Cranfield University

Alexey Larionov

Classification of endocrine resistant breast cancers from transcriptomic datasets using multi-gene signatures

Cranfield Health
MSc Applied Bioinformatics

MSc Thesis
Academic Years 2011 to 2012

Supervisors: Prof David Cameron, Dr Sarah Morgan

September 2012

This thesis is submitted in partial fulfilment of the requirements
for the degree of Master of Science in Applied Bioinformatics

Abstract

Breast cancer is the most frequent cancer in women in developed countries. Endocrine treatment is indicated to the majority of breast cancer patients. However, in some cases it does not work despite the current clinical indications. Eventually the resistance may develop in many of those who initially respond. Re-analysis of available breast cancer transcriptomic datasets using new multi-gene signatures associated with endocrine resistance may help to understand and overcome endocrine resistance. The goal of this project was to develop a bioinformatics pipeline to (i) select endocrine resistant cases from the available breast cancer datasets and (ii) classify the selected cases by multiple multi-gene signatures.

The pipeline has been successfully designed and applied for classification of endocrine-resistant samples from 9 breast cancer datasets using 7 transcriptional signatures. The obtained results have been presented in a dedicated web site. The pipeline consists of:

- Procedures for a manually curated selection of relevant datasets and signatures;
- Procedures for semi-automatic data pre-processing, allowing cross-platform analysis;
- A new, fully automated, classification algorithm (Iterative Consensus PAM).

The main features of the developed classification algorithm include:

- It is based on un-supervised partitioning;
- It allows for "non-classifiable" samples;
- The procedure does not require a training set;
- The procedure can be used in a cross-platform context (Affymetrix & Illumina).

The developed pipeline and web site may constitute a prototype for a future web-hub collecting (i) data on endocrine-resistant breast cancer specimens, (ii) collecting multi-gene signatures relevant to endocrine resistance and (iii) providing tools to apply the signatures to the data. The web-repository could provide a tool to integrate the data and signatures and to produce new clinical and biological knowledge about endocrine resistance in breast cancer.

Contents

List of Figures

List of Tables

List of Abbreviations

AFFY – abbreviation indicating that a product/algorithm is related to Affymetrix arrays

ANN – artificial neural network

CEL –extension of a file type to store data from Affymetrix array

E2F3 - transcription factor E2F3

ER / ESR1 – oestrogen receptor / oestrogen receptor gene

FFPE – formalin fixed paraffin embedded (tissue sample)

FISH – fluorescence in situ hybridization

FNA – fine-needle aspiration biopsy

HER2 – human epidermal growth factor receptor 2

HTML – hyper text markup language

IHC – immuno histo chemistry

LDA – linear discriminant analysis

MYC – MYC gene

P53 – P53 gene

PCA – principal component analysis

PIK3CA – PIK3CA gene

RAS – RAS gene

CSS – cascading style sheets

SOM – self-organising maps

SRC – SRC gene

SVM – support vector machines

U133A / U133-Plus-2 / HT12 – names of microarray platforms

Acknowledgements

I thank my supervisors for trust and support.

I thank my wife and brother: without their help I would never be able to complete this bioinformatics course.

1 Introduction and Literature Review

The reanalysis of publicly available bioinformatics datasets may provide an important source of new knowledge. Modern biological methods produce vast amounts of data that can be analysed from different perspectives. Authors originally conducting a study usually focus their analysis on a specific question that can be addressed using bioinformatics resources available at the time. New bioinformatics tools may open new ways to re-analyse the same data. New datasets, collected within similar context, may allow comparison between the previously available and newly published studies. Further development of biology may generate new biological questions that can be answered using the old data.

While having a great potential, the comparison and re-analysis of already published datasets has its challenges. First of all, the re-analysis requires either an appearance of new questions that may be addressed using the old data or availability of new methods and datasets that may be used in re-analysis. Second, but equally important, re-analysing someone else's data requires good understanding of these data. This includes a range of questions starting from the general biological context (e.g. criteria for patient selection or response assessment) through to the technicalities of the lab methods employed (e.g. procedure for tumour biopsy collection or nucleic acid extraction). Finally, the complexity of multiple datasets and data analysis features requires special attention when presenting the results: ideally the results shall be presented in a concise and transparent way, clear for users with clinical or biological backgrounds.

This project re-analyses available transcriptomic datasets on endocrine-resistant breast cancers. These datasets come from studies focused on the development of prognostic or predictive signatures for endocrine treated patients. While deriving the signatures, authors considered the resistant (poor prognosis) patients as a single entity opposed to the responsive (good prognosis) patients. However, breast cancer is well known for its molecular diversity. Endocrine resistance may be caused by different mechanisms with distinctive molecular signatures. Therefore, it may be interesting to re-analyse these datasets focusing on the molecular diversity of endocrine resistance, instead of considering resistant tumours as a homogeneous group. The aim of this project is to

classify endocrine-resistant tumours from publicly available datasets using known multi-gene signatures for different mechanisms of endocrine resistance. This may allow us to suggest the mechanisms causing resistance in individual tumours and to see how different mechanisms of resistance are represented in different datasets.

1.1 Overview of Breast Cancer

1.1.1 Impact of breast cancer

Breast cancer affects millions of lives worldwide [1]. About 48,000 women are diagnosed with breast cancer and about 11,000 women are dying from breast cancer in the UK each year, making it the most common cancer in women [2].

Average cost of breast cancer treatment in developed world vary between GBP 7.000 - 35.000 per patient, depending on the country, stage and calculation method [3-5]. Given the incidence of breast cancer, even the modest estimate amounts to 243 million pounds per year spent in the UK for breast cancer care [3]. The total losses, including absence from work, production loss and early retirement may result to much higher numbers [6].

1.1.2 Causes of breast cancer

Breast cancer is caused by a combination of life-style, environmental, inherited and stochastic genetic factors, which differ in each individual patient. The main established risk factors for breast cancer are summarised in Table 1 [7-9] and discussed below.

1.1.2.1 Age

Age is a common risk factor for all major malignancies. The mechanism of this association is not clear. However, accumulation of DNA damage was associated with both ageing and carcinogenesis [10-12]. Taken together with age-associated decline of immune response [13], this may explain the higher incidences of cancer in elderly people. It may be noted that rates of the most cancers keep accelerating till the age of 70. In contrast, the breast cancer rate declines after 60 years. This may be explained by reduced oestrogen levels and by breast involution in post-menopause. Alternatively, one could consider the opposite: that endocrine disturbance associated with menopause may lead to earlier development of breast cancers. For instance, cessation of the cycle

may lead to a prolonged acyclic expression of oestrogen receptors (ERs) in normal breast ducts in contrast to their cyclic expression in the reproductive age [14]. Potentially this could make breast epithelium more susceptible to oestrogen-associated tumour promoting events despite the general fall of oestrogens during the menopause.

1.1.2.2 Reproductive factors and oestrogens

Risk of breast cancer is strongly associated with a number of reproductive and oestrogen-related factors including pregnancy, breast feeding, age of menarche and menopause, hormonal contraception and hormone-replacement therapy (Table 1). Breast feeding and pregnancy reduce risk of breast cancer through a complex and not yet understood endocrine effect on breast tissue rearrangements [15]. In contrast, the breast cancer risk associated with early menarche, late menopause and oestrogen-containing pills may be explained as a direct result of increased exposure of the breast to oestrogens.

Oestrogens play an important role in the development and function of normal breast. Specifically, they stimulate proliferation of breast epithelium [16]. Intriguingly, in

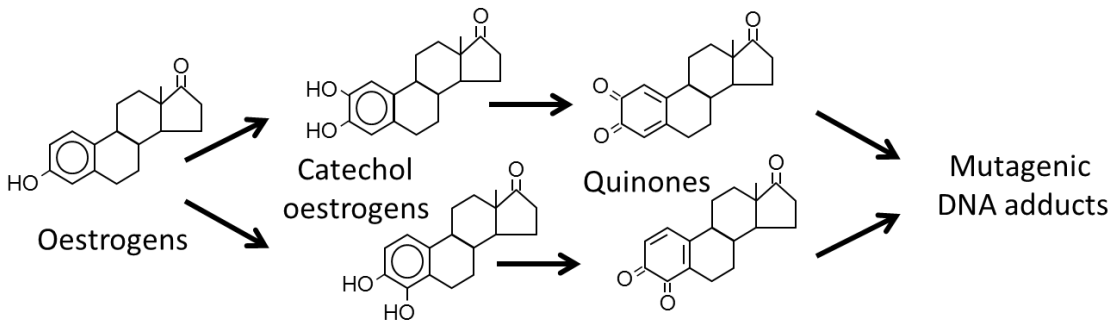
Table 1: Main risk factors for breast cancer

Factor	Relative risk
Elderly age	> 10
High breast density on mammogram	6
Atypical hyperplasia or cancer in other breast	> 4
High free estradiol in serum	3.6
Exposure to ionising radiation	3
First child after 40s	3
Menarche before age 11	3
Menopause after age 54	2
Breast cancer in a first degree relative	2
Obesity (post-menopausal)	2
High intake of saturated fat	1.5
Alcohol consumption	1.3
Hormone replacement therapy for >10 years	1.3
Current use of oral contraceptives	1.2
Obesity (pre-menopausal)	0.7

normal breast the cells carrying ERs do not proliferate themselves [16-18]. This led to a hypothesis that oestrogen-stimulated ER-positive epithelial cells induce proliferation in adjacent ER-negative epithelial neighbours [16-18]. Alternatively, one may suggest that normal ER-carrying cells lose ERs when enter proliferation after stimulation by oestrogens. The dissociation between ER-positivity and proliferation is lost during breast cancer development: about 75% of breast tumours preserve oestrogen receptors on the proliferating cancer cells [19]. This group of tumours is commonly referred as oestrogen receptor positive (ER+ve) breast cancer; they have a number of distinctive clinical features, such as better prognosis and high responsiveness to endocrine treatment. Noteworthy, the expression of estrogen receptors in ER+ve breast cancers is often higher than in normal breast epithelium [20-22]. In 2 to 20% of cases this may be explained by the receptor's gene (ESR1) amplification, depending on the method used for amplification detection [23,24]. However, the exact mechanisms regulating oestrogen receptor overexpression in the remaining majority of ER+ve breast cancers remain unknown yet [20].

In addition to the proliferative effect through oestrogen-receptor signalling, it has been suggested that oestrogens can contribute to breast carcinogenesis through a direct mutagenic effect by formation of DNA adducts. Oestrogens can be converted to catechol-oestrogens by p450-mediated hydroxylation in A-ring. In turn, the catechol-estrogens may be converted to quinones, which directly bind purines' residues in DNA, resulting in mutagenic DNA adducts (Figure 1, [25]).

Figure 1: Genotoxic effects of oestrogens



Theoretically, there is no obvious reason why this genotoxic effect of oestrogens shall be limited to breast tissue. Therefore, if the effect was strong, it might be expected that higher life exposure to oestrogens may be associated with higher risk of other, non-breast malignancies, which has not been reported (except for uterus, which is an endocrine-dependent tissue). At the same time, the risks of the life-long oestrogen exposure may be under-studied for methodical reasons. Measuring life-long exposure to oestrogens is not a trivial task: oestrogens fluctuate during the cycle in reproductive age and drop below the sensitivity of most commercially available tests in post-menopause [26]. In addition, the level of bio-available estrodiol depends on concentrations of sex-hormone-binding globulin [27]. Even with regard to the breast cancer, the methodical difficulties originally led to contradicting results whether the blood estrogen is related to the cancer risk [28]. Only measuring of free estradiol in large cohorts of patients allowed to detect the link of oestrogens in blood with risk of breast cancer [8].

1.1.2.3 Inheritance

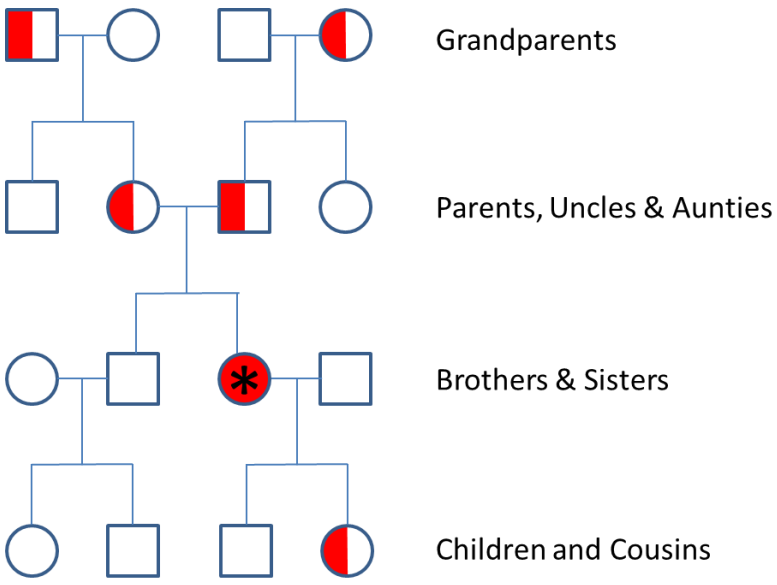
Familial cases constitute ~10% of all breast cancers [29,30]. However, familial history alone does not reveal the whole contribution of inheritance to breast cancer [31]. Criteria for familial cancer include a number of the affected 1st or 2nd degree relatives [32]. This is appropriate for detection of dominant high penetrance alleles. For instance: BRACA1/2 breast cancers have familial history in 27-66% cases, depending on the country [31]. At the same time, cancers caused by rare recessive alleles, low penetrance variants or complex multi-gene heritable traits will not affect close relatives; thus they will manifest as sporadic cases despite having the hereditable nature (

Figure 2 [31,33]).

Apart from familial cancers, the hereditable component may be prevalent in multiple and bilateral breast cancers, cancers in tweens and in early onset breast cases [30,33-35]. Estimates for the total contribution of inheritance into breast cancer incidence are still controversial. An analysis of a large tween dataset derived the hypothesis that “a high proportion, and perhaps the majority, of breast cancers arise in a susceptible minority of women” [36]. However, a later detailed analysis of the same data concluded that “the proportion of all breast cancer represented by heritable disease

exceeds 15%" [33]. Whichever estimate is correct, it is clear that even in women with established heritable predisposition, the genetic component alone is not sufficient to develop the cancer: only 20-30% of the identical tweens will have breast cancer if the sister is affected [36,37]. Having two first-degree relatives diagnosed with breast cancer increases the individual's risk by only 13.3% [38]. Carrying any known risk allele is neither necessary nor sufficient for breast cancer development (arguably, except for a rare combination of several high penetrance genes). Taken together these observations suggest that the inheritance shall be considered as a predisposing rather than a causal factor, and that an additional exposure to environmental factors and some additional somatic mutations are necessary to develop breast cancer even in women inherited the high risk genes.

Figure 2: Example of a homozygous recessive inherited condition without familial history



Note: The figure shows an example of typical family affected by homozygous recessive disease caused by a rare allele. Star [*] indicates the affected family member. Complete or partial red shading indicates homo- or heterozygosity for the risk allele. No first or second degree relatives are affected despite the heritable nature of the disease.

Most of the heritable breast cancer susceptibility genes fall into two major categories:

- genome maintenance / tumour suppressor genes or
- endocrine / steroid metabolism related genes.

Alterations in the genome maintenance genes are usually of high penetrance, which may lead to the familial history. The genes may be involved in DNA damage reception (ATM [39]), DNA repair (BRCA1/2, BLM, [40,41]) or response to DNA damage (e.g. halting cell cycle or triggering apoptosis: CHK2, P53, [42,43]).

The endocrine and steroid metabolism genes include genes related to oestrogen production and signalling, e.g.: CYP19 (estrogen-synthetase, [44]), COMT (catechol-estrogen inactivation, [45]) and ESR1 (estrogen receptor alpha, [46,47]). The effect of each steroid metabolism variant taken separately is usually small and limited to ER-positive tumours [48]. Despite the low penetrance of the individual variants, some tween studies suggest that multi-gene endocrine-related traits may constitute a major part in breast cancer heritable susceptibility [33].

There are epidemiological data, indicating that currently known high penetrance predisposition genes are responsible for only ~20% of all inheritable risk of breast cancer. The remaining 80% may be caused by a combined effect of multiple low-penetrance variants. Linkage studies based on family history cannot detect such genes. Genome-wide association studies (GWAS) on large cohorts of patients have been suggested to address this issue [49-52]. Interestingly, the GWAS may also be used to search for heritable protective traits, not only for the predisposing genes.

1.1.2.4 Other factors

Many of the remaining specific risk factors, mentioned in Table 1, may be considered as derivatives from the discussed above age, endocrine influences and inheritance. For instance, high breast density may be inherited and may be indicative of exposure to endocrine factors and pre-existing breast conditions [53]. Similarly, postmenopausal obesity is associated with increased oestrogen exposure through the peripheral synthesis of oestrogens in adipose tissue [54]. Interestingly, the pre-menopausal obesity may have an opposite effect [55].

Marked geographical and social differences in breast cancer rates have not yet been satisfactorily explained. However, at least partially, they may be related to low number of children and tendency for later first childbirth in developed world. Exposure to ionising radiation and other established carcinogens increases breast cancer incidences in a way similar to their effect on the other cancers.

1.1.3 Diversity of breast cancer

Several types of tumours can originate from the breast [56-59]. These types have distinctive clinical, pathological and molecular features summarised in Table 2 and described in more details below.

The major clinical sub-types of breast cancer are early breast cancer (including locally advanced) [60] and advanced breast cancer [61]. Most of the cases are diagnosed in the early stage, when cancer does not spread beyond the regional (axillary) lymph nodes. Early breast cancer is subdivided depending on lymph node involvement into lymph-node-positive (LN+ve) and lymph-node negative (LN-ve) disease, which have different clinical management and prognosis. A small proportion of breast cancers are diagnosed at the advanced stage, which is characterised by distant dissemination; treatment and prognosis of the advanced breast cancer depends on the degree of dissemination and locations of metastases. Most common locations of breast cancer metastases include bones (better prognosis) and viscera (liver, lung or brain), which have less favourable outcomes.

The pathological classifications most widely adopted in clinical practice include

- assessment of invasiveness (invasive vs in-situ cancer),
- histological grading by Bloom-Richardson [62] and
- histological typing of breast tumours developed by the World Health Organization [56].

Invasiveness is based on detection of cancer cells breaking through basal membrane. Invasive cancer requires more aggressive treatment, than non-invasive tumours. The Bloom-Richardson score is based on three components: disruption/preservation of breast ducts, nuclear morphology and mitotic index. The score is expressed numerically as 1 to 3: grade 1 having most favourable prognosis (preserved ductal structure, good

nuclear morphology and low mitotic index) and grade 3 having the poor outcome (no ductal architecture, disfigured nuclei and many mitoses). The WHO histological typing is based on integral morphological assessment. The most common type is invasive ductal carcinoma; the other types include non-invasive ductal carcinoma (DCIS), lobular, tubular, mucinous cancers and other rare histological types.

There are some correlations between histological types and molecular features of cancer [58]. However, development of targeted treatments requires more direct molecular markers informative for activity of specific pathways. The most useful molecular marker in breast cancer is oestrogen receptor (ER). It has been introduced in 1970th [19]. ERs are present in the majority, up to 75%, of breast cancers. Importantly, in many (but not in all) ER+ve cases signalling through the oestrogen receptor is required to maintain the tumour growth. Progesterone receptor (PgR) is used to evaluate functional status of oestrogen receptor signalling. Expression of PgR in breast is stimulated by oestrogens. Thus, presence of PgR on breast cancer cells indicates that

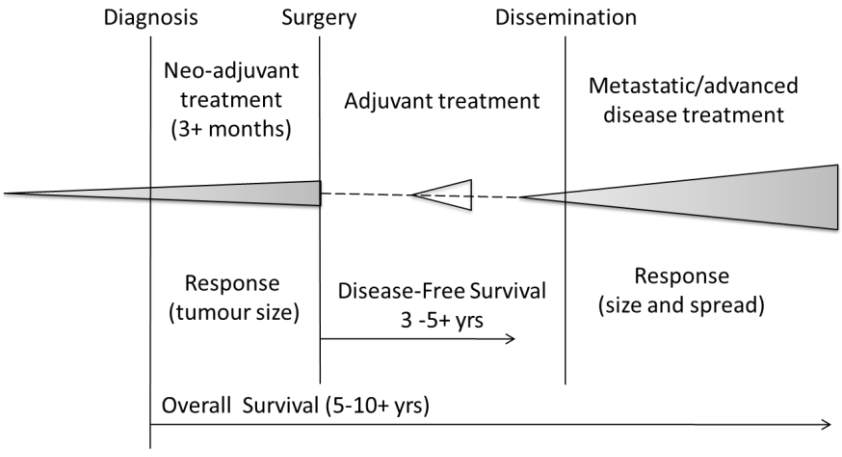
Table 2: Classifications of breast cancer

Clinical types	Early and Locally Advanced Breast Cancer Main subtypes according to lymph node involvement (LN+/-) Advanced Breast Cancer Main subtypes according to location of distant metastases
Pathological classifications	Invasiveness Main sub-types: Invasive, In Situ Histological type Main subtypes: Ductal, Lobular, Tubular etc Grade by Bloom Richardson Grades 1-3 based on disruption of glandular structure, nucleolar morphology and mitotic index
Molecular classifications	Traditional markers Main subtypes: ER+/-, PgR+/-, HER2+/-, Triple-negative Intrinsic subtypes Luminal A/B, Basal, HER2-Like, Normal-like

oestrogen signalling is active. In contrast, absence of PgR in ER+ve tumours suggests that oestrogen signalling may be dysfunctional despite the presence of oestrogen receptors. Majority of hormonal-receptor (ER and PgR) positive tumours respond well to endocrine treatment (such as tamoxifen or aromatase inhibitors); none of hormonal-receptor negative tumours respond to these drugs [63]. HER2 (Human Epidermal growth factor Receptor 2) is the last molecular marker that has been incorporated in standard clinical practice. It can be used to guide targeted treatments by Herceptin (trastusumab) or other drugs targeting this receptor. Ki67 is a proliferation marker that is currently being proposed for clinical use to complement ER, PgR and HER2 [64].

Most recently a number of multi-gene biomarkers have been suggested to further characterise molecular basis of breast cancers. One of the most developed molecular classifications identifies five major "intrinsic sub-types" with different clinical and pathological features: luminal A and B (correspond to ER+ve tumours), basal, HER2-Like and normal-like types (the latter three correspond to ER-ve breast cancers) [65,66]. While being considered an important milestone in breast cancer research, the intrinsic subtypes are yet of limited clinical utility. Just a few of the multi-gene signatures have been approved for clinical use, such as Oncotype Dx and Mammaprint [67,68]. At the same time, many studies are being carried out to bring translational multi-gene signatures into clinical practice.

Figure 3: Clinical history of breast cancer



Note: Modified from A.Larionov & W.Miller (2009) with author's permission [69]

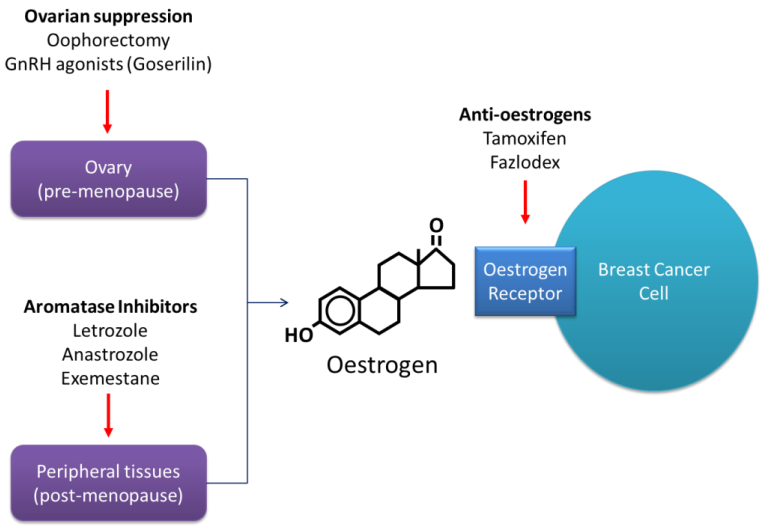
Importantly, the cancer's clinical, pathological and molecular features are not static. They change in time along with the cancer treatment and progression. Figure 3 illustrates the major clinical events and treatment regimens in clinical history/evolution of breast cancer. Different pathways are involved in tumour progression at each step; different treatment regimens and different response/progression criteria are applied at different stages of tumour progression [69].

1.2 Endocrine Treatment and Resistance in Breast Cancer

1.2.1 Endocrine treatment

Oestrogen receptors are present in ~75% of breast cancers [19]. Growth of these ER+ve tumours usually depends on oestrogen signalling. Endocrine treatment disrupts or prevents this oestrogenic stimulation. The first example of successful endocrine treatment, oophorectomy, has been reported by Beatson in 1896, decades before the discovery of oestrogens or oestrogen receptors [70]. Remarkably, oophorectomy is successfully used to treat pre-menopausal breast cancer patients till now. In addition, several other modalities of endocrine treatment have been developed during the last century; the major modalities of endocrine treatment are illustrated on Figure 4.

Figure 4: Endocrine treatment in breast cancer



Note: Reproduced from A.Larionov & W.Miller (2010) with author's permission [71]

Ovaries are main source of oestrogens in pre-menopause. Ovarian production of oestrogens can be ceased either by surgical removal of ovaries or pharmacologically (by GnRH agonists [72]). Ovarian irradiation is not recommended nowadays because it is less reliable and may be associated with adverse side-effects [73,74]. After the menopause, ovaries stop producing oestrogens and their blood level dramatically falls. However, even the residual low level of oestrogens still is sufficient to support grows of ER+ve breast cancers. In post-menopause, the primary site of oestrogen production moves to peripheral tissues, first of all – to adipose tissue [75]. Adipose tissue expresses very low levels of aromatase (the key enzyme of oestrogen biosynthesis). However, because of the bulk of the tissue in the body it can produce sufficient amount of oestrogens to stimulate growth of breast cancer. Aromatase inhibitors (AIs) are used to block the peripheral oestrogen production in post-menopause.

Instead of preventing oestrogen production, the alternative approach is to block oestrogen signalling through oestrogen receptors. For instance, Tamoxifen, the first successful targeted treatment in oncology, inhibits breast cancer growth by competing with oestrogens for binding to oestrogen receptors [76].

1.2.2 Endocrine resistance

Despite the success of endocrine treatment in general, its effectiveness vary in individual patients. About 30% of ER+ve cases do not respond to endocrine treatment despite the presence of oestrogen receptors (primary endocrine resistance). Many of those who initially respond develop the resistance later (acquired resistance) [77,78]. Clinical management of endocrine resistant cases usually includes an attempt to administer another modality of endocrine treatment and/or add cytotoxic and other targeted agents [71]. In fact this tactics is rather *ex-juvantibus* trial-and-error approach than a rational attempt to overcome the resistance basing on a knowledge of the molecular mechanisms underlying growth of the resistant tumour.

Multiple causes for endocrine resistance have been suggested. Causes residing outside of the tumour may include inaccuracy in ER assessment [79], poor adherence to treatment [80] and adverse drug metabolism [81]. Many specific molecular mechanisms acting within the tumour cell have also been suggested, which will be

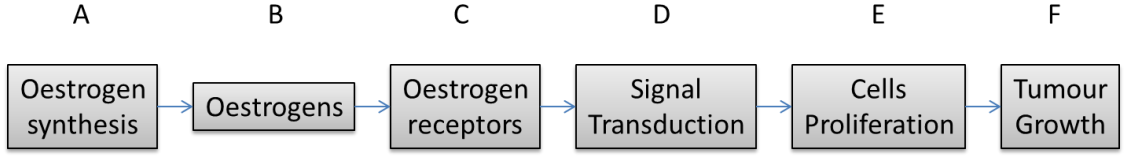
discussed later [77,82,83]. Figure 5 illustrates the main steps in oestrogen-stimulated tumour growth. Resistance to treatment can develop at each step, for instance:

A) Inhibition of oestrogen biosynthesis by AIs may be inefficient because of inherited polymorphisms in aromatase gene [84,85];
B) Effective inhibition of oestrogen biosynthesis may be compromised by exogenous oestrogenic compounds (e.g. dietary phytoestrogens or oestrogenic industrial phenolic pollutants) [86,87];
C) Aberrations and ligand-independent activation of oestrogen receptors may influence response to endocrine treatment [88,89]
D) Cross-talk with growth factors may enhance ER-signalling and ER-driven proliferation [90];
E) ER-driven proliferation may co-exist with ER-independent proliferation mechanisms in ER+ve breast cancers [83];
F) Apart of the proliferation, tumour growth depends on apoptosis, vascularisation and other processes, which may contribute to endocrine resistance and response [91,92].

1.2.3 Intratumoral molecular mechanisms of endocrine resistance

The last four steps on the above figure refer to intratumoral mechanisms of endocrine resistance. Because this project deals with molecular profiles of tumour biopsies, the intratumoral mechanisms of endocrine resistance require a detailed attention. Examples of the major intratumoral molecular events that may cause endocrine resistance are highlighted in Figure 6 and discussed below.

Figure 5: Steps in oestrogen-stimulated tumour growth

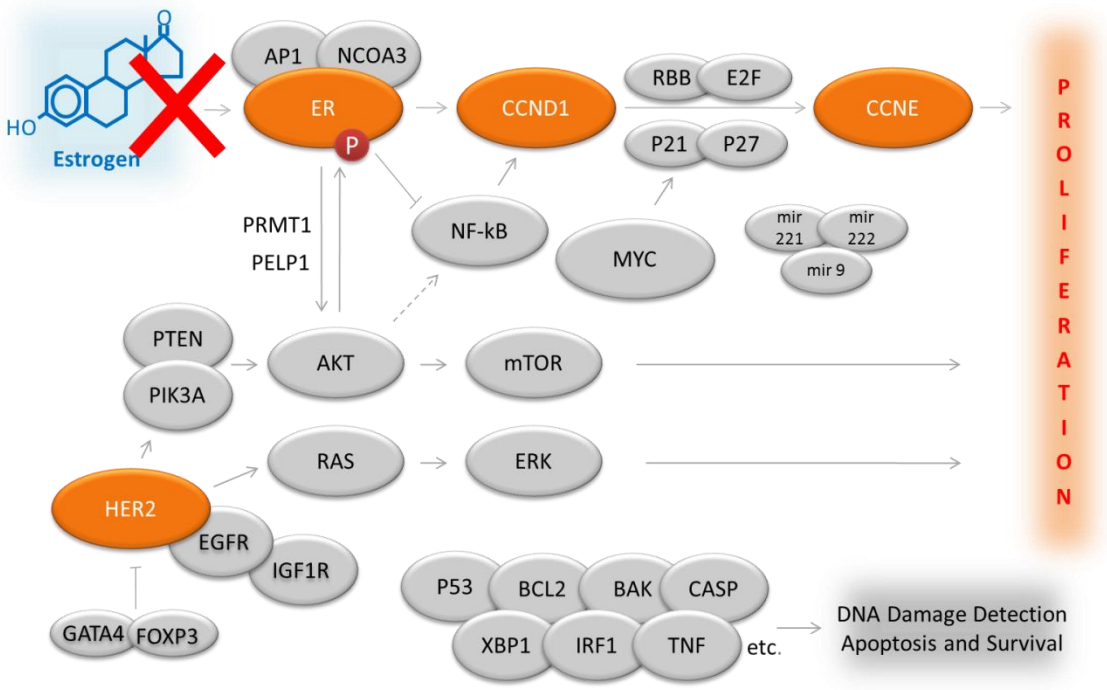


Note: Modified from W.Miller & A.Laronov (2012) with author's permission [77]

1.2.3.1 Ligand-independent activation of ER and hyper-sensitivity to low concentrations of oestrogens

Upon binding to oestrogens ERs undergo dimerization and nuclear translocation. Within the nucleus ER act as a nuclear factors binding to oestrogen-regulated elements (ERE) and changing expression of the oestrogen-regulated genes. Binding to EREs requires co-regulators (AP1, NCOA1-4 and others). It has been suggested that overexpression of these co-regulators may lead to hyper-sensitivity to low concentrations of oestrogens or even may cause oestrogen-independent activation of the ER- signalling [93]. Alternatively, oestrogen-independent activation of ERs may occur because of phosphorylation of ERs, caused by growth-factors dependent intra-cellular kinases as a part of crosstalk between growth factors and oestrogen signalling [90,94].

