, pp. 839–852, 2014.

[66] J. Füllgrabe, G. Ghislat, D.-H. Cho, and D. C. Rubinsztein, “Transcriptional regulation of mammalian autophagy at a glance”, *Journal of Cell Science*, vol. 129, no. 16, pp. 3059–3066, 2016.

[67] J. Kaur and J. Debnath, “Autophagy at the crossroads of catabolism and anabolism.”, *Nature reviews. Molecular cell biology*, vol. 16, no. 8, pp. 461–472, 2015.

[68] *The Nobel Prize in Physiology or Medicine 2016.* [Online]. Available: `http://www.nobelprize.org/nobel_prizes/medicine/laureates/2016/`.

[69] Z. Xie and D. J. Klionsky, “Autophagosome formation: core machinery and adaptations.”, *Nature cell biology*, vol. 9, no. 10, pp. 1102–1109, 2007.

[70] Z. Yang and D. J. Klionsky, “Mammalian autophagy: Core molecular machinery and signaling regulation”, *Current Opinion in Cell Biology*, vol. 22, no. 2, pp. 124–131, 2010.

[71] C. He and D. J. Klionsky, “Regulation Mechanisms and Signaling Pathways of Autophagy”, *Annual Review of Genetics*, vol. 43, no. 1, pp. 67–93, 2009.

[72] J. Kim, W.-P. Huang, P. E. Stromhaug, and D. J. Klionsky, “Convergence of multiple autophagy and cytoplasm to vacuole targeting components to a perivacuolar membrane compartment prior tode novo vesicle formation”, *Journal of biological chemistry*, vol. 277, no. 1, pp. 763–773, 2002.

[73] R. Zoncu, A. Efeyan, and D. M. Sabatini, “mTOR: from growth signal integration to cancer, diabetes and ageing”, 2011.

[74] C. A. Lamb, T. Yoshimori, and S. A. Tooze, “The autophagosome: Origins unknown, biogenesis complex”, *Nature reviews Molecular cell biology*, vol. 14, no. 12, pp. 759–774, 2013.

[75] Y. Zhou, E. B. Rucker, and B. P. Zhou, “Autophagy regulation in the development and treatment of breast cancer”, *Acta Biochimica et Biophysica Sinica*, vol. 48, no. 1, pp. 60–74, 2015.

[76] F. Nazio, F. Strappazzon, M. Antonioli, P. Bielli, V. Cianfanelli, M. Bordi, C. Gretzmeier, J. Dengjel, M. Piacentini, G. M. Fimia, and F. Cecconi, “mTOR inhibits autophagy by controlling ULK1 ubiquitylation, self-association and function through AMBRA1 and TRAF6.”, *Nature cell biology*, vol. 15, no. 4, pp. 406–16, 2013.

[77] *Autophagy Signaling Interactive Pathway — Cell Signaling Technology.* [Online]. Available: https://www.cellsignal.com/contents/science-pathway-research-autophagy/autophagy-signaling-pathway/pathways-autophagy.

[78] A. J. Meijer and P. Codogno, “Regulation and role of autophagy in mammalian cells”, *The international journal of biochemistry & cell biology*, vol. 36, no. 12, pp. 2445–2462, 2004.

[79] Z. Yang and D. J. Klionsky, “Eaten alive: a history of macroautophagy.”, *Nature cell biology*, vol. 12, no. 9, pp. 814–22, 2010.

[80] Y. Wei, S. Pattingre, S. Sinha, M. Bassik, and B. Levine, “Jnk1-mediated phosphorylation of bcl-2 regulates starvation-induced autophagy”, *Molecular cell*, vol. 30, no. 6, pp. 678–688, 2008.

[81] F. Colotta, P. Allavena, A. Sica, C. Garlanda, and A. Mantovani, “Cancer-related inflammation, the seventh hallmark of cancer: Links to genetic instability”, *Carcinogenesis*, vol. 30, no. 7, pp. 1073–1081, 2009.

[82] K. Degenhardt, R. Mathew, B. Beaudoin, K. Bray, D. Anderson, G. Chen, C. Mukherjee, Y. Shi, C. Gélinas, Y. Fan, *et al.*, “Autophagy promotes tumor cell survival and restricts necrosis, inflammation, and tumorigenesis”, *Cancer cell*, vol. 10, no. 1, pp. 51–64, 2006.

[83] K. L. Cook, A. N. Shajahan, and R. Clarke, “Autophagy and endocrine resistance in breast cancer.”, *Expert review of anticancer therapy*, vol. 11, no. 8, pp. 1283–1294, 2011.

[84] J. M. Zarzynska and J. Magdalena, “The importance of autophagy regulation in breast cancer development and treatment.”, *BioMed research international*, vol. 2014, p. 710 345, 2014.

[85] P. Maycotte and A. Thorburn, “Targeting autophagy in breast cancer.”, *World journal of clinical oncology*, vol. 5, no. 3, pp. 224–40, 2014.

