Project Thesis in partial fulfilment for the degree of Masters in Applied Science (Bioinformatics)

Gene Expression Profiling of Melanoma: Correlation of RNA-Seq Gene Expression Data with Clinical Records in Melanoma

Student Name: David Cormican

Supervisor: Dr. Paul Walsh (NSilico Life Science)

Funded by EU H2020 MSCA RISE SAGE-CARE

Grant agreement ID: 644186

Table of Contents:

Abbreviations:

CTLA4 - Cytotoxic T-Lymphocyte Antigen 4

DM - Distant Metastases

GOE - Gene Ontology Enrichment

PD1 - Programmed cell Death 1 (cell surface receptor)

PDL1 - Programmed cell Death Ligand 1

RLN - Regional Lymph Nodes

SEER - Surveillance Epidemiology and End Results (cancer epidemiology surveillance program)

TCGA - The Cancer Genome Atlas

TF - Transcription Factor

TNM - Tumour, Nodes and Metastasis (cancer staging algorithm)

WHO - World Health Organisation

XP - Xeroderma Pigmentosum

Acknowledgements:

I would like to thank all the research and development staff of NSilico Lifescience for their support in conducting this research. I owe particular gratitude to my supervisor, Dr. Paul Walsh, for his guidance and feedback in planning and carrying out this work and Dr. Michael Bekaert for his advice on the intricacies of programming in R. Finally, I would like to thank Dr. Pat Forde of UCC, who also generously offered his time and expertise to help in the completion of this project.

Abstract:

Melanoma is a form of human cancer arising in melanocytes of the major problem in health care globally. Although recent advances in understanding of the genetic mutations and altered gene expression have led to the development of novel therapeutics which can extend life, stage IV, metastatic melanoma remains an incurable and invariably fatal condition. It is to be hoped that further advances in understanding of melanoma genomic and transcriptomic features may lead to further improvements in treatment. To assist in this, the Cancer Genome Atlas has made sequence data available for download for bioinformatic analysis by all interested researchers, along with matched anonymous clinical data for the majority of patients from whom cancer tissue was obtained for sequencing. Our work focused on RNA-seq gene expression data in this data set and we sought to investigate correlation between clinical features of melanoma and gene expression signatures, using the R bioconductor DESeq2 and topGO, focusing on the 10% of genes with greatest variation in expression levels (as reflected by the standard deviation) due to constraints on time and processor power available for computation. We also sought evidence of concordance between TCGA's RNA-seq based data and older studies of gene expression in patient melanoma samples using microarray data. Our research was hampered by the fact that not all sequence data was based on biopsies obtained at time of melanoma diagnosis, while much of the clinical data provided referred to disease status at that time. Thus much of the clinical data did not accurately reflect disease status at the time of sequencing. In spite of these issues, we were able to demonstrated differences in gene expression profile in samples from distant metastases compared to primary tumour and in involved regional lymph nodes compared to primary tumour. Gene ontology enrichment analysis demonstrated that gene functional categories which were over-represented in those differentially expressed in primary versus metastases included regulators of cell proliferation, cell adhesion, cell signalling to the immune system and signal transduction, and cellular dedifferentiation with loss of normal specialised function. Concurrence with previous microarray studies of gene expression differences between distant metastases and primary tumour was surprisingly poor, at less than 10%. We speculate that this may have been due to known issues with imperfect concordance between RNA-seq and microarray methodologies, as well as smaller samples in previous studies and the fact that primary tumour sequence data from the Cancer Genome Atlas was disproportionately generated from large, advanced primary tumours, which may have had expression profiles more similar to distant metastases than smaller, less aggressive tumours.

Aims:

1. To assess feasibility of using freely available RNA-Seq gene expression data and matched clinical data to conduct research into gene expression signatures in human melanoma.

2. To assess differences in transcription behaviour between primary melanoma and melanoma from distant metastases and between primary melanoma and melanoma involving regional lymph nodes.

3. To assess functional gene categories enriched in the list of genes differentially expressed in primary versus metastatic melanoma and thereby to gain insight into the mechanism by which metastasis takes place.

4. To assess concordance between genes differentially expressed in distant metastases versus primary tumour in our data and those reported in previous comparisons of gene expression between these sites.

5. To evaluate differences in gene expression in patients with better or worse outcomes according to information on survival provided in the clinical data.

