AN INTEGRATED BIOINFORMATICS APPROACH FOR THE IDENTIFICATION OF MELANOMA-ASSOCIATED BIOMARKER GENES

W. LIU

PhD

UNIVERSITY OF BRADFORD

2014

An Integrated Bioinformatics Approach for the Identification of Melanoma-Associated Biomarker Genes

A Ranking and Stratification Approach as a New Meta-Analysis Methodology for the Detection of Robust Gene Biomarker Signatures of Cancers

Wanting Liu

BSc (Hons)

Submitted for the degree of

Doctor of Philosophy

School of Engineering and Informatics

University of Bradford

2014

ABSTRACT

Wanting Liu

AN INTEGRATED BIOINFORMATICS APPROACH FOR THE IDENTIFICATION OF MELANOMA-ASSOCIATED BIOMARKER GENES

A ranking and stratification approach as a new meta-analysis methodology for the detection of robust gene biomarker signatures of cancers

Keywords:

Melanoma, Microarray, Meta-analysis, Biomarker, Integrated Analysis

Genome-wide microarray technology has facilitated the systematic discovery of diagnostic biomarkers of cancers and other pathologies. However, meta-analyses of published arrays using melanoma as a test cancer has uncovered significant inconsistences that hinder advances in clinical practice. In this study a computational model for the integrated analysis of microarray datasets is proposed in order to provide a robust ranking of genes in terms of their relative significance; both genome-wide relative significance (GWRS) and genome-wide global significance (GWGS).

When applied to five melanoma microarray datasets published between 2000 and 2011, a new 12-gene diagnostic biomarker signature for melanoma was defined (i.e., *EGFR, FGFR2, FGFR3, IL8, PTPRF, TNC, CXCL13, COL11A1, CHP2, SHC4, PPP2R2C,* and *WNT4*). Of these, *CXCL13, COL11A1, PTPRF*

and *SHC4* are components of the MAPK pathway and were validated by immunocyto- and immunohisto-chemistry. These proteins were found to be overexpressed in metastatic and primary melanoma cells *in vitro* and in melanoma tissue *in situ* compared to melanocytes cultured from healthy skin epidermis and normal healthy human skin.

One challenge for the integrated analysis of microarray data is that the microarray data are produced using different platforms and bio-samples, e.g. including both cell line- and biopsy-based microarray datasets. In order to address these challenges, the computational model was further enhanced the stratification of datasets into either biopsy or cell line derived datasets, and via the weighting of microarray data based on quality criteria of data. The methods enhancement was applied to 14 microarray datasets of three cancers (breast, prostate, and melanoma) based on classification accuracy and on the capability to identify predictive biomarkers. Four novel measures for evaluating the capability to identify predictive biomarkers are proposed: (1) classifying independent testing data using wrapper feature selection with machine leaning, (2) assessing the number of common genes with the genes retrieved in independent testing data, (3) assessing the number of common genes with the genes retrieved in across multiple training datasets, (4) assessing the number of common genes with the genes validated in the literature.

This enhancement of computational approach (i) achieved reliable classification performance across multiple datasets, (ii) recognized more significant genes into the top-ranked genes as compared to the genes detected by the independent test data, and (iii) detected more meaningful genes than were validated in previous melanoma studies in the literature.

Declaration

I can confirm that the work of the thesis, including all computational study and laboratory-based experiments, was conducted by myself.

Wanting Liu

CONTENTS:

Figures:

Tables:

List of Abbreviations

ANOVA	Analysis of Variance
BLC	Basal-like cancer
CDK4	cyclin-dependent kinase 4
cDNA	complementary DNA
CIB-DDBJ	Centre for Information Biology and DNA Data Bank of Japan
CIBEX	Center for Information Biology Gene Expression database
DOPA	Dihydroxyphenlalanine
DS	Donkey Serum
EBI	European Bioinformatics Institute
EM	Epidermal melanocyte
FAK	Focal adhesion kinase
FoxM1	Forkhead Bax M1
GenoMEL	Melanoma Genetics Consortium
GEO	Gene Expression Omnibus
GPL	GEO Platform
GSE	GEO Series format file
GSM	GEO Soft format Sample file
GWGS	Genome-Wide Global Significance
GWRS	Genome-Wide Relative Significance
GXA	Gene Expression Atlas
HMEC	Human Mammary Epithelial Cells
HPC1	Hereditary Prostate Cancer gene 1

ICC	Immunocytochemistry
IDC	Invasive Ductal Carcinomas
IHC	Immunohistochemistry
ILC	Invasive Lobular Carcinomas
KNN	K-nearest neighbors algorithm
LOWESS	Locally Weighted Scatterplot Smoothening
MGED	Microarray Gene Expression Data
MGED	Microarray Gene Expression Data Society
MM	Metastatic Melanoma
NCBI	National Center for Biotechnology Information
Non-BLC	non-Basal Like Cancer Samples
PSA	Prostate-Specific Antigen
PTEN	Phosphatase and Tensin Homolog
RGP	Radial Growth Phase
SAM	Significance Analysis of Microarrays
SD	Standard Deviation
SMD	Stanford Microarray Database
spp1	Osteopontin
SS_B	Sum of Squared Residuals Between
SS_W	Sum of Squared Residuals Within
SVM	Support Vector Machine
TNM	Tumour, Node, Metastases
UVR	Ultraviolet Radiation
VGP	Vertical Growth Phase

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest appreciation to my two supervisors Dr Yonghong Peng and Prof. Desmond J Tobin, both of them have been tremendous mentors for me, without their supervision this dissertation and paper would not have been possible.

Dr Peng who has great expertise in computing science, continually and persuasively conveyed a spirit of adventure in regard to research. He has always encouraged my research and allowed me to grow as a researcher.

Prof. Tobin undertook to act as my supervisor by his professional commitments in biological science. His advice, brilliant comments and suggestions have been priceless. Without his contribution, I would not be able to validate genes in wet-lab and attend the IID conference.

A special thanks to my family. Words cannot express how grateful I am to my parents (Chuanwei Liu & Huimin Zhang) and my husband (Wanwei Wang) for their strong support during the PhD course.

I would also like to thank my friend Mohamad Hilal Al-Shammari, he helped me a lot during the writing-up period. I would also like to thank my friends Chongye Guo and Tianlin Zhao, we have encouraged and helped each other during the overseas study time.

1. INTRODUCTION

This chapter provides background and review on the subject areas of this thesis. This study was focused on the development of an integrated bioinformatics approach for the discovery of diagnostic biomarkers of melanoma, based on microarray genome-wide expression data. In this chapter I will introduce the topic of melanoma, including its definition and characteristics, how it develops, its related staging, genetic organization and other relevant information. To evaluate the effectiveness of bioinformatics method, developed in this project an enhancement of this method also to other microarray datasets including breast cancer and prostate cancer was applied (Chapter 2), and so provide a brief review of breast and prostate cancer here too. At the end of this chapter, I discuss to topics of DNA microarray and meta-analysis.

1.1. Skin and Melanoma

1.1.1. Human skin structure and function

Many people do not immediately think of skin as a functional organ like the heart or the liver, instead believing that skin simply protects them from different external stimuli, like excessive heat or cold (Poole & Guerry IV, 2005).

Actually, human skin is considered to be the largest organ in the body (Tobin, 2006) and plays a unique role in not only providing the main barrier between the

internal and external environment, but also maintains our internal homeostasis (Slominski & Wortsman, 2000) by exhibiting all the requirements of a classic endocrine organ. It is proposed to be an independent peripheral endocrine organ (Slominski & Mihm, 1996). The barrier effect of skin is not only evident by the physical function of its epidermis, but skin also has a chemical or biochemical role through its array of lipids, enzymes, defense-related cells and so on (Proksch *et al*, 2008). The structure of skin is shown in **Figure 1** below.

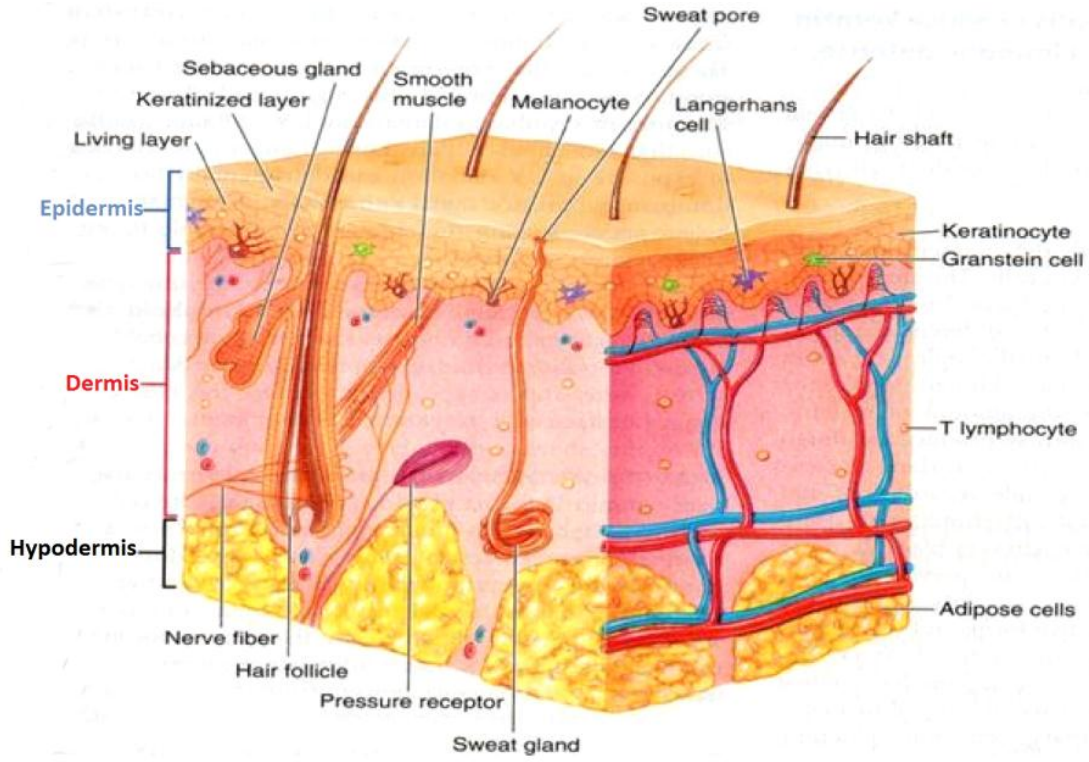


Fig.1: Anatomy of human skin (Sherwood, 2007)

As shown in above figure, human skin includes two main layers; the epidermis (outer layer) and dermis (inner layer). The epidermis can be further divided to two layers: the outer 'dead' but biochemically-active keratinized layer (stratum corneum) and the inner living layer 'nucleated epidermis' (Proksch *et al*, 2008;

Sherwood, 2007). The keratinized layer is made from dead or peeling cells, and is the main barrier to prevent ingress by microbes, as well as chemical and mechanical stress (Madison, 2003).

The living layer of the epidermis is further stratified and consists of four cell types; keratinocytes, melanocytes, Langerhans cells and Merkel cells (Sherwood, 2007). The function of nucleated epidermis is to contribute to forming the main barrier, which is present to prevent loss of water and invasion of noxious substances and materials (Honari, 2004). If trauma causes damage of the epidermis, the barrier function will be severely disturbed (Elias *et al*, 1977). And repair of the barrier needs increased DNA synthesis (for cell proliferation) and lipid synthesis (Proksch *et al*, 1993).

The dermis can also be separated into two parts: the upper papillary dermis and the lower reticular dermis (Poole & Guerry IV, 2005). The dermis contains hair follicles, exocrine glands like sweat glands and sebaceous glands (holocrine) and various cell types including adipocytes, monocytes, mast cells, and fibroblasts (Sherwood, 2007). The skin is both the source and target of various hormones, growth factors and binding proteins, steroid hormones and vitamin D etc. (Zouboulis, 2000).

Most skin cancers originate in the epidermis (Poole &Guerry IV, 2005). For melanoma, the originating cell is the melanocyte or its precursor cell the melanoblast, which is located mainly in the epidermis layer (see below) and the upper hair follicle.

1.1.2. Melanocytes

1.1.2.1. Melanocytes and Melanoblasts

A significant component of a functional skin barrier is its complement of pigment (i.e., melanin)-producing cells (Lin & Fisher, 2007). The term 'melanocyte' was first introduced by Meyerson in 1889 (Westerhof, 2006), and represents a dendritic cell which originates from the neural crest. They exist as a minor cell subpopulation in the basal layer of skin's epidermis, mucous membranes, striavascularis of ear, hair matrix, various locations in the eye, and central nervous system (Sulaimon & Kitchell, 2003). The distribution of melanocytes in skin epidermis is shown in **Figure 2**.

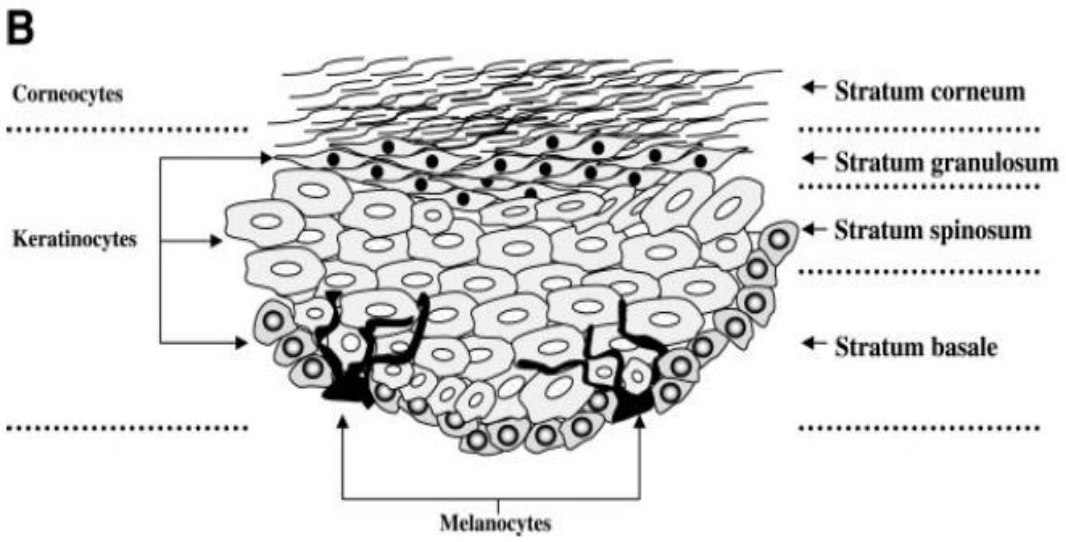


Fig. 2: Melanocytes in the epidermis of human skin (Robins, 1991)

Melanoblasts are precursors of the melanocytes and may contain pre-melanin granules (Westerhof, 2006). After melanoblasts are formed, they proliferate, differentiate, migrate, and finally arrive in the correct destinations including hair follicles and the base of the epidermis (Lin & Fisher, 2007).

Melanocytes exist in very close proximity to keratinocytes, so that they have ability to easily deliver their produced melanin to keratinocytes. The delivered melanin accomplishes a primary function of melanocytes, i.e., to provide skin pigmentation (Poole & Guerry IV, 2005; Sulaimon & Kitchell, 2003).

While pigment cells of the retina originate in the optic cup of the forebrain, all other melanocytes are neuroectodermal in origin (LaBonne & Bronner-Fraser, 1998). During human embryogenesis, melanocyte development begins when the melanoblasts migrate from the neural crest. The melanoblasts are induced to migrate by other neural crest cells and travel to various body sites, and finally form mature melanocytes (Sulaimon & Kitchell, 2003). Melanoblasts are generated at the second month of intra-uterine life in humans (Costin & Hearing, 2007). The earliest skin-homing melanoblasts are found in the dermis at the 10^{th} to 12^{th} week of gestation, and then about 2 weeks late, they leave the dermis to enter the epidermis. After approximately 50 days gestation, melanocytes can be detected in the epidermis (Westerhof, 2006; Costin & Hearing, 2007). When melanoblasts reach their terminal destinations, most will differentiate into melanocytes. The whole process is completed by month 6 of human intra-uterine life. In parallel there is a gradual decrease in the number of melanocytes in dermis, such that when the baby is born most melanoblasts /

melanocytes have disappeared from the dermis. The mature melanocyte locates in the basal layer of the epidermis and produce melanin (Costin & Hearing, 2007).

1.1.2.2. Melanosomes and melanin

Once the mature melanocyte is formed, it starts to produce its unique organelle, the melanosome. Melanosomes are generated from the endoplasmic reticulum of the melanocyte and are of lysosomal origin (Westerhof, 2006). Melanosome are the place where melanin is produced, and prepared for ultimate transfer into keratinocytes, where they play an important role in skin photo-protection (Meyskens *et al*, 2001).

A wide range of colours in skin and hair is possible, and is due to tonal variations in the pigment melanin, which is synthesized as a high density, insoluble biopolymeric pigment produced via a complex tyrosine redox reaction pathway (Costin & Hearing, 2007). However, the visual effect of melanin is seen on the surface of skin or in the hair fiber. It is separated into two broad types called eumelanin and pheomelanin according to colour, shape, and size of granules and item subsequent packaging and processing in skin (Sulaimon & Kitchell, 2003; Lin & Fisher, 2007). Brown/black eumelanin is a high density, insoluble and dark pigment, and is contained in the eumelanosome and is the major source of pigmentation of skin. Pheomelanin is yellow or red in colour, is soluble in alkali. The synthesis of this latter type of pigment is influenced by the intercellular concentration of cysteine (Potterf *et al*, 1998). Eumelanin is more photoprotective than pheomelanin, due to the latter's greater photo-instability

associated with its chemical structure, and also relate to eumelanin's preferential binding of cations, anions, chemicals and so on (Costin & Hearing, 2007).

1.1.2.3. Melanogenesis

Melanogenesis is a biochemical pathway that synthesizes melanin (either eumelanin or pheomelanin) in the melanosome. The balance of variable enzyme expression and other pigment genes can influence the ratio of these two melanin types. During melanogenesis, at least 3 kinds of enzymes are needed to synthesis different melanins. Tyrosinase is produced by ribosomes before being transported into maturing melanosomes (Westerhof, 2006) and is the rate-limiting enzyme in melanogenesis for eumelanin or pheomelanin (Lin & Fisher, 2007; Sulaimon & Kitchell, 2003). Melanocytes with lower levels of active tyrosinase produce (eu)melanin more slowly than do melanocytes with low tyrosinase activity (Lin & Fisher, 2007).

Eumelanin synthesis also involves tyrosinase-related proteins-1 and -2 (TRP1 and TRP2/DCT). These two enzymes have 40-45% homology with tyrosinase (Lin & Fisher, 2007; Sulaimon & Kitchell, 2003) and can contribute to tonal variations in brown vs black melanin production. Pheomelanin appears to be more photo- and chemically- unstable, for example to sources of radiation exposure and oxidative stress. The formation of pheomelanin requires a cysteine and/or glutathione supply (Lin & Fisher, 2007).

Human evolution only success has depended on the ability of melanocytes to transfer their melanin product efficiently to neighboring keratinocytes and so protect the skin from harmful ultraviolet radiation (UVR). In humans, one melanocyte can make intimate contact with 37 viable keratinocytes to form, so called ‘epidermal melanin unit’ (Fitzpatrick & Breathnach, 1963; Eisenger & Marko, 1982). The correlate in the hair follicle has been termed the ‘follicular melanin unit’ (Tobin, 2008). The synergistic relationship of the melanocyte and keratinocyte helps protect skin from outside-in damage, e.g. UVR (Sulaimon & Kitchell, 2003). Melanin pigment transfer ensures melanin is distributed throughout the epidermis including its upper layers (Agar & Young, 2005), though the melanin granule itself is largely degraded in the keratinocyte during this upward movement differentiation of cells. While individual melanin granule can be seen in the stratum corneum of African skin, melanin is largely degraded in the upper layers of Caucasian skin (Sulaimon & Kitchell, 2003). **Table 1** outlines the melanocyte life history.

Table 1: Steps in melanogenesis (Sulaimon & Kitchell, 2003).

Step I	Melanoblasts migrate from the neural crest
Step II	Melanoblast differentiates to melanocyte. Clonal population of skin by melanocytes
Step III	Melanosome organelle biogenesis and matrix formation
Step IV	Melanogenic genes for tyrosinase, tyrosinase-related proteins and melanosomal matrix components are induced
Step IV	Tyrosinase and related melanogenic proteins are synthesized
Step V	Post-translational processing and glycosylation of tyrosinase
Step VI	Fusion of vesicles to form melanosomes and initiation of melanogenesis
Step VII	Control of tyrosinase activity
Step VIII	Control of the activity of tyrosinase-related protein
Step IX	Post-tyrosinase modification of biosynthesis
Step X	Modification of melanin
Step XI	Melanosome transfer to keratinocytes
Step XII	Melanosome degradation
Step XIII	Melanin removal with loss of cornified cell (i.e., stratum corneum)

The process of melanin synthesis can itself result in the production of intermediates of hydrogen peroxide and quinine (Meyskens *et al*, 2001). Thus, deficient handling of melanin intermediates could cause epidermal melanocytes damage (Sulaimon & Kitchell, 2003).

1.1.2.4. Melanocyte response to UVR and melanogenesis

In 1917, Bloch reported on role of 3, 4-dihydroxyphenylalanine (DOPA) in melanocytes, and indicated that DOPA was a special substrate for the DOPA-oxidase activity of tyrosinase. Then, his co-worker Lutz found that the activity of DOPA-oxidase was increased in pigment cells by UVR, and concluded that increased melanin production would offer protection against UVR (Westerhof, 2006).

While pigmentation gene expression can decide the constitutive pigmentation of skin and hair; this level of pigment production can be increased further (i.e., facultative) by a range of stimuli including UVR exposure (i.e. tanning), hormones, and growth factors etc. (Sulaimon & Kitchell, 2003). However, exposure to UVR can cause both inflammation and pigmentation (Heenen *et al*, 2001; Gledhill *et al*, 2010). Stimulated or facultative pigmentation exhibits both 'immediate' and 'delayed' subphases (Costin & Hearing, 2007; Sulaimon & Kitchell, 2003).

Immediate 'tanning' pigmentation develops within seconds and minutes of UVA exposure, inducing pre-existing melanosomes to darken and their movement from the perinuclear area to the dendritic area in melanocytes. However, the number of the melanosomes is not increased. Immediate tanning peaks 1-2 hours after UVR (Costin & Hearing, 2007; Sulaimon & Kitchell, 2003).

By contrast, delayed pigmentation begins 2 to 3 days after UVB exposure (and to a minor extent by UVA and visible light) (Costin & Hearing, 2007). Maximum

tanning is reached approximately three weeks later. Remarkably, delayed pigmentation can take up to ten months to return back to original pigmentation level. Delayed pigmentation depends on the changing of melanocytes in quality (i.e. more eumelanin, more melanosomes) and to a limisted extent in quantity (via some very limited cell proliferation) (Costin & Hearing, 2007; Sulaimon & Kitchell, 2003).

Both immediate and delayed tanning can produce pigment that can protect the skin against further damage and also against skin cancer, as UVR exposure can induce DNA damage (Fitzpatrick, 1988; Ortonne & Prota, 1993). UVR exposure of normal skin can increase melanin levels 10 to 15 fold even in the darkest individuals. However, melanin levels can be increased to 500 to 1000 fold in paler skin, including those as risk of skin cancer induction (Kaidbey *et al*, 1979).

1.1.3. Melanoma

In 1996, the American Academy of Dermatology reported that one melanoma patient dies each hour in the United States (Poole & Guerry IV, 2005). In the USA, human malignant melanoma has the second highest mortality rate of all cancers, second only to lung cancer. Rates of malignant melanoma are also increasing rapidly in other countries, like the UK, Germany, Canada, and Australia (Sulaimon & Kitchell, 2003). Indeed, since the middle of 1960s, the incidence of melanoma has increased by 3% to 8% per year in many countries (Lens, 2008). However, thankfully the survival rate has also increased from 11% to 39%. This increased survival ratio depends however on early detection and

atypical nevi removal (Gremel *et al*, 2009).

Melanoma appears to be produced in two ways. In one melanocytes are activated and subsequently transformed by UVR after unprotected sunlight exposure; the other way is via transformation of benign nevi when benign nevi including after unprotected sun exposure (Lejeune, 1997).

1.1.3.1. Melanoma from a historical perspective

Melanoma was mentioned as a 'black pigmented human tumour' at the time of Hippocrates (460-370 B.C.E.) (Balch *et al*, 2003). John Hunter, an English physician, published the first report of melanoma in Western medical literature in 1787, describing it as "soft and black" and as a "cancerous fungous excrescence". He removed a melanoma tumour from the lower jaw of a 35y old man and preserved it. In 1968, the specimen was confirmed as a melanoma (Balch *et al*, 2003).

1.1.3.2. Definition of melanoma

Melanoma is a cancer that arises from the genetic transformation and uncontrolled growth of the melanocyte or melanoblast (i.e., melanocyte stem cell) (Miller & Mihm, 2006). Melanoma may form in skin, mucosa, uvea of eye, and leptomeningitis (Eigentler & Garbe, 2006). The most common type is cutaneous melanoma, and once melanoma cells have reached the dermis they have the potential to enter blood and lymph vessels and then to spread to other areas of the body (Poole & Guerry IV, 2005). Even though melanoma only

accounts for less than 5% of all skin diseases, it induces nearly 80% of skin disease mortality. Only 14% of patients with metastatic melanoma can expect to gain 5 years survival. The cure rate depends on early detection and removal by surgery. Some reports showed more than a 90% cure rate if the tumour is removed when it is less than 1mm in depth (Bolognia *et al*, 2003).

Cutaneous melanoma can be classified into four types: superficial spreading melanoma, nodular melanoma, lentigo maligna melanoma, and acral lentiginous melanoma (Porras & Cockerell, 1997).

The most common melanoma subtype is the superficial spreading type, which occurs in 70-80% of the all melanoma (Poole & Guerry IV, 2005). It can appear anywhere on the body and is commonly found on the upper arms, thigh and back of body, even though this skin is somewhat paradoxically not routinely exposed to the sun. The melanoma may begin from an abnormal nevus with irregular contour and various colours (Brannon, 2004). An existing nevus usually is the starting point of the development of a superficial spreading melanoma. However, melanoma can also develop from apparently unblemished skin. The diameter of nevus is commonly more than half a centimeter, it is more likely to be at risk of becoming a melanoma (Poole & Guerry IV, 2005).

About 15% to 20% of the all melanoma occur as the nodular form. This type can appear on any part of the body, and locates at the same body sites as seen with superficial spreading melanoma (Brannon, 2004). Like superficial spreading melanoma, nodular melanoma often arises from existing nevi. It is slightly more

common in men than in women, is commonly dome-shaped and sometimes it looks like a blood blister (Poole & Guerry IV, 2005).

The least frequent type of melanoma is lentigo maligna melanoma, which accounts for only 5% of all melanoma. It develops from a small abnormal patch on sun exposed skin in older people, usually on the regions of head and neck (Brannon, 2004). This melanoma results from many years of intense unprotected sun exposure, and does not arise from pre-exisiting nevi. The melanoma first appears on the skin as a dark irregular stain (Poole & Guerry IV, 2005).

The above three types of melanoma appear most frequently in Caucasians. However, another type of melanoma called acral lentiginous melanoma has lower frequency for white people and is the most common subtype in people with darker skins, e.g., xanthoderm (Poole & Guerry IV, 2005). The acral lentiginous melanoma is less than 5% of all melanomas, and is often on the palmar-plantar surface of hands, feet and the fingers or toes and under the nails (Brannon, 2004). When under the nails, the melanoma will appear on the base of the nails as a streak then reaches out to the tip. It most often appears on thumbs and great toes. The frequency of acral lentiginous melanoma is also reported to be associated with extent of sun exposure (Poole & Guerry IV, 2005).

Mucosal melanoma has been found on the mucous membranes inside the mouth and on the anal-genital region. The carcinogen involved is unknown, as

this type of melanoma is unlikely to be related to sun exposure (Poole & Guerry IV, 2005). Rarely, melanoma can present without melanin pigment, are pink or red, and so these melanomas can be very difficult to detect. (Poole & Guerry IV, 2005).

1.1.3.3. Melanoma risk factors

Sun-sensitive skin has greater risk of all skin cancers than darker skin types because the photosensitive skin usually produces lower levels of the protective pigment eumelanin (Lin & Fisher, 2007; Mille & Mihm, 2006). A suppressed skin immune response and excess UVR exposure are enhancer risk factors of melanoma (Mile & Mihm, 2006). It is thought that a single risk factor is not on its own sufficient to develop melanoma, but that multiple risk factors together in the same individuals can markedly increase the risk for melanoma (Poole & Guerry IV, 2005), e.g. certain MCIR variants (Healy *et al*, 2014).

Ultraviolet Radiation

Three kinds of electromagnetic radiation from sunlight can influence us physiologically including visible, infrared and ultraviolet light/irradiation (UVR). The greatest potential comes from the invisible UVR. UVR can be divided into UV-A (320-400 nm), UV-B (290-320 nm) and UV-C (200-290 nm). Only UV-A and UV-B reach the surface of the earth. They are carcinogens and play a role as both promoters and enhancers of melanoma (Setlow *et al*, 1993). UV-B is described as the main carcinogenic factor at the formation stage of melanoma and other skin cancers. UV-A and UV-B both induce skin damage, e.g., sunburn

and skin aging. Melanin pigment plays an important protective role in absorbing UVR and via a detoxifying function (Mile & Mihm, 2006). Actually, the sunburn sensitivity of different skin phototypes is related to the packaging and distribution of the melanin in the skin cells. Melanin type is under genetic control, and is the major factor to determine type and colour of the skin (Poole & Guerry IV, 2005).

White Caucasian skin is the most susceptible skin type for melanoma, except for acral lentiginous melanoma. Even a brief strong sun exposure episode on sensitive skin can cause it to become inflamed and red. Commonly, this kind of individual has red(-ish) hair, freckles and porcelain white skin (e.g., Celtic people) (Poole & Guerry IV, 2005), a phenotype associated with a small number of MCIR variants (Healy *et al*, 2014).

Is all sun exposure bad?

For most people, mild sun exposure is not deleterious. In fact, mild sun exposure can be beneficial e.g., for the production of vitamin D. It has also been reported some tanning (without burning) may actually aid melanoma prevention (Poole & Guerry IV, 2005; Zouboulis, 2000). While excessive exposure to UVR has been indicated as the main etiological factor in melanoma, except is rare hereditary cases (Sulaimon & Kitchell, 2003; Halaban, 1996), the precise relationship is complex. For example, primary melanoma commonly appears first on non-exposed body sites (Tronov *et al*, 2010). Still, many *in vitro* studies have demonstrated that UV radiation induces DNA damage, gene mutations, induction of reactive species oxygen, inhibition of the skin immune system and

increasing growth factor production (Halaban, 1996). All these changes can lower the resistance of skin to protect itself from the ill-effects of UVR (Tronov *et al*, 2010).

Nevi (moles)

The average number of nevi or moles on their body surface of a white Caucasian adult is about 25. Most children appear to be born without nevi. However, these children develop nevi by approximately 3 years old (Gallagher & Mclean, 1995). It has shown that sun exposure induces moles. Commonly, when a mole is produced on a child's skin, it is a flat dark brown and pinhead size dot. Then, the mole will slowly grow to a round or oval shape, flat or domed. Some moles are not influenced by sun exposure. For example, 1% of newborn babies have a dark brown, flat or shortly domed single moles, and some of these moles may be more that 1 centimetre diameter. Some doctors may advise to remove these moles to prevent melanoma. However, the chance of evolution to melanoma is rare (Pools & Guerry IV, 2005). Almost all people have some normal round and small moles. However, in 15% of white skinned persons these can show dysplastic features. Dysplastic moles are larger than normal mole (>5mm in diameter), and have domed centres. They appear more or less on sun-exposed parts of the body (Pools & Guerry IV, 2005).

Both dysplastic moles and normal moles deserve attention, because people with large numbers of moles are at relatively higher risk of melanoma than those without. The presence of numerous on normal moles and a few dysplastic moles can also help to warn individuals that they may have 2-10 times greater

risk to develop melanoma during their lives than people who have few or no moles (Pools & Guerry IV, 2005). Still, the prevention of melanoma by removing all dysplastic moles and ordinary moles may be viewed as an extreme precaution. The key is to know the moles which are dysplastic and to watch out for changes to them (Pools & Guerry IV, 2005). For example, if a person has an isolated dysplastic mole and has already had melanoma, the mole should be removed. Similarly, if someone's mole is changing in phenotype, the mole should be removed.

Melanoma genetics— Inherited and mutated genes

Only approximate 10% of melanoma patients have familial disease with family members expressing relevant and associated gene mutations. From 1976, the National Cancer Institute and the University of Pennsylvania's Pigmented Lesion Clinic have detected gene mutation frequency in 23 suspicious families. Strikingly, their results show that 90% of close blood relatives did not express these genes. However, if two or more melanoma patients exist in the same family, doctors would hypothesize that these patients will have inherited similar melanoma-associated genes (Pools & Guerry IV, 2005). Though the number of these patients still constitutes only 10% of the total, a family history of melanoma is a strong risk factor of melanoma. The members of a family with atypical nevi or previous melanoma should also be assessed for possible positive melanoma family history, even though these changes have not yet developed to malignant melanoma (Miller & Mihm, 2006).

The cell growth regulator p16 or *CDKN2A* have been identified as a relevant

mutant gene on chromosome 9. Multiple studies have determined that mutation of *CDKN2A* plays a key role for the development of melanoma (Miler & Mihm, 2006). Experimental findings have shown that the loss of *CDKN2A* expression and function increases the chances of transformation of dysplastic nevi to melanoma, or increases the probability to develop melanoma (Miler & Mihm, 2006). *CDKN2A* mutations were detected in the GenoMEL (Melanoma Genetics Consortium) and found that relatively low mutation detection rate of *CDKN2A* is not based on a failure to detect rather it implies other high penetrance melanoma genes exist (Harland *et al.*, 2008). *BRAF* is a second dominant gene associated with melanoma that encodes for a protein called B-Raf that can activate MEK. *BRAF* mutations are the most common mutations found in melanoma to date. *BRAF* mutations induce the melanocyte population to become senescent by inhibiting proliferation; it also cooperates with *p16* mutations to induce high penetrance and short latency of melanoma (Delmas *et al.*, 2007).

1.1.4. From melanocyte to melanoma

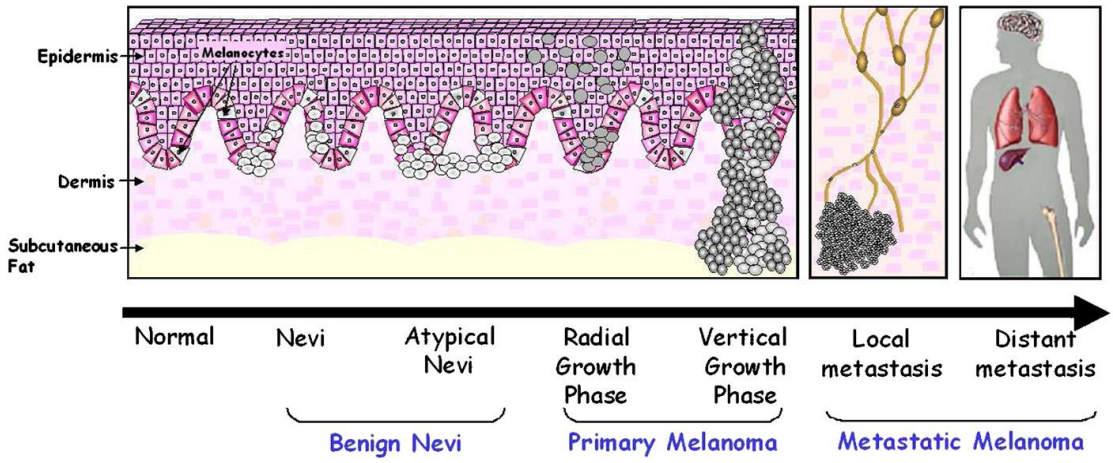


Fig.3: The evolution of melanoma (Gremel *et al*, 2009).

As shown in **Figure 3**, the melanocytes are located in the basal layer of the epidermis, where malignant transformation of an affected melanocyte starts to develop melanoma (Bittener *et al.*, 2000). However, the melanoma is not only transformed from epidermis-associated melanocytes, it also can be from cells in the dermis. Studies have proposed that melanoma may also originate via the transformation of melanocyte stem cells in the epidermis (Grichnik *et al.*, 2006) and dermis-derived stem cells (Zabierowski *et al.*, 2011).

The commonest methods used in clinical practice to classify melanoma for diagnosis rely on the histologic thickness of melanoma, the degree of invasion, and melanoma with or without ulcerations (Balch *et al.*, 2001). These are discussed in the following section.

1.1.4.1. The transition of melanocytes to melanoma: five phases

The transition of melanocytes to melanoma could be separated into five steps (Shown as the above **Figure 3**).

1. Melanocytes in normal skin or benign nevi (Gremel *et al*, 2009).

2. Dysplastic and atypical nevi: These dysplastic and atypical nevi are formed from normal nevi, which may mean melanoma changes have begun (Gremel *et al*, 2009).

3. Radial growth phase: Atypical cells emerge to form the radial growth phase (RGP), which is the initial stage of melanoma. The newly-formed melanoma does not spread to other organs of the body, until it moves to successive growth phases (Pool & Gerry IV, 2005). The radial growth phase melanoma invades the epidermis and metastasizes to the upper dermis. However, it may not be easy to detect using a diagnostic test at this stage (Gremel *et al*, 2009). The radial growth phase consists of two steps: The first is where the melanoma cells are still *in situ* i.e., still in the epidermis. The second is where the melanoma cells have ability to invade to the dermis. However, the number of melanoma cells still remains low. From the second step, the melanoma cells break free of their epidermis *in situ* containment. During the radial growth phase, the melanoma cells cannot yet spread to other parts of the body. Moreover, if the melanoma cells are not detected and removed, they will proceed to the vertical growth phase (Poole & Gerry IV, 2005).

4. Vertical growth phases: When the growing RGP melanoma invades into the dermis and subcutaneous fat layer, the stage is called vertical growth phase (VGP) (Gremel *et al*, 2009). It is also called the tumourigenic phase. The melanoma exists as a group of spreading abnormal melanocytes that move to the dermis where they start to form a tumour mass. In this phase, the melanoma has a chance to expand to other parts of body. Removing the melanoma at the radical growth stage i.e., before the beginning of the vertical growth phase, can stop the tumour from becoming potentially fatal (Poole & Gerry IV, 2005).

5. Metastatic melanoma: After the vertical growth phases of melanoma, the tumour cells metastasize to the blood vessels, lymph nodes, lung, liver, brain and others parts of the body. This stage is metastatic melanoma (Gremel *et al*, 2009).

1.1.4.2. Other schemes for the classification of melanoma

Other methods have been used to describe the stages of melanoma, like the 'Clark' and 'Breslow' staging. The Clark model describes the development process of melanoma from dysplastic nevus to metastatic melanoma (Clark *et al*, 1984). The details of the different measure methods are shown on the **Table 2**.

Table 2: Staging of malignant melanoma http://chorus.rad.mcw.edu/doc/00955.html, (Kahn, 2006)

Malignant melanoma: staging
Clark staging (Clark *et al*, 1984) • Level I: all tumour cells are in the epidermis above basement membrane (i.e., in situ) • Level II: tumour extends to the upper papillary dermis • Level III: tumour extends to interface between the papillary and reticular dermis • Level IV: tumour extends between bundles of collagen in the reticular dermis • Level V: tumour invasion of subcutaneous tissue **Breslow staging (Breslow, 1970)** • Thin: < 0.75 mm depth of invasion • Intermediate: 0.76-3.99 mm depth of invasion • Thick: >4 mm depth of invasion

Clark level I is melanoma *in situ* as classified by WHO (Leboit *et al*, 2006). At is first level, abnormal cells of dysplastic nevi appear with a reduced ability of DNA repair. However, the cells cannot produce colonies in agar (Tronov *et al*, 2010). Clark level II is microinvasive melanoma within the upper papillary dermis, which corresponds to the radial growth melanoma. Clark level III to level V and the Breslow staging are equivalent to histological features from the vertical

growth phase of melanoma to metastasis melanoma (Leboit *et al*, 2006; Tronov *et al*, 2010). At these stages, the malignant melanocytes have the ability to grow in a relatively unrestrictive way, and can form colonies on agar (Tronov *et al*, 2010). The histopathological appearance of the 5 steps of the Clark model is presented in **Figure 4**.

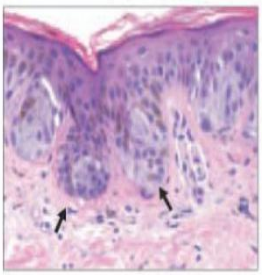
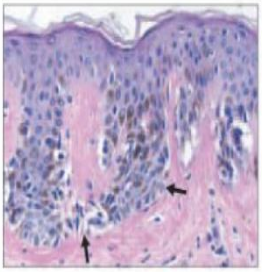
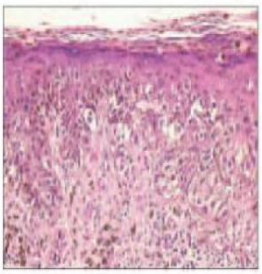
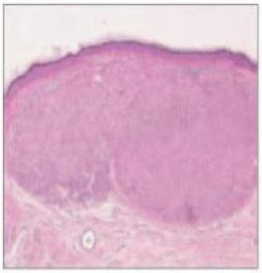
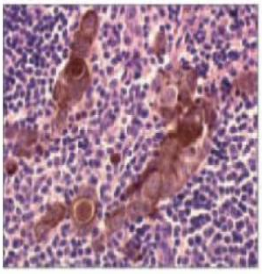
Histopathological Appearance	Description	Histologic Features
 Benign nevus	**Step 1** The first event is a proliferation of structurally normal melanocytes leading to the benign nevus. Clinically, these nevi present as flat or slightly raised lesions with either uniform coloration or a regular pattern of dot-like pigment in a tan or dark brown background. Histologically, such lesions have an increased number of nested melanocytes along the basal layer (arrows).	Proliferation of melanocytes Benign lesions
 Dysplastic nevi	**Step 2** The next step is the development of aberrant growth. This may occur within a preexisting benign nevus or in a new location. Clinically such lesions may be asymmetric, have irregular borders, contain multiple colors, or have increasing diameters. Histologically, such lesions have random and discontiguous cytologic atypia (arrows).	Dysplastic cells Random atypia
 Radial-growth phase	**Step 3** During the radial-growth phase, cells acquire the ability to proliferate intraepidermally. Clinically, they sometimes present as raised lesions. These lesions no longer display random atypia and instead show cytomorphologic cancer throughout the lesion. In addition to the intraepidermal cancer, the cells can penetrate the papillary dermis singly or in small nests but fail to form colonies in soft agar.	Intraepidermal growth Continuous atypia
 Vertical-growth phase	**Step 4** Lesions that progress to the vertical-growth phase acquire the ability to invade the dermis and form an expansile nodule, widening the papillary dermis. The cells can also extend into the reticular dermis and fat, are capable of growth in soft agar, and have the capacity to form tumor nodules when implanted in nude mice.	Dermal invasion
 Metastatic melanoma	**Step 5** The final step in the model is the successful spread of cells to other areas of the skin and other organs, where they can successfully proliferate and establish a metastatic focus. These cells can grow in soft agar and can form tumor nodules that may metastasize when implanted in nude mice.	Metastasis

Figure 4: The 5 steps of melanoma development (Clark model) (taken from Miller & Mihm, 2006)

1.1.4.3. Melanoma tumour invasion and metastasis

The extent of local invasion and metastasis of tumour cells directly influence the morbidity and mortality of melanoma, which mainly occur in the melanoma's vertical growth phase (Miller & Mihm, 2006). The tumour is said to be metastatic when it spreads and invades to blood vessels, lymph, and some surrounding stroma (Haass *et al*, 2005). Some studies have demonstrated that melanoma invasion is related to changes in cell adhesion (Miller & Mihm, 2006). Normally, cell adhesion controls cell migration, and also organizes tissue and organogenesis. Once cell adhesion is disturbed, the tumour cell can respond to stimulate different tumour-associated signaling pathways to establish new relationships between tumour and stroma that assists the tumour cells to invade (Miller & Mihm, 2006).

1.1.5. Pathogenesis

The role of genetics in melanoma development has been demonstrated by several epidemiology studies, like gene changes that are associated with a disturbed cell cycle mechanism and via other transcriptional mechanisms (Halachmi & Gilchrest, 2001). A major gene that is considered to be implicated in melanoma is *CDKN2A*, which locates on chromosome 9p21, and it also is known as *INK4a*. The exons 1α, 2, and 3 of *CDKN2A* are used for encoding the protein p16. The protein $p14^{ARF}$ is transcribed from exon 1β and the shared exons 2 and 3. Both gene products play an important role as negative regulators in cell cycle progression, and Daniel et al determined that the $p14^{ARF}$

is more commonly inactivated than p16 by genetic and epigenetic analyses (Freedberg *et al*, 2008). A view of the *CDKN2A* locus is shown as **Fig. 5**.