[86] R. Mathew and E. White, “Autophagy, stress, and cancer metabolism: What doesn’t kill you makes you stronger”, *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 76, pp. 389–396, 2011.

[87] K. Jain, K. S. Paranandi, S. Sridharan, and A. Basu, “Autophagy in breast cancer and its implications for therapy.”, *American journal of cancer research*, vol. 3, no. 3, pp. 251–65, 2013.

[88] E. Mowers, M. Sharifi, and K. Macleod, “Autophagy in cancer metastasis”, *Oncogene*, vol. 36333, no. 10, pp. 1619–1630, 2017.

[89] Y. Kondo, T. Kanzawa, R. Sawaya, and S. Kondo, “The role of autophagy in cancer development and response to therapy”, *Nature reviews. Cancer*, vol. 5, no. 9, p. 726, 2005.

[90] J. Debnath, “The multifaceted roles of autophagy in tumors - Implications for breast cancer”, *Journal of Mammary Gland Biology and Neoplasia*, vol. 16, no. 3, pp. 173–187, 2011.

[91] P. V. Schoenlein, S. Periyasamy-Thandavan, J. S. Samaddar, W. H. Jackson, and J. T. Barrett, “Autophagy facilitates the progression of er$\alpha$-positive breast cancer cells to antiestrogen resistance”, *Autophagy*, vol. 5, no. 3, pp. 400–403, 2009.

[92] A. Vazquez-Martin, C. Oliveras-Ferraros, and J. A. Menendez, “Autophagy facilitates the development of breast cancer resistance to the anti-her2 monoclonal antibody trastuzumab”, *PloS one*, vol. 4, no. 7, e6251, 2009.

[93] E. White, “The role for autophagy in cancer”, *Journal of Clinical Investigation*, vol. 125, no. 1, pp. 42–46, 2015.

[94] M. Rahman, L. K. Jackson, W. E. Johnson, D. Y. Li, A. H. Bild, and S. R. Piccolo, “Alternative preprocessing of RNA-Sequencing data in the Cancer Genome Atlas leads to improved analysis results”, *Bioinformatics*, vol. 31, no. 22, pp. 3666–3672, 2015.

[95] University of North Carolina (UNC) Center for Bioinfromatics, “TCGA mRNA-seq Pipeline for UNC data”, 2013, [Online]. Available: https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/UNC_mRNAseq_summary.pdf.

[96] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, *et al.*, "Mapsplice: Accurate mapping of rna-seq reads for splice junction discovery", *Nucleic acids research*, vol. 38, no. 18, e178–e178, 2010.

[97] B. Li and C. N. Dewey, "Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome", *BMC bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[98] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "Gc-content normalization for rna-seq data", *BMC bioinformatics*, vol. 12, no. 1, p. 480, 2011.

[99] M. D. Robinson, D. J. Mccarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics*, vol. 26, no. 1, pp. 139–14 010, 2010.

[100] C. W. Law, M. Alhamdoosh, S. Su, G. K. Smyth, and M. E. Ritchie, "Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger", *F1000Research*, vol. 5, 2016.

[101] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: precision weights unlock linear model analysis tools for RNA-seq read counts", *Genome Biology*, vol. 15, R29, 2014.

[102] M. Robinson, D. McCarthy, Y. Chen, and G. Smyth, "edgeR vingette: differential expression analysis of digital gene expression data", 2010. [Online]. Available: https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/%20edgeRUsersGuide.pdf.

[103] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data", *Genome Biology*, vol. 11, R25, 2010.

[104] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[105] M. I. Love, S. Anders, V. Kim, and W. Huber, "RNA-Seq workflow: gene-level exploratory analysis and differential expression", *F1000Research*, vol. 4, no. 1070, p. 1070, 2015.

[106] H. Wickham, *Ggplot2: Elegant graphics for data analysis*. Springer New York, 2009.

[107] H. Wickham and R. Francois, "Dplyr: A grammar of data manipulation", *R package version 0.4*, vol. 1, p. 20, 2015.

[108] J. T. Leek, R. B. Scharpf, H. Corrada Bravo, D. Simcha, B. Langmead, W. Evan Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data", *Nature Publishing Group*, vol. 11, no. 8, pp. 101–113, 2010.

[109] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis", *PLoS Genetics*, vol. 3, no. 9, pp. 1724–1735, 2007.

[110] T. Metsalu and J. Vilo, "ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap", *Nucleic Acids Research*, vol. 43, 2015.

[111] R. Kolde, "Pheatmap: Pretty heatmaps", *R package version*, vol. 61, 2012.

[112] S. Anders and W. Huber, "Differential expression analysis for sequence count data", *Genome biology*, vol. 11, no. 10, R106, 2010.

[113] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2", *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[114] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies", *Nucleic Acids Research*, vol. 43, no. 7, 2015.

[115] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments", *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–25, 2004.

[116] G. K. Smyth, M. Ritchie, N. Thorne, J. Wettenhall, W. Shi, and Y. Hu, "limma: Linear Models for Microarray and RNA-Seq Data User's Guide", [Online]. Available: https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf.