Part 1: Introduction

1.1 Melanoma in Clinical Context

Melanoma is a human malignant neoplastic condition of melanocytes (the cell subpopulation responsible for the production of the dark pigment melanin). Like all neoplastic conditions, the primary lesion consists of a proliferation of cells which have become capable of evading normal mechanisms for limiting cell multiplication in multicellular organisms to levels which are appropriate to the functioning of the organism as a whole. A malignant neoplasm, or cancer, has the potential to spread to distant secondary sites via either the blood or lymph system in a process known as metastasis. Melanoma has pronounced metastatic potential.[1]

Melanocytes and hence melanoma occur in a number of body organ systems, but the vast majority are cutaneous melanoma, in which the primary tumour arises in the skin. A small minority occur in melanocytes present in smaller numbers in other tissue types, including the lining of the eye (uveal melanoma), ear, gastrointestinal tract and leptomeninges of the central nervous system.[2, 3] The are significant differences in the behaviour of melanomas arising in the skin compared to other sites in terms of metastatic potential, chemotherapy responsiveness and genetic factors.[3, 4] In view of these differences and the relative rarity of melanoma at these alternative sites, our research focussed exclusively on cutaneous type melanomas.

The impact of melanoma in human health is considerable. According to the most recent World Health Organisation (WHO) estimates, 160 000 new cases were diagnosed in 2002 with 41 000 deaths attributable to the condition in the same year.[1] In the United States, the Surveillance, Epidemiology and End Results (SEER) programme for collection of cancer data currently provides epidemiological estimates of the rate of melanoma in the American population up to and including the year 2012. SEER indicates that in 2012 the incidence of malanoma in the United States was approximately 21.6 per 100,000 and there were 2.7 melanoma-related deaths per 100,000, with women and men approximately equally affected.[5]

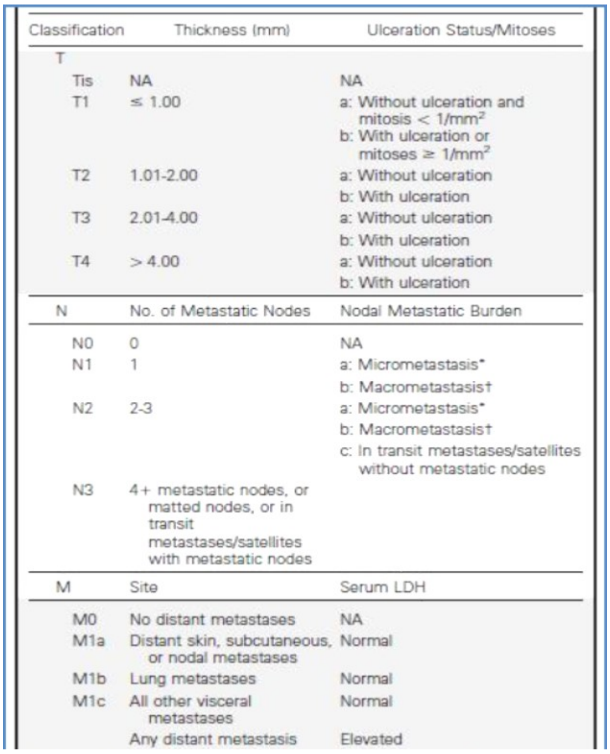
Outcomes in melanoma vary widely; many patients are cured by surgery and die of unrelated causes after many years while others rapidly progress to death from melanoma within months of diagnosis. The process of disease staging is of immense value in determining likelihood of achieving cure and optimal treatment strategy. Staging is based on principal that melanoma, like all malignant disease, begins as a small isolated primary lesion, which increases in size as it becomes more advanced. Spread beyond the primary lesion (the process defined as metastasis above) becomes more likely as the primary tumour increases in size. In most cases, the regional lymph nodes which drain fluid from the region of the primary tumour are the first site of disease spread. Evidence of malignant cells in these nodes confers a more advanced stage and hence worse prognosis. Disease which has metastasised to distant organs (most commonly bone, liver and brain[6]) is the most advanced form of the disease.[7]

Stage is determined by use of the TNM (Tumour, Nodes and Metastases) system. The TNM algorithm, presented in full in Tables 1.1 and 1.2, is complex and a detailed understanding of its implementation is not necessary in the research context. In brief however, three subscores are initially calculated; the T-stage, N-stage and M-stage (see Figure 1.1). A look up table (Figure 1.2) then allows the overall stage to be calculated on the basis of the score obtained for each of the individual components.[7, 8]