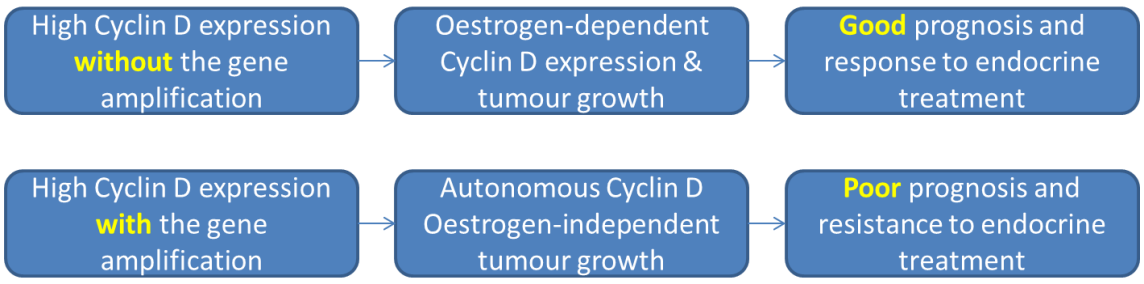
Figure 6: Selected molecular mechanisms of endocrine resistance



1.2.3.2 Cyclins and other cell cycle regulators (CCND1, CCNE)

A number of cell cycle regulators acting downstream of ERs have been linked to endocrine resistance [95]. Cyclin D (CCND1) is a commonly known oncogene often amplified in breast cancer and other cancers [96]. Cyclin D gene (CCND1) amplification is present in ~20% of breast cancer cases; the overexpression of Cyclin D protein is observed in about a half of breast tumours [97,98]. Cyclin D plays an important role in G0-G1-S transition during the cell cycle. Therefore its increased activity may be directly associated with high proliferation and poor outcome. However, attempts to verify this hypothesis in breast cancer were inconclusive. While there are observations supporting this model [99], there are also observations apparently contradicting to it: when overexpression of Cyclin D protein was associated with ER-positivity and good prognosis [97]. This controversy may be explained by the fact that CCND1 is a known ER target [100]. Therefore, high expression of Cyclin D1 protein in ER+ve cancers in absence of the gene amplification may be indicative of the estrogen-dependent growth, which is likely to respond to endocrine treatment and have a good prognosis. In contrast, the autonomous (ER-independent) Cyclin D1 activity may cause endocrine resistance [101]. The latter case is likely to be observed if high Cyclin D protein expression follows CCND1 gene amplification. Indeed, amplification of CCND1 gene is linked with poor prognosis and poor results on endocrine treatment [102,103]. Therefore, CCND1 provides a good example illustrating how proteomic, transcriptomic and genomic data shall be analysed together to decipher molecular mechanisms of endocrine resistance (Figure 7). To observe the whole picture one shall also take into account several other genes, which may be co-amplifyed with CCND1 [104,105].

Figure 7: Proposed interpretation of Cyclin D expression and amplification in breast cancer



Endocrine resistance can also be associated with other cell cycle regulators, acting downstream of Cyclin D. For instance it has been shown that overexpression of Cyclin E (CCNE1) or it's truncation by a specific protolithic cleavage can cause endocrine resistance by bypassing cell cycle arrest induced by endocrine treatment [106-109].

1.2.3.3 Cross-talk between ER and growth factors signalling (HER2)

A number of molecular pathways commonly associated with carcinogenesis may interact (cross-talk) with oestrogen receptor signalling. HER2 (Human Epidermal growth factor Receptor 2) pathway is the most studied example of such interaction because HER2 is amplified in a noticeable fraction of ER+ve tumours and there are drugs targeting HER2 signalling [110,111]. HER2 amplification in ER+ve tumours is associated with poorer results on endocrine treatment [110,112]. This can be explained either (i) by a direct effect of HER2 on proliferation (through PI3K-AKT-mTOR or RAS-ERK/MAPK cascades) or (ii) by ER-HER2 interaction [90]. The interaction is bidirectional. On one hand, AKT can activate ER by phosphorylation; on the other hand, the rapid effects of oestrogens mediated by plasma membrane ERs can cause EGFR - AKT cascade activation [83].

1.2.3.4 Other mechanisms

A number of other molecular mechanisms have also been implicated in endocrine resistance. These mechanisms involve nuclear factors (e.g. NFkB, MYC [113-115]), micro-RNAs (e.g. mir9, 221, 222 [116,117]) and molecular determinants of apoptosis (e.g. P53, BCL2, CASP8 [118,119]). Importantly, the endocrine resistance mechanisms closely interact with each other: micro-RNAs being in control of ER or cell cycle regulators, apoptosis being regulated by oestrogen signalling, etc. The combination of mechanisms may differ in each individual tumour.

1.3 Transcriptional signatures in endocrine resistance

Despite the bulk of experimental and observational data on molecular mechanisms of endocrine resistance, there are yet no clinically useful biomarkers to predict endocrine resistance in ER+ve patients and there is no rational approach to overcome the resistance. One of the strategies to address this shortage is to seek multi-gene

transcriptional signatures associated with specific mechanisms of resistance and with clinical outcomes.

1.3.1 Sample collection

The signatures can be derived from high-throughput transcriptomic studies carried out on either clinical samples or on experimental models. The experimental models use cell cultures and laboratory animals. The cell lines studied are usually ER+ve cell lines (e.g. MCF-7, T47D, BT474 or ZR-75, [120,121]) incubated with tamoxifen or long-term oestrogen deprived (to model resistance to aromatase inhibitors, [122]). Alternatively, cell lines may be transfected with genetic constructs to monitor oestrogenic signalling or to modify cell growth or production of oestrogens [123,124]. Experimental animals may be used as hosts for xenografts [125]. Alternatively these may be animals with induced breast carcinomas or genetically modified animals, e.g. mice with conditional knockout or overexpression of aromatase [126]. The main advantage of experimental models is that they allow functional interventions to study causal relations at the molecular level. The main disadvantage is that experimental findings may be of low relevance to clinical tumours and treatments. For instance, the experimental models poorly reflect clinical treatment dosages and settings, specifically – the biology of most common adjuvant setting, when treatment is directed at micro-metastatic and dormant disease.

The collection of tumour samples often accompanies breast cancer clinical trials or treatment audits [127]. Findings based on these clinical tumour biopsies may be directly translated to the clinic. However, a series collection may take years, a biopsy size is limited and no experimental interventions are possible to study the causal relations in molecular findings. When comparing transcriptomic datasets obtained in different clinical studies it is important to pay attention to the technical details, including studied populations, treatment settings and dosages, criteria for response assessment, biopsy techniques and microarray platforms. The main features characterising the studied population, treatment and response assessment have been discussed above (Table 2, Figure 3). In addition, it may be important to evaluate age, ethnicity and reproductive status of patients.

The biopsy techniques used in breast cancer transcriptomic studies include fine needle aspirates (FNA), core biopsies and excision biopsies. The tissue may be preserved by freezing in liquid nitrogen or by fixation in formalin and paraffin-embedding (FFPE blocks). Because FNA samples provide an extremely small amount of material, they may often be non-informative and/or be poorly representative for the intratumoral heterogeneity. Core biopsy is a common procedure in the breast cancer clinic; usually it is well tolerated and can be taken sequentially. Core biopsies provide sufficient material for modern transcriptomic methods (up to 25-100mg of tissue). However, it may not be enough for a repeated analysis, if the initial attempt has failed. Excision biopsies or tumour samples obtained at surgery usually are large (up to 1 gram and more); usually excision biopsies cannot be collected sequentially, e.g. before and after certain treatment.

Until recently, most transcriptional studies were conducted on frozen samples, as RNA is severely degraded in FFPE blocks. Recent progress in molecular techniques has allowed PCR analysis on FFPE blocks; however, fresh frozen samples are still preferable for the high-throughput microarray techniques. FFPE samples may be stored in archives for decades. Thus, when analysing transcriptional data obtained on FFPE blocks it is important to be aware of the age of blocks and of the storage conditions.

1.3.2 Microarray platforms

Several microarray platforms have been used in transcriptomics studies in breast cancer. Early seminal studies were conducted more than a decade ago using in-house spotted microarrays [65,66]. The experiments included several steps. First RNA was extracted from studied samples, labelled (e.g. by fluorescent labels) and hybridised to the arrays. Then the arrays are scanned: if a specific mRNA was present in the sample, then the corresponding spot showed fluorescence. This general experimental workflow is still used in present-day microarray studies. However, the array manufacturing, sample preparation and labelling have been significantly improved. Nowadays the in-house spotted arrays would be considered sub-standard because of relatively low number of spots and low accuracy of the in-house spotting. Contemporary studies use commercial arrays, with Affymetrix and Illumina being the leading microarray manufacturers.

1.3.2.1 Affymetrix arrays

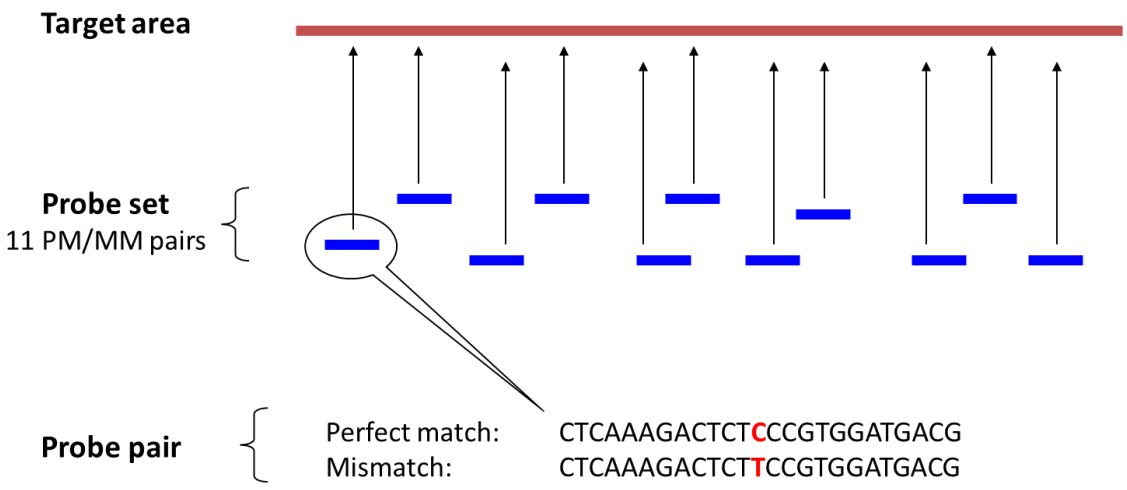
Instead of spotting the probes to arrays Affymetrix synthesises the probes in-situ. A combination of photolithography and oligonucleotide chemistry allows manufacturing of very high density arrays (10-20 microns per "spot") with precise location of each "spot" [128]. Because the in-situ synthesis becomes less accurate for longer oligonucleotides, Affymetrix arrays use multiple short (~25 nucleotides) oligos to overlap within the larger target area (Figure 8). Importantly, each oligo is designed in two versions: perfect match and single mismatch. To ensure specificity of the results the original Affymetrix algorithms for data analysis (MAS4 and MAS5) recommend comparison of signals obtained from the perfect match and mismatch probes.

1.3.2.2 Illumina arrays

The Illumina technology is a "BeadArray". In contrast to Affymetrix, Illumina does not synthesise the probes in-situ; neither Illumina spots probes on the array. Instead Illumina attaches probes to small beads (~ 2-3 microns in diameter, ~800k of oligo copies per bead). This design allows the use of long probes (~50nucleotides). The beads are spread over the array surface, which has special wells for regular accommodation of the beads. The beads allocation is random. However, it is decoded

---

Figure 8: Design of Affymetrix array probe sets



after the array manufacturing [129]. The decoded beads map is supplied in a DMAP file that is unique for each the chip. In addition the decoding procedure provides an individual quality control for each chip manufacturing. Illumina technology allows placing many tens of thousands of beads per array, which is higher density than achieved by Affymetrix. Such high number of features allows for redundancy: allocating several identical bead types per array (~15 on average) increases the reliability of measurements.

1.3.2.3 Comparison of data from different micro-array platforms

Apart of Affymetrix and Illumina the commercial microarray manufacturers include Agilent (spotted arrays), Nimblegen (Roch, in-situ synthesised arrays) and others. While the results obtained by different microarray platforms usually are similar [130,131], the direct comparison or integration of microarray data obtained on different platforms requires special precautions (see section on batch-correction and cross-platform integration below [132]).

1.4 Bioinformatics pipeline in transcriptomics

The raw data produced in microarray experiments include images generated by the array scanners, experimental annotations and meta-data. Interpretation of the raw data relies on complex bioinformatics procedures, which include a large number of relatively independent steps. Multiple legitimate options are available for each step of analysis. These options need to be tuned to specific dataset and study design. The robust result shall be confirmed using different alternative options applied to the same dataset. This section will describe the common bioinformatic tasks performed during gene expression microarray data analysis.

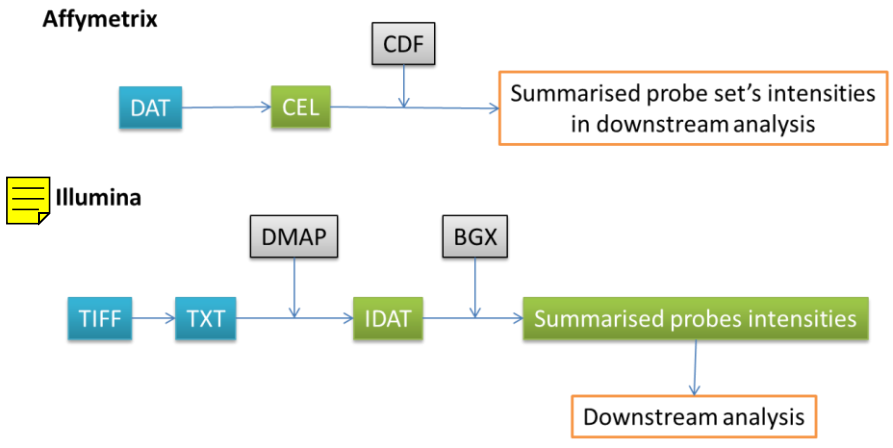
1.4.1 Microarray scanning and source data file types

Prior the further analysis, the scanned microarray images have to be converted to numerical values representing the intensity of spots [133,134]. Affymetrix scanners save data in a proprietary file format (DAT files), which can be rendered as image using specialised software including own Affymetrix tools and some R- or Matlab packages. A single probe spot on Affymetrix raw images is ~ 10x10 pixels, only the central 8x8 being used for intensity measurement. The DAT files are converted to CEL files for

further analysis. CEL files summarise information for each spot on the image. CEL files may also be rendered as pseudo-images. However, the pseudo-images generated from CELs represent each probe as a single pixel, instead of the true images available only from DAT files. Conversion of CELs to probesets' expression values requires information about chip design, provided in CDF (Chip Description File) files [135].

Illumina scanners also produce several file types. The original images are saved in TIFF files, containing ~17x17 pixels window around each bead. Only the central 9 pixels are used for intensity measurement, the peripheral pixels are used as local background. The corrected spots intensity data are stored in TXT files, which are translated to the bead type intensities (IDAT files) using mapping, available in DMAP files. Finally the bead intensity data are translated to probes intensities, using the manifest files, available from Illumina (e.g. BGX files, [136]). It may be noted that folders with "raw" data generated by Illumina GenomeStudio software may contain different files sets, depending on the user-customised settings. The folder often may contain JPEG images for each array. However, the size of JPEG files is quite small, suggesting that they are just thumbnails based on processed data, like the pseudo-images generated from Affymetrix CEL-files.

Figure 9: Affymetrix and Illumina source data files



Notes: Blue boxes show raw image data files, green boxes show processed image data files, grey boxes show files with additional information about chip design, provided by the manufacturer. Because of constant technology development, some figure details (e.g. file extensions) may be different for different Affymetrix and Illumina products.

The summary of Affymetrix and Illumina microarray dataflow from scanning to the typical files used in the downstream analyses is illustrated on Figure 9.

Usually the summarised intensity values are obtained using the manufacturer's proprietary software supplied with the scanner. In contrast, the downstream analysis of the summarised intensities is often performed using appropriate R-packages, which provide greater flexibility and transparency. For instance, Affymetrix CEL files can be read by Affy R-package [137]; text files exported by Illumina's GenomeStudio can be read by Lumi R-package [138]. Currently Illumina encrypts IDAT files to encourage generating of summarised probe intensities by GenomeStudio. However, there are R-packages that can read the encrypted IDAT files (for instance IDATreader). Alternatively there are R-packages able to read the true image-level data; for instance, Beadarray R-package can import TIFF/TXT Illumina files [139].

1.4.2 Microarray data repositories and reporting standards

As well as our own datasets, this project re-analyses several publicly available datasets. It is a common academic practice to share the raw data of microarray experiments. This is required for by most of the journals publishing results of such studies. There are several publicly available and publicly maintained repositories, which are used for the microarray data sharing. The two most popular repositories are Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/ [140]) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). In addition to storing the data, these repositories provide convenient interface for datasets searching and some basic analyses. The repositories exchange the information between each other: thus a dataset submitted to GEO will be soon available through the ArrayExpress too. The repositories accept only data satisfying to the Minimum Information About a Microarray Experiment (MIAME) requirements [141]. These include not only the data themselves, but also information on the main factors influencing interpretation of the data: design of the experiment, description of samples, array design, hybridisation and normalisation procedures.

1.4.3 Pre-processing

A typical Affymetrix or Illumina microarray experiment includes several samples, each of them hybridised on a separate chips. To compare measurements made on the

different chips they need to be pre-processed. Pre-processing includes background correction, normalisation, summarising, batch-correction and filtering.

Background correction at this step accounts for global (RMA) or local (Loess) biases in expression values. This does not substitute the background correction performed during the image analysis, which accounts for peripheral pixels in the spots.

Affymetrix probesets consist of multiple probes (Figure 8), which intensities have to be summarised to produce the overall probeset expression value. Additionally, Affymetrix provides the perfect match/ mismatch probes, which may also be integrated at the summarisation step (an example of a platform-specific pre-processing). Illumina arrays carry multiple beads carrying the same probe. Their intensities also shall be summarised to produce a single expression value for each bead type.

The basic assumption underlying the normalisation step is that the average expression over all genes shall be similar on each array. In practice, the normalisation procedures are based on more advanced assumptions, e.g. that the distribution of genes expressions shall be similar between arrays (quantile normalisation) and may include some empirically justified corrections (MAS algorithms).

Multiple R-packages can be used to perform the array pre-processing. The most popular R-packages for Affymetrix data are Limma and Affy [137,142]. Beadarray and Lumi R-packages can be used for Illumina arrays pre-processing [138,139]. In practice the background correction, summation and normalisation are often performed simultaneously, using integrated functions available in the chosen R package [143]. Thus, *mas5()* function from Affy package can be used to perform all pre-processing steps according to Affymetrix's MAS5 algorithm. Beside to the MAS5 algorithm, *expresso()* function from the same package may also perform Robust Multichip Average (RMA) background correction, Loess or Quantile normalisation and Medianpolish or Liwong summation algorithms [144]. Similar options are available in other R packages for Affymetrix and Illumina pre-processing.

The filtration step in pre-processing is used to remove non-informative probes, for instance: probes that are not expressed at all or do not change noticeably between the

arrays. Another example of filtration is selecting 500 most variable probes in the array (this is based on assumption that the most variable probes are the most informative).

Finally, the pre-processing may include batch-correction. Large transcriptomic studies associated with breast cancer clinical trials may collect samples over several years and process them in batches. It has been shown that even after baseline correction, summation and normalisation the data may still keep strong batch-specific bias [145]. There are several methods to minimise the batch effect. The simplest method uses median-centring [146]. Importantly, it is applied in a different dimension than in the normalisation: the assumption is that average expression of each gene shall be similar in each batch. Like in the normalisation, this simple principle may be developed into more sophisticated algorithms, including empirical Bayesian calculations (ComBat correction [147]). It shall be noted that any batch-correction method is removing differences between the batches. Thus, to avoid removal of legitimate meaningful differences the composition of batches shall be balanced. For instance, developing a signature for endocrine resistance in breast cancer, each batch shall include approximately similar proportion of resistant and responsive tumours.

An important specific case is when different batches are studied using different microarray platforms. In addition to a specialised batch-correction procedures (e.g. Cross-Platform-Normalisation, XPN [132]) the inter-platform integration requires probes matching between the platforms, which may not be a trivial procedure [148,149].

1.4.4 Tumour classification using multi-gene signatures

This project is focused on the application of multi-gene signatures for classifications of endocrine-resistant breast tumours. Development and application of multi-gene classifiers involve several typical steps illustrated on Figure 10 [150]. Initial un-supervised exploratory analysis is needed to acquire familiarity with the data. It may also include additional quality control checks. The intrinsic sub-classes may be related to known clinical and pathological parameters. The next common step is to derive lists of features differentially expressed between the studied groups (e.g. responders and non-responders to treatment). Finally, these features are used to construct a classification algorithm, which can be used to predict the class of the newly collected tumours.

1.4.4.1 Exploratory analysis

Exploratory analysis includes descriptive statistics, quality controls and un-supervised class discovery procedures, like Hierarchical Clustering Analysis (HCA) or Principal Component Analysis (PCA).

Descriptive statistics provide important information for quality control purposes. For instance, percentage of "detected" genes on array or average intensity of top and bottom 5% of genes can be used as quality control metrics. Labelling procedures utilised on older arrays often were sensitive to RNA degradation, which could be monitored by special control probes located at 3' and 5' regions of the genes.

Clustering is a group of methods that allocate similar cases close to each other. The degree of similarity may be calculated using different distance measures, the actual allocation of similar cases into clusters can be done using different agglomeration/linkage algorithms. Selection of the distance measure and linkage algorithm may drastically influence the clustering result. Most common distance measures include Euclidian distance, Manhattan distance or Correlation coefficient between samples. Influence of distance measures on HCA is illustrated on Figure 11A. Examples of the linkage algorithms include Complete, Average or Single linkages, as illustrated on Figure 11B. Clustering in transcriptomics is usually coupled with heatmap figures that show genes expressions in studies cases. One of the main advantages of HCA is that combining bi-clustering of genes and cases with the heatmap allows quick visual assessment of what genes are up- or down- regulated in different

---

Figure 10: Development of a multi-gene classifier

tumour groups. The disadvantage is that clustering is sensitive to the noise originating from low-informative variables (unless low weights are assigned to such variables during the clustering).

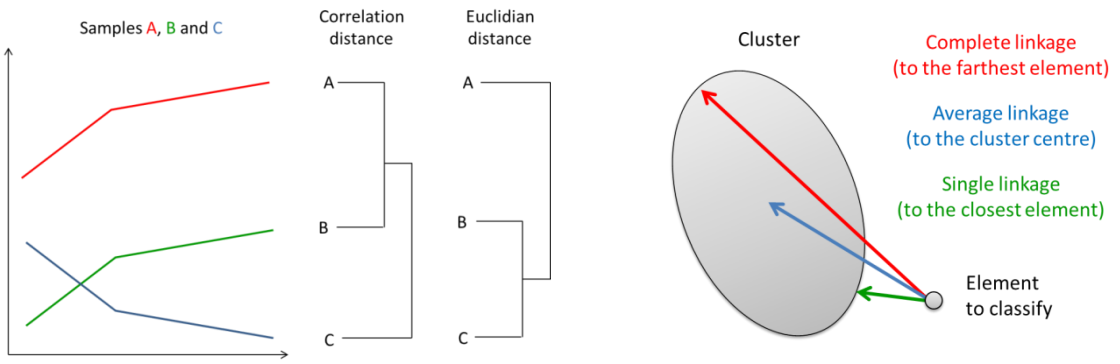
Principal component analysis is an alternative to HCA class discovery technique. Its main advantage over HCA is that PCA effectively deals with the redundant or low-informative genes reducing the multidimensional space of all initial variables to a smaller number of highly-informative principal components (PCs). Plotting cases within the space of 2 or 3 most informative principal components allows visualising sub-groups within the studied dataset. The disadvantage of PCA is that it hides biological identities of the genes, contributing to the groups' separation.

1.4.4.2 Informative features selection

Exploratory analysis is useful for acquainting with the data and for familiarising with the data inner structure. However, it is not directly informative for derivation of the multi-gene signatures, which can be used for tumours classification.

In the present study we will use signatures associated with different mechanisms of endocrine resistance, for instance: transcriptional signatures associated with P53 mutations, PTEN loss or HER2 amplification. These signatures include genes, which expressed differently between the tumours with and without the studied feature (e.g. with and without HER2 amplification).

Figure 11: Effects of Distance Measure and Linkage Algorithms on Clustering



A: Distance measures B: Linkage algorithms

The simplest methods for selection of differentially expressed genes are based on classical statistics: (i) first expression of each gene is compared in the studied groups (using for instance t-test or Wilcoxon-Mann test); (ii) then the genes are ranked according to the p-value for the difference; (iii) finally a certain number of the top differentially expressed genes are taken further to design a classifier.

Having sufficient number of observations, application of the classical statistical methods produces proper ranking of the informative genes. However, the actual p-values may be misleading because the classical tests have been developed for single experiments. The microarrays measure many thousands of genes at a time. Applying $p<0.05$ criteria to such number of measurements will produce tens or hundreds of "significantly" changed genes merely by chance. A number of multiple testing corrections have been suggested to address this problem. The simplest method is the Bonferroni correction, which merely multiplies the classical p-value by the number of tested genes. This is a very strict correction, which may exclude many significantly changed genes for not to include any false-discovered ones. An alternative approach is to explicitly allow some specific false-discovery rate (FDR, e.g. 20%) for the sake of keeping all genuinely changed genes for downstream analysis.

Apart of the misleading p-values, the direct application of classical statistics to microarray data may have some other limitations. To overcome these limitations, a large number of specialised and highly sophisticated methods have been suggested for selection of differentially expressed genes in microarray experiments [151-156]. The specialisation comes at a price of transparency. Different methods produce different lists of genes. Even repetition of the same method may produce different results because of randomisation incorporated in some methods. Theoretically, it is legitimate to have multiple equally informative multi-gene signatures [152]. However, high complexity and lack of transparency may lead to sub-optimal tuning of the sophisticated methods. It was observed that some differentially expressed genes derived by a specialised procedure may be of low median fold change and of low consistency of changes (Figure 12 [157,158]). Thus, to ensure the quality of gene lists produced by highly specialised methods, it is recommended to explore genes using conventional descriptive statistics, prior taking them to the downstream analyses.

1.4.4.3 Classification algorithms

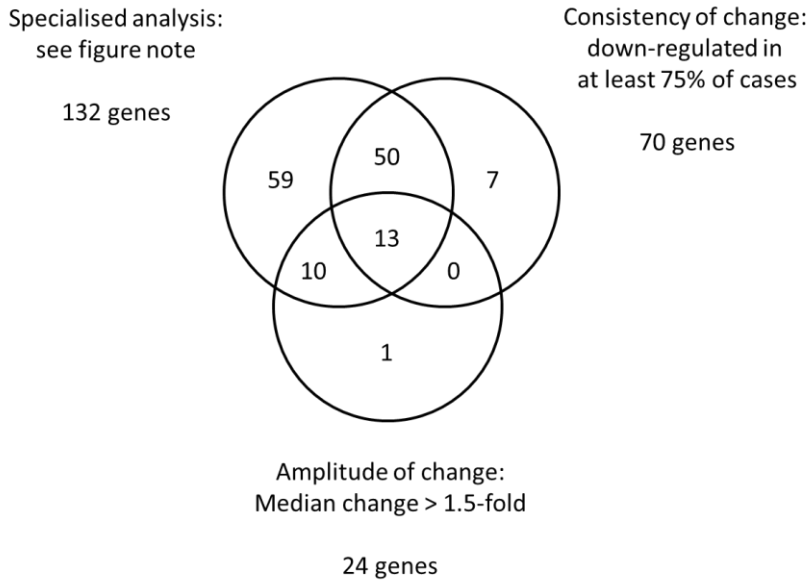
To classify tumours using informative features selected at the previous step the features shall be interpreted by a classification algorithm. The same features can be used in different classification algorithms [159]. Prior to classification of the new cases, the algorithms shall be trained on the dataset with known allocation of cases.

Some of the algorithms are building on the methods described earlier for exploratory analysis. Thus, Figure 13 illustrates principles of Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classification algorithms, which build on the PCA analysis.

LDA evaluates position of the training cases within the space of the most informative latent variables (principal components) and draws a linear border, which best separates

---

Figure 12: Comparison of a specialised analysis with consistency and amplitude of change



Notes: Modified from Miller W, Larionov A. et al 2010 with author's permission [158]. Figure shows genes down-regulated on treatment. The specialised analysis was based on statistical significance of changes assessed by linear modelling in paired samples with empirical Bayesian adjustment for multiple testing.

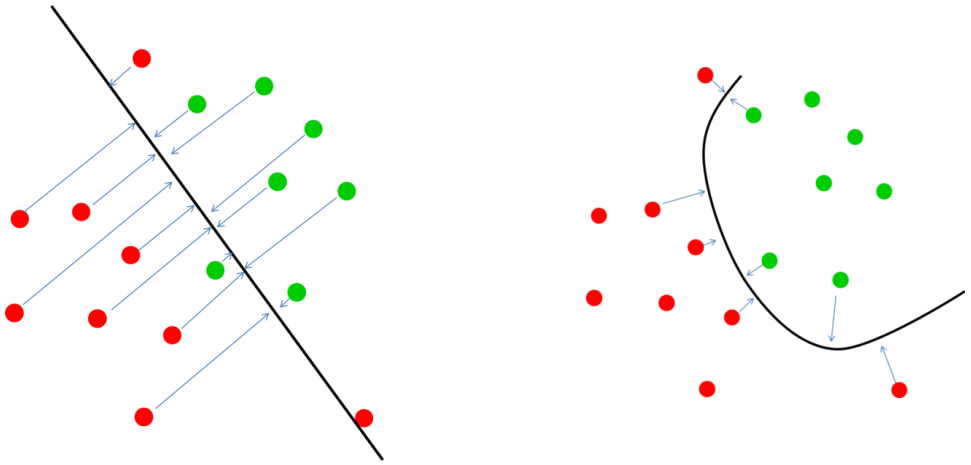
the groups. Classification of the new case is decided depending on which side of the border it lays. In practice, the border may not be linear. It also may be that cases closest to the border are more important for the exact demarcation of the borderline. These considerations are taken into account by SVM algorithm, which draws non-linear borders basing only on selected cases from the training set (so called "support vectors"). Examples of other classification algorithms include nearest neighbour method, network-based classifications (including Bayesian networks), Hidden Markov models, pattern recognition, clustering around centroids and stepwise classification methods [159-162].

It may be noted that all these algorithms produce discrete classifications assigning each case to one of the groups. The quality of the discrete classification algorithm can be assessed by overall accuracy of classification or by its sensitivity and specificity; in some cases, receiver operating characteristic (ROC) plot can be used for assessment and tuning of discrete (binary) classifiers.

Along with the allocation of the case to a class it may be important to provide the degree of confidence for the allocation, e.g. the probability of the assigned outcome. This probability is naturally available when classification is based on logistic regression

---

Figure 13: Principles of LDA and SVM classification algorithms

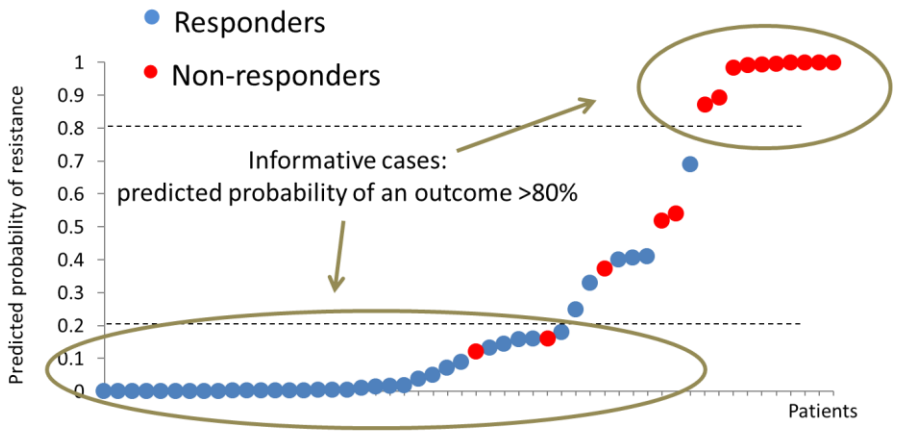


A: Linear Discriminant Analysis (LDA) B: Support Vector Machine (SVM)

(Figure 14). Similarly, a probability of outcome may be based on empirical data, as implemented in Oncotype-DX [67]. An advantage of the probability-based classification algorithms is that they allow assigning of the "non-classifiable" call, when the probability of either outcome is not high enough.

When a classification algorithm is trained to successfully classify cases in the training set it may be over-fitted to artificial or random features of the training set, instead of recognising important biological determinants underlying the classes differences. The risk of over-fitting is specifically high for classifications based on microarray data, where number of features (genes) is much higher than number of cases (tumours). A common approach to avoid this over-fitting is to train the algorithm on sub-sets drawn from the training set, using the excluded cases for assessment of the algorithm (e.g., leave-one-out test or bootstrapping [163]). Another way to reduce the risk of over-fitting is to reduce the number of features (genes), agglomerating them or using only the most informative ones, which are additionally supported by biological evidences. Because of the over-fitting, the accurate assessment of a multi-gene signature can be performed on an independent validation dataset.

Figure 14: Example of a probabilistic classification based on logistic regression



Exp(p/(1-p)) = 1.294 - 3.97691*__UCRC__ - 4.85412*__IMPDH2__ + 5.06282*__NDEL1__ + 10.4834*__RBL2__

Note: Reproduced from A.Larionov & W.Miller (2010) [71] with author's permission

Several software tools are available to perform clustering and classification. Combination of the Cluster-3 and Java-Tree-View tools provides a simple yet powerful entry level suite for clustering [164,165]. These are open source programs. They have a graphical user interface friendly to a user with a biological background. Cluster-3 allows the user to select from multiple alternative options for data pre-processing, a wide range of distance metrics and agglomeration algorithms, hierarchical clustering, K-means partitioning, SOM and PCA analyses. Cluster-3 outputs the results into text files, which can be opened by Java-Tree-View, which has rich user-friendly interface for drawing heatmaps, dendrograms and other illustrations, commonly used to represent the results of clustering and partitioning analysis.