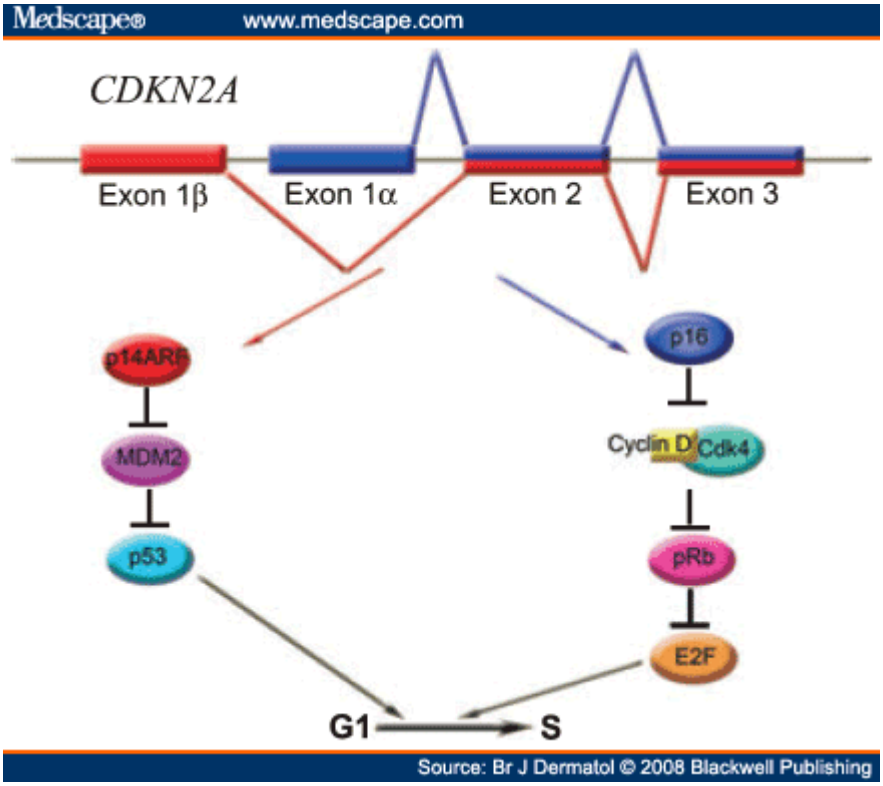


Figure 5: *CDKN2A* locus (Lin *et al*, 2008)

As shown in **Figure 5** above, p16 protein competitively inhibits of CDK4 (cyclin-dependent kinase 4). CDK4 and cyclin D can activate Rb (retinoblastoma protein) by phosphorylation, and the phosphorylated Rb can arrest cell cycle S phase, cell division and proliferation. A *CDKN2A* mutation leads to loss of the p16 protein function, and the functional loss of the p16 influences the mutagenic DNA repair before cell division. The incidence of an abnormal p16 protein in melanoma is between 30-50% in familial melanoma and between 25-40% in sporadic melanoma (Nestle & Kerl, 2003).

The *p14ARF* gene interacts with *MDM2* to regulate melanocyte growth via the p53 pathway (Piepkorn, 2000). The activated p53 pathway inhibits the cell cycle and regulates the apoptosis. So loss of p14ARF function, due to the mutation of CDKN2A, can increase cell growth or survival (Nestle & Kerl, 2003).

Many studies indicate advocate that melanoma development involves several additional genes located in chromosome regions 1p, 6q, 7p, 11q, 9p, 10q. These high mutation regions can be analyzed by new mutation analyses techniques, like cDNA and tissue microarrays (Pollock & Trent, 2000).

1.2. Breast cancer

In the current study, microarray datasets for other cancers were used for the evaluation of the proposed bioinformatics approach and included Breast Cancer and Prostate Cancer.

Breast cancer is generated in breast tissue, and commonly occurs in the associated ducts and lobules, called ductal carcinomas and lobular carcinomas respectively. The incidence of breast cancer is 100 times higher in females than males. (Sariego, 2010), Breast cancer occupies 22.9% of all female cancer (excludes non-melanoma skin cancers), and around 13.7% cancer deaths in women in 2008 (Buchholz, 2009).

Risk factors of breast cancer include smoking, high levels of estrogen hormone, diet and obesity. Negative risk factors include young age and breastfeeding. However, whether breastfeeding has a relationship with breast cancer is still an open question, some studies found the positive associations between them, but others did not (Yang & Jacobsen, 2008). Smoking increases the risk of breast cancer especially for heavy smokers who started smoking at a young age, or for those with a long term smoking habit (Johnson *et al.*, 2011). Higher hormone levels have been associated with breast cancer e.g. estrogen-related drugs increased the risk of breast cancer (Johansen *et al.*, 2010) as do high fat diets and obesity (Blackburn & Wang, 2007).

Mutations in *BRCA1* and *BRCA2* are significant biomarkers for breast cancer. Mutations in these two gene mutations have been found in up to 90% of breast cancers with a demonstrated genetic influence. Other significant gene associations with breast cancer include *p53, PTEN, STK11, CHEK2, ATM, BRIP1* and *PAL2* (Gage *et al*, 2012).

Staging of breast cancer is based on the Tumour, Node, Metastases (TNM) system, i.e. the size of the tumour, whether the tumour spreads to the lymph nodes in the armpits, and whether the tumour starts to metastasize. The stage has been classified from 0 to 4, viz. stage 0 is carcinoma *in situ* (including ductal or lobular carcinoma); stage 1 – 3 represent a tumour that is still within the breast or within the regional lymph nodes; and stage 4 exhibits tumour metastases outside the breast (Johansen *et al.*, 2010).

Survival rates for breast cancer depend on several factors: type, stage, treatment and location of the patient. The major treatment methods are surgery (the physical removal of the tumour, surrounding tissue and sentinel lymph node), medications including hormonal therapy (drugs have been used for blocking the estrogen receptors or production of estrogen), chemotherapy (causing DNA damage to proliferating cancer cells), radiotherapy (usually given after surgery to reduce the risk of recurrence) and immunotherapy (Florescu *et al.*, 2010).

1.3. Prostate cancer

This tumour originates in the prostate, a gland of the male reproductive system, and grows as a classical adenocarcinoma and glandular cancer. Although it develops relatively slowly, some prostate cancers can come aggressive. Metastatic prostate cancer spreads to other parts of body, especially to the lymph nodes and bones (Lister, 2009). This cancer most commonly affect men over fifty years old (Siegel, 2011), and ranks sixth in cancer deaths in men (Baade *et al.*, 2009).

As with breast cancer, the stages of prostate cancer are also measured by the four stages TNM system, viz. based on tumour size, nodal involvement and metastases. Stage I and II of prostate cancer means the tumour is limited to the prostate. When the cancer cells spread to lymph nodes and other organs the cancer is graded stage III and IV (Makarov *et al.*, 2012). Mortality of prostate cancer has not decreased over the last 10 years, due largely to unawareness of disease, even though prostate-specific antigen (PSA) testing has increased cancer detection (Djulbegovic *et al.*, 2010). Like breast cancer, treatment of prostate cancer includes surgery, radiation therapy, hormonal therapy and chemotherapy (Dimitropoulou *et al.*, 2009).

Obesity, age and family history are the main risk factors of prostate cancer. Whether diet is associated with prostate cancer is still unclear; with some studies reporting that meat intake has little relationship with higher risk of

prostate cancer (Alexander *et al.*, 2010), and that fruit or vegetable also have little positive effect (Key, 2011). An elevated testosterone level in blood may increase prostate cancer risk (Gann *et al.*, 1996). Age is a well-reported risk factor, and the cancer is very unusual in men who are younger than forty five. The average age of diagnosis is seventy (Hankey *et al.*, 1999). Family history also is a key, as first-degree relatives of prostate cancer patient have twice the background risk. If two or more prostate cancer patients exist in one family, other first-degree relatives have a five-fold risk than men with family history (Steinberg *et al*, 1990). Some other factors influence prostate cancer risk. Statins (cholesterol lowering drug) decrease the risk of prostate cancer (Shannon *et al.*, 2005), while infection or inflammation in prostate may increase risk (Dennis *et al.*, 2002).

At the gene mutation level, *BRCA1* and *BRCA2* are biomarkers (like in breast cancer) for prostate cancer (Struewing *et al.*, 1997). Hereditary prostate cancer gene 1 (*HPC1*), the androgen receptor and vitamin D receptor have also been reported as genes linked to prostate cancer (Gallagher & Fleshner, 1998). Mutations in P53, PTEN and KAI1 based on loss of the corresponding suppressor genes also play a role in prostate cancer (Beuzeboc *et al.*, 2009).

1.4. DNA MICROARRAY AND ANALYSIS

1.4.1. Introduction to DNA microarray

While the *Human Genome Project* has reported the location of all known human genes (Wren, 2009), the function of each gene has not yet been determined. While the functions of some of genes are frequently reported in the literature, still 37% of human genes have no published functional information (Wren, 2009). Conversely, protein domain analysis could help us to speculate on the corresponding functions of some genes. For example, a DNA binding function can be inferred when the gene's encoded protein contains zinc finger domains, and a protein-protein interaction can be inferred when the protein contains coiled-coil domains (Cahan *et al.*, 2007). Similarly, the cellular location of an unknown, cytosolic or membranous protein can be inferred if the protein contains trans-membrane domains. This 'guesstimating' should however be done in a biological context, because most gene expression is highly regulated by context and circumstance (Wren, 2009).

To deal with these challenges, the microarray technique has gained popularity in many biomedical areas, such as cancer, inflammation, cardiovascular disease, alcohol consumption, and stem cell differentiation and so on (Cahan *et al*, 2007).

DNA microarray is a high throughout technique, which has application in detecting and quantifying mRNA (gene) expression (Gremel *et al.*, 2009).

Although the microarray technique was first devised in 1975 (Mlakar & Glavac, 2007), the mature Microarray was built only in the early 1990's, and has rapidly developed since then. The technique was first used for measurement of DNA fragments, so it is named as DNA microarray (Kunz, 2008). The technique involves small pieces of discrete single strand DNA fragments with inherent properties as probes, which are used for binding their complementary and unique gene sequences (Gremel *et al.*, 2009). So the quantified amount of the detected gene sequences should be measureable with high accuracy due to complementary binding (Kunz, 2008).

Microarray technology has been used as an advanced high-throughput strategy for the discovery of diagnostic gene signatures of human diseases on a genome-wide scale. The genome-wide discovery of signatures enables one to gain important insights into the underlying biological mechanisms driving tumourigenesis. A significant amount of microarray data has been deposited in publically-available data repositories over the past decade, e.g., the Gene Expression Omnibus (GEO) (Barrett *et al*., 2011), the ArrayExpress Archive (Parkinson *et al*., 2011), CIBEX (Kodama *et al.*, 2010), and SMD (Hubble *et al.*, 2009). These repositories enable scientists to advance the discovery of diagnostic and prognostic gene signatures by means of data integration and integrated bioinformatics analysis. For example, one group constructed a global map of human gene expression by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines (Lukk *et al.*, 2010).

Although the microarray experiments are performed on different platforms, all platforms include a step of hybridization. Hybridization is used for binding the DNA probe to the solid support and then binding the probe to the fluorescent dyed target nucleotides of the samples (Mlakar & Glavac, 2007). The basic steps of microarray are shown on **Fig. 6**.

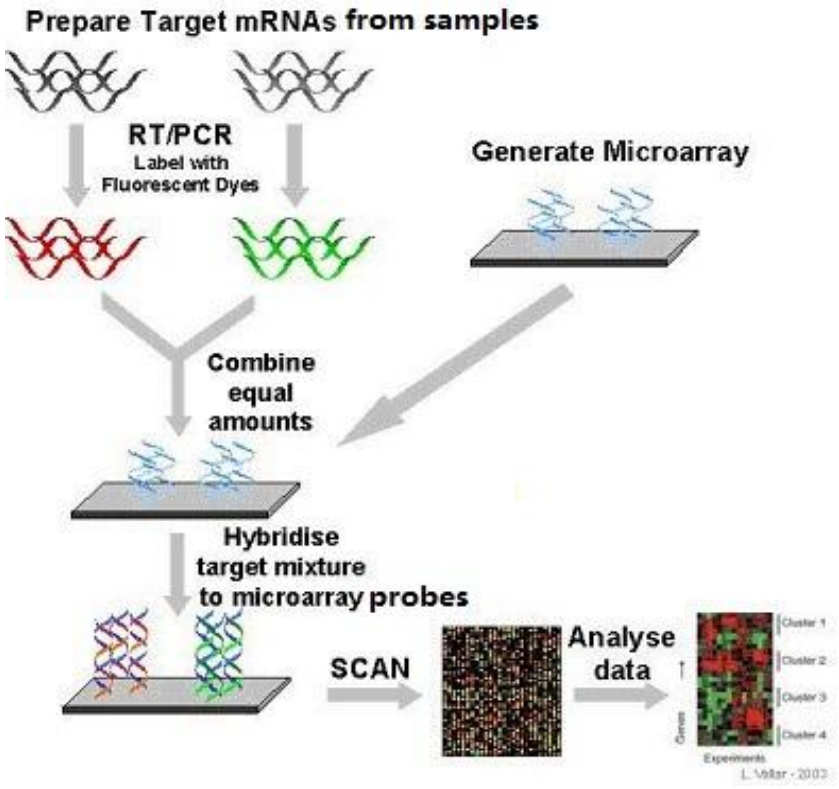


Figure 6: Basic steps of microarray analysis (Mlakar & Glavac, 2007)

1.4.2. DNA microarray experiments

There exist currently two main techniques for DNA microarray: Oligonucleotide DNA microarray and cDNA microarray (Gremel *et al*, 2009). An oligonucleotide DNA microarray can detect 25 to 70 base length short DNA or RNA sequences. cDNA microarray is used to detect 200 to 2000 base length sequences. Normally, standard PCR is used for the amplification step (Mlakar & Glavac, 2007).

Total RNA is extracted from tissue samples or cell lines. Usually, the target mRNA sequences are amplified by PCR. The total RNA are reverse transcribed into cDNA and stained with a fluorescent dye, and then hybridized to probes of microarray (Kunz *et al.*, 2004; Yang *et al.*, 2002).

DNA microarrays have been used to detect the gene expression of different disease phases, e.g., from normal skin or normal cells to metastatic tumour (Smith *et al.*, 2005). Researchers have noticed that the analysis of disease phase diversity can offer insights into how these genes are associated with enhanced melanoma cell survival (Smith *et al.,* 2005; Jaeger *et al*., 2007).

1.4.3. Microarray databases

There are four of main repositories for microarray data including; GEO (Gene Expression Omnibus), ArrayExpress, and CIBEX (Center for Information

Biology gene EXpression database) and SMD (Stanford Microarray database). The first three have been recommended by the Microarray Gene Expression Data (MGED) Society for storing public available microarray datasets (Parkinson *et al*, 2005).

1.4.3.1. GEO (http://www.ncbi.nlm.nih.gov/geo/)

The Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI) was built in 2000, which provides free access to the published gene expression data sources. The GEO project provides the simplest procedure and free-download mechanisms of high-throughput gene expression data (Barrett *et al.*, 2008). At Sept 2013 the database contained microarray datasets of over 300,000 samples of over 10,000 experiments from around the world.

The GEO includes three types of submission entity; 'platform' (a record for a summary of array-based platform, e.g., Agilent, Affymetrix), 'series' (a record for linking a group of related samples), and 'sample' (a record for conditions of each individual sample). The 'platform' is the parent platform which should be defined first (Edgar *et al.*, 2002). The 'platform' set contains a microarray data table and a brief introduction of these data to present the main features of the array e.g. cDNA microarray, oligonucleotides microarray and so on. Every platform begins with 'GPL (GEO Platform)' followed by a unique GEO accession number (Barrett *et al.*, 2008). A 'sample' record contains a data table with the experimental material and method protocols. Each 'sample' matches one

'platform'; however, the sample should be contained in multiple 'series' (Barrett *et al.*, 2008). Every sample record begins with 'GSM (GEO Soft format Sample file)' followed by a unique GEO accession number (Barrett *et al.*, 2008). A 'series' record is a group of related 'samples'; it may include one or more tables. Every 'series' begins with 'GSE (GEO Series format file)', followed by a unique GEO accession number (Barrett *et al.*, 2008).

The GEO deposit options and formats were shown on below table.

Table 3: GEO deposit options and formats (Barrett *et al.*, 2008)

Option	**Formation**	**Key features**
Web deposit	Web forms	Deposit of individual records. Simple step-by-step interactive web forms.
GEO archive	Spreadsheets (e.g. Excel)	Batch deposit. Good choice for most users who have many samples to submit.
SOFT (Simple Omnibus Format in Text)	Plain text	Batch deposit. A simple, line-based, tab-delimited format that can be readily generated, particularly if the data are already in a database.
MINiML (MIAME notation in Markup Language)	XML	Batch deposit. Basically an XML rendering of SOFT format, and similarly suitable if data are already in a database. The XML schema definition is available at the GEO website.

1.4.3.2. ArrayExpress (http://www.ebi.ac.uk/arrayexpress/)

The ArrayExpress database is available as an international microarray data repository of the European Bioinformatics Institute (EBI) since 2002 (Brazma *et al*, 2003). The ArrayExpress database contains gene names, gene function and other information related to the gene (Parkinson *et al*, 2005). The data comes from two sources: one is users' submission, the other one is mapping from the NCBI Gene Expression Omnibus. GXA (Gene Expression Atlas) is a separate database available from the ArrayExpress interface (Parkinson *et al*, 2011). The ArrayExpress and GEO have an agreement for data exchange that is ArrayExpress connects to GEO to import all GDS and GES data. Therefore all of the submitted high-throughput and HTP sequencing data in GEO will appear in ArrayExpress (Parkinson *et al.*, 2011).

1.4.3.3. CIBEX (Center for Information Biology gene EXpression database) (http://cibex.nig.ac.jp/)

CIBEX is a public microarray expression database constructed for organizing MIAME (Minimum Information about Microarray Experiment), and began storing microarray data in 2004. The CIBEX is one of the primary databases run by the Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ) and this database is now organized by the Microarray Gene Expression Data Society (MGED) (Kodama *et al.*, 2010). Most of submitted data comes from Japan, and is automatically presented in the related tables (Ikeo *et al.*, 2003). The submitted data contains its own CIBEX accession number, started with

"CBX". CIBEX has links to other databases, like PubMed (Ikeo *et al.*, 2003). From 2009, the CIB-DDBJ exchanges GEO and ArrayExpress data to join the International Database Plan (Kodama *et al.*, 2010).

1.4.3.4. SMD: Stanford microarray database (http://smd.stanford.edu)

The SMD provides a platform for biological researchers to analyze, share, view and annotate microarray data of more than 60 organisms. Over 70,000 microarrays have been stored in the SMD by Sept 2013. There are about 9,000 sets of *Homo sapiens*-specific data and have been used in over 400 published research articles. The SMD stores microarray data generated by multiple platforms, like spotted cDNA microarray, oligonucleotide microarrays, Affymetrix, Agilent (Hubble *et al.*, 2009). SMD provides a biological annotation of genes and sequences in each organism, and also has some annotation tools for users dealing with their own data. For example, MAGE-ML files writing tools (Parkinson *et al.*, 2007), directly with the ArrayExpress and Gene Expression Omnibus (Wheeler *et al.*, 2007). Registration is required to use of these tools for selected, downloaded, analyzed data etc. (Hubble *et al.*, 2009).

1.4.4. Microarray data analysis

1.4.4.1. Data pre-processing

Two kinds of microarray techniques have been used widely for comparative studies in order to detect changes in gene expression, including the spotted complementary DNA (cDNA) and the oligonucleotide microarray. Each single microarray experiment can detect the expression level of thousands of genes. There exist experimental noise and systematic errors in the raw data, which can greatly influence understanding of the true biological information (Bilban *et al*, 2002). Pre-processing and normalization is thus always necessary in microarray data analysis.

Data pre-processing of microarrays is done by the following steps: gene mapping, gene filtering, and data normalization. For gene mapping, the relationship between probe-sets (unique and identifiable set of individual probes) and genes is arranged. When multiple probe-sets map to one gene in the dataset, the expression value of that gene should be extracted from the associated probe-sets. Otherwise, when one probe-set is mapped to multiple genes, these genes will be assigned with the same expression value. Details are shown in section 2.2.

Gene filtering focuses on the genes that have significant expression changes across samples and on the availability of gene controlling DNA sequences (Yue *et al*, 2001). Several factors are frequently used for gene filtering, including standard deviation (SD) and p-value. Genes with associated large SD indicate that their expression difference between, for example, normal and disease

condition is significant, whereas genes with a relatively smaller SD will be filtered out. Furthermore, there often exists missing values in experiments. If this involves missing values in considerably large number of samples, it will be excluded in the analysis. The ratio of missing value for a gene determines the gene filter using a predefined threshold value (Hackstadt & Hess, 2009). For a gene with a small number of missing values, a non-parametric method called "KNN" (k-nearest neighbors algorithm), which is a type of instance-based learning approach, is applied to fill the missing values of genes (Altman, 1992). The 'knn' 'R package' was used in this study. K is a user-defined parameter for counting the numbers of neighbors to be applied to estimate the expression of the missing one, i.e. the values of k nearer neighbors are used to estimate the missing value, by weighting the associated neighbors as 1/*d*, where *d* is the distance of the neighbor (Altman, 1992).

Normalization of microarray data is done by fitting (filtering or smoothing) the raw data in order to enable the microarray data is comparable. The classical approaches for normalizing of expression data include linear regression analysis (Chatterjee & Price, 1991), rank invariant methods (Tseng *et al*, 2001), log centering and so on. But, none of these can deal adequately with the possible systematic bias caused by microarray noise. To address this, some studies suggested that the $\log_2$ (ratio) values should be applied to microarray analysis, which a dependence relationship with intensity value in microarray. LOWESS (Locally Weighted Scatterplot Smoothening) analysis was proposed and applied to remove the effect of intensity dependence in a binary logarithm ratio (Yang *et al.*, 2002). Non-linear normalization is adopted for high-density-

oligonucleotides (Affymetrix) and cDNA microarrays. Normalization factors (like median value or mean value) are used to process the expression value of genes (Welsh *et al.*, 2001).

Comparative analysis of microarray data

1. ***Fold change:***

The comparison of expression levels of genes across samples is one of the main steps in microarray studies. It is intended to determine differences in expression between genes in various biological samples. For example, fold change, widely used in microarray analysis, is defined by:

$$T_i = \frac{R_i}{G_i}$$

where T_i is the ratio of *i*th gene, and R_i and G_i represent the expression levels of the associated gene.

The formula is not limited to any microarray platform, the measurement of R and G can be taken from either one array or two arrays (Quackenbush, 2002).

2. ***Statistical tests:***

Microarray studies can identify differentially-expressed genes across different samples to reveal biomarker genes. To make project with this aim, statistical

tests are widely used, including three main statistical tests as described below.

2.1. t-test

The t-statistic measures the distance between the two samples in units of standard deviation, based on comparing the between- or within-group differences (Cui & Churchill, 2003). The calculation shows the significant difference of a gene from the mean expression level of group A and group B:

$$t = \frac{\overline{X_A} - \overline{X_B}}{S}$$

where $\overline{x_A} = \frac{\sum_{k=1}^{n_A} x_k}{n_A}$; $\overline{x_B} = \frac{\sum_{k=n_A+1}^{n_A+n_B} x_k}{n_B}$; $S = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$;

$S_A = \sqrt{\frac{\sum_{k=1}^{n_A}(X_k - \bar{X}_A)^2}{n_A - 1}}$; $S_B = \sqrt{\frac{\sum_{k=n_A+1}^{n_A+n_B}(X_k - \bar{X}_B)^2}{n_B - 1}}$

x_k, k=1~n_A+n_B, is the kth *expression value* of a gene in the independent study.

2.2. Analysis of Variance (ANOVA)

ANOVA is used for microarray analysis of variance where multiple groups (e.g. normal group, disease group…) are involved. The significant genes retrieved by ANOVA (F-statistics) between two groups should be the same as genes retrieved by t-statistics, because two group-ANOVA is formally

equivalent to the t-test. The ANOVA test is based on calculating the Sum of squared residuals within (SS_w) or between (SS_B) groups and should be calculated separately (Hinkelmann & Kempthorne, 2008).

$$SS_w = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 \qquad (1)$$

$$SS_B = \sum_{j=1}^{m} n_j\left(\bar{X}_j - \bar{X}\right)^2 \qquad (2)$$

$$\mathrm{F}_{m-1,N-m} = \frac{Mean\ squared\ between}{Mean\ squared\ within} = \frac{SS_w/(m-1)}{SS_B/(N-m)} \qquad (3)$$

where X_{ij}, *i*=1~n, is the *j*th independent study. m is the number of groups; N is the total number of cases. $\bar{X}$**:** Ground mean=sum of all values/N.m-1or N-m: the degrees of freedom.

2.3. Significance analysis of microarrays (SAM)

The SAM is a statistical technique for microarray data analysis for measuring the changes in gene expression as statistic *d* (Tusher *et al*, 2001). The SAM analysis is defined by the difference between two groups and the standard deviation adjusted by an exchangeability factor s_0 which is as denominator of test statistic, default is automatic choice (Tusher *et al*, 2001).

$$\boldsymbol{d = \frac{r}{s + s_o}}$$

where, $s = \sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\frac{\sum_{k=1}^{n_A}(X_k - \bar{X}_A)^2 + \sum_{k=n_A+1}^{n_A+n_B}(X_k - \bar{X}_B)^2}{n_A + n_B - 2}}$; $r = \bar{X}_A - \bar{X}_B$;

x_k, k=1~n_A+n_B, is the kth *expression value* of a gene in an independent study, s_0 is the exchangeability factor, and can be any percentile of the s value. When s_0=0, the d-statistic is equal to t-statistic.

1.5. Meta-analysis of Microarray Data

1.5.1. Introduction of meta-analysis

The definition of meta-analysis was first given by Glass in 1976 as "the statistical analysis of a large collection of analysis results for the purpose of integrating the findings" (Glass, 1976). Meta-analysis is a systematic approach and combines results of multiple studies using a set of statistical techniques, in order to increase the reliability of results (Ramasamy *et al.*, 2008; Hong & Breitling, 2008). It has been demonstrated that meta-analysis has the ability to increase the effect of statistics by amending the false negative rate of an individual study (Choi *et al*, 2003). Meta-analysis is a useful tool not only for individual primary studies the may contain inconsistent results, but also for extracting valuable information from individual primary studies with inadequate sample sizes (Cochran, 2007). Meta-analysis to analyse microarray data involves 7 steps, shown in below Table 4 (Ramasamy *et al.*, 2008).

Table 4: The seven steps of meta-analysis (Ramasamy *et al*, 2008)

Steps	Meta-analysis
1	Identify suitable microarray studies
2	Extract the data
3	Prepare individual datasets
4	Annotate individual datasets
5	Resolve the many-to-many relationship between probes and genes (like one gene contains multiple probe-sets, or one probe-set represents more than one gene)
6	Combine the study-specific estimates
7	Analyze, present and interpret results

1.5.2. A review on meta-analysis of microarray data

Published microarray studies have been produced for the same biological topic, e.g., they focus on the same disease or biological phase. Integrated analysis across multiple microarray datasets was hoped to yield more robust research interpretations than would individual studies. To integrate multiple microarray datasets and enhance the reproducibility of research findings, two approaches have been applied:

1. Merging multiple studies: through the combination of raw data of primary studies. The merits and complications of such primary datasets are reviewed, here, different meta-analysis methods in different disease datasets have been reported (Larsson *et al*, 2006).

2. Comparing and analysing the results of different published studies. Microarray data may contain some level of 'noise' ('unreal' data produced in experimental process, e.g., the background noise) (Cahan *et al*, 2007). The quality of meta-analysis depends on the quality of each individual microarray dataset. It is known that not all of microarray raw data deposited in public databases have equally quality thus contain unequally meaningful information (Larsson & Sandberg, 2006). The accuracy and reproducibility of microarray data have therefore, presented a big challenge (Severgnini *et al*, 2006), not least when different microarray datasets are produced by different platforms, backgrounds, and with different samples types (e.g., cell lines vs. biopsy).

While special care is needed when processing raw microarray data, researchers have successfully gained much valuable information from microarray studies. For example, a cross-species comparable analysis of yeast and human has determined a common transcriptional profile in aging (McCarroll *et al.*, 2004), and a common host transcriptional response to pathogens (Janner & Young, 2005). Wennmalm et al found the similarity between the expression pattern of aging and cellular senescence in mice (Wennmalm *et al*., 2005). An aging database was built for collecting data of microarray studies (Pan *et al*., 2007). A smaller robust gene signature for acute myeloid leukemia diagnoses has been revealed by comparing potential diagnostic genes reported in multiple studies (Cahan *et al*, 2005).

As alluded to above, enhanced insights can be gained by merging data from

multiple studies. In 1999, researchers realized that discoveries could be made through the synthesis of related but individual experiment microarray studies (Khan *et al*., 1999; Normand, 1999). It is hard to compare directly microarray datasets obtained in different array experiments (Kuo *et al.*, 2002). However, using this approach, Rhodes and his colleagues identified differentially-expressed genes between benign and localized prostate cancer tissue by meta-analysis of four prostate cancer datasets (Rhode *et al.*, 2002). As discussed in Choi et al study, meta-analysis is able to reduce false negatives of individual analysis and thus increased the effectiveness of the statistical analysis of microarray data (Choi *et al.*, 2003). When comparing results of meta-analysis with that of independent studies, the reproducibility was significantly improved in meta-analysis studies (Hong *et al.*, 2006). Similarly, Park and Stegall investigated gene expression level of cytokines through the combination of data from open source and their own microarray datasets (Park & Stegall, 2007).

However, two contradictory results were showed on two recent meta-analyses of melanoma microarray studies (Tímár *et al.* 2010, Schramm *et al.* 2011). The first compared gene signatures derived from 4 microarray datasets of human melanoma tissues, and found very little overlap between these signatures (Tímár *et al.* 2010). They considered this lack of congruence, (which is also discussed in this study, see 1.5.4 section) may be induced by sample heterogeneity (like different platforms, different types of samples etc.). By adding 5 additional studies, another team demonstrated some significantly over-represented functions among the melanoma-associated gene signatures (Schramm *et al.* 2011). For example, most of the genes signatures were related

to the immune response. Moreover, a 'leave-one-out' cross validation of the data showed that a low average error rate (28%) was achieved across all validation expression data of the gene signature (Mann *et al.* 2010). A gene signature proposed by Jönsson et al. (2010) reported an error rate of 45%.

To deal with the heterogeneity of datasets (including various types biological sample, platforms, and quality of datasets), a new methodology is needed. In this study, a new method is proposed which takes into considerations of two factors ('stratification' and 'weighting') (see section 2.2).

1.5.3. Methods used for microarray meta-analysis

Typically, three typical methodologies have been applied in the meta-analysis of integrated analysis of multiple studies, including t-test, ranking and Fisher's inverse Chi-square approaches. The t-test based approach was adopted in Choi et al, 2003, and is implemented in the *GeneMeta* package of Bioconductor (R package). The ranking based approach (a non-parametric statistic method), first applied for the analysis a single dataset (Breitling *et al.*, 2004) was then applied to measure each gene in combined multiple studies as implemented in the *Rank Prod* Bioconductor package (Hong *et al.*, 2006). Finally, Fisher's inverse method is based on an improved Fisher's inverse X^2 test to combine the individual studies (Zaykin *et al.*, 2002) based on their p-values. The p-values can be calculated by the t-test or ranking-based approaches (Hong & Breitling, 2008).

The three approaches above were compared using two datasets, and evaluated for their advantages and disadvantages (Hong & Breitiling, 2006). As Fisher's product is based on p-values, which can be obtained from both *t*-test-based and ranking-based approaches, the authors excluded it and instead focused on comparing the *t*-test-based approach and the Ranking-based approach. The *t*-test based approach, which was improved from Student's *t*-test, directly compares the magnitude of gene expression under different biological conditions. The ranking based approach provides the relative importance of a gene compared to other genes. When studies have a small sample size (large sample size in biological experiments is uncommon), the ranking based approach has its important advantage compared to the *t*-test based approach. The latter can suffer unreliable error estimates when sample sizes are small (Hong & Breitling, 2008).

1.5.4. The challenges of meta-analysis on microarray datasets

One of the main issues affecting individual microarray studies is that the testing power is often weakened by a large number of hypotheses being applied to a small sample size database. Thus, when the false positive rate is 0.05, in 24,000 transcripts of a microarray study, there are 1,200 genes (i.e., 0.05 × 24,000) that could be random fluctuations i.e., false positive genes. Meta-analysis has the potential of reducing these false positives, such that the truly significant genes will be determined by combining different studies. Similarly,

meta-analysis also can reduce the influence of different experimental conditions (Cahan *et al*., 2007).

Another important issue related to microarray study is the poor congruence between gene signatures identified by different individual microarray-based studies, e.g., non-overlapping melanoma signatures (John *et al.*, 2008; Bittner *et al.*, 2000; Tímár *et al.*, 2010). Microarray technology has also been applied to the comparative analyses of different melanoma stages and has identified various gene signatures (Hoek, 2007). However, microarray-based melanoma gene markers have thus far had less than optimal translation to the clinical situation, and the diagnosis of melanoma is still largely based on the clinical and histopathological features of the tumour (Schramm *et al.*, 2011).

In order to explore this poor congruence issue, the current project conducted a comparison on 16 independent microarray-based signatures of metastatic melanoma published from 2000 to 2011. These 16 studies of melanoma only reported a metastatic melanoma gene signature in all melanoma published microarray studies. The start year is 2000, which was the earliest melanoma microarray study in the literature. These gene signatures involved various numbers of genes ranged from 5 to 589 genes. Remarkably pair-wise comparison showed that the microarray-based gene signatures shared very few common genes (Table 5). For example, only 84 genes were common to two of the signatures (Scatolini *et al.*, 2010, Jaeger *et al.*, 2007), 14 to three studies (Scatolini *et al.*, 2010, Jaeger *et al.*, 2007 and Riker *et al.*, 2008), and strikingly only 2 common genes (*KRT15, RORA*) appeared in four of the 16 studies

(Scatolini *et al.*, 2007, Jaeger *et al.*, 2007, Riker *et al.*, 2008 and Smith *et al.*, 2005). Remarkably no gene was found to be common in five or more of the independent studies (Table 6). This finding suggests that some fundamental issues exist, for either the individual microarray studies designed, or the suitability of direct comparison meta-analyses.

Table 5: The 16 individual studies from 2000 to 2011 used in this study

Study	Gene No of metastatic melanoma signatures	Summary	Data Access No of GEO/ArrayExpress
Hoek *et al*, 2004	589	normal melanocytes compare to metastatic melanoma cell	GSE4570
Haqq *et al*, 2005	19	early stages (skin, novi and primary) compare to metastases type I & type II	
Smith *et al*, 2005*	94	early melanoma(NS,BN,MIS) compare to advanced melanoma(VGP,MGP,MGP positive lymph nodes)	GSE4587
Okamoto *et al*, 2005	20	Five malignant melanoma cell lines	GSE1845
Mandruzzato *et al*, 2006	70	stage III & stage IV melanoma to identify genes associated with patient survival	E-TABM-1,2,4
Winneperninckx *et al*, 2006*	235	CMM without metastasis compare to CMM with metastasis	
Alonso *et al*, 2007	243	CMM without metastasis compare to CMM with metastasis	
Jaeger *et al*, 2007	308	primary melanomas compare to melanoma metastases	
John *et al*, 2008	21	stage IIIb and stage IIIc melanoma poor-prognosis compare to good-prognosis	E-TABM-403
Riker *et al*, 2008	66	primary melanoma compare to metastatic melanoma	GSE7553
Bogunovic *et al*, 2009	209	genes of 44 metastatic melanoma samples are significantl associated with post-recurrence survival	E-TABM-403
Jeffs *et al*, 2009	96	higher MITF in melanoma progression compare to lower MITF in melanoma cell lines	GSE16249 & GSE16404
Koh *et al*, 2009**	14 of 36	nevus compare to melanoma	
Kashani-Sabet *et al*, 2009	5	nevus compare to melanoma	
Kabbarah *et al*, 2010	30	primary melanoma compare to metastatic melanoma	GSE7606
Scatolini *et al*, 2010	455	vertical growth phase melanoma compare to metastatic melanoma	GSE12391
* 94 and 235 are numbers of gene, respectively corresponding to 100 and 254 microarray probe-sets.			
**14 of 36 signatures were listed in Kohn et al. (2009) though it indicated 36 genes were identified.			
CMM: cutaneous malignant melanoma			

Table 6. Pair-wise comparisons of 16 independent studies of melanoma and their association distribution of common genes*.

	Alonso *et al* 2007 (243)	Bogunovic *et al* 2009 (209)	Haqq *et al* 2005 (19)	Hoek *et al* 2004 (589)	Jaeger *et al* 2007 (308)	Jeffs *et al* 2009 (96)	John *et al* 2008 (21)	Kabbarah *et al* 2010 (30)	Kashani-sabet *et al* 2009 (5)	Koh *et al* 2009 (14)	Mandruzzato *et al* 2006 (71)	Okamoto *et al* 2005 (20)	Riker *et al* 2008 (65)	Scatolini *et al* 2010 (455)	Smith *et al* 2005 (94 of 100)	Winnepenninckx *et al* 2006 (235)
Alonso *et al* 2007 (243)		2	1	12	4	2	0	1	1	0	0	0	1	5	1	7
Bogunovic *et al* 2009 (209)			1	9	3	2	0	2	0	0	6	0	2	1	0	6
Haqq *et al* 2005 (19)				7	1	0	0	0	0	0	0	0	0	1	1	1
Hoek *et al* 2004 (589)					34	19	0	7	1	2	5	2	8	17	6	4
Jaeger *et al* 2007 (308)						3	0	1	2	0	1	1	27	84	9	14
Jeffs *et al* 2009 (96)							0	0	0	0	2	2	0	0	2	1
John *et al* 2008 (21)								0	0	0	0	0	0	0	0	1
Kabbarah *et al* 2010 (30)									0	0	0	0	2	1	0	3
Kashani-sabet *et al* 2009 (5)										0	0	0	1	0	1	0
Koh *et al* 2009 (14)											0	0	0	0	1	0
Mandruzzato *et al* 2006 (71)												0	0	3	0	3
Okamoto *et al* 2005 (20)													0	0	0	0
Riker *et al* 2008 (65)														16	7	4
Scatolini *et al* 2010 (455)															4	14
Smith *et al* 2005 (94 of 100)																1
Winnepenninckx *et al* 2006 (235)																

*The numbers bracketed are the numbers of genes in the original signature. Studies of four microarray datasets applied in our study were highlighted with underline.

The aims of the study:

The aim of this study is to develop a robust model for meta-analysis of multiple microarray data, including:

1. Develop a bioinformatics approach to identify robust gene biomarker signature from multiple microarray data.
2. Identify a gene signature for melanoma, and validate them by computational and laboratory-based studies.

In this study, I propose a new model that measures the genome-wide relative significance (GWRS) and genome-wide global significance (GWGS) of gene expression. As will be described in this thesis, this new model provides an important advantage for the integrative analysis of microarray datasets produced by different platforms and protocols. Using this method, 200 top genes of melanoma were identified based on the integrated analysis of five melanoma data. Based on the relationship of those 200 genes with melanoma driver genes, 12 genes were defined as a metastatic melanoma biomarker. 4 genes were validated by wet-lab experiments, showing clearly differential expression in melanoma cells than normal cells.

To enhance the computational model further, and to make the method suitable to the datasets of other diseases, I later take into account the concept of 'Stratification' (classify microarray datasets by different features, e.g., types of platforms, types of samples etc.) and Dataset quality (ω) to deal with the datasets in the integrated analysis, and call it the 'Stratification with weighting' method. Extensive evaluations on this new integrated analysis approach were

performed.

Computational experimental results, on melanoma, breast cancer and prostate cancer, demonstrated the enhanced of performance of this method which could generate more robustly associated genes. As shown in the evaluation against of 180 known metastatic melanoma biomarkers, the ranking positions of robust genes were higher than the genes identified by previous method. This indicates that the enhanced method could identify more ‘stable’ and ‘reliable’ gene associations (i.e., the genes may be considered as biomarker).

2. MATERIAL AND METHODS

2.1. Microarray data

Fourteen datasets were used for this project across in three kinds of cancers (i.e. melanoma, breast and prostate) for research on gene association.

Five melanoma microarray datasets were selected (see table 7, melanoma part) to investigate a robust biomarker of genes using GWRS (Genome-wide relative significance) which investigates the signature within individual microarray dataset, and GWGS (Genome-wide global significance) which investigates the signature across multiple microarray datasets. These five datasets were selected as they containing gene expression values of normal skin and/or benign nevi, and metastatic melanoma. Four of them were contained in the 16 metastatic melanoma studies published between 2000 and 2011 (see table 5) and included open access databases as shown on table 8. The microarray data were extracted from the GEO database (GEO access number: GSE7553, GSE4587, GSE4579, and GSE12391). An additional GSE22301 dataset was extracted from Rose et al. (2011). Rose et al. did not provide the gene signature of metastatic melanoma in the paper, thus it was not included in the meta-analysis of 16 studies. However, it does include 14 samples of metastatic melanoma data and so the microarray data was included in our integrative analysis. As a result a total of five microarray datasets of normal and/or benign nevi and metastatic melanoma were used in this study (Table 7 - the melanoma

datasets).

A further aim was to enhance the computational method beyond melanoma for generalized use by including two groups of datasets (i.e., breast and prostate cancer) to evaluate the effectiveness of the methods overall. Details of the three groups of datasets are shown on table 7 and described in the following sections.

Table 7: Details of the 3 groups of datasets used in the project

Data number in our study	Study	Microarray data access number	Platform	Sample size of control	Sample size of disease
Breast Cancer					
Data1	Mecham *et al*, 2004	GSE1299	Affymetrix Human Genome U133A Array	2	4
Data2	Richardson *et al*, 2006	GSE3744	Affymetrix Human Genome U133 Plus 2.0 Array	7	40
Data3	Casey *et al*, 2009	GSE10797	Affymetrix Human Genome U133A 2.0 Array	10	56
Data4	Turashvili *et al,* 2007	GSE5764	Affymetrix Human Genome U133 Plus 2.0 Array	20	10
Data5	Liu *et al*, 2007	GSE6883	Affymetrix Human Genome U133A Array	3	9
Prostate Cancer					
Data1	Tomlins *et al*, 2007	GSE6099	Chinnaiyan Human 20K Hs6	34	52
Data2	Chandran *et al*, 2007	GSE6919	Affymetrix Human Genome U95B Array	17	91
Data3	Nanni *et al*, 2006	GSE3868	Affymetrix Human Genome U133A Array	4	20
Data4	Varambally *et al*, 2005	GSE3325	Affymetrix Human Genome U133 Plus 2.0 Array	6	14
Melanoma					
Data1	Hoek *et al*, 2004	GSE4570	Affymetrix Human Genome U133A Array	2	5
Data2	Smith *et al*, 2005	GSE4587	Affymetrix Human Genome U133 Plus 2.0 Array	2	5
Data3	Riker *et al*, 2008	GSE7553	Affymetrix Human Genome U133 Plus 2.0 Array	5	40
Data4	Scatolini *et al*, 2010	GSE12391	Agilent-012391 Whole Human Genome Oligo Microarray G4112A	18	5
Data5	Rose *et al,* 2011	GSE22301	Affymetrix Human Genome U133A 2.0 Array	2	14

Table 8: Publically available microarray datasets of metastatic melanoma 2000-2011 used in this study

study	microarray data access number	normal skin	benign nevi	dysplastic nevi	melanoma in situ	vertical growth phase	metastatic melanoma
Mandruzzato *et al*, 2006	E-TABM-1				83		
	E-TABM-2				17		
	E-TABM-4				9		9
John *et al*, 2008	E-TABM-403						58
Riker *et al*, 2008	GSE7553	5			2	14	40
Jeffs *et al*, 2009	GSE16040						34
Kabbarah *et al*, 2010	GSE7606	ratio value of comparing human melanoma with normal skin for 123 samples					
Rose *et al*, 2011	GSE22301	2			4		14
Smith *et al*, 2005	GSE4587	2	2	2	2	2	5
Hoek *et al*, 2004	GSE4570	2			1		5
Scatolini *et al*, 2010	GSE12391		18	11	8	15	5

2.1.1. Melanoma microarray datasets used in this project

2.1.1.1. Hoek Data (GSE4570, 2004)

Hoek Data (GSE4570)			
Number of samples	Normal melanocyte	Primary melanoma	Metastatic melanoma
8	2	1	5

The Hoek dataset was obtained using Affymetrix U133A. It contains 8 samples: 2 from normal melanocyte, a primary melanoma and 5 metastatic melanoma. In the original study of Hoek et al (2004), 589 genes were identified to be significantly differentially-expressed between normal melanocytes and melanoma (>2.5 fold change). Of these genes, 315 were up-regulated and 274

were down-regulated. They reported some novel pathways in melanoma cells, including NOTCH activation, as well as an altered expression in embryonic development and epidermal transcriptional regulators, activation of cancer antigens, and down-regulation of growth suppressors (e.g. NECDIN).

2.1.1.2. Smith data (GSE4587, 2005)

Smith data (GSE4587)					
Number of samples	Normal	benign nevus	atypical nevus	primary melanoma	metastatic melanoma
18	4	2	2	4	6

The Smith dataset (GSE4587) is a mixed sample type dataset and was derived from 15 frozen tissues and 3 cell lines, RNA was extracted from 2 normal biopsies, 2 benign nevi, 2 atypical nevi, 2 melanoma in situ, 2 vertical growth phased melanoma, 2 metastatic growth phase melanoma, 3 lymph node metastasis melanoma tissues, 2 normal epidermal cell lines, and 1 metastatic melanoma cell line. These authors conducted a comprehensive study on the different stages of malignant melanoma development, based on whole genome expression profiles, and investigated the top 50 up-regulated and 50 down-regulated genes in advanced-stage melanoma compared to early-stage melanoma.

2.1.1.3. Riker data (GSE7553, 2008)

Riker Data (GSE7553)							
Number of samples	Normal	melanocytes	melanoma in situ	primary melanoma	metastatic melanoma	basal cell carcinoma	squamous cell carcinoma
87	4	1	2	14	40	15	11

The Riker dataset (GSE7553) contains 87 samples. In addition to 15 basal cell carcinoma samples and 11 squamous cell carcinoma samples, the sample set also included 56 melanoma samples, 2 samples of melanoma in situ, 14 primary melanoma, 40 metastatic melanoma samples, and 4 normal skin and one from melanocytes. The authors compared 40 metastatic melanoma (MM) samples to 16 primary melanoma samples (14 primary melanoma samples plus 2 melanoma in situ samples). They identified that the expression of *SPRR/A/B*, *KRT16/17*, *CD24*, *LOR*, *GATA3*, *MUC15*, and *TMPRSS4* were dramatically higher in primary basal cell carcinomas, squamous cell carcinomas and thin melanomas compared to metastatic melanoma. Conversely, expression of *MAGE*, *GPR19*, *BCL2A1*, *MMP14*, *SOX5*, *BUB1*, and *RGS20* was higher in metastatic melanoma than in the other sample types. They also identified 65 differentially-expressed genes by comparing normal human epidermal melanocytes to thin primary cutaneous and metastatic melanoma samples.