As can be seen in Table 1.1, T-stage of the tumour is primarily based on its thickness in millimetres as determined at biopsy (also referred to as the Breslow depth). This has long been established as the most important factor in determining prognosis in melanoma.[8] Clark depth is an alternative system for assessing thickness of melanoma primary tumour, in which tumour is assigned to one of five categorical values (level 1 to 5) depending on the deepest anatomic layer of skin into which it extends. Although Breslow depth is considered simpler in clinical practice and is used in most guidelines, the information conveyed by Clark level is similar.[1] In addition to thickness of the lesion, the presence of ulceration (obvious inflammation and degeneration of skin in the vicinity of the melanoma) has been shown to be an independent predictor of worse prognosis in a large study.[9] Furthermore, cells undergoing mitosis (cell division) can be easily identified on light microscopy inspection of melanoma biopsy specimen and the number of 'mitotic figures' identified per mm^2 of tissue is a third feature of primary tumour with independent prognostic value.[10] As a result, the most recent update of TNM staging guidelines allow T stage to be increased in some circumstances if ulceration of surrounding skin or high mitotic figure count are identified on inspection and biopsy of the primary.[7]

N and M staging rely on fewer factors than T staging. N-stage is based on the number of lymph nodes invaded by melanoma and whether or not this invasion is identifiable on biopsy of the node only (micrometastasis), or causes symptoms such as enlargement and pain in the area of the node (macrometastasis).[7, 8] Finally, M-stage depends primarily on presence and pattern of distant organ metastases. However, in metastatic disease only, blood levels of the enzyme lactate dehydrogenase (LDH) is also an independent predictor of prognosis and M the current TNM algorithm increases M stage to its highest possible value if any distant metastasis is identified and serum LDH is elevated, regardless of the pattern of organ involvement.[7, 8, 11]

In the final part of the staging process, the melanoma is assigned a stage between stage I

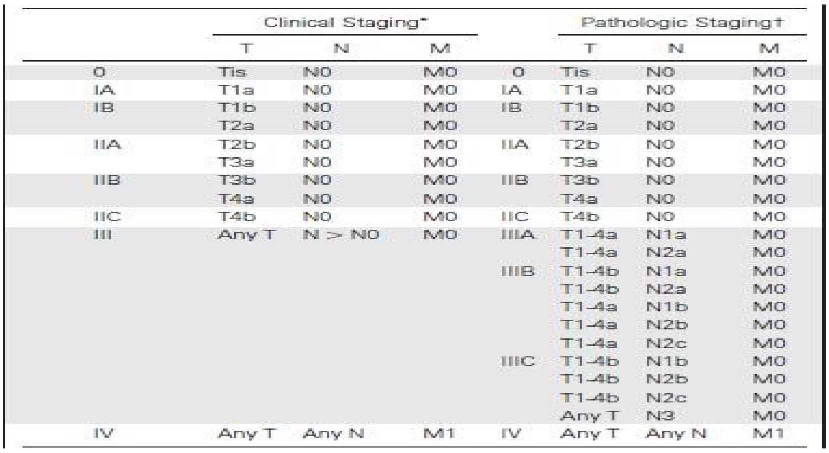


Classification	Thickness (mm)	Ulceration Status/Mitoses
T		
Tis	NA	NA
T1	≤ 1.00	a: Without ulceration and mitosis < 1/mm² b: With ulceration or mitoses ≥ 1/mm²
T2	1.01-2.00	a: Without ulceration b: With ulceration
T3	2.01-4.00	a: Without ulceration b: With ulceration
T4	> 4.00	a: Without ulceration b: With ulceration
N	No. of Metastatic Nodes	Nodal Metastatic Burden
N0	0	NA
N1	1	a: Micrometastasis* b: Macrometastasis†
N2	2-3	a: Micrometastasis* b: Macrometastasis† c: In transit metastases/satellites without metastatic nodes
N3	4+ metastatic nodes, or matted nodes, or in transit metastases/satellites with metastatic nodes	
M	Site	Serum LDH
M0	No distant metastases	NA
M1a	Distant skin, subcutaneous, or nodal metastases	Normal
M1b	Lung metastases	Normal
M1c	All other visceral metastases	Normal
	Any distant metastasis	Elevated

Table 1.1: Details of algorithm for assigning separate T, N and M stages to melanoma. Table 1.2 illustrates how overall stage is calculated from these components. (Source: Melanoma of the skin. In: Edge, Byrd & Compton eds. AJCC Cancer Staging Manual, 7th ed. New York, NY: Springer, 2010)[8]

e of

invasion of regional lymph nodes or distant metastases. Stage III disease has invaded regional lymph nodes but has not yet metastasised further to distant organs. Stage IV melanoma has progressed to a stage of involving deep organs other beyond the skin.[7]



