Gene expression data analysis for identifying crucial gene markers and subtype classification in breast cancer

A thesis submitted in partial fulfilment

of the requirements for the degree of

Master of Applied Science

at

Lincoln University

by

Aiman Hassan Almazroey

Lincoln University

2012

Abstract & Keywords:

An abstracts of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Applied Science (M.Appl.Sc.)

Gene expression data analysis for identifying crucial gene markers and subtype classification in breast cancer

By

Aiman Almazroey

Breast cancer is one of the leading causes of death in women. Even with advances in early-stage breast cancer treatment, physicians are still lacking the ability to precisely predict which patients would benefit from adjuvant chemotherapy. Gene expression profiling studies have been used to provide us with insights into the heterogeneity of breast cancer. Therefore, patterns of gene activity can be identified by genome-wide measures of gene expression to subclassify tumours. This might provide a better means, than is currently available, of treating patients with breast cancer, and to help physicians find the accurate treatment (Wang, Klijn et al. 2005). This study uses a combination of three different clustering methods: Hierarchical clustering, Self-organizing maps (SOM) and Ward method to further explore and validate the characters of previously identified, novel, 306 intrinsic genes thought to discriminate five types of breast cancer (LumA, LumB, Normal-like, Basal-like and Her2). It is also used to derive improved cluster characterisations for accurate subtype identification from independent gene expression data analyses. Implementation of these methods, in widespread clinical practice at present, remains limited. From an exploratory pilot study in this research, it was found that one or more of the few most highly active genes in one subtype can be active in one or more other subtypes indicating that several gene markers are essential for subtype discrimination. Nevertheless, this study identified one or two potential genes for some

subtypes that may be useful as markers in their identification. In the main part of the investigation, the originally selected whole gene set was assessed for their efficacy in subtype discrimination using Hierarchical clustering and SOM, both in conjunction with Ward method that indicates the optimum number of clusters (subtypes). Hierarchical-Ward method found 6 clusters and SOM-Ward method found 7 clusters as optimum compared to the 5 subtypes reported by the original authors from whose work the gene set used in this study was extracted. This indicated the heterogeneity of subtypes. In both methods, second optimum number of clusters was 2. Our clusters revealed interesting results: for example, closer examination of the 2 cluster structure from both Herarchical-Ward and SOM-Ward indicated that 3 subtypes (LumA, LumB and Normal-like) always cluster together and the other 2 subtypes (Basal and Her2) make up the second cluster.

The six and 7 optimum clusters from the two respective methods did indicate that most clusters contain patients from more than one subtype and revealed which subtypes are more likely to cluster together. These results indicate subtype overlap. Although not featured highly in SOM-Ward results, the 6 cluster format from this method was explored to compare with Herarchical-Ward results and outcomes from the two methods were identical. This gives validity to the results in the study. Interestingly, two out of the 7 clusters from SOM consisted of only one subtype each (LumA or Basal-like), and 1 out 6 clusters from Heirarchical-Ward also contained only one subtype (LumA); but these clusters did not contain all of the patients originally thought to be belonging to each particular subtype. However, these clusters containing only one subtype each may indicate the core behaviour of the respective subtype and is worth exploring further. The results overall points to the complexity of discriminating the subtypes due to their heterogeneity and overlap, when viewed through the selected set of 306 intrinsic genes; this study has shed light on these characteristics in a reliable and predictable way.

Keywords: gene expression, gene inhibition, clusters, Hierarchical clustering, Self-organizing maps (SOM), Ward method, intrinsic genes, LumA, LumB, Normal-like, Basal-like and Her2.

Acknowledgements

I would like to extend my deepest thanks and appreciation to my supervisor Associate Professor Sandhya Samarasinghe for her valuable guidance and patience. I would also like to thank my co- supervisor Prof. Don Kulasiri for his help and assistance. I would like to thank the Saudi Arabian Ministry of Higher Education for granting me with the scholarship to do my Master degree at Lincoln University – New Zealand. I express my gratitude to my wife Alaa Almazroey for her valuable support during the course of my study; and to my parents for their moral support during my stay away from home. I extend my thanks to all my fellow students who offered me support and help during my study. Finally, I would like to convey my warm thanks to all staff and administrative secretaries of the Centre for Advanced Computational Solutions (C-fACS) and Department of Environmental Management, for their lovely reception, help and sense of humor.

Contents

List of Figures

List of Tables

Chapter One

1 BACKGROUND

Breast cancer is a disease in which some cells multiply without control to form an abnormal mass. The most common form of breast cancer originates in the cells that line the ducts that carry milk to the nipple (ductal carcinoma). Other forms of breast cancer originate from the glands that produce milk (lobular cancer) or from other parts of the breast. Breast cancer is a common cause of death in women in New Zealand and worldwide (Ref. Global Cancer Statistics). Although Breast cancer occurs in both men and women, statistics show that this disease occurs 100 times more frequently in women than in men (Sasco, Lowenfels et al. 1993).

Globally, breast cancer incidence rates are highest in North America and northern Europe although incidence rate is decreasing in these parts of the world. The incidence rate is usually considered the lowest in Asia and Africa but there has been an increase in the rate in the recent years. These international differences are thought to be related to societal changes occurring during industrialization (e.g. changes in fat intake, body weight, age at menarche, and/or lactation, and reproductive patterns such as fewer pregnancies and later age at first birth) (Parkin, Bray et al. 2005). Studies of increasing breast cancer among first-generation daughters of Japanese Americans suggest that environmental and lifestyle factors are of greater significance than are genetic factors in explaining international differences in breast cancer risk (Johnson 2001; Deapen, Liu et al. 2002; Lacey, Devesa et al. 2002). Since 1975, mortality rates of Breast cancer have declined. These were because of the improvements of adjuvant therapies and attributed to the increased use of screening mammography (Berry, Cronin et al. 2005).

1.1 Aetiology and risk factors

The frequency of this disease in women has prompted an intensive study of risk factors for developing breast cancer to gain clues as to its aetiology as well as to identify modifiable risk factors that would be helpful for prevention strategies. Risk factors of breast cancer include age and gender. These are considered as one of the strongest risk factors (Jemal, Siegel et al. 2010). It is rarely found before the age of 25 except in some certain familial cases. The incidence rises throughout women's lifetime (Robbins and cotrans), Race and ethnicity is another factor. In the US, although breast cancer is common in women of every major ethnic group there are interracial differences (Pike, Spicer et al. 1993; Peto and Mack 2000; Bradley, Given et al. 2002; Palmer, Wise et al. 2003; Jatoi, Chen et al. 2007; Jemal, Thun et al. 2008; Society 2009 - 2010). Much of these ethnic differences are attributable to factors associated with lifestyle and socioeconomic status (e.g. access to diagnosis and treatment).

Several lifestyle factors including body size, physical activity and dietary factors can contribute to the risk of breast cancer (van den Brandt, Spiegelman et al. 2000; Morimoto, White et al. 2002; Feigelson, Jonas et al. 2004; Lahmann, Hoffmann et al. 2004; Eliassen, Colditz et al. 2006; Ahn, Schatzkin et al. 2007).

Reproductive and hormonal factors such as prolonged exposure to higher concentrations of endogenous estrogen increases the risk of breast cancer (Kelsey, Gammon et al. 1993; Parkin, Bray et al. 2005).

Family history is an important risk factor for breast cancer where strong genetic mutations with little relation to environmental differences have been reported (Lichtenstein, Holm et al. 2000; Peto and Mack 2000; Ahn, Schatzkin et al. 2007).

Exposure to ionizing radiation of the chest at a young age, as occurs with radiotherapy treatment or in survivors of atomic bomb or nuclear plant accidents, is associated with an increased risk of breast cancer (Guibout, Adjadj et al. 2005; Pukkala, Kesminiene et al. 2006; Ostroumova, Preston et al. 2008).

1.2 Pathology

More than 95 percent of breast malignancies arise from the breast epithelial elements and are therefore carcinomas. The term "breast carcinoma" encompasses a diverse group of lesions that differ in microscopic appearance and biologic behaviour, although these disorders are often discussed as a single disease.

The invasive breast carcinomas consist of several histological subtypes. These include infiltrating ductal (76 %), invasive lobular (8 %), ductal/lobular (7%), mucinous (colloid) (2.4%), tubular (1.5%), medullary (1.2%), and papillary (1%).

1.3 Symptoms and Signs

Most women with symptomatic rather than screen-detected breast cancer present with a painless increasing mass which may also be associated with nipple discharge, skin tethering, ulceration and, in inflammatory cancers, oedema and erythema. In developing countries, 80% are likely to be present with advanced disease and metastases. (Clark 2009)

Breast cancer can also present as a subtle change in the general appearance of the breast, such as an increase or decrease in size, a change in symmetry. There are conditions that may resemble the symptoms of breast cancer such as those related to menstrual cycles (cyclic or non-cyclic), and whether these symptoms are aggravated or alleviated by any activities or medications (Blake Cady 1998; Monica Morrow 2000).

1.4 Diagnostic investigations

The diagnostic evaluation of a patient suspected of having breast cancer includes screening and diagnostic breast imaging and breast biopsy. The majority of breast cancers are diagnosed as a result of an abnormal mammogram, but not all mammographic findings represent cancer (Stomper 2000; Ostroumova, Preston et al. 2008). It has been reported that fewer than 10 percent of cancers can be detected solely by physical examination and over 90 percent mammographically (Ostroumova, Preston et al. 2008). Positive mammographic findings mainly represent soft tissue masses and microcalcifications.

Contrast-enhanced breast magnetic resonance imaging (MRI) may complement mammographic staging, because the latter is more sensitive and is assumed that it would estimate the extent of disease more accurately than conventional imaging. However, the MRI may sometimes produce falsely positive result and can delay treatment by necessitating additional biopsies (Berg, Gutierrez et al. 2004; Bleicher, Ciocca et al. 2009).

Targeted ultrasonography is a useful diagnostic test to evaluate a palpable mass or an area of abnormality detected on mammogram. It is particularly useful for assessing whether a mass is solid or cystic in nature.

Chest radiography, abdominal/pelvic computed tomography (CT) scanning, and bone scans are other techniques commonly used in evaluating different body parts for the presence of metastatic disease.

Abnormal screening mammogram often needs further diagnostic evaluation with magnification views, spot compression views, targeted ultrasonography, tissue sampling or biopsy (Barlow, Lehman et al. 2002). Biopsy helps obtaining sufficient diagnostic material using the least invasive approach and to avoid surgical excision of benign lesions (Barlow, Lehman et al. 2002).

Fine needle aspiration (FNA) is a classical method to determine the histopathological features of breast lumps by obtaining samples from cellular lesions and metastatic lymph nodes. A major advantage of FNA is that it can be easily and quickly performed at the time of a diagnostic study, with potential for an immediate preliminary interpretation (Pisano, Fajardo et al. 2001). However, the FNA has some disadvantages and produce false negative results in inexperienced hands (Pisano, Fajardo et al. 2001).

Core needle biopsy (CNB) is a procedure that removes small but solid samples of tissue using a hollow "core" needle that has a special cutting edge. Compared to FNA, CNB offers a more definitive histological diagnosis, avoids inadequate samples and may permit the distinction between invasive versus in situ cancer (Verkooijen 2002).

Tumour markers are used to assess the estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status, and should be made on every primary invasive breast cancer to identify patients likely to benefit from endocrine and/or anti-HER2 therapy (Verkooijen 2002).

1.5 Treatment

1.5.1 Early breast cancer

The treatment of early breast cancer includes the treatment of locoregional disease with surgery, radiation therapy, or both, and the treatment of systemic disease with one or a combination of chemotherapy, endocrine therapy, or biologic therapy. The need for, timing, and selection of therapy are based upon tumour variables such as histology, stage, and tumour markers; patient variables such as age, menopausal status, and comorbid conditions; as well as patient preference, such as a desire for breast preservation (Clark 2009).

1.5.2 Neoadjuvant systemic therapy

Neoadjuvant systemic therapy prior to local treatment may be required to improve surgical outcomes and options. For operable breast cancer, the aim is to increase the chance of breast conserving surgery in patients who would otherwise require mastectomy. For inoperable locally advanced breast cancers, the aim is to achieve operability.

Surgery as a local treatment may vary from wide local excision or segmental mastectomy and breast conservation for masses < 4 cm in diameter, to simple mastectomy with or without reconstruction. The choice is dictated by the location and extent of the breast mass in relation to the breast size, and patient preferences (Clark 2009). Breast cancer surgery has changed dramatically over the past several decades and continues to evolve (Cuzick, Stewart et al. 1994).

Radiotherapy is another local treatment that is given to the conserved breast after wide local excision to reduce local recurrence and to the chest wall after mastectomy if there are risk factors such as proximity to surgical margins or lymph node metastases, to complete the local control measures.

Adjuvant systemic therapy is also a local treatment that refers to the administration of endocrine therapy, chemotherapy, and/or biologic therapy after definitive local therapy for breast cancer. Adjuvant systematic therapy has made a significant advancement in breast cancer patients with regards to disease-free survival and overall survival. (Carlson 2005; Goldhirsch, Glick et al. 2005). Using computer-based software, Adjuvant, is proved to be helpful in objectively estimating survival and benefit from adjuvant therapy, for early stage breast cancer patients, based on the clinico pathlogic features mentioned above (Ravdin, Siminoff et al. 2001). However, Adjuvant does not benefit very young women (Olivotto, Bajdik et al. 2005). Although, Adjuvant is useful in predicting the outcome for an individual patient when the clinico pathologic features are known, those clinic-pathologic features do not take into account the biological complexity of an individual's tumour. Therefore, more reliable and precise prognostic and predictive models are needed to estimate what Adjuvant therapy should be offered to an individual patient.

1.5.3 Advanced breast cancer

Patients with established metastatic (i.e. advanced) disease may require endocrine therapy, chemotherapy and radiotherapy. The treatment is not curative but may be of great palliative benefit and consistent often with many years of good-quality life.

High levels of oestrogen receptors (ER) and progesterone receptors (PR) in their tumour have greater chance of responding to endocrine treatments. Endocrine therapy is usually tried first in patients with characteristics indications to respond and who do not have immediately life-threatening disease (Clark 2009). On the other hand, chemotherapy is used for patients who are unlikely to respond to hormonal treatment or who fail to respond to endocrine therapy or who require a rapid response if at risk such as in liver or respiratory failure cases. If chosen carefully, chemotherapy can provide good-quality palliation and prolongation of life (Clark 2009).

1.6 Molecular intrinsic subtypes of breast cancer

Breast cancer is a heterogeneous and phenotypically diverse disease. Breast tumours are characterised by cellular and molecular heterogeneity and large number of genes potentially involved in controlling cell growth, death, and differentiation. Some cancers such as Lymph node metastases (Fisher, Costantino et al. 1993), histologic grade (Elston and Ellis 1991), expression of steroid and growth factor receptors (Vollenweiderzerargui, Barrelet et al. 1986; Torregrosa, Bolufer et al. 1997), estrogen-inducible genes like cathepsin D (Foekens, Look et al. 1999), protooncogenes like ERBB2 (Slamon, Godolphin et al. 1989), and mutations in the TP53 gene (Bergh, Norberg et al. 1995; Borresen, Andersen et al. 1995) have been correlated to prognosis. However, individual prognostic factors provide limited information about the cellular changes induced by the disease. For example, removal of ovaries used to be considered as therapeutic in the past, which was successful in some patients, but not others. This shows that the prognostic value of many of these parameters may mislead the interpretation (Howat, Barnes et al. 1983; Battaglia, Scambia et al. 1988). It has been shown that correlating tumour cell gene's cDNA microarrays with specific features of phenotypic variation might provide better understating of the taxonomy of cancer (Golub, Slonim et al. 1999; Alizadeh, Eisen et al. 2000; Perou, Sorlie et al. 2000; Hedenfalk, Duggan et al. 2001). These observations emphasize the importance of studying genetic basis of this disease to gain a better understanding.

Characterization (profiling) of breast cancers has advanced significantly since the turn of the millennium due to the development of sophisticated technologies. These include gene expression arrays, which permit simultaneous measurement of thousands of genes to create a molecular portrait of the tumour. Gene-expression profiling has helped in identifying breast cancer molecular subtypes that have distinct behaviour and response to therapy; and in the development of prognostic and predictive molecular signatures. These in turn, have resulted in a better understanding of the biological heterogeneity of breast cancer. Microarray analyses on breast cancers have identified gene expression profiles able to separate tumour classes associated with patient survival (Sorlie, Tibshirani et al. 2003; Wang, Klijn et al. 2005).

Several studies on genome-wide expression patterns in cancer types, including lymphoma, breast, lung, liver, ovarian and soft tissue sarcomas have been carried out over the years. A common feature of these studies has been the emergence of tumour subtypes with distinct gene expression patterns for each subtype. The differences in gene expression patterns among these subtypes are likely to reflect basic differences in the cell biology of the tumours (Sorlie, Tibshirani et al. 2003; Sotiriou, Neo et al. 2003). Gene expression profiling, using microarrays to which cDNA or oligonucleotide probes are affixed, allows simultaneous measurement of the activity (expression) of thousands of genes in a breast cancer cell.

Gene expression arrays represent the level of expression of a group of genes in a semi-quantitative manner by comparing the level of messenger RNA (mRNA) to that of the mRNA of the same gene from a reference sample.

The two main types of molecular profiling techniques commonly used in the laboratory are “supervised” and “unsupervised” analyses. In supervised analyses, the gene sets are designed to differentiate tumours by a defined clinical endpoint; as such, the subtypes of cancer can be identified based exclusively on the clinical data. Prognostic molecular profiles (prognosis signatures) are examples of supervised analyses (Bair and Tibshirani 2004).

Unsupervised analysis (clustering) permits examination of gene expression patterns regardless of clinical endpoints and reflects inherent biologic differences. This approach does not use any of the clinical information about the patient and therefore the subgroups are identified only by using the gene expression data (Bair and Tibshirani 2004).

In addition, there is the semi-supervised analysis which combines both gene expression data and clinical data. Clinical data is used to identify a list of genes that correlate with the clinical variables of interest and then unsupervised clustering techniques are applied to this subset of the genes (Bair and Tibshirani 2004). An example for the semi-supervised is the "intrinsic" subtypes.

1.6.1 **Intrinsic subtypes of breast cancer**

The intrinsic subtypes can generally be classified into two groups in relation to the expression of hormone receptor-related genes (Sorlie, Perou et al. 2001). Gene expression studies have identified several distinct breast cancers within these intrinsic subtype groups. The main subtypes are as follow: estrogen receptor (ER)-negative tumours: basal-like, human epidermal growth factor receptor-2 (HER2)-enriched and normal-like; and ER-positive tumours: luminal A and luminal B. A sixth breast cancer subtype, termed Claudine-low, has also been identified (Carey, Dees et al. 2007) as well as a Luminal C subtype (Sorlie et at. 2001). These subtypes differ markedly in prognosis and in the therapeutic targets they express. It is agreed that the first three are the major subtypes (Carey, Dees et al. 2007). The list of genes that differentiates the different subtypes is called the intrinsic list. It is made up of several clusters of genes relating to ER expression.

It has been found that ER-positive and ER-negative are two biologically different breast cancers and may derive from different progenitor cells.

1.6.1.1 The luminal cancers, A and B

These are so called because they are characterized by expression of genes similar to that expressed by normal breast luminal epithelial cells, and may overlap with ER-positive breast cancers. They typically express luminal cytokeratins 8 and 18. These are the most common subtypes that make up the majority of ER-positive breast cancers, and are characterized by expression of ER, PR, and other genes associated with ER activation.

Luminal A and luminal B have some important molecular and prognostic distinctions.

- Luminal A tumours, which probably make up about 40 percent of all breast cancers, usually have high expression of ER-related genes, low expression of the HER2 cluster of genes, and low expression of proliferation-related genes (Fan, Oh et al. 2006; Hu, Fan et al. 2006). Luminal A tumours carry the best prognosis of all breast cancer subtypes (Voduc, Cheang et al. ; Sorlie, Perou et al. 2001; Sorlie, Tibshirani et al. 2003; Sotiriou, Neo et al. 2003; Loi, Haibe-Kains et al. 2007).
- The less common (about 20 percent) luminal B tumours have relatively lower expression of ER-related genes, variable expression of the HER2 cluster, and higher expression of the proliferation cluster. Luminal B tumours carry a worse prognosis than luminal A tumours (Voduc, Cheang et al.) and with poor prognostic of 70-genes signature (Fan, Oh et al. 2006). Most luminal B cancers have high Recurrence Scores.

1.6.1.2 HER2-enriched

The HER2-enriched subtype (previously known as HER2+/ER-) makes up about 10 to 15 percent of breast cancers. It is characterized by high expression of the HER2, proliferation gene clusters, and low expression of the luminal cluster. For this reason, these tumours are typically negative for ER and PR, and positive for HER2. It is important to note that this subtype comprises only about half of clinically HER2-positive breast cancer. The other half has high expression of both the HER2 and luminal gene clusters and fall in a luminal subtype. In the era before HER2-targeted therapy, this subtype carried a poor prognosis (Voduc, Cheang et al.). This adverse natural history has been markedly affected by therapeutic advances in HER2-directed therapy.

1.6.1.3 Basal-like

The basal-like subtype makes up about 15 to 20 percent of breast cancers. It is so called because of some expression similarities to that of the basal epithelial cells. It is characterized by low expression of the luminal and HER2 gene clusters. For this reason, these tumours are typically ER-, PR-, and HER2-negative on clinical assays. It is therefore, nicknamed "triple-negative" (Olopade and Grushko 2001; Sorlie, Tibshirani et al. 2003; Foulkes, Brunet et al. 2004).

Basal-like breast cancer has unique risk factors. The most intriguing is that over 80 percent of cancers arising in women born with a mutation in the breast cancer gene 1 (BRCA1), early onset gene, are basal-like (Olopade and Grushko 2001; Sorlie, Tibshirani et al. 2003; Foulkes, Brunet et al. 2004). Nevertheless, most basal-like breast cancers are sporadic, and the BRCA1 gene and protein appear intact in these tumours. A commonly held assumption, but unproven, is that the BRCA1 pathway is abnormal in sporadic basal-like breast cancer. This may have therapeutic implications since this pathway is important in DNA repair.

There is a notable association between the basal-like subtype, race and age. Population-based studies suggest that the basal-like subtype is overrepresented in breast cancer in African-American women, and developing during the premenopausal years (Fan, Oh et al. 2006; Millikan, Newman et al. 2008; Lund, Trivers et al. 2009; Parker, Mullins et al. 2009). This may explain its prevalence among young African-American women. Basal-like breast cancer is known to have poor prognosis; and this may contribute in part to the worse outcomes experienced by African-American women with breast cancer (Voduc, Cheang et al.). Fortunately, studies suggest that basal-like breast cancers are sensitive to modern chemotherapy (Rouzier, Perou et al. 2005; Carey, Dees et al. 2007). However, the absence of targeted therapy, of known effectiveness, to this subtype remains a real obstacle to improving outcomes.

1.6.1.4 Claudin-low

This newly described subtype is found in non-basal triple-negative breast cancers, which is uncommon but interesting because of its expression of epithelial-mesenchymal transition genes and characteristics reminiscent of stem cells.

This subtype comprises a minority of triple-negative breast cancers. It is characterized by low to absent expression of epithelial cell-cell adhesion genes (claudin 3, 4 and 7, E-cadherin), differentiated luminal cell surface markers (EpCAM and MUC1) and enrichment for epithelial-to-mesenchymal transition markers, immune response genes and cancer stem cell-like features (CD44+/CD24-, high ALDH1A1) (Carey, Dees et al. 2007). In contrast to basal-like breast cancers, claudin-low tumours appear to be slower growing and, with features of mesenchymal and mammary stem cells, of different oncogenic origin.

1.6.1.5 Normal-like

This subtype was one of the first subtypes identified by gene expression arrays and consistently appears in breast cancer clusters. It is typified by similar gene expression pattern as normal breast cells, and thus remains enigmatic as to whether it represents a separate subtype or a technical artefact introduced by low tumour cell composition of the sampled specimen.