The advanced classification and clustering tasks can be performed using R-scripting and specialised R-packages. R has a wide variety of packages for clustering and classifications tasks. Basic clustering functions are embedded into the R Stats package. Advanced clustering and classification tasks can be performed using the packages Cluster, Amac, Cclust, Clue, Dcluster, E1071, Pvclust, Splancs, Hmisc, Gclus, Fpc and Flexclust. The current project intensively uses the *pam()* function implemented in the R Cluster package [166,167] as the core element of our classification algorithm. The PAM algorithm will be discussed in details below (section 3.4). Additionally, during the results visualisation, we use hierarchical clustering features called by the *heatmap()* function implemented in the R Stats package.

1.5 Aim and objectives

Endocrine resistance remains an important issue in the treatment of breast cancer. There is no clinical marker to predict endocrine resistance in ER+ve patients or to guide a targeted treatment to overcome the resistance. Transcriptomic profiling is one of the most promising approaches to study the markers and mechanisms of endocrine resistance. Multiple transcriptomic datasets are publicly available that contain endocrine resistant samples. Multiple multi-gene expression signatures were published to interrogate the molecular mechanisms of endocrine resistance. At the same time, there are no transparent and user-friendly bioinformatics algorithms and tools for selection of endocrine resistant specimens from the available published datasets and for classification of these endocrine-resistant samples using relevant multi-gene signatures.

1.5.1 Aims

The aim of this study is therefore to design a bioinformatic pipeline to classify publicly available breast cancer transcriptomic datasets according to the relevant multi-gene signatures. It will also attempt to perform the classification of endocrine-resistant samples from the public datasets using multiple transcriptional signatures relevant to the mechanisms of endocrine resistance.

1.5.2 Objectives

- The identification of appropriate endocrine resistant cases in available public transcriptomic datasets
- Through literature review, determine a set of transcriptional signatures associated with endocrine resistance
- Pre-process the selected datasets for further analysis
- Translate signatures to the namespaces of relevant datasets
- Classify the resistant cases according to the molecular mechanisms represented by the transcriptional signatures
- Design a simple to use web-interface to present the pipeline and results

2 Bioinformatics analysis

The bioinformatics analysis has been performed according to the project's objectives.

2.1 Selection of datasets

The datasets were searched in GEO and ArrayExpress public repositories. At the time of the project preparation (March 2012) querying GEO datasets for "breast" produced 103 Datasets and 1309 Series. The search was repeated by including only the series with more than 50 samples annotated as "transcriptional profiling". This narrowed the results to 271 series. The GEO annotation system has not allowed for more detailed automatic selection of the relevant datasets. Therefore the remaining 271 candidate datasets were reviewed manually according to the following criteria:

1) Series contain endocrine-treated samples;
2) Annotation includes response to endocrine treatment;
3) Series are not selected on the basis of lymph-nodes and HER2 status;
4) Authors used frozen excision- or core- biopsies;
5) Data was obtained on Affymetrix or Illumina microarray platforms.

In addition, two proprietary datasets collected in the Edinburgh Breast Unit were included to the project. The datasets finally selected for analysis are shown in Table 3 and described in detail below.

2.1.1 Edinburgh datasets

Over the last 20 years the Edinburgh Breast Research Unit collected endocrine-treated specimens for multiple different studies. Some of these samples have come from endocrine-resistant tumours. All samples included in the current projects were collected with the patient's informed consent. The studies were conducted with local ethics committee's approval and supervision (initial LREC 2001/8/80, LREC 2001/8/81 and later amendments, including amendment of 2007 number 06/S1103/65).

Table 3: Transcriptional datasets containing data on endocrine resistant tumours

Dataset	Microarray platform(s)	Samples			Endocrine treatment setting	Time of biopsy	Criteria for endocrine resistance
		total	treated	resistant			
Edinburgh RS dataset	Illumina HT12	55	55	55	Mixed endocrine treatments	On treatment: at time of relapse/ progression	Relapse or progression on treatment
Edinburgh L23 dataset	Affymetrix U133A Illumina HT12	167	167	27	Neo-adjuvant letrozole	On treatment: at time of progression	<50% reduction within 3 months of treatment
U133A subset of GSE2990, GSE6532 and GSE9195	Affymetix U133A	327	190	49	Adjuvant tamoxifen	Primary tumours before treatment	Relapse within 3 years on endocrine treatment
U133-Plus-2 subset of GSE6532 and GSE9195	Affymetrix U133-Plus-2	163	163	15	Adjuvant tamoxifen	Primary tumours before treatment	Relapse within 3 years on endocrine treatment
GSE17705	Affymetrix U133A	298	298	36	Adjuvant tamoxifen	Primary tumours before treatment	Relapse within 3 years on endocrine treatment
GSE4922	Affymetrix U133A/B	289	66	23	Adjuvant endocrine	Primary tumours before treatment	Relapse within 3 years on endocrine treatment
GSE16391	Affymetrix U133-Plus-2	48	48	30*	Adjuvant tamoxifen or letrozole	Primary tumours before treatment	Relapse within 3 years on endocrine treatment

Two datasets available in the Edinburgh Breast Research Unit were selected for this project.

2.1.1.1 Edinburgh RS dataset

55 breast cancer biopsies [168] were taken from:

- Primary tumours growing on endocrine treatment in pre-operative settings (neo-adjuvant or advanced-disease treatments)
- Local relapses developed during adjuvant endocrine treatment.

Endocrine treatment included either tamoxifen or an aromatase inhibitor of the 3rd generation (letrozole, exemestane or anastrozole). Core- or excision- biopsies were snap-frozen and kept in liquid nitrogen until analysis.

Whole genome transcriptional profiles were obtained using Illumina HT-12 chips in the Wellcome Trust Clinical Research Facility at the Edinburgh Western General Hospital. Background correction, probes summation and quantile normalization were performed using Illumina Genome Studio v2011.1 (gene expression module version 1.9.0). The pre-processed gene expression values were exported to a tab-delimited file for further analysis in R-packages.

All 55 samples in this dataset were collected at the time of endocrine resistance, providing the largest single series analysed in this project.

2.1.1.2 Edinburgh L3 dataset

This is another internal dataset collected in the Edinburgh Breast Research Unit. The sequential biopsies from breast cancers treated with neo-adjuvant letrozole were taken at diagnoses (core biopsies), after 2-3 weeks of treatment (core biopsies) and during surgery after 3-6 months of treatment (excision biopsies). The samples were snap-frozen and stored in liquid nitrogen until analysis. Micro-array profiling was performed similarly to the RS dataset, except the Illumina HT-12 chips were run in the genomics laboratory of Roslin Institute (the University of Edinburgh).

A subset of 13 endocrine-resistant tumours from this series was suitable for the current project. These were biopsies taken after 3 months of treatment.

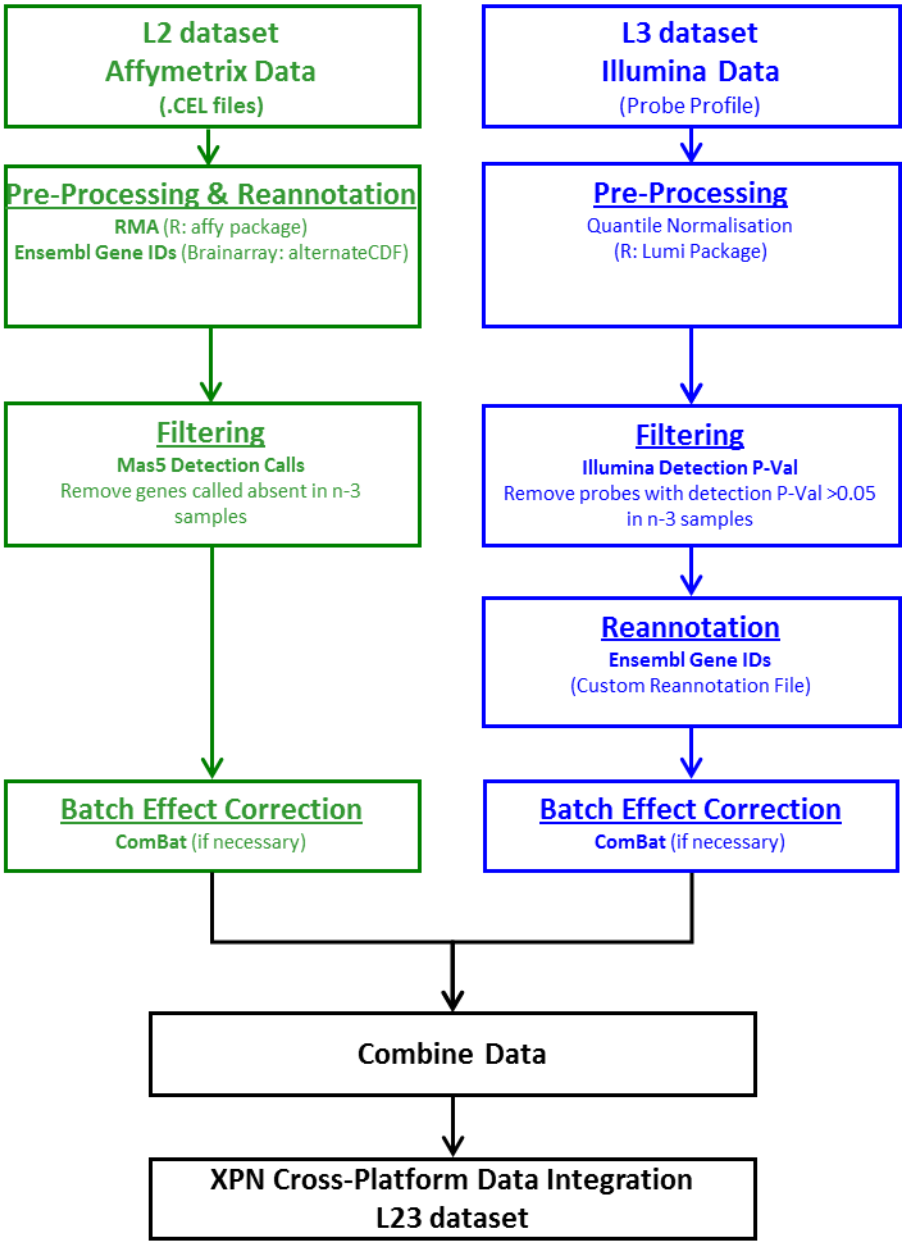
2.1.1.3 Edinburgh L2 dataset (GSE20181)

This dataset was generated in the Edinburgh Breast Unit in a way similar to the L3 dataset described above, except the transcriptomic profiles were obtained using Affymetrix HG-U133A chips. A detailed description of the dataset is provided in earlier publications [158,169]. The source CEL files and clinical annotations are available in GEO (GSE20181). 14 samples from this series were selected for the current project: all of them represented biopsies taken from resistant breast cancers after 3-6 month of neo-adjuvant treatment with letrozole.

2.1.1.4 Combining L2 and L3 datasets

The L2 and L3 series were derived from very similar cohorts of patients. However, they were profiled using different micro-array platforms. The different microarray platforms were used intentionally, because the primary purpose of the L3 series was to provide a platform-independent dataset for validation of the results obtained earlier using the L2 series. The total numbers of samples in each series was sufficient for analysis based on the whole group [78,157,158,169]. However, most of the tumours in the groups were responsive to treatment and the subsets of resistant cases were small (13 and 14 cases for L2 and L3 datasets respectfully). Taken separately, these small subsets could only be used for observational analysis [78]. To facilitate using the L2 and L3 series in the current project it was decided to join them using the pipeline for Affymetrix and Illumina data integration [170], which is summarised on Figure 15. Briefly, the procedure included the following. In addition to the standard pre-processing steps, the Affymetrix-Illumina integration pipeline includes (i) probe re-annotation, (ii) cross-platform probe mapping and (iii) data integration at the signal level. The cross-platform probe mapping was mediated by Ensembl gene IDs. Affymetrix probes were re-annotated to the Ensemble IDs using a custom CDF file to include only accurately annotated probe sets [149]. Similarly, the Illumina IDs were mapped to Ensembl genes using robust re-annotation based on a composite list from ReMOAT [148], BioMart and a custom BLAST search performed using the probe sequences [170]. Finally, the data integration at signal level was done with the cross-platform-normalisation (XPN) procedure developed by Shabalin *et al* [132].

Figure 15: Combining L2 and L3 datasets



Note:

Modified from Turnbull, Kitchen, Larionov *et al* (2012) with authors' permission [170]

Integration of the L2 and L3 datasets resulted in the L23 dataset consisting of 26 samples resistant to neo-adjuvant letrozole. The main drawback of the integration was that it reduced the number of the available genes to only 7,160 genes. The gene number reduction occurs during the probe re-annotation and cross-platform mapping. Importantly, the remaining 7,160 genes included large parts of the selected transcriptional signatures. Therefore, the gene number reduction has not precluded the downstream classification analysis.

2.1.2 Tamoxifen-treated dataset from Oxford, Uppsala and London

This collection of datasets was designed to study adjuvant treatment with tamoxifen [171] and has been extensively re-analysed in a number of influential studies [172-174]. It presents an impressive example of international collaboration, combining efforts of research groups from Oxford, London, Brussels, Sweden and Singapore. Their microarray data is publicly available from the GEO in the GSE2990, GSE6532 and GSE9195 series. Each series combines samples from different centres. Clinical annotations are included in supplementary files. The response assessment and identification of resistant cases can be done using relapse-free-survival, provided in the clinical annotations.

2.1.2.1 GSE2990 series

This was the first series published by the consortium in 2006 [171]. The series includes 189 samples: 64 treated with adjuvant tamoxifen and 125 non-treated. The biopsies were collected during surgery and frozen until analysis on Affymetrix U133A arrays. 101 samples (40 treated and 61 untreated) were collected in Oxford (UK) and microarrayed in the Jules Bordet Institute in Brussels, Belgium. 88 samples (24 treated and 64 un-treated) were collected in Uppsala (Sweden) and analysed in Genome Institute of Singapore.

2.1.2.2 GSE6532 series

This series extends the GSE2990 tamoxifen-treated dataset. GSE6532 was published in 2007 [172]. It includes 414 samples: 137 un-treated and 277 tamoxifen treated breast cancers. First, it adds U133B (GPL97) profiles to the samples reported in GSE2990.

Then it provides many new samples collected in Oxford, Uppsala and London (Guys Hospital) profiled with U133-Plus-2.0 Affymetrix arrays (GPL570).

2.1.2.3 GSE9195 series

This series expands the previous datasets with an additional 77 tamoxifen-treated tumours collected in Guys Hospital in London.

2.1.2.4 Combining tamoxifen-treated datasets

Similarly to the Edinburgh datasets described earlier, the tamoxifen-treated series (GSE2990, GSE6532 and GSE9195) share their clinical features but use different micro-array platforms (U133A/B and U133-Plus-2.0). Again, the majority of the tumours in these datasets were responsive to endocrine treatment. Thus we had to combine the resistant samples from different series to obtain sufficient numbers for the downstream analysis.

In this case, the array platforms share many design features because they came from the same manufacturer. In fact, the U133B array just complements the U133A by adding new targets, and the U133-Plus-2.0 array just allocates all the U133A and U133b probes onto a single chip (by increasing the density of printing). Importantly, the arrays share the same probe names, so there is no need for probe re-annotation and mapping during the cross-platform integration. At the same time, it has been observed that technical differences between U133A and U133 Plus 2.0 may require adjustments in the classification procedures [175]. To simplify these series' integration, while avoiding the platform-dependent influences, it was decided to merge the data from different series that were generated using the same microarrays. Thus, the samples were combined in the following way:

- U133A CEL files from GSE2990/GSE6532 were combined into one dataset (Tam-U133A set)
- U133-Plus-2.0 CEL files from GSE6532 and GSE9195 were combined into the second dataset (Tam-U133Plus2 set).

As the result, the combined Tam-U133A dataset included 49 resistant tumours (34+15). The combined Tam-U133Plus2 dataset included 15 tumours (9+6) resistant to adjuvant tamoxifen.

2.1.3 GSE17705 dataset

This dataset accompanies a recent study that searched for transcriptional signature of ER-signalling [175]. The study analyses several cohorts of patients who received different modalities of endocrine treatment. While reviewing GSE17705 data in GEO it was found that the submission is split into two sub-sets (255 and 43 cases), both treated with adjuvant tamoxifen for 5 years and profiled by Affymetrix HG-U133A arrays. Samples used for this series were described as fresh frozen tissue. No human readable clinical annotation was included in supplementary files. However, the annotation was available in the standard GEO submission files (SOFT, MINiML or Series Matrix files). The clinical annotation included data on disease-free survival, which was used for selection of 36 patients resistant to adjuvant tamoxifen.

2.1.4 GSE4922 dataset

This dataset is reported in a paper that aimed to develop a transcriptional signature for histological grades in breast cancer [176]. The frozen excision biopsies were profiled on Affymetrix U133A&B gene chips. Review of the dataset's data submitted to the GEO has shown that the whole set of 289 samples consist of two parts: 249 samples from Sweden and 40 samples from Singapore. Only Swedish samples are accompanied with information about treatment and response; of those only 67 have received endocrine treatment; only 23 of them could be classified as endocrine-resistant (basing on the relapse-free survival data available in clinical annotations).

2.1.5 GSE16391 dataset

This dataset includes 55 tumours collected in a large multi-centre trial of Letrozole and Tamoxifen in an adjuvant setting [177]. Excision biopsies were collected at surgery and kept frozen until analysis. The tumours were profiled using Affymetrix U133 Plus 2.0 chips. Clinical data included relapse-free-survival, which allowed the assessment of response and the selection of 30 resistant ceases. The extremely high proportion of

resistant cases may be explained by a bias in sample selection: the original trial enrolled many thousands of patients, only 55 of them were selected for the microarraying.

2.1.6 Examples of non-included datasets

A number of other studies have been identified that could potentially provide sufficient numbers of transcriptional profiles from endocrine resistant tumours. Most of these studies were excluded because of insufficient clinical annotations (e.g. GSE22219, GSE1456, GSE2034, GSE1456). Usually, the authors did not provide clinical annotations which were not necessary for the data analysis performed in the original studies. Some studies provided apparently sufficient clinical information, which yet lacked some important details. For instance, GSE12093 series provides relapse-free survival expressed in units, which cannot be confidently identified either as days, weeks, months or years. Other studies provided data with technical glitches, e.g. GSE26971 series' annotation file had irregularities that required manual corrections, which eliminated this series from the project. Finally, some of the very promising datasets just did not provide actual primary microarray data. For instance, E-MTAB-520 dataset (publicly available at Array Express) accompanies an interesting recent neo-adjuvant study designed similarly to the Edinburgh L2 and L3 studies. Unfortunately, the authors only deposited data to the Array Express on 205 Illumina probes, relevant to the data analysis presented in their paper [178].

2.2 Selection of signatures

The next objective of the project was to identify transcriptional signatures, which could be used for classification of the selected resistant samples. This was done by manual literature mining. The preference was given to the following signatures:

1) The signatures associated with pathways of importance for endocrine resistance;
2) The signatures developed or tested using breast cancer clinical samples;
3) The signatures providing probe IDs for their genes;
4) The signatures designed to classify tumours into no more than two classes;
5) The signatures with data about direction of the gene changes indicating to high activity of the associated pathway.

The selected signatures are summarised in Table 4 and described in detail below.

Table 4: Transcriptional signatures associated with endocrine resistance

Pathway	Publication	N of genes	Microarray Platform(s)	Signature source
ESR1	Symmans 2010 [175]	165	U133A	Genes co-expressed with ESR1 in clinical specimens of breast cancer
PIK3A	Loi 2010 [174]	278	U133A	Genes co-expressed with PIK3A mutations in clinical breast cancer specimens
MYC	Bild 2006 [124]	248	U133 Plus-2	Genes activated in transfected primary mammary epithelial cell cultures
E2F3	Bild 2006 [124]	298	U133 Plus-2	Genes activated in transfected primary mammary epithelial cell cultures
RAS	Bild 2006 [124]	348	U133 Plus-2	Genes activated in transfected primary mammary epithelial cell cultures
Beta-Catenin	Bild 2006 [124]	98	U133 Plus-2	Genes activated in transfected primary mammary epithelial cell cultures
SRC	Bild 2006 [124]	73	U133 Plus-2	Genes activated in transfected primary mammary epithelial cell cultures
Hypoxia	Buffa 2010 [179]	58	U133A	Meta-analysis of multiple transcriptomic studies
Invasiveness	Shats 2011 [180]	100	U133A	Mixed analysis of cell line models and clinical breast cancer specimens
Stemness	Shats 2011 [180]	100	U133A	Genes activated in normal fibroblasts reprogrammed to pluripotency

2.2.1 Activity of ESR1 signalling

Oestrogen receptor signalling, continuing its activity despite the endocrine treatment, is one of the most studied mechanisms of endocrine resistance. It was suggested that ESR1 signalling on endocrine treatment may be maintained (i) through ligand-independent activation of oestrogen receptors or (ii) through hyper-sensitivity of oestrogen receptors to low concentrations of oestrogens or (iii) through over-expression of oestrogen-receptor co-activators or through other mechanisms [83].

The selected oestrogen signalling signature includes genes derived on the basis of co-expression with ESR1 in clinical specimens of breast cancer [175]. 106 of the selected genes are positively correlated and 59 genes are negatively correlated with oestrogen signalling. Authors provide U133 Affy probe IDs for each gene. The signature had been extensively validated on a number of breast cancer datasets. Unfortunately, the manuscript does not provide a direct reference to the original training cohort. This precludes implementation of many classification algorithms requiring a training dataset (such as LDA, SVM of ANN algorithms).

2.2.2 PIK3CA activation

Activation of PIK3CA signalling is one of the established mechanisms supporting malignant growth [181]. Activating PIK3CA mutations are found in ~30% of ER+ve breast tumours, being one of the most frequent somatic mutations in this type of breast cancer [182]. It was hypothesised that active PIK3A signalling can support breast cancer growth in ESR1-independent manner, despite the effective inhibition of oestrogen signalling by endocrine treatment (Figure 6).

The selected PIK3CA signature was derived from analysis of a very large number (1800) of clinical breast cancer specimens [174]. The signature was validated on two independent datasets. Authors provide Affy probes IDs and directions of changes associated with active and inactive state of PIK3CA signalling. Similarly to the ESR1 signature, the GEO datasets referred in the publication does not provide information on the PIK3A status. Thus the signature cannot be used in the classification algorithms requiring training.

2.2.3 Signatures for oncogenic pathways from Bild et al 2010

Bild et al (2010) developed an experimental pipeline, which allowed them to generate transcriptional signatures for several common oncogenic pathways, including MYC, E2F3, RAS, SRC and Beta-Catenin signalling [124]. The signatures were derived from primary human mammary epithelial cell cultures, transfected with recombinant adenoviruses, leading to activation of specific target pathways. In addition to being derived from mammary epithelial models, all signatures were applied to a large clinical dataset of breast cancers (amongst the other cancers studied in the paper). The studied pathways were selected on a basis of their frequent involvement in carcinogenesis. Theoretically, activation of these pathways in ER+ve tumours during endocrine treatment may support tumour growths and lead to resistance. Some of the studied pathways were directly linked to endocrine resistance in breast cancers by others (e.g. MYC- associated signalling [114]). Authors provide Affy IDs and the direction of changes associated with the signalling activation. Again, no training dataset accompanies the publication.

2.2.4 Hypoxia

Insufficient vascularisation and intra-tumoural hypoxia are important hallmarks of malignant growth [91,92]. It has been shown that hypoxia is strongly associated with proliferation in endocrine-treated breast cancers [178]. The selected signature is based on a large meta-analysis of different cancers, which was designed to reveal a compact and robust consensus list of genes associated with intra-tumoral hypoxia [179]. The prognostic value of the signature has been validated on several datasets, including breast cancer. The signature includes 58 genes: 49 of positively associated and 9 negatively associated with hypoxia. The genes are provided with their Affymetrix IDs.

2.2.5 Stemness and Invasiveness

The last group of signatures included in the project was developed to detect stem-cells and invasive properties of the tumours. Both of these pathways are important to maintenance of malignant growth in general [91,92] and were linked to endocrine response in breast cancers [183]. Two signatures were selected for this project from the paper of Shats *et al* [180]. However, analysing their gene identities, it was difficult to draw a clear line between stemness and invasiveness. There is a tight inter-play between

mammary epithelial de-differentiation and epithelial-mesenchyme transition. The first being a way to acquire a stem-cell-like phenotype and the second required for the acquisition of invasive properties. For the purpose of this project, one of the signatures was assigned to be indicative to "stemness" and the other to "invasiveness". The "stemness"- associated signature was derived from a cell line model, where several transcriptional factors ectopically expressed in human fibroblasts reprogrammed the cells to pluri-potency. The "invasiveness"-associated signature was developed mainly through meta-analysis. Both signatures were tested on of several experimental datasets, including datasets based on breast cancer. Both signatures provide AFFY IDs and directions of change indicating to the activity of the associated pathway. Again, no training datasets are publicly available with the paper.

2.2.6 Examples of non-included signatures

The multi-gene signatures for intrinsic breast cancer sub-types [65,66,184,185] and transcriptional signatures to HER2 amplification [186,187] may be mentioned amongst the most relevant omitted pathways. The first has not been included because it splits tumours into more than two classes, which would require the algorithms implementation to be different from the most of other selected signatures. The second was not selected because there are non-transcriptional methods (FISH and IHC) widely used in clinical practice to detect HER2 amplification. Several candidate signatures were excluded because they do not provide probes IDs for the genes [188-190]. Many of the other highly relevant signatures could potentially be added to the project. However, the primary task of the project was not to analyse a large number of signatures, but to develop a bioinformatics tool for their analysis. Therefore, the list of already selected signatures was considered sufficient for the task.

2.3 Datasets import and pre-processing

Prior to downstream analysis, the selected datasets had to be imported from repositories to a local computer and pre-processed.

The Edinburgh datasets did not need to be imported because they were generated locally. Pre-processing of the Edinburgh L23 dataset has been described earlier. The Edinburgh RS dataset (generated on Illumina HT12 chips) was background-corrected,

summarised to gene level and quantile normalised using the Illumina Genome Studio suite. The pre-processed gene signal values were imported to R and log-transformed. Importantly, the Genome Studio may generate negative background-corrected gene signal values. Negative values were converted to zeroes without the log transformation (Appendix: A.1.1.).

Import and pre-processing of the datasets selected from the GEO have been conducted using specialised R packages available from the Bioconductor [191]. First, the data was downloaded using GEOquery R package [192]. Then data was log-transformed, background-subtracted, summarised and normalised using Affy and affyPLM packages [137,193]. For consistency, all GEO datasets were pre-processed using RMA.2 background correction, median polish summation and quantile normalisation. Examples of the R scripts used for data import and pre-processing are given in Appendix (A.1.2 and A.1.3). Finally, all the pre-processed and log-transformed data was median-centred and scaled to a range [-1 to 1] before use in the downstream classification algorithm.

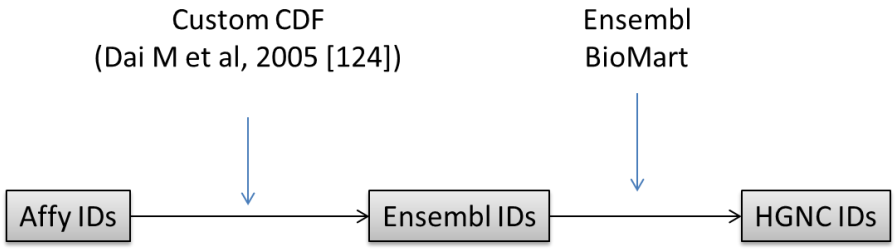
2.4 Signatures translation and pre-processing

To use signatures in dataset classification, the signatures shall use the same gene IDs as the datasets.

All selected signatures contained Affymetrix U133 IDs for their genes (Table 4). These IDs were directly compatible with the datasets imported from GEO (GSE2990, GSE6532, GSE9195, GSE17705, GSE4922 and GSE16391; Table 3). In this case the signatures pre-processing included only unified coding for the directions associated with high activity of the pathway: the up-regulated genes were coded as 1, the down-regulated genes were coded as -1. The composition of signatures used for classification of the GEO datasets is available in the Appendix (A.2).

More pre-processing had to be applied to the Edinburgh datasets. The genes in the Edinburgh RS dataset were identified by HGNC gene IDs (as exported from Illumina Genome Studio). To ensure accuracy of translation between the Affymetrix and HGNC IDs the conversion was mediated by a custom CDF annotation file (Figure 16, [149]).

Figure 16: Custom CDF-mediated conversion of Affy IDs to HGNC IDs



The custom CDF files for Affymetrix chips are maintained and regularly updated by Michigan University. For this project we used the latest available version 15.1.0 released on 18th January 2012, which is available at http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/15.1.0/ensg.asp.

Comparing the HGNC gene lists generated by custom CDF files to the lists supplied by the signatures' authors, it was evident that the re-annotation does not change the genes identities; however, it removes a noticeable number of the probe sets, whose mapping is not reliable.

Affymetrix chip design is known for its redundancy, when multiple AFFY probes may target the same HGNC gene. After translation of U133 codes to HGNC IDs the redundant genes were collapsed. The collapsed genes were additionally checked for consistency of the constituting AFFY probesets. Inconsistent changes of the AFFY probesets targeting the same gene were extremely rare: only two genes were excluded for being inconsistent (Stemness signature, Tables 5 and 6). Then the signatures' genes (presented as HGNC IDs) were mapped to the HGNC IDs present in the RS dataset. Finally, only the genes informative in the dataset (detected with $p<0.05$ in at least 10% cases) were used for the downstream classification. The signatures translation and pre-processing for the Edinburgh RS dataset is summarised in Table 5.

The genes in the Edinburgh L23 dataset were coded by Ensembl IDs, as generated by the pipeline for Affymetrix and Illumina integration (Figure 15, [170]). Translation and

Table 5: Pre-processing of signatures for Edinburgh RS dataset

Signature	Affy IDs	Ensemble IDs	Redundant HGNC IDs	Non-Redundant HGNC IDs	Consistently directed non-redundant HGNC IDs	Present in RS dataset	Informative in RS dataset*
Beta-Catenin	98	81	80	73	73	58	54
E2F3	298	259	256	220	220	197	137
ESR1	165	152	152	150	150	136	115
Hypoxia	58	48	48	39	39	38	38
Invasiveness	100	99	99	89	89	84	81
MYC	248	190	189	167	167	148	113
PIK3A	278	250	248	211	211	200	166
RAS	348	298	297	234	234	219	133
SRC	73	58	56	53	53	49	44
Stemness	100	89	89	79	77	68	59

* Detected with $p<0.05$ in at least 10% cases

Table 6: Pre-processing of signatures for Edinburgh L23 dataset

Signature	Affy IDs	Redundant Ensembl IDs	Non-Redundant Ensembl IDs	Consistent non-redundant Ensembl IDs	Present in L23 dataset
Beta-Catenin	98	81	74	74	58
E2F3	298	259	223	223	112
ESR1	165	152	150	150	133
Hypoxia	58	48	39	39	34
Invasiveness	100	99	89	89	66
MYC	248	190	168	168	96
PIK3A	278	250	212	212	184
RAS	348	298	235	235	123
SRC	73	58	55	55	38
Stemness	100	89	79	77	54

pre-processing of signatures for the L23 dataset has been performed similarly to the procedure applied for the RS dataset, except (i) there was no need for custom-Ensembl-to-HGNC conversion and (ii) all the genes in the dataset were considered informative (which was assured by the dataset shrinkage during the Affy-Lumi integration). The signatures translation and pre-processing for the Edinburgh L23 dataset is summarised in Table 6.