	Clinical Staging*				Pathologic Staging†		
	T	N	M		T	N	M
0	Tis	N0	M0	0	Tis	N0	M0
IA	T1a	N0	M0	IA	T1a	N0	M0
IB	T1b	N0	M0	IB	T1b	N0	M0
	T2a	N0	M0		T2a	N0	M0
IIA	T2b	N0	M0	IIA	T2b	N0	M0
	T3a	N0	M0		T3a	N0	M0
IIB	T3b	N0	M0	IIB	T3b	N0	M0
	T4a	N0	M0		T4a	N0	M0
IIC	T4b	N0	M0	IIC	T4b	N0	M0
III	Any T	N > N0	M0	IIIA	T1-4a	N1a	M0
					T1-4a	N2a	M0
				IIIB	T1-4b	N1a	M0
					T1-4b	N2a	M0
					T1-4a	N1b	M0
					T1-4a	N2b	M0
					T1-4a	N2c	M0
				IIIC	T1-4b	N1b	M0
					T1-4b	N2b	M0
					T1-4b	N2c	M0
					Any T	N3	M0
IV	Any T	Any N	M1	IV	Any T	Any N	M1

Table 1.2: Details of the algorithm for overall staging of melanoma on the basis of T-, N- and M-stages. (Source: Melanoma of the skin. In: Edge, Byrd & Compton eds. AJCC Cancer Staging Manual, 7th ed. New York, NY: Springer, 2010)[8]

1.2 The Hallmarks of Cancer

Melanoma cells have the ability, common to all cancer, to proliferate indefinitely and spread beyond their site of origin in the form of metastases. In two seminal papers, Hanahan and Weinberg outlined the phenotypic 'hallmarks' of cancer, which are essential to the profound dysregulation of overall body function by which they induce serious illness and death:[12, 13]

1. Proliferation independent of external growth factors: Normal human cells undergo replication by mitosis only when stimulated by external signalling molecules, known as growth factors, secreted by other tissue. In malignancy, the involved cells are capable of undergoing mitosis in the absence of growth factors.

2. Insensitivity to anti-growth stimuli: In addition to positive regulation by growth factors, replication of most tissue types is also negatively regulated by other external signals. Cancer cells become insensitive to these.

3. Evasion of apoptosis: Apoptosis is also known as 'programmed cell death'. It occurs in normal cells in response to various signals which ensure elimination of damaged, redundant or abnormal cells. Cancer cells do not undergo normal programmed death response to these signals, enabling them to survive and proliferate in spite of their presence.

4. Limitless replicative potential: Malignant tumours continue to grow indefinitely if adequate nutrients are available, partly in view of the insensitivity of constituent cells to external signals which might curb their replication, as described in the previous three points.

5. Stimulation of angiogenesis: Angiogenesis is the formation of new blood vessels. To ensure that malignant cells do indeed have access to adequate nutrients, an enlarging tumour signals to surrounding normal blood vessels to promote their growth. In the absence of this hallmark, tumours outgrow their vascular supply and die by a process of necrosis.

6. Invasion and metastasis: Metastasis was defined in the previous section. Invasion of tumours beyond their site of origin may depend on their ability to disrupt the extracellular 'scaffolding' tissue in their environment which normal functions partly to keep cell proliferation organised and confined.

7. Abnormality of metabolic pathways: Malignant tumour frequently show more avid uptake and higher turnover of glucose and other nutrients to support their rapid growth

8. Immune evasion: Cells of the normal immune system, especially the natural killer (NK) cells destroy many abnormal cells, probably including many with malignant potential. In order for proliferating tumour cells to survive in the long term, they require mechanisms to avoid this kind of targeted destruction.

1.3 Genome Instability in Melanoma

In general, cells acquire the hallmarks of cancer due to the accumulation of mutations in their genomes, though other factors such as epigenetic modification may also play a role (see Section

1.5). Factors which increase the risk of DNA mutation cause a corresponding increase in the risk of cancer development. In general, such factors may be categorized as environmental or hereditary.[15]

In most melanoma cases, environmental factors are of greater significance. In particular, exposure to ultraviolet radiaiton (UVR) has been consistently shown to increase rates of human melanoma.[16, 17] UVR is believed to damage DNA by both direct and indirect mechanisms. In the direct effect, free energy is transferred from the photons of UVR to the DNA macromolecule. The energy thus imparted leads to aberrations in DNA structure which are normally impossible as they require the DNA molecule assuming energetically unfavourable intermediate states. Two kinds of abnormality are particularly typical of UVR-exposed DNA: 6–4 photoproducts, generated between adjacent pyrimidine residues, and pyrimidine or cyclobutane dimers, formed specifically between adjacent pairs of thymine (T) or cytosine (C) residues.[15] Eukaryotic cells possess mechanisms by which to repair damage of this kind but, they are imprecise and can give rise to genome alterations, i.e. mutations. Furthermore, the free energy transmitted by UVR can lead to the generation of free radicals such as reactive oxygen species in exposed cells. These radicals attack and damage several of the cells macromolecules including DNA, with repair mechanisms once again having a certain error rate, thereby leading to mutations.[17]