The challenge is how to define the intrinsic subtypes of breast cancer. They have been identified using cluster analysis of gene expression data from frozen banked tissue. This method uses fixed tissues and has poor reproducible classification. Recently, a new assay (PAM50) has been developed. It is a 50-gene reverse transcription-polymerase chain reaction (RT-PCR) based method, and is derived from a 500-gene intrinsic list. It can be performed on fixed tissue but has not yet been sufficiently validated for clinical use (Parker, Mullins et al. 2009). Others methods include immunohistochemical surrogates for the intrinsic subtypes. These are reasonable but also have not been validated, and the simplest schema using clinical assays for ER, progesterone receptor (PR), and HER2 misclassify a significant proportion of tumours (Carey, Perou et al. 2006).

1.6.2 Prognosis of intrinsic subtypes

The intrinsic subtypes were developed to identify relevant biology, not for prognosis. However, in multiple independent datasets, these subtypes correlate with prognosis. In general, patients with the luminal A subtype have the best prognosis; patients with the other major hormone receptor-positive subtype, luminal B, suffers a significantly worse outcome. In a population-based study of nearly 500 tumours using immunohistochemical proxies for the subtypes, the best outcome was observed among patients with luminal A tumours compared with the other subtypes, and the worst outcome was seen among basal-like breast cancers (Carey, Perou et al. 2006). Both the basal-like and HER2-enriched subtypes have the worst survival chances, at least until recently. The HER2-targeting has altered the outcome for the HER2-enriched subtype and HER2-positive luminal cancers.

In summary, clustering or unsupervised classification approaches are used to build groups of genes, with related expression patterns, and separate genes into highly similar groups without predefined group labels. Clustering is a common method of gene expression data analyses. Most researchers use clustering methods based on similarity/dissimilarity, distance measure and hierarchical clustering. Several studies have concluded that using gene-expression profiling would be potentially crucial as a new prognostic and predictive tool. However, validation of the clinical use of this technology, at the moment, is still a major challenge for microarray studies, particularly with clinical implications.

The conflicting number of clusters reported by different studies may be due to the number of samples analysed and/or the method of analysis in identifying the overlapping between cluster classes. Almost all previous studies applied Hierarchical clustering method, and no attempts have been made to use alternative clustering methods that might be more helpful in clearly identifying the clusters, and the present study aimed at exploring the possibility of using combined clustering analysis methods to enhance the identification accuracy.

Chapter Two

2 OBJECTIVES OF THE STUDY

Breast cancer patients with the same diagnostic and clinical prognostic profile can have markedly different clinical outcomes. Classification of breast cancer based on gene expression profiling captures the molecular complexity of tumours. Patterns that distinguish subtypes further can provide a more refined stratification of the patients compared to individual tumour markers Hu et al., (2006). Current classifications assign a patient's gene expression signature to a single class (i.e. subtype). This means that all patients in a subtype should show similar expression patterns. Although this has been considered generally true, it is now recognised that a reason for relapse or an adverse response to therapy is that some patients normally do not clearly show typical symptoms of one subtype. Such patients may be on the borderline of a subtype, may show symptoms of more than one subtype, or there are other subtypes currently unknown.

This study aims at further exploring the characters of novel breast cancer genes (intrinsic genes) that have been previously identified and to refine the classification of breast cancer subtypes through advanced gene expression data analyses. The primary aim of this study is to do a new analysis by combining two known different clustering methods; (i.e. Hierarchical clustering and Self-organizing maps (SOM)). The data used in the current analysis method were those reported by Hu et al (2006). The training data contains 259 samples (patients) representing five subtypes of cancer (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like); and the 306 genes identified by Hu et al., (2006) and used to obtain a corresponding set of five class centroids using hierarchical clustering (Table 3.1).

The proposed new approach may help: (i) Study characteristic spread of gene expression patterns within a predefined subtype. This will provide more clarity into the variability within a subtype. (ii) Identify similarity/closeness and/or overlapping of subtypes which may help identify patients who cannot be clearly categorised into a specific subtype. This can help in choosing the appropriate treatment.

Chapter Three

3 MATERIALS AND METHOD

Methods used previously by four main investigating groups were adopted as a base line for the analysis in the present study. A brief description of the work of these groups is as follows:

Sorlie et al (2001) studied 85 tissue samples in which tumour specimens contained more than 50% tumour cells and the mRNA microarrays for those cells were processed. They selected 456 cDNA clones from an 8,102 intrinsic gene list as representatives for Basal-like, ERBB2+, Normal Breast-like, Luminal Subtype C, Luminal Subtype B and Luminal Subtype A. Those genes were with significantly greater variation in expression between different tumours than between paired samples from the same tumour. This measure was taken to represent inherent properties of the tumours themselves rather than just differences between different samplings. They concluded that relating gene expression patterns to clinical outcomes is a key issue in understanding the biological diversity of the tumours, and this will improve the chances of choosing the suitable type of treatment.

Later, Sorlie et al., (2003) conducted a study to further refine the previously defined (Sorlie et al., 2001) subtypes of breast tumours that have distinct patterns of gene expression. They used Hierarchical (unsupervised) clustering analysis based on expression pattern of 534 intrinsic genes. These genes were selected on the basis of the similarity of the expression level. The results of the study supported the concept that the studied breast tumour subtypes represent biologically distinct disease entities.

Van't Veer et al. (2002) used DNA microarray analysis on primary breast tumours of 117 patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases in patients without tumour cells in local lymph nodes at diagnosis.

Approximately 5,000 genes were selected out of 25,000 genes because they are significantly regulated in more than 3 tumours out of 78. These selected genes were clustered on the basis of their similarities. They suggested that the tumours can be divided into two types on the basis of this set of significant genes. The first type is a good prognostic (i.e. related to low metastasis risk) and the second type is a poor prognostic (i.e. associated with a high metastasis risk). Their finding of separating tumour classes associated with patient survival supported those previously reported by Sorlie et at. (2001), but they extended the work to gene expression (or genetic profile) associations with survival in an untreated, node negative cohort. In their work to identify reliable good and poor prognostic tumours they used a powerful three-step supervised classification method, similar to those used previously by (Gruvberger, Ringnér et al. 2001; He and Friend 2001; Khan J 2001). The correlation coefficient of the expression for each gene with disease outcome was calculated. It was found that 231 genes were significantly associated with disease outcome. The 231 genes were then rank-ordered on the basis of the magnitude of the correlation coefficient. This was followed by optimizing the number of genes in the 'prognosis classifier' by sequentially adding subsets of 5 genes from the top of this rank-ordered list and evaluating its power for correct classification using the 'leave-one-out' method for cross-validation. The accuracy improved until the optimal number of marker genes was reached, which was 70 genes. Their results showed that the gene expression profile outperformed the available clinical parameters, at the time, in predicting disease outcome. Their findings provided a strategy to select patients who would benefit from adjuvant therapy.

Sotiriou et al. (2003) selected 99 patients out of 700 patient population and the overall survival for this group was adjusted for standard prognostic factors of tumour size and nodal status comparable to the population. They conducted a study of comprehensive gene expression patterns generated from cDNA microarrays obtained with a 7,650-feature microarray (using unsupervised hierarchical clustering approach) and correlated that with detailed clinico-pathological characteristics and clinical outcome in the group of 99 node-negative and node-positive breast cancer patients. Gene expression patterns were found to be strongly associated with estrogen receptor (ER) status and to be the most important discriminator of expression subtypes, but moderately associated with grade, and not associated with menopausal status, nodal status, or tumour size. Hierarchical cluster analysis segregated the tumours into two main groups based on their ER status, which correlated well with basal and luminal characteristics. Their findings were in strong agreement with the findings of Perou et at. (2000), Sorlie et at. (2001) and Van't Veer et al. (2002) that the ER biology plays a central role in breast carcinogenesis defining the configuration of the final tumour.

The conditions and outcomes of the studies mentioned above have been used by Hu et al. (2006) who conducted breast cancer microarray analyses and subtype clarification study on a set of 259 breast cancer samples, represented by 306 expressed genes. This dataset was collected from the above four different studies (Table 3.1). These are ‘Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications’ (Sorlie, Perou et al. 2001), ‘Gene expression profiling predicts clinical outcome of breast cancer’ (van 't Veer, Dai et al. 2002), ‘Repeated observation of breast tumour subtypes in independent gene expression datasets' (Sorlie, Tibshirani et al. 2003) and 'Breast cancer classification and prognosis based on gene expression profiles from a population-based study'(Sotiriou, Neo et al. 2003).

The collected data was pooled and analysed by Hu et al. (2006) to define intrinsic subtypes by utilising pathological/clinical data as well as gene expression data. They first evaluated the datasets independently, and then combined them to find 306 breast genes which were analysed using the hierarchical clustering method. Hu et al. (2006) created a single data table from these four sets as follow:

i. Firstly, identifying the common genes present across all four microarray datasets (2800 genes).

ii. Secondly, Using Distance Weighted Discrimination (DWD) to combine these four datasets (Benito, Parker et al. 2004).

    - DWD is a multivariate analyses tool that performs statistical corrections to reduce systematic biases resulting in separate datasets. It then makes a global adjustment to compensate for these biases: in essence, each separate dataset is a multi-dimensional cloud of data points. It takes two point clouds and shifts one, such that it more optimally overlaps the other.

iii. Finally, they determined that 306 of the 1300 unique intrinsic genes were present in the combined test set and performed a hierarchical clustering analysis.

Table 3.1: The combined datasets of Sorlie et al. (2001; 2003), Van't Veer et al. (2002) and Sotiriou et al. (2003) arranged by Hu et al (2006).

Microarray type	*Validation*	*Informative genes*	*Metastasis determinants*	*References*
cDNA	Independent datasets	306 genes	Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like	Hu et al., (2006)
cDNA	Independent training set	99 genes	invasive ductal carcinomas; 46 individuals were node negative and 53 were node positive	**Sotiriou,** et al., (2003).
cDNA	Crossvalidation	534 intrinsic genes	Repeated finding in independent datasets	Sorlie et al., 2003
Oligonucleotide	Independent training set	70 genes	'Good signature' is related to low metastasis risk; a 'poor signature' is associated with a high metastasis risk.	Van't Veer et al., (2002)
cDNA		456 'intrinsic' genes	'Luminal A' tumours have a better outcome than 'luminal B' tumours. Worst outcome is for 'basal-like' and 'ERBB2+' tumours	Sorlie et al., (2001)

The work of Hu et al., (2006) identified the five main subtypes corresponding to the previously defined HER2+/ER-, Basal-like, LumA, LumB and Normal Breast like tumour groups (Perou, Sorlie et al. 2000; Sorlie, Perou et al. 2001). This type of analysis provides more statistical power to perform multivariate analyses. It would also yield more meaningful results because any findings would need to be shared/present across all four datasets.

The finding of Hu et al., (2006) is considered as the starting point for this proposed study. The Matlab 5 Toolboxes used to investigate the datasets in this study are: SOM Toolbox for SOM and Statistics Toolbox for both Ward and Hierarchical clustering. SOM Toolbox can be downloaded free of charge under the GNU General Public License from http://www.cis.hut.fi/projects/somtoolbox.

3.1 Methods of analyses

In this study, after a pilot study to explore the most highly influential genes (potential markers) in each subtype, we used a new approach of analyses by combining two pairs of analyses methods (Hierarchical-Ward and SOM-Ward clustering) to:

- Diagnostic markers for each subtype (pilot study).
- Validate the clusters found by the original authors Hu et al., (2006).
- Find relationship or similarity between subtypes to identify subtypes that can mask each other.
- Identify possible new cancer subtypes.

Three specific methods are going to be used in this study in an in-depth investigation of tumour subtypes: Hierarchical clustering, Self-organizing maps (SOM) and Ward method and with these three methods, it will be possible to display differences more clearly between the subtypes than using only one method. These methods are employed after an initial pilot study involving the most significant genes. A summary of these methods is as follows:

3.1.1 Ward Method

In Statistics, Ward method uses an analysis of variance approach to evaluate the distances between clusters. Ward's minimum variance method is a special case of the objective function approach originally presented by (Ward 1963). Ward suggested a general agglomerative hierarchical clustering procedure that, in general, is regarded as very efficient; however, it tends to create clusters of small size, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. To illustrate the procedure, Ward used the example where the objective function is error sum of squares, and this example is known as Ward's method or more precisely Ward's minimum variance method. In brief, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step (Ward 1963). Ward method works as follow:

- ❑ Uses an analysis of variance approach to evaluate the distances between clusters.
- ❑ Attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step.

$$d_{rs} \coloneqq \frac{n_r . n_s}{n_r + n_s} . \left\| \overline{Xr} - \overline{Xs} \right\|^2$$

Where r and s denote two specific clusters n_r and n_s denote the number of data points in the two clusters and $\bar{x}_r$ and $\bar{x}_s$ denote the centres of gravity of the clusters; | . | is the Euclidean norm.

- ❑ The mean and cardinality of the new cluster built as a product of the merger step is computed as follows

$$\bar{x}_r^{(new)} \coloneqq \frac{1}{n_r + n_s} . (n_r . \bar{x}_r + n_s . \bar{x}_s)$$

$$n_r^{(new)} \coloneqq n_r + n_s$$

When Ward method used on a trained SOM (Section 3.1.3), the homogeneity of clusters within the map is achieved by merging only neurons and clusters that are neighbours in the map (Samarasinghe 2007). Ward method helps to find the optimum cluster structure that indicates the best level of clustering. This method computes the likelihood of various numbers of clusters from which the most appropriate number of clusters can be obtained based on a likelihood index defined as

$$Ward\ Index = \frac{1}{NC}\left(\frac{d_t - d_{t-1}}{d_{t-1} - d_{t-2}}\right) = \frac{1}{NC}\left(\frac{\Delta d_t}{\Delta d_{t-1}}\right)$$

> Where $\mathrm{d_t}$ is the distance between the centres of the two cluster to be merged in the current step, and $\mathrm{d_{t-1}}$ and $\mathrm{d_{t-2}}$ are the distance between merged clusters in the previous step and the step earlier than the previous. NC is the number of cluster left.

3.1.2 Hierarchical clustering method

Hierarchical clustering (connectivity based clustering) is one of the most straightforward methods among numerous ways in which clusters can be formed. It is based on the core idea of objects being more related to nearby objects than to objects farther away. It can be either agglomerative (aggregating single patient into clusters) or divisive (dividing complete data set into partitions) (Johnson's 1967).

Agglomerative hierarchical clustering considers each case as a cluster. Any two objects that have the smallest value on the distance measure (or largest value if similarities are used) are joined into a single cluster. Either a third patient is added to this cluster that already containing the two objects, or a new cluster is formed by merging two new other objects. At each step, individual objects are added to existing clusters, two individuals are combined, or two existing clusters are combined.

The distance between two clusters with more than one patient in a cluster can be defined in different ways. For example, average distances between all pairs of patients formed by taking one member from each of the two clusters in turn and calculating the distance to other members to find the average distance. Or the largest or smallest distance between two patients that are in different clusters can be taken. Computing methods to measure the distance between clusters are available; however, different methods suffer from producing different solutions.

The distance between sets of observations as a function of the pairwise distances between observations is determined by linkage criteria.

The linkage function uses the distance information generated to determine the proximity of objects to each other. Once the proximity between objects in the data set has been computed the objects are paired into binary clusters, and the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed. Different distances, different clusters will form, what is known as the dendrogram. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by the hierarchical clustering. It is a useful approach to illustrate the clustering of genes or samples in Computational Biology.

To form clusters using a hierarchical cluster analysis, the following must be identified: a criterion to determine the similarity (or distance) between patients (such as Euclidean distance); a criterion to determine the merging of identified clusters at consecutive steps (such as minimum variance); and selecting the number of clusters to represent a particular dataset. The Ward method can be applied with hierarchical clustering to find the optimal number of clusters in the dendrogram.

3.1.3 Self-organizing maps (SOM)

The self-organizing map (SOM) is an artificial neural network algorithm used to visualize and interpret large high-dimensional datasets. This helps us to understand these high dimensional data by organising them on a map of reduced dimensions. The learning process is competitive and unsupervised. The components of SOMs are nodes or neurons that are usually arranged in a two-dimensional hexagonal or rectangular grid. A neuron is represented by a weight vector with the same dimension as the input vector and neuron weights are initially set to small random value. Neurons in a trained map represent the centre of gravity of a cluster input vectors. Self-organizing maps operate in two modes: training to build the map using input examples, and mapping to automatically classify a new input vector. The Euclidean distance is the most commonly used method to determine the neuron with the shortest distance to an input, called “the winner”. SOMs preserve topological properties of the input space. They use a neighbourhood feature to adjust neuron weights, not only of the winner neurons with the closest distance to an input vector, but also of their neighbour neurons. Furthermore, it presents mapping from a higher dimensional input space to a lower dimensional map space. This makes it possible to visualise the organisation of input vectors. Learning in self-organizing maps aims to specialise parts of the network to respond similarly to certain input patterns.

Figure 1 gives a pictorial example of a trained SOM; where the darker colour represents the area of larger distance between neurons (i.e., between weights). Smaller distances are represented by the lighter colour. This shows that the white areas represent different clusters and the black lines represent the boundaries between them.

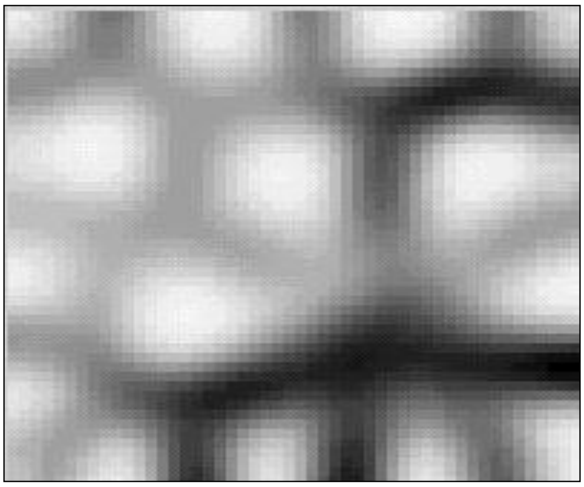


Figure 1: Self-organizing maps

3.1.3.1 Self-Organizing Map algorithm

Assume that some sample datasets have to be mapped onto an SOM. The set of input samples is described by a real vector

$$\mathbf{x(t)} \in \mathbf{R}^{\mathbf{n}} \qquad [5]$$

Where (*t)* is the index of the sample, or the discrete-time coordinate. Each node (*i)* in the map is represented by a weight vector

$$\mathbf{m_i(t)} \in \mathbf{R}^{\mathbf{n}} \qquad [6]$$

which has the same number of elements as the input vector **x(t).** The SOM algorithm performs an iteration process, where the initial values of the components of the weight vector, **$\mathbf{m_i}$(t)**, may be selected at random. In practical applications, however, the model vectors are more profitably initialized in some orderly fashion, e.g., along a two-dimensional subspace, spanned by the two principal eigenvectors of the input data vectors (Kohonen 1995c). Additionally, a batch version of the SOM algorithm may also be used (Kohonen

1995c). The $\mathbf{m_i(t)}$ which matches best with $\mathbf{x(t)}$ in some metric, is thought to map that input into its location. The self-organizing algorithm creates the ordered mapping as a repetition of the following basic tasks:

i. An input vector $\mathbf{x(t)}$ is compared with all the weight vectors $\mathbf{m_i(t)}$. The best-matching unit (node) on the map is selected. It is often called the winner.

ii. The weight vectors of the winner and a number of its neighbouring nodes in the array are changed towards the input vector according to the learning principle specified below.

The basic idea behind the learning process of SOM is that, for each sample input vector $\mathbf{x(t)}$, the winner and the nodes in its neighbourhood are brought closer to $\mathbf{x(t)}$ in the input data space. During the learning process, individual changes may be contradictory. However, the net outcome in the process is that ordered values for the mi(t) emerge over the map. If the number of available input samples is restricted, the samples must be presented iteratively to the SOM algorithm.

Adaptation of the weight vectors in the learning process may take place according to the following equations:

$$\mathbf{m_i(t + 1) = m_i(t) + \alpha(t)\ N_c(t)\ [x(t) - m_i(t)] \qquad \text{for each } i \in N_c(t),} \qquad [7]$$

Otherwise,

$$\mathbf{m_i(t + 1) = m_i(t)} \qquad [8]$$

where *t* is the discrete-time index of the variables, the factor $\boldsymbol{\alpha(t) \in [0,1]}$ is a scalar that defines the relative size of the learning step (learning rate), and $\boldsymbol{N_c(t)}$ specifies the strength of the weight adjustment for the *neighbourhood* around the winner in the map array.

The radius of the neighbourhood, at the beginning of the learning process, is fairly large. However, it is made to shrink during learning. This ensures that the global order is obtained already at the beginning, whereas towards the end, as the radius gets smaller, the local corrections of the model vectors in the map will be more specific. Gaussian neighbour strength as a function of radius is commonly used. The learning rate $\alpha(t)$ also decreases during learning.

Chapter Four

4 RESULTS

4.1 Pilot investigation

An initial analysis was conducted on 306 genes involving 5 types of cancer across 249 samples (patients). The set of data was the same as that used in the study of Hu et al. (2006). The purpose of the analysis was to identify genes that might be significantly related to each individual subtype, by using a preliminary exploration of the 306 genes (described below). If a gene is highly significant in one subtype but not significant in any other subtype, this can be used as a unique marker exclusive to that subtype.

The 306 genes were plotted against the gene expression level of each patient under each subtype. The value of the expression level is represented by ratio of (Abnormal/Normal). These values can be classified into normal, highly expressed, or highly depressed. Both the highly expressed and the highly depressed are indicators of cancer. The data were analysed using Matlab software to identify the median expression level of the most significantly expressed genes.

The results of the analysis are illustrated in Figure 2. The figure shows that the most significantly expressed gene in LumB subtype is Hs.79136 gene, at +1.888 expression level (Figure 2a, black colour). The expression level of the same gene (green colour) was +2.225 in LumA (Figure 2b), -0.402 in Normal Like (Figure 2c), -1.638 in Her2 (Figure 2d), and -1.063 in Basal Like (Figure 2e), respectively.

From these results it is evident that the Hs.79136 gene has significant expression levels, in different degrees, across almost all studied subtypes. For example, this gene is more highly expressed in LumA (+ 2.225) than in LumB (+1.888). However, as it is expressed in both it is not possible to consider this gene as an indicator of LumA or LumB subtype specifically. Same applies to subtypes in which this gene is highly expressed.