2.5 Development and implementation of classification algorithm

Currently, most of the published multi-gene signatures relevant to the endocrine resistance come without the original training dataset [124,174,175]. In many cases, when the "consensus" signatures are developed via meta-analysis the "original" training dataset may not exist at all [179,180]. The classification algorithms, provided with some of the signatures [124,185], are often complicated, non-transparent and tuned to specific dataset or micro-array platform [175,184]. To address these practicalities, it was decided to develop a new classification procedure that

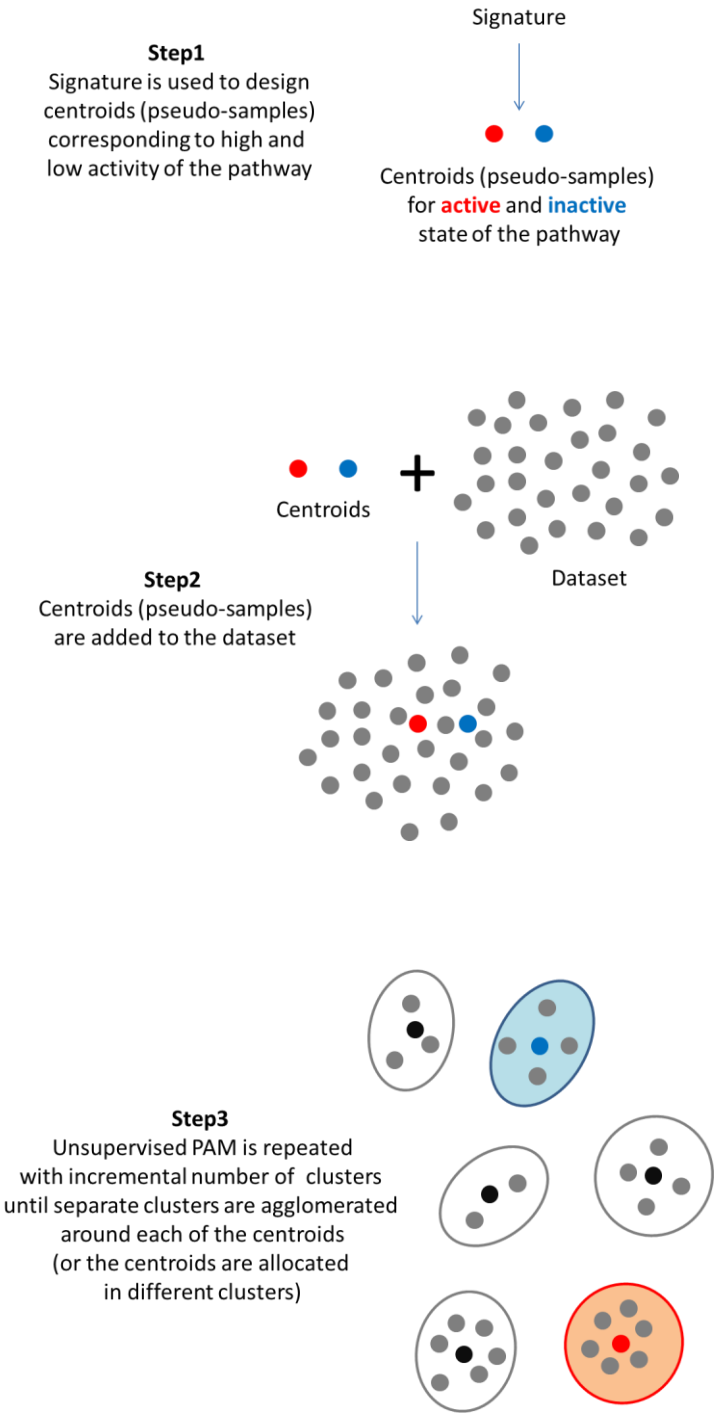
1) does not require a training dataset
2) takes into account only up- or down- regulation of genes.

The first condition was selected by necessity; the second was a conscious choice, aimed to reduce platform-dependence. Indeed, the signal values vary significantly between platforms. At the same time, the direction of changes of the most up- or down-regulated genes are expected to hold independently of the used microarray platform.

The developed procedure is based on addition of signature-based pseudo-samples to the dataset followed by unsupervised Partitioning Around Medoids (PAM, [167,194]). The suggested implementation includes iterations of PAM with incremental numbers of clusters and different distance metrics, hence we call it Iterative Consensus PAM. The main steps of the procedure are illustrated on Figure 17:

Pre-processing (not shown on the picture): Datasets were pre-processed using commonly used procedures, as described earlier. Then datasets were median-centred

Figure 17: Classification algorithm: Iterative Consensus PAM



Step 4
PAM is repeated with alternative distance metrics:
Euclidian, Manhattan and Spearman correlation.
Cases are classified by consensus between the PAMs
repeated with different distance metrics.

and scaled to a range [-1 to 1]. Only the sub-set of all data containing the signature's genes was taken for classification. For instance, if a signature contains 250 genes and the dataset contains 50 cases, then a 250x50 matrix of centred and scaled values was prepared for the classification procedure.

Step 1: A pathway's signature is used to construct two pseudo-samples (centroids): one centroid corresponding to the high activity of the pathway and the other centroid corresponding to the low activity.

Importantly, the actual values assigned to the centroids are adjusted to the dataset's median absolute values. For instance, if the median absolute value in the pre-processed dataset is 0.3, then the signature's up-regulated genes are assigned a value of +0.3 and the down-regulated genes are assigned -0.3. The resultant vector is used as the pseudo-sample corresponding to the active state of the pathway. The pseudo-sample corresponding to low activity of the pathway is designed by multiplying the "high-activity" vector by -1.

Step 2: The centroids are added to the datasets.

Step 3: The datasets are subjected to unsupervised partitioning using PAM (as implemented in Cluster R-package, [166]). Unsupervised PAM is repeated incrementing number of clusters until separate clusters are agglomerated around each of the centroids (or the centroids are allocated in different clusters). The iterations start from 3 clusters and may continue up to the total number of samples in the dataset.

Step 4: The above procedure is repeated with three different distance metrics: Euclidian, Manhattan and Spearman correlation. Consensus between all three metrics is used for the final class allocations.

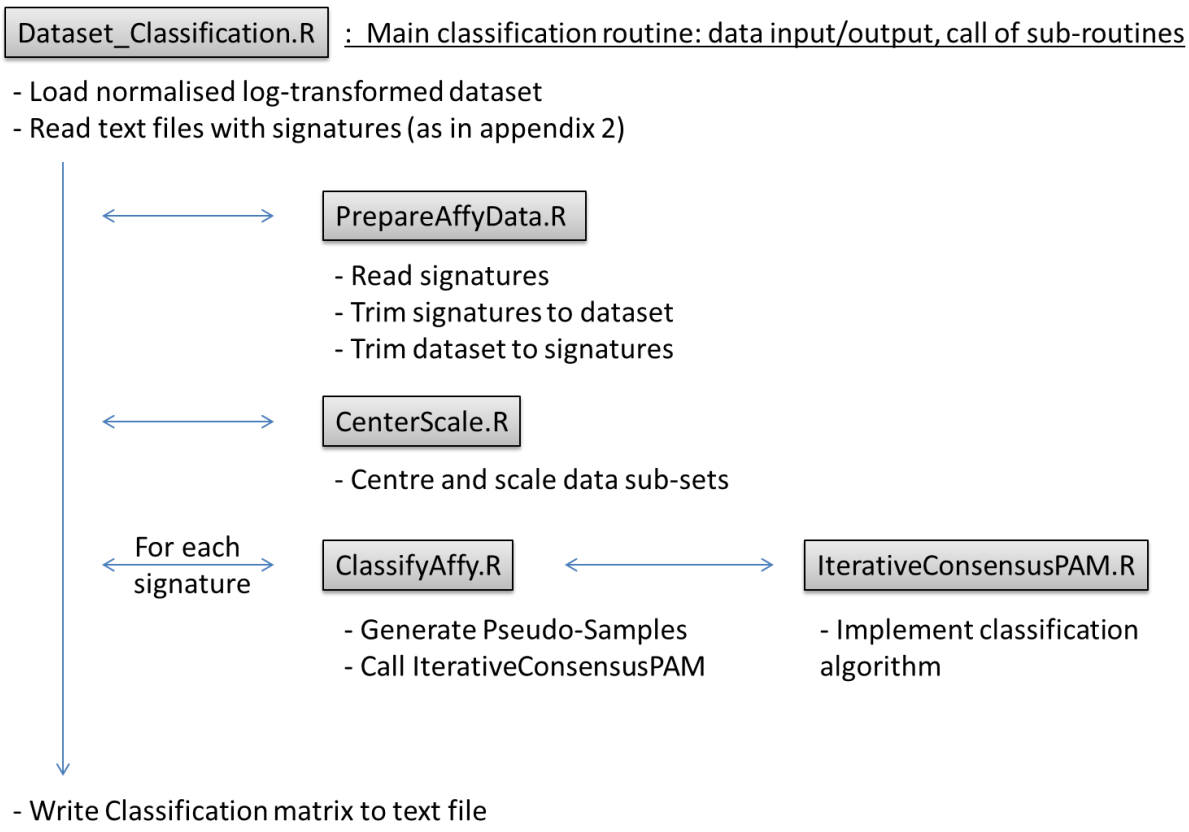
Importantly, the algorithm does not force each sample into any category. The number of clusters starts from 3. Even if the desired classification is achieved after the first round of partitioning (which is often the case), one cluster includes "high" activity

samples, one “low activity” and one – the “inconclusive” samples. If more than one iteration are needed to separate “high” and “low” clusters, then the “inconclusive” samples are split into several groups. Finally, some samples are classified as inconclusive during the consensus step.

The classification procedure was implemented in R as shown on Figure 18. Briefly: one execution of the main procedure classifies one dataset by multiple selected signatures. The main procedure performs data input, calls sub-routines and writes out the classification matrix. Examples of the R-scripts are available in the Appendix (A.1.4). Finally, an additional script was used to visualise the classification matrix as a heatmap (Appendix A.1.5). This script was run separately, after the main routine, allowing to adjust the resolution and proportion of the heatmaps for required purposes.

The classification was performed for all datasets and signatures selected for the project. The resulting heatmaps are available in the Appendix (section A.3).

Figure 18: Implementation of classification in R



2.6 WEB-site presenting the results of analysis

The current project is a part of a larger series of studies in endocrine resistance of breast cancer that is carried out in the Edinburgh Breast Unit. Within this broader framework, the developed pipeline may constitute a prototype for a future web-hub collecting (i) data on endocrine-resistant breast cancer specimens, (ii) collecting multi-gene signatures relevant to endocrine resistance and (iii) providing tools to apply the signatures to the data.

Because of the current lack of commonly accepted standards for reporting clinical data in breast cancer datasets and the lack of commonly accepted ways to report multi-gene signatures, such a repository would need manual curation and the analysis procedures would involve many interactive semi-automatic steps. Therefore, developing an integrated fully automatic web-tool would not be feasible at this stage. At the same time, presenting the results of this project on the web was considered important to

illustrate a concept of a web-hub for signatures, datasets and tools focused on endocrine resistance.

The web site was written in plain HTML, using Framesets to organise the content and CSS (Cascading Style Sheets) to maintain unified formatting throughout all of the pages.

The site content includes one menu and 11 pages:

- Introduction
- Summary of selected signatures with links to corresponding publications
- Summary of the bioinformatics pipeline and the classification procedure
- Summary of selected datasets and links to the datasets in the GEO repository
- Pages with results of classifications: a separate page for each analysed dataset

The examples of HTML code and a web-page screen-shot are available in the Appendix (A.4). At the time of the project presentation the web site is available at the address: http://larionov.co.uk/ .

3 Discussion

Endocrine resistance is an important clinical issue in treatment of breast cancer [69]. It can be caused by different mechanism in different tumours [78] . These mechanisms are not yet fully understood [83]. Conventional therapy of endocrine resistance is still semi-empirical: usually it includes a change of the hormonal drug and/or addition of a cytotoxic treatment [71]. Several recent attempts to add targeted agents to endocrine treatment have not shown significant success because of absence of the biomarkers needed to guide the targeted treatment [195]. Transcriptional profiling is one of the most promising strategies that may be used for studying of markers and mechanisms of endocrine resistance in breast cancer. Multiple breast cancer transcriptional datasets and multi-gene signatures have been published over the recent years [124,168,174,175,177,179,180,188]. However, applying signatures is not yet a straight-forward process. It is complicated by several factors. First of all, analysis of the available datasets and signatures revealed lack of commonly accepted standards for reporting multi-gene signatures and for reporting clinical information on breast cancer. No microarray repository provides a standardised interface to capture and organise clinical information required to analyse endocrine resistance. For instance, endocrine treatment modality and results are often missed in phenotype annotations or presented in an arbitrary way. Similarly, there is no standardised way of reporting a multi-gene transcriptional signature. Different authors report it in different formats. The most convenient format yet is an Excel file in the supplementary materials [179,180]. At the same time, many authors still report the genes identities in supplementary PDF files or even in texts of their publications or on figures illustrating the signature performance [124,185,196]. Importantly, the genes IDs in signatures are often different from the genes IDs in the datasets of interest. Frequently there is no clarity about the training data used during the signature(s) development. Finally, there is lack of commonly accepted generic algorithms and software that can be used to apply a multi-gene signature to new datasets across micro-array platforms.

The main goal of this project was to classify endocrine-resistant tumours from publicly available datasets using multiple published multi-gene signatures for different

Table 7: Pipeline to apply multi-gene signatures to public datasets

Review of datasets, signatures and classification algorithms
• Assert biological meaningfulness of classification • Confirm computational feasibility of classification • Select of optimal/suitable algorithms for classification
Data download
• GeoQuery [192] • Manual download
Data import and pre-processing
• Standard data pre-processing - Background correction, Summation, Normalisation and Log-transfrormation Affy, affyPLM, Lumi, Beadarray, Illumina Genome studio [136-139,197] • Optional data pre-processing - Batch-correction (ComBat, XPN, [132,147]) - Filtering of genes - Selection of cases
Data integration (optional)
• Re-annotation and matching probes between datasets Custom CDF, REMoat [148,149] • Signals integration XPN [132]
Signature pre-processing
• Re-annotation and mapping probes to dataset (optional) Custom CDF, REMoat [148,149] • Trimming signature to dataset: - Removing probes absent in the dataset - Collapsing redundant gene IDs - Removing genes non-informative in the dataset
Data centring and scaling
Reduces platform-specific bias
Design of signature-specific centroids (“Pseudo-samples”)
• Direction of changes defined in signature • Median signal amplitude in the dataset
Classification
• Iterative consensus PAM [167,194] Custom algorithm implemented in R (Figures 17, 18; uses the R-package Cluster[166])
Results visualisation
• Heatmaps using appropriate R functions

mechanisms of endocrine resistance. The above complicating factors have been addressed within the context of project. A bioinformatics pipeline and a new classification algorithm have been developed to achieve the project goal.

3.1 Summary of the pipeline

The main steps of the pipeline are summarised in Table 7 and discussed below.

3.2 Datasets and signatures review

The pipeline starts with review of the datasets and signatures of interest. This is an important step prior to the computation because it assures biological meaningfulness of the analysis. It is also necessary to assess the computational feasibility and to select the most suitable classification algorithms and software.

3.2.1 Datasets review

Selection and review of the datasets, which may be used to study endocrine resistance in breast cancer, constituted a significant part of this project. Except for the series of resistant tumours being recently collected in Edinburgh, no datasets focused on collecting endocrine resistant samples have been published yet. At the same time, there is a number of large series on endocrine-treated patients, which can be used to extract information on the resistant patients.

More than 30,000 series were available in the GEO at the time of the project preparation; more than 1,000 of them were related to breast cancer. Including only "transcriptional profiling" series with more than 50 samples reduced this number to 271. No further meaningful reduction was possible using the automated mining of datasets available in the GEO. Annotations for all remaining series had to be reviewed manually. First we selected only the studies which contain endocrine-treated samples and provide information about responses to the treatment. Then, series selected by lymph-node, HER2 status or other special biological sub-types were excluded because they could be biased with regard of the diversity of endocrine resistance mechanisms. Only Affymetrix or Illumina platforms were included to assure high quality of the microarray platform and availability of reliable software packages for pre-processing. To assure high quality of the source material used for micro-arraying it was decided to

exclude series that used FNA and FFPE specimens, leaving only those, which profiled frozen core- or excision biopsies. None of the above parameters was available for automatic search in the annotations. Applying all these criteria allowed the identification of 5 top series which were included into the project: GSE6532 (including GSE2990 samples), GSE9195, GSE17705, GSE4922 and GSE 16391 (Table 3). Two more datasets were included because they were generated in the unit hosting the project (Edinburgh RS and L23 datasets; part of L23 is deposited in GEO as GSE20181).

An important part of the datasets' review was to select the endocrine resistant cases. All available series (except for the Edinburgh RS dataset) contained a mix of resistant, responding or even non-treated samples. Selection of resistance criteria depended on modality of treatment and numbers of available specimens. For the neo-adjuvant study (L23) the resistance was defined as less than 50% reduction in tumour volume within 3 months of treatment, as suggested by the authors that collected the series [169]. Response to adjuvant studies was assessed by the relapse-free survival after surgery. No threshold for resistance was set by the authors of the used datasets. Thus the resistance was arbitrarily defined as a relapse within 3 years of treatment. This is within the commonly accepted range of practices for adjuvant response assessment in breast cancer. Applying a 2 year threshold would significantly reduce the number of specimens available in the adjuvant datasets, precluding analysis on some of them.

3.2.2 Signatures review

There is no public repository or common standards to report a multi-gene signature. Therefore the signatures selection was based entirely on the literature search. Biological criteria for signature selection (outlined earlier in section 2.2.) were self-evident, including relevance of the signature to endocrine resistance and breast cancer. Initially, technical criteria for the signatures selection included availability of a clearly described training dataset. However, the signatures' review showed that this requirement would eliminate most of the published signatures, elsewhere available for the analysis. Therefore it was decided to exclude this requirement and accept signatures that at least provide the directions of gene changes corresponding to activity of the pathway in question. This decision precluded the use of many common classification algorithms (such as SVM, ANN, LDA or logit-regression), which was one of the

reasons for development of the new classification procedure. On the other hand, this forced decision had its positive side: it enhances platform-independence and safeguards against over-fitting.

The other criteria for signature selection were (i) the signature separates tumours into two classes and (ii) the authors provide probe IDs rather than gene names. Splitting tumours into only two classes was intended to simplify the classification algorithm. Requirement for the probes IDs was necessary to ensure accurate cross-platform translation of the signatures. The author's assignment of the gene names is usually based on the manufacturer's chip description (CDF) available at the time of the study. It was observed that a noticeable proportion of manufacturer's probe annotations may be inaccurate, or may become inaccurate because of the constantly changing human genome annotation [149]. A case of misleading annotations may be illustrated by ESR1 gene in U133 array platforms. 9 different probesets are annotated for ESR1 on U133 arrays (211235_s_at, 211234_x_at, 211233_x_at, 205225_at, 211627_x_at, 217190_x_at, 215552_s_at, 217163_at and 215551_at). Only one of these correlates well with oestrogen receptor protein in breast cancer (205225_at). The others represent alternatively spliced variants, mutated variants or other versions that do not reflect ESR1 expression in most of breast cancers. Availability of the original probes IDs in the published signatures allows controlling for the inaccuracies by using independently verified and updated custom versions of CDFs [148,149].

Within the signatures satisfying the above criteria, the further signatures selection was subjective. In general, an additional preference was given to the signatures derived or validated on clinical specimens, rather than cell lines, to account for heterogeneity of the tumour tissue. The 9 signatures summarised in Table 4 were considered sufficient for the purpose of this project.

Overall, the datasets' and signatures' reviews in this project were laborious and time-consuming because of the non-uniformity in reporting of gene signatures and clinical annotations.

3.3 Pre-processing of datasets and signatures

Affymetrix data download from the GEO, the data import to R and their pre-processing were performed using standard procedures and packages. Pre-processing of Illumina arrays in the RS dataset was also performed using standard Illumina's recommendations. The data Affymetrix-Illumina integration for L23 dataset has been described in details in the previous chapter and discussed in details elsewhere [170]. It may be interesting to study effects of different pre-processing options on the classification results. However, it was considered out of scope of the current project.

Importantly, the datasets were median-centred and ranged prior to the classification. This was intended to enhance platform-independence of the down-stream classification. The absolute range signal values may be dataset- or platform- specific. The median-centring and ranging of the data translated these platform-specific values toward generic interpretation in terms of the up- and down- regulation.

The steps required for signatures pre-processing included mapping signature's probes to the dataset and trimming signatures (removing probes absent in the dataset, collapsing redundant gene IDs and removing genes non-informative in the dataset). Except for the cross-platform probes/genes translations, these steps are intuitive and self-explanatory. For the cross-platform translations we used an annotation-based approach mediated by custom CDF files (Figure 16). The alternative approaches might be

- Utilising manufacturer's CDFs (directly from Affymetrix/Illumina or through Ensembl Biomart)
- In-house sequence-based translation, e.g. cross-platform co-alignment of the probe sequences against the current version of genome.

The first alternative was considered less accurate than the custom-CDF approach. The second based procedure was considered too laborious and partially duplicating the work being done by the groups maintaining custom-CDF files.

The accurate signature translation and rigorous trimming may remove a large portion of the genes originally included in the signature (Tables 5, 6, 8). This may complicate

Table 8: Trimming signatures during pre-processing

Signature	Original size	Trimmed translated signatures						
		Edinburgh RS	Edinburgh L23	Tam U133A	TamU133 Plus2	GSE17705	GSE16391	GSE4922
BCatenin	98	54	58	54	98	54	98	54
E2F3	298	137	112	173	298	173	298	173
ESR1	165	115	133	165	165	165	165	165
Hypoxia	58	38	34	58	58	58	58	58
Invasiveness	100	81	66	100	100	100	100	100
MYC	248	113	96	154	229	154	229	154
PIK3A	278	166	184	278	278	278	278	278
RAS	348	133	123	228	348	228	348	228
SRC	73	44	38	46	73	46	73	46
Stemness	100	59	54	100	100	100	100	100

applying such procedure to short signatures. However, most of the reviewed multi-gene signatures (including all signatures selected for this project) comprised of many tens or hundreds of genes, allowing for significant size reduction. For these large signatures, it was considered that it is better to shorten the signature, then to include an irrelevant gene.

3.4 Classification procedure

Design and testing of the Iterative Consensus PAM classification algorithm is the key computational element of the project. It is based on the Partitioning Around Medoids (PAM).

PAM was suggested by Kaufman and Rousseeuw in 1990 [194]. This partitioning algorithm has several useful properties [167], which make it a very attractive choice for our classification procedure:

1) In an un-supervised manner it selects elements within the partitioned set, which are the best centres (medoids) for the given number of partitions;
2) It provides lists of elements agglomerated around each medoid;
3) It was originally designed for use with any distance measure (in contrast to K-mean partitioning, that is native to Euclidian distance);
4) It is robust and it is good at recognising small clusters, when there is true agglomeration between the elements.

Recognising of medoids is particularly useful, when there is a need to distinguish whether a particular element is the agglomeration centre and an authentic member of the cluster, or if this element is just a spurious peripheral member of this cluster.

Having an element representative to a class of interest (e.g. a centroid representative for high activity of a pathway) iterative PAM may be used to look for other elements in the dataset, which co-cluster with this centroid. Indeed, repeating PAM with incremental number of partitions, it may happen that at a certain number of partitions PAM produces a cluster agglomerated around the centroid. Interestingly, if the element is not authentic to any of the sub-classes, then it will be included to partitions as a non-medoid member or placed in a separate class of its own (with a sufficiently high number of partitions).

It is well known, that clustering results may strongly depend on the choice of distance metric. This is why we perform classification with three different metrics (Euclidian, Manhattan and Spearman correlation distances), taking forward only the consensus results. To further minimise possibility of spurious agglomerations, we construct two opposite centroids (Figure 17) and set the initial number of partitions to three.

While testing and tuning the algorithm it was found that absolute values in centroids should be adjusted to the typical amplitudes observed in the dataset. I.e. if the median absolute value in the dataset was 0.3, then the up-regulated genes in the centroid should be given the value +0.3, and the down-regulated genes should be given the values of -0.3. The initial classifications based on correlation distance were very sensitive to the differences between pseudo-samples and the rest of the dataset. To obtain meaningful classifications using correlation distance, we had to relax criteria for class allocation during the PAM's iterations: the iterations stopped as soon as the pseudo-samples were placed in different clusters (i.e. they did not have to be the medoids of their clusters) when using correlation distances.

The algorithm has been tested using the datasets and signatures selected for the project. The results are shown in the supplementary figures (Appendix A3). Table 8 shows the numbers of partitions that were needed to achieve the classification. In 90% of the cases the desired classifications were achieved within the first 7 iterations (=10 partitions). In 85% of cases the initial iteration with 3 clusters produced the result, confirming the robustness of PAM and suitability of the centroids for partitioning. Classification success rate was different in different signatures. Signatures used for PIK3A, ESR1 and Hypoxia were successful in 100% of cases; the lowest success rate was in the "Stemness" signature: 67%.

It was also noticeable, that Euclidian distance had low success rate in Tam-U133-Plus2 dataset, which was the smallest dataset in the project. In terms of classification calls, the high number of iterations indicates to a low number of classified cases. Thus, only 2 cases were classified by E2F3 signature in the Tam-U133-Plus-2 dataset, and no cases were called either "positive" or "negative" by MYC, RAS and "Stemness" signatures (see heatmap in Appendix A.3.4).

Table 9: Numbers of PAM partitions in tested classifications

Signature	RS (55)			L23 (27)			TamU133A (49)			TamU133Plus (15)			GSE17705 (36)			GSE16391 (30)			GSE4922 (23)		
	m	e	c	m	e	c	m	e	c	m	e	c	m	e	c	m	e	c	m	e	c
BCatenin	3	3	3	3	3	3	41	5	3	3	3	3	15	30	3	3	3	3	3	7	3
E2F3	3	3	3	3	3	3	3	3	3	3	14	3	37	3	3	3	22	3	3	3	3
ESR1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Hypoxia	3	3	3	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Invasiveness	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	25	6	3	3	3	3
MYC	3	3	3	3	3	3	3	4	3	3	16	3	3	3	3	3	3	3	3	3	3
PIK3A	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
RAS	3	3	3	3	3	3	3	3	3	3	16	3	9	5	3	3	3	3	3	3	3
SRC	3	3	3	25	3	3	3	3	3	3	3	3	3	3	3	3	4	3	3	3	3
Stemness	49	54	3	3	24	3	28	3	3	3	15	3	21	31	3	31	31	3	3	5	3

Yellow highlights classifications with 4 to 10 partitions; red highlights classifications with more than 10 partitions

Numbers in brackets show total number of elements in the set (before adding centroids)

m: Manhattan distance, e: Euclidian distance, c: Correlation distance

The initial intent at the beginning of the project was to evaluate several classification approaches, such as SVM, ANN, LDA and logistic regression. Unfortunately, many of the available breast cancer signatures are not accompanied by the original training datasets. This excluded the possibility of using most of these approaches because they rely on a training set for building the algorithm. Recognising this challenge, a number of authors have suggested signature-specific algorithms based on similarity of samples to the signature- based centroids [184,185,198-200]. In general, this approach worked well. However, most of these published algorithms were tuned for a specific signature(s) and micro-array platform [175]; some of the published algorithms were quite complicated and relatively non-transparent [124]. Exploiting the useful features of PAM together in combination with rigorous data and signature pre-processing allowed us to suggest a new algorithm, overcoming these shortcomings.

Further development and testing of our algorithm may include:

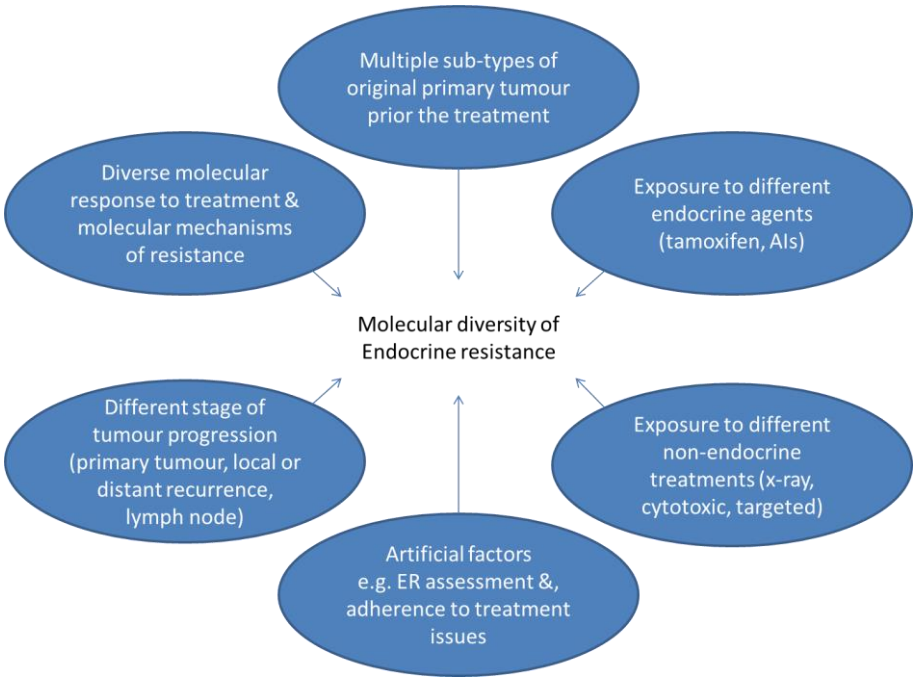
- Comparison with other published algorithms for selected datasets and signatures;
- Assessment of the algorithm's robustness using bootstrapping-like strategies;
- Evaluation the effect of alternative procedures that can be used at data and signature pre-processing stages

Ultimate correctness of the classifications can be checked only experimentally. The result of such testing will depend not only on the algorithm, but equally on the quality of the signature.

3.5 Molecular diversity of endocrine resistance

The developed classification pipeline was tested on the datasets and signatures selected to explore molecular diversity of endocrine resistant breast cancers. An important advantage of the tested classification algorithm is that it is designed for simultaneous profiling by multiple signatures. This opens a possibility for assessment of interactions between the pathways. For instance, it may be noted that in the Edinburgh RS dataset only two tumours show evidence of simultaneous activation of MYC and PIK3A (Appendix A.3.1). In contrast, co-activation of PIK3A and oestrogen receptor

Figure 19: Sources of molecular diversity in endocrine resistant breast cancers



signalling is much more common (10 tumours). It may also be noted that high ESR1 signalling rarely coincides with high invasiveness in this dataset. At the same time, specific biological interpretation of the obtained classifications is still challenging and should take into account all available additional information about endocrine resistance in breast cancers.

The molecular profiles of endocrine resistant breast cancers are very diverse. When interpreting this diversity it is important to consider that this diversity is caused by the interplay of many different factors. The main sources of molecular diversity in endocrine resistant breast cancers are illustrated in Figure 19. Phenotype of each individual endocrine-resistant tumour is shaped by a combination of multiple factors, including

- Molecular mechanism of resistance
- Molecular response to applied treatments (including all endocrine and non-endocrine interventions)

- The sub-type of original primary tumour before the treatment
- The current stage of tumour progression (primary tumour, local recurrence, lymph node or distant metastatic lesion)

To decipher the combinations of these factors there is a need in bioinformatics tools that can simultaneously apply different molecular signatures to the same dataset, which was the main goal of the current project.

4 Conclusions and Further Directions

4.1 Conclusions

The aims and objectives of the project have been fully achieved. A bioinformatics pipeline has been designed and used to classify endocrine-resistant samples from publicly available transcriptomic datasets by multiple multi-gene signatures associated with mechanisms of endocrine resistance.

The pipeline includes (i) procedures and criteria for selection of relevant datasets and signatures, (ii) procedures for rigorous data and signature pre-processing allowing cross-platform analysis and (iii) a classification algorithm specifically tailored to the signatures and datasets typically published in studies in endocrine resistance of breast cancer.

Rigorous manual review of a large number of published datasets and signatures allowed formulating of typical requirements for the down-stream data analysis and selection of datasets and signatures that were used in this project.

The procedures for data and signature pre-processing were developed to enable the pipeline to be used in a cross-platform context, including Illumina and Affymetrix microarrays. The Illumina/Affymetrix signature translation and data merging are based on custom description files developed and maintained for these microarray platforms by the academic community. This assures that only probes with reliable annotations are used for the cross-platform translation or integration. Addition of other platforms depends on reliable chip description files for probe translation.

The classification algorithm (Iterative Consensus PAM) is transparent and has an intuitive mechanism of class allocation. The algorithm does not require a training set. Instead, the class assignment is based on binary centroids generated from the signatures informing only on up- and down- regulation of the genes. The class allocation is based on un-supervised partitioning: only samples that are agglomerated with the centroids in the unsupervised manner are included in the classes. It does not force samples to choose between "high" or "low" clusters: samples that are not naturally agglomerated with the centroids are left in the "non-classifiable" area (up to all the samples in the dataset, if no sample is found similar to a centroid). Finally, the classification is distance-metric independent because is based on a consensus between different distance metrics. The classification algorithm was implemented in a series of R-scripts, using specialised R and Bioconductor packages for cluster analysis and for micro-array data analysis.

The pipeline has been successfully applied for classification of 7 publicly available datasets using 9 multi-gene signatures. This allowed tuning of the algorithm and analysis of its performance with real data. The obtained results have been presented in a dedicated web site.

4.2 Further directions

The further development of the project may go along several directions.

First, it may include further evaluation and development of the Iterative Consensus PAM classification algorithm. It may be interesting to assess the effect of alternative pre-processing steps on the results of classification (e.g. the effect of different modalities of background subtraction, summation and normalisation). It may also be interesting to compare the results of Iterative Consensus PAM with the results of other published algorithms, tuned for specific datasets and signatures. It may be possible to envisage an experimental validation of the Iterative Consensus PAM; however, this would depend on performance of both: the algorithm and the signatures employed for such validation.

Another direction of development for this project may include the development of a web-repository collecting endocrine resistant datasets, relevant signatures and tools for

their analyses. In view of the present lack of standards for reporting breast cancer datasets and signatures, this direction would include manual review of current updates to the GEO and Array Express and manual review of new publications reporting multi-gene signatures relevant to endocrine resistance. In the first instance, if a new dataset or signature is found then it may be added to those, already calculated and presented on the web site accompanying this project. A forum/feed-back section may be added to the site. If the site attracts attention, then it may be used to discuss/suggest the lacking standards for reporting breast cancer clinical annotations and multi-gene signatures.