Approximately 10% of melanoma cases occur in individuals affected by hereditary 'melanoma syndromes'.[18] In these cases, the cause genome instability is a defect in the complex enzymatic system for maintenance of DNA fidelity. Such hereditary defects result in mutation accumulation in the melanocytes from early in life and melanoma occurs at young age in affected individuals. The best described germ line mutations leading to early onset melanoma in affected family members are listed below:

- Xeroderma pigmentosum (XP) is a rare inherited syndrome in which one of seven genes (denoted XPA to XPG) involved in the nucleotide excision repair (NER) pathway is non-functional. Both copies of the affected gene must be defective; hence the inheritance pattern is autosomal recessive. NER is the main mechanism by which directly induced UVR-induced DNA damage is corrected in an orderly manner, to give a corrected DNA sequence faithful to the original. In the absence of normal functioning of this mechanism, UVR DNA damage leads to disorder in the genome of skin cells exposed to UVR, most commonly leading to cell death and severe blistering of the skin when exposed to direct sunlight. Surviving cells, including melanocytes, have severely rearranged and mutated genomes, due to chaotic attempts at damage repair. These genome alterations lead to very high rates of young-onset melanoma in XP patients compared to the normal population.[19]
- CDKN2A inactivating mutation: CDKN2A is a gene on human chromosome 9 with two alternative reading frames (ARF) meaning transcription of this gene can produce two different RNA products. The alpha transcript encodes protein p16, which has a key role in inhibiting cell division when DNA errors are present, giving time for DNA repair mechanisms to correct the error before it is copied in daughter cells. In a subset of melanoma-prone families CDKN2A is mutated to a functionally attenuated form in affected members. This has the predictable effect of increased melanoma risk. Compared to XP this mutation causes far fewer non-melanoma related skin problems.[18]
- CDK4 activating mutation: One of the mechanisms by which CDKN2A inhibits ongoing mitosis is by downregulation of CDK4. Thus it is unsurprising that mutations in CDK4 which

lend it greater activity or resistance to CDKN2A inhibition give rise to a syndrome similar to CDKN2A inactivation, although CDK4 mutation appear to be rarer.[18]

- BAP1 mutation: The BAP1 gene is located on chromosome 3 and has a similar effect to CDKN2A in its wild type form, to prevent replication of cells with abnormal genome. The pathway by which it achieves this is however separate to that via which the previous two genes act. Germ line mutations in BAP1 have been infrequently reported as a cause of hereditary melanoma.[18]

1.4 Melanoma Driver Mutations

As outlined above, any factor which increases instability in the genome of melanocytes confers an increased risk of melanoma. Nonetheless it is important to emphasise that genome instability in and of itself is an insufficient explanation of how cells acquire the hallmarks of cancer. There is no direct mechanism to explain why cells with more mutations should be able to proliferate in the absence of growth factor signals, induce angiogenesis or exhibit any of the other typical cancer behaviours. Instead an indirect mechanism comes into play; genome instability increases the risk of mutations affecting the function of gene products which are capable of inducing these behaviours - so called 'driver mutations'. The identification and classification of driver mutations in melanoma has been a very active research in recent years. However, this task poses significant challenges. Firstly, the driver mutations differ between melanomas; no mutation has yet been identified which is universal to all cases. Firstly, driver mutations differ between melanomas. The complex cell processes which must be perturbed for cancer to develop generally rely on several different genes for normal function and mutations affecting the function of any one of these genes can give rise to similar malignant behaviour to any other single mutation affecting the process. Thus, even the most widely reported mutations in melanoma are not relevant in more than two thirds of cases. Secondly, the random process of mutation is accelerated in melanocytes prior to the onset of melanoma. This high mutation rate gives high risk of acquiring driver mutations causing melanoma, but other mutations, which are not relevant to melanoma progression, also accumulate at a more rapid rate than normal. Thus the driver mutations in melanoma can be lost in the noise of many background mutations, sometimes referred to as 'passenger' mutations.[20] In spite of these problems, good evidence exists to implicate either activating or inactivating mutations in several genes in the development of melanoma. The best studied of these appear in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database pathway for melanoma, a diagram of which is shown in Figure 1.1.

The best established and clinically most important driver mutation in melanoma is BRAF, which as can be seen in Figure 1.1, promotes proliferation of cells through a number of intermediate gene products. BRAF mutation has been reported to be present in up to 66% of melanoma.[21] These mutations are usually cause alteration in the structure of BRAF protein so that rather than being activate only briefly when the cell receives 'pro-growth' signals, it becomes constantly active (constitutively active) forcing melanocytes to continue increasing in number even in the absence of external growth factors.[22, 23] Crucially, BRAF inhibitors have been shown to inhibit growth of both melanoma cell lines in vitro and tumours in patients.[24] The therapeutic use of these agents is discussed further in section 1.6.