2.1.1.4. Scatolini data (GSE12391, 2010)

Scatolini data (**GSE12391**)				
Number of samples	benign nevus	atypical nevus	primary melanoma	metastatic melanoma
57	18	11	23	5

The Scatolini data (GSE12391) used RNA isolated from a total of 57 freshly-harvested patient tissues from 5 stages of progression from normal to metastatic melanoma were involved. Agilent oligonucleotide glass arrays were used for hybridization, and the data were presented by 10 base logarithms after processing and normalization using the Agilent Human Whole Genome platform-specific error model. The log ratio value of gene expression was treated using the LIMMA package, involving different functions of LIMMA incl. “backgroundcorrect”, “normalizeWithinArrays” and “normalizeBetwwenArrays”. The authors determined the 5 different stages from normal to melanoma by pair comparisons. They found 36 annotated transcripts were differently expressed between benign nevi and primary radial growth phase melanomas, including *GDF15*, *AMICA1*, and *GLA*.

2.1.1.5. Rose data (GSE22301, 2011)

Rose data (GSE22301)			
Number of samples	melanocytes	primary melanoma	metastatic melanoma
22	4	4	14

Here, there are 22 samples separated into three types. Four samples represent normal melanocyte controls, 4 primary melanoma and 14 metastatic melanoma cell lines. By investigating the differences between superficial spreading melanoma and nodular melanoma, Rose et al identified 8 significant genes (i.e., *DIS3*, *FGFR1OP*, *G3BP2*, *GALNT7*, *MTAP*, *SED23IP*, *USO1* and *ZNF668*) that were differentially expressed.

2.1.2. Breast cancer microarray datasets using in this project

2.1.2.1. Mecham data (GSE1299, 2004)

Mecham data (**GSE1299**)		
Number of samples	normal breast epithelium	breast cancer cells
6	2	4

This microarray dataset was derived from 6 samples including 4 human breast tumour cell lines and 2 human mammary epithelial cells cell lines (HMEC). The authors assessed gene expression in these samples across 4 platforms (i.e. Affymetrix U133A, U133B, U95 Version 2 Arrays, and Agilent Human 1 cDNA microarray [G4100A] array). They introduced RNA aliquots to a cross-platform analysis, and found they could significantly improve the consistency of platforms. They concluded that their cross-platform analysis strategy yielded more efficient results from different cDNA microarray and Affymetrix gene-chip platforms than single platform of one.

2.1.2.2. Richardson data (GSE3744, 2006)

Richardson data (**GSE3744**)				
Number of samples	Normal	Non-BLC	BLC	BTCA1 associated cancer
47	7	20	18	2

This dataset contains 47 samples, including 7 normal breast samples, 20 non-basal like cancer samples (non-BLC), 2 BRCA1-associated cancer samples, and 18 basal-like cancer (BLC) samples. Analysis of the microarray data revealed that the active X chromosome had been duplicated and the inactive X

chromosome had been lost in almost half of BLC samples. While the abnormalities of the X chromosome did not change the global X chromosome transcription, it was associated with overexpression of a small group genes located on this chromosome. Because the association between abnormal X chromosome and gene changes did not show in the non-BLC samples, these results suggest that the X chromosome abnormalities may have been contributed in BLC.

2.1.2.3. Casey data (GSE10797, 2009)

Casey data (**GSE10797)**		
Number of samples	normal	invasive breast cancer tissues
33	5	28

Here, gene expression was assessed using total RNA of epithelial and stromal cells from 5 normal breast specimens and 28 invasive breast cancer tissues by Affymetrix U133A 2.0 GeneChips. According to comparisons of gene expression of different cell types (epithelial or stromal) and diagnosis (normal or cancer), the researchers determined that the transcriptome of epithelial cancer was enriched for proliferative, motility and ECM gene ontologies when compared with normal epithelial tissue. The transcriptomes also showed that genes were overexpressed in ECM and proteolytic ontologies in invasive breast cancer compared to epithelial and stromal cancer tissues.

2.1.2.4. Turashvili data (GSE5764, 2007)

Turashvili data (**GSE5764)**				
Number of samples	normal ductal cells	normal lobular cells	invasive ductal carcinomas	invasive lobular carcinomas
30	10	10	5	5

In this database, Turashvili et al. focused on invasive ductal and lobular carcinomas; the two most common histological types of breast cancers. They examined 30 samples including 10 normal ductal cells, 10 normal lobular cells, 5 invasive ductal carcinomas (IDC) cells and 5 invasive lobular carcinomas (ILC) cells. When comparing the different samples via algorithm and rank products, they identified 84 significantly differentially-expressed genes between ILC and normal cells, 74 significant genes between IDC and normal cells, 78 significant genes between normal ductal and lobular cells, and 28 differentially expressed genes between IDC and ILC. When these changes were combined, the authors extracted seven differentially-expressed genes (i.e., *CDH1*, *EMP1*, *DDR1*, *DVL1*, *KRT5*, *KRT6*, and *KRT17*) as novel biomarkers of breast cancer, and validated these by PCR and immunohistochemistry on tissue microarrays.

2.1.2.5. Liu data (GSE6883, 2007)

Liu data (GSE6883)			
Number of samples	normal breast epithelium	non-tumourigenic breast cancer	tumourigenic breast cancer
12	3	3	6

This microarray dataset was compiled from samples of tumourigenic breast cancer cells and normal breast epithelium cells. The samples set included 3 normal breast epithelium cells, 3 non-tumourigenic breast cancer cells and 6 tumourigenic breast cancer cells. The authors compared gene expressions in the 12 samples, and generated 186 genes as a signature for invasiveness. They found that these 186 genes were significantly associated with overall survival and metastasis-free survival in breast cancer patients. When they compared the 186 genes with the prognostic criteria of the National Institutes of Health, the invasiveness gene signature could be used to separate high-risk from early breast cancer, and also for prognosis in medulloblastoma, lung cancer, and prostate cancer.

2.1.3. Prostate cancer microarray datasets using in this project

2.1.3.1. Tomlins data (GSE6099, 2007)

Tomlins data (GSE6099)					
Number of samples	benign epithelia	atrophic lesions	localized prostate cancer	prostatic intraepithelial neoplasia	metastatic prostate cancer
101	34	5	32	13	17

Total RNA was isolated from 101 specific cell populations of 44 individuals to examine genes associated with the progression of prostate cancer. These included 34 benign epithelia, 5 atrophic lesions, 32 localized prostate cancers, 13 prostatic intraepithelial neoplasia, and 17 metastatic prostate cancers. Through analyzing the gene signatures of over 14,000 resultant probe-sets, the authors generated a model of prostate cancer progression that included protein biosynthesis, E26 transformation-specific family transcriptional targets, androgen signaling, and cell proliferation. From the model, the signature of androgen signaling was found to be similar in high-grade prostate cancer and metastatic prostate cancer relative to the low-grad prostate cancer. This may explain the clinical grade of the tumour with its prognosis. In this study, the researchers emphasized that the integrative analysis of gene expression signatures is a useful tool to understand cancer biology.

2.1.3.2. Chandran data (GSE6919, 2007)

Chandran data (GSE6919)				
Number of samples	normal	normal adjacent to tumour	primary prostate tumour	metastatic prostate tumour
152	23	41	64	24

This database included 152 human samples of normal prostate tissue, normal prostate tissue adjacent to tumour, and primary and metastatic prostate cancer tissues. Gene expression profiles of 24 androgen ablation-resistant metastatic samples (4 patients) and 64 primary prostate tumour samples were analyzed to investigate differences between primary and metastatic prostate tumours. The authors found at least a 2 fold over-expression change in 415 genes, 364 of which were down-regulated in metastasis samples. These genes were associated with some androgen ablation pathways and other networks (e.g. cell adhesion, bone remodeling and cell cycle), including transcription factor Forkhead Box M1 (*FOXM1*) and cell adhesion molecule Osteopontin (*SPP1*).

2.1.3.3. Nanni data (GSE3868, 2006)

Nanni data (**GSE3868**)					
Number of samples	normal	benign	basaloid tissue	primary tumour	metastatic tumour
30	2	2	3	22	1

This *in vitro*-based dataset includes 30 samples representing cell lines established from 2 normal tissue, 2 benign hyperplasia tissues, 3 basaloid tissues, and 1 metastatic tumour sample, and from 22 primary prostate tumour samples and 1 metastatic prostate tumour sample. The goal of this study was to

generate a model with signatures of tumours based on gene expression profiling. The authors examined the model and suggest that it can be used for studying primary prostate cancer biology and also could characterize tumours for prognostic and predictive purposes.

2.1.3.4. Varambally data (GSE3325, 2005)

Varambally data (**GSE3325**)			
Number of samples	benign	primary tumour	metastatic tumour
19	6	7	6

This dataset includes 19 samples representing 6 benign prostate, 7 primary and 6 metastatic prostate cancers. During the analysis of cancer progression, which included comparing different stages of disease via high-throughput immuno-blotting, and also integrated analysis with transcriptomic data (i.e., gene microarray), the authors identified 64 proteins that were altered relative to benign prostate, and 156 proteins which were altered relative to metastatic prostate cancer. These differential alterations of protein expression were considered by the authors as possible predictors of clinical outcome in prostate cancer.

2.2. Data preparation

One important pre-processing step for microarray data is to extract the expression value for each individual gene from the associated probe-sets using a gene mapping approach. When a probe-set is mapped to multiple genes, all the genes are assigned with the same expression value. For example, '209994_s_at' associated to two genes '*ABCB1 / ABCB4*' in GSE4570, both genes are given the expression value of the '209994_s_at' probe-set. However, in many cases, a gene is associated with multiple probe-sets. In this case the expression value achieving the highest significance in differentially expression sample classes (e.g., normal vs. disease) is assigned to this gene. For example, for a gene associated with probe-sets 1, 2, 3, if the probe-set 1 has achieved the highest p-value in a statistical test in the comparative study then the expression value of probe-set 1 is assigned as the expression value of the gene in this study. The impact of selecting the expression value of a probe-set that has the mean-, or median-value in statistical testing was also tested in this study. I found that using the highest differentially-expressed probe-set (maxim-based method), it was possible to retrieve the most significant probe-set of a gene. Thus, the aim of this study to extract the most differentially-expressed genes across multiple studies was attempted.

As a result of the above approach, a list of genes (G) from each datasets was retrieved. The number of datasets was denoted by n. For n datasets, the total genes across n datasets were retrieved by taking all the gene sets together (G). Thus, the number of genes of these n datasets is denoted by m, i.e. m=|G|. The value 'NA' was applied in cases where a gene is absent from an individual

study. A gene is also removed from G where NA is greater than a pre-defined value δ (δ=2 in this study, i.e., keeping the genes which were included in more or equal to 3 of 5 datasets). This means a gene was removed for further analysis if it is absent for more than two of five datasets. This resulted in a total of m=24,097 genes and n=5 of datasets for this study.

2.3. Melanoma biomarker detection

2.3.1. The application of melanoma datasets

This study focused on investigating the differential gene expression between normal skin and/or benign nevi and metastatic melanoma. The melanoma datasets was described in section 2.1 and table 7. Figure 7 shows the experimental protocol for melanoma biomarker detection and validation.

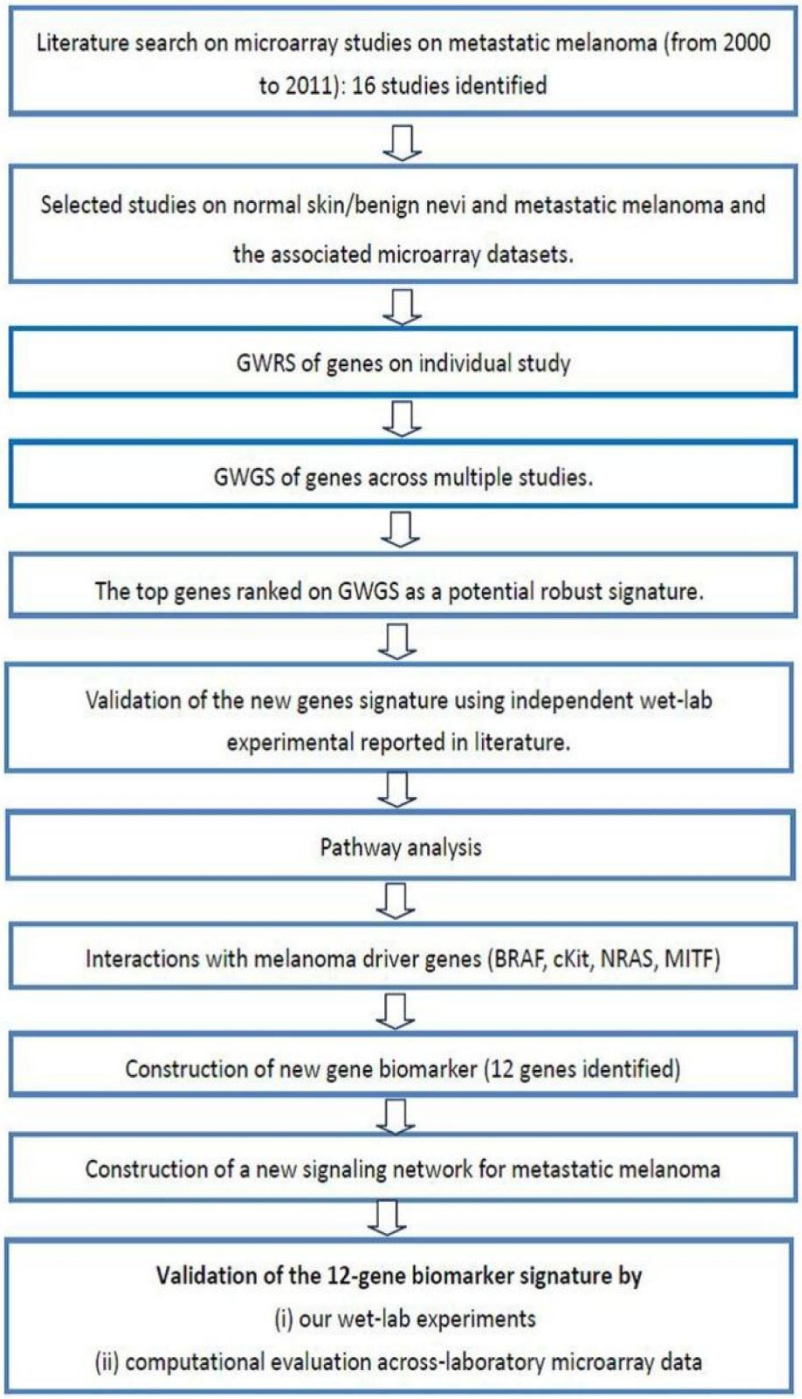


Figure 7: Experimental protocol for Melanoma Biomarker Detection and Validation

2.3.2. Genome-wide relative significance (GWRS) and Genome-wide global significance (GWGS) for integrated analysis of cross-laboratory (between independent studies) microarray data

A relatively simple method of integrative meta-analysis was proposed by Rhodes et al in 2002. This method recognizes the significance of each individual gene, based on the p-value of the corresponding gene, in each individual microarray data:

$$s_p = -\sum_{i=1}^{n} 2\log(p_i)$$

where p_i, i=1~n, is the *p*-value of a gene in the i-th independent study. This method has at least two significant limitations in its application to microarray data: (1) many microarray studies contain a small number of samples, for which the *p*-value can therefore be problematic, and (2) the p-values of a gene across different studies may have large variation. Thus, the smallest p-value may determine the outcome of S_p (effective significance of p-value).

A new approach is proposed here to overcome these limitations based on measuring the genome-wide relative significance (GWRS) and genome-wide global significance (GWGS) of genes (Liu et al. 2013). The GWRS measures the significance of a gene based on its ranking position on a genome-wide scale (*r* value) (Jurman et al, 2008). The ranking position of genes can be determined based on a differential expression measure, such as fold change, t-test p-value, SAM (Significance Analysis of Microarray data) p-value etc. Most meta-analysis methods in the literature focus on the top-k (k is the number of genes) genes (e.g. Jurman et al, 2008). In contrast to these, the method proposed here instead counts the ranking at the genome-wide scale. Compared to the model

of Rhodes et al this new approach has two important enhancements: (1) it can apply multiple different methods for measuring the degree of differential expression of a gene (e.g. fold change, t-test, Anova or SAM p-values) and (2) it uses a ranking *r* value instead of the test statistic (i.e., fold change, or p-value) to avoid the influence of a high variation of test statistics. The details are shown in the following 2.3.3 section.

2.3.3. Measuring the GWRS of genes in each single microarray database

For each gene in the gene list (G), the degree of differential expression can be measured by the commonly used methods such as fold-change, t-test (p-value), ANOVA, SAM or other statistical test. These four methods were used for this study to test the suitability of method to be used in the GWRS. However, the numbers of samples in individual datasets of this study are quite different: some datasets contained very small number of samples, for which the p-value based method may not suitable. For this reason, the fold-change is chosen and used in this study. The computational evaluation indicated that the use of fold-change produced more reliable results, likely due to the limited number of samples in some of the datasets. Each gene in the gene list G is assigned a rank number (in descending order starting from 1 to *m*) according to their corresponding degree of differential expression. Thus, a gene with a high degree of differential expression is ranked more highly and so on with a smaller ranking number. An m*n matrix (R) was created for m genes across n datasets, in which r_{ij} is the ranking number of the i-th gene in the j-th dataset. Thus, the GWRS of the i-th gene in the j-th dataset is measured by:

$$s_{ij} = -2\log\left(\frac{r_{ij}}{m}\right)$$

where r_{ij}, i=1~m, j=1~n, is the rank number of the i-th gene in the j-th study. The range of GWRS value (s_{ij}) is between 0 and -2log(1/m). For a gene with 'NA' value the s_{ij} is set to be 'NA'.

As an example, GSE3189 and GSE12391 were applied to these two different methodologies. The gene association with a small p value is considered as being an expressed gene. The r values of a gene in different studies are relevant to the p values. The smallest p value leads to the smallest r value, and the biggest sr value. Thus, genes with large sr values are considered to be significant genes in meta-analysis. The top 50 genes sorted by decreasing sr value have been displayed below.

Table 9: The Sr values of combined two microarray datasets

Sort	Gene	r(GSE3189)	r(GSE12391)	s_r
1	AHNAK	0.003596491	0.001008772	25.05364
2	ABLIM1	8.77E-05	0.043640351	24.94628
3	HLF	0.000614035	0.008508772	24.32423
4	ALDH2	0.000701754	0.012807018	23.23938
5	GLA	0.224298246	8.77E-05	21.67229
6	RPL30	0.079780702	0.000307018	21.23416
7	CYP4F12	0.002982456	0.009912281	20.85798
8	TMEM30B	0.003903509	0.008070175	20.73092
9	NFIB	0.000789474	0.040350877	20.70857
10	KIAA1305	0.000526316	0.063070175	20.62623

11	KRT15	0.000877193	0.041315789	20.45059
12	PHACTR1	0.029342105	0.001710526	19.79937
13	RPL10A	0.004342105	0.011842105	19.75098
14	MAOA	0.000175439	0.301578947	19.69389
15	PHYHIP	0.001403509	0.046578947	19.27077
16	ARF1	0.017763158	0.003903509	19.15302
17	RPL34	0.073070175	0.001008772	19.03071
18	PTPRF	0.004561404	0.01622807	19.02228
19	CHST6	0.020263158	0.004254386	18.71751
20	CTSB	0.001754386	0.049210526	18.71457
21	NACA	0.000394737	0.222105263	18.68379
22	LCP2	0.091929825	0.001008772	18.5715
23	GSTO1	0.02627193	0.003684211	18.48591
24	PGRMC2	0.002192982	0.044649123	18.46283
25	CYP4B1	0.046403509	0.002192982	18.38575
26	DNAJC15	0.101754386	0.001008772	18.36843
27	ARMC9	0.102807018	0.001008772	18.34785
28	LRRC1	0.002587719	0.043245614	18.19568
29	NUDT11	0.112982456	0.001008772	18.15909
30	CHST11	0.041798246	0.002894737	18.03952
31	TIMP2	0.018026316	0.007149123	17.91338
32	IFITM2	0.739649123	0.000175439	17.8996
33	ITM2B	0.055657895	0.002412281	17.83143
34	EPHX2	0.00122807	0.109473684	17.82876
35	GDF15	0.081973684	0.001710526	17.74462
36	PDZD2	0.021666667	0.006491228	17.73857
37	RBM35B	0.036929825	0.003903509	17.68923
38	HYOU1	0.002587719	0.059736842	17.54957

39	GATM	0.156491228	0.001008772	17.50755
40	IL11RA	0.000394737	0.403421053	17.49013
41	ACADL	0.004342105	0.036710526	17.48818
42	LDOC1	0.000263158	0.61745614	17.44981
43	NEBL	0.002017544	0.082982456	17.39
44	CTDSPL	0.009561404	0.017850877	17.35145
45	IFITM3	0.595526316	0.000307018	17.21383
46	MICAL1	0.092719298	0.002017544	17.16811
47	RPS6	0.002982456	0.066447368	17.05271
48	AP3D1	0.00754386	0.026666667	17.02272
49	MRCL3	0.212017544	0.001008772	16.90022
50	KLK1	0.001140351	0.191140351	16.86233

The sp value formula is referred to the meta-analysis method proposed by Rhodes et al (Rhoders et al, 2002), which applies the p-values to calculate sp value. The top 50 genes were selected by sp value and are shown on below table.

Table 10: The Sp values of combined two microarray datasets

Sort	Gene	P (GSE3189)	P (GSE12391)	s_p
1	ABLIM1	1.35E-25	0.040286149	120.9525406
2	MAOA	4.73E-19	0.260988884	87.07693805
3	HLF	2.18E-17	0.010400721	85.8610037
4	LDOC1	6.36E-19	0.573431223	84.91041133
5	ALDH2	4.32E-17	0.016320781	83.59201451
6	KIAA1305	1.77E-17	0.057930257	82.84296498

7	NFIB	5.06E-17	0.038033905	81.58371476
8	NACA	1.02E-17	0.194366519	81.52430717
9	KRT15	5.77E-17	0.039293992	81.25591629
10	IL11RA	1.02E-17	0.355474641	80.31689064
11	PHYHIP	1.44E-16	0.043497529	79.22353904
12	EPHX2	1.11E-16	0.094551745	78.191219
13	CTSB	2.74E-16	0.046175334	77.81742617
14	KLK1	9.63E-17	0.166285548	77.34622431
15	AHNAK	4.67E-15	0.003824143	77.12807642
16	CYP4F12	2.63E-15	0.012688143	75.87775978
17	EMP2	1.90E-16	0.190351823	75.71677764
18	PGRMC2	9.65E-16	0.04141238	75.51715796
19	NET1	7.94E-17	0.621639439	75.09485668
20	TMEM30B	5.20E-15	0.010151958	74.96041291
21	NEBL	8.41E-16	0.073364552	74.64850883
22	PLOD3	7.99E-17	0.896441924	74.35015518
23	LRRC1	2.25E-15	0.040206057	73.8831676
24	PALMD	1.20E-16	0.754188741	73.88230511
25	EFS	9.43E-16	0.097825304	73.84407479
26	KLHDC2	1.57E-16	0.615120104	73.75244722
27	RPL10A	7.30E-15	0.015301682	73.46138914
28	HYOU1	2.25E-15	0.055514925	73.23789911
29	NDRG2	6.78E-16	0.188237382	73.19487166
30	AZGP1	1.57E-16	0.834299864	73.14289652
31	PTPRF	7.71E-15	0.017745281	73.05578773
32	IL22RA1	1.69E-15	0.093503497	72.76760862
33	RPS6	2.63E-15	0.060770378	72.74489072
34	ACADL	7.30E-15	0.034017829	71.86354511

35	TMEM16A	2.87E-15	0.09115912	71.75922619
36	ACOT7	5.41E-15	0.051465837	71.63472871
37	GLRX2	3.76E-15	0.076989495	71.55688747
38	SATB1	2.63E-15	0.122586759	71.34145762
39	RNF43	6.78E-16	0.523433661	71.14945873
40	MAPKBP1	5.20E-15	0.10459419	70.29557009
41	EPS8L2	2.25E-15	0.242105326	70.29245719
42	ATP5C1	2.25E-15	0.247238885	70.25049289
43	CNIH3	1.16E-14	0.058623632	69.84877737
44	PSD3	2.35E-15	0.307256738	69.72886534
45	KRT1	1.10E-15	0.884100805	69.13330081
46	KCNK7	4.67E-15	0.23615853	68.88173857
47	PKM2	1.67E-14	0.081697968	68.45618765
48	ABCA5	1.84E-14	0.077299144	68.37299625
49	RP6-213H19.1	2.20E-14	0.067266132	68.2936647
50	GATA3	4.16E-15	0.40217291	68.04826896

The reason for using the sr value, instead of the sp value, is because the sr value is relevant to the p-value but not dominated by the smaller p-value. The sp values, unlike sr values, always incline to the smallest one of p-value groups. As shown in above sp value table (Table 10), the order of sp values is largely controlled by the p-values of GSE3189, because the p-values of GSE3189 are much lower than the p-values of GSE12391. For example, for gene no.19 (NETI), the p-value in GSE12391 is 0.621639439. This is many fold greater than the p-values of its above or below genes (gene no.18 and gene no.20) in GSE12391. Comparing the p-value (7.94E-17) of NETI in GSE3189 with its above and below genes (sr value table), there are not much difference. This

imbalance is caused by the huge gap of the order of magnitudes existing in the two datasets. The use of the sr value, however, is able to copy with this issue. As shown in the sr value table (Table 9), no such difference exists in the sr value formula by rank number of p-values.

2.3.4. Measuring the GWGS of a gene across multiple microarray datasets

We estimated the GWGS (genome-wide global significance s_i^r) of a gene based on its GWRS (genome-wide relative significance) across *n* datasets, by

$$s_i^r = \sum_{j=1}^{n} \omega_j s_{ij}$$

where ω_j represents the relative importance/weight of the j-th dataset, and $\sum_{j=1}^{n} \omega_j = 1$. The value of weight ($\omega_j$) can normally be assigned based on the data quality of the j-th datasets. However, it is important to note that the value of ω_j can also be used to reflect the differential importance of biopsy-derived versus cell line-derived samples, which biological scientists may wish to take into account. In this study, all the datasets are treated as equally important, thus the weight of each datasets is set equally to be 1/n for j=1~n. The top 200 genes were selected from the full gene list G for further analysis (i.e. selected genes with the greatest s^r value). To determinate the number of top genes, an empirical evaluation of the classification performance (accuracy ratio) was performed. This was performed using the 'wrapper-feature selection' after multiple rounds of gene addition to the classification performance model

(ranging from 20 genes up to 500 genes). This was done in order to achieve the maximum accuracy in distinguishing melanoma from normal skin/benign nevus. Using more than 200 genes did not improve further this classification. Thus, 200 genes have been considered as the optimal gene set i.e., the smallest number of genes that can achieve the highest level of classification performance.

2.3.5. Pathway analysis: functional relevance of 200 gene set

A pathway analysis was performed to assess the functional relevance of the new 200 gene signature based on the DAVID database (Hosack et al. 2003). DAVID provides a useful tool to analyze large gene lists, including via gene ontology and pathway analysis. The top 200 genes were applied to the DAVID database in order to detect potentially over-represented KEGG pathways. Before inputting into the DAVID database, the corresponding probe-sets of the 200 genes were extracted from each corresponding dataset. In order to compare the 200 genes to the original gene signatures of 16 studies, the probe-sets of these original gene signatures were also extracted. Thirty-one KEGG pathways were retrieved. Twelve genes (i.e., *EGFR, FGFR2, FGFR3, IL8, PTPRF, TNC, CXCL13, COL11A1, CHP2, SHC4, PPP2R2C,* and *WNT4*) from this 200-gene signature were identified and were found to closely interact with 4 melanoma driver genes (see Results section).

2.3.6. Immunocytochemistry (ICC)

A primary epidermal melanocyte (EM) culture (donor-Female 44y), moderately pigmented human melanoma cells (FM55), and highly pigmented human melanoma cells (FM94) (melanoma cells were a gift of Dr Janis Ancans, University of Latvia) were cultured as previously described (Gledhill, 2010). The cells were fixed in ice-cold methanol (Sigma, Poole, Dorset, UK) for 10 min before air drying and rehydration in PBS. The cells were blocked with 10% donkey serum (DS) for 1 h, washed with PBS before incubation with their respective primary antibodies to four test antigens taken from this 12-gene signature. These antibodies included: COL11A1 (Abcam, ab64883), CXCL13 (R&D Systems, AF801), PTPRF (NeuroMab, 75-193), SHC4 (Proteintech, 12641-1-AP), and were incubated overnight at 4 ^{o}C followed by secondary antibody (1:300) for 1h (donkey anti-goat (Invitrogen, A11055), donkey anti-mouse (Invitrogen, A21202), donkey anti-rabbit (Invitrogen, A21206), Alexa green). The optimal dilutions of antibodies were selected after titration (i.e. dilution tests). Following manufacturers' data sheets, the highest and lowest dilutions of these antibodies were tested respectively. According the results (i.e. most specific signal with least background) the optimal dilutions were applied.

The slides were cover-slipped by VECTASHIELD mounting medium with DAPI and viewed using a Nikon Eclipse 80i fluorescence microscope and photographed with a Nikon Digital Sight DS-U1 camera. A full assessment of all 12 proteins in our melanoma signature was beyond the scope of the current study, but will be assessed in detail in a follow-up studies.

2.3.7. Double immunohistochemistry (IHC)

Paraffin-embedded primary melanoma *in situ* (nose) and metastatic melanoma (lower leg), obtained from the Bradford university ethical tissue bank (ethical tissue.org), deparaffinized and boiled 10 mins in sodium citrate buffer (10 mM, 0.05% Tween 20, pH 6.0) for antigen retrieval. Acetone-fixed cryosections of normal human facial skin (donar-Female 52yrs) were used as control samples. All tissues were blocked with 10% donkey serum (DS) for 1h, washed with PBS before 2h incubation with NKi/beteb antibody raised against the melanocyte lineage-specific marker gp100 as a positive pigment cell control (Monosan; Mon7006-1) (1:15) which were used for many laboratory (including ours) and were validated that it is effective for gp100 antigen, followed by analysis by each of the 4 test antibodies as previously described (Gledhill, 2010).

2.4. A 'Stratification-based' Enhanced Method for Microarray Data Integration

Two important factors, 'stratification' and 'weighting', were considered in order to improve the previous method using in this study. 'Stratification' is the process to stratify individual datasets based on defined conditions (like platform, sample type, number of genes…), and then processed by GWGS and GWRS. 'Weighting' is used to measure the quality of each individual dataset (see section 2.4.2). In the previously described method each microarray data was treated equally. This may not always reflect the need of analysis. For example, it may be useful to take into account differences in microarray datasets (e.g. different platforms, different types of samples). This diversity of dataset should be analysed by using appropriate values of ω in the GWGS model.

2.4.1. The enhanced integrated analysis approach for microarray datasets

The enhanced analysis method developed in this study is shown in Figure 8, and involves 4 different methods (i.e. non-stratification, non-stratification with weighting, stratification, and stratification with weighting).

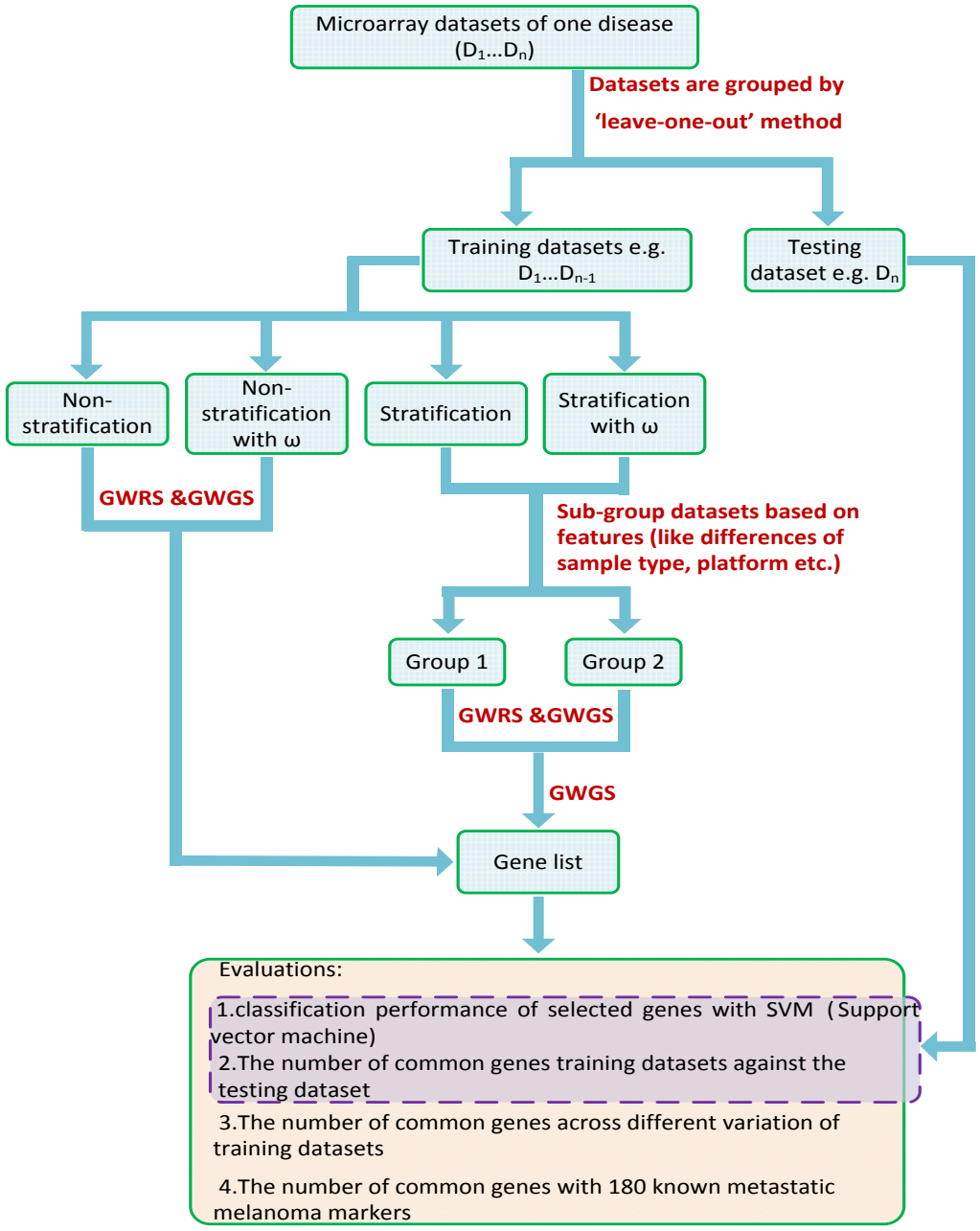


Figure 8: Procedures of stratification /non-stratification analysis

The entire process (figure 8) can be described as follows:

Assume there are n microarray datasets concerning one specific disease, and the datasets are named $D_1 \ldots D_n$. For each dataset the ranking value of each

gene in one dataset is calculated by GWRS, and the data quality is measured based on Hoeffding's D-statistic (ω) (Kauffmann *et al*, 2009). The top ranked genes for $D_1 \ldots D_n$ are denoted by $L_1 \ldots L_n$.

Step 1: Apply 'Leave-one-out' to divide the data into two groups: n-1 datasets for 'training' and the remaining one as 'testing' data. For example $D_1 \ldots D_{n-1}$ for training while D_n for testing.

Step 2: Apply the four different methods on the training dataset ($D_1 \ldots D_{n-1}$) ('non-stratification', 'non-stratification with ω', 'stratification', and 'stratification with ω') to calculate the significance of expression of a gene across multiple microarray studies respectively.

Step 3: For the two 'stratification' methods, the datasets are sub-grouped based on the features of datasets (e.g. different platforms, different type of samples etc.). In this study, the different sample types are cell line and biopsy.

Step 4: Apply GWRS and GWGS to calculate the significance of expression of genes across multiple microarray studies.

Step 5: Evaluate the methods on testing data (D_n), across the multiple studies and the 180 metastatic melanoma biomarkers (table s3 in appendix) were picked up from PubMed (based on conditions as follows: metastatic, melanoma, biomarkers) published to March of 2013.

2.4.2. Determination of the weights (ω_j) of microarray datasets

The quality of microarray data can be directly influenced by the experimental set-up of obtaining microarray, which brings inherent quality instability to the next level of analysis of the experimental procedure. The assessment of microarray dataset quality becomes a very important factor in the analysis of microarray datasets and particularly for the combination of multiple microarray datasets (Schuchhardt *et al*, 2000). In this study, an R software package called 'arrayQualityMetrics' was applied to assess the quality of microarray data for all arrays and platforms (Kauffmann *et al*, 2009). The assessment is based on the MA plot [present an M (log rations) and A (mean) average scale for visual representation of two channel DNA microarray gene expressions] with Hoeffding's D-statistic. The formula for Hoeffding's D is:

$$\mathrm{D} = 30\frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

$$D_1 = \sum_i (Q_i - 1)(Q_i - 2)$$

$$D_2 = \sum_i (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$$

$$D_3 = \sum_i (R_i - 2)(S_i - 2)(Q_i - 1)$$

where R_i and S_i are the rank of X and Y, which are the two sets of values. Q_i is the bivariate rank as 1 plus the number of points with all arrays less than the *i*th point (SAS institute, 2012). The D values of arrays in every individual dataset were assessed. The mean D values is used for describing the quality of each dataset. Based on the mean D values of a dataset, in this improved method step, the ω of GWGS formula (section 2.3.4) was instead by the ratio of D

values which is calculated as relative weighting. The sum of relative weighting (ω) is 1 as the above formula. For example, there are n datasets, thus for dataset 1 the corresponding D value is D_1 and the ratio of dataset 1 is $D_1 / (D_1 + D_2 + D_3 \ldots D_n)$.

2.4.3. Validation and selection of top genes by 'wrapper feature selection' method and machine learning for classification

Machine learning and data-mining techniques have been successfully applied in various biomedical domains. In this study, the 'wrapper-based feature selection method' was used (Peng *et al.* 2010), which applies SVM (Support vector machine) and selects significant genes that achieve the best classification performance (Figure 9). SVM is a machine learning approach based on Kernel techniques for classification or regression (Ivanciuc, 2007). The advantage of SVM is its capability to deal with high-dimensionality (larger than classical multivariate) of 'training' data (Peng, 2006), and so has been widely applied in microarray classification. It can produce the largest separation between the decision function values for instances that are located at the borderline between classes. The advantage of wrapper approaches with SVM is that they provide a high probability of identifying the feature subset (situation), and so producing better classification performances as they take into account the feature dependencies and their collective contribution to model generation (Peng, 2006).

In this study, the R package for SVM “e1071” was employed. Extract from the ranked gene lists produced by the Stratification or Non-Stratification procedure is the ‘feature’ for training SVM. The SVM was then applied on the testing dataset for the feature selection using the wrapper function, which determines which genes achieve the best predictive power in classifying the condition of the samples.

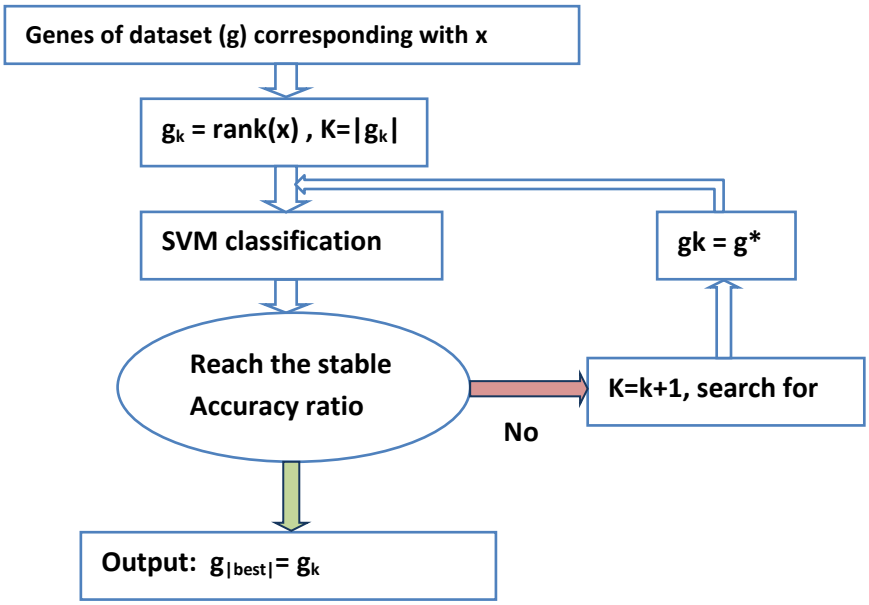


Figure 9: The wrapper feature selection method based on SVM that was used for robust signature attainment between control and disease states

2.5. Access to the associated microarray data

The microarray data used in the study were retrieved from Gene Expression Omnibus (GEO) with the access numbers shown on table 6. The 16 gene signatures of melanoma reported in the literature between 2000 and 2011 were extracted from the associated publications and are presented in SI – Table s, data as excel file stored in attached CD).

3. RESULTS

3.1. The discovery of new biomarkers of melanoma

3.1.1. A new melanoma gene signature revealed by an integrated analysis of cross-laboratory microarray datasets

A new approach has been applied here for the integrated analysis of five independent microarray studies (Hoek *et al.*, 2004, Smith *et al.*, 2005, Riker *et al.*, 2008, Scatolini *et al.*, 2010, Rose *et al.*, 2011) (see Methods) in order to determine gene associations in melanoma, rather than to look for associated gene mutations. The genome-wide 'global significance' of a gene (i.e., across multiple datasets) was measured by the GWGS (s^r) function (see Methods). A gene with a large s^r value is considered to be significant across multiple independent studies (i.e., globally significant). The 200 genes with the largest s^r values were selected as the starting point for the new proposed gene signature of melanoma, as listed in Table 11 and Table s2 of appendix. This set of 200 signature genes was empirically determined, based on classification accuracy ratios after various rounds of gene additions (using the 'wrapper feature selection' approach) in order to distinguish melanoma from normal skin cells and/or benign nevus. The classification accuracy ratio was improved very little by adding more than 200 genes, and so 200 genes achieved, computationally, the highest classification accuracy with the smallest number of genes.

Table 11: The 200 genes with largest s^r values selected as the proposed starting point for a gene signature of melanoma

NO.	Genes	NO.	Genes	NO.	Genes	NO.	Genes
1	DCD	51	GAGE7	101	DKFZP434B061	151	AQP3
2	MAGEA3	52	DGAT2	102	PPP1R14C	152	C1orf116
3	MAGEA2	53	FGFR3	103	AKR1C3	153	RGS4
4	MAGEA2B	54	MICALCL	104	C19orf33	154	GRHL3
5	CSAG3	55	KRT15	105	FGFR2	155	GPR115
6	CSAG2	56	CTAG2	106	IGL@	156	SERPINA3
7	GAGE12F	57	ANK3	107	SERPINB5	157	LAD1
8	GAGE12G	58	HMGA2	108	CYP3A5	158	FLJ37464
9	GAGE12I	59	MYOZ2	109	LEP	159	HLA-DRB4
10	GAGE2A	60	AADACL2	110	CHST6	160	TMEM79
11	GAGE2B	61	SCGB2A2	111	TF	161	ZNF750
12	GAGE2C	62	ISG20	112	MIA	162	IGHV4-31
13	GAGE2E	63	DST	113	HLA-DQB1	163	TP63
14	GAGE4	64	IL13RA2	114	GPR87	164	LOC124220
15	CTAG1B	65	APOC2	115	RHBDL2	165	RASGRF1
16	KRT77	66	TNC	116	SGPP2	166	KRT5
17	THRSP	67	FMN2	117	SCARA5	167	LAMB4
18	CTAG1A	68	SHC4	118	SAA1	168	SCML4
19	GAGE5	69	FSTL5	119	RNASE2	169	CYP4B1
20	GAGE6	70	PTPRF	120	SLAMF7	170	HLA-DRB3
21	MAGEA12	71	KRTAP19-1	121	SAA2	171	NEBL
22	MAGEA6	72	CXCL13	122	PPP2R2C	172	IGSF9
23	XAGE1A	73	GAGE1	123	GBP5	173	KLK11
24	XAGE1B	74	EYA1	124	AKR1C1	174	CHP2
25	XAGE1C	75	HLA-DRB2	125	ENTHD1	175	MAGEA10
26	XAGE1D	76	LOC100133484	126	EPHA3	176	CYP26B1
27	XAGE1E	77	LOC100133661	127	KRT6B	177	EREG
28	PRAME	78	LOC100133811	128	CCDC3	178	DLX1
29	C4orf7	79	LOC730415	129	BTBD16	179	LOC285986
30	GAGE12B	80	ZNF749	130	ANKRD35	180	TRIM7
31	GAGE12C	81	KRT14	131	HLA-DQA1	181	GAD1
32	GAGE12D	82	IGFL2	132	C10orf116	182	LOR
33	GAGE12E	83	SCEL	133	JUP	183	EXPH5
34	GAGE12H	84	GAGE3	134	IGFBP5	184	TMEM154
35	GAGE12J	85	GATA3	135	KRT25	185	LASS3
36	GAGE2D	86	DSP	136	SULF1	186	HLA-DRB5
37	GAGE8	87	WNT4	137	TKTL1	187	LOC100126583
38	WFDC5	88	TACSTD2	138	IL1F7	188	CYP4F8
39	IL8	89	CAPNS2	139	C6orf218	189	SDC1
40	COL17A1	90	MAL2	140	HEY1	190	SCGB1D2
41	FOXQ1	91	DGAT2L3	141	MGST1	191	RORA
42	ZIC1	92	PIP	142	ABCA13	192	SH3RF2
43	ELMOD1	93	AKR1C2	143	RAPGEFL1	193	LGALS7
44	ELOVL3	94	IGF2	144	TFPI2	194	MMP1
45	SERPINA12	95	MPP7	145	TRIM29	195	MAGEC1
46	DSC3	96	IGHG1	146	ALDH1A3	196	FRMD5
47	MAGEA1	97	NMU	147	ATP6V1C2	197	SERPINB7
48	DMKN	98	EGFR	148	COL11A1	198	FGF13
49	INS-IGF2	99	APOC4	149	RSPO1	199	LOC645323
50	C1orf172	100	MGP	150	PLA1A	200	COL9A3

3.1.2. Validation of a new 200-gene signature based on experimental studies reported in the literature

The 200 genes found to have genome-wide global significance (GWGS) in this study were compared with the gene signatures identified in previously-published reports (see appendix – Table s2).