Finally, the developed pipeline will be used locally, independently of the web-hub development, for analysis of the transcriptomics datasets generated in the Edinburgh Breast Unit. Taken together with the other directions of experimental and bioinformatics analysis employed in the Edinburgh Breast Unit, the developed pipeline will be used to identify pathways activated in individual endocrine resistant breast cancers. This may help to understand the resistance mechanisms and to develop individual biomarkers for tailored treatment of breast cancer patients.

Appendices

A.1 Examples of R code

A.1.1 Log-transformation of pre-processed data exported from Illumina Genome Studio

```
# File: LumiLogTransform.R
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Log-transformation of signals pre-processed by Illumina GenomeStudio.
# Positive values are log transformed, negative signals are substituted by 0.

# Function takes a numeric matrix and returns the log-transfrormed matrix
LumiLogTransform <- function(mx){

  # Get nums of rows and cols
  NumOfCols <- ncol(mx)
  NumOfRows <- nrow(mx)

  # Make matrix for output
  mx.log <- matrix(
    rep(-100,NumOfCols*NumOfRows),
    ncol = NumOfCols)

  # For each row
  for (i in 1:NumOfRows){

    # For each row element
    for (j in 1:NumOfCols){

      # Calculate log-transformed value
      x <- mx[i,j]
      if (x<=0) y <- 0
      if (x>0) y <- log(x)

      # Write log-transformed value to the output matrix
      mx.log[i,j] <- y

    } # Next element in the row

  } # Next row

  # Preserve rownames and colnames
  rownames(mx.log) <- rownames(mx)
  colnames(mx.log) <- colnames(mx)

  # Return log-transformed matrix
  return(mx.log)

}
```

A.1.2.Microarray data import from GEO

```
# File: GEO_Import.R
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Microarray data import from NCBI GEO repository

# Set Environment
rm(list=ls())
graphics.off()
setwd("C:/Documents and Settings/Breakthrough/My Documents/Alexey")

# Load GEOquery
source("http://www.bioconductor.org/biocLite.R")
biocLite("GEOquery")
library(GEOquery)

# Download raw data (~11GB in total, leave download overnight)
# Usually raw data include CEL files (for Affy arrays) and
# annotations in arbitrary format, which have to be
# manually processed.
# The data are saved in sub-folders in the working directory.
DataSetsList <- c('GSE9195', 'GSE17705','GSE6532','GSE2990','GSE4922',
                  'GSE22219','GSE16391', 'GSE26971', 'GSE12093')
for (i in DataSetsList) getGEOSuppFiles(i)

# Download processed data (~473Mb in total)
# These data may contain clinical annotations
# required for selection of resistant cases.
# Also these data  include expression values
# pre-processed by authors (not used in further analysis).
Proc_GSE9195 <- getGEO("GSE9195", GSEMatrix = TRUE)
Proc_GSE17705 <- getGEO("GSE17705", GSEMatrix = TRUE)
Proc_GSE6532 <- getGEO("GSE6532", GSEMatrix = TRUE)
Proc_GSE2990 <- getGEO("GSE2990", GSEMatrix = TRUE)
Proc_GSE4922 <- getGEO("GSE4922", GSEMatrix = TRUE)
Proc_GSE22219 <- getGEO("GSE22219", GSEMatrix = TRUE)
Proc_GSE16391 <- getGEO("GSE16391", GSEMatrix = TRUE)
Proc_GSE26971 <- getGEO("GSE26971", GSEMatrix = TRUE)
Proc_GSE12093 <- getGEO("GSE12093", GSEMatrix = TRUE)

# Save imported data in RData format
save.image("ProcessedData.RData")
```

A.1.3.Affymetrix microarry data pre-processing

```
# File: GSE16391_AffyReadPreprocess.R
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Reading and pre-processing of Affy CEL files
# CEL files shall be present in the working folder

# Environment
rm(list=ls())
graphics.off()
win.wd <- paste(
 "C:/Users/Alexey/Documents/Cranfield/Data/", dataset.name, "_CELs", sep ="")
setwd(win.wd)
rm(win.wd)

# Load packages
source("http://bioconductor.org/biocLite.R")
biocLite("affy")
biocLite("affyPLM")
library(affy)
library(affyPLM)

# Name of file containing list of resistant cases (as CEL files)
dataset.name <- "GSE16391"
file.list <- paste(dataset.name,"_Resistant.txt",sep="")

# Read list of CEL files
samples <- read.table(file = file.list, header = TRUE)
samples <- paste(as.vector(samples[,1]), ".cel.gz", sep="")

# Read CEL files into an AffyBatch object
Data.raw <- ReadAffy(filenames=samples, verbose=TRUE)

# Save AffyBatch object
save.image(paste(dataset.name,"_AffyBatch.RData", sep=""))

# Normalise data (get expression set)
Data.norm <- threestep(Data.raw,
  background.method="RMA.2",
  normalize.method="quantile",
  summary.method="median.polish")

# Remove unnecessary data
rm(Data.raw, file.list, mapCdfName, samples)

# Save expression set
save.image(paste(dataset.name, "_eset.RData", sep=""))

# Get expression's matrix
Data <- exprs(Data.norm)

# Remove expression set
rm(Data.norm)

# Save expressions matrix
save.image(paste(dataset.name,"_matrix.RData", sep=""))
```

A.1.4.Main classification routine

This is the main procedure that calls sub-routines as shown on Figure 18

```
# File: GSE16391_Classification
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Classification of a dataset by several available signatures

# Environment
rm(list=ls())
graphics.off()
setwd("C:/Users/Alexey/SkyDrive/Documents/Bix/GEO")

# Load normalised and log-transformed expression matrix
data.name <- "GSE16391"
load(paste(data.name,"_matrix.RData",sep=""))

# Read signatures, trim signatures to dataset, trim dataset to signatures
sign.names <- c("BCatenin", "E2F3", "ESR1", "Hypoxia",
  "Invasiveness", "MYC", "PIK3A", "RAS","SRC", "Stemness")
source("PrepareAffyData.R")
rm(Data)

# Center and scale subsets
source("CenterScale.R")
data.trim.cs <- data.trim
for (i in sign.names) data.trim.cs[[i]] <- CenterScale(data.trim[[i]])
rm(i, data.trim, CenterScale)

# --- Classify dataset by the signatures --- #

# Make table for all-signatures summary
Classes.all <- matrix(colnames(data.trim.cs[[1]]), ncol=1)
Iterations.all <- matrix(c("Sgn","m","e","c"), ncol=4)

# Classify by one signature at a time
source("ClassifyAffy.R")
source("IterativeConsensusPAM.R")
require(cluster)

for (i in sign.names) {

  # Get classification for the signature
  Cur.result <- ClassifyAffy(data.trim.cs[[i]], sign.trim[[i]])

  # Add classes to the summary table
  Cur.class <- Cur.result$Classes
  Classes.all <- cbind(Classes.all, Cur.class)

  # Save numbers of iterations
  Cur.iterations <- Cur.result$Iterations
  Iterations.all <- rbind(Iterations.all, c(i, Cur.iterations))
}

# Add column names to results
colnames(Classes.all) <- c("Case",sign.names)

# Save classification matrix to file (the main result)
file.name <- paste(data.name,"_Classes.txt", sep="")
write.table(Classes.all, file = file.name, row.names = FALSE, sep = "\t")
```

```
# Save iteration numbers to file (for algorithm assessmnet)
file.name <- paste(data.name,"_Iterations.txt", sep="")
write.table(Iterations.all, file = file.name,
  row.names = FALSE, col.names = FALSE, sep = "\t")

# Save sizes of trimmed signatures to file (for algorithm assessment)
file.name <- paste(data.name,"_Trimming.txt", sep="")
write.table(Sign.trimming, file = file.name,
  row.names = FALSE, sep = "\t")
```

A.1.4.1. PrepareAffyData.R

This is a fragment of sub-routine called from the main procedure as shown on Figure 18

```
# File: PrepareAffyData
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Read signatures, trim signatures to dataset, trim dataset to signatures

# Environment
library(gdata) # for trim()

# BCatenin signature
BCatenin.sgn.full <- read.table(
  file = "Bcatenin_Signature.txt", header = TRUE, sep="\t")
BCatenin.sgn.full[,1] <- trim(BCatenin.sgn.full[,1])
probe.sets <- as.vector(BCatenin.sgn.full[,1])
available.probe.sets <- probe.sets %in% rownames(Data)
BCatenin.data.trim <- Data[probe.sets[available.probe.sets],]
BCatenin.sgn.trim <- BCatenin.sgn.full[available.probe.sets,2]
names(BCatenin.sgn.trim) <- BCatenin.sgn.full[available.probe.sets,1]
sum(names(BCatenin.sgn.trim) != rownames(BCatenin.data.trim))

# BCatenin signature
BCatenin.sgn.full <- read.table(
  file = "Bcatenin_Signature.txt", header = TRUE, sep="\t")
BCatenin.sgn.full[,1] <- trim(BCatenin.sgn.full[,1])
probe.sets <- as.vector(BCatenin.sgn.full[,1])
available.probe.sets <- probe.sets %in% rownames(Data)
BCatenin.data.trim <- Data[probe.sets[available.probe.sets],]
BCatenin.sgn.trim <- BCatenin.sgn.full[available.probe.sets,2]
names(BCatenin.sgn.trim) <- BCatenin.sgn.full[available.probe.sets,1]

# To avoid manipulations with complicated and error-prone
# multy-layer nested objects each signature and dataset were
# processed separately without cycles.

# E2F3 signature
...
# ESR1 signature
...
# Hypoxia signature
...
# Invasiveness signature
...
# MYC signature
...
# PIK3A signature
...
```

```
# RAS signature
...
# SRC signature
...
# Stemness signature
...

# Make list of trimmed data for further use
data.trim <- list(BCatenin.data.trim, E2F3.data.trim,
  ESR1.data.trim, Hypoxia.data.trim, Invasiveness.data.trim,
  MYC.data.trim, PIK3A.data.trim, RAS.data.trim, SRC.data.trim,
  Stemness.data.trim)
names(data.trim) <- sign.names

# Make list of trimmed signatures for further use
sign.trim <- list(BCatenin.sgn.trim, E2F3.sgn.trim,
  ESR1.sgn.trim, Hypoxia.sgn.trim, Invasiveness.sgn.trim,
  MYC.sgn.trim, PIK3A.sgn.trim, RAS.sgn.trim, SRC.sgn.trim,
  Stemness.sgn.trim)
names(sign.trim) <- sign.names

# Save sizes of signatures
full.sgn.sises <- as.numeric(lapply(list(BCatenin.sgn.full,
  E2F3.sgn.full, ESR1.sgn.full, Hypoxia.sgn.full, Invasiveness.sgn.full,
  MYC.sgn.full, PIK3A.sgn.full, RAS.sgn.full, SRC.sgn.full,
  Stemness.sgn.full), nrow))

trim.sgn.sises <- as.numeric(lapply(list(BCatenin.sgn.trim,
  E2F3.sgn.trim, ESR1.sgn.trim, Hypoxia.sgn.trim, Invasiveness.sgn.trim,
  MYC.sgn.trim, PIK3A.sgn.trim, RAS.sgn.trim, SRC.sgn.trim,
  Stemness.sgn.trim), length))

Sign.trimming <- cbind(sign.names,full.sgn.sises,trim.sgn.sises)

colnames(Sign.trimming) <- c("Sign", "Full", "Trim")

# Remove unnecessary objects
rm(BCatenin.sgn.full, E2F3.sgn.full,
   ESR1.sgn.full, Hypoxia.sgn.full, Invasiveness.sgn.full,
   MYC.sgn.full, PIK3A.sgn.full, RAS.sgn.full, SRC.sgn.full,
   Stemness.sgn.full)

rm(BCatenin.sgn.trim, E2F3.sgn.trim,
   ESR1.sgn.trim, Hypoxia.sgn.trim, Invasiveness.sgn.trim,
   MYC.sgn.trim, PIK3A.sgn.trim, RAS.sgn.trim, SRC.sgn.trim,
   Stemness.sgn.trim)

rm(BCatenin.data.trim, E2F3.data.trim,
   ESR1.data.trim, Hypoxia.data.trim, Invasiveness.data.trim,
   MYC.data.trim, PIK3A.data.trim, RAS.data.trim, SRC.data.trim,
   Stemness.data.trim)

rm(probe.sets, available.probe.sets)
```

A.1.4.2. CenterScale.R

This is a sub-routine called from the main procedure as shown on Figure 18

```
# File: CenterScale.R
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Function takes a numeric matrix and returns matrix
# median-centered and scaled [-1 to 1] by rows
CenterScale <- function(mx){

  # Get nums of rows and cols
  NumOfCols <- ncol(mx)
  NumOfRows <- nrow(mx)

  # --- Median centering --- #
  rows.medians <- apply(mx, 1, median)
  mx.medians <- matrix(
    rep(rows.medians,NumOfCols),
    byrow = FALSE,
    ncol=NumOfCols)
  mx.centered <- mx - mx.medians

  # --- Scaling --- #

  # Make matrix for output
  mx.centered.scaled <- matrix(
    rep(-100,NumOfCols*NumOfRows),
    ncol = NumOfCols)

  # For each row
  for (i in 1:NumOfRows){

    # Get max and min elements
    RowMin = min(mx.centered[1,])
    RowMax = max(mx.centered[1,])

    # For each row element
    for (j in 1:NumOfCols){

      # Calculate scaled value
      x <- mx.centered[i,j]
      if (x<0) y <- -x/RowMin
      if (x>0) y <- x/RowMax
      if (x==0) y <- 0

      # Write scaled value to the output matrix
      mx.centered.scaled[i,j] <- y

    } # Next element in the row

  } # Next row

  # Preserve rownames and colnames
  rownames(mx.centered.scaled) <- rownames(mx)
  colnames(mx.centered.scaled) <- colnames(mx)

  # Return centered scaled matrix
  return(mx.centered.scaled)

}
```

A.1.4.3. ClassifyAffy.R

This is a sub-routine called from the main procedure as shown on Figure 18

```
# File: ClassifyAffy
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Function classifies a single matrix by a single signature
# returns table with classes and vector with nums of PAM iterations

ClassifyAffy <- function (dat, sgn) {

  # --- Prepare PseudoSamples (centroids) for classification --- #

  # Get median absolute amplitude of data
  Med <- median(abs(dat))

  # Generate PseudoSamples with high and low activity of the pathway
  c.high <- sgn * median(abs(dat))
  c.low <- c.high * -1

  # Perform classification by iterative PAM
  # source("IterativeConsensusPAM.R") # has been sourced earlier
  icp <- IterativeConsensusPAM(dat, c.high, c.low)

  # Prepare results for output
  Result <- list(icp$Classes, icp$Iterations)
  names(Result) <- c("Classes", "Iterations")

  # Return classification vector and numbers of iterations
  return(Result)
}
```

A.1.4.4. IterativeConsensusPAM.R

This is a sub-routine called from ClassifyAffy procedure as shown on Figure 18

```
# File: IterativeConsensusPAM.R
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Calssification by similarity to PseudoSamples
# - using iterative PAM with incremental number of clusters
# - using consensus of 3 distance measures

# Load library
# require(cluster) # loaded earlier in a calling procedure

# Function takes a matrix and two PseudoSamples (centroids).
# Function returns classification vector with three classes:
# two classes with samples agglomerated around each centroid
# and a third class with non-classifiable samples that do not
# agglomerate to either of centroids.
# In addition the function returns numbers of iterations
# that can be used for for algorithm performance analysis
# and tuning.
IterativeConsensusPAM <- function(pam.data, ps1, ps2){

  # Add PseudoSamples (centroids) to data
  pam.data <- cbind(ps1, ps2, pam.data)
```

```
# ------------------------------------------------#
#                  Classify cases                 #
# ------------------------------------------------#

# --- Manhattan distance --- #

# Transpose data to use in pam()
m.data <- t(pam.data)

# Get max num of clusters
max.clst.num.m <- nrow(m.data) - 1

# Increment number of clusters
for (m.i in 3:max.clst.num.m) {

  # Partition
  m.pt <- pam(m.data, m.i, metric = "manhattan")

  # Stop if an acceptable classification is acheaved
  # i.e. there is a cluster around each of the PseudoSamples
  if ("ps1" %in% rownames(m.pt$medoids)  &&
    "ps2" %in% rownames(m.pt$medoids)) break
}

# --- Euclidean distance --- #

# Transpose data to use in pam()
e.data <- t(pam.data)

# Get max num of clusters
max.clst.num.e <- nrow(e.data) - 1

# Increment number of clusters
for (e.i in 3:max.clst.num.e) {

  # Partition
  e.pt <- pam(e.data, e.i, metric = "euclidean")

  # Stop if an acceptable classification is acheaved
  # i.e. there is a cluster around each of the PseudoSamples
  if ("ps1" %in% rownames(e.pt$medoids)  &&
    "ps2" %in% rownames(e.pt$medoids)) break
}

# --- Correlation distance --- #

# Calculate distance matrix as recommented in ?dist
cor.dist.mx <- as.dist((1-cor(pam.data, method = "spearman"))/2)

# Get max num of clusters
max.clst.num.c <- ncol(pam.data) - 1

# Increment number of clusters
for (c.i in 3:max.clst.num.c) {

  # Partition
  c.pt <- pam(cor.dist.mx, c.i)

  # Stop if an acceptable classification is acheaved
  # i.e. the PseudoSamples are placed in different clusters
  if (c.pt$clustering["ps1"] != c.pt$clustering["ps2"]) break
}
```

```
# ------------------------------------------------- #
# Collapse pam() classification vectors to 3 classes #
# ------------------------------------------------- #

# --- Manhattan classification --- #

# Code 0 for samples not agglomerated with either ps1 or ps2
m.3classes <- rep(0, length(m.pt$clustering))

# Preserve the samples names
names(m.3classes) <- names(m.pt$clustering)

# Get codes of classes agglomerated around PseudoSamples
m.ps1.class <- m.pt$clustering["ps1"]
m.ps2.class <- m.pt$clustering["ps2"]

# Translate pam() classification vector 3-class vector
for (i in 1:length(m.pt$clustering)){

  # Code 1 for samples agglomerated with ps1
  if (m.pt$clustering[i] == m.ps1.class) m.3classes[i] <- 1

  # Code -1 for samples agglomerated with ps2
  if (m.pt$clustering[i] == m.ps2.class) m.3classes[i] <- -1
}

# --- Euclidean classification --- #

# Code 0 for samples not agglomerated with either ps1 or ps2
e.3classes <- rep(0, length(e.pt$clustering))

# Preserve the samples names
names(e.3classes) <- names(e.pt$clustering)

# Get codes of classes agglomerated around PseudoSamples
e.ps1.class <- e.pt$clustering["ps1"]
e.ps2.class <- e.pt$clustering["ps2"]

# Translate pam() classification vector 3-class vector
for (i in 1:length(e.pt$clustering)){

  # Code 1 for samples agglomerated with ps1
  if (e.pt$clustering[i] == e.ps1.class) e.3classes[i] <- 1

  # Code -1 for samples agglomerated with ps2
  if (e.pt$clustering[i] == e.ps2.class) e.3classes[i] <- -1
}

# --- Correlation classification --- #

# Code 0 for samples not agglomerated with either ps1 or ps2
c.3classes <- rep(0, length(c.pt$clustering))

# Preserve the samples names
names(c.3classes) <- names(c.pt$clustering)

# Get codes of classes agglomerated around PseudoSamples
c.ps1.class <- c.pt$clustering["ps1"]
c.ps2.class <- c.pt$clustering["ps2"]

# Translate pam() classification to 3-class vector
for (i in 1:length(c.pt$clustering)){
```

```
    # Code 1 for samples agglomerated with ps1
    if (c.pt$clustering[i] == c.ps1.class) c.3classes[i] <- 1

    # Code -1 for samples agglomerated with ps2
    if (c.pt$clustering[i] == c.ps2.class) c.3classes[i] <- -1
  }

  # ------------------------------------------ #
  #     Generate consensus classification      #
  # ------------------------------------------ #

  # Code 0 for samples not agglomerated with either ps1 or ps2
  Result.Classes <- rep(0, length(e.3classes))

  # Preserve the samples names
  names(Result.Classes) <- names(e.3classes)

  # For each sample in classifications
  for (i in 1:length(e.3classes)){

    # Write consensus codes to result
    if (e.3classes[i] == m.3classes[i] &&
      e.3classes[i] == c.3classes[i])
      Result.Classes[i] <- e.3classes[i]
  }

  # Remove PseudoSamples from the classification vector
  Result.Classes <- Result.Classes[c(-1,-2)]

  # Save number of iterations
  Result.Iterations <- c(m.i, e.i, c.i)

  # Generate the result
  Result <- list(Result.Classes, Result.Iterations)
  names(Result) <- c("Classes", "Iterations")

  # Return classification vector and numbers of iterations
  return(Result)

}
```

A.1.5.Heatmap generation

```
# File: GEO_Heatmap.R
# Alexey Larionov, code for personal project, 2012
# Cranfield MSc course in Applied bioinformatics

# Draw heatmap for a classification matrix

# Environment
rm(list=ls())
graphics.off()

# Names and files
data.name = "TamU133Plus2"
data.file <- paste(data.name, "_Classes.txt", sep= "" )

# Read data for heatmap
Data <- read.table(file = data.file, row.names = 1,
  header = TRUE, sep = "\t")

# Select signatures data for heatmap
Data <- t(Data)
colnames(Data) <- c(1:ncol(Data))

# Typical reverse (!) order of signatures on heatmaps:
mx <- as.matrix(Data[c("Stemness", "Invasiveness", "Hypoxia",
  "BCatenin", "RAS", "E2F3", "MYC", "PIK3A", "ESR1"),])

# It is possible to select only desired pathways
#mx <- as.matrix(Data[c("Invasiveness", "Hypoxia",
#  "RAS", "E2F3", "MYC", "PIK3A", "ESR1"),])

# Make colour palette for heatmap
require(gplots) # for colorpanel() function
cols <- colorpanel(3, "lightblue", "grey95", "red")

# Direct output to file
picture.file <- paste(data.name,"_1.jpg")
jpeg(filename=picture.file, , width = 6000, height = 6000, res=300)

# Draw heatmap
heatmap(mx, Rowv=NA, Colv=NULL, scale="none", col=cols,
        cexRow = 3, cexCol = 2)

# Stop drawing to file
dev.off()

# Direct output to file
picture.file <- paste(data.name,"_2.jpg")
jpeg(filename=picture.file, width = 6000, height = 6000, res=300)

# Draw heatmap
heatmap(mx, Rowv=NA, Colv=NULL, scale="none", col=cols, asp=1,
        cexRow = 3, cexCol = 2)

# Stop drawing to file
dev.off()
```

A.2. Signatures composition

A.2.1ESR1

209460_at	1	205862_at	1	205074_at	1	202147_s_at	-1
205355_at	1	201413_at	1	202088_at	1	206734_at	-1
213245_at	1	203628_at	1	205597_at	1	217938_s_at	-1
204497_at	1	204863_s_at	1	202752_x_at	1	204401_at	-1
209173_at	1	204686_at	1	216092_s_at	1	220239_at	-1
211712_s_at	1	203710_at	1	212956_at	1	205569_at	-1
212985_at	1	212496_s_at	1	204045_at	1	201795_at	-1
40148_at	1	217894_at	1	202371_at	1	213564_x_at	-1
202641_at	1	203144_s_at	1	205009_at	1	209205_s_at	-1
40093_at	1	212441_at	1	204623_at	1	212274_at	-1
201170_s_at	1	221874_at	1	212770_at	1	218684_at	-1
211939_x_at	1	213234_at	1	200804_at	1	206571_s_at	-1
203571_s_at	1	212442_s_at	1	203476_at	1	203636_at	-1
221823_at	1	212692_s_at	1	217979_at	1	201976_s_at	-1
218195_at	1	211596_s_at	1	210652_s_at	1	203315_at	-1
220581_at	1	208682_s_at	1	221765_at	1	203574_at	-1
203963_at	1	203929_s_at	1	218806_s_at	1	218051_s_at	-1
204811_s_at	1	209623_at	1	212637_s_at	1	200790_at	-1
41660_at	1	214077_x_at	1	200670_at	1	209791_at	-1
200810_s_at	1	218259_at	1	219741_x_at	1	201037_at	-1
219414_at	1	218211_s_at	1	215304_at	1	201397_at	-1
201754_at	1	219648_at	1	222275_at	1	218236_s_at	-1
205081_at	1	204798_at	1	213532_at	-1	204061_at	-1
219913_s_at	1	214440_at	1	209122_at	-1	204304_s_at	-1
202263_at	1	204862_s_at	1	205109_s_at	-1	200039_s_at	-1
206754_s_at	1	206197_at	1	202207_at	-1	212265_at	-1
210272_at	1	202599_s_at	1	219497_s_at	-1	213923_at	-1
205471_s_at	1	222125_s_at	1	205548_s_at	-1	221872_at	-1
218094_s_at	1	212148_at	1	219806_s_at	-1	218497_s_at	-1
218976_at	1	217770_at	1	203256_at	-1	213113_s_at	-1
205066_s_at	1	208615_s_at	1	221676_s_at	-1	210959_s_at	-1
214053_at	1	214552_s_at	1	203139_at	-1	202200_s_at	-1
217838_s_at	1	203749_s_at	1	204750_s_at	-1	202951_at	-1
218532_s_at	1	208873_s_at	1	203693_s_at	-1	221016_s_at	-1
213304_at	1	212099_at	1	201231_s_at	-1	211967_at	-1
209696_at	1	218394_at	1	212371_at	-1	202342_s_at	-1
204667_at	1	201826_s_at	1	212771_at	-1	202504_at	-1
44654_at	1	203071_at	1	213260_at	-1	208627_s_at	-1
205354_at	1	35666_at	1	221510_s_at	-1	221203_s_at	-1
209603_at	1	209443_at	1	213170_at	-1		
205696_s_at	1	200718_s_at	1	200824_at	-1		
218692_at	1	209681_at	1	206074_s_at	-1		

A.2.2.PIK3A

212415_at	1	201117_s_at	-1	206110_at	-1	214440_at	1	211578_s_at	-1
213353_at	-1	209522_s_at	1	218280_x_at	-1	202149_at	1	221523_s_at	-1
218322_s_at	1	201161_s_at	-1	214290_s_at	-1	217963_s_at	-1	208456_s_at	-1
205013_s_at	-1	206994_at	1	202708_s_at	-1	209706_at	1	212590_at	-1
212543_at	1	210835_s_at	1	221582_at	-1	212377_s_at	1	41644_at	1
204348_s_at	-1	209617_s_at	-1	208729_x_at	1	213462_at	1	213236_at	1
203608_at	-1	205472_s_at	1	217478_s_at	1	39549_at	1	206799_at	1
204446_s_at	1	205471_s_at	1	215536_at	1	204972_at	1	206378_at	1
202630_at	-1	212690_at	-1	204607_at	1	222075_s_at	1	212589_at	-1
211621_at	1	209916_at	1	213793_s_at	-1	218730_s_at	-1	212154_at	-1
211110_s_at	1	218277_s_at	-1	211548_s_at	-1	211212_s_at	-1	202376_at	1
201288_at	1	200606_at	1	203914_x_at	-1	206825_at	-1	204688_at	-1
202986_at	1	205741_s_at	-1	203913_s_at	-1	202336_s_at	1	222258_s_at	1
214553_s_at	-1	221586_s_at	-1	205543_at	-1	214130_s_at	1	213308_at	1
209788_s_at	1	219974_x_at	-1	213418_at	1	214129_at	1	221041_s_at	1
205047_s_at	-1	219850_s_at	1	202638_s_at	1	219630_at	1	218788_s_at	1
201242_s_at	1	205321_at	-1	213931_at	1	209242_at	-1	208078_s_at	-1
201171_at	-1	201340_s_at	1	209292_at	-1	217744_s_at	1	205573_s_at	-1
211944_at	1	204160_s_at	1	211406_at	-1	205361_s_at	-1	203509_at	1
203685_at	-1	206191_at	1	202421_at	1	210976_s_at	-1	219109_at	-1
202357_s_at	1	205757_at	1	203474_at	1	204992_s_at	-1	203128_at	1
210538_s_at	1	206070_s_at	-1	209185_s_at	-1	221521_s_at	-1	213562_s_at	-1
218732_at	-1	202017_at	1	209184_s_at	-1	205078_at	-1	208920_at	-1
201641_at	1	205225_at	1	204017_at	1	212240_s_at	1	202286_s_at	1
38241_at	1	217838_s_at	1	214295_at	1	212249_at	1	204654_s_at	1
218597_s_at	-1	220147_s_at	-1	213478_at	1	202743_at	1	203888_at	1
217966_s_at	1	208229_at	-1	212325_at	1	208502_s_at	1	203887_s_at	1
217967_s_at	1	218514_at	-1	212327_at	1	203649_s_at	1	203221_at	1
212875_s_at	1	220145_at	-1	212328_at	1	201860_s_at	1	204137_at	1
205248_at	1	202709_at	1	202962_at	1	203895_at	-1	220177_s_at	1
204365_s_at	1	208006_at	1	221841_s_at	1	204939_s_at	-1	214329_x_at	1
204364_s_at	1	218084_x_at	1	209016_s_at	1	204940_at	-1	202687_s_at	1
214428_x_at	1	203987_at	-1	201030_x_at	-1	202620_s_at	-1	202688_at	1
208451_s_at	1	205278_at	1	213564_x_at	-1	202619_s_at	-1	213109_at	1
218541_s_at	-1	212256_at	1	221558_s_at	1	210139_s_at	1	213107_at	1
205308_at	-1	217787_s_at	1	208949_s_at	1	212841_s_at	1	211828_s_at	1
220414_at	1	218313_s_at	1	210732_s_at	1	204566_at	-1	214774_x_at	1
212551_at	1	205280_at	-1	203236_s_at	1	218273_s_at	1	210372_s_at	1
208683_at	1	205279_s_at	-1	221194_s_at	-1	207291_at	1	203786_s_at	1
220066_at	1	206662_at	1	214791_at	1	203354_s_at	1	204352_at	1
211366_x_at	1	209276_s_at	1	216250_s_at	1	218613_at	1	202342_s_at	-1
209970_x_at	1	204875_s_at	1	219759_at	1	203355_s_at	1	210389_x_at	-1
211368_s_at	1	220108_at	1	202018_s_at	1	205961_s_at	-1	221326_s_at	-1
206011_at	1	217771_at	1	205668_at	1	202353_s_at	-1	216609_at	-1
205379_at	1	204983_s_at	1	36711_at	-1	221666_s_at	1	208997_s_at	1
212816_s_at	-1	204984_at	1	218918_at	1	201482_at	1	208998_at	1
1405_i_at	1	206204_at	1	203510_at	-1	200607_s_at	-1	204042_at	-1
201946_s_at	-1	210761_s_at	-1	214051_at	-1	209849_s_at	-1	217975_at	-1
209619_at	1	200824_at	1	203414_at	-1	204916_at	-1	205990_s_at	1
218451_at	1	204237_at	-1	203565_s_at	-1	204070_at	1	219312_s_at	-1
201884_at	1	204235_s_at	-1	210319_x_at	1	209488_s_at	1	211965_at	1
211657_at	1	213548_s_at	-1	212859_x_at	-1	205645_at	1	201367_s_at	1
203757_s_at	1	209526_s_at	-1	213693_s_at	1	205879_x_at	1	201369_s_at	1
204637_at	-1	216693_x_at	-1	207847_s_at	1	214519_s_at	-1	202028_s_at	-1
221042_s_at	1	214469_at	-1	202431_s_at	-1	219138_at	-1		
201116_s_at	-1	207156_at	-1	201976_s_at	1	221943_x_at	-1		