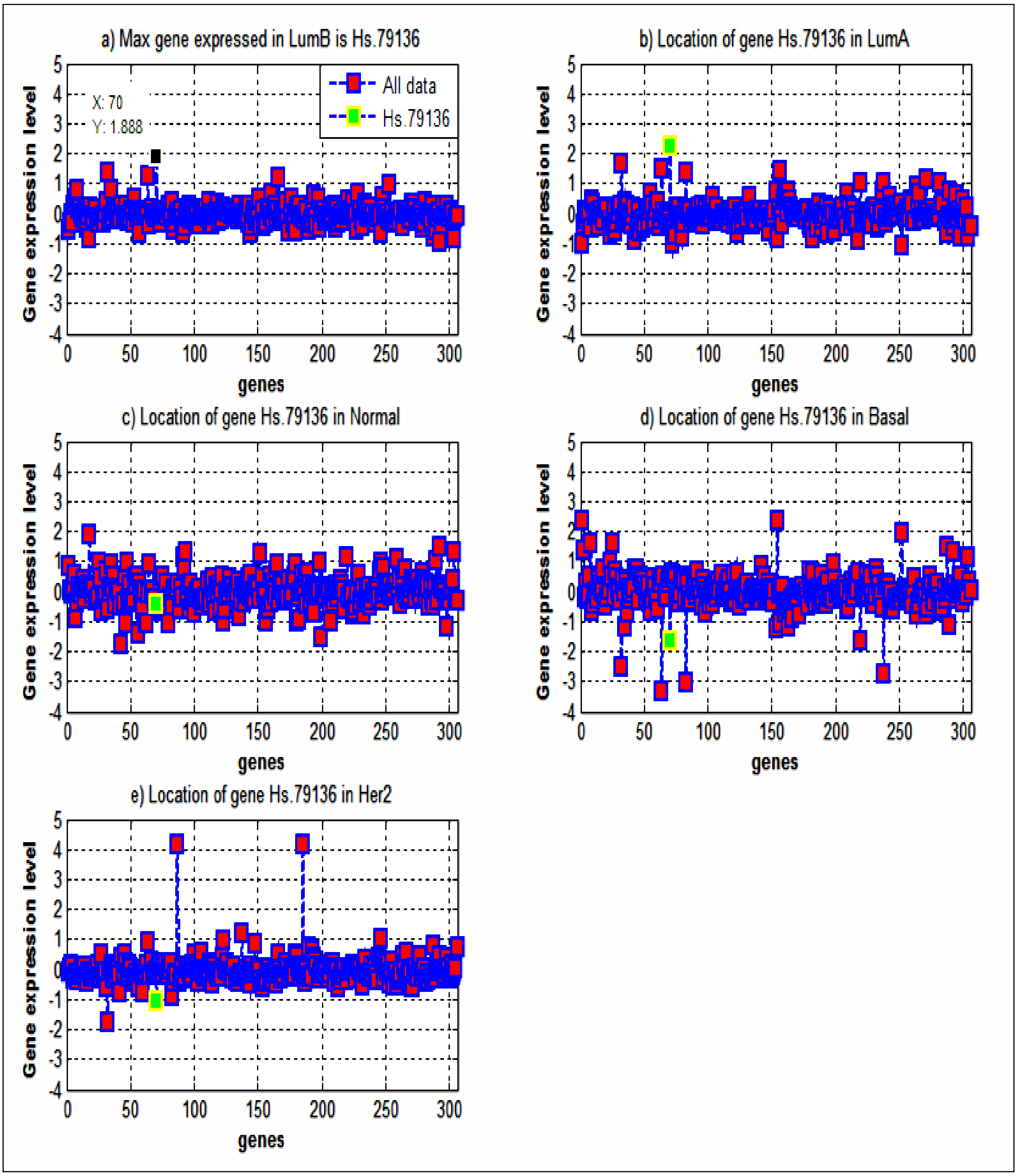


Figure 2: The median value of the highly expressed gene in LumB (Hs. 79136 in black colour) compared to its median expression levels in other subtypes (green colour).

Since using one significant gene as a marker was not successful, it was decided to involve more genes in the search. The analysis method searched for the six genes with the highest expression/inhibition in LumB, LumA, Normal Like, Basal Like, and Her2 as illustrated in Figure 3. Because the intensity of the data hindered the clarity of the outcome, a bar graph was chosen to represent the results in this case.

The six most highly expressed/inhibited genes in LumB were the previously noted Hs.79136, Hs.1657, Hs.82961, Hs.396783, Hs.80420 and Hs.425311 which are arranged here in a descending order according to their level of expression/ inhibition (Figure 3a). For example, Hs. 79136 is more highly expressed than Hs.1657 and so on.

Figure 3a also shows the expression levels of these genes in the other subtypes of cancer. It can be seen that genes Hs.79136, Hs.1657, Hs.82961 and Hs.80420 (representing genes No 1, 2, 3, and 5) are also highly expression/inhibited in subtypes other than LumB and cannot, therefore, be considered as specific markers. However, the most important genes in LumB are gene number 4 (Hs.396783) and 6 (Hs.425311) which are not significantly expressed/inhibited in the other four subtypes and could therefore be potentially useful markers for LumB.

When the same method of analysis was applied to LumA, it was observed that Hs.79136, Hs.1657, Hs.82961, Hs.89603, Hs.169946 and Hs.390163 genes were the six highly expressed/inhibited.

It was found that Hs.79136 was also highly expressed in this subtype as it was in LumB (Figure 3b and Appendix A). Thus, it was concluded that this gene cannot be considered as a useful marker to identify LumA. However, it was also observed that only gene number 4 (Hs.89603 +1.44) and possibly gene numbers 5 and 6 (Hs.169946 +1.36 and Hs.390163

+1.14) were the most significant in LumA compared to the other subtypes. These genes, unlike gene number 1 (Hs.79136), could therefore be useful as markers for LumA.

In the Normal Like subtype, the six most significantly expressed/inhibited genes were identified as Hs.81665 (+1.922), Hs.334562 (-1.75), Hs.77204 (-1.56), Hs.80420 (-1.64), Hs.69771 (-1.43) and Hs.433871 (1.34) (Figure 3c and Appendix A). The relative expression/inhibition level for these genes in the different subtypes varied. Although genes number 2 (Hs.334562) and 4 (Hs.80420) (in Figure 3c) were well expressed/inhibited, but they were also obviously present in other subtypes and cannot be considered as useful markers.

The most noticeable observation is that genes number 1 (Hs.81665), 3 (Hs.77204), 5 (Hs.69771) and 6 (Hs.433871) were very well expressed/inhibited in Normal Like than in other subtypes, particularly gene number 6 which can clearly be considered as a unique marker for this subtype.

Genes Hs.82961 (-3.35), Hs.169946 (-3.03), Hs.437638 (-2.72), Hs.1657 (-2.5), Hs.26770 (2.38) and Hs.2256 (2.37) (labelled number 1, 2, 3, 4, 5 and 6, respectively in Figure 3d) were highly expressed/inhibited in the Basal Like subtype, although they were also present in the other subtypes. The most interesting finding is that genes number 3 and 5 were clearly the most dominant in the Basal Like compared to other subtypes. It was also found that gene number 5 had no expression in Normal Like subtype. Hence, genes number 3 and especially number 5 can be confidently considered as useful markers for the Basal Like subtype. Gene number 1 is important in Basal Like subtype but since the same gene is also important in other subtypes, the third most significant gene in both LumB and LumA, this gene cannot be a useful marker for the Basal Like subtype.

In the Her2 subtype two distinctive genes were Hs.446352 (4.161) and Hs.86859 (4.130) (genes number 1 and 2 in Figure 3e) which were highly significantly expressed and can be regarded as unique markers for this subtype.

A summary of the pilot marker genes found in this analysis is gene in (Table 4.1) with colour to make it easier to find out if a gene is found important for more than one subtype.

Table 4.1: A summary of the six most significant genes with their expression levels in each subtype. (The colour indicates if a gene is significant in more than one subtype).

No	LumB		LumA		Normal Like		Basal Like		Her2	
	Gene Name	Expression	Gene Name	Expression	Gene Name	Expression	Gene Name	Expression	Gene Name	Expression
1	Hs.79136	1.888	Hs.79136	2.225	Hs.81665	1.920	Hs.82961	-3.350	Hs.446352	4.161
2	Hs.1657	1.380	Hs.1657	1.640	Hs.334562	-1.750	Hs.169946	-3.030	Hs.86859	4.130
3	Hs.82961	1.270	Hs.82961	1.460	Hs.77204	-1.560	Hs.437638	-2.720	Hs.1657	-1.754
4	Hs.396783	1.220	Hs.89603	1.440	Hs.80420	-1.640	Hs.1657	-2.500	Hs.90786	1.210
5	Hs.80420	-0.978	Hs.169946	1.360	Hs.69771	-1.430	Hs.26770	2.380	Hs.79136	-1.063
6	Hs.425311	0.976	Hs.390163	1.140	Hs.433871	1.340	Hs.2256	2.370	Hs.298654	1.000

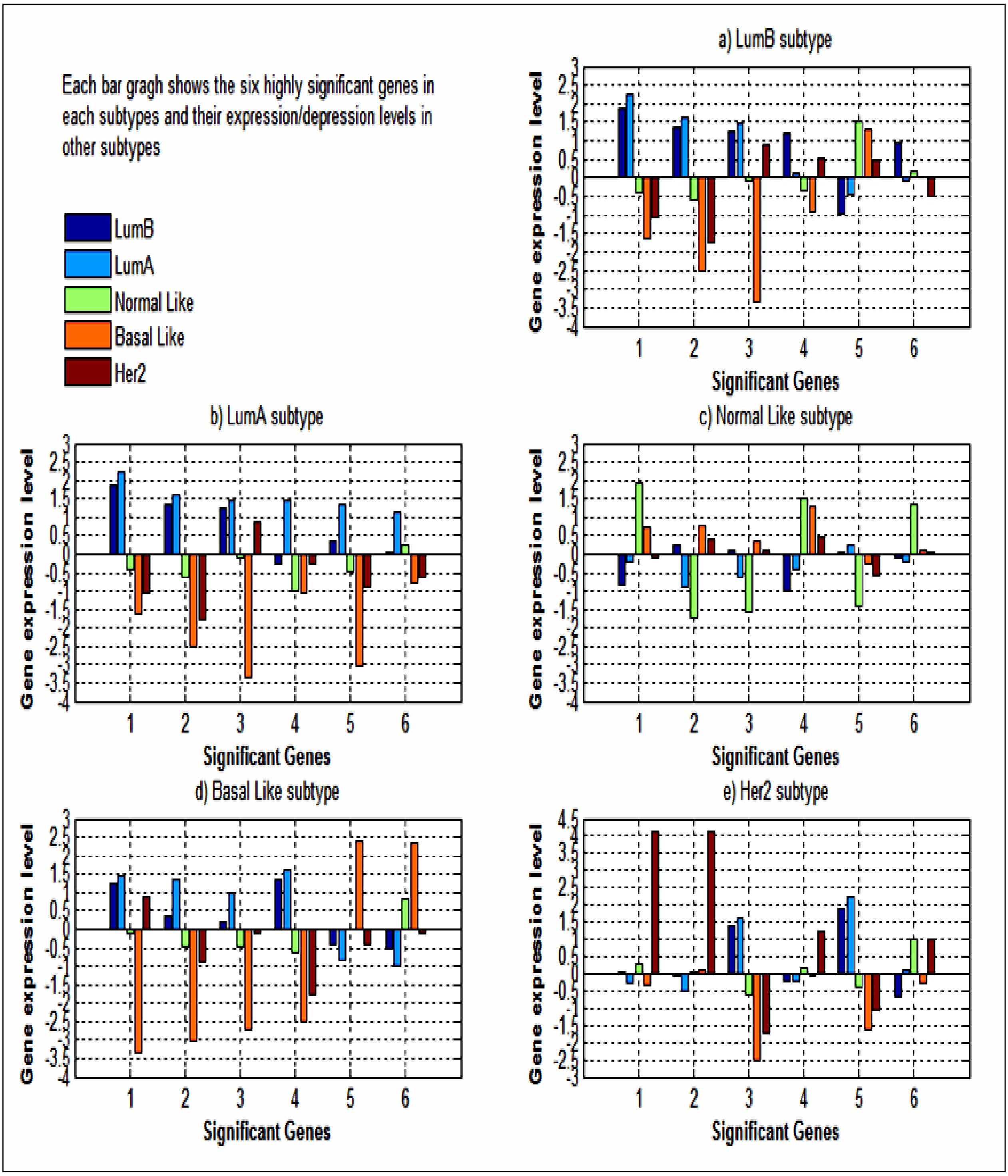


Figure 3: The median expression level of the most highly expressed/inhibited six genes in LumB, LumA, Normal Like, Basal Like, and Her2 subtypes.

4.1.1 Using Visualisation Analysis to identify potential markers

The findings of the pilot analysis were encouraging and justified the involvement of more genes that could help further identify other potential markers for the different subtypes in this study. Visualisation analysis method was carried out using GGobi software to analyse all the 306 studied genes simultaneously. The genes were plotted against their expression/inhibition levels across all subtypes and ranked from the highest positive to the highest negative value, as represented in Figure 4.

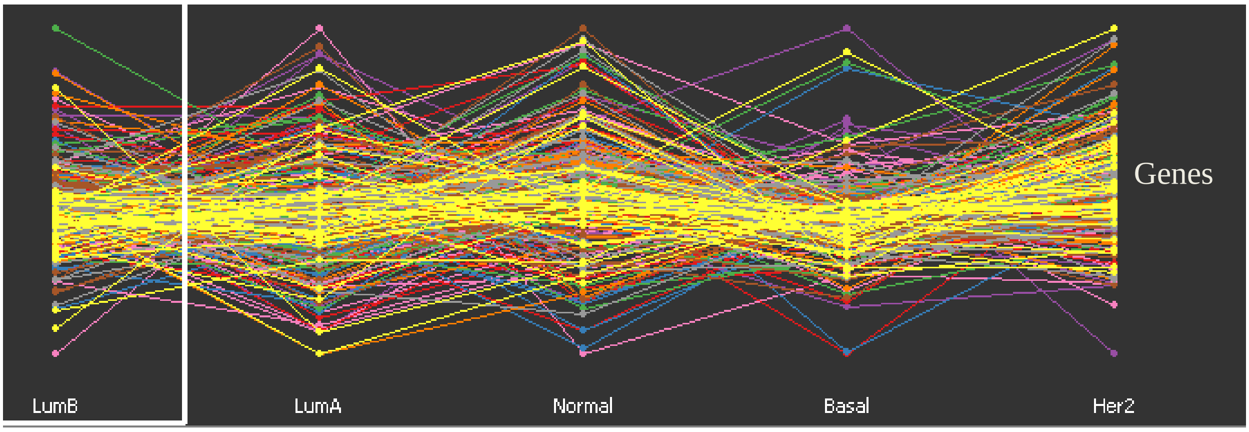


Figure 4: GGobi software based parallel coordinate plot for the 306 genes across the 249 patients representing the 5 subtypes.

The Figure 4 illustrates the complexity of plotting all the genes at the same time. Therefore, attempts were made to select the most highly expressed genes visually and label them with different colours and then follow their expression/inhibition in the different subtypes, as demonstrated in Figure 5.

Figure 5 shows that a gene signature pattern may be more clearly assessed by this method. For example, the three most highly significant genes in LumB (yellow colour) were also the most highly significant genes in LumA. The first and the second genes were insignificant in the Normal Like subtype, slightly inhibited in the Basal Like subtype but significantly inhibited in the Her2 subtype. The third significant gene in LumB and LumA was insignificant in both the Normal Like and the Her2 subtypes, but was highly significantly inhibited in the Basal Like. On the other hand, the fourth highly significant gene (also in yellow colour) in LumB was not significant in the other four subtypes. Therefore, this analysis indicates that this gene is probably a useful marker for the LumB subtype. The signature pattern of these four genes found by this method of analysis match the results observed previously when only six genes were plotted at one time (Figure 3).

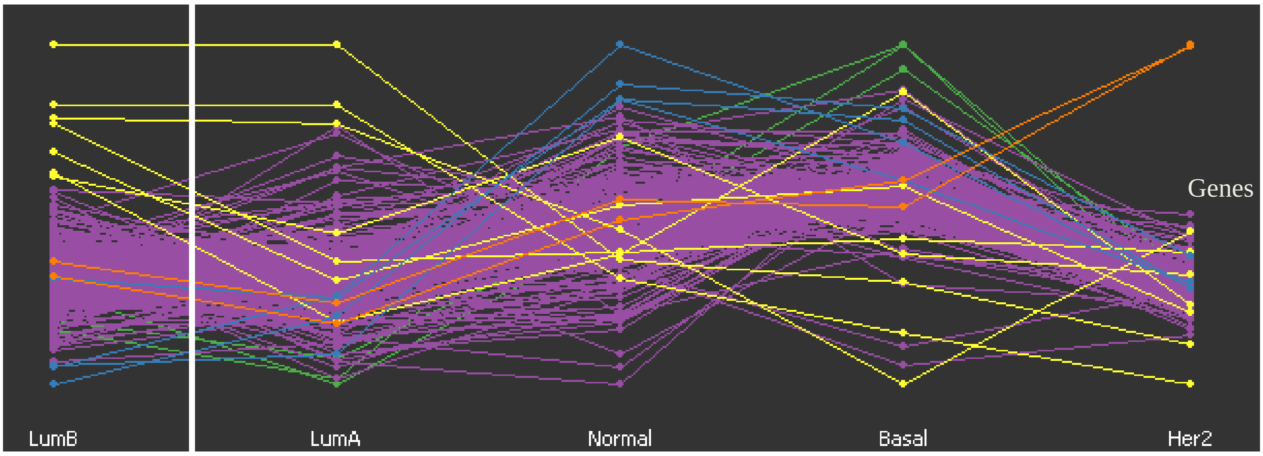


Figure 5: **GGobi software displays a parallel coordinates for the 306 genes across the 249 patients (subtypes) and highlighting only the most highly expressed genes visually and label them with different colours.**

Four significant genes were identified in the Normal Like subtype (blue colour), and the pattern of these genes' significance can be inferred as described above. Similarly, three significant genes can be identified in the Basal Like subtype (green colour line), and the pattern of significance of these genes can be traced in the other subtypes. For the Her2 subtype, two genes (orange colour line) can obviously be seen as the most significant for this subtype compared to their significance in the other subtypes.

Although, the signature pattern of the significant genes confirm the previous results recorded when only six genes were plotted at one time (Figure 3), the limitation of the visualisation analysis is that the software does not have scaling features. Therefore, care must be practiced when interrupting the results. For example, the first significant gene in LumB and LumA looks to have similar value in both subtypes, but the six genes analysis method, which is scaled, showed that the actual gene expression value was +1.888 in LumB whilst it was +2.224 in LumA. To explain further, the first significant gene in LumB can be used as a guide for its significance in other subtypes. By drawing a line across the different subtypes we can get a visual pattern of this genes' relative significance.

Another limitation of the visualisation analysis method is that it cannot identify (name) the different genes. The patterns from this method however, can be compared/matched with the patterns obtained from the six gene analysis method. Through this pattern comparison/matching we can name the genes. This study demonstrates that by combining the two methods, the patterns can be more confidently confirmed.

From the results above, it can be concluded that the two different forms of data representation, although useful, can be of limited advantage in identifying subtypes through the significance of expression/inhibition of a greater number of potential marker genes. It is therefore necessary to explore other methods of analysis using software that can isolate clusters of genes that can be important markers for the subtypes. This was not attempted in this study.

In the next sections, whole gene set (306 genes) obtained by Hu et al. (2006) is subjected to deeper investigation using Hierarchical clustering and Self Organising feature Maps (SOM) in conjunction with the Ward method to ascertain its efficacy in clustering subtypes.

4.2 Replicating original outcomes - Hierarchical clustering

The outcomes of our pilot study have encouraged us to explore the subtype classification beyond the investigation of the original researchers. The original researchers' work was first repeated by using the same method of analysis, namely hierarchical clustering method. The aim was to compare the outcomes of the current study with the original ones.

The hierarchical method uses different distance metrics. These differ from one another in how they measure the distance between clusters. These methods include: average- 'unweighted average distance', centroid- 'centroid distance', complete linkage- 'furthest distance', median- 'weighted centre of mass distance', single linkage- 'shortest distance', weighted- 'weighted average distance' and ward- 'inner square distance' (Appendix B). The Ward method was introduced to the hierarchical clustering method in this study because it employs the inner square distance (minimum variance algorithm) which helps representing the data more clearly, as illustrated in (Appendix B). The original authors had not mentioned the distance metric they used in Hierarchical clustering.

Figure 6 demonstrates that putting the data obtained from the hierarchical method through the Ward statistical analysis resulted in identifying the most likely number of clusters. The figure also shows that 6 sub-clusters are present with high probability. Furthermore, two clusters are also present but with less probability. This means that there are potentially 6 sub-clusters, are more than the five clusters found by the original authors.

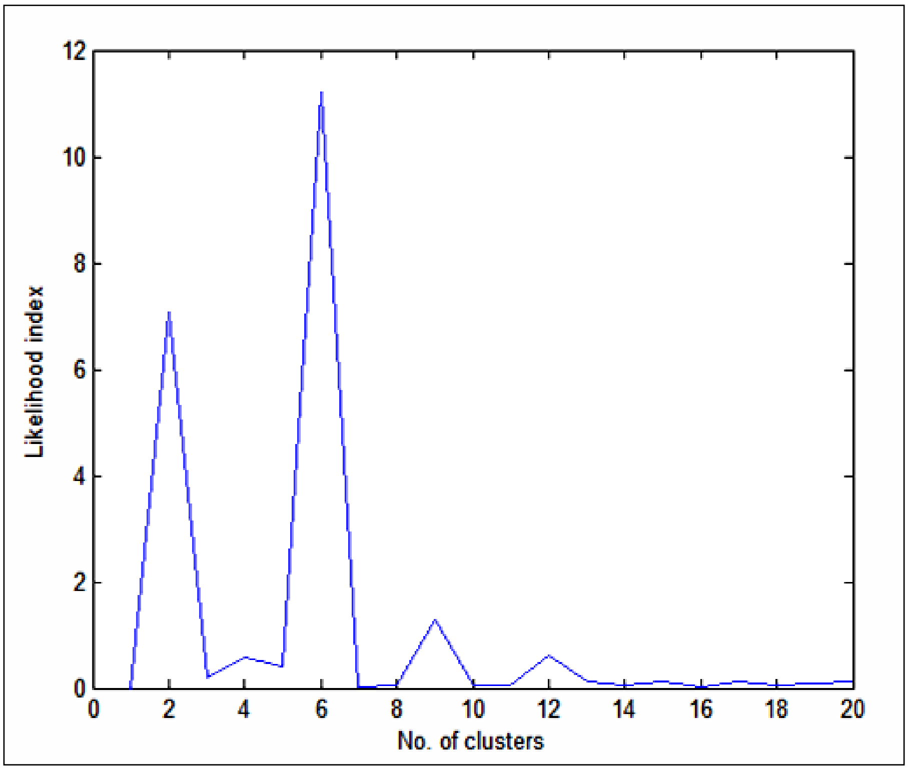


Figure 6: Ward likelihood index of hierarchical method for various numbers of clusters of patients (subtypes).

These observations were compared with those of the original authors' work. Therefore, the identified clusters were analysed further to find the distribution of the patients between the obtained clusters. The calculated distributions were compared with that of the original authors' results. Our analysis identified the sub-clusters and assigned the appropriate patients in each sub-cluster, and then identified the patients that were matched with those in the original work of the authors.

Table 4.2 shows the distribution of the patients between two clusters as identified by hierarchical method in the present study. The two clusters are shown in Figure 7. By matching the patient numbers (labels) with those numbers in the original authors' subtypes, we found that LumB, LumA and Normal Like subtypes belong to cluster 1 in this work, while the Basal Like and Her2 belong to cluster 2.

Table 4.2: Two clusters obtained from the hierarchical-ward method of the current study.

	Current Study		
	Cluster 1	Cluster 2	Total
No of Patients	161	88	249
Subtypes	LumB, LumA & Normal	Basal & Her2	

The same calculation procedure was followed to identify the distribution of patients between the subtypes. Table 4.3 shows the number of patients and their sequence (order of appearance in the dataset) in each subtype according to the original authors' data. When these numbers were compared with the numbers identified by the current hierarchical analysis (Appendix C) it was found that the patients belonging to LumB, LumA and the Normal Like subtypes (representing cluster 1) completely matched (100%) the sequence number reported in the original research. The sequence of the patients belonging to Basal Like and the Her2 subtypes (representing cluster 2) matched the sequence reported in the original research by only 98% and 83%, respectively. The remaining 2% and 17% of patient numbers were located in cluster 1 in this study.

Table 4.3: Matching percentage between original work and current study of hierarchical-ward method.