The new 200-gene signature was first validated by (i) comparing it with 16 signatures in the referred set of microarray studies (Table 6), and (ii) by checking if any existing experimental validation for the gene associations in question has been published in the literature (PubMed, last access: 16 April 2013). This analysis revealed that (a) 85 genes in our 200-gene signature were reported in at least one of the 16 microarray studies, and (b) 21 genes of the 200-gene signature were reported in both microarray studies and biological experimental-based studies (Table 12, labeled yellow background). While 38 genes of this 200-gene signature were not reported in any of the 16 reference studies, they had in fact been previously validated in independent wet-lab studies (Table 12 and discussion section). These 38 genes, including EGFR and MIA, can be considered as “validated novel genes” as they were not identified in the previous microarray studies but validated by biological experiments.

On the other hand, the new gene signature reported an additional 77 genes that were not previously reported anywhere in the literature in association with melanoma (Figure 13). The ranking positions of these 77 genes show that 39%

of these appear in the top 100 and 34% in bottom 50 of the 200 gene set (table 11). These genes may represent 'novel genes' as they were not previously identified in any previous microarray studies.

We further investigated the characteristics of the 85 genes reported in at least 1 of the 16 reference microarray studies (see Appendix - Table s2). Forty-four were reported in ≥2 studies, while 17 genes have been reported in ≥3 of the 16 studies (Appendix - Table s2). *KRT15, MAGEA6, RORA* and *SULF1* were most frequently reported, appearing in 4 of the 16 studies. It is noted that, using the method proposed in this study, we identified 4 of the 7 most frequently reported genes in the 16 studies. This suggests that the methodology used to select the top 200 genes provides a more powerful signature than these previously reported 16 published signatures (Table 6).

Table 12: Novel genes identified by the proposed method were validated by independent wet-lab experimental studies

gene symbol	studies					
ALDH1A3	Luo et al. 2012. *Stem Cells* 30:2100-2113					
COL1A1	Lin et al. 2005. *Arch Pathol Lab Med* 129:884-92					
COL17A1	Krenacs et al. 2012. *Histochem cell Biol* . 138:653-667					
CSAG2	Feller et al. 2000. *Anticancer Res* 20:4147-4151	Janjic et al. 2006. *J Immunol* 177:2717-2727				
CTAG2	Lethe et al. 1998. *Int J Cancer* 76:903-908	Usener et al. 2003. *Br J Dermatol* 149:282-288	Svobodova et al. 2011. *Eur J Cancer* 47:460-469			
CYP3A5	Zhang et al. 2006. *J Pathol* 209:213-219					
DCD	Rieg et al. 2004. *Br J Dermatol* 151:534-539					
EGFR	Huang et al. 1996. *Anticancer Res* 16:3557-3563	Scholes et al. 2001. *Arch Ophthalmol* 119:373-377	Mallikarjuna et al. 2007. *Curr Eye Res* 32:281-290	Ueno et al. 2008. *Int J Cancer* 123:340-347	Boone & Brochez. 2009.*Verh K Acad Geneeskd Belg* 71:251-294	Diaz et al. 2009. *Front Biosci* 14:159-166
	Topcu-Yilmaz. 2010. *Melanoma Res* 20:126-132	Boone et al. 2011. *J Cutan Pathol.* 38:492-502				
EPHA3	Balakrishnan et al. 2007. *Cancer Res* 67:3545-3550	Mosch et al. 2012. *Cell Adh Migr* 6:113-125	Lisabethe et al. 2012. *Biochemistry* 51:1464-1475			
FCRLA	Inozume et al. 2007. *J Invest Dermatol* 127: 2818-2822					
FGFR2	Gartside et al. 2009. *Mol Cancer Res* 7:41-54					
FGFR3	Cheng et al. 2006. *Bio Pharm Bull* 29:655-669	Yadav et al. 2012. *J Biol Chem* 287:28087-98				
GAGE family	Bazhin et al. 2007. *Cancer Lett* 251:258-267					
GBP5	Fellenberg et al. 2004. *J Invest Dermatol* 122:1510-1517					
HLA-DQB1	Lee et al. 1994. *Int J Cancer* 59:510-513	Lee et al. 1996. *Cancer* 78:758-763	Kageshita et al. 1997. *Tissue Antigens* 49:466-470	Bateman et al. 1998. *Tissue Antigens* 52:67-73		
HMGA2	Murakami et al. 2010. *Mol Biol Rep* 37:1279-86.					
IGF2/INS-IGF2	Soares et al. 2010. *Growth Horm IGF Res* 20 :295-297					
IL8	Zhang et al. 2011. *Int J Mol Sci* 12:1505-1518					
LEP	Brandon et al. 2009. *Cancer Biol Ther* 8:1871-1879	Ellerhorst et al. 2010. *Oncol Rep* 23:901-907	Amjadi et al. 2011. *J Exp Clin Cancer Res* 30:21.	Lawrence et al. 2012. *Neurol Res Int* 2012:870807: Epub 2012 Jan 4		
LUM	Vuillermoz et al. 2004. *Exp Cell Res* 296:294-306	Sifaki et al. 2006. *IUBMB Life* 58:606-610	Radwanska et al. 2008. *Life Sci* 83:651-660.	Brezillon et al. 2009. *J Physiol Pharmacol* 60:15-22		
MAGEA1	Brasseur et al. 1995. *Int J Cancer* 63:375-380	Mulcahy et al. 1996. *Int J Cancer* 66:738-742	Chen et al. 1997. *J Immunother* 20:265-275	Luyten et al. 1998. *Melanoma Res* 8:11-16	Caballero et al. 2010. *PLoS One* 5:e12773	Errington et al. 2012. *Br J Ophthalmol* 96:451-458
MAGEA2	Brasseur et al. 1995. *Int J Cancer* 63:375-380	Chen et al. 1997. *J Immunother* 20:265-275	Luyten et al. 1998. *Melanoma Res* 8:11-16	Caballero et al. 2010. *Cancer Immun* 8:2		
MAGEA3	Brasseur et al. 1995. *Int J Cancer* 63:375-380	Chen et al. 1997. *J Immunother* 20:265-275	Luyten et al. 1998. *Melanoma Res* 8:11-16	Vourc'h-Jourdain et al. 2009. *Arch Dermatol Res* 301:673-679	Errington et al. 2012. *Br J Ophthalmol* 96:451-458	
MAGEA6	Gibbs et al. 2000. *Melanoma Res* 10:259-264	Errington et al. 2012. *Br J Ophthalmol* 96:451-458				
MAGEA12	Gibbs et al. 2000. *Melanoma Res* 10:259-264					
MIA	Bosserhoff et al. 1997. *Cancer Res* 57:3149-3153	Bosserhoff et al. 1998. *Hautarzt* 49:762-769	Bosserhoff et al. 1999. *Anticancer Res* 19:2691-2693	Perez et al. 2000. *Hum Pathol* 31:1381-1388	Guba et al. 2000. *Br J Cancer* 83:1216-1222	Schmitz et al. 2000. *Anticancer Res* 20:5059-5063
	Stahlecker et al. 2000. *Anticancer Res* 60:5041-5044	Bosserhoff et al. 2001. *Recent Results Cancer Res* 158:158-168	Bosserhoff et al. 2001. *Melanoma Res* 11:417-421.	Meral et al. 2001. *Melanoma Res* 11:627-632	Juergensen et al. 2001. *Tumour Biol* 22:54-58	Hochberg et al. 2002. *Br J Dermatol* 146:244-249.
	Schaller et al. 2002. *Melanoma Res* 12:593-599	Garbe et al. 2003. *Cancer* 97:1737-1745	Faries et al. 2004. *Ann Surg Oncol* 11:85-93	Tas et al. 2004. *Am J Clin Oncol* 27:225-228	Reiniger et al. 2005. *Graefes Arch Clin Exp Ophthalmol* 243:1161-1166	Cao et al. 2007. *Anticancer Res* 27:595-599
	Lugovic et al. 2007. *Coll Antropol* 31 Suppl 1:7-11	Barak et al. 2007. *Anticancer Res* 27:1897-1900	Faries et al. 2007. *Cancer Invest* 25:285-293	Vucetic et al. 2008. *Melanoma Res* 18:201-207	Dumitrascu et al. 2009. *Roum Arch Microbiol Immunol* 68:125-135.	Kluger et al. 2011. *Clin Cancer Res* 17:2417-2425
	Klingenstein et al. 2011. *Melanoma Res* 21:352-356	Hofmann et al. 2011. *J Dermatol* 38:880-886	Diaz-Lagares et al. 2011. *Tumour Biol* 32:1155-1161	Essler et al. 2011. *PLoS One* 6:e24632	Schmidt et al. 2012. *PLoS One* 7:e37941	
PRAME	Soikkeli et al. 2007. *J Pathol* 213:180-189	Westekemper et al. 2010. *Br J Ophthalmol* 94:1322-1327				
SERPINA3	Dimberg et al. 2011. *Oncol Lett* 2:413-418					
SHC4	Fagiani et al. 2007. *Cancer Res* 67:3064-3073	Pasini et al. 2009. *Expert Opin Ther Targets* 13:93-104				
TF	Krzyminiewski et al. 2011. *Appl Magn Reson* 40:321-330					
TFPI2	Liu et al. 2008. *Pigment Cell Melanoma Res* 21:545-558	Tanemura et al. 2009. *Clin Cancer Res* 15:1801-1807				
TNC	Fukunaga-Kalabis et al. 2010. *Oncogene* 29:6115-6124	Grahovac et al. 2012. *J Invest Dermatol* doi: 10.1038/jid.2012.263				
XAGE1	Egland et al. 2002. *Mol Cancer Ther* 1:441-450	Zendman et al. 2002. *Int J Cancer* 97:195-204				

The correlations between melanoma and the genes shown in table 12 are provided below:

ALDH1A3

Luo et al 2012 detected the relationship between ALDH1A3 and melanoma cells by Aldefluor Assay and FACS, mRNA copy number analysis and microarray analysis, and found that ALDH1A3 isozyme is a key molecule to regulate the function of melanoma cells.

COL1A1

Lin et al 2005 determined that *COL1A1* was significantly increased in expression in melanoma compared with poorly invasive non-pattern–forming cells though real time PCR, and suggested *COL1A1* may be synthesized by melanoma cells.

COL17A1

Krenacs et al 2012 tested expression of *COL17A1* in benign and malignant melanoma by endodomain- and ectodomain-selective antibodies, and found it may be a potential biomarker or target for melanoma.

CSAG2

The expression of *CSAG2* was compared in several melanoma cell lines by real time PCR in Feller et al study in 2000, and also in Janjic et al. study of 2006. Both of these two studies suggest CSAG2 as a novel antigen with application for melanoma vaccines.

CTAG2

The CTAG2 protein was determined to be significantly expressed in melanoma by three studies, Lethe et al 1998 suggested the expression may be induced by demethylation, Usener et al 2003 evaluated CTAG2 and suggested it could be a new antigen in immunogenic response, and Svobodova et al 2011 investigated the expression by IHC and showed that it could be as potential therapeutic target of melanoma.

CYP3A5

CYP3A5 protein is as metabolism enzyme expressed in melanoma tissues and was determined by tissue microarray to indicate which drug pathway is dependent in melanoma, and then selected f or its potential in cancer therapy (Zhang et al, 2006).

DCD

The mRNA expression of DCD was detected in melanoma cells, but not in other kinds of epidermal cells (like keratinocytes, fibroblasts, and melanocytes). Thus, Rieg et al 2004 suggested that DCD may play an important role in preventing local and systemic pathogens invasion.

EGFR

Some studies reported *EGFR* expression in melanoma at DNA, RNA and protein levels in melanoma. Huang et al. 1996 detected overexpression of EFGR in a melanoma cell line. Scholes et al. 2001 found that 23% of melanoma samples contained EGFR immunoreactivity, and ta imilar result was reported by Mallikarjuna et al. 2007, who found that EGFR was expressed in 30% of 60

uveal melanomas. Boone and Brochez 2009 found that the immunoreactivity of EFGR was more frequently present in melanoma patients with a positive sentinel lymph node. Next, EGFR gene and protein expression were established by Fluorescence in situ hybridizaiton (FISH) and immunohistochemistry (IHC) by Boone et al 2011 in a further study with positive sentinel lymph node patients. Targeting EFGR is as a therapeutic option in melanoma. Diaz et al 2009 suggested that the immune system in melanoma patients could downregulate the increased expression level of *EFGR*. Ueno et al 2008 inhibited the tyrosine kinase activity of EGFR using PD153035 to show the promoter role of EGFR in tumour growth and metastasis melanoma. Topcu-Yilmaz et al. detected the expression of EGFR and suggested that it plays a role in uveal melanoma development and progression.

EPHA3

EPHA3 is highly mutated in melanoma as reported by Balakrishnan et al 2007, and Lisabeth et al 2012. Both studies suggested that the tyrosine kinase *EPHA3* is a tumor suppressor gene and that the mutation of this gene could be caused by somatic cancer predisposition. The up-regulated of EPHA3 in X-ray irradiation-induced metastatic property of melanoma was detected by Mosch et al 2012.

FCRLA

FCRLA was reported by Inozume et al 2007as an antigen that was specifically expressed in melanoma cells, and also was recognized by IgG antibodies from melanoma.

FGFR2

FGFR2 gene was reported to be mutated in 10% melanoma tumours and cell lines with loss of function in the receptor. Gartside et al 2009 suggested that *FGFR2* may play context-dependent opposing roles in melanoma.

FGFR3

FGFR3 was detected to be down-regulated by DNA microarray analysis by Cheng et al 2006, to suppress the biological processes of cell cycle in melanoma. The opposite result was detected by Yadav et al 2012, they found that *FGFR3* was up-regulated by microarray, and this enhanced activation corresponding to Ras and MAPK activation.

GAGE family

GAGE family proteins were up-regulated in melanoma tissues by immunohistochemical analysis, and Bazhin et al 2007 first demonstrated that GAGE family proteins might be immunotherapy targets of melanoma.

GBP5

The expression of GBP5 protein was detected by western blotting in melanoma cell lines, and Fellenberg et al. suggested that GBP5 plays an important role in proliferation and differentiation which may relate to caner functions.

HLA-DQB1

Bateman et al 1998 reported that *HLA-DQB1* expression is significantly increased in malignant melanoma in the UK population, when compared to control samples by PCR, and suggested *HLA-DQB1* plays an important role in

the risk of development and prognosis of malignant melanoma in UK population. The similar results were detected by three other studies: the increased *HLA-DQB1* was detected by PCR-RFLP when comparing with control samples in Kageshita et al study 1997, Lee et al 1994 and Lee et al 1996 determined *HLA-DQB1* was highly expressed in melanoma samples, and found this gene may influence the progression or incidence of melanoma.

HMGA2

HMGA2 was reported to be related to TGF-beta/activin, and it also was down-regulated E-cadherin by up-regulated zinc-finger transcription factors and which showed *HMGA2* may be susceptible to metastasis melanoma by Murakami et al 2010.

IGF2

Soares et al 2010 investigated *IGF2* in a 100 healthy individuals and in 21 patients with hereditary melanoma by Chi-square, Fisher's exact tests and RegRNA software, and found *IGF2* was associated with progression of melanoma.

IL8

IL8 level was significantly increased by immunological tests with a special buffer which can eliminate the interference of serum, Zhang et al 2011 suggested IL8 is as a prognosis serum marker for melanoma mortality.

LEP

High protein expression of LEP was detected in melanoma blood samples by Brandon et al 2009 and Amjadi et al 2011, and they suggested that melanoma tumour growth rate is increased by LEP. A similar increased expression level of LEP also was determined by Ellerhorst et al 2010 and Lawrence et al 2012, by western blotting, indirect immunofluorescence and real time PCR. They concluded LEP is a growth factor to contribute uncontrolled melanoma cells proliferation.

LUM

The mRNA and protein levels of *LUM* were investigated by many ways (e.g. Immunocytochemistry, confocal microscopy, western immunoblotting, Real-time Reverse Transcription-PCR) in four studies (Sifaki et al 2006, Vuillermoz et al 2004, Radwanska et al 2008, and Brezillon et al 2009.) They detected that the mRNA level of *LUM* was overexpressed in melanoma cells, and that the LUM protein was expressed in melanoma cell lines but not be expressed in normal melanocytes. They suggested that LUM may induce apoptosis of melanoma cells to control of melanoma growth and invasion.

MAGEA-1/ -2/ -3/ -6/ -12

The expression of *MAGEA1, MAGEA2* and *MAGEA3* were analysed by RT-PCR in three types of melanomas. Brasseur et al 1995 reported that the expression of these three genes was greater in thick cutaneous melanoma, Chen et al 1997 examined these increased expressions in cultured ocular melanoma cells, and Luyten et al 1998 evaluated these expression in uveal melanoma cell lines. Three studies all found that the expression of *MAGEA-1,-2*

and -*3* were significantly increased. Other studies also reported similar results of gene expressions, which were detected by RT-PCR: The expression of *MAGEA1* in metastatic uveal melanoma was determined in Mulcahy et al study 1996, Vourc'h-Jourdain et al 2009 determined that *MAGEA3* was significantly expressed in melanoma, and Gibbs et al 2000 detected the significant expressions of *MAGEA6* and *MAGEA12* in 47 melanoma biopsies and 11 melanoma cell lines. Except these mRNA level expressions, Errington et al 2012 assessed that protein expressions of MAGEA-1, -3, and -6 by CT antigens in ocular melanoma cells. The mutations of *MAGEA1* and *MAGEA2* genes were determined in 111 fresh melanoma samples by Caballero et al 2010.

MIA

The mRNA level of *MIA* was indicated by **RT-PCR** in four studies: Bosserhoff et al 2001, Guba et al 2000 and Garbe et al 2003 these three studies determined the higher expression levels in melanoma stage III and IV (metastatic melanoma) than control samples or stage I and II melanoma, and suggested that *MIA* plays an important role in metastatic melanoma. Hochberg et al 2002 ompared mRNA expressions of *MIA* between sentinel lymph nodes of melanoma and control nodes, and found that higher level is in sentinel lymph nodes melanoma, so they suggested *MIA* could help detect metastatic melanoma cells in sentinel lymph nodes to survival patients.

The serum levels of MIA protein were quantitated by **ELISA** in quite a few studies: Bosserhoff and his colleagues published four studies about MIA serum level in 1997, 1998, 1999, and 2001, they detected the higher MIA serum levels were existed in stage I or II of melanoma in 32 patients and 350 patients, and

stage II to IV of melanoma in 84 patients, they suggested MIA is a novel serum marker for malignant melanoma and used to detect metastatic melanoma disease. Then they analysed the relationship between enhanced MIA expression and progression of melanoma, suggested that MIA play a role in promoting melanoma metastasis. Similar results and suggestions were generated by Schmitz et al 2000, Juergensen et al 2001, Tas et al 2004, Lugovic et al 2007, and Stahlecker et al 2000, they detected the MIA serum levels in 87, 50, 48, 50 and 326 melanoma patients respectively, and found that the increased MIA serum levels in stage III/IV melanoma comparing to stage I/II melanoma, and suggested MIA is specificity and sensitivity to detect progression of metastatic melanoma. Three studies of uveal melanoma were reported by Schaller et al 2002, Barak et al 2007, and Reiniger et al 2005, they applied the similar experiments to detect 139, 18 and 285 patients' serum samples of uveal melanoma, and found the significantly higher level of MIA in metastatic uveal melanoma, and suggested MIA is as a serum marker for monitoring uveal melanoma for metastasis. The MIA serum levels were detected in melanoma with positive and negative sentinel lymph nodes by Vucetic et al 2008 and Hofmann et al 2011, Both teams found that there are much higher mean MIA values in positive lymph nodes melanoma than other groups with negative lymph nodes. Thus, they suggested that MIA serum level can be helpful in screening the tumour spread to sentinel lymph nodes. The MIA serum levels before or after therapy (e.g., surgical excision, radiotherapy, immunotherapy or chemotherapy) also are analysed by three studies (Meral et al 2001, Faries et al 2004, and Cao et al 2007), they all investigated that MIA levels significantly increased when melanoma progressed, and the MIA levels

significantly decreased when patients responded to systemic treatment. Thus, MIA levels play an important role in prognosis marker for survival. Comparisons between MIA and S100B these two serum markers have been done by Faries et al 2007, Diaz-Lagares et al 2011, Dumitrascu et al 2009, and Essler et al 2011 in 75, 110, 123, and 125 patients and health samples, they found that MIA and S100B both significantly increased in progressions, and suggested combined MIA and S100B serum levels together showed a better prognostic value to indicate the evolution of melanoma.

The protein levels of MIA in plasma were detected by ELISA in Kluger et al 2011 study, they compared the MIA levels of 108 metastatic melanoma patients with 108 patients with melanoma stage I or II, and found the MIA levels are much higher in metastatic melanoma samples. They suggested the levels of MIA in plasma can be applied for monitoring disease recurrence.

Except applied RT-PCR and ELISA respectively, some other techniques were applied to detect MIA. Receiver operating characteristic (ROC) analysis with ELISA was applied for analysis of MIA serum levels, and showed that MIA levels significantly higher in metastatic melanoma patients than other stages of melanoma patients by Klingenstein et al 2011. RT-PCR and immunohistochemistry were used to assay 23 melanoma and 25 nonmelanoma specimens by Perez et al 2000, MIA was over-expressed in most melanoma specimens. Schmidt et al discovered that the relationship between dimerization of MIA and functional activity, and also suggested that dodecapeptide AR71 as a MIA inhibitor to prevent MIA dimerization in melanoma therapy.

PRAME

Significant expression of PRAME protein was detected in melanoma samples by immunohistochemistry in Westekemper et al 2010 study, they applied receiver operating characteristics (ROC) analysis to assess the value of PRAME as diagnostic markers, and suggested that PRAME is marker to discriminate cutaneous melanoma and conjunctival nevi. Asimilar result also was detected by Soikkeli et al 2007, they combined immunohistochemistry and RT-PCR to test PRAME in 160 melanoma patients, and determined that PRAME was a melanoma marker to distinguish melanoma from benign nevi in the sentinel lymph nodes.

SERPINA3

SERPINA3 was determined to be significantly down-regulated in malignant melanoma by ELISA, Dimberg et al 2011 suggested that SERPINA3 could be a potential biomarker and the concentration of SERPINA3 was related to systemic inflammation.

SHC4

The expression of SHC4 was revealed at high protein in primary and metastatic melanomas, however, low levels of the protein were detected in normal melanocytes and benign nevi. Thus, Fagiani et al 2007 and Pasini et al 2009 indicated SHC4 is a specific biomarker for metastatic melanoma, and could be a potential drugable target for melanoma therapeutic strategies.

TF

The TF protein was investigated in blood samples of human melanoma patients by electron paramagnetic resonance (EPR) by Krzyminiewski et al 2011, and found that the TF saturation is correlated to the total iron ion complexes concentration.

TFR12

The increased hypermethylation of TFR12 in a melanoma cell line and in melanoma frozen samples was determined by quantitative methylation-specific PCR in Liu et al 2008, and Tanemura et al 2009 studies. Both suggested TFR12 might play an important role in malignant melanoma progression.

TNC

The protein level of TNC was strongly up-regulated in melanoma cells growth by detecting 3D spheres in Fukunaga-Kalabis et al 2010 study. They suggested that TNC may mediate protective signals in melanoma progression. Then, Grahovac and his colleagues 2012 found that up-regulated TNC is associated with EGFL domains of THC (TNCEGFL), which is expressed in melanoma cells. The authors suggested TNCEGFL may play a role in melanoma invasion through modulating ROCK signalling and cell migration.

XAGE1

XAGE1 is reported to overexpressed in metastatic melanoma by RT-PCR in Zendman et al 2002 study. The expressed protein level of XAGE1 was detected by cancer testis (CT) antigen in Egland et al 2002 study, they suggested that

XAGE1 may play a role as a potential target for melanoma and other cancers immunotherapies.

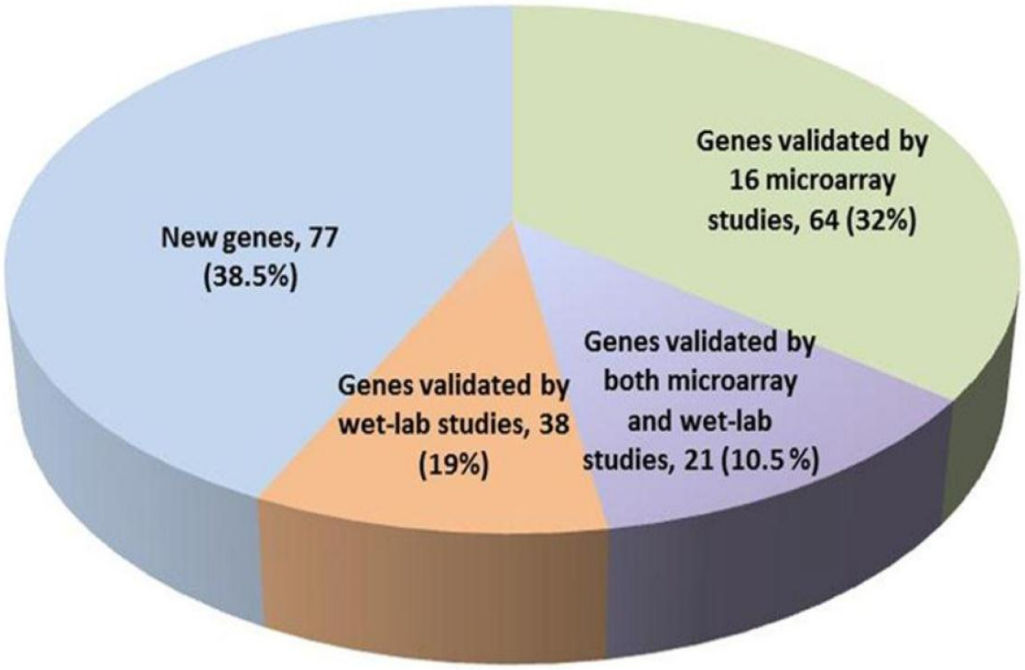


Figure 10: Validation of the proposed 200-gene signature.

The 200 signature gene set taken from the full list of genes associated with melanoma was selected for further analysis based on their classification accuracy ratio (i.e. genes with the greatest s^r value).

Table 13: The Sr values and ranking position of the 77 'new' genes not previously reported in the association with melanoma

	Genes	Sr	Ranking position		Genes	Sr	Ranking position		Genes	Sr	Ranking position
1	CTAG1B	12.10406784	5	27	DGAT2L3	8.633051674	91	53	RGS4	7.819213019	153
2	KRT77	11.99092009	6	28	MPP7	8.547483331	95	54	GRHL3	7.815176487	154
3	CTAG1A	11.56869048	8	29	IGHG1	8.542553019	96	55	GPR115	7.806607985	155
4	C4orf7	11.09096894	9	30	MGP	8.476842108	100	56	FLJ37464	7.792779686	158
5	ZIC1	10.2538162	42	31	DKFZP434B061	8.457911917	101	57	TMEM79	7.772719825	160
6	ELMOD1	10.22606747	43	32	PPP1R14C	8.449068608	102	58	ZNF750	7.772186022	161
7	SERPINA12	10.21661117	45	33	AKR1C3	8.412694702	103	59	IGHV4-31	7.759066298	162
8	DMKN	10.07185311	48	34	IGL@	8.363091942	106	60	TP63	7.75546205	163
9	C1orf172	9.988849797	50	35	SCARA5	8.198953928	117	61	LOC124220	7.751991913	164
10	DGAT2	9.823663103	52	36	SAA1	8.197432553	118	62	LAMB4	7.717425739	167
11	MICALCL	9.58335741	54	37	RNASE2	8.194555741	119	63	SCML4	7.698480948	168
12	AADACL2	9.297067696	60	38	SLAMF7	8.189374367	120	64	CYP4B1	7.689612344	169
13	IL13RA2	9.119495297	64	39	SAA2	8.183931738	121	65	HLA-DRB3	7.686568303	170
14	FMN2	8.956107891	67	40	PPP2R2C	8.164009742	122	66	CHP2	7.64842974	174
15	FSTL5	8.949959491	69	41	ENTHD1	8.129231606	125	67	MAGEA10	7.645898563	175
16	EYA1	8.80726093	74	42	BTBD16	8.076249966	129	68	LOC285986	7.616034013	179
17	ZNF749	8.80038067	75	43	ANKRD35	8.069919063	130	69	GAD1	7.582412726	181
18	LOC730415	8.80038067	75	44	HLA-DQA1	8.035157176	131	70	EXPH5	7.570222671	183
19	LOC100133811	8.80038067	75	45	C10orf116	8.025783514	132	71	TMEM154	7.567552064	184
20	LOC100133661	8.80038067	75	46	TKTL1	7.953701277	137	72	HLA-DRB5	7.555909616	186
21	LOC100133484	8.80038067	75	47	C6orf218	7.943240696	139	73	LOC100126583	7.54175814	187
22	HLA-DRB2	8.80038067	75	48	ABCA13	7.895932655	142	74	SH3RF2	7.491162328	192
23	IGFL2	8.772810711	82	49	ATP6V1C2	7.881948206	145	75	MMP1	7.482838435	194
24	WNT4	8.699939884	87	50	COL11A1	7.853686313	148	76	FRMD5	7.457195568	196
25	CAPNS2	8.689696696	89	51	RSPO1	7.836114165	149	77	LOC645323	7.427783892	199
26	MAL2	8.640387131	90	52	C1orf116	7.823842579	152				

3.1.3. Pathway analysis of the new melanoma gene signature

To validate whether the new melanoma gene signature provides a more robust association with melanoma biology, the difference in pathways associated with the new signature compared to pathways of previously-published signatures was investigated using the KEGG pathway database (Kanehisa et al. 2012). The proposed new melanoma gene signature was found to be focused largely on four clusters of pathways: (i) steroid hormone biosynthesis (Table 14, yellow), (ii) diabetes, asthma and cytochrome P450 (Table 14, blue), (iii) immune, allograft and graft versus host (Table 14, red), and (iv) cell adhesion (Table 14, amber). However, compared to the pathways of the new proposed melanoma gene signature, previously-reported gene signatures were distributed in a remarkably large number of pathways and were not significantly over-presented in any pathways of Table 14. This suggests that the new signature reveal more biological meaningful functions than previously-published signatures.

Table 14: KEGG pathways associated with the proposed new melanoma gene signature

Pathway	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
hsa04940:Type I diabetes mellitus	1.08E-04	HLA-DQB1, HLA-DRB1, INS-IGF2, HLA-DRB3, HLA-DRB4, HLA-DRB5, IGF2	18.91741071	0.00836268	0.00836268	0.11382075
hsa05310:Asthma	6.74E-04	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	21.91810345	0.05119502	0.02593379	0.71013937
hsa05330:Allograft rejection	0.0012786	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	17.65625	0.09497692	0.03271775	1.344229911
hsa05332:Graft-versus-host disease	0.00161646	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	16.29807692	0.11854888	0.03105403	1.696682455
hsa00140:Steroid hormone biosynthesis	0.00260863	AKR1C3, AKR1C2, CYP3A5, AKR1C1	13.81793478	0.18432456	0.03992872	2.725139276
hsa04672:Intestinal immune network for IgA production	0.00312723	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	12.97193878	0.21675215	0.03989985	3.258823068
hsa05320:Autoimmune thyroid disease	0.00350579	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	12.46323529	0.23961624	0.03837734	3.646716314
hsa00980:Metabolism of xenobiotics by cytochrome P450	0.00555144	AKR1C3, AKR1C2, CYP3A5, AKR1C1	10.59375	0.35222956	0.05283064	5.718545124
hsa04514:Cell adhesion molecules (CAMs)	0.00792711	HLA-DQB1, PTPRF, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	6.019176136	0.46247403	0.06665032	8.073930173
hsa05416:Viral myocarditis	0.00886261	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	8.952464789	0.50060842	0.06708061	8.986724607
hsa04612:Antigen processing and presentation	0.01356338	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	7.65813253	0.65533539	0.09229411	13.45040645
hsa05322:Systemic lupus erythematosus	0.021687	HLA-DQB1, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	6.420454545	0.81917004	0.13282677	20.69943238
hsa05219:Bladder cancer	0.02662116	EGFR, FGFR3, IL8	11.35044643	0.87810438	0.14946656	24.82936579
hsa05200:Pathways in cancer	0.04608107	EGFR, JUP, FGFR2, WNT4, FGFR3, IL8	2.906821646	0.97477132	0.23113585	39.28775204
hsa04640:Hematopoietic cell lineage	0.0959101	HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5	5.543241279	0.99961585	0.40802902	65.57966783

Color Key: Blue represented diabetes, asthma and cytochrome P450; Red represented immune, allograft and graft versus host, Yellow represented steroid hormone biosynthesis and Amber represented cell adhesion

Recently published studies have confirmed these functional associations with melanoma. For example, Morrison et al. (Morrison *et al.*, 2011) and Norval (Norval, 2011) reported that type I diabetes mellitus, cytochrome P450 and asthma have all been related to vitamin D status, which is thought to play an important role in melanoma. Carretero et al. reported that immune rejection mediated the regression of metastatic melanoma (Carretero *et al.*, 2011). A recent study by Miyamoto et al. reported that steroid hormone biosynthesis is associated with the emergence of nevi, they also showed that the increased steroid levels will increase the rate of indoor tanning, which is a risk for melanoma (Miyamoto *et al.*, 2012). Koh and his colleagues (Koh *et al.*, 2012) have reported recently that some genes differentially expressed in sentinel lymph node metastases are involved in cell adhesion, though melanoma may not mention specifically in this study.

3.1.4. Interaction of a new 200-gene signature with melanoma 'driver' genes informs a new signaling network in melanoma

The interactions between genes within this 200-gene signature and the four known melanoma 'driver' genes (i.e., *NRAS, BRAF, MITF* and *cKIT*) in the corresponding pathways were investigated. Of these driver genes, *NRAS* is mutated in up to 25% of melanoma cases (Goel *et al.*, 2006; Schubbert *et al.*, 2007), while *BRAF* (located downstream of *NRAS*) is mutated in up to 45% of malignant melanomas (Hocker & Tsao, 2007; Flaherty & McArthur, 2010). *MITF*, a master transcription factor in melanocyte function, cooperates when mutated with BRAF in melanomagenesis (Garraway et al. 2005; Taylor et al. 2011), and

recent studies show that mutant *cKIT* can activate the Ras/Raf/Mek/Erk pathway as well as *MITF* (Monsel *et al.*, 2009; Phung *et al.*, 2011). While these four well-known melanoma driver genes do not appear on the 200 gene list as identified in microarray studies, this is most likely because these four driver genes are associated with melanoma at the gene mutation level, rather than at the gene expression level.

It is interesting to note, however, that 12 of the 200 genes closely interact with the above 4 melanoma driver genes. These 12 genes are engaged in 31 pathways retrieved from the KEGG database that connect with the MAPK, Ca^{2+} and WNT signaling pathways (Table 15), and included *EGFR, FGFR2, FGFR3, IL8, PTPRF, TNC, CXCL13, COL11A1, CHP2, SHC4, PPP2R2C,* and *WNT4*. The differential expression (i.e., fold change) of these 12 genes in each of the five original studies is shown Figure 11.

Table 15: KEGG Pathways where the 12 genes closely interact with melanoma driver genes (BRAF, NRAS, c-KIT and MITF)

No.	pathways	Driver genes assocated pathway (BARF, NRAS, c-KIT, MITF)	IL8	FGFR3	PTPRF	TNC	SHC4	CXCL13	EGFR	WNT4	FGFR2	PPP2R2C	CHP2	COL11A1
1	hsa04010	MAPK signaling pathway		√					√		√		√	
2	hsa04012	ERBB signaling pathway					√		√					
3	hsa04060	cytokine-cytokine receptor interaction	√					√	√					
4	hsa04062	chemokine signaling pathway	√				√	√						
5	hsa04115	p53 signaling pathway												
6	hsa04144	endocytosis		√					√		√			
7	hsa04360	axon guidance											√	
8	hsa04370	VEGF signaling pathway											√	
9	hsa04510	focal adhesion				√	√		√					√
10	hsa04530	tight junction										√		
11	hsa04540	GAP junction							√					
12	hsa04650	natural killer cell mediated cytotoxicity					√						√	
13	hsa04660	T cell receptor signaling pathway											√	
14	hsa04662	B cell receptor signaling pathway											√	
15	hsa04720	long-term potentiation											√	
16	hsa04722	neurotrophine signaling pathway					√							
17	hsa04810	regulation of actin cytoskeleton		√					√		√			
18	hsa04910	insulin signaling pathway			√		√							
19	hsa04912	GnRH signaling pathway							√					
20	hsa04916	melanogenesis								√				
21	hsa05160	hepatitis C	√						√			√		
22	hsa05166	HTLV-1 infection								√				
23	hsa05200	pathways in cancer	√	√					√	√	√			
24	hsa05212	pancreatic cancer							√					
25	hsa05213	endometrial cancer							√					
26	hsa05214	glioma					√		√					
27	hsa05215	prostate cancer							√		√			
28	hsa05218	melanoma							√					
29	hsa05219	bladder cancer	√	√					√					
30	hsa05220	chronic myeloid leukemia					√							
31	hsa05223	non-small cell lung cancer							√					

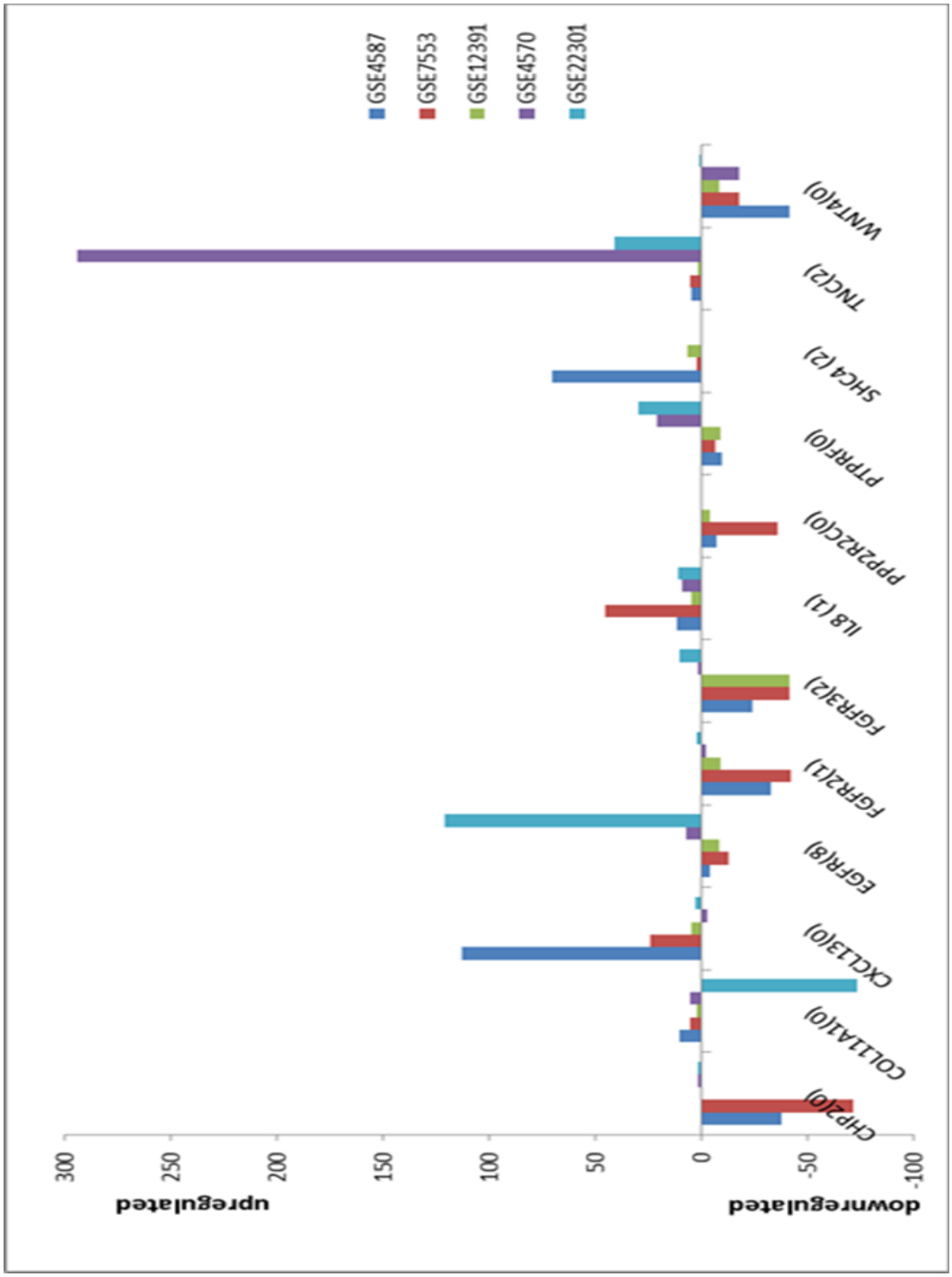


Figure 11: Differential expression (y-Axis) of genes in corresponding microarray data. The numbers in the bracket show how many wet-lab studies report on the association of these 12 genes with melanoma.

Based on the complex interactions of the 12 signature genes (labeled in red) and the 4 melanoma driver genes (*BRAF, cKit, NRAS, MITF*) in 3 signaling pathways (MAPK, Ca^{2+} and WNT), a new signaling network for melanoma is proposed, which may provide insight into melanoma development and progression (Figure 12). This new signaling network is created based on the relationship between the signature genes retrieved from the 31 KEGG pathways (shown in table 15). Most genes of the network are known and have been validated by wet-lab studies, only four genes in the network (see speculative gene pathway location with light purple hexagon background) were not reported by any previous studies. The speculated location of these four genes in the pathway below is based on where other members of their gene families are reported to be located. That being said, the 4 genes have appeared in previous pathways (e.g., WNT family was shown in WNT pathway of KEGG database, but *WNT4* in particular was). Thus, the involvement of these genes in this pathway needs formal investigation.

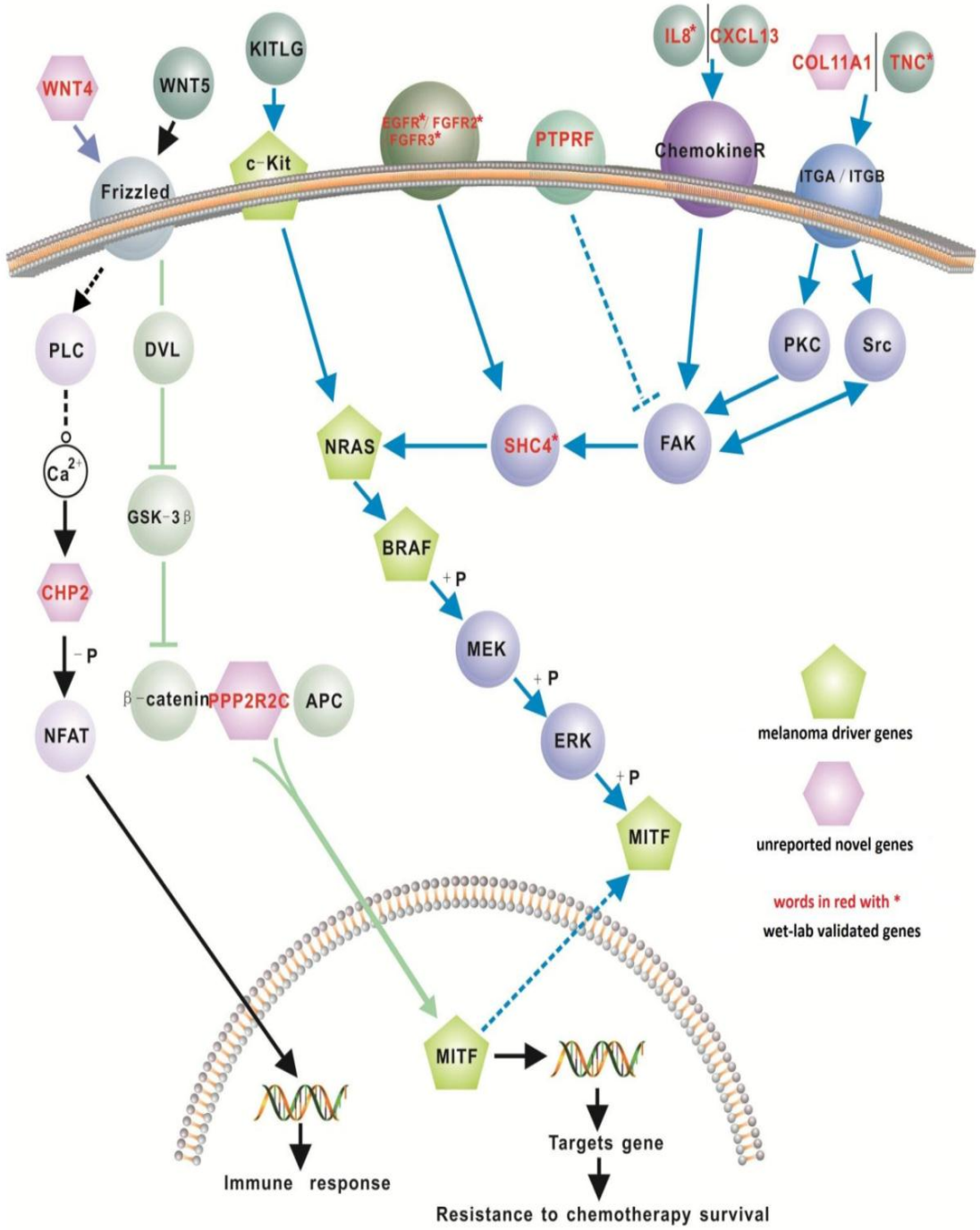


Figure 12： A proposed new signaling network for melanoma.
The signaling network is based on the complex interactions of the 12 signature genes (labeled in red) and the 4 melanoma driver genes (BRAF, cKit, NRAS, MITF) represented in 3 major signaling pathways (MAPK, Ca^{2+} and WNT). Nine of these 12 genes (i.e., EGFR, FGFR2, FGFR3, IL8, PTPRF, CXCL13, TNC, COL11A1, and SHC4) closely interact with three melanoma driver genes (i.e., NRAS, BRAF, and MITF) in the MAPK signaling pathway: The remaining 3 genes include WNT4, PPP2R2C and CHP2, which also play important roles in WNT and Ca^{2+} signaling pathways.