A.2.3.MYC

208161_s_at	-1	218779_x_at	-1	227037_at	1	204476_s_at	-1	202384_s_at	1
209641_s_at	-1	226213_at	-1	227485_at	-1	219295_s_at	1	219131_at	1
231907_at	-1	228131_at	-1	218096_at	1	218590_at	1	218605_at	1
234312_s_at	-1	202159_at	1	204682_at	-1	202212_at	1	206008_at	-1
205180_s_at	-1	226799_at	-1	212281_s_a	1	210976_s_at	1	223776_x_at	-1
227530_at	-1	227271_at	-1	212282_at	1	200658_s_at	1	202510_s_at	-1
227529_s_at	-1	226698_at	-1	212279_at	1	40446_at	-1	209118_s_at	-1
209645_s_at	1	218920_at	-1	219278_at	-1	211668_s_at	-1	213326_at	-1
207396_s_at	1	221712_s_at	1	230110_at	1	201373_at	-1	1569003_at	-1
229267_at	1	203867_s_at	1	226211_at	-1	203201_at	1	224917_at	-1
224634_at	1	220353_at	1	226210_s_a	-1	225291_at	1	218512_at	1
47069_at	1	221536_s_at	1	204027_s_a	1	212541_at	1	226938_at	1
209824_s_at	-1	223200_s_at	1	232077_s_a	-1	218273_s_at	-1	201294_s_at	-1
210971_s_at	-1	219987_at	1	224468_s_a	1	209158_s_at	-1	223055_s_at	1
224204_x_at	-1	236635_at	1	224500_s_a	1	203150_at	1	219836_at	-1
208758_at	1	210463_x_at	1	1553715_s_	1	203108_at	-1	222227_at	-1
212135_s_at	-1	203701_s_at	1	227103_s_a	1	212444_at	-1	117_at	1
205410_s_at	-1	203785_s_at	1	221637_s_a	1	222666_s_at	1	244623_at	1
207618_s_at	1	235026_at	1	203119_at	1	218686_s_at	-1	229715_at	1
220688_s_at	1	236745_at	1	204699_s_a	1	213427_at	1	65585_at	1
50314_i_at	1	222333_at	-1	218953_s_a	1	224610_at	1	1562904_s_at	1
211559_s_at	-1	223035_s_at	1	211986_at	-1	204133_at	1	212563_at	1
221520_s_at	-1	225712_at	1	235281_x_at	-1	218481_at	1	234049_at	1
211804_s_at	-1	35436_at	-1	209467_s_at	-1	210365_at	-1	216212_s_at	1
202246_s_at	1	238689_at	-1	205455_at	-1	230333_at	-1	211725_s_at	1
211862_x_at	-1	205014_at	-1	233803_s_at	1	221514_at	1	1556111_s_at	1
218732_at	1	222305_at	1	202431_s_at	1	221513_s_at	1	224603_at	1
223232_s_at	-1	209971_x_at	1	211824_x_at	-1	212268_at	-1	1568597_at	1
230656_s_at	1	1552334_at	-1	211822_s_at	-1	225143_at	1	235474_at	1
224903_at	1	1552767_a_a	1	200610_s_at	1	229236_s_at	1	225933_at	1
233986_s_at	-1	200800_s_a	1	227249_at	-1	219874_at	1	241687_at	1
202310_s_at	-1	213418_at	1	207535_s_at	-1	211576_s_at	1	202632_at	1
203325_s_at	-1	214011_s_a	1	205858_at	-1	209776_s_at	1	235501_at	-1
221900_at	-1	200807_s_a	1	218376_s_at	-1	204717_s_at	1	65521_at	-1
205076_s_at	-1	212411_at	1	202891_at	-1	202219_at	1	233493_at	-1
215537_x_at	-1	218305_at	1	214427_at	1	232481_s_at	-1	179_at	-1
202262_x_at	-1	203882_at	-1	200875_s_at	1	207390_s_at	-1	201278_at	-1
204977_at	1	202138_x_a	1	218199_s_at	1	209427_at	-1	1555673_at	-1
208895_s_at	1	212510_at	1	211951_at	1	212666_at	-1	201042_at	-1
203385_at	-1	1552257_a_	1	205895_s_at	1	201563_at	1	237591_at	-1
213632_at	1	212357_at	-1	200063_s_at	1	203509_at	-1	1562416_at	-1
213279_at	-1	212356_at	-1	212298_at	-1	215235_at	-1	238967_at	-1
201479_at	1	212355_at	-1	217850_at	1	208611_s_at	-1	229004_at	-1
226763_at	-1	36865_at	1	231785_at	-1	229952_at	-1	216971_s_at	-1
209725_at	1	227920_at	1	206376_at	1	201516_at	1	242509_at	-1
215800_at	-1	225929_s_a	-1	239352_at	1	51192_at	-1	1569150_x_at	-1
204794_at	1	221843_s_a	-1	205135_s_at	1	222557_at	-1	215071_s_at	-1
226440_at	-1	207517_at	-1	223432_at	-1	226923_at	1	1568408_x_at	-1
201325_s_at	-1	225874_at	1	208676_s_at	1	212894_at	1		
91826_at	-1	227285_at	1	201013_s_at	1	235020_at	1		

A.2.4.E2F3

223320_s_at	1	227386_s_at	1	1569796_s_at	1	228401_at	1	222227_at	-1
213485_s_at	-1	220161_s_at	1	212492_s_at	-1	222740_at	1	225382_at	1
209735_at	1	203499_at	-1	212792_at	1	218782_s_at	1	229551_x_at	1
239579_at	1	203358_s_at	1	212956_at	1	209337_at	1	204026_s_at	1
209321_s_at	1	203806_s_at	1	228051_at	1	205128_x_at	-1	59697_at	1
218697_at	1	203805_s_at	1	218829_s_at	1	201606_s_at	-1	244467_at	1
225342_at	1	212231_at	1	218418_s_at	1	219076_s_at	1	241957_x_at	1
201272_at	1	204768_s_at	1	231851_at	1	50965_at	1	241464_s_at	-1
207163_s_at	1	204767_s_at	1	228565_at	1	219562_at	1	238513_at	1
203608_at	1	206404_at	1	226796_at	1	218585_s_at	1	237187_at	1
223094_s_at	1	204379_s_at	1	227804_at	1	1553015_a_at	1	236488_s_at	1
228415_at	1	218974_at	1	229582_at	-1	213338_at	1	236289_at	1
239435_x_at	1	219760_at	1	226702_at	1	212027_at	-1	235919_at	1
37117_at	-1	228774_at	1	235391_at	1	201529_s_at	1	233364_s_at	-1
205980_s_at	-1	204365_s_at	1	235177_at	1	214291_at	-1	229899_s_at	-1
235333_at	1	204364_s_at	1	212771_at	1	238156_at	-1	229715_at	1
204966_at	1	222760_at	1	221823_at	1	221523_s_at	1	229691_at	1
225606_at	1	226487_at	1	225650_at	1	228550_at	1	229656_s_at	1
223566_s_at	1	223171_at	1	211596_s_at	1	204198_s_at	1	228955_at	1
219433_at	1	218510_x_at	1	212850_s_at	1	204197_s_at	1	228238_at	-1
231810_at	1	217899_at	1	212282_at	1	207049_at	1	228180_at	-1
225224_at	1	225139_at	1	212281_s_at	1	203453_at	-1	227193_at	1
218796_at	-1	226925_at	1	212279_at	1	1569594_a_at	-1	226618_at	1
227456_s_at	1	230137_at	1	207069_s_at	1	223283_s_at	1	226549_at	1
227455_at	1	226132_s_at	1	225478_at	1	223282_at	1	226548_at	1
232067_at	1	235144_at	1	218358_at	1	213370_s_at	1	225716_at	1
221766_s_at	1	1553986_at	1	233480_at	-1	206108_s_at	-1	225467_s_at	-1
218309_at	1	236219_at	1	226912_at	1	213649_at	-1	216843_x_at	-1
212252_at	1	244297_at	1	235005_at	1	204979_s_at	1	212693_at	-1
201700_at	1	233592_at	1	226605_at	-1	227923_at	1	209815_at	1
213523_at	1	240161_s_at	1	227764_at	1	39705_at	-1	1568597_at	1
211814_s_at	1	227475_at	1	222728_s_at	-1	229009_at	1	1568408_x_at	-1
205034_at	1	219889_at	1	218750_at	-1	230748_at	1	1556486_at	1
204440_at	1	226348_at	1	201764_at	1	203340_s_at	1	1554007_at	1
212899_at	1	204452_s_at	1	203365_s_at	1	203339_at	1		
212897_at	1	204451_at	1	225185_at	1	222217_s_at	1		
219534_x_at	1	204224_s_at	1	204798_at	1	201349_at	1		
209644_x_at	1	234192_s_at	1	201970_s_at	1	204432_at	1		
204159_at	1	229312_s_at	1	221805_at	1	225752_at	1		
204039_at	1	205280_at	1	222774_s_at	1	202308_at	-1		
205567_at	1	206355_at	1	218888_s_at	1	203016_s_at	1		
203921_at	1	214157_at	1	225921_at	1	209478_at	1		
206756_at	1	227769_at	1	209505_at	1	202260_s_at	1		
226215_s_at	1	242517_at	1	206550_s_at	1	213090_s_at	1		
211358_s_at	1	227471_at	1	227379_at	1	41037_at	1		
204662_at	1	218603_at	1	226350_at	1	212330_at	1		
209674_at	1	242890_at	1	230104_s_at	1	213135_at	1		
39966_at	1	44783_s_at	1	201202_at	1	228256_s_at	1		
218898_at	1	218839_at	1	219295_s_at	1	225388_at	1		
204190_at	-1	222996_s_at	1	212522_at	1	225387_at	1		
209570_s_at	1	205449_at	1	212094_at	1	219892_at	1		
203302_at	1	224361_s_at	1	212092_at	1	204137_at	1		
222889_at	1	224156_x_at	1	244677_at	-1	207291_at	1		
209094_at	1	219255_x_at	1	202464_s_at	1	226186_at	1		
226986_at	1	205067_at	-1	225048_at	1	216005_at	-1		
204382_at	-1	205258_at	1	219126_at	1	202644_s_at	-1		
212730_at	1	227432_s_at	1	212726_at	1	213885_at	1		
213088_s_at	1	226216_at	1	209780_at	1	239694_at	1		
221677_s_at	1	229139_at	1	202927_at	1	228956_at	1		
207267_s_at	1	222668_at	1	226299_at	1	208358_s_at	1		
201908_at	1	222664_at	1	216218_s_at	1	210021_s_at	1		
228033_at	1	238077_at	1	38671_at	1	231227_at	1		
204540_at	1	209781_s_at	1	216026_s_at	1	213425_at	1		
214805_at	-1	212057_at	1	205909_at	1	205990_s_at	1		
201313_at	1	212056_at	1	212230_at	1	203712_at	-1		
219731_at	1	206102_at	1	235266_at	1	204234_s_at	-1		

A.2.5.RAS

203504_s_at	-1	204015_s_at	1	38149_at	1	209193_at	1	1552648_a_at	1
205179_s_at	1	209457_at	1	225611_at	1	221577_x_at	1	231775_at	1
205180_s_at	1	208891_at	1	41386_i_at	1	210845_s_at	1	210405_x_at	1
219935_at	-1	208893_s_at	1	212943_at	-1	211924_s_at	1	218368_s_at	1
206170_at	1	208892_s_at	1	226808_at	-1	214866_at	1	234734_s_at	-1
231067_s_at	1	206722_s_at	1	213358_at	-1	213030_s_at	1	228834_at	1
223333_s_at	1	202711_at	1	229817_at	-1	215667_x_at	-1	208901_s_at	1
221009_s_at	1	227404_s_at	1	221778_at	1	209598_at	1	238688_at	-1
203946_s_at	1	201694_s_at	1	225582_at	1	214146_s_at	1	213293_s_at	-1
203263_s_at	-1	209039_x_at	1	209212_s_at	1	201490_s_at	1	215111_s_at	1
220658_s_at	1	221773_at	1	212408_at	1	201489_at	1	226120_at	-1
209281_s_at	1	203499_at	1	202067_s_at	1	202014_at	1	212242_at	1
212930_at	1	205767_at	1	217173_s_at	1	37028_at	1	209340_at	1
225612_s_at	1	202081_at	1	202068_s_at	1	215707_s_at	1	221291_at	1
1554835_a_at	1	210638_s_at	-1	210732_s_at	-1	227510_x_at	1	203234_at	1
228498_at	1	203639_s_at	-1	212658_at	1	231735_s_at	-1	226029_at	-1
208002_s_at	1	217943_s_at	1	205266_at	1	1554997_a_at	1	212171_x_at	1
203140_at	-1	229676_at	1	1558846_at	1	204748_at	1	210513_s_at	1
209373_at	1	219235_s_at	-1	230323_s_at	1	211756_at	1	211527_x_at	1
205289_at	1	219388_at	-1	226726_at	1	210355_at	1	210512_s_at	1
205290_s_at	1	227180_at	1	238058_at	-1	1556773_at	1	1553993_s_at	-1
219563_at	1	238063_at	1	228046_at	-1	221840_at	1	219836_at	1
1558378_a_at	-1	235390_at	1	232158_x_at	1	206157_at	1	201531_at	1
60474_at	1	1553581_s_at	1	229125_at	-1	214443_at	1	206579_at	-1
218796_at	1	230769_at	1	220317_at	1	225189_s_at	1	234608_at	1
229545_at	1	226908_at	-1	208433_s_at	1	225188_at	1	226863_at	1
1552575_a_at	1	1560017_at	-1	202626_s_at	-1	1553722_s_at	-1	228314_at	1
202241_at	1	208614_s_at	1	228846_at	1	204133_at	1	239331_at	1
207243_s_at	1	208613_s_at	1	226275_at	1	211181_x_at	-1	242509_at	1
214845_s_at	1	219250_s_at	1	223217_s_at	1	211182_x_at	-1	217608_at	1
200756_x_at	1	214701_s_at	-1	208786_s_at	1	228923_at	1	244025_at	1
227364_at	1	209189_at	1	232138_at	-1	230333_at	1	240991_at	1
206011_at	-1	227475_at	1	200797_s_at	1	201286_at	1	226034_at	1
226032_at	-1	213524_s_at	1	235374_at	-1	201287_s_at	1	230711_at	1
205476_at	1	204457_s_at	-1	235077_at	1	202071_at	1	227755_at	1
205899_at	1	215243_s_at	1	203417_at	1	234725_s_at	1	1566968_at	1
241495_at	-1	205490_x_at	1	224480_s_at	1	46665_at	1	227288_at	1
218451_at	1	206156_at	1	215239_x_at	-1	219039_at	1	208785_s_at	1
226372_at	1	215977_x_at	1	238741_at	1	212268_at	1	230973_at	1
219500_at	1	225706_at	-1	229518_at	-1	213572_s_at	1	225950_at	1
230603_at	-1	219267_at	1	220949_s_at	-1	228726_at	1	225316_at	1
208960_s_at	1	226177_at	1	203636_at	-1	204614_at	1	230778_at	1
208961_s_at	1	221050_s_at	1	1557158_s_at	-1	209720_s_at	-1	211506_s_at	1
207945_s_at	1	205014_at	1	217279_x_at	1	204855_at	1	227057_at	1
225756_at	1	208553_at	-1	202828_s_at	1	223196_s_at	1	1558517_s_at	1
202332_at	1	202934_at	1	160020_at	1	223195_s_at	1	224606_at	1
222265_at	1	209377_s_at	-1	1553293_at	1	242899_at	-1	201861_s_at	1
204470_at	1	213472_at	-1	228527_s_at	1	209260_at	1	216483_s_at	1
209774_x_at	1	206858_s_at	-1	212096_s_at	-1	203625_x_at	-1	211620_x_at	-1
207850_at	1	222881_at	1	209124_at	1	202856_s_at	1	229949_at	-1
215101_s_at	1	219403_s_at	1	204823_at	1	201920_at	1	1568513_x_at	-1
202436_s_at	-1	212983_at	1	200632_s_at	1	216236_s_at	1	215071_s_at	-1
202435_s_at	-1	201631_s_at	1	211467_s_at	-1	202499_s_at	1	232947_at	-1
205676_at	1	206924_at	1	205895_s_at	1	209453_at	1	230779_at	-1
227109_at	-1	206172_at	1	1553995_a_at	1	209427_at	1	232478_at	-1
201925_s_at	1	210118_s_at	1	203939_at	1	207390_s_at	1	241464_s_at	-1
201926_s_at	1	39402_at	1	206376_at	1	230820_at	1	229872_s_at	-1
1555950_a_at	1	205067_at	1	200790_at	1	210001_s_at	1	243712_at	-1
208151_x_at	-1	202859_x_at	1	202696_at	1	221489_s_at	1	1570425_s_at	-1
208719_s_at	-1	202794_at	1	218736_s_at	-1	1554671_a_at	-1	236656_s_at	-1
204420_at	1	223309_x_at	1	1555167_s_at	1	202440_s_at	-1	240245_at	-1
235263_at	-1	228462_at	-1	227458_at	1	204729_s_at	1	216867_s_at	-1
224215_s_at	-1	205032_at	1	223834_at	1	225544_at	1	232034_at	-1
215210_s_at	1	201188_s_at	1	217997_at	1	216035_x_at	-1	229004_at	-1
204720_s_at	-1	201189_s_at	1	218000_s_at	1	209278_s_at	1	1559360_at	-1
38037_at	1	201473_at	1	217996_at	1	205016_at	1	234951_s_at	-1
203821_at	1	204678_s_at	1	209803_s_at	1	205015_s_at	1	227449_at	-1
201041_s_at	1	204679_at	1	203691_at	1	220407_s_at	-1	209908_s_at	-1
201044_x_at	1	204401_at	1	217864_s_at	-1	201447_at	-1		
204014_at	1	204882_at	1	203879_at	1	201666_at	1		

A.2.6.SRC

213485_s_at	-1
201128_s_at	-1
215867_x_at	-1
201879_at	-1
222667_s_at	-1
218796_at	-1
206011_at	-1
213243_at	-1
221900_at	-1
229666_s_at	-1
206414_s_at	-1
213279_at	-1
203301_s_at	-1
213865_at	-1
225461_at	-1
209537_at	-1
218397_at	-1
1568680_s_at	-1
31874_at	-1
213056_at	-1
206976_s_at	-1
238933_at	-1
235392_at	-1
213352_at	-1
212492_s_at	-1
213069_at	-1
219181_at	-1
231866_at	-1
229582_at	-1
202245_at	-1
202569_s_at	-1
242082_at	1
213164_at	-1
37028_at	1
226065_at	-1
1552797_s_at	-1
1556773_at	-1
211756_at	-1
206591_at	1
212044_s_at	1
200908_s_at	1
213350_at	1
202648_at	1
209773_s_at	-1
213262_at	-1
224250_s_at	-1
204614_at	-1
204404_at	-1
212560_at	-1
1558211_s_at	1
221284_s_at	1
202506_at	-1
201737_s_at	-1
201447_at	-1
224321_at	1
202643_s_at	-1
220687_at	1
212928_at	-1
1554021_a_at	-1
219571_s_at	-1
204847_at	-1
241617_x_at	1
229101_at	-1
225640_at	-1
212435_at	-1
235423_at	-1
230304_at	-1
228955_at	-1
1556006_s_at	-1
227921_at	-1
1556499_s_at	-1
236251_at	-1
1568408_x_at	-1

A.2.7.Beta-Catenin

225098_at	-1	207700_s_at	-1
218150_at	-1	213328_at	-1
222667_s_at	-1	203304_at	1
208859_s_at	-1	211671_s_at	-1
222696_at	1	229422_at	-1
60474_at	-1	244677_at	-1
218796_at	-1	226094_at	-1
212996_s_at	-1	207002_s_at	-1
212177_at	-1	209318_x_at	-1
204048_s_at	-1	219024_at	-1
1555945_s_at	-1	210355_at	-1
1555920_at	-1	212263_at	-1
236241_at	-1	235209_at	1
211343_s_at	-1	212044_s_at	1
221900_at	-1	213350_at	1
215646_s_at	-1	202648_at	1
209257_s_at	-1	224250_s_at	-1
206504_at	1	222747_s_at	-1
223139_s_at	-1	1569594_a_at	-1
229115_at	-1	244287_at	-1
209457_at	-1	213850_s_at	-1
212420_at	-1	206108_s_at	-1
200842_s_at	-1	210057_at	-1
203255_at	-1	203509_at	-1
226799_at	-1	212560_at	-1
225021_at	-1	222122_s_at	-1
235388_at	-1	212994_at	-1
222760_at	1	202643_s_at	-1
232094_at	-1	208901_s_at	-1
227475_at	1	208900_s_at	-1
210178_x_at	-1	203147_s_at	1
222834_s_at	-1	214814_at	-1
225097_at	-1	222227_at	-1
225116_at	-1	1555673_at	1
210118_s_at	-1	241617_x_at	1
208953_at	-1	241464_s_at	-1
212355_at	-1	217277_at	1
213352_at	-1	228315_at	-1
1554260_a_at	-1	233204_at	-1
216563_at	-1	244075_at	-1
212492_s_at	-1	201865_x_at	-1
213478_at	-1	229958_at	-1
212794_s_at	-1	1557081_at	-1
235009_at	-1	1560318_at	-1
223380_s_at	-1	228180_at	-1
212692_s_at	-1	1568408_x_at	-1
1558173_a_at	-1	1562416_at	-1
229846_s_at	-1	232231_at	1
222728_s_at	-1	213637_at	-1

A.2.8.Stemness

206442_at	1	211778_s_at	1
206286_s_at	1	203917_at	1
210905_x_at	1	208755_x_at	1
214791_at	1	204224_s_at	1
218319_at	1	209864_at	1
204294_at	1	220536_at	1
213721_at	1	211331_x_at	1
203449_s_at	1	209168_at	1
218338_at	1	218536_at	1
206857_s_at	1	213947_s_at	1
206424_at	1	201578_at	1
210074_at	1	204391_x_at	1
203286_at	1	219823_at	1
216623_x_at	1	203298_s_at	1
218261_at	1	203448_s_at	1
203638_s_at	1	205938_at	1
211401_s_at	1	210852_s_at	1
205309_at	1	202683_s_at	1
215145_s_at	1	213828_x_at	1
208939_at	1	208940_at	1
219121_s_at	1	204890_s_at	1
214532_x_at	1	202551_s_at	1
202003_s_at	1	204832_s_at	1
210758_at	1	210265_x_at	1
204836_at	1	204807_at	1
209757_s_at	1	219743_at	1
201413_at	1	215707_s_at	1
220668_s_at	1	202889_x_at	1
204084_s_at	1	41577_at	1
203129_s_at	1	200096_s_at	1
203453_at	1	220285_at	1
214023_x_at	1	211214_s_at	1
204154_at	1	218186_at	1
208899_x_at	1	206012_at	1
208358_s_at	1	221605_s_at	1
205742_at	1	205640_at	1
216266_s_at	1	212919_at	1
201839_s_at	1	209489_at	1
220184_at	1	212180_at	1
217988_at	1	210029_at	-1
208286_x_at	1	202822_at	-1
213467_at	1	221245_s_at	-1
209170_s_at	1	211940_x_at	-1
219301_s_at	1	202911_at	-1
205350_at	1	202956_at	-1
214397_at	1	213722_at	-1
212750_at	1	213050_at	-1
203020_at	1	201266_at	-1
203639_s_at	1	213283_s_at	-1
209169_at	1	213301_x_at	-1

A.2.9.Invasiveness

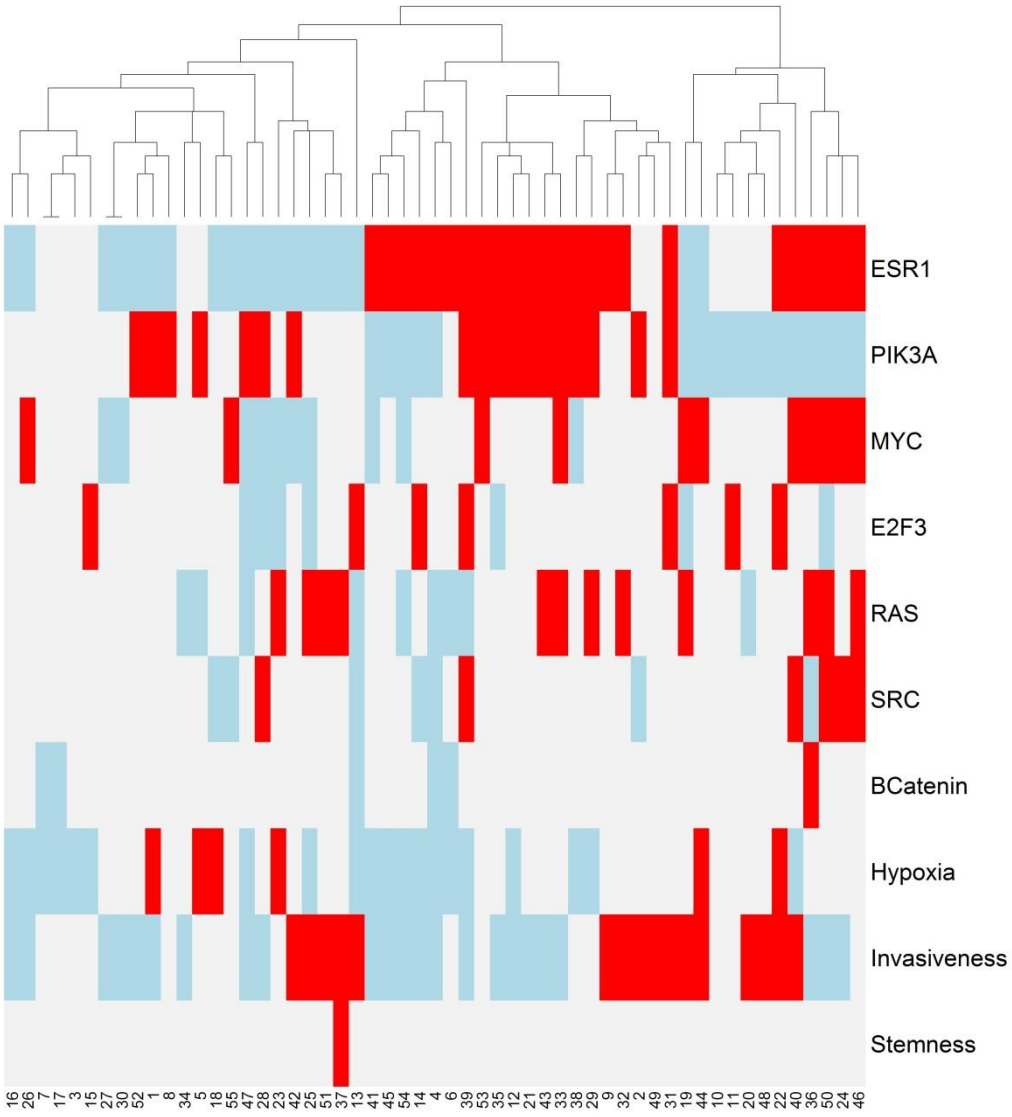
204822_at	1	209172_s_at	1
218542_at	1	204240_s_at	1
219918_s_at	1	203554_x_at	1
202870_s_at	1	218662_s_at	1
219148_at	1	218726_at	1
209642_at	1	206364_at	1
203764_at	1	201014_s_at	1
204962_s_at	1	202503_s_at	1
204641_at	1	213007_at	1
201292_at	1	200783_s_at	1
210052_s_at	1	209421_at	1
203362_s_at	1	204023_at	1
207828_s_at	1	204767_s_at	1
204444_at	1	222037_at	1
204162_at	1	204026_s_at	1
202095_s_at	1	201663_s_at	1
218355_at	1	203625_x_at	1
201890_at	1	218875_s_at	1
204825_at	1	201897_s_at	1
214710_s_at	1	220865_s_at	1
202954_at	1	201664_at	1
210559_s_at	1	219004_s_at	1
218009_s_at	1	204709_s_at	1
203755_at	1	214061_at	1
218883_s_at	1	208808_s_at	1
209773_s_at	1	209709_s_at	1
218755_at	1	218239_s_at	1
219787_s_at	1	212949_at	1
202580_x_at	1	209825_s_at	1
203213_at	1	212141_at	1
204033_at	1	201930_at	1
218039_at	1	206499_s_at	1
203418_at	1	205024_s_at	1
219000_s_at	1	211042_x_at	1
207165_at	1	218984_at	1
204170_s_at	1	203095_at	1
218782_s_at	1	38158_at	1
202705_at	1	201463_s_at	1
209714_s_at	1	206055_s_at	1
218585_s_at	1	204817_at	1
204244_s_at	1	219481_at	1
203214_x_at	1	212766_s_at	1
209408_at	1	221685_s_at	1
205240_at	1	202808_at	-1
204146_at	1	208611_s_at	-1
213226_at	1	200811_at	-1
222077_s_at	1	216264_s_at	-1
213599_at	1	204863_s_at	-1
219306_at	1	201360_at	-1
204510_at	1	201508_at	-1

A.2.10. Hypoxia

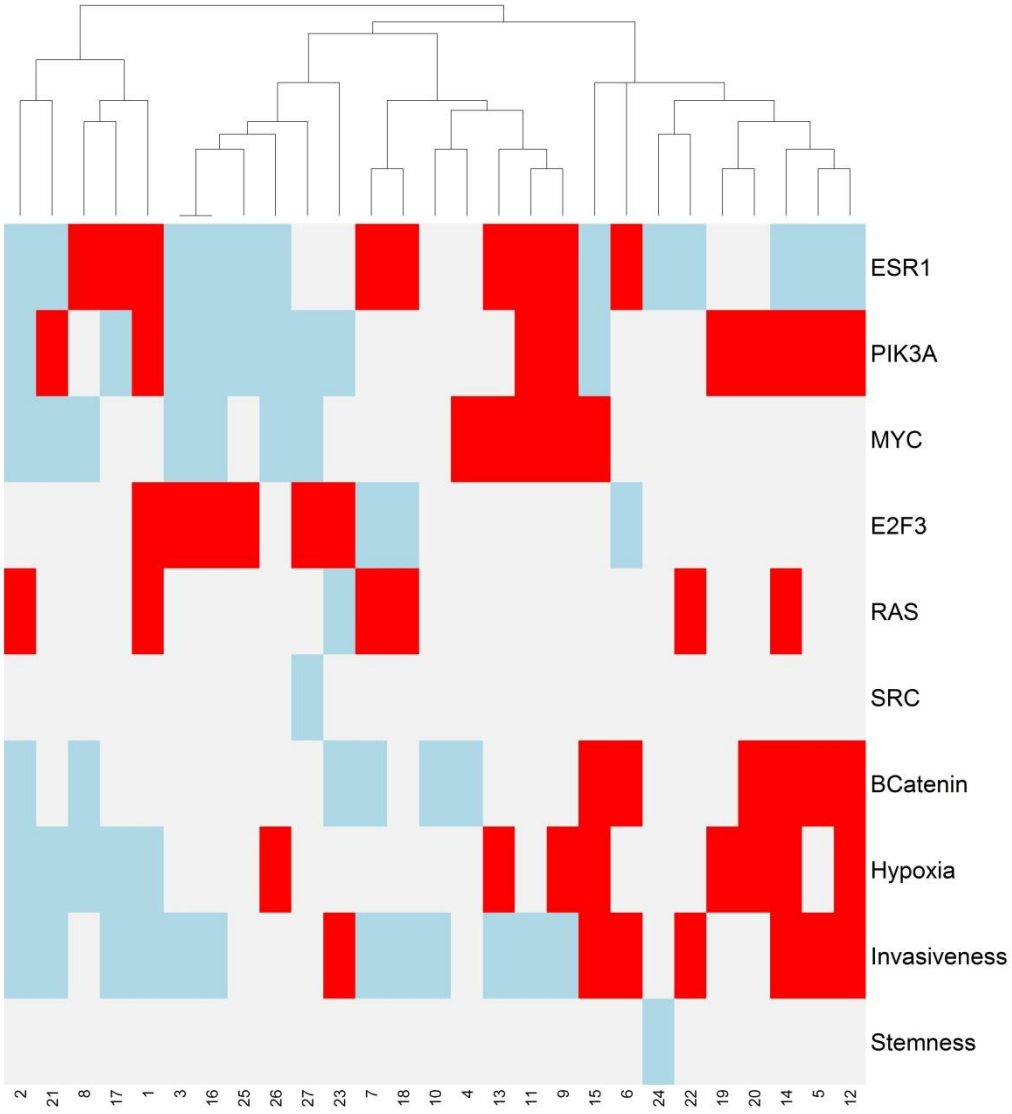
207332_s_at	1	212271_at	1
201231_s_at	1	212980_at	-1
217772_s_at	1	214683_s_at	-1
217294_s_at	1	208699_x_at	1
AFFX-HUMGAPDH/M33197_5_at	1	213453_x_at	1
203746_s_at	1	218982_s_at	1
218163_at	1	203080_s_at	-1
200039_s_at	1	213612_x_at	-1
AFFX-HUMGAPDH/M33197_M_at	1	201298_s_at	1
221263_s_at	1	203484_at	1
201923_at	1	219449_s_at	1
217398_x_at	1	207507_s_at	1
200886_s_at	1	201317_s_at	1
200750_s_at	1	201629_s_at	1
203207_s_at	1	218516_s_at	1
212153_at	-1	202856_s_at	1
213011_s_at	1	205583_s_at	-1
200737_at	1	215227_x_at	1
218482_at	1	213593_s_at	-1
221676_s_at	1	217720_at	1
200738_s_at	1		
200822_x_at	1		
211762_s_at	1		
202511_s_at	1		
217356_s_at	1		
217627_at	-1		
206550_s_at	1		
207668_x_at	1		
200889_s_at	1		
202483_s_at	1		
201321_s_at	-1		
216640_s_at	1		
208799_at	1		
218027_at	1		
201199_s_at	1		
213696_s_at	1		
202929_s_at	1		
208691_at	1		