[117] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth, "Empirical bayes in the presence of exceptional cases, with application to microarray data", *Technical Report*, 2013.

[118] G. Smyth, "limma: Linear Models for Microarray Data", in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 2005, 2005, pp. 397–420. [Online]. Available: http://link.springer.com/chapter/10.1007/0-387-29362-0_23%5Cnhttp://dx.doi.org/10.1007/0-387-29362-0_23.

[119] S. B. Pounds, "Estimation and control of multiple testing error rates for microarray studies", *Briefings in bioinformatics*, pp. 25–36, 2006.

[120] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[121] L. Kumar and M. Futschik, "Mfuzz: A software package for soft clustering of microarray data", *Bioinformation*, vol. 2, no. 1, pp. 5–7, 2007.

[122] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* Boston, MA: Springer US, 1981. [Online]. Available: http://link.springer.com/10.1007/978-1-4757-0450-1.

[123] V. Schwammle and O. N. Jensen, "A simple and fast method to determine the parameters for fuzzy c-means cluster analysis", *Bioinformatics*, vol. 26, no. 22, pp. 2841–2848, Nov. 2010.

[124] M. Futschik, *Mfuzz tool homepage*, 2007. [Online]. Available: http://w3.ualg.pt/~mfutschik/software/R/Mfuzz/index.html.

[125] G. Chen, N. Han, G. Li, X. Li, G. Li, Z. Li, and Q. Li, "Time course analysis based on gene expression profile and identification of target molecules for colorectal cancer", *Cancer cell international*, vol. 16, no. 1, p. 22, 2016.

[126] J. J. Goeman and P. Bü Hlmann, "Analyzing gene expression data in terms of gene sets: methodological issues", vol. 23, no. 8, pp. 980–987, 2007.

[127] C. Clapham and J. Nicholson, *Fisher's exact test.* [Online]. Available: //www.oxfordreference.com/10.1093/acref/9780199235940.001.0001/acref-9780199235940-e-1143.

[128] *Pathway Commons, a web resource for biological pathway data. Fisher's Exact Test - Guide.* [Online]. Available: https://pathwaycommons.github.io/guide/primers/statistics/fishers_exact_test/.

[129] H. Chen and P. C. Boutros, "Venndiagram: A package for the generation of highly-customizable venn and euler diagrams in r", *BMC bioinformatics*, vol. 12, no. 1, p. 35, 2011.

[130] J. Larsson, *eulerr: Area-proportional euler diagrams*, R package version 1.0.0, 2016.

[131] H. He, Y. Dang, F. Dai, Z. Guo, J. Wu, X. She, Y. Pei, Y. Chen, W. Ling, C. Wu, *et al.*, "Post-translational modifications of three members of the human map1lc3 family and detection of a novel type of modification for map1lc3b", *Journal of Biological Chemistry*, vol. 278, no. 31, pp. 29 278–29 287, 2003.

[132] M. Zhao, P. Kim, R. Mitra, J. Zhao, and Z. Zhao, "TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes", *Nucleic Acids Research*, vol. 44, 2016.

[133] J. Zhao, J. J. Brault, A. Schild, P. Cao, M. Sandri, S. Schiaffino, S. H. Lecker, and A. L. Goldberg, "Foxo3 coordinately activates protein degradation by the autophagic/lysosomal and proteasomal pathways in atrophying muscle cells", *Cell metabolism*, vol. 6, no. 6, pp. 472–483, 2007.

[134] Y. Zhao, J. Yang, W. Liao, X. Liu, H. Zhang, S. Wang, D. Wang, J. Feng, L. Yu, and W.-G. Zhu, "Cytosolic foxo1 is essential for the induction of autophagy and tumour suppressor activity", *Nature cell biology*, vol. 12, no. 7, p. 665, 2010.

[135] D. Chen, F. Gao, B. Li, H. Wang, Y. Xu, C. Zhu, and G. Wang, "Parkin mono-ubiquitinates bcl-2 and regulates autophagy", *Journal of Biological Chemistry*, vol. 285, no. 49, pp. 38 214–38 223, 2010.

[136] M. Renna, C. F. Bento, A. Fleming, F. M. Menzies, F. H. Siddiqi, B. Ravikumar, C. Puri, M. Garcia-Arencibia, O. Sadiq, S. Corrochano, *et al.*, "Igf-1 receptor antagonism inhibits autophagy", *Human molecular genetics*, vol. 22, no. 22, pp. 4528–4544, 2013.

# Appendices

**Appendix A.** One dimensional PCA plots for year and sample source data, based on cancer samples only.

**Appendix B.** Raw counts of differentially expressed genes in the models with a single main effect (PAM50, stages, morphology).

**Appendix C.** Enrichment analysis results on differentially expressed genes data.

**Appendix D.** Enrichment analysis results on soft-clustering data for individual PAM50 subtypes with the cluster figures.

**Appendix E.** Venn Diagrams of Luminal B, HER2-enriched, Normal-like subtypes showing the overlap between downregulated autophagy genes in different cancer stages.