It is interesting to note that nine of these 12 genes (i.e., *EGFR, FGFR2, FGFR3, IL8, PTPRF, CXCL13, TNC, COL11A1,* and *SHC4*) closely interact with three driver genes (*NRAS, BRAF, and MITF*) in the MAPK signaling pathway as follows:

- *EGFR/FGFR2/FGFR3* Epidermal growth factor receptor and Fibroblast growth receptor: These cell surface receptors are activated by multiple ligands with multiple signaling trajectories.
- *PTPRF*: Protein tyrosine phosphatase receptor type F membrane receptor can regulate cell growth, differentiation, mitotic cycle, and oncogenic transformation. It can inhibit FAK (Focal adhesion kinase) by tyrosine-phosphorylation to process MAPK signaling (Medley et al, 2003), and can interact with β-catenin in melanoma (Bonvini et al. 2001).
- *CXCL13*: A 'novel' melanoma-associated gene detected in this study - is a small cytokine belonging to the CXC chemokine family and is selectively chemotactic for B cells of both B-1 and B-2 subsets.
- *IL8*: A 'novel' melanoma-associated gene detected in this study - also known as interferon-γ inducing factor- is a proinflammatory cytokine.
- *COL11A1*: This gene encodes one of the two α chains of type XI collagen, a minor fibrillar collagen.
- *TNC*: Encodes for Tenascin C, an extracellular matrix glycoprotein with anti-adhesive properties by binding to fibronectin and so blocking fibronectin's interactions with specific syndecans. The expression of tenascin-C in the stroma of certain tumours is associated with a poor prognosis. Tenascin-C is highly expressed in melanoma and promotes tumour progression

(Fukunaga-Kalabis et al. 2010).

- *SHC4*: A novel melanoma-associated gene detected in this study - forms a hub to connect *EGFR/FGFR2/FGFR3, PTPRF, CXCL13, COLL1A,* and *TNC* directly or indirectly, to work with *NRAS, BRAF, MITF* driver genes in the MAPK pathway.

The remaining 3 of these 12 genes in this proposed new melanoma signature included *WNT4, PPP2R2C* and *CHP2*, which also play important roles in WNT and Ca^{2+} signaling pathway as follows:

- *WNT4* binds to Frizzled receptor on the cell surface to promote the WNT signaling pathway.
- *WNT4* and *PPP2R2C* interact with *MITF* in the WNT signaling pathway. The deregulation or activation of the WNT pathway may lead to melanoma formation (Larue and Delmas 2006).
- A complex is formed by β-catenin with *PPP2R2c* and *APC* (adenomatosis polyposis coli protein) to inhibit the degradation of β-catenin.
- *CHP2* is activated by calcium ions in a Ca^{2+} signaling pathway and has a role in the immune response.

Seven genes of our 200 signature including *EGFR, FGFR2, IL8, IGF2, TACSTD2, PPP1R14C* (*KEPI*) and *LEP* are reported to be associated with *PTEN* (Phosphatase and Tensin Homolog). *PTEN* is a tumour suppressor gene that is mutated in many cancers including in many cases of melanoma (Guntur *et al.*, 2011). The loss of *PTEN* protein function was found in melanoma cell

spreading, migration as well as in growth factor-stimulated MAPK signaling. *PTEN* loss has been reported as a late event in melanoma (Guntur *et a*l., 2011). Indeed, this has been confirmed for 3 of these 7 genes (*IL8, IGF2, LEP*) in independent melanoma studies. The other four genes (*EGFR, FGFR2, TACSTD2, PPP1R14C*) have not been previously reported to be associated to *PTEN* in melanoma.

3.1.5. Experimental validation of a MAPK pathway-associated subset in our 12-gene melanoma signature

Based on the expression values of the 12 genes (figure 11), it was clearly seen that the expression of *TNC, PRPTF, SHC4, COL11A1, IL8, EGFR, CXCL13* were up-regulated to a high ratio between nevi and melanoma in 5 melanoma datasets. However, *PTPRF, COL11A1* and *CXCL13* have not yet been reported to be associated with melanoma (Figure 11). *SHC4* plays an important role as a hub in the MAPK pathway (Figure 12). As COL11A1, CXCL13, PTPRF and SHC4 appear in the MAPK signaling pathway (KEGG pathway), I further attempted to validate their expression via wet-lab analysis, including in human melanoma cell lines and melanoma tissues. Using this approach, COL11A1, CXCL13, PTPRF, and SHC4 were found to be over-expressed in two human melanoma cell lines (i.e., FM55 and FM94) compared to normal human epidermal melanocytes *in vitro* (Figure 13). A significant degree of heterogeneity in the expression pattern for these markers was observed. For example, COL11A1, a secreted collagen protein, was observed at low levels in the cytoplasm of normal melanocytes, but much more intensely in the perikayon of moderately-pigmented FM55 melanoma cells, and unexpectedly exhibited a

nuclear/nuclear membrane association in the pigmented FM94 melanoma cells. Similarly, a weak cytoplasmic localization of CXCL13 in normal melanocytes appeared to shift towards the perikayon and nucleus of FM55 and FM94 melanoma cells respectively, as evidenced by co-localization with DAPI staining. Low level PTPRF expression in normal epidermal melanocytes contrasted with higher expression (both cytoplasmic and nuclear) in melanoma cells. Finally, SHC4 expression was membranous in normal melanocytes, contrasting with some punctuate nuclear membrane expression in melanoma cells (Figure 13).

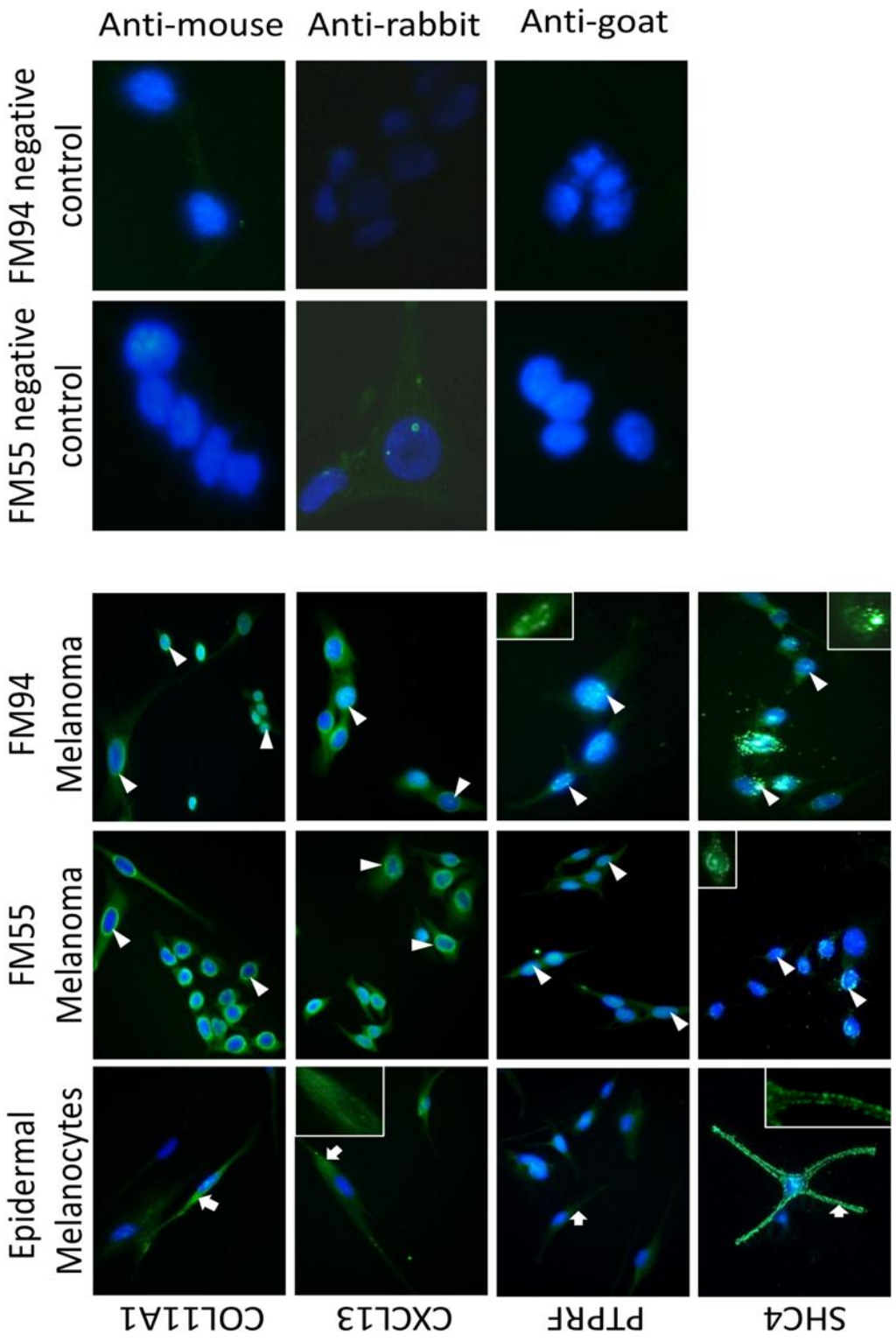


Figure 13: Immunocytochemical analysis of human epidermal melanocytes and melanoma cells *in vitro*.

The expression of COL11A1, CXCL13, PTPRF and SHC4 proteins was upregulated (green fluorescence) in melanoma cells. Inserts show higher power views of expression, including when associated with the perinuclear region of the cell.

The expression of these proteins encoded for by these four genes in the proposed 12-gene signature for melanoma was also assessed in normal human healthy skin and in melanoma patient tissue (both primary and metastatic melanoma). Using double immunofluorescence with a melanocyte lineage marker gp100, is expression profile of these four test proteins in melanocytes or melanoma cells within these tumour biopsy tissues was assessed. We included primary melanoma in addition to metastatic melanoma in a immunohistochemistry-based validation study because the expression levels for the 12 genes in the gene signature exhibited a several fold change between primary melanoma and normal skin/benign nevi across five microarray datasets (Table 16).

Table 16: Fold change in the expression of 12 genes between normal skin/benign nevi and primary melanoma in five microarray datasets

Gene	GSE4570	GSE4587	GSE7553	GSE12391	GSE22301
EGFR	-2.806	-25.666	-4.750	-1.548	22.344
FGFR2	3.168	-1068.810	-6.010	-1.824	1.907
FGFR3	-1.398	-232.953	-6.730	-1.555	17.991
CXCL13	-1.058	60.146	18.980	1.436	2.722
COL11A1	-37.162	4.720	10.980	2.037	1.296
WNT4	-36.157	-39.452	-5.310	-1.985	1.103
PTPRF	3.811	-3.445	-3.010	-1.871	32.920
PPP2R2C	NA	-14.128	-11.340	-1.207	NA
TNC	191.111	4.232	8.740	4.000	41.612
IL8	-1.748	130.984	13.570	7.148	527.484
CHP2	1.658	-352.380	-8.340	NA	1.750
SHC4	NA	26.420	2.240	2.238	NA

This experiment showed that COL11A1, CXCL13 and PTPRF were not expressed in normal human epidermal melanocytes *in situ* (Figure 14a), while

some low level expression of SHC4 was detected in normal pigment cells. By contrast, COL11A1 was expressed intensely by melanoma cells located in the dermis of both primary and metastatic melanoma (Figure 14b and 14c). CXCL13 was strongly expressed in a minor subpopulation of tumour cells in primary melanoma, while a greater fraction of cells in metastatic melanoma tissue expressed this protein. By contrast, PTPRF was intensely expressed in the majority of tumour cells of both primary and metastatic melanoma cells. Finally, SHC4 was found to be expressed in only a minor fraction of gp100-positive primary melanoma, but in most gp100-positive metastatic melanoma cells.

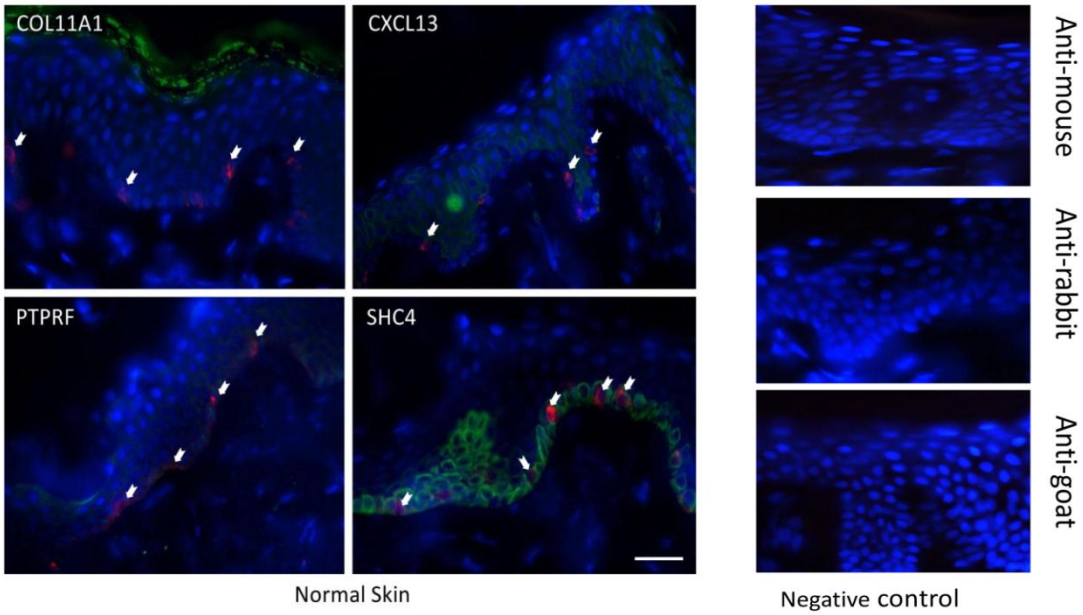


Figure 14:
(a) Immunohistochemical analaysis of COL11A1, CXCL13, PTPRF and SHC4 in normal human skin epidermis (frozen donor-Female 52yrs sections).

Melanocytes were detected using an antibody (NKi/beteb) raised against the melanocyte-specific marker gp100 (red, arrows). COL11A1, CXCL13, PTPRF (green) were not detected in normal epidermal melanocytes. SHC4 was expressed sronglty in proliferating keratinocytes in the basal layer on the epidermis, but only very weakly in melanocytes (i.e. double positive cells in orange-yellow).

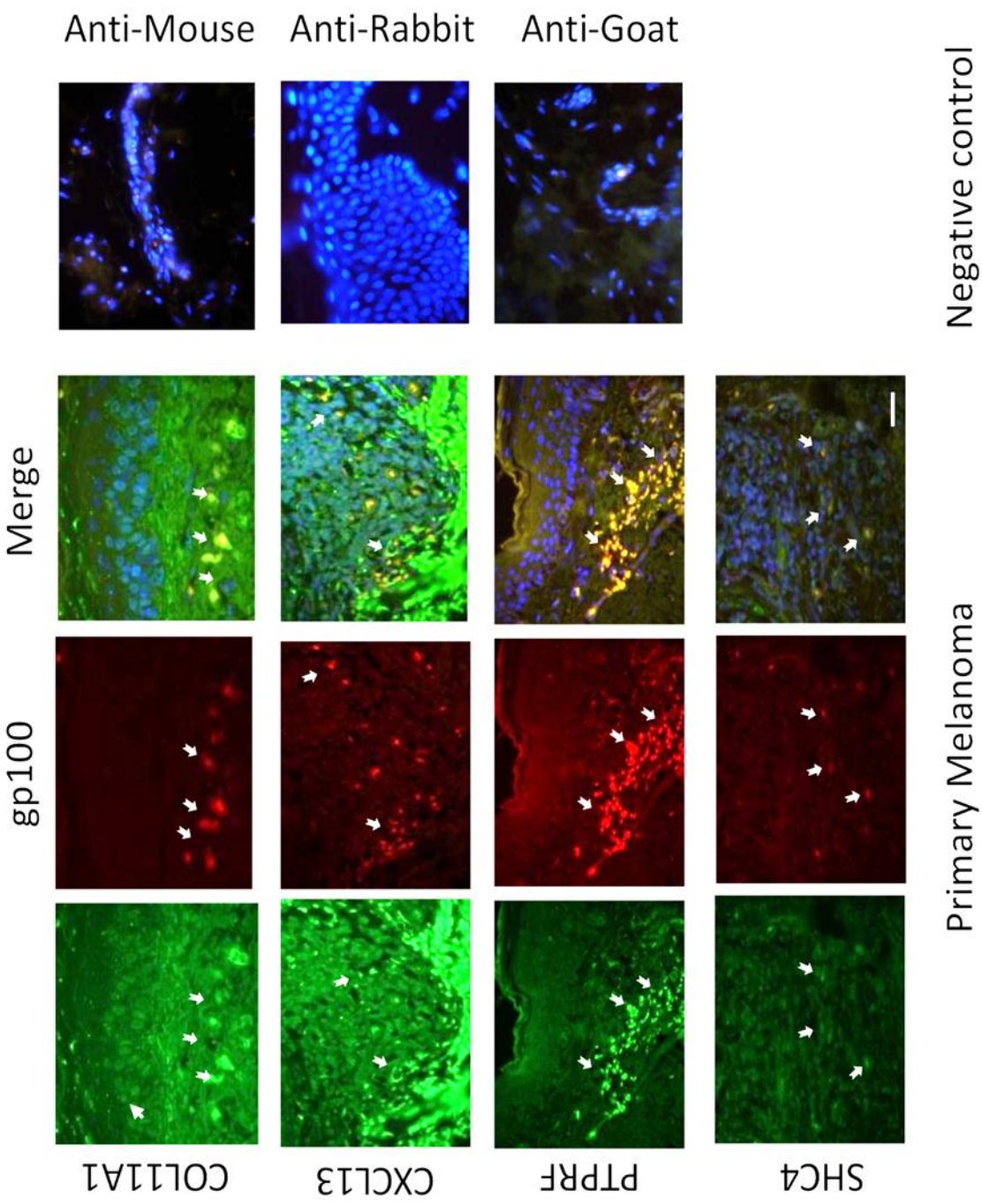


Figure 14:

(b) Immunohistochemical analaysis of COL11A1, CXCL13, PTPRF and SHC4 in primary melanoma. Double staining of test protein (shown in green) and pigment cell lineage-specific marker gp100 (in red, arrows). Both immunoreactivites were merged with yellow/orange fluorescence indicating co-localization of these protiens in melanoma cells.

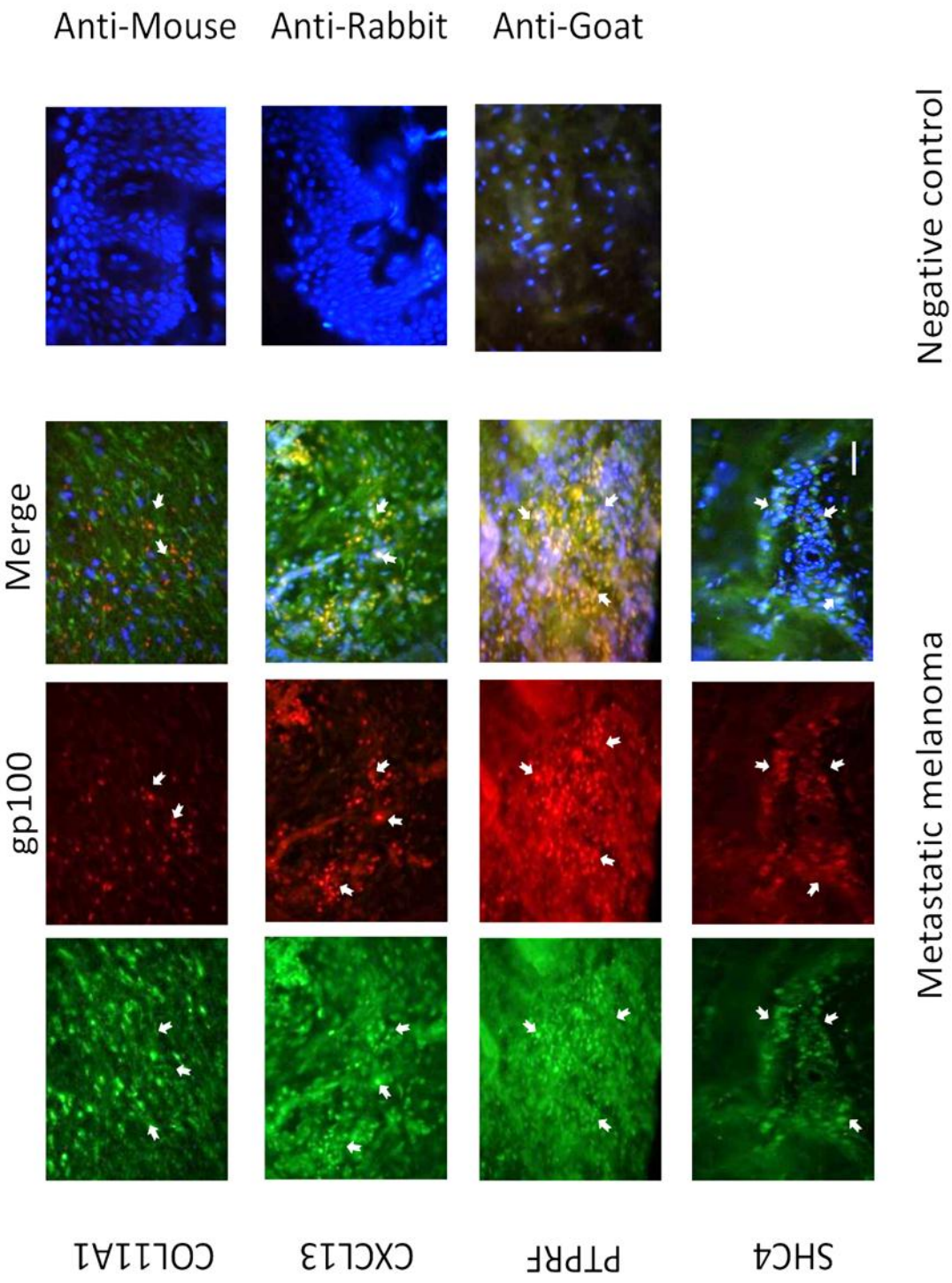


Figure 14:

(c) Immunohistochemical analaysis of COL11A1, CXCL13, PTPRF and SHC4 in metastatic melanoma. Double staining of test protein (shown in green) and pigment cell lineage-specific marker gp100 (in red, arrows). Both immunoreactivites were merged with yellow/orange fluorescence indicating co-localization of these protiens in melanoma cells.

3.1.6. Computational evaluation of the robustness of a proposed 12-gene biomarker signature to distinguish melanoma from normal skin and/or benign nevi

A computational evaluation of the robustness of the proposed 12-gene signature was performed with the aim of distinguishing melanoma from normal skin and/or benign nevi using cross-laboratory published microarray data. This data evaluation to assess the robustness of a new biomarker may yield information that may have potential diagnostic application and/or possible therapeutic development. The Support Vector Machine classification model (known as the SVM model) (Brown et al. 2000) and the 'leave-one-out method' were used to classify the microarray datasets in this project (Hoek et al. 2004, Smith et al. 2005, Riker et al. 2008, Scatolini et al. 2010, and Rose et al. 2011). Results showed that these 12 genes achieved excellent classification accuracy ratios across all five datasets (i.e., averaging at 99.1%, Table 17). It is noted that the new 12-gene biomarker signature achieved a much better performance, on average, than the signatures of Smith et al. 2005, Riker et al. 2008, and Scatolini et al. 2010, and only very slightly less (0.44% less) classification accuracy than the signature of Hoek et al. 2004. It should be noted that the signature of Hoek et al. (2004) consisted of 589 genes, while this project's biomarker signature is very much shorter at just 12 genes. Thus, the 12-gene biomarker achieved the best classification accuracy, compared to the original signature of the individual studies, with a much smaller number of genes.

Table 17：Classification accuracy of four original signatures and our proposed 12-gene signature using across-laboratory data

Original signatures	GSE4570 (2004)	GSE4587 (2005)	GSE7553 (2008)	GSE12391 (2010)	GSE22301 (2011)	Average
Hoek *et al*, 2004 (589)	100%	100%	97.78%	100%	100%	99.556%
Smith *et al*, 2005 (100)	71.43%	100%	97.78%	100%	100%	93.842%
Riker *et al*, 2008 (65)	71.43%	100%	95.56%	100%	100%	93.398%
Scatolini *et al*, 2010 (455)	85.71%	100%	97.78%	100%	100%	96.698%
New 12-gene biomarkers	100%	100%	95.56%	100%	100%	99.11%

3.2. Improvement of the computational method

One of the potential weaknesses of the above method of generating the 12-gene signature was the inclusion of data from both melanoma cell line and skin biopsy material. It was considered therefore move appropriated to attempt to 'stratify' these data to analyse separately by sample type. Thus, the method for integrated analysis of microarray data was further improved to consider the gene differences on each individual dataset by platform, sample types etc. to make the method more biological meaningful. The improved method was called the "stratification based method".

In order to evaluate the performance of the "Stratification based method" on the integrated analysis of multiple microarray datasets, the method was applied on a broader set of 14 datasets (table 6) covering three distinct diseases (i.e., Breast cancer, prostate cancer, as well as melanoma).

For each disease, the microarray datasets were prepared for the 'leave-one-out' validation. This involved separating the microarray datasets into two groups: one group was for 'training' to generate the ranked gene list, while the other group was for 'independent testing' (see methods). In order to understand how two main factors may lead to performance enhancement (i.e., influencing dataset quality, and the sub-grouping of datasets), results obtained by using four different methods (see Figure 8 and methods) were compared. These 4 methods included: **(1)** datasets treated equally (**non-stratification**); **(2)** datasets treated with appropriate weighting (**non-stratification with weighting**). **(3)** datasets first sub-grouped and then treated equally (**stratification**), **(4)** datasets sub-grouped, but then each dataset treated with corresponding weighting (**stratification with weighting**). To evaluate the performance of the above four methods, the associated classification accuracy ratios were compared. The data emerging from these gene ranking analyses above were further evaluated to determine whether this stratification procedure, for detecting significantly-ranked and disease-associated genes, was more robust than those previously published. In brief, I checked whether there were good classification accuracy ratios across different test data, or whether more common genes occurred between the 'training' studies and the 'independent testing' dataset in the following ways:

a) Classification accuracy of genes using 'Wrapper feature selection' and Machine Learning (SVM method): This is a computational measurement performed by both the 'support vector machine classification' model and the 'leave-one-out' method with Wrapper feature (e.g., Peng *et al*., 2010) and previously described earlier in this thesis (see Section 2.2.5; Liu *et al.*, 2013).

The validation method was applied to detect the classification accuracy ratio between these top-gene ranking lists derived after combining multiple 'training' datasets and the 'individual testing' dataset.

b) Evaluation of the number of common genes between the training datasets and the independent test study: This was an analysis of the genes common between the top-gene lists by combining 'training' datasets and the top gene lists from the corresponding 'individual testing' study.

c) Evaluation of the number of common genes across different gene ranking lists of combined 'training' studies: This was an analysis of the common genes between these gene ranking lists that were derived from different 'training' datasets combinations.

d) Evaluation of the number of common genes between the combined melanoma studies and 180 known metastatic melanoma biomarkers: This was an analysis of the common genes between the gene lists were derived from different melanoma 'training' datasets combinations and the 180 known metastatic melanoma biomarkers (the 180 biomarkers were picked up from PubMed based on conditions as follows: metastatic, melanoma, biomarkers published to March of 2013, Appendix-table s3).

3.2.1. Classification performance validated by wrapper feature selection and machine learning

A top 500 gene list was generated using each of the above methods (i.e., 1-4) for each disease. These lists were further analysed with the 'Wrapper feature selection' method by a one-gene incremental addition approach (figure 15).

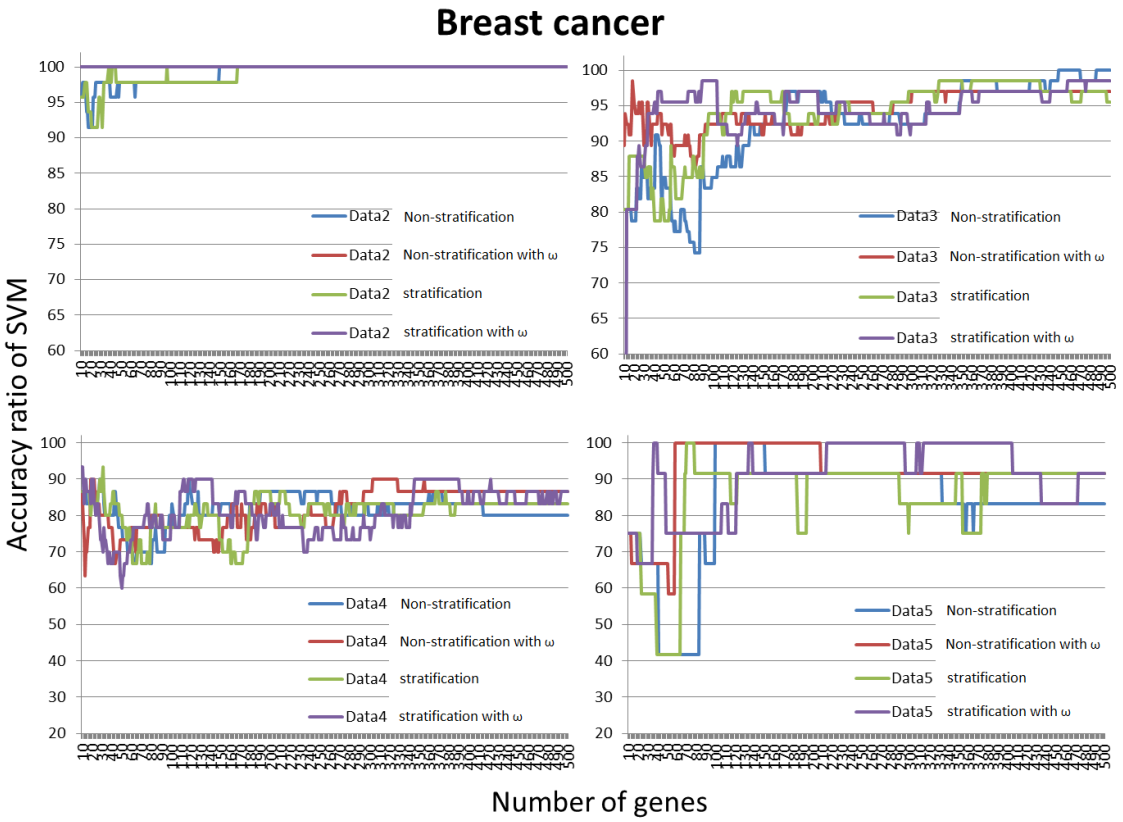


Figure 15a: Performance curve of the four methods when applied to **Breast cancer** data.

Horizontal axis: Number of genes between 10 and 500. Vertical axis: Accuracy ratio in % of SVM. The different line colors indicate the four different methods as indicated in the embedded key. The data number (e.g., Data4) shows which dataset was applied as the test data in the 'leave one out' study. (See table 7 datasets for the corresponding disease datasets).

Note: Data 1, cell line-type data only and so was not used for testing data.

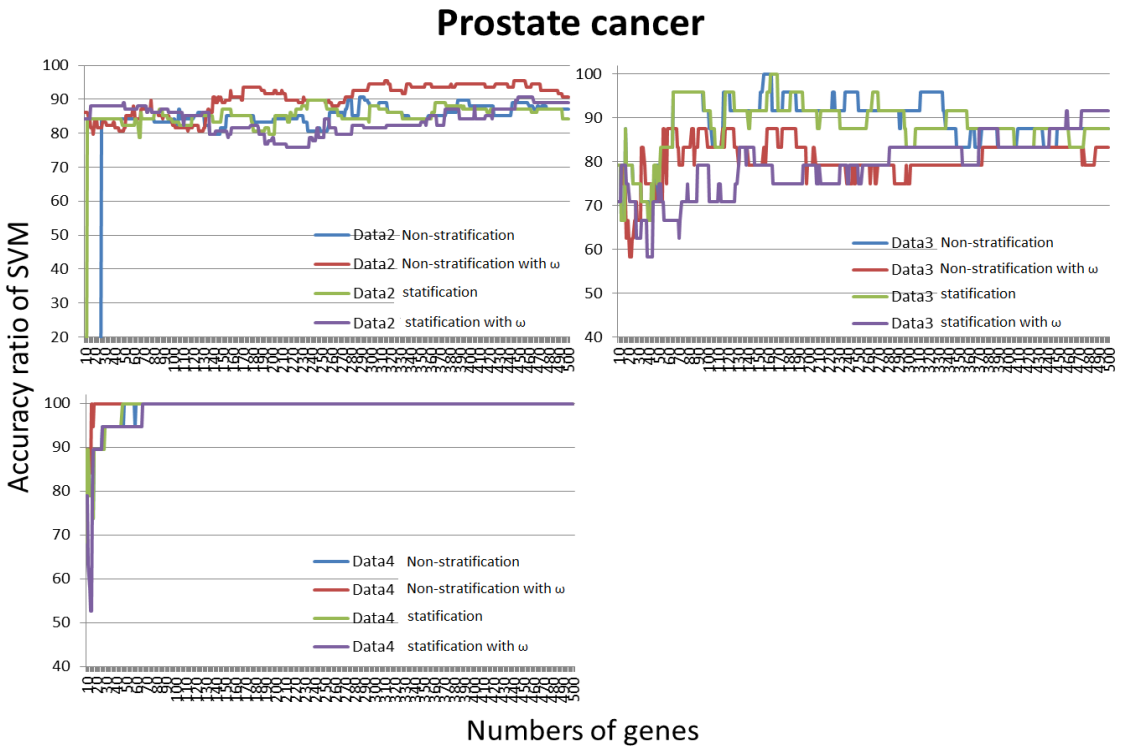


Figure 15b: Performance curve of the four methods when applied to **Prostate cancer** data.

Horizontal axis: Number of genes between 10 and 500. Vertical axis: Accuracy ratio in % of SVM. The different line colors indicate the four different methods as indicated in the embedded key. The data number (e.g., Data4) shows which dataset was applied as the test data in the 'leave one out' study (See table 7 datasets for the corresponding disease datasets).

Note: Data 1, cell line-type data only and so was not used for testing data.

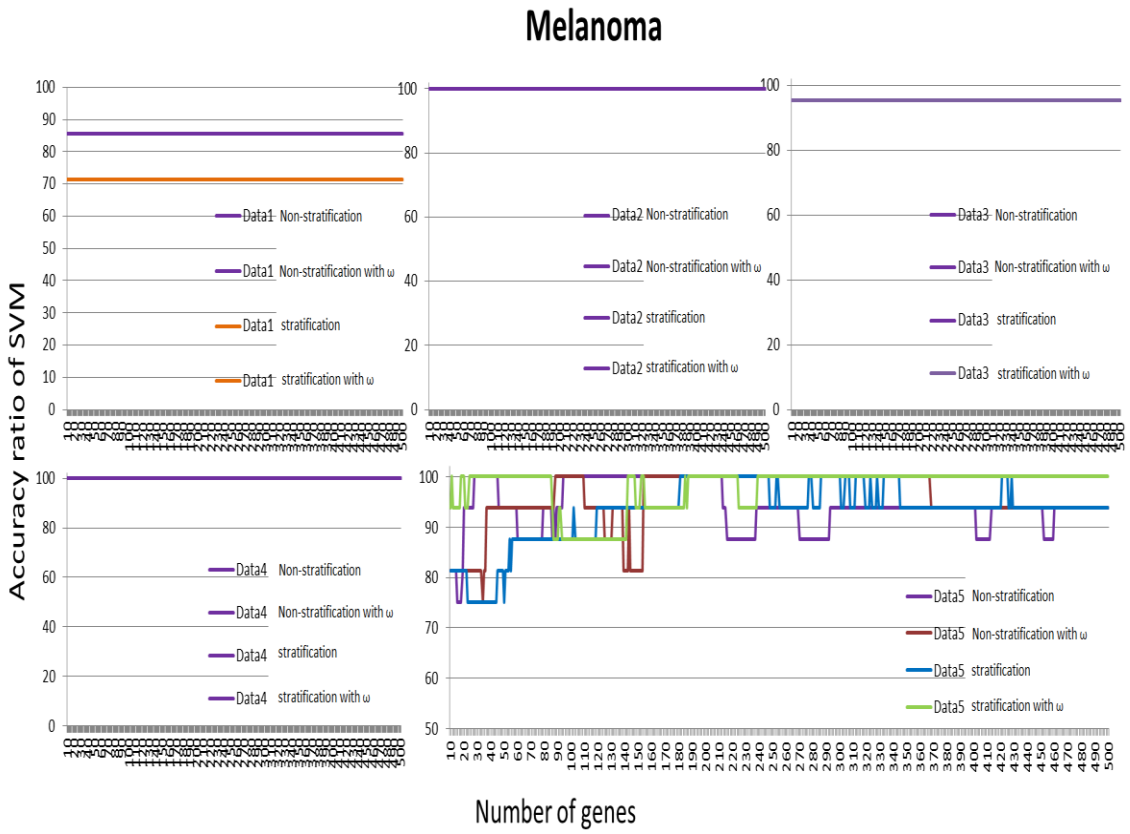


Figure 15c: Performance curve of the four methods when applied to **Melanoma** data.

Horizontal axis: Number of genes between 10 and 500. Vertical axis: Accuracy ratio in % of SVM. The different line colors indicate the four different methods as indicated in the embedded key. The data number (e.g., Data4) shows which dataset was applied as the test data in the 'leave one out' study (See table 7 datasets for the corresponding disease datasets).

In Breast Cancer, the accuracy ratios of ‘stratification with weighting’ method (purple lines) were highest in data2 and data5 independent test studies, and second highest in data3 and data4 independent test studies. The next best accuracy ratios for breast cancer data was generated using the ‘non-stratification with weighting’ method (red lines), the ratios showed the second highest position in data3 and data5 independent test studies. The other two methods did not generate good ratios in this disease (figure 15a).

In Prostate Cancer, the highest ratios were produced by ‘non-stratification with weighting’ method (red lines) in data2 independent test studies, followed by ‘stratification with weighting’, ‘non-stratification’, and ‘stratification’ for this dataset. In data3 and data4 independent test studies, the ‘stratification with weighting’ method produced the highest ratios (purple lines), but the accuracy ratio of ‘non-stratification with weighting’ was not good in data3 (red lines). The other two methods always produced middle level ratios; however, that said the ratios of ‘stratification’ method (green lines) were always higher than the ratios of ‘non-stratification’ method (blue lines) (Figure 15b).

In Melanoma, the ratios are good for 3 of 5 total independent test studies. In data1, the accuracy ratios of ‘non-stratification’ with or without ‘weighting’ methods (purple lines) were higher than the ratios of two ‘stratification’ with or without ‘weighting’ methods (orange lines). However, the two highest ratios in data5 were generated by ‘stratification with weighting’ method (green lines) and ‘stratification’ method (blue lines) (Figure 15c).

When the accuracy ratios of all three diseases were combined, the best accuracy ratios were obtained with the 'stratification with weighting' method (purple lines mostly; green lines in melanoma data5). This method yielded good results regardless of whether the other 3 methods did or did not individually, e.g. ratios in testing data2 of breast cancer or ratios in testing data1 of melanoma. The next best method was the 'non stratification with weighting' method. Thus, these two methods showed that the 'weighting' plays an important role, reflecting the benefit of measurement of each dataset, with qualities adjusted by *D* values calculated by the arrayQualityMetrics' R package (see section 2.4.2). Comparison of the method with or without 'stratification' showed that the separation of data derived from 'biopsy' or 'cell line' produced more robust data, compared to applying the non-stratified combined cell line plus biopsy datasets. Thus, the two factors 'stratification' and 'weighting' were valuable method enhancements.

3.2.2. Evaluation of the number of common genes between the combined training studies and the independent test study

In this study, the numbers of common genes between the combined 'training' studies and the 'independent test' study were detected 50 times (or 50 rounds), starting from applying 10 genes of top-gene lists up to the maximum 500 genes (using a 10 gene incremental addition approach). As gene numbers were increased, the numbers of common genes would be expected to increase (Figure 16). Figure 15 and figure 16 include results in breast cancer, prostate cancer and melanoma respectively.

Clearly, the stronger will be to enhance that those common genes are truly associated with the disease state. Thus, under one method, if the top genes are retrieved regardless of the datasets used, the more robust that gene selection is likely to be (Figure 16).

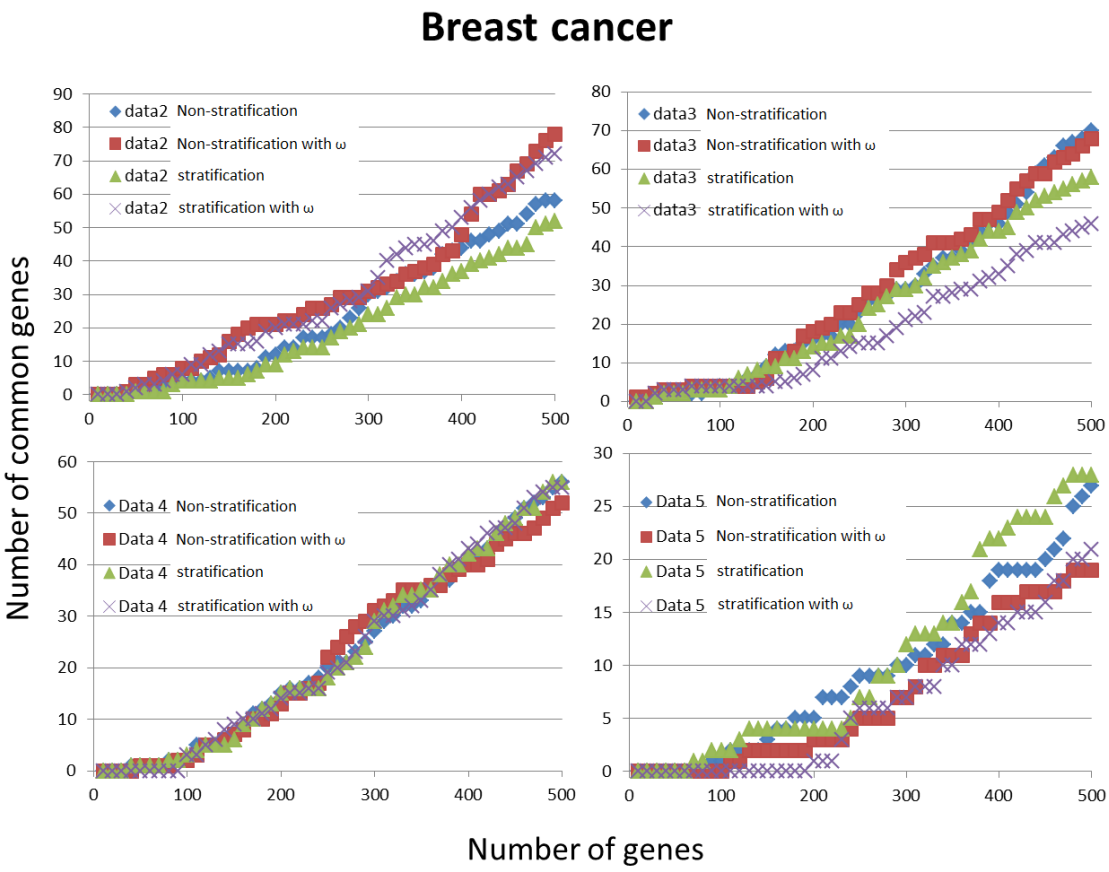


Figure 16a: Common genes between combined studies and individual test study for **breast cancer**

Horizontal axis: Number of genes between 10 and 500. Vertical axis: the number of common genes. The different color symbols indicate the four different methods. The data number (e.g., Data4) shows which dataset was applied as the test data in the 'leave one out' study. The datasets are showed on table 7 for the corresponding disease.

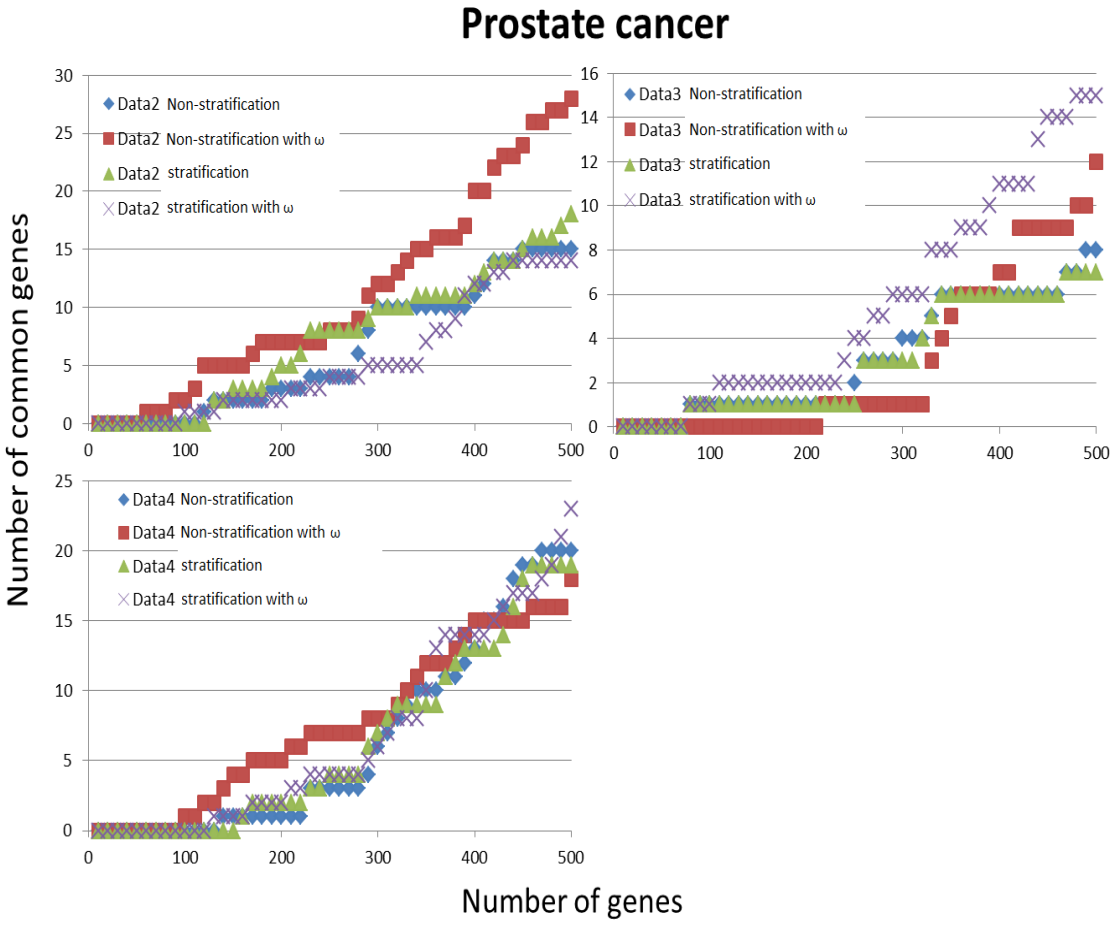


Figure 16b: Common genes between combined studies and individual test study for **prostate cancer**.