BRAF exerts it effect on cell division via second messenger downstream signalling molecules. The most proximal of these is MEK1, with which BRAF interacts directly.[22] Activating mutations in

MEK1 are also believed to act as driver mutations in some melanoma but are present in a much lower proportion of tumours (perhaps 6 - 7%).[23] Cyclin D1 (CCND1), another downstream component of the BRAF pathway is also frequently abnormally active in melanoma; in the case of this gene the genomic lesion often consists of excessive copies of a structurally normal gene.[24]

NRAS is not involved in BRAF signalling but can activate some of the same pathways to promote cell proliferation. Furthermore, activation of NRAS induces activity in a second protein kinase cascade involving PI3K, PKBA1 and BAD which counteracts pro-apoptotic signals, thereby conferring insensitivity to anti-growth signalling as well. Activating mutations of NRAS have a driver role in up to 25% of cutaneous melanoma.[24, 25] PTEN, a gene whose product downregulates the NRAS signalling pathway may also have a role in melanoma development. PTEN mutations inducing melanoma lead to a defective or inactive final gene product, in contrast to the previously discussed driver mutations. Loss of the normal suppression of NRAS signalling leads to overall effects similar to mutations which lead to overly active NRAS. In transgenic mouse melanoma models, genetic knockout of PTEN appears to have a synergistic effect with BRAF mutant expression to promote growth and metastasis of melanoma as the combined effects of these two genetic lesions cause excessive activity of both BRAF and NRAS signalling; related but distinct mechanisms for promoting malignant melanocyte proliferation and survival.[26]

While mutations of BRAF, NRAS and related genes are clearly very important for unchecked proliferation and resistance to cell death in many melanoma cases, they do not explain the acquisition of all the hallmarks of cancer by these cells. Understanding of the mechanisms by which these other features arise is less detailed at this time. However, the mechanisms by which immune system evasion have been explored to some degree and deserve special mention as recent evidence has emerged that their disruption by pharmacological means can be effective in the treatment of melanoma patients. Clinical and pre-clinical evidence from recent years suggest that melanoma cells, unlike normal melanocytes, have molecules on their surfaces which interact with and activate receptors on the surface of white blood cells of the immune system. Activation of these white blood cell receptors, which include cytotoxic T-lymphocyte associated antigen 4 (CTLA-4) and the programmed death receptor 1 (PD-1), dampens the white cells' cytotoxic activity, i.e. the mechanisms via which it kills potentially malignant abnormal cells which have not developed means to evade it. Discussion of the clinical impact of monoclonal antibodies which prevent this reduced white cell cytotoxicity by blockade of CTLA-4 and PD-1 is presented in section 1.6.[27]