A.3. Heatmaps representing results of classification

A.3.1. Edinburgh RS dataset

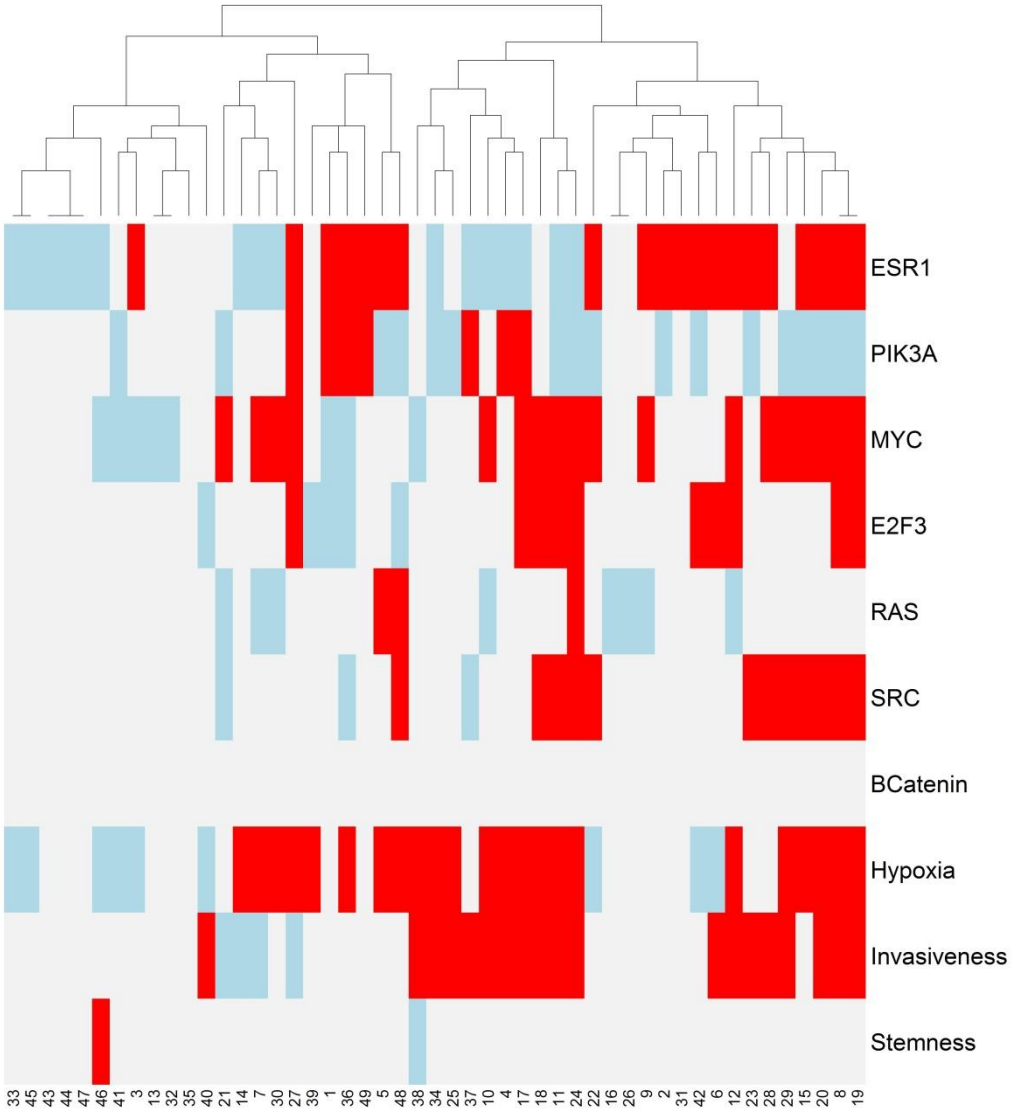


A.3.2. Edinburgh L23 dataset



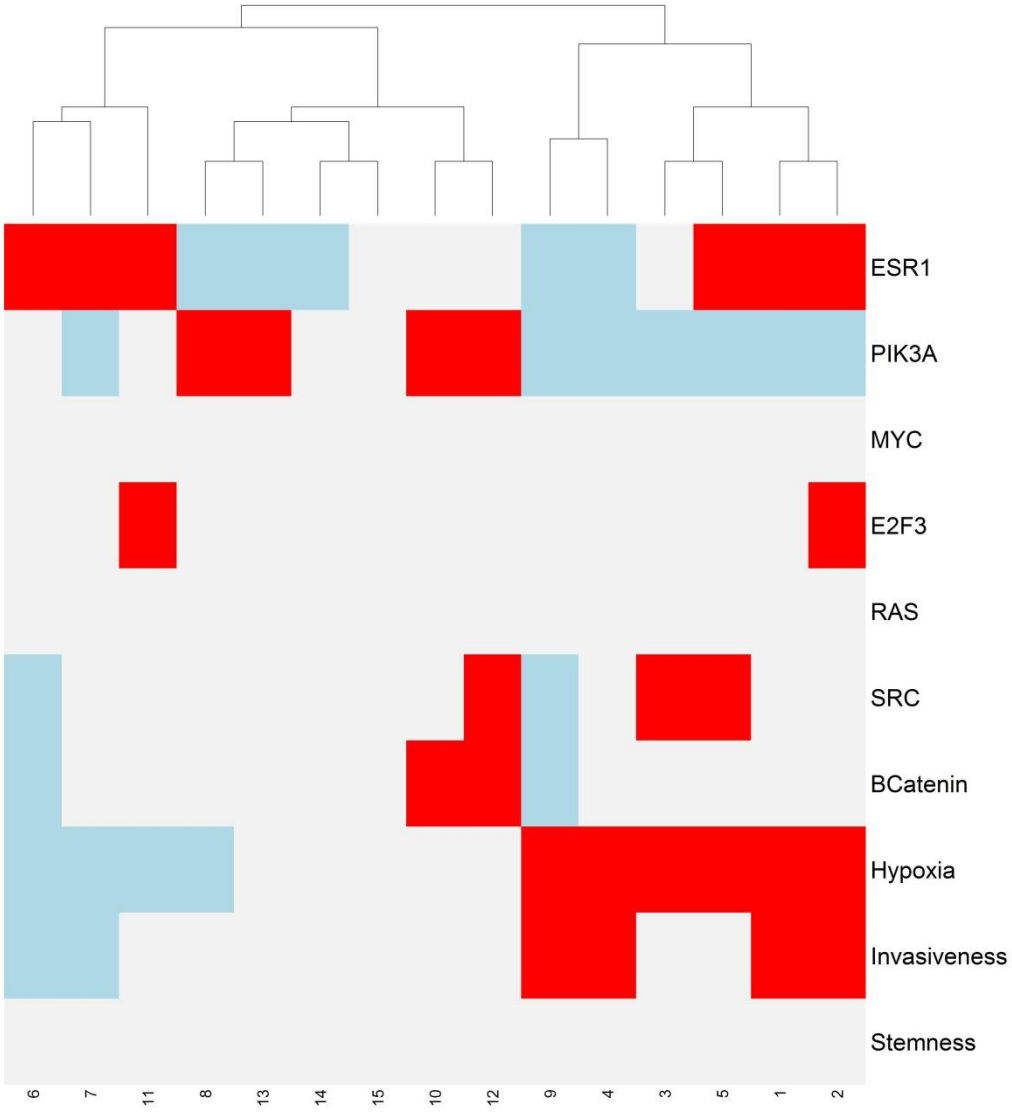
A.3.3. Tam-U133A Series

U133A CEL files from GSE2990/GSE6532 GEO series

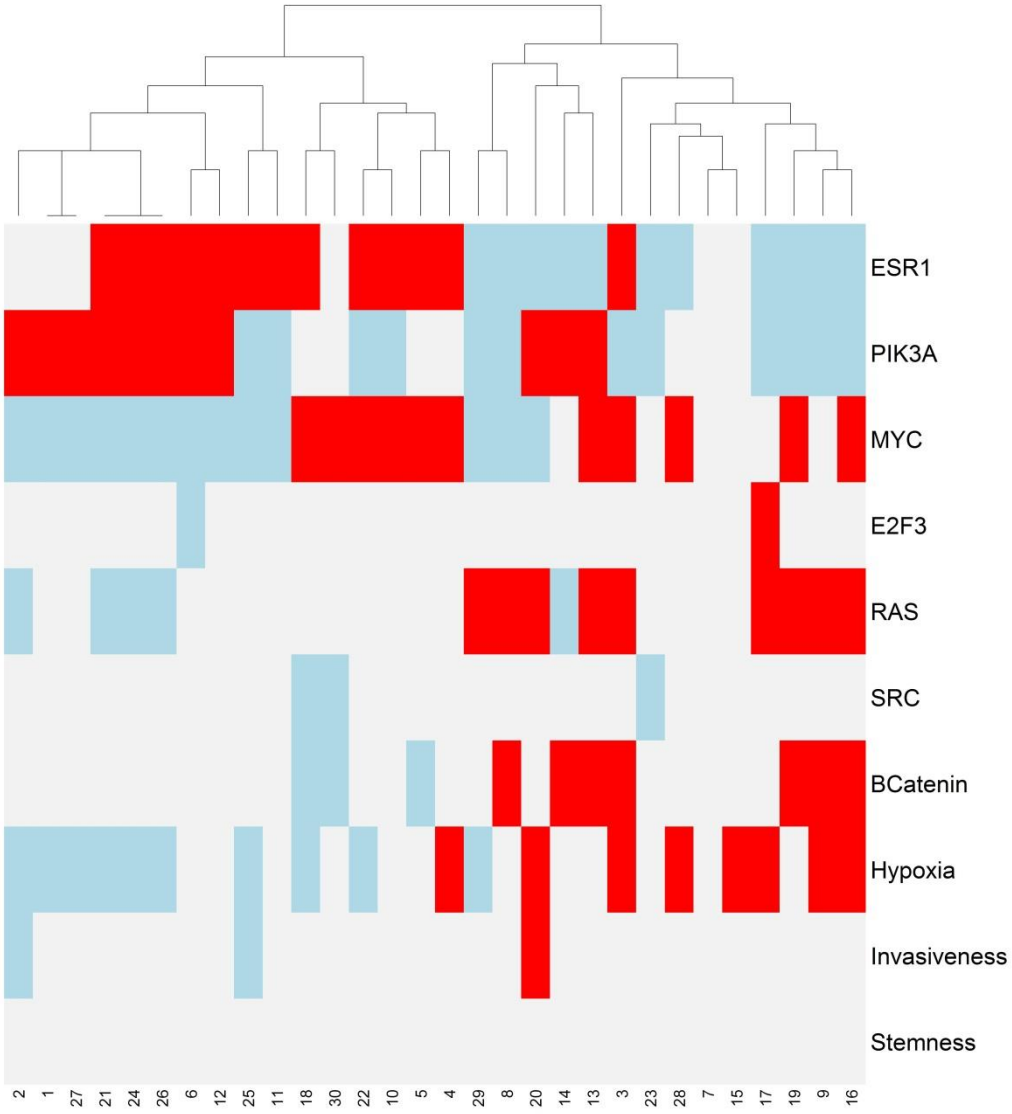


A.3.4. TamU133Plus2 Series

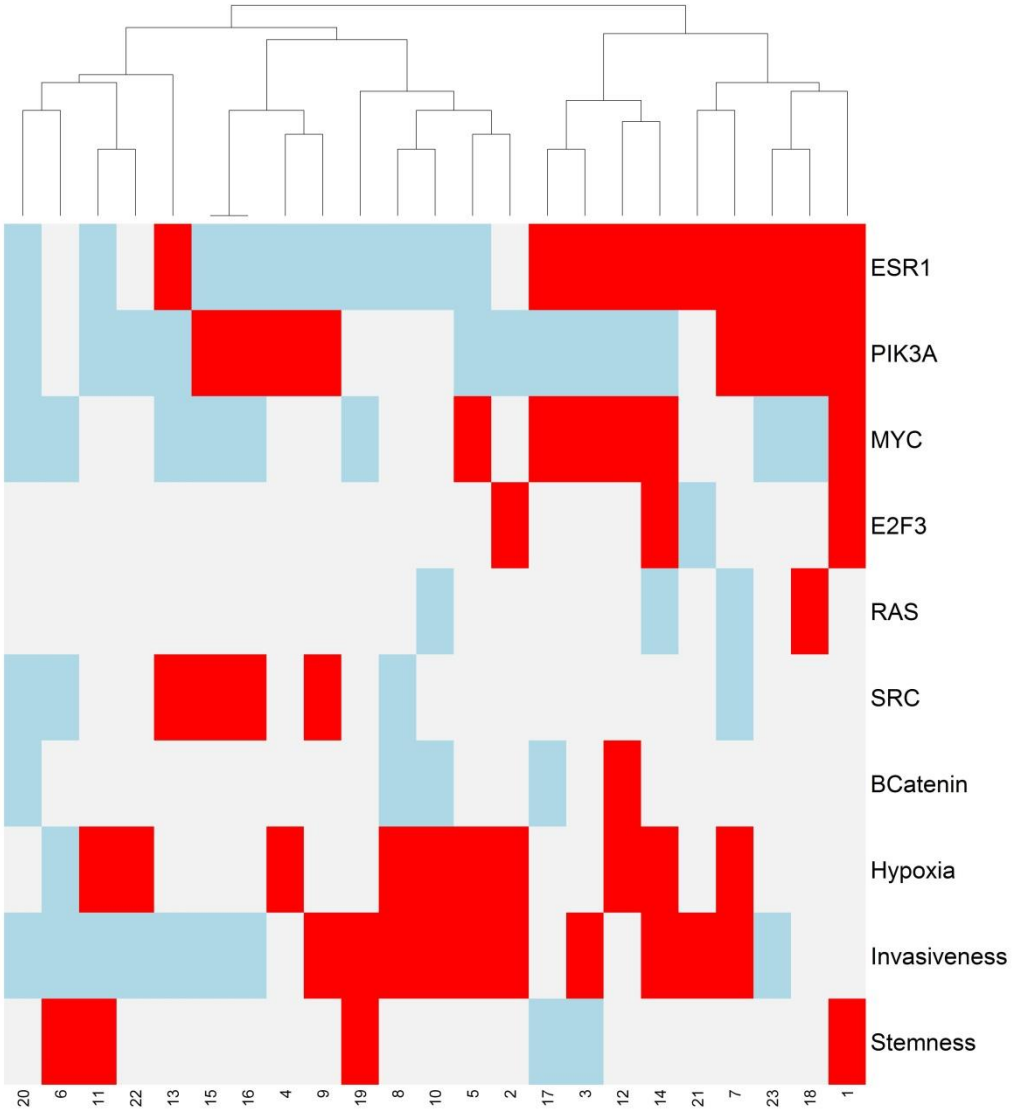
U133-Plus2.0 CEL files from GSE6532 and GSE9195 GEO series



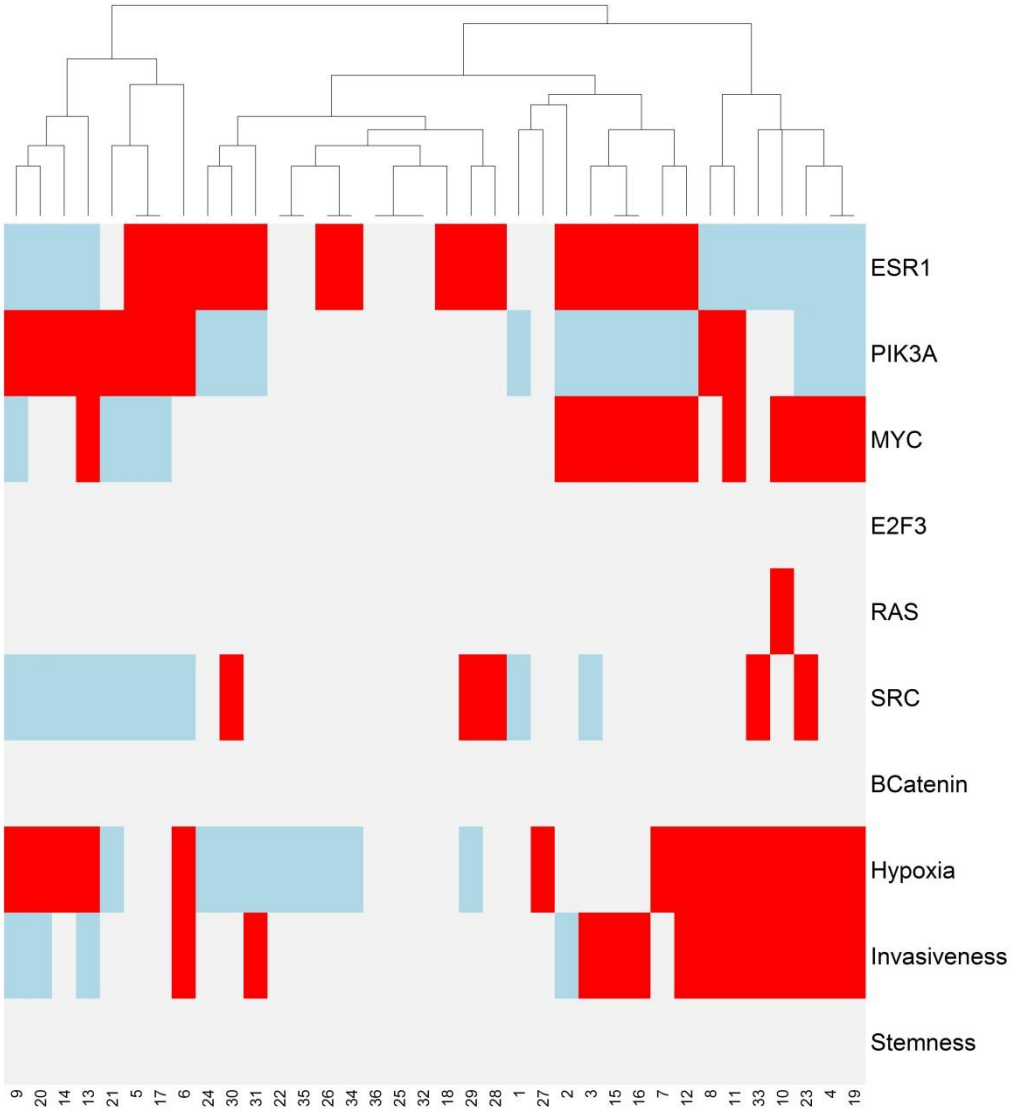
A.3.5. GSE16391 GEO Series



A.3.6. GSE4922 GEO Series



A.3.7. GSE17705 GEO Series



A.4. Examples of web site code and screenshots

A.4.1. Examples of code

A.4.1.1. Index.html

Frame-based index file

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Frameset//EN"
    "http://www.w3.org/TR/html4/frameset.dtd">
<HTML  lang="en">
<HEAD>
    <TITLE> Molecular Diversity of Endocrine Resistance </TITLE>
    <META http-equiv="Content-Type"
      content="text/html; charset=ISO-8859-1">
    <META name="Keywords" content=
      "Breasr cancer, Endocrine resistance,
      Transcriptional signatures">
    <META name="Description" content=
      "Applying transcriptional signatures to endocrine-
      resistant breast cancer datasets">
    <META name="Identifier-URL" content="http://larionov.co.uk/">
    <META HTTP-EQUIV="Pragma" CONTENT="no-cache">
    <META HTTP-EQUIV="Expires" CONTENT="-1">
</HEAD>
<FRAMESET rows="100,*,20" border="0">
    <FRAMESET cols="180,*">
        <FRAME name="Logo" src="Logo.htm"
            marginwidth="10" marginheight="10" scrolling="no"
            frameborder="0" border="none">
        <FRAME name="Top" src="Top.htm"
            marginwidth="10" marginheight="10" scrolling="no"
            frameborder="0" border="none">
    </FRAMESET>
    <FRAMESET cols="180,*">
        <FRAME name="Left" src="Menu.htm"
            marginwidth="10" marginheight="10" scrolling="auto"
            frameborder="0" border="none">
        <FRAME name="Main" src="Introduction.htm"
            marginwidth="10" marginheight="10"
            scrolling="auto" frameborder="0" border="none">
    </FRAMESET>
    <FRAMESET cols="180,*">
        <FRAME name="LeftFooter" src="LeftFooter.htm"
            marginwidth="10" marginheight="5" scrolling="no"
            frameborder="0" border="none">
        <FRAME name="MainFooter" src="MainFooter.htm"
            marginwidth="10" marginheight="5" scrolling="no"
            frameborder="0" border="none">
    </FRAMESET>
</FRAMESET>
</HTML>
```

A.4.1.2. Styles.css

Fragment of the style sheet

```
<STYLE type="text/css">

P.TopTitle
{
    FONT-FAMILY: Arial;
    FONT-SIZE: 24pt;
    TEXT-ALIGN: center;
      COLOR: blue;
      PADDING-BOTTOM : 0pt;
      PADDING-TOP : 0pt;
      MARGIN-BOTTOM : 0pt;
      MARGIN-TOP : 0pt;
}
P.TopSubTitle
{
    FONT-FAMILY: Arial;
    FONT-SIZE: 14pt;
    TEXT-ALIGN: center;
}
P.MainTitle
{
    FONT-FAMILY: Arial;
    FONT-SIZE: 16pt;
    TEXT-ALIGN: center;
}
```

More paragraph styles go here …

```
P.Footer
{
      FONT-FAMILY: Arial;
      FONT-SIZE: 8pt;
      TEXT-ALIGN: center;
}
A
{
    TEXT-DECORATION: none;
}
A:hover
{
      COLOR: red;
}
```

Etc…

A.4.1.3. Menu.html

HHML code for the Menu frame

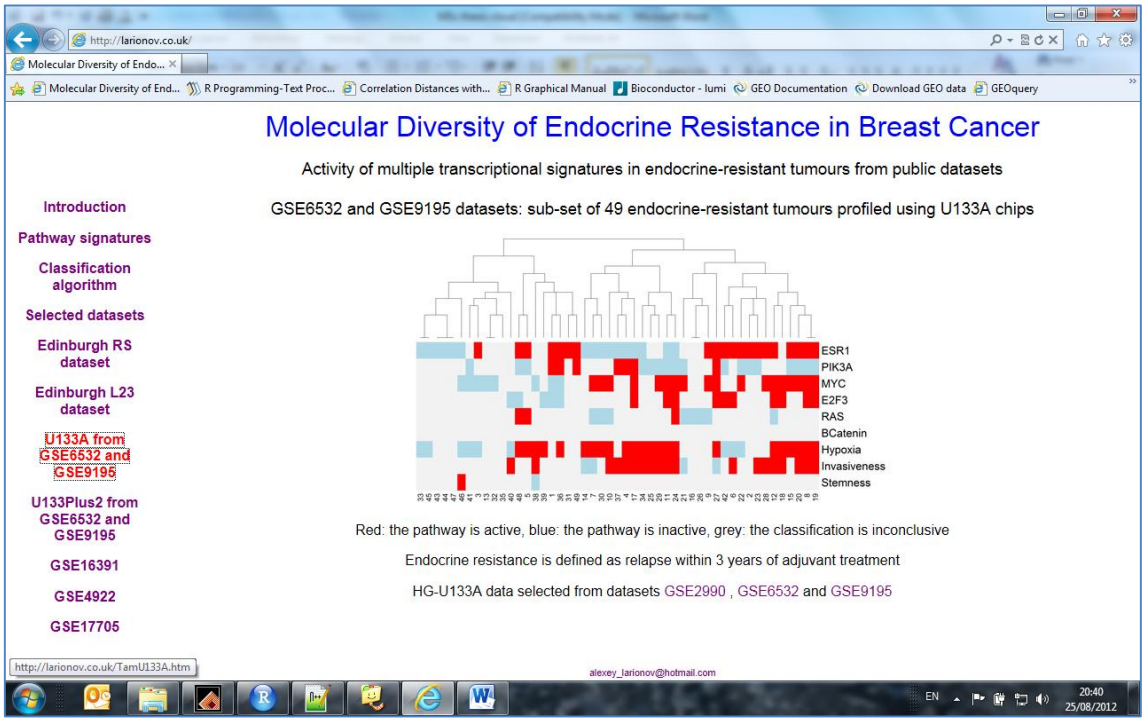
```
<HTML>
<HEAD>
    <TITLE> Molecular Diversity of Endocrine Resistance </TITLE>
    <META http-equiv="Content-Type"
      content="text/html; charset=ISO-8859-1">
    <META http-equiv="Content-Style-Type" content="text/css">
    <LINK rel=stylesheet type="text/css" href="Styles.css">
</HEAD>
<BODY>
      <P class="Menu">
            <A href="Introduction.htm" target="Main">Introduction</A>
      </P>
      <P class="Menu">
            <A href="Signatures.htm" target="Main">Pathway
            signatures</A>
      </P>
      <P class="Menu">
            <A href="Classification.htm" target="Main">Classification
            algorithm</A>
      </P>
      <P class="Menu">
            <A href="Datasets.htm" target="Main">Selected datasets</A>
      </P>
      <P class="Menu">
            <A href="RS.htm" target="Main">Edinburgh RS dataset</A>
      </P>
      <P class="Menu">
            <A href="L23.htm" target="Main">Edinburgh L23 dataset</A>
      </P>
      <P class="Menu">
            <A href="TamU133A.htm" target="Main">U133A from GSE6532
            and GSE9195</A>
      </P>
      <P class="Menu">
            <A href="TamU133Plus2.htm" target="Main">U133Plus2 from
            GSE6532 and GSE9195</A>
      </P>
      <P class="Menu">
            <A href="GSE16391.htm" target="Main">GSE16391</A>
      </P>
      <P class="Menu">
            <A href="GSE4922.htm" target="Main">GSE4922</A>
      </P>
      <P class="Menu">
            <A href="GSE17705.htm" target="Main">GSE17705</A>
      </P>
</BODY>
</HTML>
```

A.4.1.4. TamU133Plus2.html

Example of a page presenting results of analysis