	Original Author's data					
	LumB	LumA	Normal	Basal	Her2	
No of patients	46	90	19	65	29	249
Sequence order of patient number	1 to 46	47 to 136	137 to 155	156 to 220	221 to 249	
match percentage	100%	100%	100%	98%	83%	

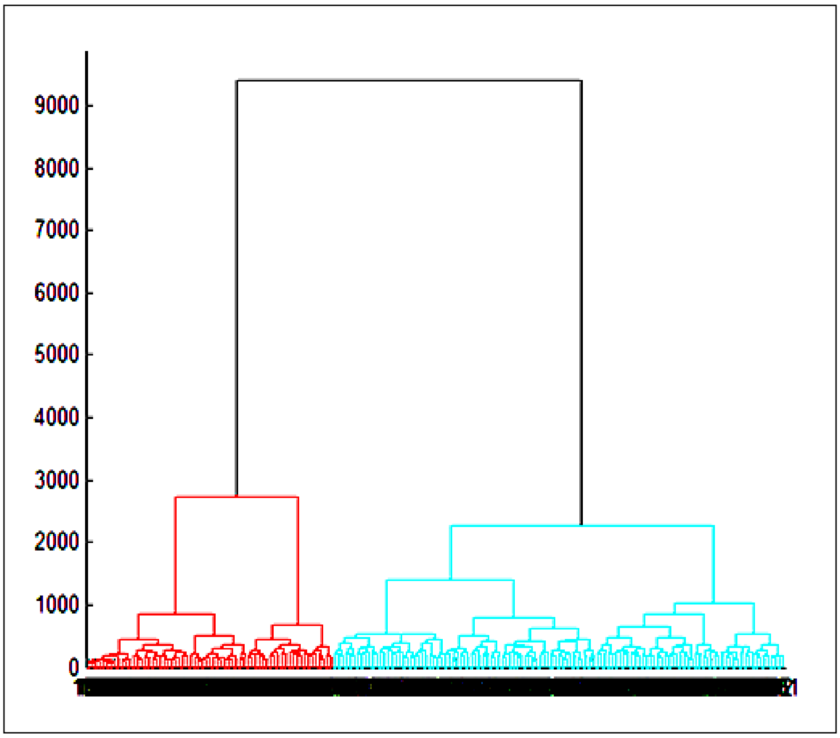


Figure 7: A dendrogram specifying the two clusters of patients (at cut off point 2727) according to their correlation strength.

Although not a strong outcome of our study, five sub-clusters were selected based on cut off distance 1050 in order to compare them with the original five subtypes reported by the original researchers (Figure 8). It was found that 67% of the patients of LumB were present in sub-cluster 1, 30% in sub-cluster 2 and 2% in other sub-clusters of this work (Table 4.4, Table 4.5 and Appendix C). Similarly, 50% of the LumA was in sub-cluster 3, 43% in sub-cluster 2 and 7% in sub-cluster 1 of the present work. Likewise, 95% of the Normal Like subtype was present in sub-cluster 3 and 5% only in sub-cluster 2. Sub-cluster 4 contained 86%, sub-cluster 5 contained 12% and the other sub-clusters contained 2% of the Basal Like subtype respectively. Finally, 79%, 10% and 10% of the Her2 subtype were in sub-cluster 5, sub-cluster 3 and other sub-clusters, respectively.

This clearly indicates that the present results have similarities, to varying degrees, to those of the original researchers. However, the results also indicate that there is a clear overlap between the subtypes, especially between LumB, LumA and Normal Like subtypes. As shown in Table 4.5, patients from sub-cluster 1 and sub-cluster 3 (belonging to cluster 1 in Table 4.2) can overlap with sub-cluster 4 and sub-cluster 5. This confirms the earlier finding that were explained in Table 4.2 and Table 4.3 regarding the overlapping of patients between the clusters.

Table 4.4: Five sub-clusters obtained from the hierarchical-ward method of the current study.

Hierarchical data analysis						
	Sub-cluster 1	Sub-cluster 2	Sub-cluster 3	Sub-cluster 4	Sub-cluster 5	Total No of patients
No of patients	40	54	67	57	31	249

Table 4.5: Matching percentage between original work and current study of the 5 sub-clusters obtained by hierarchical-ward method.

Original authors' subtype analysis						
	LumB	LumA	Normal	Basal	Her2	Total No of patients
No of patients	46	90	19	65	29	249
Sequence order of patient number	1 to 46	47 to 136	137 to 155	156 to 220	221 to 249	
Our Hierarchical as % of the original	Sub-clu 1	Sub-clu 3	Sub-clu 3	Sub-clu 4	Sub-clu 5	
	67%	50%	95%	86%	79%	
	Sub-clu 2	Sub-clu 2	Sub-clu 2	Sub-clu 5	Sub-clu 3	
	30%	43%	5%	12%	10%	
	Others	Sub-clu 1	Others	Sub-clu 1	Sub-clu 1	
	2%	7%	0%	2%	10%	

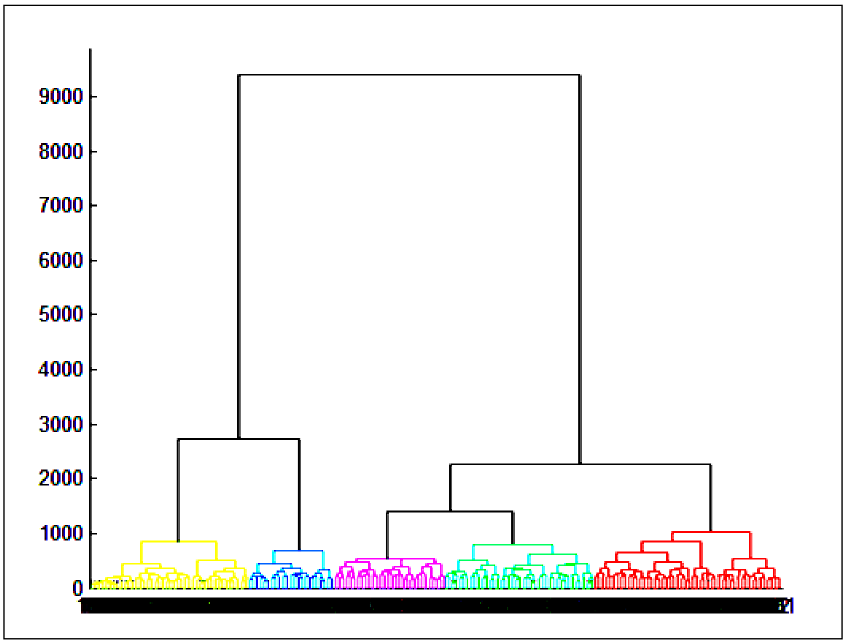


Figure 8: A dendrogram specifying the five clusters of patients (at cut off point 1050) according to their correlation strength

Our Hierarchical-Ward analysis showed the possibility of the presence of six sub-clusters rather than five. Examining these sub-clusters in depth (Table 4.6, Figure 9 and Appendix C) showed that 67%, 30% and 2% of the LumB subtype were in sub-cluster 3, sub-cluster 4 and sub-cluster 1 respectively. Further, 43%, 24%, 26% and 7% of the LumA subtype were in sub-cluster 4, sub-cluster 1, sub-cluster 2 and sub-cluster 3, respectively. Ninety five percent of the Normal Like subtype was in sub-cluster 1 and the remaining 5% was in sub-cluster 4. In addition, 86%, 12% and 2% of the Basal Like subtype were found sub-cluster 5, sub-cluster 6 and sub-cluster 3 in that order. Finally, 79%, 10%, 7% and 3% of the Her2 subtype were respectively in sub-cluster 6, sub-cluster 1, sub-cluster 3 and sub-cluster 5.

This analysis shows that there is considerable overlapping between the original five subtypes, plus the presence of one extra sub-cluster that has not been identified before. The extent of overlapping was characterised by all subtypes except the Basal Like subtype being present in sub-cluster 1. Similarly, all subtypes except the Normal Like were present in sub-cluster 3.

Furthermore, it was noticed that sub-cluster 1 and sub-cluster 3 were less frequent in the Basal Like and the Her2 subtypes than in the LumB, LumA and Normal Like subtypes.

Cluster consisting of the Basal Like and the Her2 subtypes in the two cluster format (Table 4.2) overlaps with different sub-clusters in the six sub-cluster format. However, the overlapping involving sub-cluster 5 and sub-cluster 6 was unique to Basal Like and Her2 subtypes. By contrast, overlapping with sub-cluster 4 was unique to the group containing LumB, LumA and the Normal Like subtypes.

The most significant observation was that sub-cluster 2 contained LumA subtype only.

Table 4.6: Matching percentage between original work and current study of the 6 sub-clusters obtained by hierarchical-ward method.

	Original Author's subtypes				
	LumB	LumA	Normal	Basal	Her2
No of Patients	46	90	19	65	29
Sequence order of patient numbers	1 to 46	47 to 136	137 to 155	156 to 220	221 to 249
Our Hierarchical as % of the original	Sub-clu 3	Sub-clu 4	Sub-clu 1	Sub-clu 5	Sub-clu 6
	67%	43%	95%	86%	79%
	Sub-clu 4	Sub-clu 1	Sub-clu 4	Sub-clu 6	Sub-clu 1
	30%	24%	5%	12%	10%
	Sub-clu 1	Sub-clu 2		Sub-clu 3	Sub-clu 3
	2%	26%		2%	7%
		Sub-clu 3			Sub-clu 5
		7%			3%

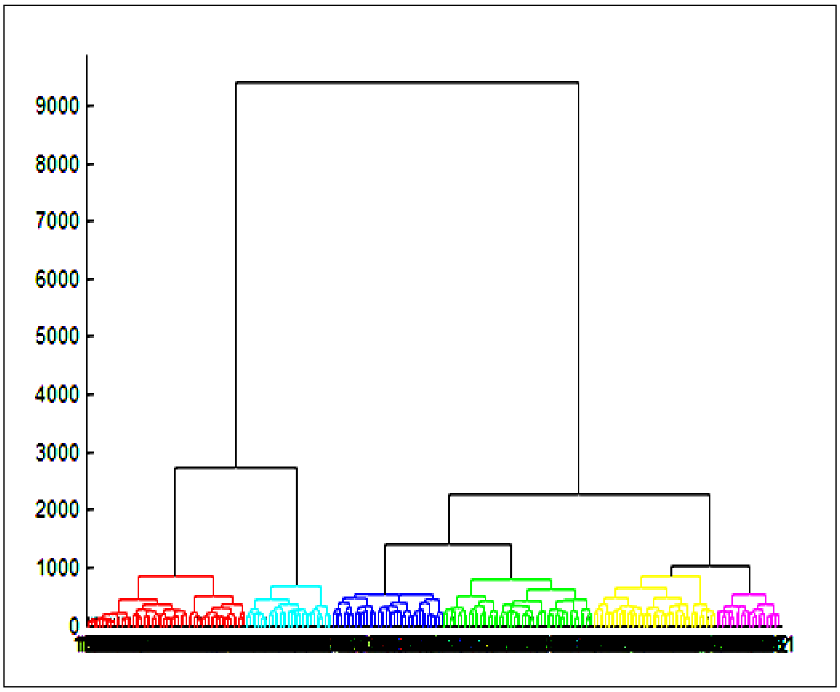


Figure 9: A dendrogram specifying the six clusters of patients (at cut off point 847) according to their correlation strength

The above results are now presented in the reverse format to highlight which sub-clusters represent which subtype(s). Table 4.7 shows results for the 5 cluster case. When these sub-clusters were compared with the 5 original authors' sub-clusters, it was found that 78%, 15%, 5% and 3% of LumB, LumA, Her2 and Basal subtypes, respectively were contained in Sub-cluster 1. LumA, LumB and Normal Like subtypes were contained in sub-cluster 2 at 72%, 26% and 2%, respectively. In sub-cluster 3 LumA, Normal Like, Her2 and LumB subtypes were present at 67%, 27%, 4% and 1%, respectively. Sub-cluster 4 98% contained Basal Like subtype and 2% Her2 subtypes. On the other hand, sub-cluster 5 contained 74% Her2 and 26% Basal like subtypes.

As shown previously, when the Ward analysis method was applied to the results of our Hierarchical analysis of the original dataset, it produced a strong likelihood for the presence of two clusters that were again divided into six sub-clusters (Figure 9) with the strongest likelihood. When the 6 clusters were analysed in terms of the subtypes they contain. It was interesting to observe that cluster 2 (Table 4.2) was divided into two sub-clusters (5 and 6 in Table 4.8) that contained the same subtypes that were present in sub clusters 4 and 5 (Table 4.7), and more interestingly, with the exact percentages.

Cluster 1 (Table 4.2) was divided into four sub-clusters (1, 2, 3 and 4, Table 4.8). Sub-clusters 3 and 4 contained the same subtypes and percentages that were present in sub-clusters 1 and 2 in the 5 clusters analysis (Table 4.7). Similarly, sub-clusters 1 in the 6 sub-clusters analysis contained the same subtypes that were contained in sub-cluster 3 in the 5 sub-clusters analysis, but with different percentages (Table 4.7). Sub-clusters 2 in the 6 sub-clusters analysis contained LumA subtype only. This could indicate that sub-clusters 1 and 2 resulted from the division of sub-cluster 3 in the 5 sub-clusters analysis. These results highlight the intricate relationship between LumA, LumB and Normal subtypes on one hand and that of Basal like and Her2 subtypes on the other, and clustering cannot separate the 5 subtypes uniquely. However, it can be said a patient falling into sub-cluster 2 could be identified as LumA with certainty. Similarly, a patient falling into sub-cluster 5 would highly likely be of Basal like subtype. Sub-cluster 6 would identify a patient to be more likely to be of Her2 but there is 25% chance for it to be of Basal-like. Overall, six clusters seems to offer an advantage over 5 clusters.

Table 4.7: The distribution comparison between the original five subtypes and the five sub-clusters obtained from the hierarchical-ward method in the current study.

Hierarchical Current data - 5 sub-clusters						
	Sub-Cluster 1	Sub-Cluster 2	Sub-Cluster 3	Sub-Cluster 4	Sub-Cluster 5	Total No of patients
No of patients	40	54	67	57	31	249
	LumB (78%)	LumA (72%)	LumA (67%)	Basal (98%)	Her2 (74%)	
	LumA (15%)	LumB (26%)	Normal (27%)	Her2 (2%)	Basal (26%)	
	Her2 (5%)	Normal (2%)	Her2 (4%)			
	Basal (3%)		LumB (1%)			

Table 4.8: The distribution comparison between the original five subtypes and the six sub-clusters obtained from the hierarchical-ward method in the current study.

Hierarchical Current data - 6 sub-clusters						
	Sub-Cluster 1	Sub-Cluster 2	Sub-Cluster 3	Sub-Cluster 4	Sub-Cluster 5	Sub-Cluster 6
No of patients	44	23	40	54	57	31
	LumA (50%)	LumA (100%)	LumB (78%)	LumA (72%)	Basal (98%)	Her2 (74%)
	Normal (41%)		LumA (15%)	LumB (26%)	Her2 (2%)	Basal (26)
	Her2 (7%)		Her2 (5%)	Normal (2%)		
	LumB (2%)		Basal (3%)			

4.3 Analysis with Self-Organizing Maps (SOM):

In the SOM analysis, patients are clustered to assess their separability into groups. A patient is represented by an input vector that contains expression ratios for the 306 genes. A 25 neuron map was trained in this study in batch mode. Figure 10 shows a partial pictorial view of SOM with only 2 inputs (2 genes) with green dots indicating the location of inputs and the blue indicating the corresponding components of the weight vectors. Results shown are after 200 iterations of the batch algorithm and indicate that the map is reasonably distributed through the input space. The SOM has spread across the input space as indicated by the map neurons connected to their neighbours with red lines.

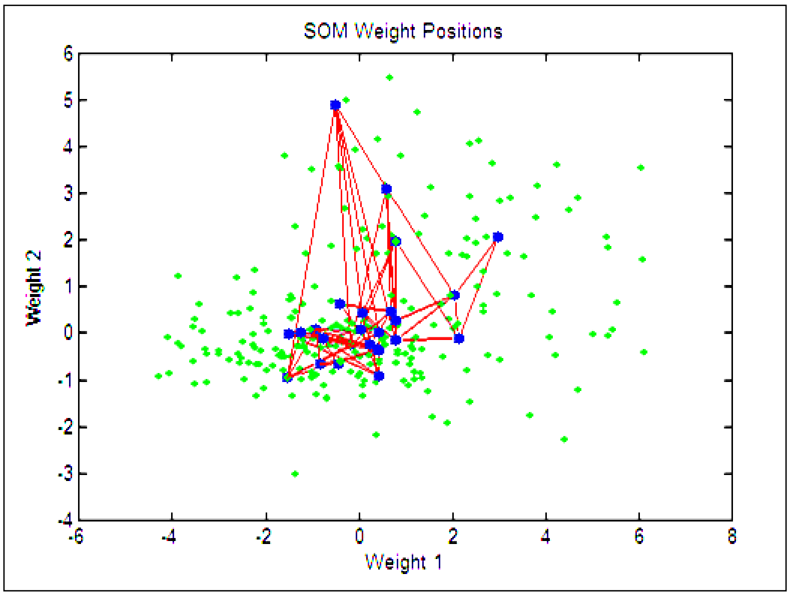


Figure 10: SOM plot of the inputs (green dots) and weights (blue) showing how the SOM spans the input space. Neighbour neurons are connected by red lines.

Figure 11 shows the U-matrix representation of the Self-Organizing Map that visualizes the SOM Neighbour Distances between the neurons (clusters) of the 249 samples. The blue hexagons represent the neurons and the red lines connect neighbouring neurons. The distance between the adjacent clusters is calculated and presented with different colours between the adjacent nodes. A darker colour between the clusters corresponds to a larger distance and thus a larger gap between the weights in the input space. A lighter colour between the neurons signifies that the weights vectors are closer to each other in the input space. Lighter areas can be thought of as clusters and dark areas as cluster separators. This can be a helpful presentation when one tries to find clusters in the input data without having any *a priori* information about the clusters.

According to Figure 11, there can be potentially 2 to 3 solid clusters that can be further broken down to smaller clusters which could possibly represent subtypes. Figure 12 shows how many of the training data are associated with each of the neurons (cluster centres). The maximum number of hits associated with any neuron is 35. Thus, there are 35 inputs or patients in that cluster.

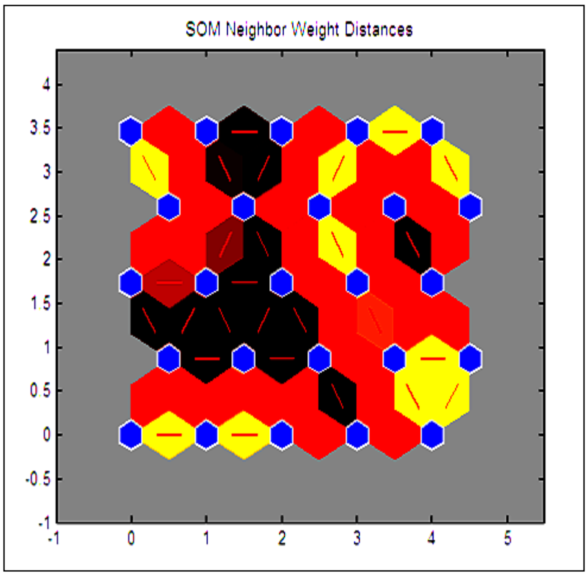


Figure 11: SOM layer showing neurons as blue and their direct neighbour relations with red lines. The neighbour patches (potential clusters) are coloured from black to yellow to show how close each neuron's weight is to that of its neighbours.

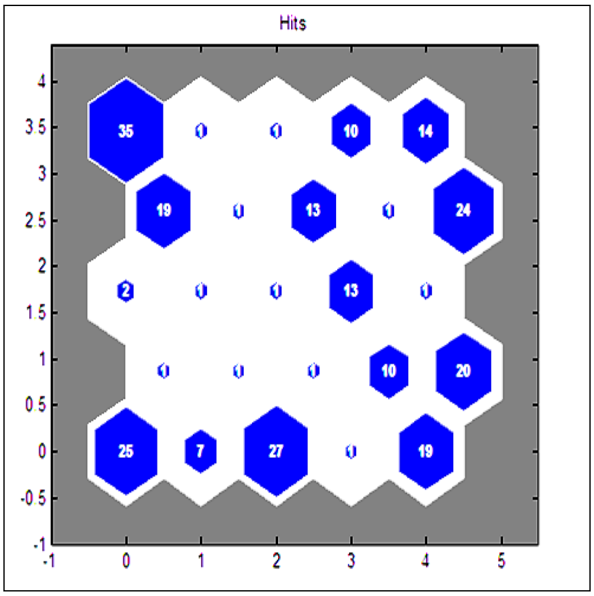
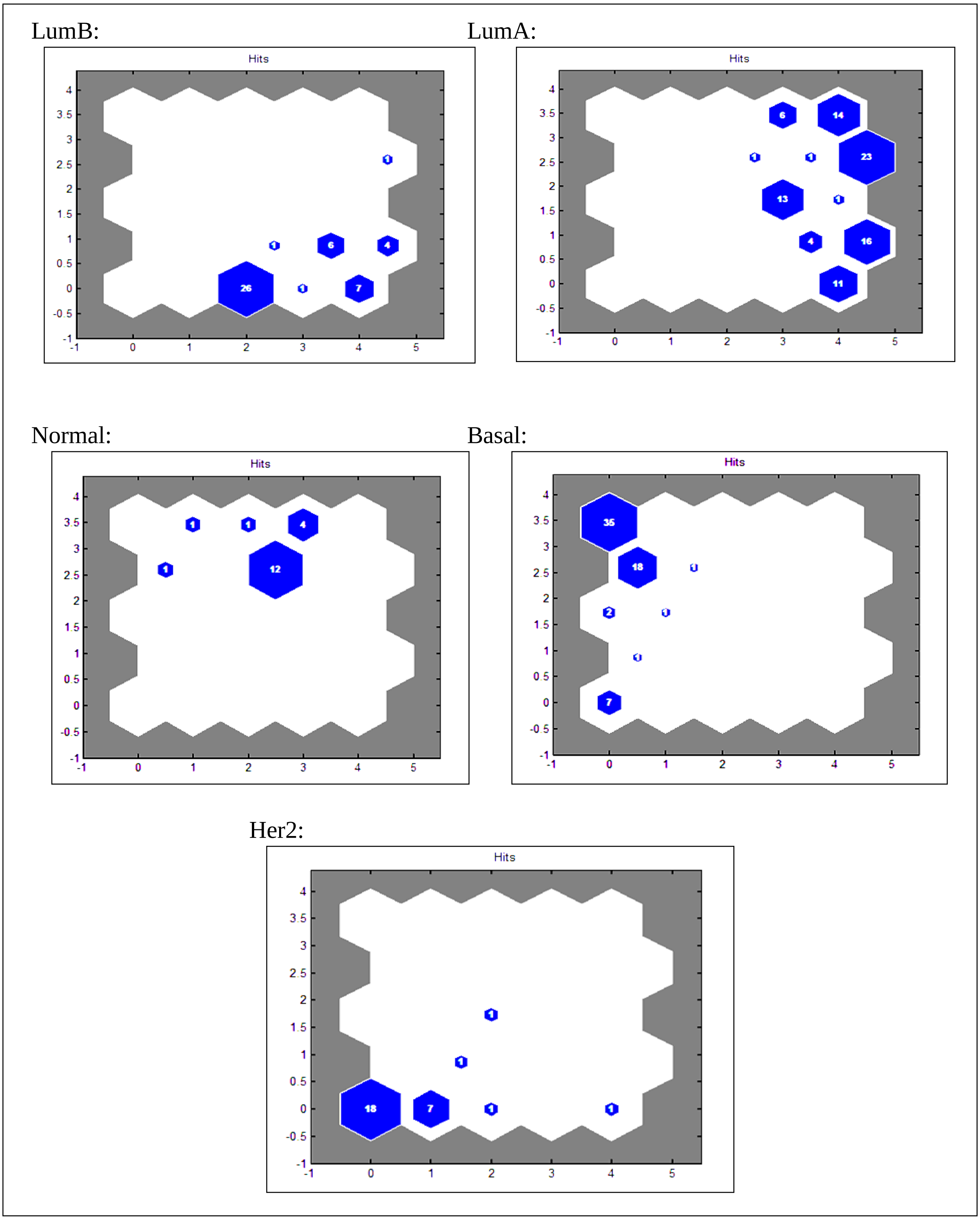


Figure 12: SOM layer with each neuron showing the number of inputs (patients) that it classifies. The relative number of patients for each neuron is shown via the size of a coloured patch.


LumB:

LumA:

Normal:

Basal:

Her2:

Figure 13: SOM layer with each neuron showing the number of inputs (patients) that it classifies from each subtype (LumB, LumA, Normal Like, Basal Like and Her2).