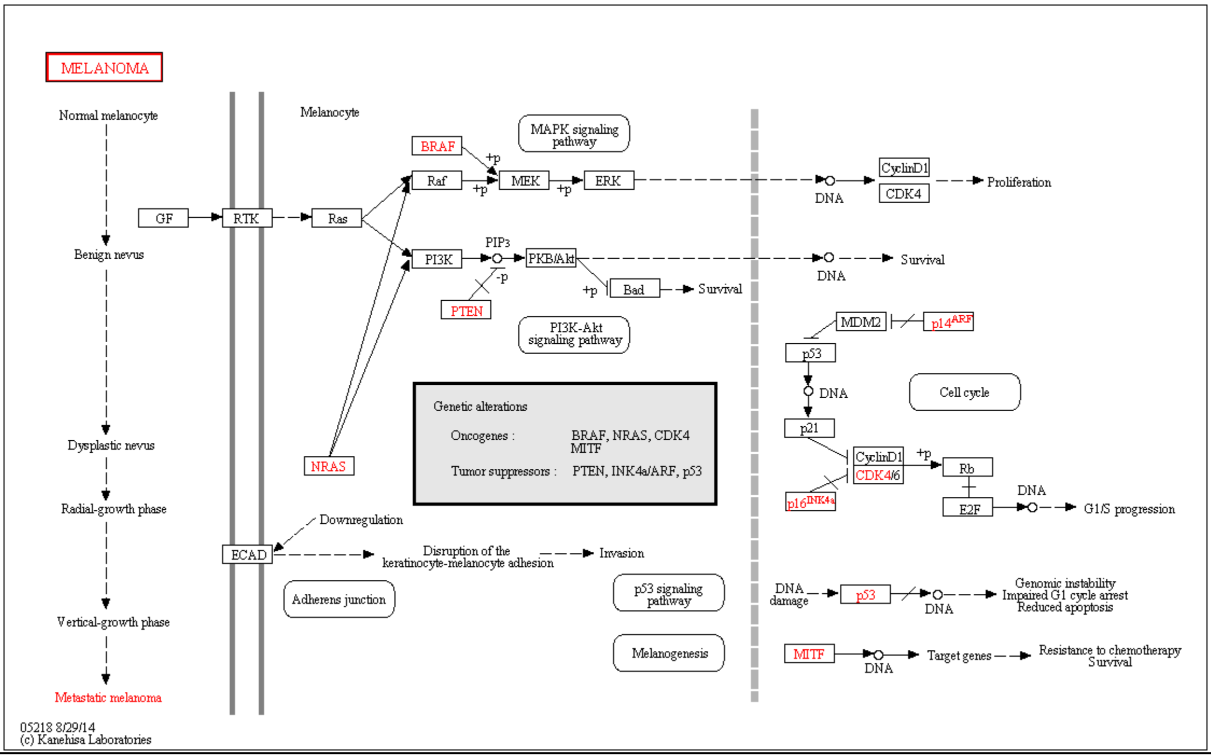


Figure 1.1: KEGG pathway for melanoma. Genes in red are commonly mutated. Obtained from: http://www.kegg.jp/kegg-bin/highlight_pathway?scale=1.0&map=map05218&keyword=melanoma

1.5 Altered Gene Expression in Melanoma

Although study of the DNA genome has lead to many interesting insights into the process underlying melanoma development, the RNA transcriptome also offers huge opportunities for insight into the causes of malignancy. RNA expression is a function of the interplay among a large number of transcription factors (TF's) which are active to varying degrees in the cell nucleus. TF's are protein products of certain genes, which regulate the expression of a large number of other genes. It is notable that many of the driver mutations described in the previous section are themselves TF's or else indirectly influence the behaviour of TF's. As a result many genes show altered expression in melanoma (i.e. RNA transcripts of the gene are present in the cell at higher or lower levels than at baseline) and expression of genes can be altered in the malignant cell even if the DNA gene itself is completely normal.

In addition to TF-driven changes in expression, genes can experience alterations in transcription due to alterations in epigenetic regulatory mechanisms such as methylation, chromatin remodelling and interference with transcription by small non-coding RNA molecules. There is some evidence that these mechanisms also play a role in development and progression of melanoma and may explain some part of the alterations in gene expression seen in melanocytes during the acquisition of the hallmarks of cancer.[28]

Classically the study of the transcriptome was difficult in cell biology owing to the huge amount of data involved; approximately 20 000 genes all undergoing distinct dynamic changes in expression level in response to changes in the cell's environment or extracellular signals being received. However, the advent of microarray technology in the mid-1990's facilitated more detailed analysis of RNA transcription and was applied to a number of human malignancies, including melanoma. Later still, massive parallel sequencing of RNA, for example by RNA-Seq technology, has become possible. Microarray has been widely employed in the analysis of transcriptomic features of clinical melanoma samples but studies utilizing the newer RNASeq technology have to date been limited.

Early studies of gene expression in melanoma focussed on laboratory cultured cell lines and compared populations of cells grown from normal, non-malignant melanocytes to populations derived from melanoma cells.[29, 30] More recently, other studies have used tissue removed from patients at biopsy or surgery. Many of these reports describe differences in expression between samples from malignant melamona compared to others from melanocytic naevi. Melanocytic naevi represent abnormal aggregates of melanocytes, visible as a dark area of skin, present either from birth or as a consequence of a requirement for increased melanin production due to sun exposure. Despite a superficial similarity to melanoma they are benign lesions and with normal appearance at the microscopic level. Several groups have reported high degrees of success in separating naevi from melanoma samples on the basis using clustering techniques for microarray expression data.[30, 31, 32]

Other investigators have sought to investigate differences in expression in more advanced compared to earlier stage melanomas by comparing gene expression in melanoma primaries to expression in distant metastases using microarray data.[31, 32] Mauerer et al[33] compared gene expression data in melanocytic naevi (n = 18), primary melanomas (n = 20) and melanoma metastases (n = 20). Arrays used in that study contained 22000 probe sets representing 14500 genes. A total of 189 genes were differentially expressed in metastases versus primary melanomas and 247 were differentially expressed in primary melanoma versus melanocytic naevi. The study noted that several MAP kinase genes, which are regulated by both BRAF and NRAS pathways known to play a role in melanoma carcinogenesis (see section 1.4), were expressed significantly more strongly in more advanced disease. Jaeger et al[34] carried out a similar study comparing expression in 22 metastases to 19 primary tumours. They reported 308 genes which showed differences in gene expression between the two groups and proceeded to demonstrate that cases could be separated into primary tumours and metastases on the basis of gene expression with 85% accuracy using a support vector machine-based classification learner. It should be noted that both the study of Mauerer et al and that of Jaeger et al were potentially limited relatively small sample sizes.