Horizontal axis: Number of genes between 10 and 500. Vertical axis: the number of common genes. The different color symbols indicate the four different methods. The data number (e.g., Data4) shows which dataset was applied as the test data in the 'leave one out' study. The datasets are showed on table 7 for the corresponding disease.

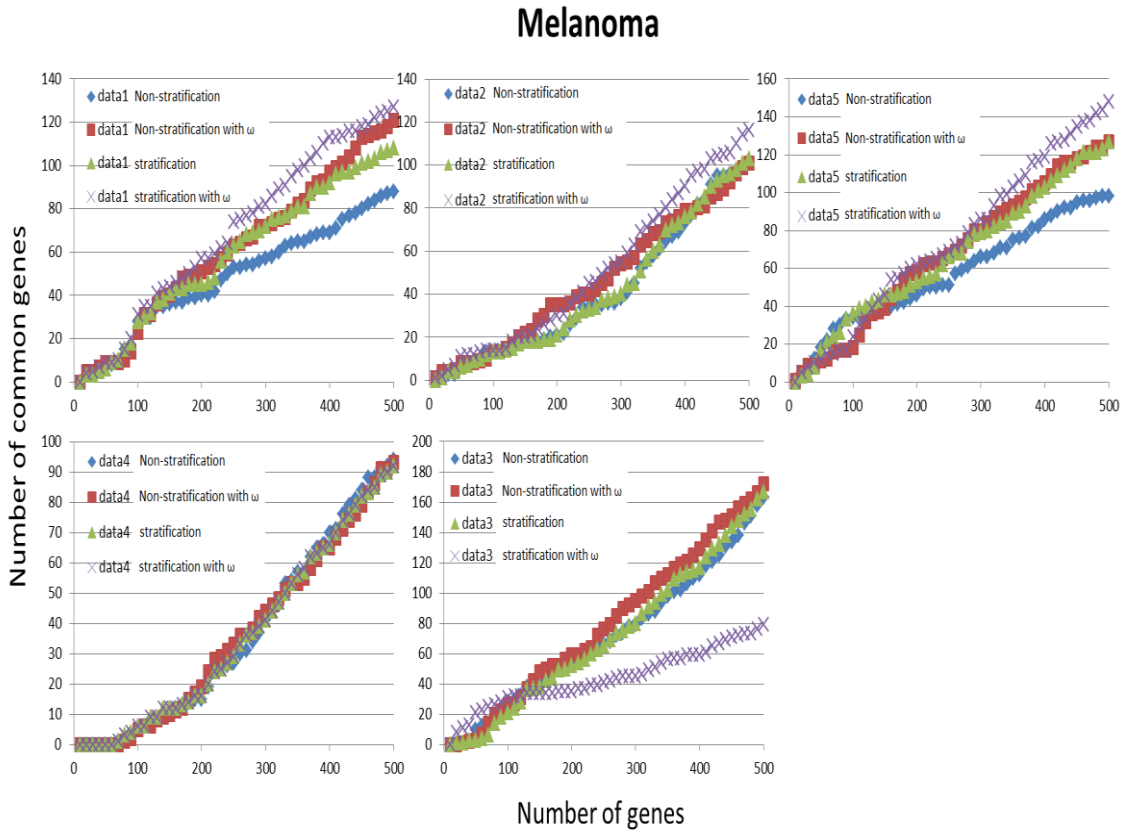


Figure 16c: Common genes between combined studies and individual test study for **melanoma**.

Horizontal axis: Number of genes between 10 and 500. Vertical axis: the number of common genes. The different color symbols indicate the four different methods. The data number (e.g., Data4) shows which dataset was applied as the test data in the 'leave one out' study. The datasets are showed on table 7 for the corresponding disease.

For Breast Cancer, the best two methods were ‘stratification’ and ‘stratification with weighting’ methods, even though they did not generate a good number of common genes for data2 and data3 respectively; these methods yield good results for the other individual datasets, i.e., ‘stratification’ method in data4 and data5, and ‘stratification with weighting’ method in data2 and data4. Thereafter, ‘non-stratification with weighting’ generated the best results in data2 and date 3, but also produced the poorest results for data4 and data5. The worst method was ‘non-stratification’, which yielded a middle range for common genes (figure 16a).

For Prostate Cancer, more common genes were obtained by the ‘stratification with weighting’ method, although this method was poor for data2. The next best method was ‘non-stratification with weighting’, with more common genes generated using this method for data2 and data3 (Figure 16b).

For Melanoma, the best method was ‘stratification with weighting’ method, which generated the highest number of common genes in 4 of 5 individual studies. The next best was the ‘non stratification with weighting’ method, followed by ‘stratification’ method and then the ‘non-stratification’ method (Figure 16c).

When combined the three disease results together, the ‘stratification with weighting’ method yielded the highest number of common genes for all three

diseases (shown as purple cross). Only in data3 of melanoma was there a large difference between the number of common genes of ‘stratification with weighting’ method and common genes of the other three methods. The reason is the large difference may be induced by the weighting (ω) (i.e., the quality of datasets). The ‘non-Stratification with weighting’ method also gave good results on most of datasets (green triangle), but the linear trend of numbers of common genes in this method was quite different from the trends of other three methods. And the numbers of common genes of this method also were not stable when applied it to different ‘training studies’, e.g. especially in breast cancer and prostate cancer (Figure 16).

3.2.3. Evaluation of the number of common genes across different gene ranking lists of the combined training studies

In order to identify the stability of each method, method stability is defined as the ability to retrieve similar significant genes regardless of changes in different combined datasets in each test time. Thus, the number of common genes detected across different gene ranking lists of combined studies was a measurement standard. The higher number of common genes means the more stable method. Like the above evaluation approach, the numbers of common genes were detected via 50 rounds, staring by applying the top 10 genes to the maximum top 500 genes, with 10 genes incrementally added approach each time. The best method should generate the highest number of common genes, and the linear trend of common genes in the figure 17 should reflect increasing stability of the method.

The results obtained by the ‘non-Stratification’ method (blue diamond) indicated that this method performed least well compared to the others, followed by ‘Stratification’ method (green triangle). The trends in results of ‘non-Stratification with weighting’ (red square) and ‘stratification with weighting’ (purple cross) methods were similar, the former one was lower in prostate cancer, and the latter was lower in melanoma, while the ‘stratification with weighting’ method was the best method in both breast and prostate cancer. This indicates that this method is relatively more stable than the others (Figure 17).

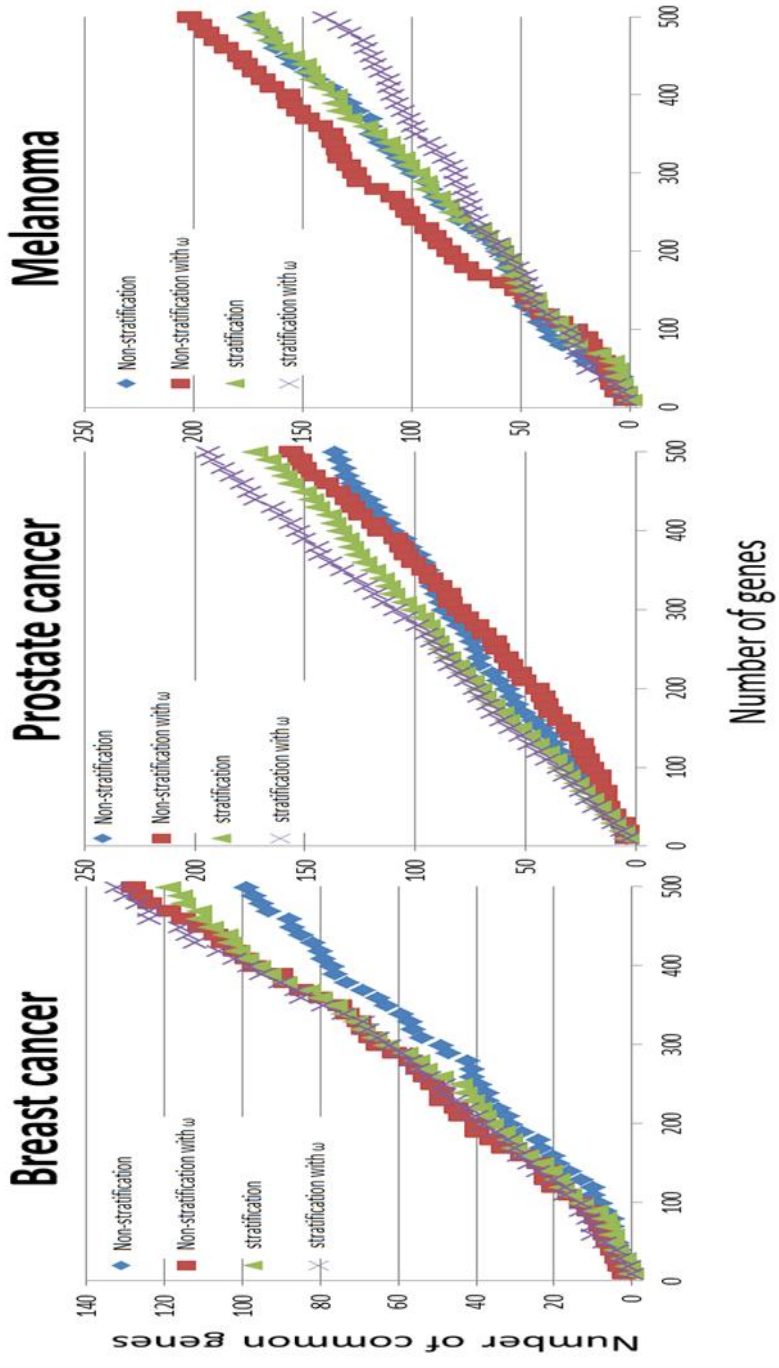


Figure 17: Common genes across different top gene ranking lists of combined studies.

Horizontal axis: Number of genes between 10 and 500. Vertical axis: the number of common genes. The different embedded color symbols indicate the four different methods.

3.2.4. Evaluation of the number of common genes between the combined melanoma studies and 180 known metastatic melanoma biomarkers

The enhanced method was then checked to determine whether it could retrieve more gene biomarkers for melanoma than had been identified so far in the literature. A PubMed search (query: metastasis, melanoma, biomarkers, last accessed March 2013) retrieved 180 metastatic melanoma biomarkers (see table s3 in appendix). These 180 genes are defined here as 'true biomarkers' for melanoma, and were compared with the different top gene ranking lists that were selected based on the different methods using the 5 melanoma datasets (Hoek *et al.*, 2004, Smith *et al.*, 2005, Riker *et al.*, 2008, Scatolini *et al.*, 2010, Rose *et al.*, 2011), four of the 5 studies were combined as 'training' studies at each stage of analysis. Similar to the above approach, the 10 genes incremental addition approach was applied between 10 and 500 genes. The number of common genes was used to measure the effectiveness of the method in identifying true biomarkers. The 'more common genes' outcome means that the method was able to identify true melanoma biomarkers (i.e., true positives). With results shown in figure 18, it was clearly seen that the 'non-stratification with weight' method (red square) revealed the greatest number of the 180 metastatic melanoma biomarkers, while the next best method is 'stratification with weight values' (purple cross).

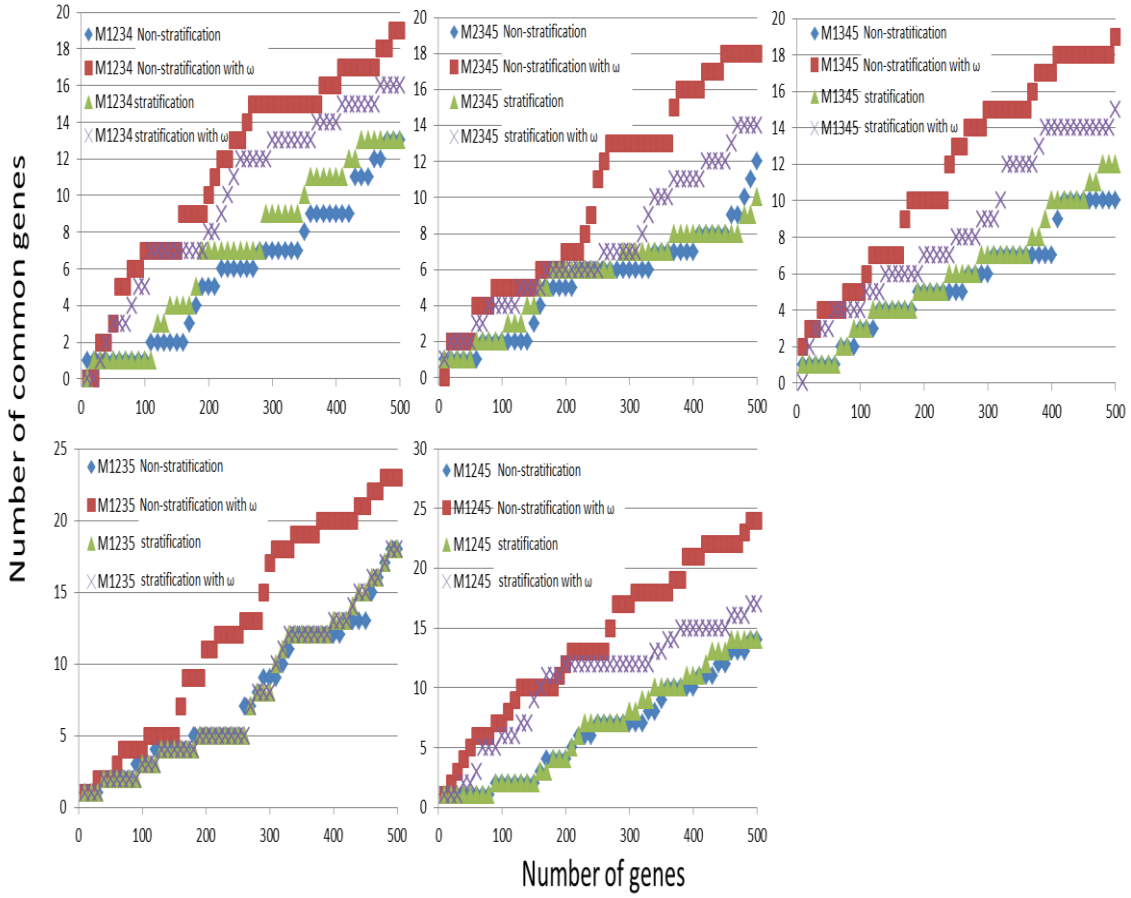


Figure 18: Common genes between the top gene ranking lists of combined studies and the 180 metastatic melanoma biomarkers

Horizontal axis: Number of genes between 10 and 500. Vertical axis: the number of common genes. The different color symbols indicate the four different methods. The data number (e.g., M1234) shows which four datasets were applied as combined studies in the 'leave one out' study.

Summarizing all four above evaluations, it was seen that the 'stratification with weighting' method produced more reliable results than did the other three methods. In order to forward validate this viewpoint, a top gene list were derived from all 5 datasets of melanoma, and then compared with the known 180 metastatic melanoma biomarkers (five datasets shown on table 6). It was found that the combined all 5 datasets together could generated the most common gene outcome when it compared to the other outcomes of 'stratification with weighting' in Figure 18 (purple cross). The final number of common genes was 18 (figure 19). However, in figure 18, the highest numbers in each common gene outcome by 'stratification with weighting method' were approximate 14 or 16. The improved number may provide an evidence for the more microarray datasets using in integrated analysis the higher number of reliable true biomarker can be identified (Figure 19).

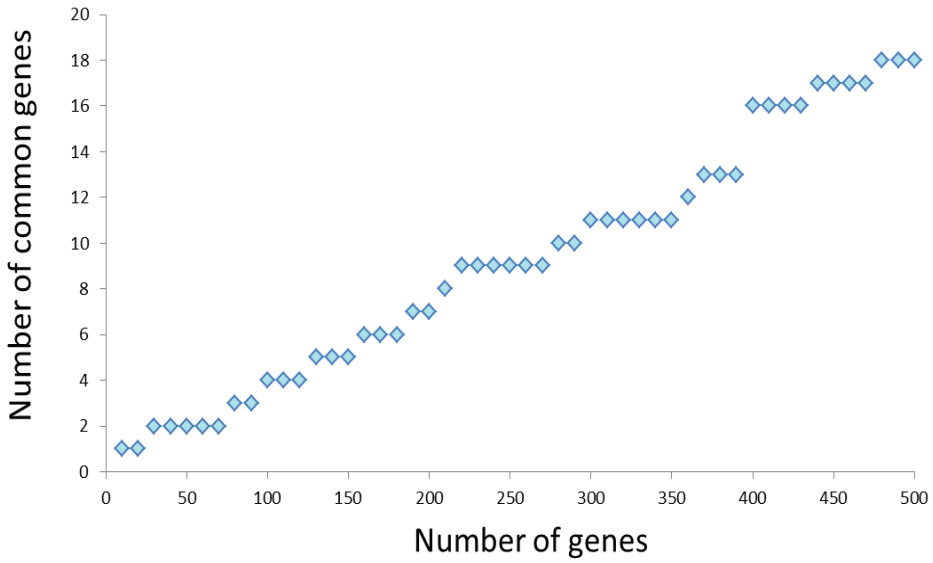


Figure 19: Common gene numbers between top gene ranking lists of all 5 melanoma datasets and 180 metastatic melanoma biomarkers

3.2.5. A new 200 gene signature of melanoma derived from the application of 'stratification with weighting' method

In order to check if 'stratification' and 'weighting' are two factors which are able to enhance the biomarker discovery, I applied the 'stratification with weighting' method to reanalyze the original 5 microarray melanoma datasets to produce the new ranking gene list. Table 18 shows the top 200 genes of the new list.

Table 18: The 200 genes with highest ranking position after selection by 'stratification with weighting' method. Source dataset (Hoek et al., 2004, Smith et al., 2005, Riker et al., 2008, Scatolini et al., 2010, Rose et al., 2011)

No.	Genes	No.	Genes	No.	Genes	No.	Genes
1	DCD	51	IGFL2	101	BTBD16	151	TMEM154
2	MAGEA2B	52	CAPNS2	102	ANKRD35	152	COL11A1
3	KRT77	53	IL13RA2	103	GAGE12J	153	HLA-DRB3
4	THRSP	54	DST	104	GAGE2D	154	LASS3
5	CSAG2	55	APOC2	105	GAGE8	155	AQP3
6	C4orf7	56	MAL2	106	KRT25	156	RASGRF1
7	MAGEA3	57	DGAT2L3	107	HLA-DQB1	157	SH3RF2
8	MAGEA6	58	SCGB2A2	108	AKR1C3	158	HLA-DRB5
9	CTAG1A	59	ISG20	109	FGFR2	159	INS-IGF2
10	MAGEA2	60	MPP7	110	SERPINB5	160	PLA1A
11	WFDC5	61	CXCL13	111	CYP3A5	161	RAPGEFL1
12	MAGEA12	62	PTPRF	112	LEP	162	CYP26B1
13	FOXQ1	63	GAGE5	113	C6orf218	163	LAD1
14	ELMOD1	64	GAGE6	114	GPR87	164	FRMD5
15	ELOVL3	65	EYA1	115	RNASE2	165	KRT5
16	CSAG3	66	GAGE1	116	IGFBP5	166	C1orf116
17	SERPINA12	67	DKFZP434B061	117	TF	167	IL1F7
18	PRAME	68	XAGE1A	118	MGST1	168	LOC645323
19	DMKN	69	XAGE1B	119	SAA1	169	TRIM29
20	COL17A1	70	XAGE1C	120	ABCA13	170	MAGEA10
21	C1orf172	71	XAGE1D	121	EPHA3	171	HTN1
22	DGAT2	72	XAGE1E	122	ATP6V1C2	172	GAD1
23	IL8	73	PPP1R14C	123	SAA2	173	ACER1
24	DSC3	74	C19orf33	124	MIA	174	RDH12
25	MICALCL	75	GAGE3	125	AKR1C1	175	OLFM3
26	ZIC1	76	GATA3	126	RSPO1	176	HS6ST2
27	AADACL2	77	TACSTD2	127	SLAMF7	177	SPINK6
28	GAGE7	78	WNT4	128	GRHL3	178	FGF13
29	MAGEA1	79	KRT14	129	GPR115	179	COL9A3
30	FGFR3	80	CHST6	130	HEY1	180	LCP2
31	CTAG2	81	DSP	131	FLJ37464	181	NEBL
32	CTAG1B	82	RHBDL2	132	KRT6B	182	ZNF750
33	KRT15	83	SGPP2	133	TMEM79	183	LUM
34	GAGE12F	84	PIP	134	HLA-DQA1	184	SPRR2G
35	GAGE12G	85	SCARA5	135	SERPINA3	185	MAGEC1
36	GAGE12I	86	PPP2R2C	136	C10orf116	186	MMP1
37	GAGE2A	87	NMU	137	LOC124220	187	EREG
38	GAGE2B	88	SCEL	138	HLA-DRB4	188	SERPINB7
39	GAGE2C	89	GBP5	139	JUP	189	CXCL1
40	GAGE2E	90	ENTHD1	140	TKTL1	190	HLA-DRB1
41	GAGE4	91	GAGE12B	141	SCML4	191	XG
42	TNC	92	GAGE12C	142	SULF1	192	KIAA1881
43	FMN2	93	GAGE12D	143	IGSF9	193	KIAA1199
44	MGP	94	GAGE12E	144	TFPI2	194	LAMB4
45	SHC4	95	GAGE12H	145	COL1A1	195	LOXL3
46	FSTL5	96	APOC4	146	DLX1	196	NGFR
47	KRTAP19-1	97	IGF2	147	ALDH1A3	197	SDC1
48	HMGA2	98	IGHG1	148	LOC285986	198	DSCR8
49	ANK3	99	CCDC3	149	TRIM7	199	EXPH5
50	MYOZ2	100	EGFR	150	RGS4	200	LOR

3.2.6. Re-analysis of the previous 200 gene significant with the enhanced 'stratification with weighting' gene list

The new 200 genes list (table 18) generated by the 'stratification with weighting' method was compared with the old 200 genes list generated by modified method GWGS + GWRS (table 11, section 3.1.1), 91% of the genes were similar between the 2 lists. Thus, only 18 genes (or 9%) were found that were not common between the old and the enhanced lists. These 18 genes are displayed with green background on table 18. Interestingly, these 18 genes were all located in the lower ranks of the gene list. The genes with yellow background were transcription factors (table 18). The 18 genes of the old 200 gene list (table 19) have not been reported by any microarray or wet-lab studies to be associated with melanoma. Of the 18 new genes (see table 18 with green background) in the enhanced list, 6 were reported by at least one of the 16 microarray studies referred to previously (see table 21), and 7 were reported by wet-lab studies (table 22). A further four genes (i.e., *LUM, HLA-DRB1, CXCL1* and *NGFR*) have been validated in both microarray and wet-lab studies and are shown with yellow background in table 22. Another worthy of attention was that 17 of 18 new genes were located in the top 300 genes of previous ranking list, however, only 12 of 18 old genes were in the top 300 of new ranking list (Table 19 and 20). These illustrated that new genes derived from the 'stratification with weighting' method were still in higher positions of old list, but the old genes did not have the similar situation in the new list.

Table 19: The 18 genes present in the previous 200 gene list that are not included on the enhanced gene list

No.	OLD genes	old ranking position	new ranking position
1	HLA-DRB2	75	279
2	LOC100133484	76	250
3	LOC100133661	77	251
4	LOC100133811	78	252
5	LOC730415	79	253
6	ZNF749	80	254
7	AKR1C2	93	300
8	IGL@	106	465
9	IGHV4-31	162	651
10	TP63	163	532
11	CYP4B1	169	203
12	KLK11	173	207
13	CHP2	174	907
14	LOC100126583	187	769
15	CYP4F8	188	206
16	SCGB1D2	190	230
17	RORA	191	221
18	LGALS7	192	1259

Table 20: The positions of new 18 genes in previous and enhanced 200 gene lists

No.	NEW genes	new ranking position	old ranking position
1	COL1A1	145	247
2	HTN1	171	208
3	ACER1	173	205
4	RDH12	174	206
5	OLFM3	175	207
6	HS6ST2	176	209
7	SPINK6	177	211
8	LCP2	180	230
9	LUM	183	239
10	SPRR2G	184	220
11	CXCL1	189	331
12	HLA-DRB1	190	254
13	XG	191	237
14	KIAA1881	192	240
15	KIAA1199	193	227
16	LOXL3	195	243
17	NGFR	196	222
18	DSCR8	198	245

Table 21: The new genes validated by 16 independent microarray studies

Genes	In 16 Microarray Studies		
KIAA1199	Hoek *et al* , 2004		
NGFR	Hoek *et al* , 2004		
LCP2	Hoek *et al* , 2004	Koh *et al* , 2009	
HLA-DRB1	Hoek *et al* , 2004	Mandruzzato *et al* , 2006	
LUM	Hoek et al, 2004	Haqq *et al* , 2005	Alonso *et al* , 2007
CXCL1	Hoek *et al* , 2004	Haqq *et al* , 2005	Bogunovic *et al* , 2009

Table 22: The new genes validated by independent wet-lab experimental studies

Genes	**Studies**			
LUM	Vuillermoz *et al*, 2004	Sifaki *et al*, 2006	Radwanska *et al*, 2008	Brezillon *et al*, 2009
COL1A1	Lin *et al*, 2005			
HLA-DRB1	Luongo *et al*, 2004			
CXCL1	Dhawan & Richmond, 2002	Di Cesare *et al*, 2007	Botton *et al*, 2011	Sapoznik *et al*, 2012
NGFR	Radfar *et al*, 2006	Chan & Tahan, 2010		
LOXL3	Kirschmann *et al*, 2002			
MMA1	de Wit *et al*, 2002	de Wit *et al*, 2005		

The relative ranking positions of genes in these two gene lists was investigated using two comparisons. The first assessed the ranking position of common genes of the 2 gene lists, the second assessed the ranking position of the 12 melanoma significant genes in the new signaling network. The ranking positions of 18 common genes (at the enhanced list) that were reported by both microarray and wet-lab studies are shown in table 23. Table 24 shows the ranking position of 12 proposed biomarkers of melanoma in both gene lists.

Table 23 shows that there was no significant change in gene ranking position for 7 of the 18 common genes. Indeed there were identical ranking positions for 3

genes (*DCD, HMGA2,* and *TFPI2*), and 4 genes (*MAGEA2, MAGEA3, MIA,* and *ALDH1A3*) exhibited a slightly changed ranking position. Interestingly, 11 genes (>60% of total genes) were ranked at higher positions in the new list, suggesting the positions of meaningful genes were pushed up by 'stratification with weighting' method.

The similar situation was observed with genes of the 12 biomarkers for melanoma (Table 24), except that *CHP2* did not appear at all in new 200 genes list. While nine genes (75%) were ranked higher in the new 200 genes list. three genes (*EGFR, FGFR2, COL11A1*) were ranked slightly lower in the new gene list (2 to 4 positions lower). Importantly, of the four wet-lab validated genes (i.e., *SHC4, CXCL13, PTPRF*, and *COL11A1*), three of them were ranked higher by at least 8 positions. Thus, the melanoma network represented in Figure 12 was not changed by the substituted of these new 18 genes in the new list.

When combining the results of table 23 and 24, it was clear that these changes in ranking positions allowed us to see that the new 'stratification with weighting' method ranked the meaningful pathway genes at higher positions than previously. In this sense the new method appears to be more robustly able to detect reliable biomarkers of melanoma.

Table 23: Ranking position for the 18 substituted genes in the two 200 gene lists for melanoma

Genes	Ranking position in previous 200 gene list	Ranking position in new 200 gene list
DCD	1	1
MAGEA3	2	7
MAGEA2	3	10
CSAG2	6	5
MAGEA12	21	12
MAGEA6	22	8
PRAME	28	18
IL8	39	23
COL17A1	40	20
MAGEA1	47	29
HMGA2	58	58
TNC	66	42
MIA	112	124
HLA-DQB1	113	107
EPHA3	126	121
TFPI2	144	144
ALDH1A3	146	147
SERPINA3	156	135

Table 24: Ranking position of 12 melanoma biomarker genes in two 200 gene lists

Genes	Ranking position in previous 200 gene list	Ranking position in new 200 gene list
IL8	39	23
FGFR3	53	30
TNC	66	42
SHC4	68	45
PTPRF	70	62
CXCL13	72	61
WNT4	87	78
EGFR	98	100
FGFR2	105	109
PPP2R2C	122	86
COL11A1	148	152
CHP2	174	not included

4. DISCUSSION

This study first of all attempted to develop a new integrated bioinformatics approach for the identification of gene biomarkers based on multiple microarray data. The main contributions of this study to my original aim are summarized as follows:

1. I developed a new bioinformatics approach by applying ranking of gene significance to increase the application range.
2. I defined a new signature of 200 genes in human melanoma, and through the associations with melanoma driver genes, 12 melanoma biomarkers were identified. 4 of them were validated by laboratory experiments.
3. I developed a general framework approach to enhance the integrated analysis method by adding two factors ('stratification' and 'weighting'). The computational evaluation showed that the framework approach can identify a more meaningful signature than other methods.

The following sections describe the main discovery of this study.

4.1. Discovery and validation of a 12-genes biomarker of melanoma

A review of the literature reveals poor congruence between gene signatures generated by different microarray-based melanoma studies (John et al. 2008; Bittner et al. 2000; Tímár et al. 2010). Unsurprisingly therefore, microarray-

based melanoma gene biomarkers have had poor translation to clinical practice, and melanoma diagnosis is still based on clinical and histopathological features of the tumour (Schramm et al. 2011). To perform a meta-analysis on microarray gene expression data, Rhodes *et al.* (2002) introduced a model for combination of differentially-expressed genes based on their p-value in a statistical test (Rhodes et al. 2002). However, there are two significant limitations in its application to microarray data: (1) many microarray studies can contain small numbers of samples, for which the *p*-value can therefore be problematic, and (2) the p-values of a gene across different studies may have large variations. Thus, the smallest p-value may determine the outcome of S_p (effective significance of p-value). Here a new and universally-applicable method has been proposed to overcome some limitations of the Rhodes model. The core mathematic model proposed in this study measures firstly the significant gene ranking list in an individual dataset by the 'genome-wide relative significance' (GWRS), and then assesses the significant gene ranking list across multiple datasets by the 'genome-wide global significance' (GWGS).

In the GWRS, fold-change was used for this study. This can be changed to different test methods when datasets are suitable for the using situation. In this study, T-test, ANOVA, SAM, fold-change (formulas were showed in section 1.4.4) were considered as the methods for use in GWRS. After testing in five melanoma microarray studies, only fold-change is suitable for the microarray datasets situation. The other three methods depend on p-values, and so the smaller samples could not generate reliable p-values. However, fold-change relies on either fold-change increase or decrease, and because this method

does not request the smallest sample numbers it can be used for solving the limitation of using small number samples.

A comparison of my GWRS with the Rhodes et al method was conducted by applying it to two datasets (GSE3189 and GSE12391) (section 2.3.2). From the Sr values and Sp values of the corresponding top 50 genes, it was clear that Sp values could be controlled by the smallest p value group, risking so data imbalance can be easily produced. However, GWRS depends on *ranking positions* of genes and so avoids the imbalance problem and treats the two datasets more equally. Another advantage of GWRS is that applying 'fold-change' and ranking position instead p-value could avoid the influences of sample number limitation. For example the sample number limitation also appeared in this study; only two control samples are in someone dataset. Key to this situation, 'fold-change' was applied instead of p-values to show differences of gene expressions, and then ranking them to get sr values. Thus, comparing to previous p-value-dependent methods, GWGS and GWRS could be applied more widely.

The comparison between metastatic melanoma and normal skin was done to reveal the significant associated gene expression in melanoma to define the biomarker of metastatic melanoma. The effectiveness of this new approach can be supported by several lines of evidence and validation.

First, a considerable number of novel genes (e.g., *GTAG1A/1B/2, GAGE1-8/12B-J, XAGE1A-E, IL8, IGF2/INS-IGF2, SHC4, LEP, TF, CYP3A5*, TP63 and *GBP5*) revealed by our method were not identified as significant genes in the set of previous 16 melanoma microarray studies published between 2000-2011 used in this study, but none the less have still been confirmed as being melanoma-associated by independent 'wet-lab' studies in the literature (table 11).

Second, the new method identified a core signature of 12 genes (*i.e., EGFR, FGFR2, FGFR3, IL8, PTPRF, TNC, CXCL13, COL11A1, SHC4, CHP2, PPP2R2C* and *WNT4*) that are closely associated with known melanoma driver genes. Of note, however, six of these signature genes (i.e., *IL8, SHC4, COL11A1, CHP2, PPP2R2C* and *WNT4*) were not reported previously by microarray-based melanoma studies, although two (i.e. *IL8* and *SHC4*) have been identified in independent wet-lab studies (Zhang et al. 2011, Fagiani et al. 2007 and Pasini et al. 2009). This leaves *WNT4, CHP2, PPP2R2C* and *COL11A1* as genes which have not been previously reported to be associated with melanoma either via microarray or wet-lab studies. However, Fedida-Metula recently suggested a relationship between Ca2+ signaling members and *PP2A* and melanoma tumour growth (Fedida-Metula et al. 2012). Moreover *CHP2* (full name 'calcineurin-like EF hand protein') is involved in calcium signaling, while *PPP2R2C* is a member of the *PP2A* family.

Third, the expression of gene encoding the MAPK-associated members (i.e., COL11A1, CXCL13, PTPRF, SHC4) of the 12-gene biomarkers have been

validated in a comparative analysis of normal melanocytes and melanoma cells *in vitro* and in primary versus metastatic melanoma biopsy tissue *in situ* in this project. All four markers were found to be preferentially associated with melanoma, being differentially expressed in primary and metastatic melanoma. Strikingly, COL11A1, CXCL13, and PTPRF were not detectable in epidermal melanocytes of normal healthy human skin epidermis. SHC4 was expressed at only very low levels in normal epidermal melanocytes, as previously shown (Fagiani et al, 2007).

The over-expression of COL11A1, CXCL13, PTPRF, and SHC4 in melanoma cells *in vitro* and *in situ* may reflect the observed over-expression of the associated genes in our microarray meta-analysis results. The considerably higher level of SHC4 expression in the perikaryon of melanoma cells is of note, and concurs with other studies showing restricted expression in melanomas, while only weak expression in normal melanocytes and benign nevi (Fagiani et al, 2007). There is evidence that SHC4 is highly expressed at the transition from radial growth phase to vertical growth phase and metastatic melanomas, contemporaneous with the acquisition of melanoma migratory competence and invasive potential (Fagiani et al. 2007; Pasini et al. 2009). This protein tyrosine phosphatase acts as a signaling molecule to regulate cell growth, differentiation, mitotic cycle, and oncogenic transformation (Junta et al. 2008). PTPRF is usually expressed in the cell membrane (i.e. is a receptor-type protein tyrosine phosphatase) where it interacts with β-catenin. Like β-catenin, it may be translocated to the nucleus upon activation. The over-expression of COL11A1, CXCL13, PTPRF and SHC4 in our melanoma cell lines and primary and

metastatic tissue, and their potential association with MAPK signalling suggests these could be biomarkers for melanoma and so potential therapeutic targets. Based on the limitation of melanoma tissue samples, only these 4 genes were validated. The role of the others in the signature will be the focus of follow-on work from this thesis.

The computational evaluation conducted in this project also indicates that this new 12-gene biomarker signature achieved excellent diagnostic power in distinguishing primary and metastatic melanoma from normal skin. The integrated analysis of these five microarray datasets has identified a robust 12-gene biomarker signature that includes six previously-unreported genes in melanoma. Further experimental validation of the role of these 12 signature genes in a revised signaling network (Figure 12) may provide new insights into the underlying biological mechanisms driving the progression of melanoma. Moreover, given that the source 'original signatures' in this meta-analysis involved much larger numbers of genes (e.g., 589, 100, 65, 455 genes per signature), the excellent classification accuracy ratio performance achieved by our melanoma biomarker signature with just 12 genes is of note. This supports the view that our integrated approach extracts more informative genes than do the original signatures. From a clinical perspective our 12-gene signature could therefore be a more valuable biomarker for melanoma in the clinical setting. This will need to be followed up in further studies.

The method developed in this study was focused on gene *expression* association research, and was not directed to detecting gene *mutations*. It is

noted that *BRAF*, as one of main mutated gene for melanoma, was not shown in the list of differential expressed genes. This project focused on gene association not the gene mutation.

Five melanoma microarray datasets were applied in this study, as they all contain benign nevi/normal skin and metastatic melanoma samples (including cell line and biopsy). Because of the sample limitation, some datasets do not to contain normal skin or benign nevi sample. The normal skin and benign nevi samples were pooled together and considered as control group. Some microarray studies in the literature also investigated the relationship between benign nevi/normal skin and metastatic melanoma, e.g., some articles determined that significantly-different gene expression occurs between benign nevi and metastatic melanoma and so can define diagnosis biomarkers of metastatic melanoma (Kashani-Sabet et al., 2009; Smith et al., 2005). Others determined that metastasis and survival correlate with genes based on comparing combined benign nevi/normal skin with melanoma samples (Mandruzzato et al, 2006).

Table 25: The fold-change of 12 genes in five melanoma microarray datasets

	Gene	GSE4570	GSE4587	GSE7553	GSE12391	GSE22301
1	**EGFR**	-2.80597	-25.6657	-4.75	-1.54833	22.34352
2	**FGFR2**	3.168142	-1068.81	-6.01	-1.82438	1.90731
3	**FGFR3**	-1.39801	-232.953	-6.73	-1.55501	17.9909
4	**CXCL13**	-1.05814	60.14563	18.98	1.435844	2.721906
5	**COL11A1**	-37.162	4.71954	10.98	2.036713	1.29632
6	**WNT4**	-36.1569	-39.4524	-5.31	-1.98495	1.103384
7	**PTPRF**	3.810742	-3.44544	-3.01	-1.87123	32.91988
8	**PPP2R2C**	NA	-14.1279	-11.34	-1.20658	NA
9	**TNC**	191.1111	4.232415	8.74	4.000489	41.61225
10	**IL8**	-1.74774	130.9839	13.57	7.148044	527.4839
11	**CHP2**	1.658228	-352.38	-8.34	NA	1.7495
12	**SHC4**	NA	26.42012	2.24	2.237992	NA

The table 25 shows the fold change in expression for 12 marker genes between benign nevi/normal skin with metastatic melanoma in five individual microarray studies (green and yellow show the significant fold each gene). No one individual fold value could influence the finial meta-analysis result, no matter whether control sample is nevi or normal skin. This demonstrates that combining the significance of genes between benign nevi /normal skin with metastasis melanoma across multiple studies lead to the identification of gene biomarker of metastasis melanoma.

4.2. Enhancement for the computational method

As the microarray technique becomes increasingly popular, meta-analysis is been frequently applied to extract more information. However, meta-analysis still faces a critical concern i.e., that not all microarray data are of the same quality due to their use of various platforms and experimental set-ups. The quality of meta-analyses mainly depends on the quality of each individual microarray dataset (Larsson & Sandberg, 2006). Thus, the accuracy and reproducibility of microarray datasets have been a subject of some debate (Severgnini *et al.*, 2006).

To address this challenge, four meta-analysis methods were assessed in this study to investigate the potential impact of some of these intrinsic microarray weaknesses. Experimental results revealed that:

(1). The genes selected using the 'stratification with weighting' method (i.e., separating biopsy data from cell line data, and measuring the quality of datasets by *D* value, the ratio of each dataset's D value is as weighting of this dataset when calculated by GWRS across multiple datasets, see section 2.4.2) always achieved the good classification accuracy. This method also yielded the greatest number of common genes versus the other methods.

(2). The second next most effective enhancement method was 'non-stratification with weighting' method. This method also can retrieve relatively good classification accuracy and a high number of common genes.

(3). When one compares the results of 'stratification' and 'non-stratification' methods without weighting, it is clear that the results of 'stratification' method were much better than after 'non-stratification'.

(4). The new 200 ranked gene list which was detected by the 'stratification with weighting' method was also more meaningful. Although there were only 18 (of 200) genes that were different compared with the previous list, the new list generated by 'stratification with weighting' method did not lose any meaningful genes of the previous gene list, and also added some new validated genes into the gene list. The old 18 genes of previous list that did not appear in the new list furthermore have never been reported in any melanoma studies. However, some genes of newly added 18 genes have been reported in microarray studies or wet-lab studies or both melanoma studies. The other advantage of the improved method is that the ranking positions of the most meaningful genes were increased comparing with the previous positions.

In conclusion, the GWRS and GWGS methodologies could help deal with the limitation of sample numbers of datasets, and also merging individual datasets together to retrieve the higher relevant melanoma signature. However, the method is limited by that they did not fully recognize the influence of differences in datasets, like different type of platform, type of samples, quality of dataset etc. Based on these, the new stratification-based methodology provided a more robust set of gene associations in this integrated microarray analysis. In addition, combining either 'stratification' or 'non-stratification', with 'weighting' i.e., ω, always yielded better results versus the unweighted methodology. This means that it is very important to treat microarrays differently according to their data quality (the ratio of D value, see section 2.4.2). Both 'weighting' and 'stratification' are two important factors for enhancing the robustness of meta-analysis of microarray data. Combining the strengths of GWRS, GWGS, 'stratification' and 'weighting', a new framework for others in the field to consider was proposed, which combines these two useful factors together in an integrated analysis of microarray studies (Figure 20). In the new framework, the 'stratification with weighting' method was applied to a combined multi-study scenario, and when the gene list was screened by wrapper feature selection method with machine learning (figure 20) for getting robust genes. These proposed algorithm is defined in Fig.20 below with associated test following.

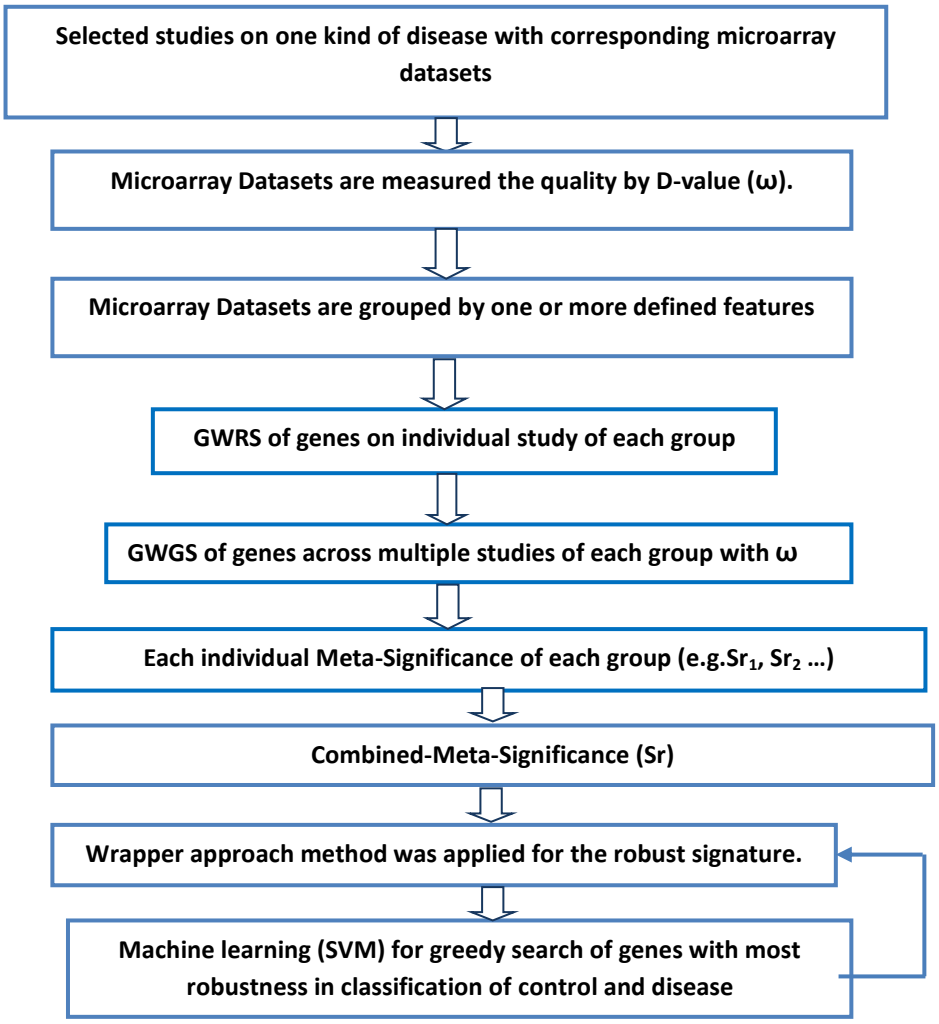


Figure 20: The proposed procedure of microarray meta-analysis to yield robust significant gene associations. This consists of two steps for characterizing the genome-wild global significance of genes, and an additional step for integrating the gene significance from biopsy and cell line samples, followed by a machine learning approach for the searching of robust genes.