When the SOM representation is broken down to show results for samples from a particular subtype – LumA, LumB, Normal, Basal, and Her2 – we can see (Figure 13) that there are discrete patterns observable, suggesting that the subtypes are, for the most part, separate entities. LumB mostly occupies the lower right quadrant; LumA covers most of the right-hand side of the map; Normal accounts for the upper central region; Basal on the left, mostly the upper-left corner; and Her2 represents mostly the lower left-hand-side of the map. However, there is some degree of overlap; for example, the bottom-right neuron has a total of nineteen patients, eleven of them coming from LumA, seven from LumB, and one from Her2. Also, the bottom-left neuron has a total of twenty-five patients; these are shared between Basal and Her2, with the latter having the majority. There are also some less pronounced overlaps. Because this method only partially confirms the existence of the five subtypes, with some overlap being apparent between them, we need to find a method of clearly and unambiguously identifying and distinguishing between the subtypes.

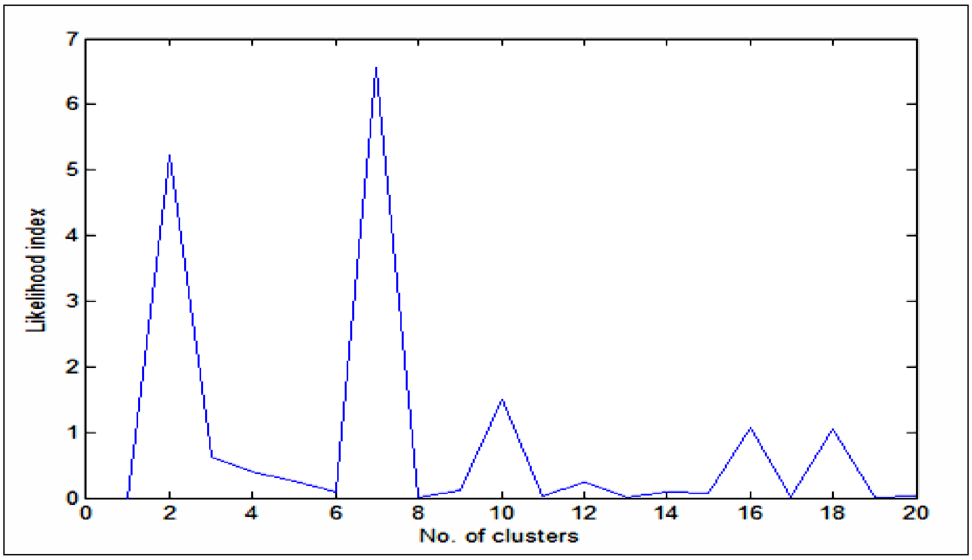


Figure 14: Ward likelihood index for various numbers of clusters of SOM map neurons.

Figure 14 demonstrates that putting the data (trained neuron weights) obtained from the SOM method (on selecting the suitable set of criteria of SOM method (i.e. patch learning tools) to obtain reproducible result) through the Ward statistical analysis can help in identifying the most likely number of clusters. The figure shows that 7 clusters are present with high probability. This means that there are two other clusters that have not been identified by the original researchers using the same set of data. The figure also shows the possibility of the presence of two and with much less probability of ten clusters.

To test the significance of these observations in relation to the original authors' work, the clusters identified by the current analysis were analysed further to find the distribution of the patients between these clusters, and compare the calculated distributions with that of the original authors'. The analysis identified the clusters and assigned the appropriate patients in each cluster. The identified patients were matched with those identified by the original authors.

Table 4.9 shows the distribution of the patients between two clusters as identified in the present study and further highlighted in Figure 15. By matching the patient numbers with those numbers in the clusters that were identified by the original authors, we found that LumB, LumA and Normal Like subtypes belong to cluster1 in this work, while the Basal Like and Her2 belong to cluster2. These results are identical to those obtained from the Hierarchical-Ward analysis (Table 4.1).

Table 4.9: Two clusters obtained from the SOM-Ward method including the number of patients in each cluster.

Current Study			
	Cluster 1	Cluster 2	Total
No of Patients	161	88	249
Subtypes	LumB, LumA & Normal	Basal & Her2	

The same calculation procedure was followed to identify the distribution of patients between the subtypes. Table 4.10 shows the number of patients and their sequence in each subtype according to the original authors' data. When those patients numbers were matched with the number sequence identified by the current analysis (Appendix D) it was found that the sequence of the patients belonging to LumB, LumA and the Normal Like subtypes (representing cluster 1) matched (100%) the sequence reported in the original research.

On the other hand, the sequence of the patients belonging to Basal Like and the Her2 subtypes (representing cluster 2) matched the sequence reported in the original research by 98% and 83%, respectively. The remaining 2% and 17% of patient numbers were located in cluster 1. These results are identical to the findings from Hierarchical-Ward analysis (Table 4.2).

Table 4.10: Comparison of the two clusters obtained from the current SOM-Ward with the original subtypes

	Original Author's data					
	LumB	LumA	Normal	Basal	Her2	
No of patients	46	90	19	65	29	249
Sequence order of patient number	1 to 46	47 to 136	137 to 155	156 to 220	221 to 249	
SOM to the original %	100%	100%	100%	98%	83%	

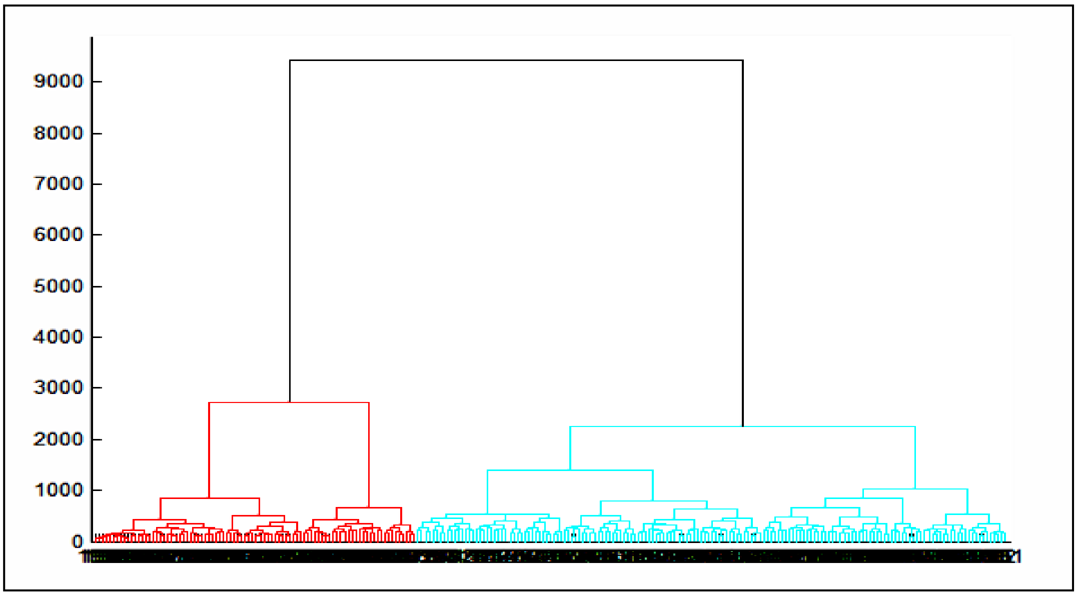


Figure 15: A dendrogram specifying the two clusters of patients (at cut off point 3000) according to their correlation strength

Based on a cut off point 1050 in Figure 16, five clusters were identified for the purpose of comparison with the original study because the original research work reported the presence of 5 clusters only. These clusters were matched with the original 5 subtypes to examine the accuracy of the method used here.

It was found that 67% of the patients of LumB were present in sub-cluster 1, 30% in sub-cluster 2 and 2% in other sub-clusters of this work (Table 4.11, Table 4.12 and Appendix D). Similarly, 50% of the LumA was in sub-cluster 3, 43% in sub-cluster 2 and 7% in sub-cluster 1 of the present work. Likewise, 95% of the Normal Like subtype was present in sub-cluster 3 and 5% in sub-cluster 2 and none in the other subtypes. Sub-cluster 4 contained 86%, sub-cluster 5 contained 12% and the other sub-clusters contained 2% of the Basal Like subtype, respectively. Finally, 79%, 10% and 10% of the Her2 subtype were in sub-cluster 5, sub-cluster 3 and other sub-clusters, respectively.

This clearly indicates that the present results have similarities, to varying degrees, to those of the original researchers. However, the results also indicate that there is a clear overlap between the subtypes, especially between LumB, LumA and Normal Like subtypes. We noticed in Table 4.12 that patients from sub-cluster 1, sub-cluster 2 and sub-cluster 3 (belonging to cluster 1 in Table 4.9) contain all LumB, LumA and Normal Like subtypes. Sub-cluster 4 and sub-cluster 5 represent majority of Basal Like and Her2, with sub-cluster 1 and sub-cluster 3 sharing the rest. This confirms the earlier finding as explained in Table 4.9 and Table 4.10 regarding the overlapping of patients between the sub-clusters. Such overlapping would in turn lower the confidence in using the genes as exclusive markers for the subtypes.

Table 4.11: Current data of SOM five sub-clusters including the number of patients in each sub-cluster .

SOM analysis data						
	Sub-cluster 1	Sub-cluster 2	Sub-cluster 3	Sub-cluster 4	Sub-cluster 5	Total No of patients
No of patients	40	54	67	57	31	249

Table 4.12: compression of the five clusters from SOM with the original dataset 5 subtypes (Hu et al. 2006).

Original authors' data						
	LumB	LumA	Normal	Basal	Her2	Total No of Clusters
No of Samples	46	90	19	65	29	249
from	1 to 46	47 to 136	137 to 155	156 to 220	221 to 249	
SOM to the original %	Clu1	Clu3	Clu3	Clu4	Clu5	
	67%	50%	95%	86%	79%	
	Clu2	Clu2	Clu2	Clu5	Clu3	
	30%	43%	5%	12%	10%	
	Clu3	Clu1		Clu1	Clu1	
	2%	7%		2%	10%	

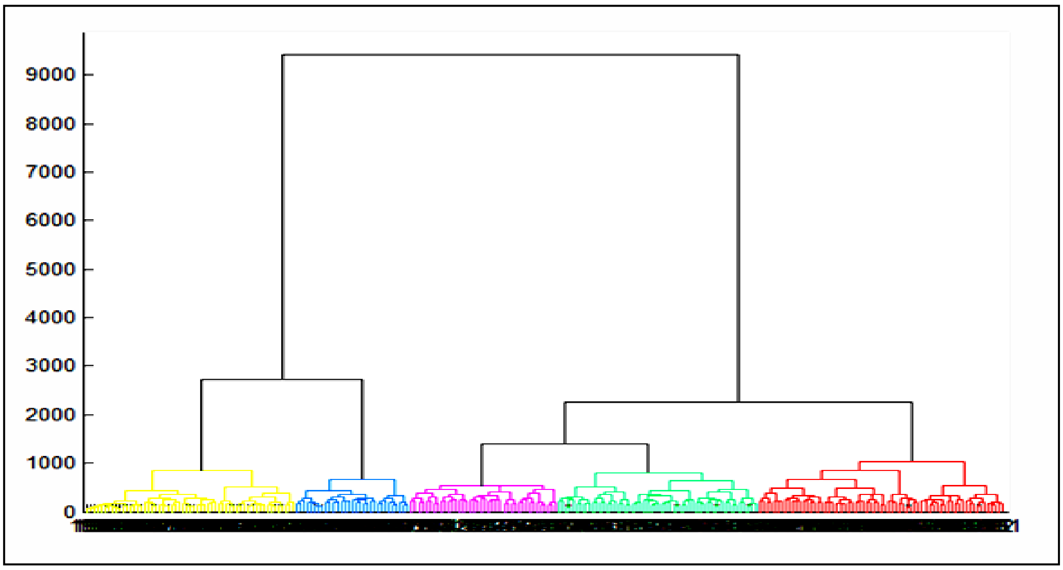


Figure 16: A dendrogram specifying the five clusters of patients (at cut off point 1050) according to their correlation strength

The overlapping in the five SOM-Ward sub-clusters raised the question of whether there are more unidentified sub-clusters. Our SOM-Ward analysis showed the possibility of the presence of seven sub-clusters (Figure 17). Examining these clusters (Table 4.14, Figure 17 and Appendix D) showed that 67%, 30% and 2% of the LumB subtype were in sub-cluster 5, sub-cluster 6 and sub-cluster 3, respectively. Further, 43%, 24%, 26% and 7% of the LumA subtype were in sub-cluster 6, sub-cluster 3, sub-cluster 4 and sub-cluster 5, respectively. Ninety five percent of the Normal Like subtype was in sub-cluster 3 and the remaining 5% was in sub-cluster 6. In addition, 55%, 31%, 12% and 2% of the Basal Like subtype were found sub-cluster 1, sub-cluster 2, sub-cluster 7 and sub-cluster 5 in that order. Finally, 79%, 10%, 7% and 3% of the Her2 subtype were respectively in sub-cluster 7, sub-cluster 3, sub-cluster 5 and sub-cluster 2.

The seven sub-clusters analysis clearly indicates that there is a significant overlapping between the original five subtypes, plus the presence of two extra clusters that have not been identified before.

Studying the overlap characteristics (between the original subtypes) showed that sub-cluster 3 contained patients from all subtypes except Basal Like. On the other hand, sub-cluster 5 did not contain the Normal Like subtype. Furthermore, it was noticed that sub-cluster 3 and sub-cluster 5 represented more of LumB, LumA and Normal Like subtypes than Basal Like and Her2 subtypes.

It was found that although in the 2 cluster format (Table 4.9) the Basal Like and the Her2 subtypes are represented by one cluster, in the 7 sub-cluster format, there is some overlap with LumB, LumA and Normal Like subtypes. However the overlapping involving sub-cluster 2 and sub-cluster 7 was unique as they only shared Basal Like and the Her2 subtypes and did not contain LumB, LumA and Normal Like subtypes. By contrast, sub-cluster 6 was unique to the group containing LumB, LumA and the Normal Like subtypes.

The most significant observation was that sub-cluster 1 contained Basal Like subtype only, and sub-cluster 4 contained LumA subtype only.

Table 4.13: SOM seven clusters including the number of patients in each cluster.

	Sub-cluster 1	Sub-cluster 2	Sub-cluster 3	Sub-cluster 4	Sub-cluster 5	Sub-cluster 6	Sub-cluster 7
No of Patients	36	21	44	23	40	54	31

Table 4.14: Comparison of the seven clusters from SOM with the original 5 subtypes (Hu et al. 2006).

	Author's data				
	LumB	LumA	Normal	Basal	Her2
No of Patients	46	90	19	65	29
Sequence order of patient number	1 to 46	47 to 136	137 to 155	156 to 220	221 to 249
SOM as % of the original	Clus 5	Clus 6	Clus 3	Clus 1	Clus 7
	67%	43%	95%	55%	79%
	Clus 6	Clus 3	Clus 6	Clus 2	Clus 3
	30%	24%	5%	31%	10%
	Clus 3	Clus 4		Clus 7	Clus 5
	2%	26%		12%	7%
		Clus 5		Clus 5	Clus 2
		7%		2%	3%

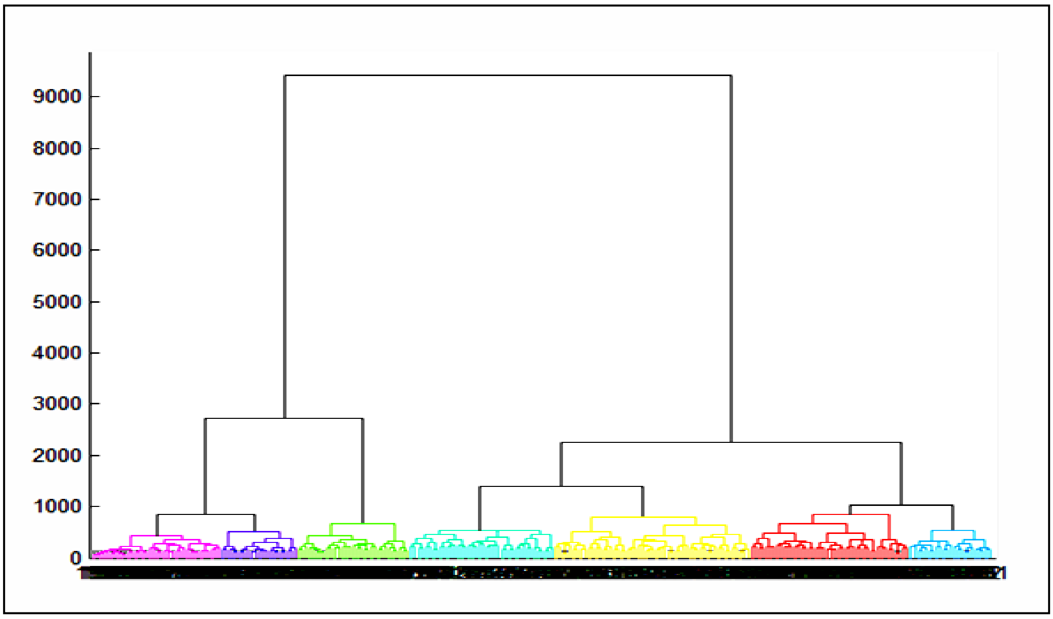


Figure 17: A dendrogram specifying the seven clusters of patients (at cut off point 845) according to their correlation strength.

The turning now to the results of clusters from the perspective of subtypes, Table 4.15 indicates that repeating the analysis of the same data used by the original authors with SOM-Ward method produced identical (100%) results to those obtained by the Hierarchical-Ward method for the case of five sub-clusters. This provides validity to our analysis methods.

Introducing Ward analysis to our SOM analysis produced the likelihood of the presence of two clusters that divided into seven sub-clusters (Figure 14). Cluster 1 (Table 4.2) was divided into four sub-clusters (3, 4, 5 and 6, Table 4.16). Sub-clusters 5 and 6 in this analysis contained the same subtypes and percentages that were present in sub-clusters 1 and 2 of the 5 sub-clusters analysis (Table 4.15). Sub-cluster 3 contained the same subtypes as those contained in sub-cluster 3 of the 5 sub-clusters analysis, but with different percentages (Table 4.15). Unlike other sub-clusters, sub-cluster 4 contained LumA subtype only. This indicates that sub-cluster 4 could identify LumA patients with certainty.

The analysis indicates that cluster 2 (Table 4.2) was divided into three sub-clusters (1, 2 and 7 in Table 4.16). Sub-cluster 7 contained the same subtypes that were present in sub-clusters 5 (the 5 sub-clusters analysis, Table 4.15) and with the exact percentages. Sub-cluster 1 in the 7 sub-clusters analysis contained the Basal Like subtype only. This indicates that sub-cluster 1 could identify Basal like subtype with certainty. Furthermore, sub-cluster 2 would identify Basal-like subtype with high likelihood as 95% of it is this subtype.

Table 4.15: The distribution comparison between the original five subtypes and the five sub-clusters obtained from the SOM-Ward method in the current study.

SOM Current study - 5 sub-clusters						
	Sub-cluster 1	Sub-cluster 2	Sub-cluster 3	Sub-cluster 4	Sub-cluster 5	Total No of patients
No of patients	40	54	67	57	31	249
	LumB (78%)	LumA (72%)	LumA (67%)	Basal (98%)	Her2 (74%)	
	LumA (15%)	LumB (26%)	Normal (27%)	Her2 (2%)	Basal (26%)	
	Her2 (5%)	Normal (2%)	Her2 (4%)			
	Basal (3%)		LumB (1%)			

Table 4.16: The distribution comparison between the original five subtypes and the seven sub-clusters obtained from the SOM-Ward method in the current study.

SOM Current data -7 clusters								
	Sub-cluster 1	Sub-cluster 2	Sub-cluster 3	Sub-cluster 4	Sub-cluster 5	Sub-cluster 6	Sub-cluster 7	Total No of patients
No of patients	36	21	44	23	40	54	31	249
	Basal (100%)	Basal (95%)	LumA (50%)	LumA (100%)	LumB (78%)	LumA (72%)	Her2 (74%)	
		Her2 (5%)	Normal (41%)		LumA (15%)	LumB (26%)	Basal (26%)	
			Her2 (7%)		Her2 (5%)	Normal (2%)		
			LumB (2%)		Basal (3%)			

Chapter Five

5 DISCUSSION

In this study new analysis based on two known clustering methods, namely, Hierarchical clustering and Self-Organizing Maps (SOM), has been applied. Unlike the work of the previous researchers (Sorlie et al. 2001 & 2003; Van't Veer et al. 2002; Sotiriou, et al. 2003; Hu et al. 2006; and Build et al. 2009) who used the Hierarchical clustering method only, the present study paired two different clustering methods with the Ward method. This is to investigate whether more details about the clusters found by Hu et al., (2006) can be obtained. Another aim is to ascertain the degree of overlap of subtypes, and to find the most significant reliable diagnostic markers for the known subtypes.

Hu et at. (2006), stated that genes under study should be selected on the basis of their consistent expression when individual tumours are examined, but that vary in expression across different tumours. The pilot work analysis in this study aimed at finding the most significant genes in each of the five subtypes reported by Hu et al. (2006) study. The analysis identified genes that can be possibly considered as significant markers for the different subtypes, and some genes were exclusive to certain subtypes. These exclusive genes have not been identified as specific to each subtype previously. These genes are as follows: LumB (Hs.79136, Hs.1657, Hs.82961, Hs.396783, Hs.80420 and Hs.425311), LumA (Hs.79136, Hs.1657, Hs.82961, Hs.89603, Hs.169946 and Hs.390163), Normal Like (Hs.81665, Hs.334562, Hs.77204, Hs.80420, Hs.69771 and Hs.433871), Basal Like (Hs.82961, Hs.169946, Hs.437638, Hs.1657, Hs.26770 and Hs.2256) and Her2 (Hs.446352, Hs.86859, Hs.1657, Hs.9078, Hs.7913 and Hs.2986).

The results of the pilot work encouraged us to repeat the original work of Hu et al. (2006) but by introducing the Ward clustering to the Hierarchical clustering methods instead of the Hierarchical clustering on its own. The Hierarchical clustering produces a dendrogram that consists of many U-shaped lines connecting objects in a hierarchical tree of clusters, but it does not provide the optimum partitioning required by the user. Partitioning can be achieved by cutting the dendrogram at certain levels (cut off distance) (Samarasinghe 2007). Our approach was to use Ward and Hierarchical clustering methods simultaneously to merge clusters and to find the optimum cluster structure. As seen in the results, this analysis approach showed the presence of two main clusters that in turn are divided into six possible sub-clusters.

These six sub-clusters contained significant percentages of the original patients' distribution pattern which indicate that our analytical method is credible. The rest of the original patients' were distributed among these six sub-clusters. This may explain the overlapping that was observed in the pilot analysis. Thus introducing the Ward method to the Hierarchical clustering method could help in obtaining more detailed description about the possible number of clusters and the distribution of the patients between clusters.

Gene expression analysis using hierarchical clustering has categorized breast cancers into at least five main groups or subtypes (McCafferty, Healy et al. 2009). It is evident from the findings reported in the literature that integrating different analysis methods may produce more detailed outcomes. Huber et al. (2009) concluded that the effectiveness of target-specific therapies are still limited, but understanding of the heterogeneous biology of breast cancer can give clear direction for future research.

The new approach in this study takes into account the above observations and results are in agreement with those of Sorlie et al (2001) and Sotiriou, et al. (2003) who identified six subtypes, compared with the findings of Sorlie et al. (2003), Hu et al. (2006) and Build et al. (2009) who identified five subtypes only.

The self-organizing map (SOM) helps in visualizing and interpreting large high-dimensional datasets. The introduction of the Ward method to the SOM in this study was to investigate whether the result obtained from the Hierarchical clustering method alone could be verified. The results showed high probability of two as well as seven clusters. The two clusters were identical to those found by the Hierarchical-Ward clustering method. The inspection of the seven sub-clusters showed that a unique overlap existed between the Basal Like and Her2 subtypes, namely sub-clusters number 7 and 2 which shared these subtypes only (Table 4.14). It was also evident that sub-cluster number 1 represented only the Basal Like subtype.

This study shows that sub-cluster number 6 was unique to LumB, LumA and Normal Like subtypes only, and did not represent the other subtypes. A significant outcome is that cluster number 4 represented LumA subtype only.

Hu et al. (2006) selected genes that are consistently expressed when individual tumours are examined, but that vary in expression across different tumours. The results of the present study showed that certain genes that are significantly expressed in one subtype may also be significantly expressed in other subtypes too. This means that such genes cannot be considered as specific markers as indicated by Hu et al. (2006).