```
<HTML>
<HEAD>
    <TITLE> Molecular Diversity of Endocrine Resistance </TITLE>
    <META http-equiv="Content-Type"
      content="text/html; charset=ISO-8859-1">
    <META http-equiv="Content-Style-Type" content="text/css">
    <LINK rel=stylesheet type="text/css" href="Styles.css">
    <META HTTP-EQUIV="Pragma" CONTENT="no-cache">
    <META HTTP-EQUIV="Expires" CONTENT="-1">
</HEAD>
<BODY>
    <P class = "MainSubTitle">
            GSE6532 and GSE9195 datasets: sub-set of 15 endocrine-
            resistant tumours profiled using U133-Plus2 chips
      </P>
      <P class = "MainFigure">
            <img border="0" alt="TamU133 Plus 2 picture"
            src="TamU133Plus2_web.jpg"/>
      </P>
    <P class = "MainTextCenter">
            Red: the pathway is active, blue: the pathway is
            inactive, grey: the classification is inconclusive
      </P>
    <P class = "MainTextCenter">
            Endocrine resistance is defined as relapse within
            3 years of adjuvant treatment
      </P>
    <P class = "MainTextCenter">
            HG-U133 Plus-2 data selected from datasets
            <a
href="http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532"
                  target="_blank"> GSE6532 </a> and
            <a
href="http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9195"
                  target="_blank"> GSE9195 </a>
      </P>
</BODY>
</HTML>
```

A.4.2. Example of a screenshot



References

1. W.H.O. (2008) *Ten leading causes of death in 2008*. Available from: http://gamapserver.who.int/gho/interactive_charts/mbd/cod_2008/graph.html (accessed 30th August 2012)
2. Mistry, M., Parkin, D.M., Ahmad, A.S. and Sasieni, P. (2011) Cancer incidence in the United Kingdom: projections to the year 2030. *British journal of cancer*, **105**, 1795-1803.
3. Dolan, P., Torgerson, D.J. and Wolstenholme, J. (1999) Costs of breast cancer treatment in the United Kingdom. *Breast*, **8**, 205-207.
4. Taplin, S.H., Barlow, W., Urban, N., Mandelson, M.T., Timlin, D.J., Ichikawa, L. and Nefcy, P. (1995) Stage, age, comorbidity, and direct costs of colon, prostate, and breast cancer care. *Journal of the National Cancer Institute*, **87**, 417-426.
5. Will, B.P., Berthelot, J.M., Le Petit, C., Tomiak, E.M., Verma, S. and Evans, W.K. (2000) Estimates of the lifetime costs of breast cancer treatment in Canada. *European journal of cancer*, **36**, 724-735.
6. Lidgren, M.W., N.; Jönsson, B.; Rehnberg, C. (2007) Resource use and costs associated with different states of breast cancer. *International Journal of Technology Assessment in Health Care*, **23**, 223–231.
7. McPherson, K., Steel, C.M. and Dixon, J.M. (2000) ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. *Bmj*, **321**, 624-628.
8. Cauley, J.A., Lucas, F.L., Kuller, L.H., Stone, K., Browner, W. and Cummings, S.R. (1999) Elevated serum estradiol and testosterone concentrations are associated with a high risk for breast cancer. Study of Osteoporotic Fractures Research Group. *Ann Intern Med*, **130**, 270-277.
9. Clemons, M. and Goss, P. (2001) Estrogen and the risk of breast cancer. *The New England journal of medicine*, **344**, 276-285.
10. Jackson, S.P. and Bartek, J. (2009) The DNA-damage response in human biology and disease. *Nature*, **461**, 1071-1078.
11. Lombard, D.B., Chua, K.F., Mostoslavsky, R., Franco, S., Gostissa, M. and Alt, F.W. (2005) DNA repair, genome stability, and aging. *Cell*, **120**, 497-512.
12. Reinhardt, H.C. and Schumacher, B. (2012) The p53 network: cellular and systemic DNA damage responses in aging and cancer. *Trends Genet*, **28**, 128-136.
13. Dorshkind, K. and Swain, S. (2009) Age-associated declines in immune system development and function: causes, consequences, and reversal. *Curr Opin Immunol*, **21**, 404-407.
14. Markopoulos, C., Berger, U., Wilson, P., Gazet, J.C. and Coombes, R.C. (1988) Oestrogen receptor content of normal breast cells and breast carcinomas throughout the menstrual cycle. *Br Med J (Clin Res Ed)*, **296**, 1349-1351.
15. Yang, L. and Jacobsen, K.H. (2008) A systematic review of the association between breastfeeding and breast cancer. *J Womens Health (Larchmt)*, **17**, 1635-1645.
16. Clarke, R.B. (2006) Ovarian steroids and the human breast: regulation of stem cells and cell proliferation. *Maturitas*, **54**, 327-334.

17. Russo, J., Ao, X., Grill, C. and Russo, I.H. (1999) Pattern of distribution of cells positive for estrogen receptor alpha and progesterone receptor in relation to proliferating cells in the mammary gland. *Breast Cancer Res Tr*, **53**, 217-227.
18. Clarke, R.B., Howell, A., Potten, C.S. and Anderson, E. (1997) Dissociation between steroid receptor expression and cell proliferation in the human breast. *Cancer research*, **57**, 4987-4991.
19. McGuire, W.L., Carbone, P.P., Vollmer, E.P. and United States. National Cancer Institute. Breast Cancer Treatment Committee. (1975) *Estrogen receptors in human breast cancer*. Raven Press, New York.
20. Fowler, A.M. and Alarid, E.T. (2007) Amping up estrogen receptors in breast cancer. *Breast cancer research : BCR*, **9**, 305.
21. Fabris, G., Marchetti, E., Marzola, A., Bagni, A., Querzoli, P. and Nenci, I. (1987) Pathophysiology of estrogen receptors in mammary tissue by monoclonal antibodies. *Journal of steroid biochemistry*, **27**, 171-176.
22. Petersen, O.W., Hoyer, P.E. and van Deurs, B. (1987) Frequency and distribution of estrogen receptor-positive cells in normal, nonlactating human breast tissue. *Cancer research*, **47**, 5748-5751.
23. Holst, F., Stahl, P.R., Ruiz, C., Hellwinkel, O., Jehan, Z., Wendland, M., Lebeau, A., Terracciano, L., Al-Kuraya, K., Janicke, F. *et al.* (2007) Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nat Genet*, **39**, 655-660.
24. Moelans, C.B., Monsuur, H.N., de Pinth, J.H., Radersma, R.D., de Weger, R.A. and van Diest, P.J. (2011) ESR1 amplification is rare in breast cancer and is associated with high grade and high proliferation: a multiplex ligation-dependent probe amplification study. *Cell Oncol (Dordr)*, **34**, 489-494.
25. Cavalieri, E.L., Stack, D.E., Devanesan, P.D., Todorovic, R., Dwivedy, I., Higginbotham, S., Johansson, S.L., Patil, K.D., Gross, M.L., Gooden, J.K. *et al.* (1997) Molecular origin of cancer: catechol estrogen-3,4-quinones as endogenous tumor initiators. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 10937-10942.
26. Lee, J.S., Ettinger, B., Stanczyk, F.Z., Vittinghoff, E., Hanes, V., Cauley, J.A., Chandler, W., Settlage, J., Beattie, M.S., Folkerd, E. *et al.* (2006) Comparison of methods to measure low serum estradiol levels in postmenopausal women. *The Journal of clinical endocrinology and metabolism*, **91**, 3791-3797.
27. Fortunati, N., Catalano, M.G., Boccuzzi, G. and Frairia, R. (2010) Sex Hormone-Binding Globulin (SHBG), estradiol and breast cancer. *Mol Cell Endocrinol*, **316**, 86-92.
28. Thomas, H.V., Key, T.J., Allen, D.S., Moore, J.W., Dowsett, M., Fentiman, I.S. and Wang, D.Y. (1997) A prospective study of endogenous serum hormone concentrations and breast cancer risk in premenopausal women on the island of Guernsey. *British journal of cancer*, **75**, 1075-1079.
29. Rowell, S., Newman, B., Boyd, J. and King, M.C. (1994) Inherited predisposition to breast and ovarian cancer. *Am J Hum Genet*, **55**, 861-865.
30. Turnbull, C. and Rahman, N. (2008) Genetic predisposition to breast cancer: past, present, and future. *Annu Rev Genomics Hum Genet*, **9**, 321-345.
31. Cui, J. and Hopper, J.L. (2000) Why are the majority of hereditary cases of early-onset breast cancer sporadic? A simulation study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, **9**, 805-812.
32. Gadzicki, D., Evans, D.G., Harris, H., Julian-Reynier, C., Nippert, I., Schmidtke, J., Tibben, A., van Asperen, C.J. and Schlegelberger, B. (2011) Genetic testing for familial/hereditary breast cancer-comparison of guidelines and recommendations from the UK, France, the Netherlands and Germany. *J Community Genet*, **2**, 53-69.

33. Mack, T.M., Hamilton, A.S., Press, M.F., Diep, A. and Rappaport, E.B. (2002) Heritable breast cancer in twins. *British journal of cancer*, **87**, 294-300.
34. Imyanitov, E.N. and Hanson, K.P. (2003) Molecular pathogenesis of bilateral breast cancer. *Cancer letters*, **191**, 1-7.
35. Claus, E.B., Risch, N.J. and Thompson, W.D. (1990) Age at onset as an indicator of familial risk of breast cancer. *Am J Epidemiol*, **131**, 961-972.
36. Peto, J. and Mack, T.M. (2000) High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet*, **26**, 411-414.
37. Hamilton, A.S. and Mack, T.M. (2003) Puberty and genetic susceptibility to breast cancer in a case-control study in twins. *The New England journal of medicine*, **348**, 2313-2322.
38. (2001) Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet*, **358**, 1389-1399.
39. Goldgar, D.E., Healey, S., Dowty, J.G., Da Silva, L., Chen, X., Spurdle, A.B., Terry, M.B., Daly, M.J., Buys, S.M., Southey, M.C. *et al.* (2011) Rare variants in the ATM gene and risk of breast cancer. *Breast cancer research : BCR*, **13**, R73.
40. Sokolenko, A.P., Iyevleva, A.G., Preobrazhenskaya, E.V., Mitiushkina, N.V., Abysheva, S.N., Suspitsin, E.N., Kuligina, E., Gorodnova, T.V., Pfeifer, W., Togo, A.V. *et al.* (2012) High prevalence and breast cancer predisposing role of the BLM c.1642 C>T (Q548X) mutation in Russia. *International journal of cancer. Journal international du cancer*, **130**, 2867-2873.
41. Wang, F., Fang, Q., Ge, Z., Yu, N., Xu, S. and Fan, X. (2012) Common BRCA1 and BRCA2 mutations in breast cancer families: a meta-analysis from systematic review. *Mol Biol Rep*, **39**, 2109-2118.
42. Nevanlinna, H. and Bartek, J. (2006) The CHEK2 gene and inherited breast cancer susceptibility. *Oncogene*, **25**, 5912-5919.
43. Liu, J., Desai, K.V., Li, Y., Banu, S., Lee, Y.K., Qu, D., Heikkinen, T., Aaltonen, K., Muranen, T.A., Kajiji, T.S. *et al.* (2009) Germ-line variation at a functional p53 binding site increases susceptibility to breast cancer development. *Hugo J*, **3**, 31-40.
44. Haiman, C.A., Stram, D.O., Pike, M.C., Kolonel, L.N., Burtt, N.P., Altshuler, D., Hirschhorn, J. and Henderson, B.E. (2003) A comprehensive haplotype analysis of CYP19 and breast cancer risk: the Multiethnic Cohort. *Hum Mol Genet*, **12**, 2679-2692.
45. Huang, C.S., Chern, H.D., Chang, K.J., Cheng, C.W., Hsu, S.M. and Shen, C.Y. (1999) Breast cancer risk associated with genotype polymorphism of the estrogen-metabolizing genes CYP17, CYP1A1, and COMT: a multigenic study on cancer susceptibility. *Cancer research*, **59**, 4870-4875.
46. Li, N., Dong, J., Hu, Z., Shen, H. and Dai, M. (2010) Potentially functional polymorphisms in ESR1 and breast cancer risk: a meta-analysis. *Breast Cancer Res Tr*, **121**, 177-184.
47. Gold, B., Kalush, F., Bergeron, J., Scott, K., Mitra, N., Wilson, K., Ellis, N., Huang, H., Chen, M., Lippert, R. *et al.* (2004) Estrogen receptor genotypes and haplotypes associated with breast cancer risk. *Cancer research*, **64**, 8891-8900.
48. Dunning, A.M., Healey, C.S., Baynes, C., Maia, A.T., Scollen, S., Vega, A., Rodriguez, R., Barbosa-Morais, N.L., Ponder, B.A., Low, Y.L. *et al.* (2009) Association of ESR1 gene tagging SNPs with breast cancer risk. *Hum Mol Genet*, **18**, 1131-1139.
49. Ponder, B.A., Antoniou, A., Dunning, A., Easton, D.F. and Pharoah, P.D. (2005) Polygenic inherited predisposition to breast cancer. *Cold Spring Harb Symp Quant Biol*, **70**, 35-41.

50. Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nature reviews. Genetics*, **2**, 91-99.
51. Gold, B., Kirchhoff, T., Stefanov, S., Lautenberger, J., Viale, A., Garber, J., Friedman, E., Narod, S., Olshen, A.B., Gregersen, P. *et al.* (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 4340-4345.
52. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087-1093.
53. Pinsky, R.W. and Helvie, M.A. (2010) Mammographic breast density: effect on imaging and breast cancer risk. *J Natl Compr Canc Netw*, **8**, 1157-1164; quiz 1165.
54. Bulun, S.E., Chen, D., Moy, I., Brooks, D.C. and Zhao, H. (2012) Aromatase, breast cancer and obesity: a complex interaction. *Trends Endocrinol Metab*, **23**, 83-89.
55. Carmichael, A.R. (2006) Obesity and prognosis of breast cancer. *Obes Rev*, **7**, 333-340.
56. (1982) The world Health Organization Histological Typing of Breast Tumors--Second Edition. The World Organization. *Am J Clin Pathol*, **78**, 806-816.
57. Gatza, M.L., Lucas, J.E., Barry, W.T., Kim, J.W., Wang, Q., Crawford, M.D., Datto, M.B., Kelley, M., Mathey-Prevot, B., Potti, A. *et al.* (2010) A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 6994-6999.
58. Weigelt, B., Horlings, H.M., Kreike, B., Hayes, M.M., Hauptmann, M., Wessels, L.F., de Jong, D., Van de Vijver, M.J., Van't Veer, L.J. and Peterse, J.L. (2008) Refinement of breast cancer classification by molecular characterization of histological special types. *J Pathol*, **216**, 141-150.
59. Bertos, N.R. and Park, M. (2011) Breast cancer - one term, many entities? *J Clin Invest*, **121**, 3789-3796.
60. Harnett, A., Smallwood, J., Titshall, V. and Champion, A. (2009) Diagnosis and treatment of early breast cancer, including locally advanced disease--summary of NICE guidance. *Bmj*, **338**, b438.
61. Murray, N., Winstanley, J., Bennett, A. and Francis, K. (2009) Diagnosis and treatment of advanced breast cancer: summary of NICE guidance. *Bmj*, **338**, b509.
62. Bloom, H.J. and Richardson, W.W. (1957) Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, **11**, 359-377.
63. Jensen, E.V., Block, G.E., Ferguson, D.J. and DeSombre, E.R. (1977) Estrogen receptors in breast cancer. *World J Surg*, **1**, 341-342.
64. Cuzick, J., Dowsett, M., Pineda, S., Wale, C., Salter, J., Quinn, E., Zabaglo, L., Mallon, E., Green, A.R., Ellis, I.O. *et al.* (2011) Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **29**, 4273-4278.
65. Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747-752.
66. Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, **98**, 10869-10874.

67. Sparano, J.A. and Paik, S. (2008) Development of the 21-gene assay and its application in clinical practice and clinical trials. *J Clin Oncol*, **26**, 721-728.
68. Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A.M., d'Assignies, M.S., Bergh, J., Lidereau, R., Ellis, P. *et al.* (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst*, **98**, 1183-1192.
69. Larionov, A.A. and Miller, W.R. (2009) Challenges in defining predictive markers for response to endocrine therapy in breast cancer. *Future oncology*, **5**, 1415-1428.
70. Beatson, G.T. (1896) On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment, with illustrative cases. *Lancet*, **2**, 104-107, 162-165.
71. Larionov, A.A. and Miller, W.R. (2010) Tailoring of Endocrine Treatment in Breast Cancer. *Treatment Strategies - Oncology*, **1**, 61-67.
72. Robertson, J.F. and Blamey, R.W. (2003) The use of gonadotrophin-releasing hormone (GnRH) agonists in early and advanced breast cancer in pre- and perimenopausal women. *Eur J Cancer*, **39**, 861-869.
73. Leung, S.F., Tsao, S.Y., Teo, P.M., Choi, P.H. and Shiu, W.C. (1991) Ovarian ablation failures by radiation: a comparison of two dose schedules. *The British journal of radiology*, **64**, 537-538.
74. Prowell, T.M. and Davidson, N.E. (2004) What is the role of ovarian ablation in the management of primary and metastatic breast cancer today? *The oncologist*, **9**, 507-517.
75. Simpson, E.R., Mahendroo, M.S., Nichols, J.E. and Bulun, S.E. (1994) Aromatase gene expression in adipose tissue: relationship to breast cancer. *Int J Fertil Menopausal Stud*, **39 Suppl 2**, 75-83.
76. Jordan, V.C. (2003) Tamoxifen: a most unlikely pioneering medicine. *Nat Rev Drug Discov*, **2**, 205-213.
77. Miller, W.R. and Larionov, A.A. (2012) Understanding the mechanisms of aromatase inhibitor resistance. *Breast cancer research : BCR*, **14**, 201.
78. Miller, W.R. and Larionov, A. (2010) Changes in expression of oestrogen regulated and proliferation genes with neoadjuvant treatment highlight heterogeneity of clinical resistance to the aromatase inhibitor, letrozole. *Breast Cancer Research*, **12**, -.
79. Nadji, M. (2008) Quantitative immunohistochemistry of estrogen receptor in breast cancer: "much ado about nothing!". *Appl Immunohistochem Mol Morphol*, **16**, 105-107.
80. Gotay, C. and Dunn, J. (2011) Adherence to long-term adjuvant hormonal therapy for breast cancer. *Expert Rev Pharmacoecon Outcomes Res*, **11**, 709-715.
81. Goetz, M.P., Knox, S.K., Suman, V.J., Rae, J.M., Safgren, S.L., Ames, M.M., Visscher, D.W., Reynolds, C., Couch, F.J., Lingle, W.L. *et al.* (2007) The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. *Breast Cancer Res Treat*, **101**, 113-121.
82. Osborne, C.K. and Schiff, R. (2011) Mechanisms of endocrine resistance in breast cancer. *Annu Rev Med*, **62**, 233-247.
83. Musgrove, E.A. and Sutherland, R.L. (2009) Biological determinants of endocrine resistance in breast cancer. *Nat Rev Cancer*, **9**, 631-643.
84. Colomer, R., Monzo, M., Tusquets, I., Rifa, J., Baena, J.M., Barnadas, A., Calvo, L., Carabantes, F., Crespo, C., Munoz, M. *et al.* (2008) A single-nucleotide polymorphism in the aromatase gene is associated with the efficacy of the aromatase inhibitor letrozole in advanced breast carcinoma. *Clin Cancer Res*, **14**, 811-816.

85. Garcia-Casado, Z., Guerrero-Zotano, A., Llombart-Cussac, A., Calatrava, A., Fernandez-Serra, A., Ruiz-Simon, A., Gavila, J., Climent, M.A., Almenar, S., Cervera-Deval, J. *et al.* (2010) A polymorphism at the 3'-UTR region of the aromatase gene defines a subgroup of postmenopausal breast cancer patients with poor response to neoadjuvant letrozole. *BMC Cancer*, **10**, 36.
86. Patisaul, H.B. and Jefferson, W. (2010) The pros and cons of phytoestrogens. *Front Neuroendocrinol*, **31**, 400-419.
87. Zava, D.T., Blen, M. and Duwe, G. (1997) Estrogenic activity of natural and synthetic estrogens in human breast cancer cells in culture. *Environ Health Perspect*, **105 Suppl 3**, 637-645.
88. Barone, I., Brusco, L. and Fuqua, S.A. (2010) Estrogen receptor mutations and changes in downstream gene expression and signaling. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **16**, 2702-2708.
89. Herynk, M.H. and Fuqua, S.A. (2007) Estrogen receptors in resistance to hormone therapy. *Adv Exp Med Biol*, **608**, 130-143.
90. Osborne, C.K., Shou, J., Massarweh, S. and Schiff, R. (2005) Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **11**, 865s-870s.
91. Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57-70.
92. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646-674.
93. Girault, I., Bieche, I. and Lidereau, R. (2006) Role of estrogen receptor alpha transcriptional coregulators in tamoxifen resistance in breast cancer. *Maturitas*, **54**, 342-351.
94. Williams, C.C., Basu, A., El-Gharbawy, A., Carrier, L.M., Smith, C.L. and Rowan, B.G. (2009) Identification of four novel phosphorylation sites in estrogen receptor alpha: impact on receptor-dependent gene expression and phosphorylation by protein kinase CK2. *BMC Biochem*, **10**, 36.
95. Sutherland, R.L. and Musgrove, E.A. (2004) Cyclins and breast cancer. *J Mammary Gland Biol Neoplasia*, **9**, 95-104.
96. Arnold, A. and Papanikolaou, A. (2005) Cyclin D1 in breast cancer pathogenesis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **23**, 4215-4224.
97. Barnes, D.M. and Gillett, C.E. (1998) Cyclin D1 in breast cancer. *Breast Cancer Res Tr*, **52**, 1-15.
98. Dickson, C., Fantl, V., Gillett, C., Brookes, S., Bartek, J., Smith, R., Fisher, C., Barnes, D. and Peters, G. (1995) Amplification of chromosome band 11q13 and a role for cyclin D1 in human breast cancer. *Cancer letters*, **90**, 43-50.
99. Aaltonen, K., Amini, R.M., Landberg, G., Eerola, H., Aittomaki, K., Heikkila, P., Nevanlinna, H. and Blomqvist, C. (2009) Cyclin D1 expression is associated with poor prognostic features in estrogen receptor positive breast cancer. *Breast Cancer Res Tr*, **113**, 75-82.
100. Balasenthil, S., Barnes, C.J., Rayala, S.K. and Kumar, R. (2004) Estrogen receptor activation at serine 305 is sufficient to upregulate cyclin D1 in breast cancer cells. *FEBS Lett*, **567**, 243-247.
101. Wilcken, N.R., Prall, O.W., Musgrove, E.A. and Sutherland, R.L. (1997) Inducible overexpression of cyclin D1 in breast cancer cells reverses the growth-inhibitory effects of antiestrogens. *Clin Cancer Res*, **3**, 849-854.

102. Bostner, J., Ahnstrom Waltersson, M., Fornander, T., Skoog, L., Nordenskjold, B. and Stal, O. (2007) Amplification of CCND1 and PAK1 as predictors of recurrence and tamoxifen resistance in postmenopausal breast cancer. *Oncogene*, **26**, 6997-7005.
103. Jirstrom, K., Stendahl, M., Ryden, L., Kronblad, A., Bendahl, P.O., Stal, O. and Landberg, G. (2005) Adverse effect of adjuvant tamoxifen in premenopausal breast cancer with cyclin D1 gene amplification. *Cancer research*, **65**, 8009-8016.
104. Lundgren, K., Holm, K., Nordenskjold, B., Borg, A. and Landberg, G. (2008) Gene products of chromosome 11q and their association with CCND1 gene amplification and tamoxifen resistance in premenopausal breast cancer. *Breast cancer research : BCR*, **10**, R81.
105. Brown, L.A., Johnson, K., Leung, S., Bismar, T.A., Benitez, J., Foulkes, W.D. and Huntsman, D.G. (2010) Co-amplification of CCND1 and EMSY is associated with an adverse outcome in ER-positive tamoxifen-treated breast cancers. *Breast Cancer Res Tr*, **121**, 347-354.
106. Akli, S., Bui, T., Wingate, H., Biernacka, A., Moulder, S., Tucker, S.L., Hunt, K.K. and Keyomarsi, K. (2010) Low-molecular-weight cyclin E can bypass letrozole-induced G1 arrest in human breast cancer cells and tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **16**, 1179-1190.
107. Berglund, P. and Landberg, G. (2006) Cyclin e overexpression reduces infiltrative growth in breast cancer: yet another link between proliferation control and tumor invasion. *Cell Cycle*, **5**, 606-609.
108. Caldon, C.E. and Musgrove, E.A. (2010) Distinct and redundant functions of cyclin E1 and cyclin E2 in development and cancer. *Cell Div*, **5**, 2.
109. Hunt, K.K. and Keyomarsi, K. (2005) Cyclin E as a prognostic and predictive marker in breast cancer. *Semin Cancer Biol*, **15**, 319-326.
110. Rasmussen, B.B., Regan, M.M., Lykkesfeldt, A.E., Dell'Orto, P., Del Curto, B., Henriksen, K.L., Mastropasqua, M.G., Price, K.N., Mery, E., Lacroix-Triki, M. *et al.* (2008) Adjuvant letrozole versus tamoxifen according to centrally-assessed ERBB2 status for postmenopausal women with endocrine-responsive early breast cancer: supplementary results from the BIG 1-98 randomised trial. *The lancet oncology*, **9**, 23-28.
111. Shou, J., Massarweh, S., Osborne, C.K., Wakeling, A.E., Ali, S., Weiss, H. and Schiff, R. (2004) Mechanisms of tamoxifen resistance: increased estrogen receptor-HER2/neu cross-talk in ER/HER2-positive breast cancer. *Journal of the National Cancer Institute*, **96**, 926-935.
112. Ahnstrom, M., Nordenskjold, B., Rutqvist, L.E., Skoog, L. and Stal, O. (2005) Role of cyclin D1 in ErbB2-positive breast cancer and tamoxifen resistance. *Breast Cancer Res Tr*, **91**, 145-151.
113. Fujiwara, K., Yuwanita, I., Hollern, D.P. and Andrechek, E.R. (2011) Prediction and genetic demonstration of a role for activator E2Fs in Myc-induced tumors. *Cancer research*, **71**, 1924-1932.
114. Musgrove, E.A., Sergio, C.M., Loi, S., Inman, C.K., Anderson, L.R., Alles, M.C., Pinese, M., Caldon, C.E., Schutte, J., Gardiner-Garden, M. *et al.* (2008) Identification of functional networks of estrogen- and c-Myc-responsive genes and their relationship to response to tamoxifen therapy in breast cancer. *PLoS One*, **3**, e2987.
115. Zhou, Y., Eppenberger-Castori, S., Eppenberger, U. and Benz, C.C. (2005) The NFkappaB pathway and endocrine-resistant breast cancer. *Endocr Relat Cancer*, **12 Suppl 1**, S37-46.

116. Zhou, X., Marian, C., Makambi, K.H., Kosti, O., Kallakury, B.V., Loffredo, C.A. and Zheng, Y.L. (2012) MicroRNA-9 as Potential Biomarker for Breast Cancer Local Recurrence and Tumor Estrogen Receptor Status. *PLoS One*, **7**, e39011.
117. Yoshimoto, N., Toyama, T., Takahashi, S., Sugiura, H., Endo, Y., Iwasa, M., Fujii, Y. and Yamashita, H. (2011) Distinct expressions of microRNAs that directly target estrogen receptor alpha in human breast cancer. *Breast Cancer Res Tr*, **130**, 331-339.
118. Yamashita, H., Toyama, T., Nishio, M., Ando, Y., Hamaguchi, M., Zhang, Z., Kobayashi, S., Fujii, Y. and Iwase, H. (2006) p53 protein accumulation predicts resistance to endocrine therapy and decreased post-relapse survival in metastatic breast cancer. *Breast cancer research : BCR*, **8**, R48.
119. Nehra, R., Riggins, R.B., Shajahan, A.N., Zwart, A., Crawford, A.C. and Clarke, R. (2010) BCL2 and CASP8 regulation by NF-kappaB differentially affect mitochondrial function and cell fate in antiestrogen-sensitive and -resistant breast cancer cells. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, **24**, 2040-2055.
120. Holliday, D.L. and Speirs, V. (2011) Choosing the right cell line for breast cancer research. *Breast cancer research : BCR*, **13**, 215.
121. Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F. *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, **10**, 515-527.
122. Yue, W., Fan, P., Wang, J., Li, Y. and Santen, R.J. (2007) Mechanisms of acquired resistance to endocrine therapy in hormone-dependent breast cancer cells. *The Journal of steroid biochemistry and molecular biology*, **106**, 102-110.
123. Millour, J., Constantinidou, D., Stavropoulou, A.V., Wilson, M.S., Myatt, S.S., Kwok, J.M., Sivanandan, K., Coombes, R.C., Medema, R.H., Hartman, J. *et al.* (2010) FOXM1 is a transcriptional target of ERalpha and has a critical role in breast cancer endocrine sensitivity and resistance. *Oncogene*, **29**, 2983-2995.
124. Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353-357.
125. Chumsri, S., Sabnis, G.J., Howes, T. and Brodie, A.M. (2011) Aromatase inhibitors and xenograft studies. *Steroids*, **76**, 730-735.
126. Czajka-Oraniec, I. and Simpson, E.R. (2010) Aromatase research and its clinical significance. *Endokrynol Pol*, **61**, 126-134.
127. Loi, S., Symmans, W.F., Bartlett, J.M., Fumagalli, D., Van't Veer, L., Forbes, J.F., Bedard, P., Denkert, C., Zujewski, J., Viale, G. *et al.* (2011) Proposals for uniform collection of biospecimens from neoadjuvant breast cancer clinical trials: timing and specimen types. *The lancet oncology*, **12**, 1162-1168.
128. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, **14**, 1675-1680.
129. Gunderson, K.L., Kruglyak, S., Graige, M.S., Garcia, F., Kermani, B.G., Zhao, C., Che, D., Dickinson, T., Wickham, E., Bierle, J. *et al.* (2004) Decoding randomly ordered DNA arrays. *Genome Res*, **14**, 870-877.
130. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**, 1151-1161.

131. Barnes, M., Freudenberg, J., Thompson, S., Aronow, B. and Pavlidis, P. (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res*, **33**, 5914-5923.
132. Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M. and Nobel, A.B. (2008) Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, **24**, 1154-1160.
133. Sasik, R., Woelk, C.H. and Corbeil, J. (2004) Microarray truths and consequences. *J Mol Endocrinol*, **33**, 1-9.
134. Timlin, J.A. (2006) Scanning microarrays: current methods and future directions. *Methods Enzymol*, **411**, 79-98.
135. Affymetrix. *Affymetrix Data File Formats* Available from: http://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/index.html (accessed 30th August 2012)
136. Illumina. (2008) Genome Studio Gene Expression Module User Guide, v1. *Illumina propietary part # 11319121*.
137. Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307-315.
138. Du, P., Kibbe, W.A. and Lin, S.M. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**, 1547-1548.
139. Dunning, M.J., Smith, M.L., Ritchie, M.E. and Tavare, S. (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183-2184.
140. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*, **35**, D760-765.
141. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29**, 365-371.
142. Smyth, G. (2005) In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. and Huber, W. (eds.), *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, New York, pp. 397--420.
143. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.
144. Florido, J.P., Pomares, H., Rojas, I., Calvo, J.C., Urquiza, J.M. and Claros, M. (2009) In Cabestany, J. (ed.), *Bio-inspired systems : computational and ambient intelligence : 10th International Work-Conference on Artificial Neural Networks, IWANN 2009, Salamanca, Spain, June 10-12, 2009 : proceedings*. Springer-Verlag, Berlin, pp. 845–852.
145. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K. and Irizarry, R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, **11**, 733-739.
146. Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J. and Clarke, R.B. (2008) The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC Med Genomics*, **1**, 42.
147. Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118-127.

148. Barbosa-Morais, N.L., Dunning, M.J., Samarajiwa, S.A., Darot, J.F., Ritchie, M.E., Lynch, A.G. and Tavare, S. (2010) A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res*, **38**, e17.
149. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, **33**, e175.
150. Simon, R. (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol*, **23**, 7332-7341.
151. Dougherty, E.R., Hua, J. and Sima, C. (2009) Performance of feature selection methods. *Curr Genomics*, **10**, 365-374.
152. Dougherty, E.R. and Brun, M. (2006) On the number of close-to-optimal feature sets. *Cancer Inform*, **2**, 189-196.
153. Peng, Y., Wu, Z. and Jiang, J. (2010) A novel feature selection approach for biomedical data classification. *J Biomed Inform*, **43**, 15-23.
154. Chandra, B. and Gupta, M. (2011) An efficient statistical feature selection approach for classification of gene expression data. *J Biomed Inform*, **44**, 529-535.
155. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.
156. Cutler, A. and Stevens, J.R. (2006) Random forests for microarrays. *Methods Enzymol*, **411**, 422-432.
157. Miller, W.R., Larionov, A.A., Renshaw, L., Anderson, T.J., White, S., Murray, J., Murray, E., Hampton, G., Walker, J.R., Ho, S. *et al.* (2007) Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole. *Pharmacogenet Genom*, **17**, 813-826.
158. Miller, W.R., Larionov, A., Anderson, T.J., Evans, D.B. and Dixon, J.M. (2010) Sequential changes in gene expression profiles in breast cancers during treatment with the aromatase inhibitor, letrozole. *Pharmacogenomics J*.
159. Ressom, H.W., Varghese, R.S., Zhang, Z., Xuan, J. and Clarke, R. (2008) Classification algorithms for phenotype prediction in genomics and proteomics. *Frontiers in bioscience : a journal and virtual library*, **13**, 691-708.
160. Obulkasim, A., Meijer, G.A. and van de Wiel, M.A. (2011) Stepwise classification of cancer samples using clinical and molecular data. *Bmc Bioinformatics*, **12**, 422.
161. Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L. *et al.* (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**, 96.
162. Lusa, L., McShane, L.M., Reid, J.F., De Cecco, L., Ambrogi, F., Biganzoli, E., Gariboldi, M. and Pierotti, M.A. (2007) Challenges in projecting clustering results across gene expression-profiling datasets. *Journal of the National Cancer Institute*, **99**, 1715-1723.
163. Molinaro, A.M., Simon, R. and Pfeiffer, R.M. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301-3307.
164. Saldanha, A.J. (2004) Java Treeview--extensible visualization of microarray data. *Bioinformatics*, **20**, 3246-3248.
165. de Hoon, M.J., Imoto, S., Nolan, J. and Miyano, S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453-1454.
166. Maechler, M., Rousseeuw, P., Struyf, A. and Hubert, M. (2005), unpublished: cited as recommended by citation() finction in cluster R package.
167. Van der Laan, M., Pollard, K. and Bryan, J. (2003) A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, **73**, 575–584.

168. Larionov, A., Faratian, D., Caldwell, H., Sims, A.H., Fawkes, A., Murphy, L., Renshaw, L. and Dixon, J.M. (2009) Gene Expression Profiles of Endocrine Resistant Breast Tumours. . *Cancer Research*, **69**, 5132.
169. Miller, W.R., Larionov, A., Renshaw, L., Anderson, T.J., Walker, J.R., Krause, A., Sing, T., Evans, D.B. and Dixon, J.M. (2009) Gene expression profiles differentiating between breast cancers clinically responsive or resistant to letrozole. *J Clin Oncol*, **27**, 1382-1387.
170. Turnbull, A.K., Kitchen, R.R., Larionov, A.A., Renshaw, L., Dixon, J.M. and Sims, A.H. (2012) Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. *BMC Med Genomics*, **5**, 35.
171. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B. *et al.* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, **98**, 262-272.
172. Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J.A. *et al.* (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **25**, 1239-1246.
173. Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Ryder, K., Reid, J.F. *et al.* (2008) Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, **9**, 239.
174. Loi, S., Haibe-Kains, B., Majjaj, S., Lallemand, F., Durbecq, V., Larsimont, D., Gonzalez-Angulo, A.M., Pusztai, L., Symmans, W.F., Bardelli, A. *et al.* (2010) PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 10208-10213.
175. Symmans, W.F., Hatzis, C., Sotiriou, C., Andre, F., Peintinger, F., Regitnig, P., Daxenbichler, G., Desmedt, C., Domont, J., Marth, C. *et al.* (2010) Genomic index of sensitivity to endocrine therapy for breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **28**, 4111-4119.
176. Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H. *et al.* (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, **66**, 10292-10301.
177. Desmedt, C., Giobbie-Hurder, A., Neven, P., Paridaens, R., Christiaens, M.R., Smeets, A., Lallemand, F., Haibe-Kains, B., Viale, G., Gelber, R.D. *et al.* (2009) The Gene expression Grade Index: a potential predictor of relapse for endocrine-treated breast cancer patients in the BIG 1-98 trial. *BMC Med Genomics*, **2**, 40.
178. Ghazoui, Z., Buffa, F.M., Dunbier, A.K., Anderson, H., Dexter, T., Detre, S., Salter, J., Smith, I.E., Harris, A.L. and Dowsett, M. (2011) Close and stable relationship between proliferation and a hypoxia metagene in aromatase inhibitor-treated ER-positive breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **17**, 3005-3012.
179. Buffa, F.M., Harris, A.L., West, C.M. and Miller, C.J. (2010) Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British journal of cancer*, **102**, 428-435.

180. Shats, I., Gatza, M.L., Chang, J.T., Mori, S., Wang, J., Rich, J. and Nevins, J.R. (2011) Using a stem cell-based signature to guide therapeutic selection in cancer. *Cancer research*, **71**, 1772-1780.
181. Vivanco, I. and Sawyers, C.L. (2002) The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nature reviews. Cancer*, **2**, 489-501.
182. Ellis, M.J., Lin, L., Crowder, R., Tao, Y., Hoog, J., Snider, J., Davies, S., DeSchryver, K., Evans, D.B., Steinseifer, J. *et al.* (2010) Phosphatidyl-inositol-3-kinase alpha catalytic subunit mutation and response to neoadjuvant endocrine therapy for estrogen receptor positive breast cancer. *Breast Cancer Res Tr*, **119**, 379-390.
183. Creighton, C.J., Li, X., Landis, M., Dixon, J.M., Neumeister, V.M., Sjolund, A., Rimm, D.L., Wong, H., Rodriguez, A., Herschkowitz, J.I. *et al.* (2009) Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 13820-13825.
184. Kao, K.J., Chang, K.M., Hsu, H.C. and Huang, A.T. (2011) Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer*, **11**, 143.
185. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **27**, 1160-1167.
186. Bertucci, F., Borie, N., Ginestier, C., Groulet, A., Charafe-Jauffret, E., Adelaide, J., Geneix, J., Bachelart, L., Finetti, P., Koki, A. *et al.* (2004) Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene*, **23**, 2564-2575.
187. Sircoulomb, F., Bekhouche, I., Finetti, P., Adelaide, J., Ben Hamida, A., Bonansea, J., Raynaud, S., Innocenti, C., Charafe-Jauffret, E., Tarpin, C. *et al.* (2010) Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer*, **10**, 539.
188. Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L.H., Borg, A., Ferno, M., Peterson, C. and Meltzer, P.S. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Research*, **61**, 5979-5984.
189. Saal, L.H., Johansson, P., Holm, K., Gruvberger-Saal, S.K., She, Q.B., Maurer, M., Koujak, S., Ferrando, A.A., Malmstrom, P., Memeo, L. *et al.* (2007) Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 7564-7569.
190. Ertel, A., Dean, J.L., Rui, H., Liu, C., Witkiewicz, A.K., Knudsen, K.E. and Knudsen, E.S. (2010) RB-pathway disruption in breast cancer: differential association with disease subtypes, disease-specific prognosis and therapeutic response. *Cell Cycle*, **9**, 4153-4163.
191. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80.
192. Davis, S. and Meltzer, P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846-1847.
193. Bolstad, B. (2004) Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. *Dissertation*, University of California, Berkeley.
194. Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, New York.

195. Fedele, P., Calvani, N., Marino, A., Orlando, L., Schiavone, P., Quaranta, A. and Cinieri, S. (2012) Targeted agents to reverse resistance to endocrine therapy in metastatic breast cancer: Where are we now and where are we going? *Critical reviews in oncology/hematology*.
196. Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T. *et al.* (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13550-13555.
197. Ritchie, M.E., Dunning, M.J., Smith, M.L., Shi, W. and Lynch, A.G. (2011) BeadArray expression analysis using bioconductor. *PLoS Comput Biol*, **7**, e1002276.
198. Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S. *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 8418-8423.
199. Calza, S., Hall, P., Auer, G., Bjohle, J., Klaar, S., Kronenwett, U., Liu, E.T., Miller, L., Ploner, A., Smeds, J. *et al.* (2006) Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast cancer research : BCR*, **8**, R34.
200. Chia, S.K., Bramwell, V.H., Tu, D., Shepherd, L.E., Jiang, S., Vickery, T., Mardis, E., Leung, S., Ung, K., Pritchard, K.I. *et al.* (2012) A 50-Gene Intrinsic Subtype Classifier for Prognosis and Prediction of Benefit from Adjuvant Tamoxifen. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **18**, 4465-4472.