The proposed new framework was applied for integrated analysis of microarray datasets. When applied to datasets of one disease, firstly each individual dataset should be measured for quality based its D-values, the quality value is used for distributing the importance ratios of all datasets. Then the datasets are grouped based on one or more different features, like sample type, microarray technique, number of genes or samples. After grouping, each individual study of

each group is calculated by GWRS, and combined by GWGS with weighting to generate the significant gene list of each group. Applying GWGS again to combine the significant gene lists of all groups, allows them to generate a final gene list for the disease. Wrapper feature selection with SVM method is then applied to detect the disease signature from the final list. The genes are inputted based on the in ranking order of the list and using a one-gene incremental addition approach. When the most robustness classification accuracy ratios are generated, the inputted genes are considered as the robust signature of the disease.

In summary, this study suggests that:

1. The proposed method combining individual studies together is able to reduce the false negatives comparing to individual analysis, and increase the effectiveness of statistical analysis of microarray datasets.
2. A proposed new signaling network for melanoma, which involve 12 new biomarker genes. 4 of them were validated in this study.
3. The new method adding stratification and wighting shows to be able to suit general applications of integrated analysis of microarray data.

It is, however, noted that this study is limited by two factors:

1. The number of samples of microarray datasets used in this study were small, which might have impact to the discovery of robust biomarkers .
2. 4 of 12 genes were validated in the biological experiments. Further experiments are needed for the validation of the remaining 8 genes.

5. CONCLUSION AND FUTURE WORK

Two significant contributions on the integrated analysis of microarray data were made in this thesis.

First, a new method to address the limitations of Rhodes' meta-analysis method was proposed. This new method was applied to the integrated analysis of five melanoma datasets and generated a new signature of melanoma containing 200 genes. Based on their interactions with four melanoma driver genes (*NRAS, BRAF, c-Kit,* and *MITF*), a new signaling network based on pathway analysis was created. This new signaling network includes 12 core genes from the 200 gene signature (*i.e., EGFR, FGFR2, FGFR3, IL8, PTPRF, TNC, CXCL13, COL11A1, SHC4, CHP2, PPP2R2C* and *WNT4*). These genes belong to three main signaling pathways (MAPK pathway, WNT pathway and Ca^{2+} pathway). Four of the 12 genes (i.e., *SHC4, CXCL13, COL11A1*, and *PTPRF*) link to the MAPK pathway and have been validated in wet-lab validations. In this study, the four genes all showed having strong relationship in biopsy and cell with melanoma, the other 8 genes will be focused in future work.

Second, the computational method was enhanced integrating the two factors, these factor are 'stratification' by sample type and 'weighting' by data quality to provide more flexibility in treating microarray data based on the biological nature of the samples and on the quality of dataset.

Stratification was the improved integrated analysis, and differences between each individual dataset in terms of platform, sample types, number of genes or probe-sets etc. were considered by this factor.

Weighting, which is the ratio of the D value of one dataset to the sum of D values for all datasets, improved the analysis as represented the quality of individual microarray datasets. The assessment of microarray dataset quality is very important part in combined analysis of multiple microarray datasets, because the quality of microarray data influences microarray experimental procedure directly, and will bring instability for the next level of the analysis.

The evaluations performed in this thesis clearly showed that the method including 'stratification' and 'weighting' together produces more robust results in biomarker discovery. It is noted that the 'stratification with weighting' method has produced different results than the 'non-stratification' or 'without weighting' method:

(1) The best accuracy ratios were obtained by 'stratification with weighting' method in Wrapper Feature selection with SVM evaluation. The next best method is 'Non-stratification with weighting'. It clearly showed that 'weighting' (the quality of dataset) is as the important factor in combined multiple datasets analysis.

(2) When evaluated the number of common genes between top gene lists of training datasets and the top gene list of individual testing dataset, the

'stratification with weighting' method was also the best method. It means that whatever the combined multiple datasets were changed, the method still can retrieve the same top genes, and these genes were more reliable genes in corresponding disease.

(3) When evaluated the number of common genes across different variations of combined training studies, the higher number of common genes also represented how stable the method and the top genes are. The 'stratification with weighting' method is still the best method in this evaluation.

(4) When applied the top gene lists which generated across multiple melanoma microarray datasets by the four methods to compare with 180 known metastatic melanoma biomarkers, found that 'stratification with weighting' was the second best method.

(5) Apply the 'stratification with weighting' method to combine the all five original melanoma datasets, and then the generated gene list compared to 180 metastatic melanoma biomarker, the number of common genes was the highest in the comparisons of the other gene lists with the 180 biomarkers. It also evidenced that more microarray datasets were integrated in the analysis more reliable biomarker can be identified.

(6) The 'stratification with weighting' method identified 18 new genes when compared the new 200 genes with the previous 200 gene list, and brought more validation genes into the new 200 gene list. The old 18 genes in previous gene list were never reported by any study, however, in the

enhanced list, 6 genes were reported by at least one microarray studies, and 7 genes were reported by wet-lab studies. Four genes were validated in both microarray and wet-lab studies.

(7) The most meaningful genes are ranked at higher positions. There were 18 common genes between previous and new gene list which were reported by both microarray and wet-lab studies. >60% of them (11 genes) presented higher positions in new list, and other 7 of them had identical or slightly changed ranking positions.

(8) Even though the 200 genes were changed, the network of melanoma was not influenced. The 11 of proposed 12 biomarkers based on the previous method were still in the top 200 genes list of the 'stratification with weighting' method. Of them, 9 genes (75%) were increased the positions. And the three of four wet-lab validated genes (*SHC4, CXCL13, PTPRF,* and *COL11A1*) were higher at least 8 positions than previous.

The results demonstrated that the 'stratification with weighting' method is able to increase the detection of true biomarker (true positives) more effectively.

Future work

In the future, the validation of 12 biomarkers of melanoma in wet-lab is in process, based on only 4 of them in MAPK pathway were validated by IHC and ICC and the limitations of samples, the four genes only were validated in normal, primary and metastatic samples, and the sample number was a few, the

gene progression have been not done. The next plans about the 12 biomarkers are:

(1) Before wet-lab validation, simulation of pathway for the 12 biomarkers would be done for computational validation. The method also be applied for the enhanced gene list.

(2) Continually validate the 12 biomarkers of melanoma in vitro studies, exam the progression of melanoma (normal → benign nevi → primary melanoma → metastatic melanoma) based on one donor ideally, but from mixed donor also possible if the number is large.

(3) The significance of the 12 genes based on the differential expression, no matter the expression is up or down. Especially for the down-regulated genes, the IHC and ICC may not help in validation. Based on that, knock down each gene of 12 biomarkers should be important validation. Through knock down each of them one by one to analysis whether the gene may influence others and how the gene works in the whole network.

For the new ‘stratification with weighting’ method and new proposed approach, future works could include the following.

(1) In this study, the datasets were stratified and analyzed based on differences in sample types (cell line vs. biopsy). The method still can be evaluated by

other types of stratification, for example, based on differences in platforms, differences in number of genes or probe-sets etc. And the stratification could be improved, like combined multiple stratified factors together, i.e., consider the differences of platform and sample types at the same time. The multiple factors combination could derive more elaborative classification and could improve robustness of microarray outcomes.

(2) The 'weighting' could be measured by different measurements. In this study, the *D*-value was applied to assess the quality of dataset, and it may be beneficial to include instead by other algorithms to represent different quality of datasets.

(3) The new 200 gene list will be analyzed more deeply, not only to compare with the previous 200 gene lists and the literature based 180 biomarkers. They will be analyzed in other ways (like pathway analysis) to indicate whether meaningful genes exist in the new list.

(4) ChIP-on-chip analysis should be considered for the biomarkers of melanoma in further research, based on this enhanced meta-analysis microarray method. ChiP-on-chip is a technique that can detect the interactions of proteins and DNA by combining chromatin immunoprecipitation (ChIP) and microarray technology (chip).

Publication and Presentation

1. Liu W., Peng Y., Tobin D.J. 2013. A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis. PeerJ 1:e49 http://dx.doi.org/10.7717/peerj.49

2. Liu W., Tobin, D.J., Peng, Y. 2013. Intefrated microarray analysis reveals *COL11A1*, *CXCL13,* and *PTPRF* as three new potential biomarkers of metastatic melanoma. Poster presented at the international investigative dermatology (IID) Edinburgh

REFERENCES

Agar, N., and Young, A. R. (2005). Melanogenesis: a photoprotective response to DNA damage? Mutation Research *571*, 121-132.

Alexander, D.D., Mink, P.J., Cushing, C.A., Sceurman, B. (2010). A review and meta-analysis of prospective studies of red and processed meat intake and prostate cancer. Nutrition journal *9*, 50.

Alonso, S.R., Tracey, L, Ortiz, P., Perez-Gomez, B., Palacios, J., Pollan, M., Linares, J., Serrano, S., Saez-Castillo, A.I., Sanchez, L, Pajares, R., Sanchez-Aguilera, A., Artiga, M.J., Piris, M.A., Rodriguez-Peralto, J.L. (2007). A high-throughput study in melanoma identifies epithelial-mesenchymal transition as a major determinant of metastasis. Cancer Res *67*, 3450-3460.

Altman, N.S. (1992). An introduction to kernel and nearest neighbor nonparametric regression. The American Statistician *46*, 175-185.

Amjadi, F., Javanmard, S.H., Zarkesh-Esfahani, H., Khazaei, M., Narimani, M. (2011). Leptin promotes melanoma tumour growth in mice related to increasing circulating endothelial progenitor cells numbers and plasma NO production J Exp Clin Cancer Res *30*, 21.

Baade, P.D., Youlden, D.R., Krnjacki, L.J. (2009). International epidemiology of prostate cancer: geographical distribution and secular trends. Molecular nutrition & food research *52*, 171-184.

Balakrishnan, A., Bleeker FE, Lamba S, Rodolfo M, Daniotti M, Scarpa A, van Tilborg AA, Leenstra S, Zanon C, Bardelli A. (2007). Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. Cancer Res *67*, 3545-3550.

Balch, C., Houghton, A., Sober, A., and Soong, S. (2003). Cutaneous Melanoma, 4th edn (St.Louis: Quality Medical Publishing).

Balch, C.M., Soong, S.J., Gershenwald, J.E., Thompson, J.F., Reintgen, D.S., Cascinelli, N., Urist, M., McMasters, K.M., Ross, M.I., Kirkwood, J.M., Atkins, M.B., Thompson, J.A., Coit, D.G., Byrd, D., Desmond, R., Zhang, Y., Liu, P.Y., Lyman, G.H., Morabito, A. (2001). Prognostic factors analysis of 17,600 melanoma patients: validation of the American joint committee on cancer melanoma staging system. J Clin Oncol *19*, 3622-3634.

Barak, V., Frenkel S, Kalickman I, Maniotis AJ, Folberg R, Pe'er J (2007). Serum markers to detect metastatic uveal melanoma. Anticancer Res *27*, 1897-1900.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H.,

Sherman, P.M., Muertter, R.N., Edgar, R. (2008). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Research *37*, D885-D890.

Bateman, A., Turner SJ, Theaker JM, Howell WM. (1998). HLA-DQB1*0303 and *0301 alleles influence susceptibility to and prognosis in cutaneous malignant melanoma in the British Caucasian population. Tissue Antigens *52*, 67-73.

Batton, T., Puissant, A., Cheli, Y., Tomic, T., Giuliano, S., Fajas, L., Deckert, M., Ortonne, J.P., Bertolotto, C., Tartare-Deckert, S., Ballotti, R., Rocchi, S. (2011). Ciglitazone negaribely regulates CXCL1 signalling throught MITF to suppress melanoma growth. Cell Death Differ *18*, 109-121.

Bazhin, A.V., Wiedemann, N., Schonolzer, M., Schadendorf, D., Eichmuller, S.B. (2007). Expression of GAGE family proteins in malignant melanoma. Cancer Lett *251*, 258-267.

Berwick, M. (2002). challenges in Melanoma (oxford: Blackwell Scientific Publishers).
Beuzeboc, P., Soulie, M., Richaud, P., Salomon, L, Staerman, F., Peyromaure, M., Mongiat-Artus, P., Cornud, F., Paparel, P., Davin, J.L., Molinie, V. (2009). Fusion genes and prostate cancer. Prog Urol *19*, 819-824.

Bilban, M., Buehler, L.K., Head, S., Desoye, G., Quaranta, V. (2002). Normalizing DNA microarry data. Curr Issues Mol Biol *4*, 57-64.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, M., Hendrix, M., Radmacher, R., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E. R., Wang, E., Marincola, F., Gooden, G., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., Trent, J. M.. Molecular (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406*, 536-540.

Blackburn, G.L. (2007). Dietary fat reduction and breast cancer outcome: resutls from the Women's Intervention Nutrition Study (WINS). The American journal of clinical nutrition *86*, s878-881.

Bogunovic, D., O'Neill, D.W., Belitskaya-Levy, I., Vacic, V., Yu, Y.L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A.C., Lonardi, S., Zavadil, J., Osman, I., Bhardwaj, N. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in prodicting patient survival. Proc Nati Acad Sci U S A *106*, 20429-20434.

Bolognia, J.L., Jorizzo, J.L., Schaffer, J.V. (2003). Dermatology, Vol 2, 2 edn (Mosby).

Bonvini, P., An, W.G., Rosolen, A., Nguyen, P., Trepel, J., Garcia de Herreros, A. (2001). Geldanamycin abrogates ErbB2 association with proteasome-resistant beta-catenin in melanoma cells, increases beta-catenin-E-cadherin

association, and decreases beta-catenin-sensitive transcription. Cancer Res *61*, 1671-1677.

Boone, B., Brochez L (2009). Clinical markers and driving mechanisms in melanoma progression: VEGF-C, RhoC, c-Ski/SnoN and EGFR. Verh K Acad Geneeskd Belg *71*, 251-294.
Boone, B., Jacobs K, Ferdinande L, Taildeman J, Lambert J, Peeters M, Bracke M, Pauwels P, Brochez L (2011). EGFR in melanoma: clinical significance and potential therapeutic target. J Cutan Pathol *38*, 492-502.

Bosserhoff, A., Dreau D, Hein R, Landthaler M, Holder WD, Buettner R (2001a). Melanoma inhibitory activity (MIA), a serological marker of malignant melanoma. Recent Results Cancer Res *158*, 158-168.

Bosserhoff, A., Echtenacher B, Hein R, Buettner R. (2001b). Functional role of melanoma inhibitory activity in regulating invasion and metastasis of malignant melanoma cells in vivo. Melanoma Res *11*, 417-421.

Bosserhoff, A., Golob M, Buettner R, Landthaler M, Hein R. (1998). MIA ("melanoma inhibitor acivity"). BIological functions and clinicl relevance in malignant melanoma. Hautarzt *49*, 762-769.

Bosserhoff, A., Kaufmann M, Kaluza B, Bartke I, Zirngibl H, Hein R, Stolz W, Buettner R. (1997). Melanoma-inhibiting activity, a novel serum marker for progression of malignant melanoma. Cancer Res *57*, 3149-3153.

Bosserhoff, A., Lederer M, Kaufmann M, Hein R, Stolz W, Apfel R, Bogdahn U, Buettner R (1999). MIA, a novel serum marker for progression of malignant melanoma. Anticancer Res *19*, 2691-2693.

Brandon, E.L., Gu, J.W., Cantwell, L., He, Z., Wallace, G., Hall, J.E. (2009). Obesity promotes melanoma tumour growth: role of leptin. . Cancer Biol Ther *8*, 1871-1879.

Brannon, H. (2004). Types of melanoma.
http://dermatology.about.com/od/melanoma/a/meltypes.htm

Brasseur, F., Rimoldi D, Lienard D, Lethe B, Carrel S, Arienti F, Suter L, Vanwijck R, Bourlond A, Humblet Y et al (1995). Expression of MAGE gene in primary and metastatic cutaneous melanoma. Int J Cancer *63*, 375-380.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P., Sansone, S.A. (2003). ArrayExpress-a public repository for microarray gene expression data at the EBI. Nucleic Acids Res *31*, 68-71.

Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett *573*, 83-92.

Breslow, A. (1970). Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. Annals of Surgery *172*, 902-908.

Brezillon, S., Zeltz C, Schneider L, Terryn C, Vuillermoz B, Ramont L, Perrau C, Pluot M, Diebold MD, Radwanska A, Malicka-Blaszkiewicz M, Maquart FX, Wegrowski Y. (2009a). Lumican inhibits B16F1 melanoma cell lung metastasis. J Physiol Pharmacol *60*, 15-22.

Brezillon, S., Zeltz, C., Schneider, L., Terryn, C., Vuillermoz, B., Ramont, L., Perrau, C., Pluot, M., Diebold, M.D., Radwanska, A., Malicka-Blaszkiewicz, M., Maquart, F.X., Wegrowski, Y. (2009b). Lumican inhibits B16F1 melanoma cell lung metastasis. J Physiol Pharmacol *60*, 15-22.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Nati Acad Sci USA *97*, 262-267.

Buchholz, T.A. (2009). Tadiation therapy for early-stage breast cancer after breast-conserving surgery. NEnglJMed *360*, 63-70.

Caballero, O., Zhao Q, Rimoldi D, Stevenson BJ, Svobodova S, Devalle S, Rohrig UF, Pagotto A, Michielin O, Speiser D, Wolchok JD, Liu C, Pejovic T, Odunsi K, Brasseur F, Van den Eynde BJ, Old LJ, Lu X, Cebon J, Strausberg RL, Simpon AJ. (2010). Frequent MAGE mutations in human melanoma. PLoS One *5*, e12773.

Cahan, P., Ahmad, A.M., Burke, H., Fu, S., Lai, Y., Florea, L., Dharker, N., Kobrinski, T., Kale, P., McCaffrey, T.A. (2005). List of lists-annotated (LOLA): A database for annotation and comparison of published microarray gene lists. Gene *360*, 78-82.

Cahan, P., Rovegno, F., Mooney, D., Newman, J.C., St. Laurent III, G., McCaffrey, T.A. (2007). Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. GENE *401*, 12-18.

Cao, M., Auge JM, Molina R, Marti R, Carrera C, Castel T, Vilella R, Conill C, Sanchez M, Malvehy J, Puig S (2007). Melanoma inhibiting activity protein (MIA), beta-2 microglobulin and lactate dehydrogenase (LDH) in metastatic melanoma. Anticancer Res *27*, 595-599.

Carretero, R., Wang, E., Rodrguez, A.I., Reinboth, J., Ascierto, M.L., Engle, A.M., Liu, H., Camacho, F.M., Marincola, F.M., Garrido, F., Cabrera, T. (2011). Regression of melanoma metastases after immunotherapy is associated with activation of antigen presentation and interferon-mediated rejection genes. Int J Cancer, doi:10.1002/ijc.26471.

Casey, T., Bond, J., Tighe, S., Hunter, T., Lintault, L., Patel, O., Eneman, J., Crocker, A., White, J., Tessitore, J., Stanley, M., Harlow, S., Weaver, D., Muss,

H., Plaut, K. (2009). Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. Breast Cancer Res Treat *114*, 47-62.

Chan, M.M., Tahan, S.R. (2010). Low-affinity nerve growth factor receptor (P75 NGFR) as a marker of perineural invasion in malignant melanomas. J Cutan Pathol *37*, 336-343.

Chandran, U.R., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., Michalopoulos, G., Becich, M., Monzon, F.A. (2007). Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer *7*, 64.

Chen, P., Murray TG, Salgaller ML, Ksander BR. (1997). Expression of MAGE genes in ocular melanoma cell lines. J Immunother *20*, 265-275.

Cheng, S., Huang Liu R, Sheu JN, Chen ST, Sinchaikul S, Tsay GJ (2006). Toxicogenomics of kojic acid on gene expression profiling of a375 human malignant melanoma cells. Bio Pharm Bull *29*, 655-669.

Choi, J.K., Yu, U., Kim, S., Yoo, O.J. (2003). Combining multiple microarray studies and modeling interstudy variation. Bioinformatics *19*, i84-i91.

Clark, W.J.J., Elder, D.E., Guerry, D.IV, Epstein, M.N., Greene, M.H., Van Horn, M. (1984). A study of tumour progression: the precursor lesions of superficial spreading and modular melanoma. Hum Pathol *15*, 1147-1165.

Cochran, J., Conn, V.S. (2008). Meta-analysis of quality of life outcomes following diabetes self-management training. The Diabetes Educator *34*, 815-823.

Constant, P. (1629). Les Oeuvres anatomiques de M. Jean Tiolan, Vol 1, 2nd edn (Paris: Denys Moreau).

Costin, G.E., Hearing, V.J. (2007). Human skin pigmentation: melanocytes modulate skin color in response to stress. The FASEB Journal *21*, 976-994.

Cui, X., Churchill, G.A. (2003). Statistical tests for differential expression in cDNA microarray experiments. Genome Biology *4*, 210.

de Wit, N.J., Cornelissen, I.M., Diepstra, J.H., Weidle, U.H., Ruiter, D.J., van Muijen, G.N. (2005). The MMA1 gene family of cancer-testis antigens has multiple alternative splice variants: characterization of their expression profile, the genomic organization, and transcript properties. Genes Chromosomes Cancer *42*, 10-21.

de Wit, N.J., Weidle, U.H., Ruiter, D.J., van Muijen, G.N. (2002). Expression profiling of MMA-1a and splice variant MMA-1b: new cancer/testis antigenes identified in human melanoma. Int J Cancer *98*, 547-553.

Delmas, V., Veermann, F., Martinozzi, S., Carreira, S., Ackermann, J., Kumasaka, M., Denat, L, Goodall, J., Luciani, F., Viros, A., Demirkan, N., Bastian, B.C., Goding, C.R., Larue, L. (2007). β-Catenin induces immortalization of melanocytes by suppressing p16INK4a expression and cooperates with N-Ras in melanoma development. Genes & Developement *27*, 2923-2935.

Dennis, L.K., Lynch, C.F., Torner, J.C. (2012). Epidemiologic association between prostatitis and prostate cancer. Urology *60*, 78-83.

Dhawan, P., Richmond, A. (2002). Role of CXCL1 in tumourigenesis of melanoma. J Leukoc Biol *72*, 9-18.

Di Cesare, S., Marshall, J.C., Logan, P., Antechka, E., Faingold, D., Maloney, S.C., Burnier, M.N. Jr. (2007). Expression and mifratory analysis of 5 human uveal melanoma cell lines for CXCL12, CXCL8, CXCL1 and HGF. J Carcinog *6*, 2.

Diaz-Lagares, A., Alegre E, Arroyo A, Gonzalez-Cao M, Zudaire ME, Viteri S, Martin-Algarra S, Gonzalez A (2011). Evaluation of multiple serum markers in advanced melanoma. Tumour Biol *32*, 1155-1161.

Diaz, A., Suarez E, Blanco R, Lopez A, Montero E (2009). Epidermal growth factor receptor modulates the tumourigenic potential of melanoma. Front Biosci *14*, 159-166.
Dimberg, J., Strom K, Logfren S, Zar N, Hugander A, Matussek A. (2011). Expression of the serine protease inhibitor serpinA3 in human colorectal adenocarcinomas. Oncol Lett *2*, 413-418.

Dimitropoulou, P., Lophatananon, A., Easton, D., Pocock, R., Dearnaley, D.P., Guy, M., Edwards, S., O'Brien, L., Hall, A., Wilkinson, R. (2009). Sexual activity and prostate cancer risk in men diagnosed at a younger age. BJU International *103*, 178-185.

Djulbegovic, M., Beyth, R.J., Neuberger, M.M., Stoffs, T.L., Vieweg, J., Djulbegovic, B., Dahm, P. (2010). Screening for prostate cancer: systematic review and meta-analysis of randomised controlled trials. BMJ *341*, c4543.

Dumitrascu, G., Constantin C, Manda G, Hristescu S, Margaritescu I, Chirta D, Neagu M (2009). Serum markers in skin melanoma--preliminary study. Roum Arch Microbiol Immunol *68*, 125-135.

Edgar, R., Domrachev, M., Lash, A.E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research *30*, 207-210.

Egland, K.A., Kumar, V., Duray, P., Pastan, I. (2002). Characterization of overlapping XAGE-1 transcripts encoding a cancer testis antigen expressed in lung, breast, and other types of cancers. Mol Cancer Ther *1*, 441-450.

Eigentler, T., Garbe, C. (2006). Malignant melanoma: classification and staging of malignant melanoma. Front Radiat Theor Oncol *39*, 149-158.

Eisenger, M., Marko, O. (1982). Selective proliferation of normal human melanocyte in vitro in the presence of phorbol ester and cholera toxin. Proceedings of the National Academy of Sciences of the USA *79*, 2018-2022.

Elias, P.M., Fritsch, P., Epstein, E.H. (1977). Staphylococcal scalded skin syndrome. Clinical features, pathogenesis and recent microbiological and biochemical development. Arch Dermatol *113*, 207-219.

Ellerhorst, J.A., Diwan, A.H., Dang, S.M., Uffort, D.G., Johnson, M.K., Cooke, C.P., Grimm, E.A. (2010). Promotion of melanoma growth by the metabolic hormone leptin. Oncol Rep *23*, 901-907.

Errington, J., Conway RM, Walsh-Conway N, Browning J, Freyer C, Cebon J, Madigan MC. (2012). Expressiong of cancer-testis antigens (MAGE-A1, MAGE-A3/6, MAGE-A4, MAGE-C1 and NY-ESO-1) in primary human uveal and conjunctival melanoma. Br J Ophthalmol *96*, 451-458.

Essler, M., Link A, Belloni B, Mirceva V, Souvatzoglou M, Thaler M, Haller B, Hein R, Krause BJ (2011). Prognostic value of [18F]-fluoro-deoxy-glucose PET/CT, S100 or MIA for assessment of cancer-associated mortality in patients with high risk melanoma. PLoS One *6*, e24632.

Fagiani, E., Giardina, G., Luzi, L., Cesaroni, M., Quarto, M., Capra, M., Germano, G., Bono, M., Capillo, M., Pelicci, P. (2007). RaLP, a new member of the Src homology and collagen family, regulates cell migration and tumour growth of metastatic melanomas Cancer Res *67*, 3064-3073.

Faries, M., Gupta RK, Ye X, Hsueh EC, Morton DL (2004). Melanoma-inhibiting activity assay predicts survival in patients receiving a therapeutic cancer vaccine after complete resection of American Joint Committee on Cancer Stage III Melanoma. Ann Surg Oncol *11*, 85-93.

Faries, M., Gupta RK, Ye X, Lee C, Yee R, Leopoldo Z, Essner R, Foshag LJ, Elashoff D, Morton DL (2007). A Comparison of 3 tumour markers (MIA, TA90IC, S100B) in stage III melanoma patients. Cancer Invest *25*, 285-293.

Fellenberg, F., Hartmann, T.B., Dummer, R., Usener, D., Schadendorf, D., Eichmüller, S. (2004). GBP-5 splicing variants: New guanylate-binding proteins with tumour-associated expression and antigenicity. J Invest Dermatol *122*, 1510-1517.

Feller, A., Duan Z, Penson R, Toh HC, Seiden MV (2000). TRAG-3, a novel cancer/testis antigen, is overexpressed in the majority of melanoma cell lines and malignant melanoma. Anticancer Res *20*, 4147-4151.

Field, S., Newton-Bishop, J.A. (2011). Melanoma and vitamin D. Mol Oncol. *5*, 197-214.

Fitzpatrick, T.B. (1988). The validity and practicality of sun-reactive skin types I althrough VI. ArchDermatol *124*, 869-871.

Fitzpatrick, T.B., Breathnach, A.S. (1963). The epidermal melanin unit system. Dermatol Wochenschr *147*, 481-489.

Flaherty, K.T., McArthur, G. (2010). BRAF, a target in melanoma: implications for solid tumour drug development. Cancer *116*, 4902-4913.

Florescu, A., Amir, E., Bouganim, N., Clemons, M. (2011). Immune herapy for breast cancer in 2010---hype or hope. Current Oncology *18*, e9-e18.

Freedberg, D.E., Rigas, S.H., Russak, J., Gai, W., Kaplow, M., Osman, I., Turner, F., Randerson-Moor, J.A., Houghton, A., Busam, K., Bishop, D.T., Bastian, B.C., Newton-Bishop, J.A., Polsky, D. (2008). Frequent p16-independent inactivation of p14arf in human melanoma. J Natl Cancer Inst *100*, 784-795.

Fukunaga-Kalabis, M., Martinez G, Nguyen TK, Kim D, Santiago-Walker A, Roesch A, Herlyn M (2010). Tenascin-C promotes melanoma progression by maintaining the ABCB5-positive side population. Oncogene *29*, 6115-6124.

Gage, M., Wattendorf, D., Henry, L.R. (2012). Translational advances regarding hereditary breast cancer syndromes. Journal of surgical oncology *105*, 444-451.

Gallagher, R.P., Fleshner, N. (1998). Prostate cancer: 3. Individual risk factors. CMAJ *159*, 807-813.

Gallagher, R.P., McLean D.I. (1995). The epidemiology of acquired melanocytic nevi. Dermoepidemiology *13*, 595-603.

Gann, P.H., Hennekens, C.H., Ma, J., Longcope, C., Stampfer, M.J. (1996). Prospective study of sex hormone levels and risk of prostate cancer. J Natl Cancer Inst *88*, 1118-1126.

Garbe, C., Leiter U, Ellwanger U, Blaheta HJ, Meier F, Rassner G, Schittek B (2003). Diagnostic value and prognostic significance of protein S-100beta, melanoma-inhibitory activity, and tyrosinase/MART-1 reverse transcription-polymerase chain reaction in the follow-up of high-risk melanoma patients. Cancer *97*, 1737-1745.

Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., Milner, D.A., Granter, S.R., Du, J., Lee, C., Wagner, S.N., Li, C., Golub, T.R., Rimm, D.L., Meyerson, M.L., Fisher, D.E., Sellers, W.R. (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117-122.

Gartside, M., Chen H, Ibrahimi OA, Byron SA, Curtis AV, Wellens CL, Bengston A, Yudt LM, Eliseenkova AV, Ma J, Curtin JA, Hyder P, Harper UL, Riedesel E, Mann GJ, Trent JM, Bastian BC, Meltzer PS, Mohammadi M, Pollock PM (2009). Loss-of-function fibroblast growth factor receptor-2 mutations in melanoma. Mol Cancer Res *7*, 41-54.

Gibbs, P., Hutchins AM, Dorian KT, Vaughan HA, Davis ID, Silvapulle M, Cebon JS. (2000). MAGE-12 and MAGE-6 are frequently expressed in malignant melanoma. Melanoma Res *10*, 259-264.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher *5*, 3-8.

Gledhill, K., Rhodes, L.E., Brownrigg, M., Haylett, A.K., Masoodi, M., Thody, A.J., Nicolaou, A., Tobin, D.J. Prostaglandin-E2 is produced by adult human epidermal melanocytes in response to UVB in a melanogenesis-independent manner. Pigment Cell Melanoma Res *23*, 394-403.

Gledhill K., R., L.E., Brownrigg, M., Haylett, A.K., Masoodi, M., Thody, A.J., Nicolaou, A., Tobin, D.J. (2010). Prostaglandin-E2 is produced by adult human epidermal melanocytes in response to UVB in a melanogenesis-independent manner. Pigment Cell Melanoma Res *23*, 394-403.

Goel, V.K., Lazar, A.J., Warneke, C.L., Redston, M.S., Haluska, F.G. (2006). Examination of mutations in BRAF, NRAS and PTEN in primary cutaneous melanoma. J Invest Dermatol *126*, 154-160.

Grahovac, J., Becker D, Wells A. (2012). Melanoma cell invasiveness is promoted at least in part by the epidermal growth factor-like repeats of tenascin-C. J Invest Dermatol *doi: 10.1038/jid.2012.263.*

Gremel, G., Rafferty, M., Lau, T.Y.K., Gallagher, W.M. (2009). Identification and functional validation of therapeutic targets for malignant melanoma. Critical Reviews in Oncology/Hematology *72*, 194-214.

Grichnik, J.M., Burch, J.A., Schulteis, R.D., Shan, S., Liu, J., Darrow, T.L., Vervaert, C.E., Seigler, H.F. (2006). Melanoma, a tumour based on a mutant stem cell. Journal of Investigative Dermatology *126*, 142-153.

Guba, M., Bosserhoff AK, Steinbauer M, Abeis C, Anthuber M, Buettner R, Jauch KW (2000). Overexpression of melanoma inhibitory activity (MIA) enhances extravasation and metastasis of A-mel 3 melanoma cells in vivo. Br J Cancer *83*, 1216-1222.

Guntur, A.R., Reinhold, M.I., Cuellar, J.Jr., Naski, M.C. (2011). Conditional ablation of Pten in osteoprogenitors stimulates FGF signaling. Development *138*, 1433-1444.

Haass, N.K., Smalley, K.S., Li, L., Herlyn, M. (2005). Adhesion, migration and communication in melanocytes and melanoma. Pigment Cell Research *18*, 150-159.

Hackstadt, A.J., Hess, A.M. (2009). Filtering for increased power for microarray data analysis. BMC Bioinformatics *10*, 11.

Haddadeen, C., Lai, C., Cho, Healy, E. (2014). Variants of the melanocortin-1 receptor: Do they matter clinically? Experimental Dermatology.

Halaban, R. (1996). Growth factors and melanomas. Semin Oncol *23*, 673-681.

Halachmi, S., Gilchrest, B.A. (2001). Update on genetic events in the pathogenesis of melanoma. Curr Opin Oncol *13*, 120-136.

Hankey, B.F., Feuer, E.J., Clegg, L.X., Hayes, R.B., Legler, J.M., Prorok, P.C., Ries, L.A., Merrill, R.M., Kaplan, R.S. (1999). Cancer surveillance series: interpreting trends in prostate cancer- part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. J Natl Cancer Inst *91*, 1017-1024.

Haqq, C., Nosrati, M., Sudilovsky, D., Crothers, J., Khodabakhsh, D., Pulliam, B.L., Federman, S., Miller, J.R. 3rd, Allen, R.E., Singer, M.I., Leong, S.P., Ljung, B.M., Sagebiel, R.W. (2005). The gene expression signatures of melanoma progression. Proc Nati Acad Sci U S A *102*, 6092-6097.

Harland, M., Goldstein, A.M., Kukalizch, K., Taylor, C., Hogg, D., Puig, S., Badenas, C., Gruis, N., Huurne, J.t., Bergman, W., Hayward, N.K., Stark, M., Tsao, H., Tucker, M.A., Landi, M.T., Scarra, G.B., Ghiorzo, P., Kanetsky, P.A., Elder, D., Mann, G.J., Holland, E.A., Bishop, D.T., Newton-Bishop, J. (2008). A comparison of CDKN2A mutation detection within the Melanoma Genetics Consortium (GenoMEL). Eur J Cancer *44*, 1269-1274.

Heenen, M., Giacomoni, P.U., and Golstein, P. (2001). Individual variations in the correlation between erythermal threshold, UV-induced DNA damage and sun-burn cell formation. JPhotochemPhotobiol B *63*, 84-87.

Hinkelmann, K., Kempthorne, O. (2008). Design and analysis of experiments. In (Wiley).

Hochberg, M., Lotem M, Gimon Z, Shiloni E, Enk CD (2002). Expression of tyrosinase, MIA and MART-1 in sentinel lymph nodes of patients with malignant melanoma. Br J Dermatol *146*, 244-249.

Hocker, T., Tsao, H. (2007). Ultraviolet radiation and melanoma: a systematic review and analysis of reported sequence variants. Hum Mutat *28*, 578-588.

Hoek, K., Rimm, D.L., Williams, K.R., Zhao, H., Ariyan, S., Lin, A., Kluger, H.M., Berger, A.J., Cheng, E., Trombetta, E.S., Wu, T., Niinobe, M., Yoshikawa, K.,

Hannigan, G.E. (2004). Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. Cancer Res *64*, 5270-5282.

Hoek, K.S. (2007). DNA microarray analysis of melanoma gene expression: a decade in the mines. Pigment Cell Res *20*, 466-484.

Hofmann, M., Schicke B, Frisch A, Biesold S, Gussmann F, Kuchler I, Voit C, Trefzer U (2011). Impact of lymph node metastases on serum level of melanoma inhibitory activity in stage III melanoma patients. J Dermatol *38*, 880-886.

Honari, S. (2004). Topical therapies and antimicrobials in the management of burn wounds. Crit Care Nurs Clin North Am *16*, 1-11.

Hong, E.J., Park, S.H., Choi, K.C., Leung, P.C., Jeung, E.B. (2006). Identification of estrogen-regulated genes by microarray analysis of the uterus of immature rats exposed to endocrine disrupting chemicals. Reprod Biol Endocrinol *4*, 49.

Hong, F., Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics *24*, 374-382.

Hosack, D.A., Dennis, G.Jr., Sherman, B.T., Lane, H.C. Lempicki, R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol *4*, R70.

Huang, T., Rauth S, Das Gupta TK (1996). Overexpression of EGF receptor is associated with spontaneous metastases of a human melanoma cell line in nude mice. Anticancer Res *16*, 3557-3563.

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., Reddy, T.B.K., Wymore, F., Zachariah, Z.K., Sherlock, G., Ball, C.A. (2008). Implementation of genepattern with the stanford microarray database. . Nucleic Acids Research *37*, D898-D901.

Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T., Tateno, Y. (2003). CIBEX: center for information biology gene EXpression database. Comptes Rendus Biologies *326*, 1079-1082.

Inozume, T., Mitsui, H., Okamoto, T., Matsuzaki, Y., Kawakami, Y., Shibagake, N., Shimada, S. (2007). Dendritic cells transduced with autoantigen FCRLA induce cytotoxic lymphocytes and vaccinate against murine B-cell lymphoma. J Invest Dermatol *127*, 2818-2822.

Institute, S. (2012). Base SAS 9.3 procedures guide: statistical procedures

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. In Reviews in computational chemistry, K.B. Lipkowitz, Cundari, T.R. (Wiley-VCH: Weinheim).

Jaeger, J., Koczan, D., Thiesen, H.J., Ibrahim, S.M., Gross, G., Spang, R., Kunz, M. (2007). Gene expression signatures for tumour progression, tumour subtype, and tumour thickness in laser-microdissected melanoma tissues. Clinical Cancer Research *13*, 806-815.

Janjic, B., Andrade P, Wang XF, Fourcade J, Almunia C, Kudela P, Brufsky A, Jacobs S, Friedland D, Stoller R, Gillet D, Herberman RB, Kirkwood JM, Maillere B, Zarour HM (2006). Spontaneous CD4+ T cell responses against TRAG-3 in patients with melanoma and breast cancers. J Immunol *177*, 2717-2727.

Janner, R.G., Young, R.A. (2005). Insights into host responses against pathogens from transcriptional profiling. NatRevMicrobiol *3*, 281-294.

Jeffs, A.R., Glover, A.C., Slobbe, L.J., Wang, L., He, S., Hazlett, J.A., Awasthi, A., Woolley, A.G., Marshall, E.S., Joseph, W.R., Print, C.G., Baguley, B.C., Eccles, M.R. (2009). A gene expression signatures of invasive potential in metastatic melanoma cells. PLoS One *4*, e8461.

Johansen Taber, K.A., Morisy, L.R., Osbahr, A.J. 3rd., Dickinson, B.D. (2010). Male breast cancer: risk factors, diagnosisand management. Oncol Rep *24*, 1115-1120.

John, T., Black, M.A., Toro, T.T., Leader, D., Gedye, C.A., Davis, I.D., GUilford, P.J., Cebon, J.S. (2008). Predicting clinical outcome through molecular profiling in stage III melanoma. Clin Cancer Res *14*, 5173-5180.

Johnson, K.C., Miller, A.B., Collishaw, N.E., Palmer, J.R., Hammond, S.K., Salmon, A.G., Cantor, K.P., Miller, M.D., Boyd, N.F., Millar, J., Turcotte, F. (2011). Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk. Tobacco control *20*, e2.

Jönsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringnér, M., Lillehaug, J.R., Borg, A., Lønning, P.E. (2010). Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. Clin Cancer Res *16*, 2256-2267.

Juergensen, A., Holzapfel U, Hein R, Stolz W, buettner R, Bosserhoff A (2001). Comparison of two prognostic markers for malignant melanoma: MIA and S100 beta. Tumour Biol *22*, 54-58.

Junta, C.M., Sandrin-Garcia. P., Fachin-Saltoratto, A.L., Mello, S.S., Oliveira, R.D., Rassi, D.M., Giuliatti, S., Sakamoto-Hojo, E.T., Louzada-Junior, P., Donadi, E.A., Passos, G.A. (2008). Different gene expression of peripheral blood mononuclear cells from rheumatoid arthritis patients may discriminate immunogenetic, pathogenic and treatment features. Immunology *127*, 365-372.

Jurman G, M.S., Barla A, Paoli S, Galea A, Furlanello C. (2008). Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics *24*, 258-264.

Kabbarah, O., Nogueira, C., Feng, B., Nazarian, R.M., Bosenberg, M., Wu, M., Scott, K.L., Kwong, L.N., Xiao, Y., Cordon-Cardo, C., Granter, S.R., Ramaswamy, S., Golub, T., Duncan, L.M., Wagner, S.N., Brennan, C., Chin, L. (2010). Integrative genome comparison of primary and metastatic melanomas. PLoS One *5*, e10770.

Kageshita, T., Naruse T, Hirai S, Ono T, Horikoshi T, Nakagawa H, Tamaki K, Hayashibe K, Ichihashi M, Nakayama J, Hori Y, Ozawa A, Miyahara M, Ohkido M, Inoko H. (1997). Molecular genetic analysis of HLA class II alleles in Japanese patients with melanoma. Tissue Antigens *49*, 466-470.

Kahn, C.E. (2006). malignant melanoma:staging (Medical College of Wisconsin).

Kaidbey, K.H., Agin, P.P., Sayre, R.M., and Kligman, A.M. (1979). Photoprotection by melanin - a comparison of black and Caucasian skin J Am Acad Dermatol *1*, 249-260.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research *40*, D109-D114.

Kashani-Sabet, M., Venna, S., Nosrati, M., Rangel, J., Sucker, A., Egberts, F., Baehner, F.L., Simko, J., Leong, S.P., Haqq, C., Hauschild, A., Schadendorf, D., Miller, J.R. 3rd (2009). A multimarker prognostic assay for primary cutaneous melanoma. Clin Cancer Res *15*, 6987-6992.

Kauffmann, A., Gentleman, R., Huber, W. (2009). arrayQualityMetrics-a bioconductor package for quality assessment of microarray data. Bioinformatics *25*, 415-416.

Key, T.J. (2011). Fruit and vegetables and cancer risk. British journal of cancer *104*, 6-11.

Khan, J., Saal, L.H., Bittner, M.L., Chen, Y., Trent, J.M., Meltzer, P.S. (1999). Expression profiling in cancer using cDNA microarrays. Electrophoresis *20*, 223-229.

Kirschmann, D.A., Seftor, E.A., Fong, S.F., Nieva, D.R., Sullivan, C.M., Edwards, E.M., Sommer, P., Csiszar, K., Hendrix, M.J. (2002). A molecular role for lysyl oxidase in breast cancer invasion. Cancer Res *62*, 4478-4483.

Klingenstein, A., Haritoglou I, Schaumberger MM, Nentwich MM, Hein R, Schaller UC (2011). Receiver operating characteristic analysis: calculation for the marker 'melanoma inhibitory activity' in metastatic uveal melanoma patients. Melanoma Res *21*, 352-356.

Kluger, H., Hoyt K, Bacchiocchi A, Mayer T, Kirsch J, Kluger Y, Sznol M, Ariyan S, Molinaro A, Halaban R (2011). Plasma markers for identifying patients with metastatic melanoma. Clin Cancer Res *17*, 2417-2425.

Kodama, Y., Kaminuma, E., Saruhashi, S., Ikeo, K., Sugawara, H., Tateno, Y., Nakamura, Y. (2010). Biological databases at DNA Data Bank of Janpan in the era of next-generation sequencing technologies. . In Advances in computational biology, J.R. Arabnia, ed., pp. 125-135.

Koh, S.S., Opel, M.L., Wei, J.P., Yau, K., Shah, R., Gorre, M.E., Whitman, E., Shitabata, P.K., Tao, Y., Cochran, A.J., Abrishami, P., Binder, S.W. (2009). Molecular classification of melanomas and nevi using gene expression microarray signatures and formalin-fixed and paraffin-embedded tissue. Mod Pathol *22*, 538-546.

Koh, S.S., Wei, J.P., Li, X., Huang, R.R., Doan, N.B., Scolyer, R.A., Cochran, A.J., Binder, S.W. (2012). Differential gene expression profiling of primary cutaneous melanoma and sentinel lymph node metastases. Mod Pathol *2012.32*

Krenacs, T., Kiszner G, Stelkovics E, Balla P, Teleki I, Nemeth I, Varga E, Korom I, Barbai T, Plotar V, Timar J, Raso E (2012). CSollagen XVII is expressed in malignant but not in benign melanocytic tumours and it can mediate antibody induced melanoma apoptosis. Histochem cell Biol *138*, 653-667.

Krzyminiewski, R., Kruczyński, Z., Dobosz, B., Zając, A., Mackiewicz, A., Leporowska, E., Folwaczna, S. (2011). EPR study of iron ion complexes in human blood. Appl Magn Reson *40*, 321-330.

Kunz, M. (2008). DNA microarray technology in dermatology. Semin Cutan Med Surg *27*, 16-24.

Kunz, M.I., S.M., Koczan, D., Scheid, S., Thiesen, H.J., Gross, G. (2004). DNA microarray technology and its applications in dermatology. Experimental Dermatology *13*, 593-606.

Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L., Kohane, I.S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics *18*, 405-412.

LaBonne, C., Bronner-Fraser, M (1998). Induction and patterning of the neural crest, a stem cell like precursor population. Journal of Neurobiology *36*, 175-189.

Larsson, O., Sandberg, R. (2006a). Lack of correct data format and comparability limits future integrative microarray research. NatBiotechnol *24*, 1322-1323.

Larsson, O., Wennmalm, K., Sandberg, R. (2006b). Comparative microarray analysis. Omics *10*, 381-397.

Lawrence, J.E., Cook, N.J., Rovin, R.A., Winn, R.J. (2012). Leptin promotes glioblastoma. . Neurol Res Int *2012:870807: Epub 2012 Jan 4.*

Leboit, P., Burg, G., Weedon, D., Sarasin, A. (2006). WHO classification of tumours, pathology & genetics, skin tumours (Lyon: IARC Press).