Hu et al. (2006) claimed that the basal subtype can be recognized as a distinct group and should be considered as a separate disease with respect to treatment and follow up. Our results may provide the explanation for why the basal subtype is more distinct. Our five sub-cluster results showed that when the Hierarchical-Ward analysis method was used, sub-cluster 4 uniquely represent the Basal Like subtype only. This finding was also confirmed by the results of SOM-Ward analysis method. This may mean that the Basal Like subtype is more distinct than other subtypes.

Sorlie et al. (2003) suggested that other subtypes are less clear, and require molecular definition refinement before they can be reliably defined and diagnosed. The results of the six sub-cluster analysis using the Hierarchical-Ward method in this study showed sub-cluster 2 was unique in LumA subtype only. Whilst the seven sub-clusters analysis using SOM-Ward method showed that sub-cluster 4 was unique in LumA subtype too. Introducing the Ward clustering method to both the Hierarchical and SOM may therefore, provide a means for more distinction between different subtypes.

Chapter Six

6 CONCLUSIONS

The ability to classify tumours by identifying recurrent gene expression patterns of hundreds or thousands of genes could enable identification of combinations of marker genes that may not be recognized by standard methods and help to get a deeper understanding of the function of gene interplay. Therefore, gene profiles that relate to prognosis may help define new therapeutic targets. The Hierarchical clustering method has been used as a standard tool to identify different clusters, but several researchers have indicated that the result obtained by this method may need refining to get a more clear result.

This study introduced a new approach by combining the Ward method with the Hierarchical and SOM clustering methods. This approach produced more refined and detailed results about the properties of different clusters. For example, the same genes that are highly expressed in one subtype can be the most influential in one or more other subtypes. Thus, the genes with more discriminating ability may not be the ones on the top of the list (ranked according to absolute expression ratio levels) but the ones found somewhere lower on the list. This study found the few top most significant genes in each subtype (Table 4.1).

This study showed that both the Hierarchical-Ward and SOM-Ward analysis produce two main clusters; one containing LumB, LumA and the Normal Like subtypes, and the other contain the Basal Like and the Her2 subtypes. These finding are in agreement with previous studies in this field of research. Our results showed that the two main clusters can be divided in to 6 sub-clusters (Hierarchical-Ward method) as has been reported in the literature, but the present study with SOM-Ward method illustrated that 7 sub-clusters can also be obtained. Hierarchical-Ward and SOM-Ward results complemented each other very strongly that provided validity to the outcomes from the analyses conducted in this study. Some of the identified sub-clusters in this study showed significant similarities to the findings of Hu et al. (2006) in terms of the distribution of samples (patients) among different sub-clusters. Both methods used in this study demonstrated that some sub-clusters contain one subtype only both found one cluster for LumA and SOM-Ward found distinct clusters for LumA and Basal like. Although these clusters containing a single subtype did not contain all the patients belonging to that subtype, it is worth investigating if these unique clusters define the core behaviour of these subtypes.

The result of this study also showed that SOM-Ward method is a better tool for refining the results found from more basic cluster analysis methods.

7 RECOMMENDATIONS

We recommend the use of the simultaneous analysis method (Hierarchical-Ward and SOM-Ward) to obtain more detailed outcomes from gene expression studies. We recommend that more experimental work be conducted with SOM for investigating any further refinements that can be made to separating the subtypes further. For example, a larger SOM map size than the 25 neurons used in this study may provide more granularity to the separation of patients and could provide more clear separation boundaries for the subtypes owing to the ability of SOM to create complex nonlinear boundaries.

The Fuzzy C Means (FCM) or a related fuzzy clustering method is recommend as an alternative clustering approach that is generally considered as a better clustering tool for solving multiclass and ambiguous clustering problems than the Hierarchical and SOM clustering methods. Its goal is to determine the intrinsic grouping in a set of unlabeled data (Li, Lu et al. 2009). This method may provide extra information on the membership of patients in all subtypes thereby enabling the profiling of the core response of the subtypes and determining the likelihood of patients belonging to a particular subtype.

REFERENCES:

Ahn, J., A. Schatzkin, et al. (2007). "Adiposity, Adult Weight Change, and Postmenopausal Breast Cancer Risk." Arch Intern Med **167**(19): 2091-2102.

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature **403**(6769): 503-511.

Bair, E. and R. Tibshirani (2004). "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data." PLoS Biol **2**(4): e108.

Barlow, W. E., C. D. Lehman, et al. (2002). "Performance of Diagnostic Mammography for Women With Signs or Symptoms of Breast Cancer." Journal of the National Cancer Institute **94**(15): 1151-1159.

Battaglia, F., G. Scambia, et al. (1988). "Epidermal growth-factor receptor in human-breast cancer - correlation with steroid-hormone receptors and axillary lymph-node involvement." European Journal of Cancer & Clinical Oncology **24**(11): 1685-1690.

Benito, M., J. Parker, et al. (2004). "Adjustment of systematic microarray data biases." Bioinformatics **20**(1): 105 - 114.

Berg, W. A., L. Gutierrez, et al. (2004). "Diagnostic Accuracy of Mammography, Clinical Examination, US, and MR Imaging in Preoperative Assessment of Breast Cancer1." Radiology **233**(3): 830-849.

Bergh, J., T. Norberg, et al. (1995). "Complete sequencing of the p53 gene provides prognostic information in breast-cancer patients, particularly in relation to adjuvant systemic therapy and radiotherapy." Nature Medicine **1**(10): 1029-1034.

Berry, D. A., K. A. Cronin, et al. (2005). "Effect of Screening and Adjuvant Therapy on Mortality from Breast Cancer." New England Journal of Medicine **353**(17): 1784-1792.

Blake Cady, G. D. S., Jr., Monica Morrow, Bernard Gardner, David P. Winchester, (1998). "EVALUATION OF COMMON BREAST PROBLEMS: A PRIMER FOR PRIMARY CARE PROVIDERS." prepared by the Society of Surgical Oncology and the Commission on Cancer of the American College of Surgeons for the Centers for Disease Control and Prevention. (Available online at: www.utmb.edu/Surgery/clerks/primer.htm) (Accessed on August 8, 2008)

Bleicher, R. J., R. M. Ciocca, et al. (2009). "Association of Routine Pretreatment Magnetic Resonance Imaging with Time to Surgery, Mastectomy Rate, and Margin Status." Journal of the American College of Surgeons **209**(2): 180-187.

Borresen, A. L., T. I. Andersen, et al. (1995). "Tp53 mutations and breast-cancer prognosis - particularly poor survival rates for cases with mutations in the zinc-binding domains." Genes Chromosomes & Cancer **14**(1): 71-75.

Bradley, C. J., C. W. Given, et al. (2002). "Race, Socioeconomic Status, and Breast Cancer Treatment and Survival." Journal of the National Cancer Institute **94**(7): 490-496.

Carey, L., E. Dees, et al. (2007). "The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes." Clin Cancer Res **13**: 2329 - 2334.

Carey, L., C. Perou, et al. (2006). "Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study." JAMA **295**: 2492 - 2502.

Carlson, R. W., Anderson B.O. and Burstein H.J. et al (2005). "Breast Cancer Clinical Practice Guidelines in Oncology." Journal of the National Comprehensive Cancer Network (JNCCN) **3**: 238-238.

Clark, P. K. a. M., Ed. (2009). Clinical Medicine, Elsevier Health Sciences.

Cuzick, J., H. Stewart, et al. (1994). "Cause-specific mortality in long-term survivors of breast cancer who participated in trials of radiotherapy." Journal of Clinical Oncology **12**(3): 447-453.

Deapen, D., L. Liu, et al. (2002). "Rapidly rising breast cancer incidence rates among Asian-American women." International Journal of Cancer **99**(5): 747-750.

Eliassen, A. H., G. A. Colditz, et al. (2006). "Adult Weight Change and Risk of Postmenopausal Breast Cancer." JAMA: The Journal of the American Medical Association **296**(2): 193-201.

Elston, C. W. and I. O. Ellis (1991). "Pathological prognostic factors in breast-cancer .1. the value of histological grade in breast-cancer - experience from a large study with long-term follow-up." Histopathology **19**(5): 403-410.

Fan, C., D. Oh, et al. (2006). "Concordance among gene-expression-based predictors for breast cancer." N Engl J Med **355**: 560 - 569.

Feigelson, H. S., C. R. Jonas, et al. (2004). "Weight Gain, Body Mass Index, Hormone Replacement Therapy, and Postmenopausal Breast Cancer in a Large Prospective Study." Cancer Epidemiology Biomarkers & Prevention **13**(2): 220-224.

Fisher, E. R., J. Costantino, et al. (1993). "Pathological findings from the national surgical adjuvant breast project (protocol-4) - discriminants for 15-year survival." Cancer **71**(6): 2141-2150.

Foekens, J. A., M. P. Look, et al. (1999). "Cathepsin-D in primary breast cancer: prognostic evaluation involving 2810 patients." British Journal of Cancer **79**(2): 300-307.

Foulkes, W. D., J.-S. b. Brunet, et al. (2004). "The Prognostic Implication of the Basal-Like (Cyclin Ehigh/p27low/p53+/Glomeruloid-Microvascular-Proliferation+) Phenotype of BRCA1-Related Breast Cancer." Cancer Research **64**(3): 830-835.

Goldhirsch, A., J. H. Glick, et al. (2005). "Meeting Highlights: International Expert Consensus on the Primary Therapy of Early Breast Cancer 2005." Ann Oncol **16**(10): 1569-1583.

Golub, T. R., D. K. Slonim, et al. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." Science **286**(5439): 531-537.

Gruvberger, S., M. Ringnér, et al. (2001). "Estrogen Receptor Status in Breast Cancer Is Associated with Remarkably Distinct Gene Expression Patterns." Cancer Research **61**(16): 5979-5984.

Guibout, C., E. Adjadj, et al. (2005). "Malignant Breast Tumors After Radiotherapy for a First Cancer During Childhood." Journal of Clinical Oncology **23**(1): 197-204.

He, Y. D. and S. H. Friend (2001). "Microarrays[mdash]the 21st century divining rod?" Nat Med **7**(6): 658-659.

Hedenfalk, I., D. Duggan, et al. (2001). "Gene-expression profiles in hereditary breast cancer." New England Journal of Medicine **344**(8): 539-548.

Howat, J. M. T., D. M. Barnes, et al. (1983). "The association of cytosol estrogen and progesterone receptors with histological features of breast-cancer and early recurrence of disease." British Journal of Cancer **47**(5): 629-640.

Hu, Z., C. Fan, et al. (2006). "The molecular portraits of breast tumors are conserved across microarray platforms." BMC Genomics **7**(1): 96.

Jatoi, I., B. E. Chen, et al. (2007). "Breast Cancer Mortality Trends in the United States According to Estrogen Receptor Status and Age at Diagnosis." Journal of Clinical Oncology **25**(13): 1683-1690.

Jemal, A., R. Siegel, et al. (2010). "Cancer Statistics, 2010." CA Cancer J Clin: caac.20073.

Jemal, A., M. J. Thun, et al. (2008). "Annual Report to the Nation on the Status of Cancer, 1975â€“2005, Featuring Trends in Lung Cancer, Tobacco Use, and Tobacco Control." Journal of the National Cancer Institute **100**(23): 1672-1694.

Johnson, J. W. C. (2001). "The millennial mark: Presidential address." American journal of obstetrics and gynecology **185**(2): 261-267.

Kelsey, J. L., M. D. Gammon, et al. (1993). "Reproductive Factors and Breast Cancer." Epidemiologic Reviews **15**(1): 36-47.

Khan J, W. J., Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. (2001). "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." Naure Med **7**: 673-679

Kohonen, T. (1995c). "Self-Organizing Maps." Springer,Berlin, Heidelberg.

Lacey, J. V., S. S. Devesa, et al. (2002). Recent trends in breast cancer incidence and mortality, Wiley Subscription Services, Inc., A Wiley Company. **39:** 82-88.

Lahmann, P. H., K. Hoffmann, et al. (2004). Body size and breast cancer risk: Findings from the European prospective investigation into cancer and nutrition (EPIC), Wiley Subscription Services, Inc., A Wiley Company. **111:** 762-771.

Li, X., X. Lu, et al. (2009). "Application of Fuzzy c-Means Clustering in Data Analysis of Metabolomics." Analytical Chemistry **81**(11): 4468-4475.

Lichtenstein, P., N. V. Holm, et al. (2000). "Environmental and Heritable Factors in the Causation of Cancer â€” Analyses of Cohorts of Twins from Sweden, Denmark, and Finland." New England Journal of Medicine **343**(2): 78-85.

Loi, S., B. Haibe-Kains, et al. (2007). "Definition of Clinically Distinct Molecular Subtypes in Estrogen ReceptorÂ–Positive Breast Carcinomas Through Genomic Grade." Journal of Clinical Oncology **25**(10): 1239-1246.

Lund, M., K. Trivers, et al. (2009). "Race and triple negative threats to breast cancer survival: a population-based study in Atlanta, GA." Breast Cancer Research and Treatment **113**(2): 357-370.

Millikan, R., B. Newman, et al. (2008). "Epidemiology of basal-like breast cancer." Breast Cancer Research and Treatment **109**(1): 123-139.

Monica Morrow, M. D. (2000). "The Evaluation of Common Breast Problems." Am Fam Physician **61**(8): 2371.

Morimoto, L. M., E. White, et al. (2002). "Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States)." Cancer Causes and Control **13**(8): 741-751.

Olivotto, I. A., C. D. Bajdik, et al. (2005). "Population-Based Validation of the Prognostic Model ADJUVANT! for Early Breast Cancer." J Clin Oncol **23**(12): 2716-2725.

Olopade, O. I. and T. Grushko (2001). "Gene-Expression Profiles in Hereditary Breast Cancer." New England Journal of Medicine **344**(26): 2028-2029.

Ostroumova, E., D. L. Preston, et al. (2008). "Breast cancer incidence following low-dose rate environmental exposure: Techa River Cohort, 1956-2004." Br J Cancer **99**(11): 1940-1945.

Palmer, J. R., L. A. Wise, et al. (2003). "Dual Effect of Parity on Breast Cancer Risk in African-American Women." Journal of the National Cancer Institute **95**(6): 478-483.

Parker, J., M. Mullins, et al. (2009). "Supervised risk predictor of breast cancer based on intrinsic subtypes." J Clin Oncol **27**: 1160 - 1167.

Parkin, D. M., F. Bray, et al. (2005). "Global Cancer Statistics, 2002." CA Cancer J Clin **55**(2): 74-108.

Perou, C. M., T. Sorlie, et al. (2000). "Molecular portraits of human breast tumours." Nature **406**(6797): 747-752.

Peto, J. and T. M. Mack (2000). "High constant incidence in twins and other relatives of women with breast cancer." Nat Genet **26**(4): 411-414.

Pike, M. C., D. V. Spicer, et al. (1993). "Estrogens, Progestogens, Normal Breast Cell Proliferation, and Breast Cancer Risk." Epidemiologic Reviews **15**(1): 17-30.

Pisano, E. D., L. L. Fajardo, et al. (2001). "Fine-Needle Aspiration Biopsy of Nonpalpable Breast Lesions in a Multicenter Clinical Trial: Results from the Radiologic Diagnostic Oncology Group V1." Radiology **219**(3): 785-792.

Pukkala, E., A. Kesminiene, et al. (2006). Breast cancer in Belarus and Ukraine after the Chernobyl accident, Wiley Subscription Services, Inc., A Wiley Company. **119:** 651-658.

Ravdin, P. M., L. A. Siminoff, et al. (2001). "Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer." J Clin Oncol **19**(4): 980-991.

Rouzier, R., C. M. Perou, et al. (2005). "Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy." Clinical Cancer Research **11**(16): 5678-5685.

Samarasinghe, S. (2007). Neural Networks for Applied Sciences and Engineering.

Sasco, A. J., A. B. Lowenfels, et al. (1993). Review article: Epidemiology of male breast cancer. A meta-analysis of published case-control studies and discussion of selected aetiological factors, Wiley Subscription Services, Inc., A Wiley Company. **53:** 538-549.

Slamon, D. J., W. Godolphin, et al. (1989). "Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer." Science **244**(4905): 707-712.

Society, A. C. (2009 - 2010). "Breast Cancer Facts & Figures 2009-2010. Atlanta: American Cancer Society, Inc."

Sorlie, T., C. M. Perou, et al. (2001). "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proceedings of the National Academy of Sciences of the United States of America **98**(19): 10869-10874.

Sorlie, T., R. Tibshirani, 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**(14): 8418-8423.

Sotiriou, C., S. Neo, et al. (2003). "Breast cancer classification and prognosis based on gene expression profiles from a population-based study." Proceedings of the National Academy of Sciences of the United States of America **100**(18): 10393-10398.

Stomper, P., Winston, JS, Proulx, GM, et al (2000). "Mammographic detection and staging of ductal carcinoma in situ: mammographic-pathologic correlation." **3**: 1.

Torregrosa, D., P. Bolufer, et al. (1997). "Prognostic significance of c-erbB-2/neu amplification and epidermal growth factor receptor (EGFR) in primary breast cancer and their relation to estradiol receptor (ER) status." Clinica Chimica Acta **262**(1-2): 99-119.

van 't Veer, L. J., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530 - 536.

van den Brandt, P. A., D. Spiegelman, et al. (2000). "Pooled Analysis of Prospective Cohort Studies on Height, Weight, and Breast Cancer Risk." American Journal of Epidemiology **152**(6): 514-527.

Verkooijen, H. M. (2002). Diagnostic accuracy of stereotactic large-core needle biopsy for nonpalpable breast disease: Results of a multicenter prospective study with 95% surgical confirmation, Wiley Subscription Services, Inc., A Wiley Company. **99:** 853-859.

Voduc, K. D., M. C. U. Cheang, et al. "Breast Cancer Subtypes and the Risk of Local and Regional Relapse." Journal of Clinical Oncology **28**(10): 1684-1691.

Vollenweiderzerargui, L., L. Barrelet, et al. (1986). "The predictive value of estrogen and progesterone receptors concentrations on the clinical behavior of breast-cancer in women - clinical correlation on 547 patients." Cancer **57**(6): 1171-1180.

Wang, Y., J. G. M. Klijn, et al. (2005). "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." The Lancet **365**(9460): 671-679.

Ward, J. H. J. (1963). "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association **58**(236).

Appendix A

Pilot results illustrating individual breast cancer subtypes.

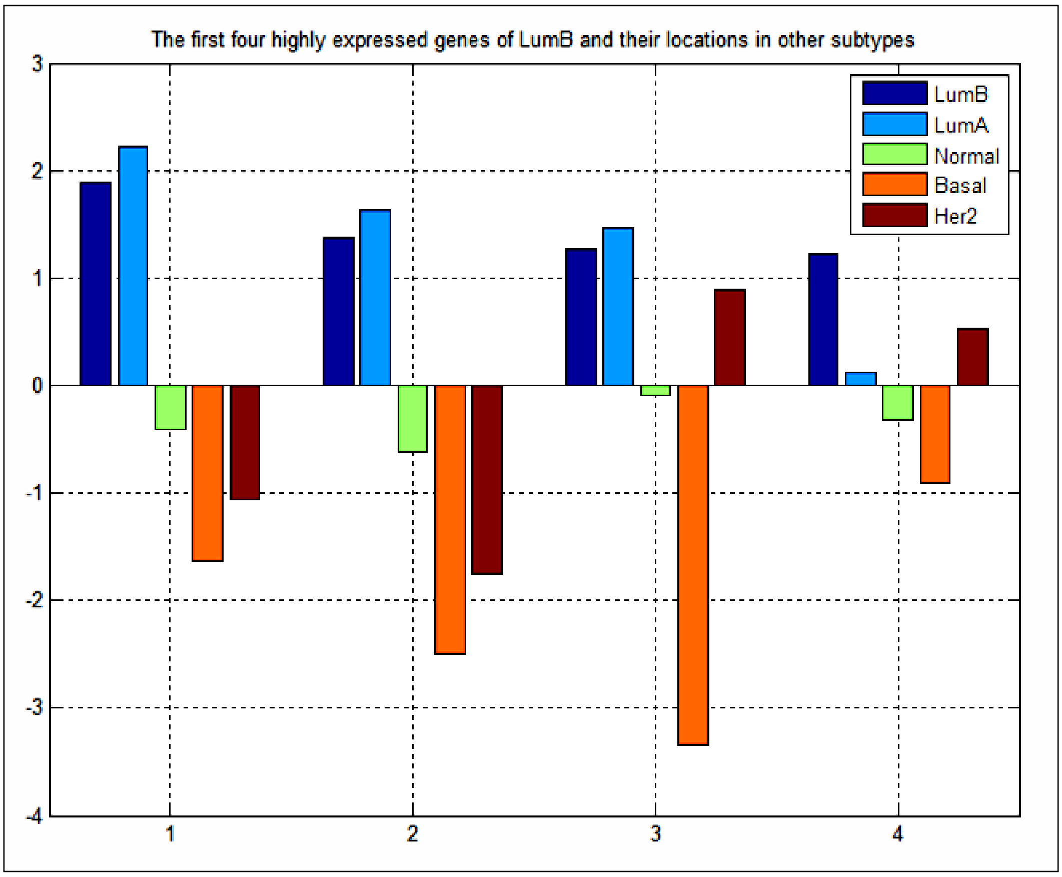


Figure 18: Bar graph illustrating the four most highly expressed genes in LumB vs. the expression of the same genes in the other subtypes based on median log ratio.

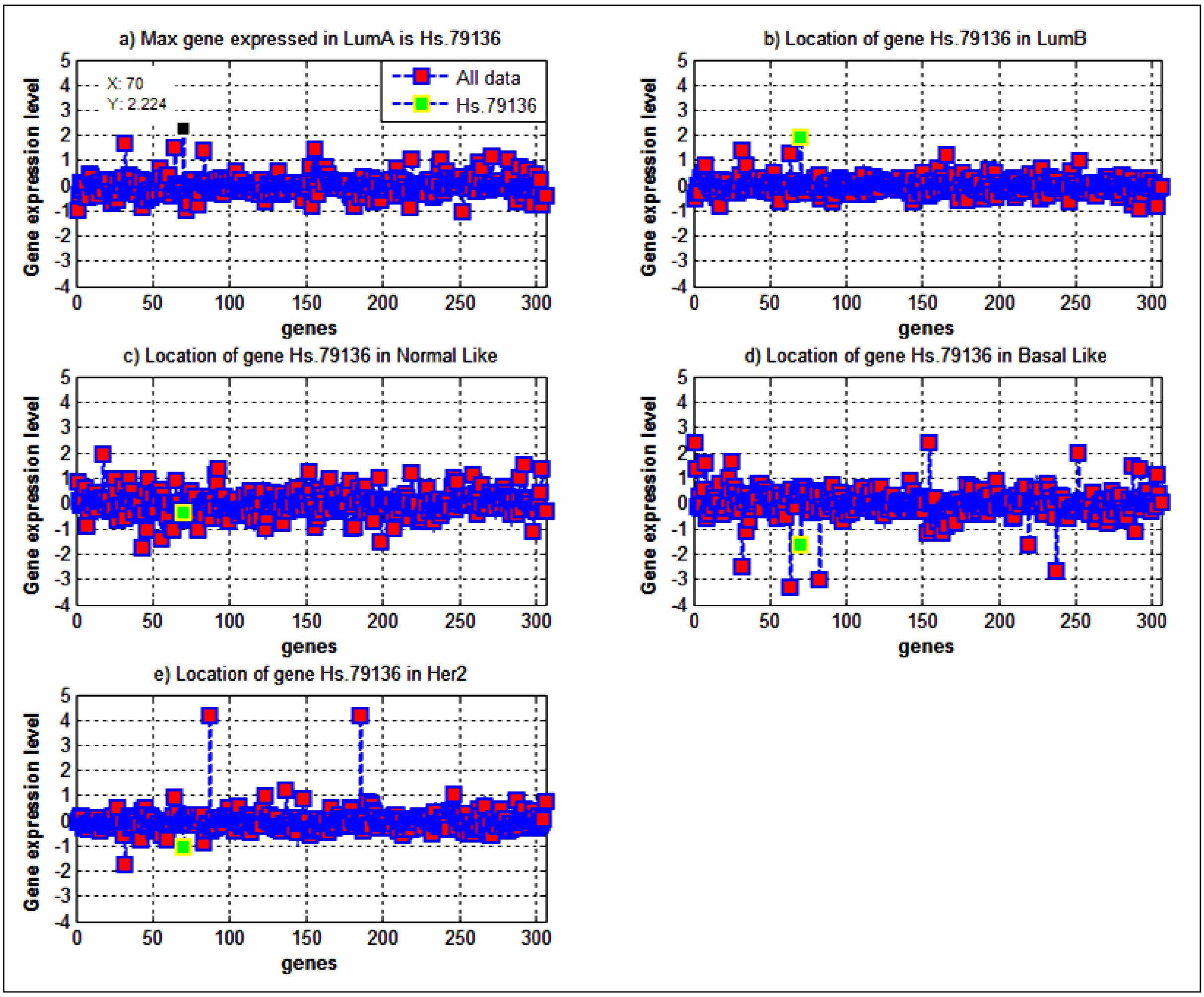


Figure 19: The median expression of the highly expressed gene in LumA (black colour) vs. its expression (green colour) in the rest of the examined subtypes.

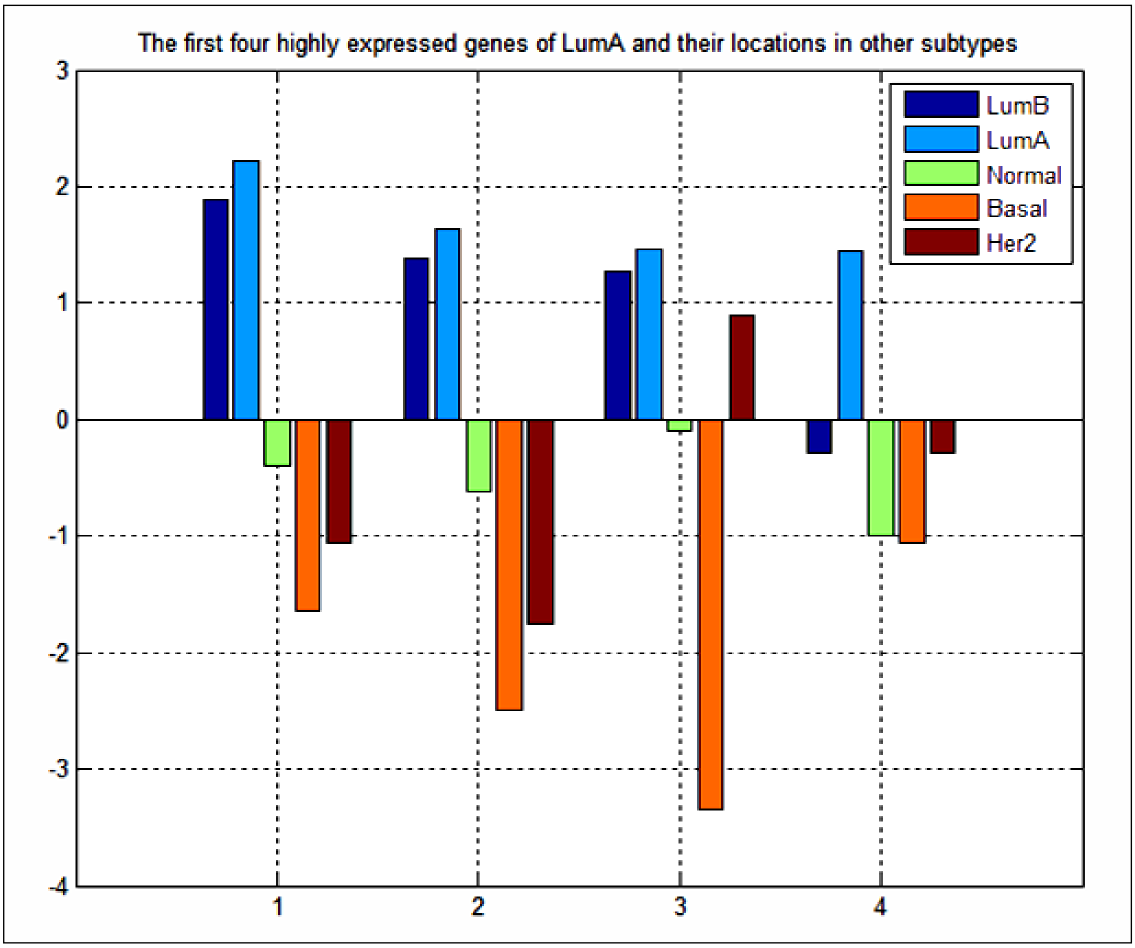


Figure 20: Bar graph illustrating the four most highly expressed genes in LumA vs. the expression of the same genes in the other subtypes based on median log ratio.

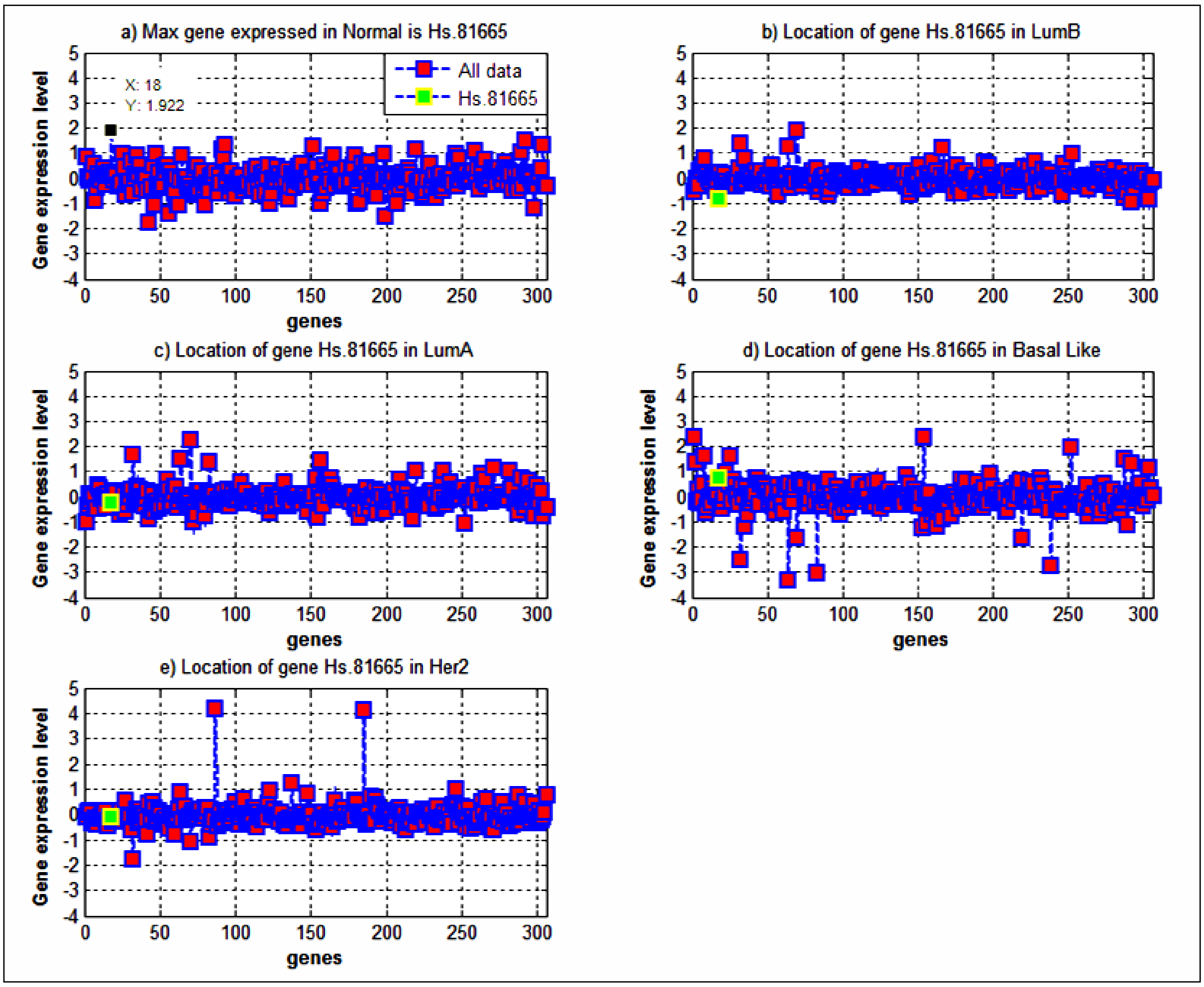


Figure 21: The median expression of the highly expressed gene in Normal (black colour) vs. the same gene expression in the rest of the subtypes (green colour).

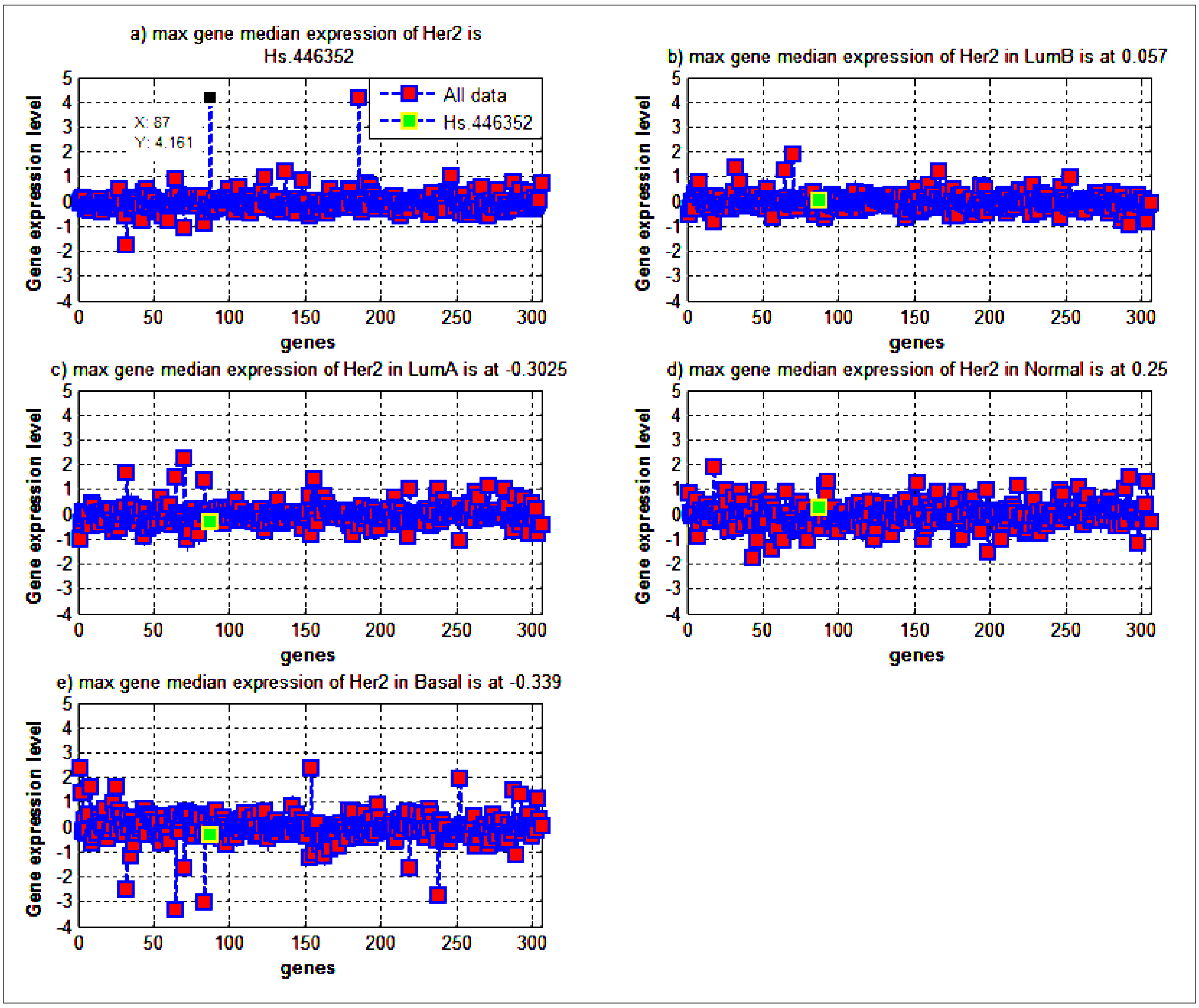


Figure 22: The median expression of the highly expressed gene in Her2 (black colour) vs. the same gene expression in the rest of the subtypes (green colour).

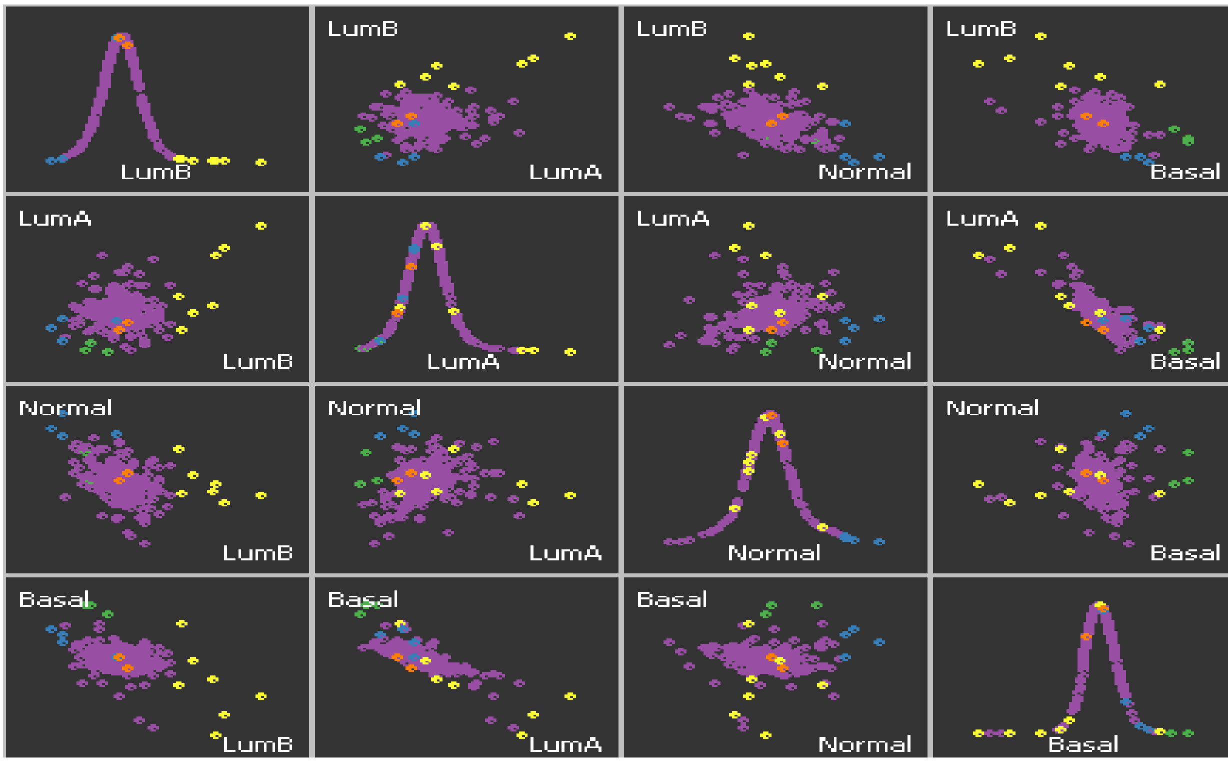


Figure 23: GGobi software based display of a scatterplot matrix for the 306 genes across the 249 patients for the subtypes highlighting only the most highly expressed genes visually and labelling them with different colours.

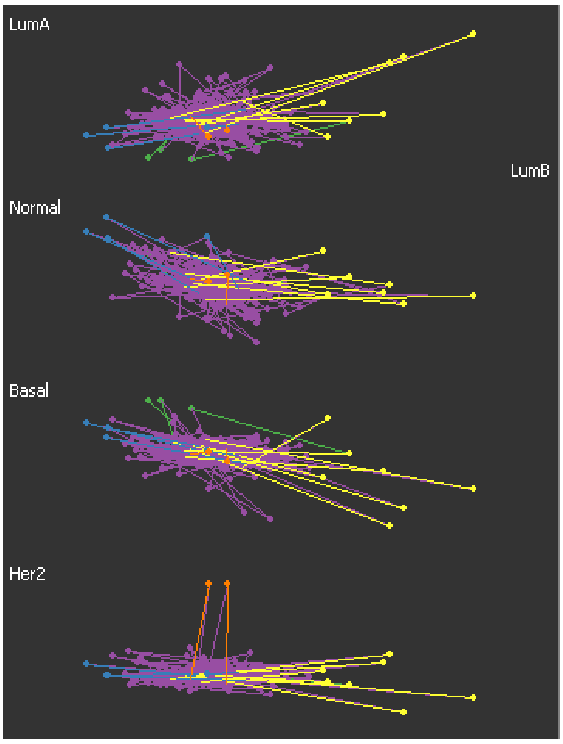


Figure 24: GGobi software based display of a time series (sequential) plot for the 306 genes across the 249 patients in the five subtypes highlighting only the most highly expressed genes visually and labelling them with different colours.

(Figure 23and Figure 24) Different format of the same data as in Figure 5 this led to the same conclusion with regard to identifying the subtypes.

Appendix B

Hierarchical clustering results with several linkage methods.

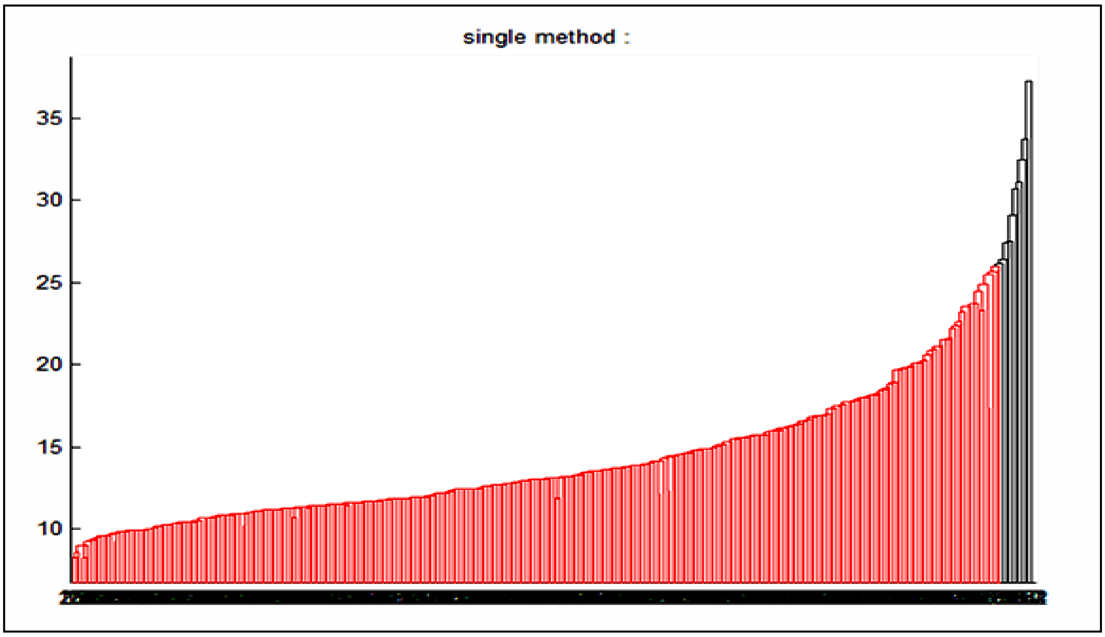


Figure 25: A dendrogram using hierarchical single linkage method representing the distribution of patients according to their correlation strength to clusters.

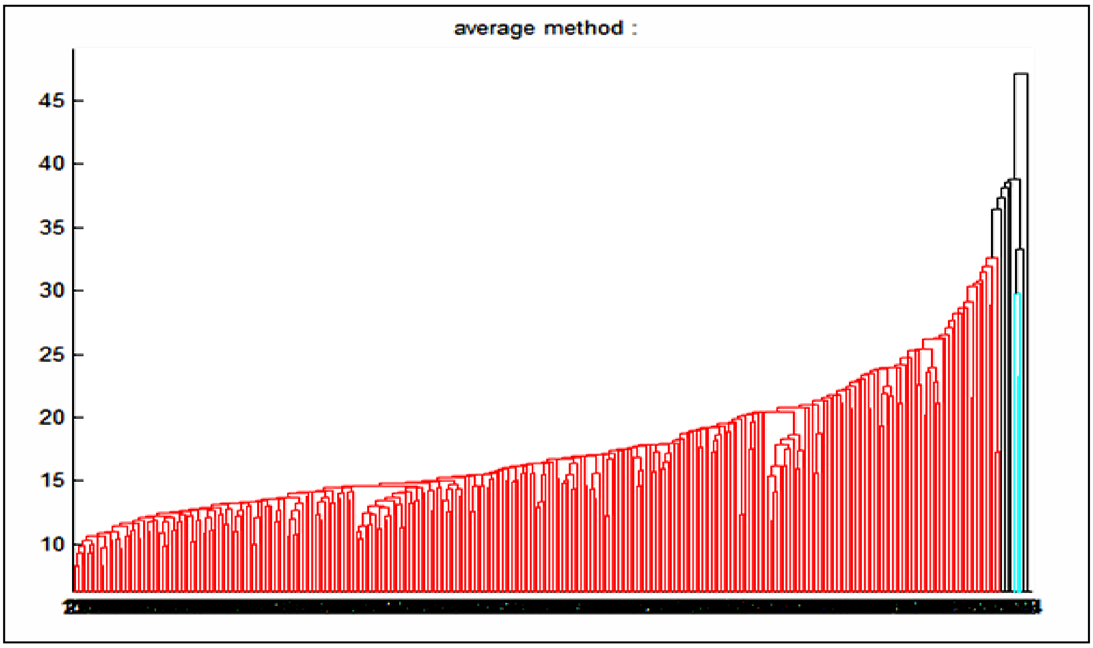


Figure 26: A dendrogram using hierarchical average method representing the distribution of patients according to their correlation strength to clusters.

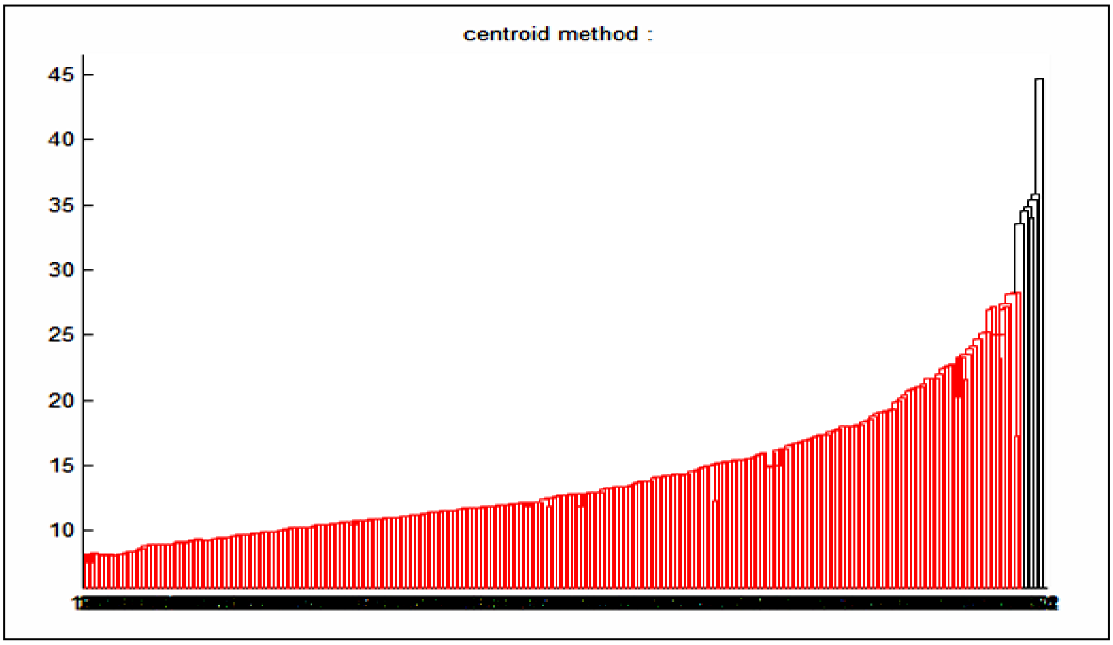


Figure 27: A dendrogram using hierarchical centroid method representing the distribution of patients according to their correlation strength to clusters.