Lee, J., Lu M, Mansfield PF, Platsoucas CD, Reveille JD, Ross MI. (1996). Malignant melanoma: relationship of the human leukocyte antigen class II gene DQB1*0301 to disease recurrence in American joint committee on cancer stage I or II. Cancer *78*, 758-763.

Lee, J., Reveille JD, Ross MI, Platsoucas CD. (1994). HLA-DQB1*0301 association with increased cutaneous melanoma risk. Int J Cancer *59*, 510-513.

Lejeune, F.J. (1997). The management of malignant melanoma revisited. Acta Chirurgica Belgica *97*, 209-214.

Lens, M. (2008). Current clinical overview of cutaneous melanoma. Br J Nurs *17*, 300-305.

Lethe, B., Lucas, S., Michaux, L., De Smet, C., Godelaine, D., Serrano, A., De Plaen, E., Boon, T. (1998). LAGE-1, a new gene with tumour specificity. Int J Cancer *76*, 903-908.

Lin, A.Y., Maniotis, A.J., Valyi-Nagy, K., Majumandar, D., Setty, S., Kadkol, S., Leach, L., Pe'er, J., Folberg, R. (2005). Distinguishing fibrovascular septa from vasculogenic mimicry patterns. Arch Pathol Lab Med *129*, 884-892.

Lin, J., Hocker, T.L., Singh, M., Tsao, H. (2008). Genetics of melanoma predisposition. The British Journal of Dermatology *159*, 286-291.

Lin, J.Y., Fisher, D.E. (2007). Melanocyte biology and skin pigmentation. NATURE *445*, 843-850.

Lisabeth, E., Fernandez C, Pasquale EB (2012). Cancer somatic mutations disrupt functions of the EphA3 receptor tyrosine kinase through multiple mechanisms. Biochemistry *51*, 1464-1475.

Lister, S. (2009). Urine test could speed treatment of prostate cancer (London: The Sunday Times).

Liu, R., Wang, XH., Chen , G.Y., Dalerba, P., Gurney, A, Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., Clarke, M.F. (2007). The prognostic role of a gene signature from tumourigenic breast-cancer cells. The NEW ENGLAND JOURNAL of MEDICINE *356*, 217-226.

Liu, S., Ren S, Howell P, Fodstad O, Riker, A.I. (2008). Identification of novel epigenetically modified genes in human melanoma via promoter methylation gene profiling. Pigment Cell Melanoma Res *21*, 545-558.

Liu, W., Peng, Y., Tobin, D.J. (2013). A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis. PeerJ *1*, e49.

Lugovic, L., Situm M, Buljan M, Poduje S, Sebetic K (2007). Results of the determination of serum markers in patients with malignant melanoma. Coll Antropol *31 Suppl 1*, 7-11.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., Brazma, A. (2010). A global map of human gene expression. Nat Biotechnol *28*, 322-324.

Luo, Y., Dallaglio, K., Chen, Y., Robinson, W.A., Robinson, S.E., McCarter, M.D., Wang, J., Gonzalez, R., Thompson, D.C., Norris, D.A., Roop, D.R., Vasiliou, V., Fujita, M. (2012). ALDH1A isozymes are markers of human melanoma stem cells and potential therapeutic targets. . Stem Cellls *30*, 2100-2113.

Luongo, V., Pirozzi, G., Caraco, C., Errico, S., de Angelis, F., Celentano, E., Paino, F., Chiofalo, M.G., Luongo, M., Mozzillo, N., Lombardi, M.L. (2004). HLA allele frequency and clinical outcome in Italian patients with cutaneous melanoma. Tissue Antigens *64*, 84-87.

Luyten, G., van der Spek CW, Brand I, Sintnicolaas K, de Waard-Siebinga I, Jager MJ, de Jong PT, Schrier PI, Luider TM. (1998). Expression of MAGE, gp100 and tyrosinase genes in uveal melanoma cell lines. Melanoma Res *8*, 11-16.

Madison, K.C. (2003). Barrier function of the skin:"La Raison d'Etre" of the epidermis. THe Jounal of Investigative Dermatology *121*, 231-241.

Makarov, D.V., Desai, R.A., Yu, J.B., Sharma, R., Abraham, N., Alvertsen, P.C., Penson, D.F., Gross, C.P. (2012). The population level prevalence and correlates of appropriate and inappropriate imaging to stage incident prostate cancer in the medicare population. The Journal of Urology *187*, 97-102.

Mallikarjuna, K., Pushparaj V, Biswas J, Krishnakumar S (2007). Expression of epidermal growth factor receptor, ezrin, hepatocyte growth factor, and c-Met in uveal melanoma: an immunohistochemical study. Curr Eye Res *32*, 281-290.

Mandruzzato, S., Callegaro, A., Turcatel, G., Francescato, S., Montesco, M.C., Chiarion-Sileni, V., Mocellin, S., Rossi, C.R., Bicciato, S., Wang, E., Marincola, F.M., Zanovello, P. (2006). A gene expression signature associated with survival in metastatic melanoma. J Transl Med *4*, 1-11.

Mann, G.J., Pupo, G.M., Campain, A.E. (2010). BRAF mutation, NRAS mutation and absence of an immune-related expressed gene profile predict poor outcome in surgically resected stage III melanoma. Pigment Cell Melanoma Res *23:6*.

McCarroll, S.A., Murphy, C.T., Zou, S., Pletcher, S.D., Chin, C.S., Jan, Y.N., Kenyon, C., Barqmann, C.I., Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet *36*, 197-204.

Mecham, B.H., Klus, G.T., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D.Z., Mariani, T.J., Kohane, I.S., Sallasi, Z. (2004). Sequence-matched probes produce increased corss-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Research *32*, e74.

Medley, Q.G., Buchbinder, E.G., Tachibana, K., Ngo, H., Serra-Pages, C., Streuli, M. (2003). Signaling between gocal adhesion kinase and trio. The journal of Biological Chemistry *278*, 13265-13270.

Meral, R., Duranyildiz D, Tas F, Camlica H, Yasasever V, Kurul S, Dalay N (2001). Prognostic significance of melanoma inhibiting activity levels in malignant melanoma. Melanoma Res *11*, 627-632.

Meyskens, F.L.J., McNulty, S.E., Buckmeier, J.A., Tohidian, N.B., Spillane, T.J., Kahlon, R.S., and Gonzalez, R.I. (2001). Aberrant redox regulation in human metastatic melanoma cells compared to normal melanocytes. Free Radical Biology and Medicine *31*, 799-808.

Miller, A.J., Mihm, M.C. (2006). Mechanisms of Diseases: Melanoma. The NEW ENGLAND JOURNAL of MEDICINE *355*, 51-65.

Miyamoto, J., Berkowitz, Z., Jones, S.E., Saraiya, M. (2012). Indoor tanning device use among male high school students in the United States. J Adolesc Health *50*, 308-310.

Mlakar, V., Glavac, D. (2007). DNA microarrays and their use in dermatology. Acta Dermatoven APA *16*, 7-12.

Monsel, G., Ortonne, N., Bagot, M., Bensussan, A., Dumaz, N. (2009). c-Kit mutants require hypoxia-inducible factor 1alpha to transform melanocytes. Oncogene *29*, 227-236.

Morrison, M.A., Silveira, A.C., Huynh, N., Jun, G., Smith, S.E., Zacharaki, F., Sato, H., Loomis, S., Andreoli, M.T., Adams, S.M., Radeke, M.J., Jelcick, A.S., Yuan, Y., Tsiloulis, A.N., Chatzoulis, D.Z., Silvestri, G., Kotoula, M.G., Tsironi, E.E., Hollis, B.W., Chen, R., Haider, N.B., Miller, J.W., Farrer, L.A., Hageman, G.S., Kim, I.K., Schaumberg, D.A., DeAngelis, M.M. (2011). Systems biology-based analysis implicates a novel role for vitamin D metabolism in the pathogenesis of age-related macular degeneration. Hum Genomics *5*, 538-568.

Mosch B, P.D., Pietzsch J (2012). Irradiation affects cellular properties and Eph receptor expression in human melanoma cells. Cell Adh Migr *6*, 113-125.

Mulcahy, K., Rimoldi D, Brasseur F, Rodgers S, Lienard D, Marchand M, Rennie IG, Murray AK, Mclntyre CA, Platts KE, Leyvraz S, Boon T, Rees RC. (1996). Infrequent expression of the MAGE gene family in uveal melanomas. Int J Cancer *66*, 738-742.

Murakami, M., Suzuki, M., Nishino, Y., Funaba, M. (2010). Regulatory expression of genes related to metastasis by TGF-beta and activin A in B16 murine melanoma cells. Mol Biol Rep *37*, 1279-1286.

Nanni, S., Priolo, C., Grasselli, A., D'Eletto, M., Merola, R., Moretti, F., Gallucci, M., De Carli, P., Sentinelli,S., Cianciulli, A.M., Mottolese, M., Carlini, P., Arcelli, D., Helmer-Citterich, M., Gaetano, C., Loda, M., Ponetecorvi, A., Bacchetti, S., Sacchi, A., Farsetti, A. (2006). Epithelial-restricted gene profile of primary cultures from human prostate tumours: a molecular approach to predict clinical behavior of prostate cancer. Mol Cancer Res *4*, 79-92.

Nestle, F.O., Kerl, H. (2003). Melanoma. In Dermatology, J.L. Golognia, Jorizzo, J., Rapini, R.P., Horn, T.D., Mascaro,J.M., Mancini,A.J., Salasche, S.J., Saurat, J.H., Stingl,G. (London: Mosby), pp. 1789-1814.

Newton-Bishop, J., Chang, Y., Elliott, F., Chan, M., Leake, S., Karpavicius, B., Haynes, S., Fitzgibbon, E., Kukalizch, K., Randerson-Moor, J., Elder, D.E., Bishop, D.T., Barrett, J.H. (2011). Relationship between sun exposure and melanoma risk for tumours in different body sites in a large case-control study in a temperate climate. Eur J Cancer *47*, 732-741.

Newton-Bishop, J., Gruis, N. (2010). Melanoma susceptibility genes. Melanoma Research *20*, 161-162.

Newton-Bishop, J., Lorigan, P. (2012). Identifying melanomas in primary care: can we do better? BMJ *345*, e4244.

Newton-Bishop, J.A., Chang, Y., Iles, M.M., Taylor, J.C., Bakker, B., Chan,M., Leake, S., Karpavicius, B., Haynes, S., Fitzgibbon, E., Elliott, F., Kanetsky, P.A., Harland, M., Barrett, J.H., Bishop, D.T. (2010). Melanocytic nevi, nevus genes, and melanoma risk in a large case-control study in the United Kingdom. Cancer Epidemiol Biomarkers Prev *19*, 2043-2054.

Normand, S.L. (1999). Tutorial in biostatistics-Meta-analysis: formulating, evaluating, combining, and reporting. Statist Med *18*, 321-359.

Okamoto, I., Pirker, C., Bilban, M., Berger, W., Losert, D., Marosi, C., Haas, O.A., Wolff, K., Pehamberger, H. (2005). Seven novel and stable translocations associated with oncogenic gene expression in malignant melanoma. Neoplasia *7*, 303-311.

Ortonne, J.P., Prota, G. (1993). Hair melanins and hair color: ultrastructural and biochemical aspects. Journal of Investigative Dermatology *101*, 82S-89S.

Pan, F., Chiu, C.H., Pulapura, S., Mehan, M.R., Nunez-Iglesias, J., Zhang, K., Kamath, K., Waterman, M.S., Finch, C.E., Zhou, X.J. (2007). Gene aging nexus: a web database and data mining platform for microarray data on aging. Nucleic Acids Res *35*, D756-759.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics *18*, 546-554.

Park, W.D., Stegall, M.D. (2007). A meta-analysis of kidney microarray datasets: investifation of cytokine gene detection and correlation with RT-PCR and detection threshold. BMC Genomics *8*, 88.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Lilja, P.,Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., Brazma, A. (2007). ArrayExpress-a public database of microarray experiments and gene expression profiles. Nucleic Acids Research *35*, D747-D750.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandiai, M., Sklyar, N., Brazma, A. (2011). ArrayExpress update - an archive of microarray and high-throughput sequencing - based functional genomics experiments. Nucleic Acids Res *39*, D1002-D1004.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G.G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S., Brazma, A. (2005). ArrayExpress-a public repository for microarray gene expression data at the EBI. Nucleic Acids Res *33*, D553-D555.

Pasini, L., Turco, M.Y., Luzi, L., Aladowicz, E., Fagiani, E., Lanfrancone, L. (2009). Melanoma: targeting signaling pathways and RaLP. Expert Opin Ther Targets *13*, 93-104.

Peng, Y.H. (2006). A novel ensemble machine learning for robust microarray data classification. International Journal of Computers in Biology and Medicine, Elsevier *36*, 553-573.

Peng, Y.H., Wu, Z.Q., Jiang, J. (2010). A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, Elsevier *43*, 15-23.

Perez, R., Zhang P, Bosserhoff AK, Buettner R, Abu-Hadid M. (2000). Expression of melanoma inhibitor activity in melanoma and nonmelanoma tissue specimens. Hum Pathol *31*, 1381-1388.

Phung, B., Sun, J., Schepsky, A., Steingrimsson, E., Ronnstrand, L. (2011). C-KIT signaling depends on microphthalmia-associated transcription factor for effects on cell proliferation. PLoS ONE *6*, e24064.

Piepkorn, M. (2000). Melanoma genetics: an update with focus on the CDKN2A(p16)/ARF tumour suppressors. . J Am Acad Dermatol *42*, 705-722.

Pollock, P.M., Trent, J.M. (2000). The genetics of cutaneous melanoma. . Clin Lab Med *20*, 667-690.

Poole, C.M., and Guerry IV, D. (2005). Melanoma: Prevention, detection, and treatment., 2nd edn (New Haven and London: Yale University Press).

Porras, B.H., Cockerell, C.J. (1997). Cutaneous melignant melanoma: classification and clinical diagnosis. Semin Cutan Med Surg *16*, 88-96.

Porrerf, S.B., Virador, V., Wakamatsu, K., Furumura, M., Santis, C., ITO, S., Hearing, V.J. (1998). Cysteine transport in melanosomes from murine melanocytes. Pigment Cell Research *12*, 4-12.

Proksch, E., Brandner, J.M., Jensen, J.M. (2008). The skin: an indispensable barrier. Experimental Dermatology *17*, 1063-1072.

Proksch, E., Holleran, W.M., Menon, G.K., Elias, P.M., Feingold, K.R. (1993). Barrier function regulates epidermal lipid and DNA synthesis. Br J Dermatol *128*, 473-482.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nature Genetics Supplement *32*, 496-501.

Radfar, A., Stefanato, C.M., Ghosn, S., Bhawan, J. (2006). NGFR-positive desmoplastic melanomas with focal or absent S-100 staining: Further evidence supporting the use of both NGFR and S-100 as a primary immunohistochemical panel for the diagnosis of desmoplastic melanomas. Am J Dermatopathol *28*, 162-167.

Radwanska, A., Baczynska D, Nowak D, Brezillon S, Popow A, Maquart FX, Wegrowski Y, Malicka-Blaszkiewicz M. (2008). Lumican affects actin cytoskeletal organization in human melanoma A375 cells. Life Sci *83*, 651-660.

Ramasamy, A., Mondry, A., Holmes, C.C., Altman, D.G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Medicine *5*, 1320-1332.

Reiniger, I., Schaller UC, Haritoglou C, Hein R, Bosserhoff AK, Kampik A, Mueller AJ (2005). "Melanoma inhibitory activity" (MIA): a promising serological tumour marker in metastatic uveal melanoma. Graefes Arch Clin Exp Ophthalmol *243*, 1161-1166.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., Chinnaiyan, A.M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Research *62*, 4427-4433.

Richardson, A.L., Wang, Z.C., Nicolo, A.D., Lu, X., Brown, M., Miron, A., Liao, X., Lalehart, J.D., Livingston, D.M., Ganesan, S. (2006). X Chromosomal abnormalities in basal-like human breast cancer. CANCER CELL *9*, 121-132.

Rieg, S., Garbe C, Sauer B, Kalbacher H, Schittek B (2004). Dermcidin is constitutively produced by eccrine sweat glands and is not induced in epidermal cells under inflammatory skin conditions. Br J Dermatol *151*, 534-539.

Riker, A.I., Enkemann, S.A., Fodstad, O., Liu, S., Ren, S., Morris, C., Xi, Y., Howell, P., Metge, B., Samant, R.S., Shevde, L.A., Li, W., Eschrich, S., Daud, A., Ju, J., Matta, J. (2008). The gene expression profiles of primary and metastatic melanoma yields a transition point of tumour progression and metastasis. BMC Med Genomics *1:13*.

Rose, A.E., Poliseno, L., Wang, J., Clark, M., Pearlman, A., Wang, G., Vega Y Saenz de Miera, E.C., Medicherla, R., Christos, P.J., Shapiro, R., Pavlick, A., Darvishian, F., Zavadil, J., Polsky, D., Hernando, E., Ostrer, H., Osman, I (2011). Integrative genomics identifies molecular alterations that challenge the linear model of melanoma progression. Cancer Res *71*, 2561-2571.

Sapoznik, S., Ortenberg, R., Galore-Haskel, G., Kozlovski, S., Levy, D., Avivi, C., Barshack, I., Cohen, C.J., Besser, M.J., Schachter, J., Markel, G. (2012). CXCR1 as a novel target for directing reactive T cells toward melanoma: implications for adoptive cell transfer immunotherapy. Cancer Immunol Immunother *61*, 1833-1847.

Sariego, J. (2010). Breast cancer in the young patient. The American surgeon *76*, 1397-1401.

Scatolini, M., Grand, M.M., Grosso, E., Venesio, T., Pisacane, A., Balsamo, A., Sirovich, R., Risio, M., Chiorino, G. (2010). Altered molecular pathways in melanocytic lesions. Int J Cancer *126*, 1869-1881.

Schaller, U., Bosserhoff AK, Neubauer AS, Buettner R, Kampik A, Mueller AJ (2002). Melanoma inhibitory activity: a novel serum marker for uveal melanoma. Melanoma Res *12*, 593-599.

Schmiddt, J., Riechers A, Stoll R, Amann T, Fink F, Spruss T, Gronwald W, Konig B, Hellerbrand C, Bosseroff AK. (2012). Targeting melanoma metastasis and immunosuppression with a new mode of melanoma inhibitory activity (MIA) protein inhibition. PLoS One *7*, e37941.

Schmitz, C., Brenner W, Henze E, Christophers E, Hauschild A. (2000). Comparative study on the clinical use of protein S-100B and MIA (melanoma inhibitory activity) in melanoma patients. Anticancer Res *20*, 5059-5063.

Scholes, A., Hagan S, Hiscott P, Damato BE, Grierson I (2011). Overexpression of epidermal growth factor receptor restricted to macrophages in uveal melanoma. Arch Ophthalmol *119*, 373-377.

Schramm, S.J., Campain, A.E., Scolyer, R.A., Yang, Y.H., Mann, G.J. (2011). Review and cross-validation of gene expression signatures and melanoma prognosis. J Invest Dermatol *132*, 274-283.

Schubbert, S., Shannon, K., Bollag, G. (2007). Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer *7*, 295-308.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H. (2000). Normalization strategies for cDNA microarrays. Nucleic Acids Res *28*, E47.

Setlow, R.B., Grist, E., Thompson, K., Woodhead, A.D. (1993). Wavelengths effective in induction of malignant melanoma. PNAS *90*, 6666-6670.

Severgnini, M., Bicciato, S., Mangano, E., Scarlatti, F., Mezzelani, A., Mattioli, M., Ghidoni, R., Peano, C., Bonnal, R., Viti, F., Milanesi, L., De Bellis, G., Battaglia, C. (2006). Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. Anal Biochem *353*, 43-56.

Shannon, J., Tewoderos, S., Garzotto, M., Beer, T.M., Derenick, R., Palma, A., Farris, P.E. (2005). Statins and prostate cancer risk: a case-control study. Am J Epidemiol *162*, 318-325.

Sherwood, L. (2007). Human physiology: from cells to systems, 7th edn (london: Cengage learning).

Siegel, R. (2011). Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths.CA Cancer J Clin *61*, 212-236.

Sifaki, M., Assouti M, Nikitovic D, Krasagakis K, Karamanos NK, Tzanakakis GN. (2006). Lumican, a small leucine-rich proteoglycan substituted with keratan sulfate chains is expressed and secreted by human melanoma cells and not normal melanocytes. IUBMB Life *58*, 606-610.

Slominski, A., and Mihm, M.C. (1996). Potential mechanism of skin response to stress. Int J Dermatol *35*, 849-851.

Slominski, A., and Wortsman, J. (2000). Neuroendocrinology of the skin. Endocr Rev *21*, 457-487.

Smith, A.P., Hoek, K., Becker, D. (2005). Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. Cancer Biology & Therapy *4*, 1018-1029.

Soares, M.R., Huber, J., Rios, A.F., Ramos, E.S. (2010). Investigation of IGF2/Apal and H19/Rsal polymorphisms in patients with cutaneous melanoma. Growth Horm IGF Res *20*, 295-297.

Soikkeli, J., Lukk M, Nummela P, Virolainen S, Jahkola T, Katainen R, Harju L, Ukkonen E, Saksela O, Holtta E (2007). Systematic search for the best gene expression markers for melanoma micrometastasis detection. J Pathol *213*, 180-189.

Stahlecker, J., Gauger A, Bosserhoff A, Buttner R, Ring J, Hein R. (2000). MIA as a reliable tumour marker in the serum of patients with malignant melanoma. Anticancer Res *60*, 5041-5044.

Steinberg, G.D., Carter, B.S., Beaty, T.H., Childs, B., Walsh, P.C. (1990). Family history and the risk of prostate cancer. Prostate *17*, 337-347.

Struewing, J.P., Hartge, P., Wacholder, S., Baker, S.M., Berlin, M., McAdams, M., Timmerman, M.M., Brody, L.C., Tucker, M.A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. N Engl J Med *336*, 1401-1408.

Sulaimon, S.S., Kitchell, B.E. (2003). The biology of melanocytes. Veterinary Dermatology *14*, 57-65.

Svobodova, S., Browning, J., MacGregor, D., Pollara, G., Scolyer, R.A., Murali, R., Thompson, J.F., Deb, S., Azad, A., Davis, I.D. (2011). Cancer-testis antigen expression in primary cutaneous melanoma has independent prognostic value comparable to that of breslow thichness, ulceration and mitotic rate. Eur J Cancer *47*, 460-469.

Tanemura, A., Terando AM, Sim MS, van Hoesel AQ, de Maat MF, Morton DL, Hoon DS (2009). CpG island methylator phenotype predicts progression of malignant melanoma. Clin Cancer Res *15*, 1801-1807.

Tas, F., Yasasever V, Duranyildiz D, Camlica H, Ustuner Z, Aydiner A, Topuz E (2004). Clinical value of protein S100 and melanoma-inhibitory activity (MIA) in malignant melanoma. Am J Clin Oncol *27*, 225-228.

Taylor, K.L., Lister, J.A., Zeng, Z., Ishizaki, H., Anderson, C., Kelsh, R.N., Jackson, I.L., Patton, E.E. (2011). Differentiated melanocyte cell division occurs in vivo and is promoted by mutations in Mitf. Development *138*, 3579-3589.

Tímár, J., Gyorffy, B., Rásó, E. (2010). Gene signature of the metastatic potential of cutaneous melanoma: too much for too litter? Clin Exp Metastasis *27*, 371-387.

Tobin, D.J. (2006). Biochemistry of human skin - our brain on the outside. Chemical Society Reviews *35*, 52-67.

Tobin, D.J. (2008). Human hair pigmentation--biological aspects. Int J Cosmet Sci *30*, 233-257.

Tomlinson, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M., Kalyana-Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J., Shah, R.B., Chinnaiyan, A.M. (2007). Integrative molecular concept modeling of prostate cancer progression. NATURE GENETICS *39*, 41-51.

Topcu-Yilmaz, P., Kiratli H, Saglam A, Soylemezoglu F, Hascelik G (2010). Correlation of clinicopathological parameters with HGF, c-Met, EGFR, and IGF-1R expression in uveal melanoma. Melanoma Res *20*, 126-132.

Tronov, V.A., Artamonov, D.N., Gorbacheva, L.B. (2010). Genetic markers of melanoma. Russian Journal of Genetics *46*, 146-156.

Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C., Wong, W.H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucl Acids Res *29*, 2549-2557.

Turashvili, G., Bouchal, J., Baumforth, K., Wei, W., Dziechciarkova, M., Ehrmann, J., Klein, J., Fridman, E., Skarda, J., Srovnal, J., Hajduch, M., Murray, P., Kolar, Z (2007). Novel markers for differentation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. BMC Cancer *7*, 55.

Tusher, V.G., Tibshirani, R., Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. PNAS *98*, 5116-5121.

Ueno, Y., Sakurai H, Tsunoda S, Choo MK, Matsuo M, Koizumi K, Saiki I (2008). Heregulin-induced activation of ErbB3 by EGFR tyrosine kinase activity promotes tumour growth and metastasis in melanoma cells. Int J Cancer *123*, 340-347.

Usener, D., Gerhardt, A., Schadendorf, D., Eichmuller, S. (2003). Seroreactivity against MAGE-A and LAGE-1 proteins in melanoma patients. Br J Dermatol *149*, 282-288.

Varambally, S., Yu, J., Laxman, B., Rhodes, D.R., Mehra, R., Tomlins, S.A., Shah, R.B., Chandran, U., Monzon, F.A., Becich, M.J., Wei, J.T., Pienta, K.J., Ghosh, D., Rubin, M.A., Chinnaiyan, A.M. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. CANCER CELL *8*, 393-406.

Vourc'h-Jourdain, M., Volteau C, Nguyen JM, Khammari A, Dreno B. (2009). Melanoma gene expression and clinical course. Arch Dermatol Res *301*, 673-679.

Vucetic, B., Rogan SA, Hrabac P, Hudorovic N, Cupic H, Lukinac L, Ledinsky M, Matejcic A, Lovricevic I, Zekan M. (2008). Biological value of melanoma inhibitory activity serum concentration in patients with primary skin melanoma. Melanoma Res *18*, 201-207.

Vuillermoz, B., Khoruzhenko A, D'Onofrio MF, Ramont L, Venteo L, Perreau C, Antonicelli F, Maquart FX, Wegrowski Y. (2004). The small leucine-rich proteoglycan lumican inhibits melanoma progression. Exp Cell Res *296*, 294-306.

Wai, W.C., Dye, D.E., Coombe, D.R. (2012). The role of immunoglobulin superfamily cell adhesion molecules in cancer metastasis. Int J Cell BIol, 2012:340296.

Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A., Frierson, H.F., Jr., Hampton, G.M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res *61*, 5974-5978.

Wennmalm, K., Wahlestedt, C., Larsson, O. (2005). The expression signature of in vitro senescence resembles mouse but not human aging. Genome Biol *6*, R109.

Westekemper, H., Karimi S, Susskind D, Anastassjou G, Freistuhler M, Meller D, Zeschnigk M, Steuhl KP, Bornfeld N, Schmid KW, Grabellus F (2010). Expression of MCSP and PRAME in conjunctival melanoma. Br J Ophthalmol *94*, 1322-1327.

Westerhof, W. (2006). The Discovery of the Human Melanocytes. Pigment cell Res *19*, 183-193.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O, Landsman, D., Lipman, D.J., Madden, T.L.., Maglott, D.R., Osteel, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., Yaschenko, E. (2007). Database resources of the national center for biotechnology information. . Nucleic Acids Research *35*, D5-D12.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S.R., Avril, M.F., Ortiz Romero, P.L., Robert, T., Balacescu, O., Eggermont, A.M., Lenoir, G., Sarasin, A., Tursz, T., van den Oord, J.J., Spatz, A. (2006). Gene expression profiling of primary cutaneous melanoma and clinical outcome. J Natl Cancer Inst *98*, 472-482.

Wren, J.D. (2009). A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. Bioinformatics *doi: 10.1093/bioinformatics/btp290*

Yadav, V., Zhang X, Liu J, Estrem S, Li S, Gong XQ, Buchanan S, Henry JR, Starling JJ, Peng SB. (2012). Reactivation of mitogen-activated protein kinase (MAPK) pathway by FGF receptor 3 (FGFR3)/Ras mediates resistance to vemurafenib in human B-Raf V600E mutant melanoma. J Biol Chem *287*, 28087-28098.

Yang, L., Jacobsen, K.H. (2008). A systematic review of the association between breastfeeding and breast cancer. Journal of women's health *17*, 1635-1645.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research *30*, e15

Yue, H., Eastman, P.S., Wang, B.B., Minor, J., Doctolero, M.H., Nuttall, R.L., Stack, R., Becker, J.W., Montgomery, J.R., Vainer, M., Johnston, R. (2001). An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucl Acids Res *20*, E41.

Zabierowski, S.E., Fukunaga-Kalabis, M., Li, L., Herlyn, M. (2011). Dermis-derived stem cells: a source of epidermal melanocytes and melanoma? Pigment Cell Melanoma Res *24*, 422-429.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., Weir, B.S. (2002). Truncated product method for combining P-values. Gene Epidemiol *22*, 170-185.

Zendman, A.J., van Kraats, A.A., den Hollander, A.I., Weidle, U.H., Ruiter, D.J., van Muijen, G.N. (2002). Characterization of XAGE-1b, a short major transcript of cancer/testis-associated gene XAGE-1, induced in melanoma metastasis. Int J Cancer *97*, 195-204.

Zhang, H., Fu, T., McGettigan, S., Kumar, S., Liu, S., Speicher, D., Schuchter, L., Xu, X. (2011). IL8 and cathepsin B as melanoma serum biomarkers. Int J Mol Sci *12*, 1505-1518.

Zhang, W., Shannon, W.D., Duncan, J., Scheffer, G.L., Scheper, R.J., McLeod, H.L. (2006). Expression of drug pathway proteins is independent of tumour type. J Pathol *209*, 213-219.

Zouboulis, C.C. (2000). Human skin: an independent peripheral endocrine organ. Horm Res *54*, 230-242.

Appendix:

This appendix includes the literature search I conducted with the terms 'microarray datasets; and melanoma' in PubMed published for the years 2000-2011. Of these 22 studies, 16 reported data on gene signatures that shared little commonality or overlap between studies.

2000

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Den-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., Trent, J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406,* 536-*540.*

Bittner et al. (2000) used 38 samples (7 controls and 31 melanomas) to detect gene expressions using an 8,150 cDNA microarray. They identified 19 melanomas as a major cluster and found the down-regulated genes of the major cluster related to cell migration, especially decreased expressions in integrin β1, β3 or α1, syndecan 4 and vinculin. Over-expressed genes were outside of the major gene cluster and were relevant to fibronectin.

2003:

Zuidervaart, W., van der Velden, P.A., Hurks, M.H., van Nieuwpoort, F.A., Out-Luiting, C.J.J., Singh, A.D., Frants, R.R., Jager, M.J., Gruis, N.A. (2003). Gene expression profiling identifies tumour markers potentially playing a role in uveal

melanoma development. British Journal of Cancer *89*, 1914-1919.

Zuidervaart et al. (2003) detected 15 highly differentially expressed genes (>1.5 fold higher) in a microarray comparing 12 human uveal melanoma cell lines with 3 uveal melanocyte cultures. Four candidate genes were selected as tumour markers to discriminate 19 primary uveal melanoma samples into two classes. This may indicate differential uveal melanoma development processes.

Dooley, T.P., Curto, E.V., Davis, R.L., Grammatico, P., Robinson, E.S., Wilborn, T.W. (2003). DNA microarray and likelihood ratio bioinformatics methods: discovery of human melanocyte biomarkers. Pigment Cell Res *16*, 245-253.

Dooley et al. (2003) investigated 25 significant biomarkers of normal melanocytes by comparing normal human epidermal melanocytes with one primary melanoma cell line (MS7) and one metastatic melanoma cell line (SKMel-28). They advised some biomarkers could be potential molecular targets for diagnostics and drug discovery in melanoma.

2004:

Hoek, K., Rimm, D.L., Williams, K.R., Zhao, H., Ariyan, S., Lin, A., Kluger, H.M., Berger, A.J., Cheng, E., Trombetta, E.S., Wu, T., Niinobe, M., Yoshikawa, K., Hannigan, G.E. (2004). Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. Cancer Res *64*, 5270-5282.

Hoek et al, 2004 detected 589 significantly expressed genes by Affymetrix U133A dataset based on a 2.5 fold change in melanoma. Of these, 315 were up-regulated and 274 were down-regulated between normal melanocytes

and melanoma cells. They reported some novel pathways and expression in melanoma cells, like NOTCH pathway activation, altered expression in embryonic development and epidermal transition transcriptional regulators, activation of cancer antigens, and down-regulated growth suppressors including NECDIN.

Mirmahammadsadegh, A., Baer, A., Nambiar, S., Bardenheuer, W., Hengge, U.R. (2004). Rapid identification of dysregulated genes in cutaneous malignant melanoma metastases using cDNA technology. Cell Tissues Organs *117*, 119-123.

Mirmohammadsadegh et al, 2004 detected a series of differential gene expressions through comparing total RNA of melanoma metastases and primary human melanocytes from 10 patients. They selected five genes (including GRB10, BAX, BAD, GSTT1, GSR) as examples and suggested that the significant genes may be used as targets to provide therapeutic guidance.

McDonald, S.L., Edington, H.D., Kirkwood, J.M., Becker, D. (2004). Expression analysis of genes identified by molecular profiling of VGP melanomas and MGP melanoma-positive lymph nodes. Cancer Biology & Therapy *3*, 110-120.

McDonald et al., 2004 identified 5 known genes (ST13, CST-8, DKC1, NESP55, and NP-C2) and 1 unknown gene (16.7 kD) which could play important roles in advanced stage melanoma by examining gene expression patterns between primary melanoma and melanoma-positive lymph node specimens.

2005:

Smith, A.P., Hoek, K., Becker, D. (2005). Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanoma. Cancer Biol Ther *4*, 1018-1929.

Haqq, C., Nosrati, M., Sudilovsky, D., Crothers, J., Khodabakhsh, D., Pulliam, B.L., Federman, S., Miller, J.R. 3rd, Allen, R.E., Singer, M.I., Leong, S.P., Ljung, B.M., Sagebiel, R.W. (2005). The gene expression signatures of melanoma progression. Proc Nati Acad Sci U S A *102*, 6092-6097.

Smith et al, 2005 and Haqq et al. (2005) conducted a comprehensive study on different stages of malignant melanoma development based on whole genome expression profiles. Smith et al, 2005 investigated the top 50 up-regulated and 50 down-regulated genes in advanced-stage melanoma in order to see the main expression changes between early-stage and advanced-stage melanoma. Haqq et al. (2005) reported 19 gene signatures between nevi and metastases. A major finding of the study of Haqq et al. (2005) was the identification of two different gene patterns found in metastases reflecting those in the vertical or radial growth phase cells of primary melanomas.

Okamoto, I., Pirker, C., Bilban, M., Berger, W., Losert, D., Marosi, C., Haas,

O.A., Wolff, K., Pehamberger, H. (2005). Seven novel and stable translocations associated with oncogenic gene expression in malignant melanoma. Neoplasia *7*, 303-311.

Okamoto et al, 2005 have identified 20 over-expressed genes which were located in tumour-related regions of chromosomes by Affymetrix U133A gene chip using five malignant melanoma cell lines. These included AKT1, BMI1, CDK6, CTNNB1, E2F1, GPNMB, GPRK7, KBRAS2, LDB2, LIMK1, MAPK1, MEL, MP1, MUC18, NRCAM, PBX3, RAB22A, RAB38, SNK and STK4. They also indicated that the down-regulated CDK6 expression can dramatically reduce the growth of all five cell lines.

2006:

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S.R., Avril, M.F., Ortiz Romero, P.L., Robert, T., Balacescu, O., Eggermont, A.M., Lenoir, G., Sarasin, A., Tursz, T., van den Oord, J.J., Spatz, A. (2006). Gene expression profiling of primary cutaneous melanoma and clinical outcome. J Natl Cancer Inst *98*, 472-482.

Winnipennickx et al, 2006 attempted to detect the significantly-expressed genes related to the progression and prognosis of melanoma. Based on gene expression the patterns in primary tumours differed between patients with a 4-year distant metastasis-free survival from those who developed metastases within this time. 254 genes associated with metastasis-free survival primary melanoma were identified.

Mandruzzato, S., Callegaro, A., Turcatel, G., Francescato, S., Montesco, M.C.,

Chiarion-Sileni, V., Mocellin, S., Rossi, C.R., Bicciato, S., Wang, E., Marincola, F.M., Zanovello, P. (2006). A gene expression signature associated with survival in metastatic melanoma. J Transl Med *4*, 1-11.

The study by Mandruzzato et al. (2006) is based on the expression profiles of 17,500 probes of 43 stage III and IV metastatic melanoma tissues in 38 patients. 80 probes corresponding to 70 genes (45 and 35 probes relevant to longer and shorter survival times respectively) were identified by significance analysis of microarrays (SAM). In further analysis, they used a survival prediction model to predict 30 relevant survival probes from the 80 probes by supervised principal components (SPC) and cross-validation.

2007:

Alonso, S.R., Tracey, L, Ortiz, P., Perez-Gomez, B., Palacios, J., Pollan, M., Linares, J., Serrano, S., Saez-Castillo, A.I., Sanchez, L, Pajares, R., Sanchez-Aguilera, A., Artiga, M.J., Piris, M.A., Rodriguez-Peralto, J.L. (2007). A high-throughput study in melanoma identifies epithelial-mesenchymal transition as a major determinant of metastasis. Cancer Research *67*, 3450-3460.

The study by Alonso et al. (2007), is based on gene-expression profiles of 34 vertical growth phase (VGP) melanoma patients (21 developed nodal metastatic disease, and all had the minimum follow-up 36 months), and identified 243 genes (206 over-expressions and 37 down-expressions). All showed comparisons with >2 fold ratio and a false discovery rate is <0.2.

Jaeger, J., Koczan, D., Thiesen, H.J., Ibrahim, S.M., Gross, G., Spang, R., Kunz, M. (2007). Gene expression signatures for tumour progression, tumour

subtype, and tumour thickness in laser-microdissected melanoma tissue. Clin Cancer Res *13*, 806-815.

Jaeger et al. (2007) analyzed the 22,283 probe expression profiles of 41 primary melanomas and metastatic melanoma using oligonucleotide microarrays, 389 probe sets corresponding to 308 genes were identified with significant differential expression (Jaeger, 2007). A predictive diagnostic model (Support vector machine, SVM) was applied to discriminate these two stages of melanoma, and it achieved >85% correct classifications in cross-validation.

2008:

John, T., Black, M.A., Toro, T.T., Leader, D., Gedye, C.A., Davis, I.D., GUilford, P.J., Cebon, J.S. (2008). Predicting clinical outcome through molecular profiling in stage III melanoma. Clin Cancer Res *14*, 5173-5180.

John et al. (2008) used oligonucleotide arrays (30,888 probe sets) to examine 29 Patients with melanoma metastases to lymph nodes (stage IIIB and stage IIIC) that were grouped as 16 'poor-prognosis' and 13 'good-prognosis' cases by tumour progression time. 21 genes that were with highly differentially expressed between these two groups were identified, and then were validated in training data by quantitative PRC. Of the 21 genes, 15 genes exhibited the highest correlation (90% and 95% predictive scores) between two independent validation sets containing 10 and 14 stage III tumour samples respectively.

Riker, A.I., Enkemann, S.A., Fodstad, O., Liu, S., Ren, S., Morris, C., Xi, Y.,

Howell, P., Metge, B., Samant, R.S., Shevde, L.A., Li, W., Eschrich, S., Daud, A., Ju, J., Matta, J. (2008). The gene expression profiles of primary and metastatic melanoma yields a transition point of tumour progression and metastasis. BMC Med Genomics *1:13*.

Riker et al. (2008) compared 40 metastatic melanoma (MM) samples with 42 primary cutaneous cancers using microarray. Of the 42 primary samples, 16 were primary melanoma. They found that many genes (including SPRR/A/B, KRT16/17, CD24, LOR, GATA3, MUC15, and TMPRSS4) were expressed dramatically higher in primary basal cell carcinomas, squamous cell carcinomas and thin melanomas compared to metastatic melanoma. By contrast, some genes were up-regulated in metastatic melanoma such as MAGE, GPR19, BCL2A1, MMP14, SOX5, BUB1, and RGS20. 65 significantly expressed genes were detected by comparing normal human epithelial melanocytes to thin primary cutaneous and metastatic melanoma samples.

2009:

Bogunovic, D., O'Neill, D.W., Belitskaya-Levy, I., Vacic, V., Yu, Y.L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A.C., Lonardi, S., Zavadil, J., Osman, I., Bhardwaj, N. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in prodicting patient survival. Proc Nati Acad Sci *106*, 20429-20434.

Bogunovic et al. (2009) analysed 44 metastatic melanoma tissue samples from 38 patients with the approximately 20 month's clinical observation and identified 266 genes which were strongly associated with post-recurrence

survival. Among these some over expressed genes were associated with the immune response (e.g. ICOS, CD3d, ZAP70, TARP, GZMK, LCK, CD2, CXCL13, CCL19, CCR7, and VCAM1), and some down-regulated genes were associated with cell proliferation (e.g. PDE4D, CDK2, GREF1, NUSAP1, SPC24). The mitotic index (MI) was recognized as the most significant predictor of outcome (hazard ratio=2.13, p=0.0008), and it has now been included as an important prognostic factor in the AJCC melanoma staging and classification system.

Kashani-Sabet, M., Venna, S., Nosrati, M., Rangel, J., Sucker, A., Egberts, F., Baehner, F.L., Simko, J., Leong, S.P., Haqq, C., Hauschild, A., Schadendorf, D., Miller, J.R. 3rd (2009). A multimarker prognostic assay for primary cutaneous melanoma. Clin Cancer Res *15*, 6987-6992.

Kashani-Sabet et al, 2009 identified 5 overexpressed markers of melanoma (ARPC2, FN1, RGS1, SPP1 and WNT2), based on which a diagnostic algorithm was built, which achieved 95% specificity and 91% sensitivity in testing (n=534) diagnosis, and 95% specificity and 97% sensitivity in nevus-derived melanoma (n=75).

Koh, S.S., Opel, M.L., Wei, J.P., Yau, K., Shah, R., Gorre, M.E., Whitman, E., Shitabata, P.K., Tao, Y., Cochran, A.J., Abrishami, P., Binder, S.W. (2009). Molecular classification of melanomas and nevi using gene expression microarray signatures and formalin-fixed and paraffin-embedded tissue. Mod Pathol *22*, 538-546.

Koh et al. (2009), a study based on 120 samples, and identified 36 significant

differentially-expressed genes between melanomas and nevi and generated a gene expression classifier that was capable of distinguishing between melanomas and nevi. Genes involved signal transduction, transcription, and cell growth were identified as high expressing genes in melanomas compared with nevi. By contrast, L1CAM expression decreased in melanomas.

Jeffs, A.R., Glover, A.C., Slobbe, L.J., Wang, L., He, S., Hazlett, J.A., Awasthi, A., Woolley, A.G., Marshall, E.S., Joseph, W.R., Print, C.G., Baguley, B.C., Eccles, M.R. (2009). A gene expression signatures of invasive potential in metastatic melanoma cells. PLoS One *4*, e8461.

Jeffs et al. (2009) investigated molecular genomic characteristics to identify new prognostic and therapeutic markers in melanoma cell lines that may aid melanoma diagnosis. Expression of MITF, EDNRB, DCT and TYR were decreased and PLAUR, VCAN, and HIF1a were increased in melanoma cells. The result validated the hypothesis that differential gene expression may drive invasive metastatic melanoma and melanoma heterogeneity, like decreased specific lineage genes in the melanoma tumour microenvironment.

2010:

Scatolini, M., Grand, M.M., Grosso, E., Venesio, T., Pisacane, A., Balsamo, A., Sirovich, R., Risio, M., Chiorino, G. (2010). Altered molecular pathways in melanocytic lesions. Int J Cancer *126*, 1869-1881.

This study by Scatolini et al, 2010 was based on the comparison of gene

expression in 18 common nevi, 11 dysplastic nevi, 8 radial growth phase melanoma (RGPM), 15 vertical growth phase melanoma (VGPM) and 5 metastasis melanoma. They found that transition from RGPM to VGPM was related to apoptosis alteration (GO function). However, the transition from common nevi to RGPM corresponded to changes in intracellular junctions.

Kabbarah, O., Nogueira, C., Feng, B., Nazarian, R.M., Bosenberg, M., Wu, M., Scott, K.L., Kwong, L.N., Xiao, Y., Cordon-Cardo, C., Granter, S.R., Ramaswamy, S., Golub, T., Duncan, L.M., Wagner, S.N., Brennan, C., Chin, L. (2010). Integrative genome comparison of primary and metastatic melanomas. PLoS One *5*, e10770.

Kabbarah et al. (2010) performed a global genome analysis between primary and metastatic melanomas to examine genes related to metastatic progression. They identified 32 genomic regions that were significantly up-regulated in metastatic melanoma and 30 genes that were significantly down-regulated. In human melanoma cell lines, MET, ASPM, AKAP9, IMP3, PRKCA, RPA3, and SCAP2 were validated to encode for pro-invasion activities.

2011:

Rose, A.E., Poliseno, L., Wang, J., Clark, M., Pearlman, A., Wang, G., Vega Y Saenz de Miera, E.C., Medicherla, R., Christos, P.J., Shapiro, R., Pavlick, A., Darvishian, F., Zavadil, J., Polsky, D., Hernando, E., Ostrer, H., Osman, I (2011). Integrative genomics identifies molecular alterations that challenge the linear model of melanoma progression. Cancer Res *71*, 2561-2571.

Rose et al, (2011) detected differences between superficial spreading melanoma and nodular melanoma and they defined 8 significant genes (DIS3, FGFR1OP, G3BP2, GALNT7, MTAP, SED23IP, USO1 and ZNF668). However, this study did not provide a signature for metastatic melanoma.

Appendix table s1-s3 were stored as excel files in CD.

Appendix table s1: Melanoma signatures of 16 original studies.

Appendix table s2: the genes were validated by independent 16 original microarray studies

Appendix table s3: 180 known metastatic melanoma biomarkers.