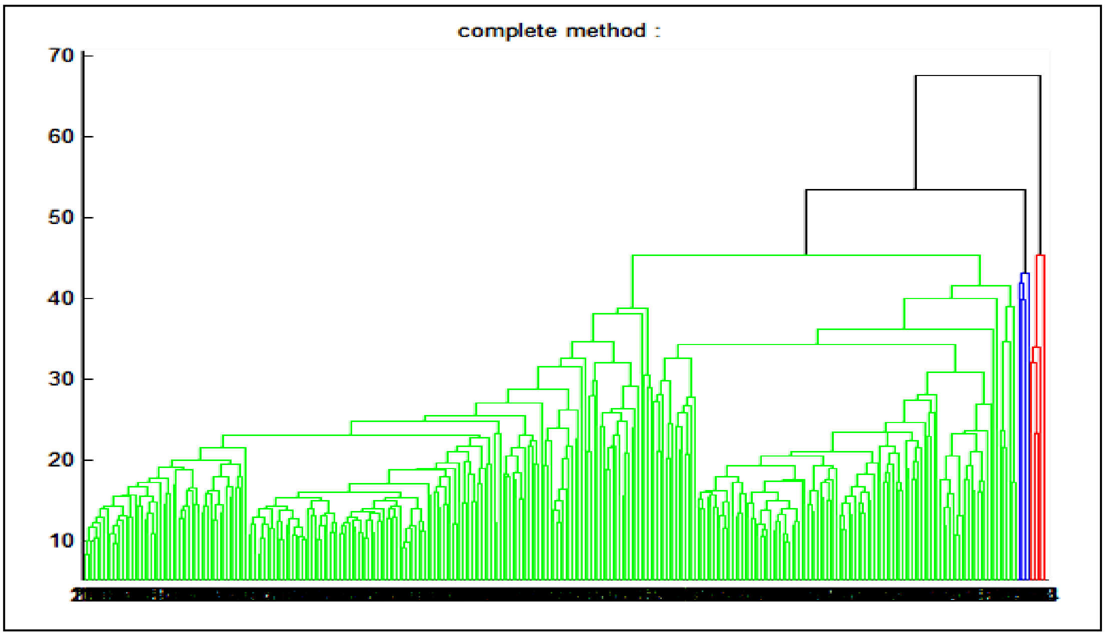


Figure 28: A dendrogram using hierarchical complete linkage method representing the distribution of patients according to their correlation strength to clusters.

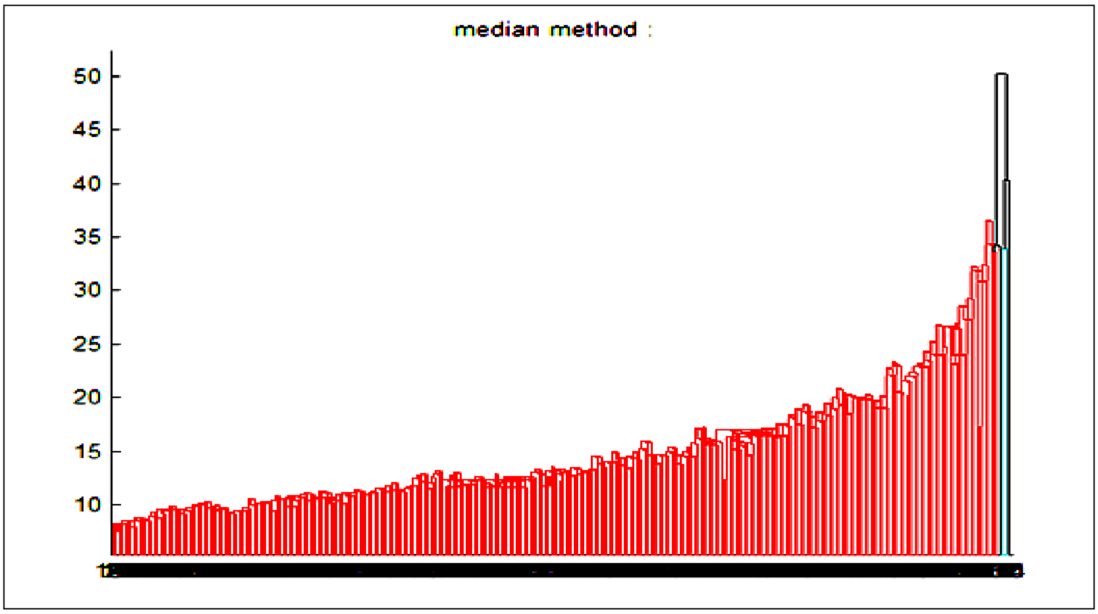


Figure 29: A dendrogram using hierarchical median method representing the distribution of patients according to their correlation strength to clusters.

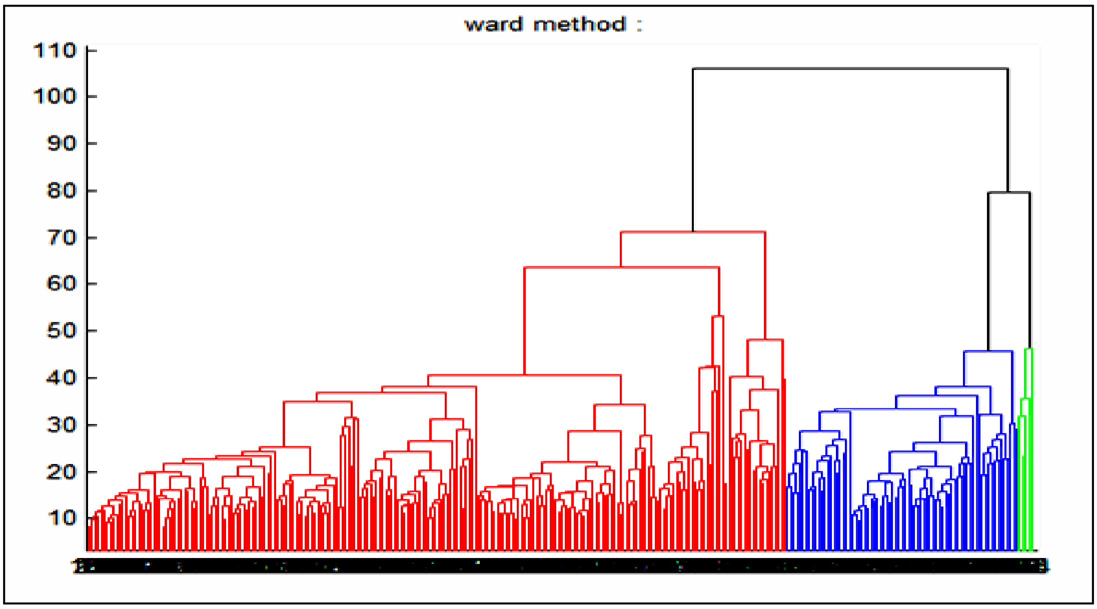


Figure 30: A dendrogram using hierarchical Ward method representing the distribution of patients according to their correlation strength to clusters.

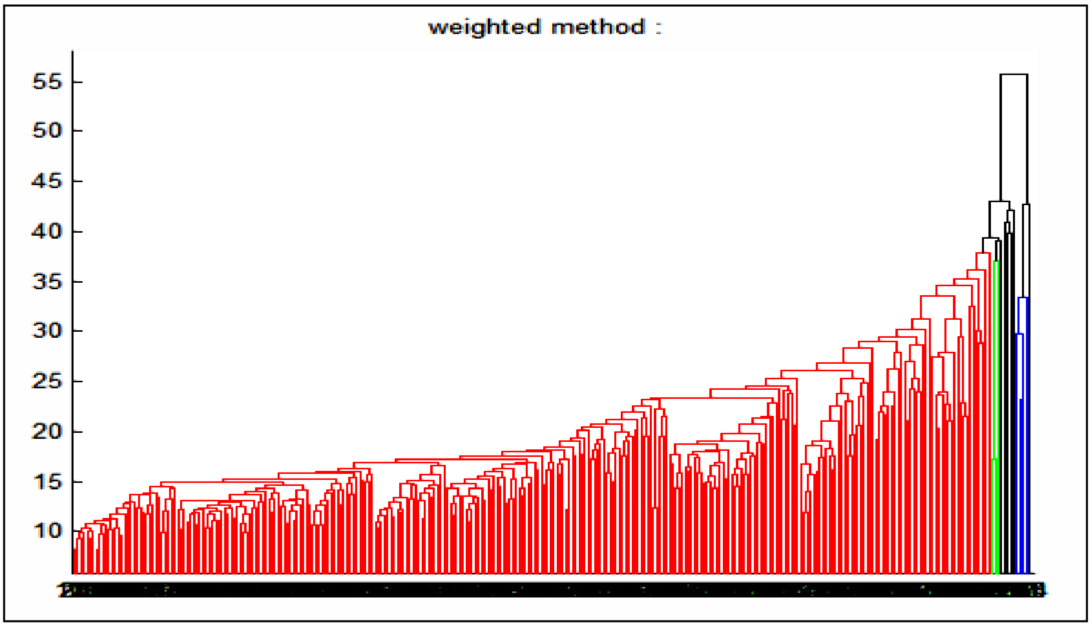


Figure 31: A dendrogram using hierarchical weighted method representing the distribution of patients according to their correlation strength to clusters.

Appendix C

Patient labels in clusters from Hierarchical-Ward clustering.

Hierarchical Two Clusters

Cluster 1	Cluster 2
1	157
2	158
3	159
4	160
5	161
6	162
7	163
8	164
9	165
10	166
11	167
12	168
13	169
14	170
15	171
16	172
17	173
18	174
19	175
20	176
21	177
22	178
23	179
24	180
25	181
26	182
27	183
28	184
29	185
30	186
31	187
32	188
33	189
34	190
35	191
36	192
37	193
38	194
39	195
40	196
41	197
42	198
43	199
44	200
45	201
46	202
47	203
48	204
49	205
50	206
51	207
52	208
53	209
54	210
55	211
56	212
57	213
58	214
59	215

Hierarchical Five Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	6	46	157	158
2	18	47	159	213
3	19	48	160	214
4	20	49	161	215
5	23	50	162	216
7	28	51	163	217
8	29	52	164	218
9	30	55	165	219
10	31	56	166	221
11	32	65	167	222
12	33	66	168	223
14	35	68	170	225
15	39	69	171	226
16	54	88	172	227
17	57	89	173	228
21	58	90	174	229
22	59	91	175	230
24	60	92	176	231
25	61	93	177	232
26	62	96	178	233
27	63	97	179	235
36	64	98	180	236
37	70	99	181	237
38	71	100	182	238
40	72	101	183	239
41	76	108	184	240
42	77	110	185	241
43	78	111	186	243
44	79	115	187	244
45	80	116	188	245
53	81	117	189	
73	82	118	190	
74	83	119	191	
75	84	120	192	
112	85	121	193	
128	86	122	194	
156	87	123	195	
242	94	124	196	
249	95	125	197	
	102	127	198	
	103	129	199	
	104	131	200	
	105	132	201	
	106	133	202	
	107	135	203	
	109	137	204	
	113	138	205	
	114	139	206	
	126	140	207	
	130	141	208	
	134	142	209	
	136	143	210	
	155	144	211	
		145	212	
		146	220	
		147	234	
		148		
		149		
		150		

60	216
61	217
62	218
63	219
64	220
65	221
66	222
67	223
68	224
69	225
70	226
71	227
72	228
73	229
74	230
75	231
76	232
77	233
78	234
79	235
80	236
81	237
82	238
83	239
84	240
85	241
86	243
87	244
88	245
89	
90	
91	
92	
93	
94	
95	
96	
97	
98	
99	
100	
101	
102	
103	
104	
105	
106	
107	
108	
109	
110	
111	
112	
113	
114	
115	
116	
117	
118	
119	
120	
121	
122	
123	
124	
125	
126	
127	
128	
129	
130	

		151		
		152		
		153		
		154		
		246		
		247		
		248		

131	
132	
133	
134	
135	
136	
137	
138	
139	
140	
141	
142	
143	
144	
145	
146	
147	
148	
149	
150	
151	
152	
153	
154	
155	
156	
242	
246	
247	
248	
249	

Hierarchical Six Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
46	52	1	6	157	158
47	55	2	18	159	213
48	56	3	19	160	214
49	88	4	20	161	215
50	89	5	23	162	216
51	90	7	28	163	217
65	91	8	29	164	218
66	92	9	30	165	219
67	93	10	31	166	221
68	96	11	32	167	222
69	97	12	33	168	223
110	98	13	34	169	224
111	99	14	35	170	225
115	100	15	39	171	226
116	101	16	54	172	227
117	108	17	57	173	228
118	121	21	58	174	229
119	122	22	59	175	230
120	123	24	60	176	231
127	124	25	61	177	232
129	125	26	62	178	233
133	131	27	63	179	235
135	132	36	64	180	236

137		37	70	181	237
138		38	71	182	238
139		40	72	183	239
140		41	76	184	240
141		42	77	185	241
142		43	78	186	243
143		44	79	187	244
144		45	80	188	245
145		53	81	189	
146		73	82	190	
147		74	83	191	
148		75	84	192	
149		112	85	193	
150		128	86	194	
151		156	87	195	
152		242	94	196	
153		249	95	197	
154			102	198	
246			103	199	
247			104	200	
248			105	201	
			106	202	
			107	203	
			109	204	
			113	205	
			114	206	
			126	207	
			130	208	
			134	209	
			136	210	
			155	211	
				212	
				220	
				234	

Appendix D

Patient labels in clusters from SOM-Ward clustering.

SOM Two Clusters

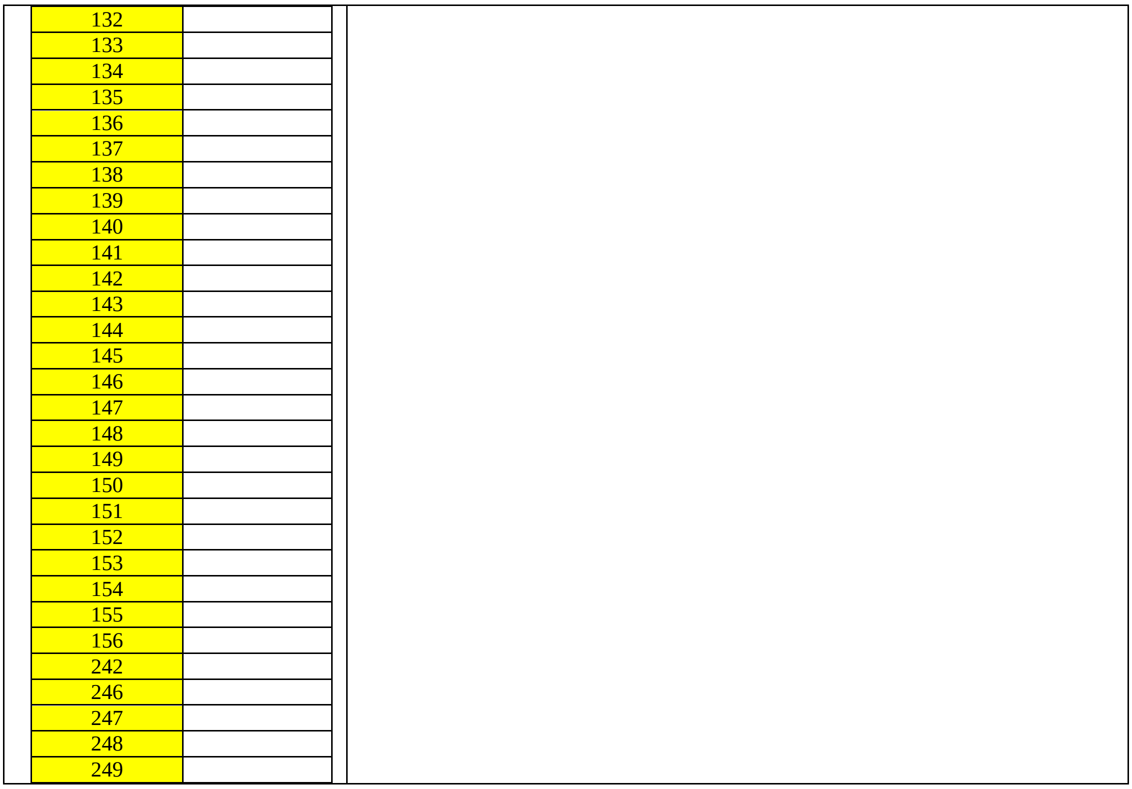
Cluster 1	Cluster 2
1	157
2	158
3	159
4	160
5	161
6	162
7	163
8	164
9	165
10	166
11	167
12	168
13	169
14	170
15	171
16	172
17	173
18	174
19	175
20	176
21	177
22	178
23	179
24	180
25	181
26	182
27	183
28	184
29	185
30	186
31	187
32	188
33	189
34	190
35	191
36	192
37	193
38	194
39	195
40	196
41	197
42	198
43	199
44	200
45	201
46	202
47	203
48	204
49	205
50	206
51	207
52	208
53	209
54	210
55	211
56	212
57	213
58	214
59	215
60	216

SOM Five Clusters

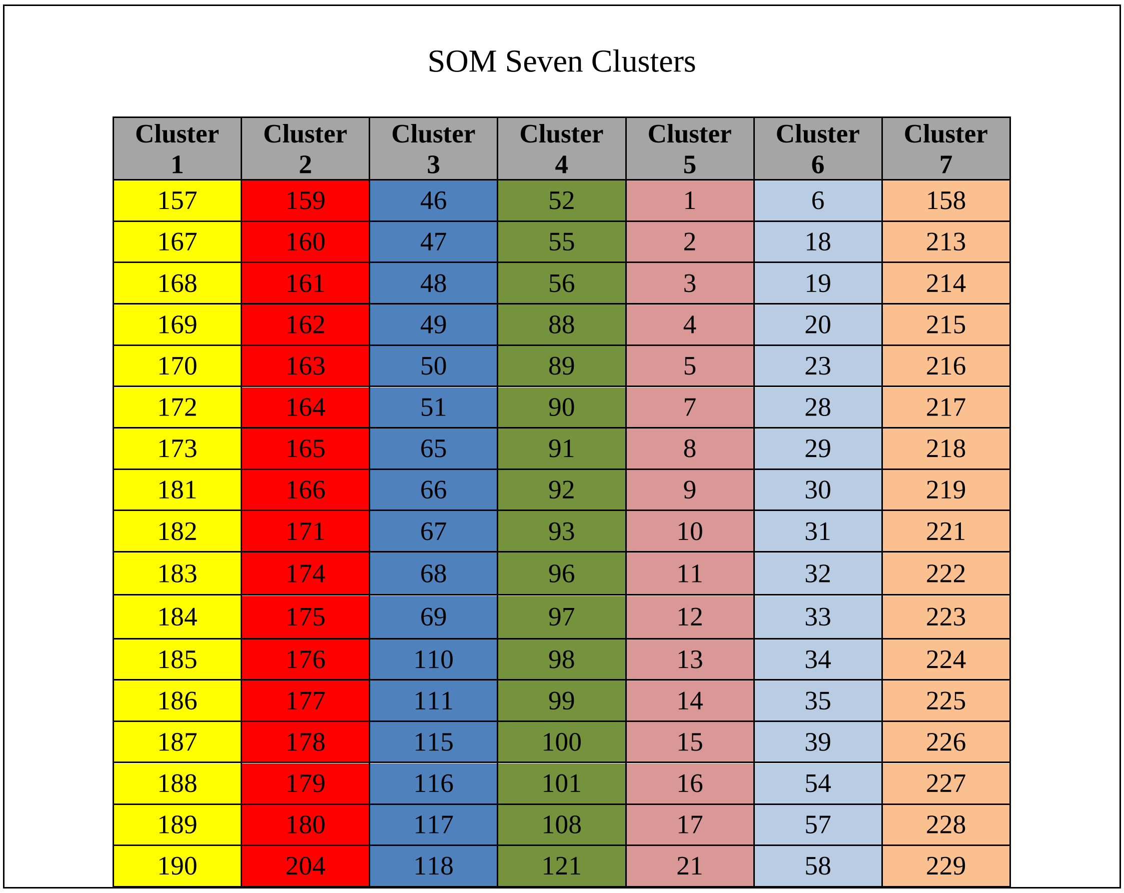
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	6	46	157	158
2	18	47	159	213
3	19	48	160	214
4	20	49	161	215
5	23	50	162	216
7	28	51	163	217
8	29	52	164	218
9	30	55	165	219
10	31	56	166	221
11	32	65	167	222
12	33	66	168	223
14	35	68	170	225
15	39	69	171	226
16	54	88	172	227
17	57	89	173	228
21	58	90	174	229
22	59	91	175	230
24	60	92	176	231
25	61	93	177	232
26	62	96	178	233
27	63	97	179	235
36	64	98	180	236
37	70	99	181	237
38	71	100	182	238
40	72	101	183	239
41	76	108	184	240
42	77	110	185	241
43	78	111	186	243
44	79	115	187	244
45	80	116	188	245
53	81	117	189	
73	82	118	190	
74	83	119	191	
75	84	120	192	
112	85	121	193	
128	86	122	194	
156	87	123	195	
242	94	124	196	
249	95	125	197	
	102	127	198	
	103	129	199	
	104	131	200	
	105	132	201	
	106	133	202	
	107	135	203	
	109	137	204	
	113	138	205	
	114	139	206	
	126	140	207	
	130	141	208	
	134	142	209	
	136	143	210	
	155	144	211	
		145	212	
		146	220	
		147	234	
		148		
		149		
		150		
		151		

61	217
62	218
63	219
64	220
65	221
66	222
67	223
68	224
69	225
70	226
71	227
72	228
73	229
74	230
75	231
76	232
77	233
78	234
79	235
80	236
81	237
82	238
83	239
84	240
85	241
86	243
87	244
88	245
89	
90	
91	
92	
93	
94	
95	
96	
97	
98	
99	
100	
101	
102	
103	
104	
105	
106	
107	
108	
109	
110	
111	
112	
113	
114	
115	
116	
117	
118	
119	
120	
121	
122	
123	
124	
125	
126	
127	
128	
129	
130	
131	

		152		
		153		
		154		
		246		
		247		
		248		



132	
133	
134	
135	
136	
137	
138	
139	
140	
141	
142	
143	
144	
145	
146	
147	
148	
149	
150	
151	
152	
153	
154	
155	
156	
242	
246	
247	
248	
249	



SOM Seven Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
157	159	46	52	1	6	158
167	160	47	55	2	18	213
168	161	48	56	3	19	214
169	162	49	88	4	20	215
170	163	50	89	5	23	216
172	164	51	90	7	28	217
173	165	65	91	8	29	218
181	166	66	92	9	30	219
182	171	67	93	10	31	221
183	174	68	96	11	32	222
184	175	69	97	12	33	223
185	176	110	98	13	34	224
186	177	111	99	14	35	225
187	178	115	100	15	39	226
188	179	116	101	16	54	227
189	180	117	108	17	57	228
190	204	118	121	21	58	229

191	205	119	122	22	59	230
192	211	120	123	24	60	231
193	212	127	124	25	61	232
194	234	129	125	26	62	233
195		133	131	27	63	235
196		135	132	36	64	236
197		137		37	70	237
198		138		38	71	238
199		139		40	72	239
200		140		41	76	240
201		141		42	77	241
202		142		43	78	243
203		143		44	79	244
206		144		45	80	245
207		145		53	81	
208		146		73	82	
209		147		74	83	
210		148		75	84	
220		149		112	85	
		150		128	86	
		151		156	87	
		152		242	94	
		153		249	95	
		154			102	
		246			103	
		247			104	
		248			105	
					106	
					107	
					109	
					113	
					114	
					126	
					130	
					134	
					136	
					155	