As mentioned previously, there have to date been few studies of gene expression in melanoma as analysed by direct RNA sequencing methods such as RNA-Seq. Previous studies suggest that using the Illumina RNA-Seq platform with use of Poisson model to capture variation across technical replicates can lead to 30% greater detection of differentially expressed genes compared to standard microarray methods, without an increase in false positives.[33] Thus, failure to make use of the newer technology in the field of melanoma research may give a less complete picture of the transcriptomic landscape of the condition.

A recent study by the Cancer Genome Atlas consortium[36] did include gene expression as assessed by RNA-Seq as one of the 'dimensions' of molecular data assessed in 331 melanoma specimens in an attempt to uncover biologically relevant subtypes. (DNA mutations, somatic copy number alterations and DNA methylation were the other dimensions included.) Clustering was carried out for each dimension using R's iCluster package, followed by integrated clustering of all dimensions. Hierarchical clustering of RNA Seq data showed evidence of three robust stable clusters. Extended follow-up data was available for most patients included in this study and this allowed for assessment of an effect on post-biopsy survival of cluster assignment. The log rank test did indicate a significant survival advantage for patients assigned to one of the clusters. The study did have some limitations however. Firstly, samples included biopsy specimens from both primary melanoma tissue and lymph node metastases, introducing a possible source of heterogeneity in the data. Furthermore, only the 1500 genes showing the greatest variation in expression between samples (as assessed by the standard deviation) were analysed in the RNA-seq expression analysis part of the study. This approach was an attempt to reduce computational resources required by the study but may have led to the loss of useful data.

1.6 Treatment of Melanoma

Treatment of melanoma depends strongly on the TNM staging that the disease is assigned at diagnosis. The process by which TNM staging is carried out is discussed in section 1.1. Stage I or II disease, which is localised to the primary tumour in the skin, can usually be treated simply by surgical excision of the melanoma. Stage III disease, which also involves nearby lymph nodes, is traditionally treated with a more radical disease excision, with lymph nodes removed as well as the primary tumour. Prognosis for stage I and stage II patients undergoing excision of melanoma is generally good, although a small minority will experience recurrence of the disease. The risk of recurrence is higher with stage III melanoma. Stage IV disease was previously defined as melanoma with metastasis to distant deep organs. Such advanced disease cannot be treated surgically and is generally fatal. Until the last decade, indeed, no effective treatment was available for stage IV melanoma and median survival was as low as 9 months.[1] However, survival has been extended in recent years with a range of new and effective therapies becoming available, many of which were developed thanks to improved understanding of the underlying genome alterations which lead to melanoma development.[7]

The new agents effective in stage IV disease include the small molecule agents, vemurafenib, dabrafenib and trametinib and the monoclonal antibodies ipilimumab and nivolumab. The choice of agent depends on genome features of the malignancy, with the status of BRAF (i.e. whether it is mutated or normal, also referred to as 'wild type') being particularly important.[7, 37] For patients with mutated BRAF, clinical trial data now strongly supports treatment with BRAF inhibitors (vemurafenib or dabrafenib) as a means of extending survival.[38] More recently, newer data has suggested that survival may be extended further still by co-administration of trametinib, an inhibitor of MEK, a downstream target of BRAF.[39] For approximately 50% with normal BRAF, the monoclonal antibody ipilimumab provides a means of blocking the white cell CTLA-4 receptor. As melanoma cells binding to this receptor to deactivate white cells which would otherwise destroy the malignant cell, blockade of this receptor in theory improves immune targeting of melanoma. Improved survival with ipilimumab administration has provided support for this principle.[39] Nivolumab, another monoclonal antibody blocking another white cell receptor, PD1, involved in the immune 'switching off' mechanism, showed evidence of synergistic effect with ipilimumab in another recent trial.[40]

1.7 The Cancer Genome Atlas

The success of therapies such as the BRAF inhibitors, whose discovery was critically reliant on exploration of the genomic and transcriptomic landscape of melanoma, underlined the potential benefit of making more sequence data available to as wide a range of researchers as possible, to encourage further discoveries about the underlying biology of the condition which could lead to important breakthroughs in treatment. The Cancer Genome Atlas (TCGA) is a research concern which has worked towards this objective.

TCGA is a collaborative program set up by the National Cancer Institute in association with the National Human Genome Research Institute with the aim of promoting research into and cataloguing of the molecular alterations which play a role in cancer biology. The biospecimen core resource (BCR) and its associated data portal are part of the resources that TCGA maintains to further this goal. The BCR is a central site using uniform protocols to receive and process all tissues and clinical data, thereby generating matched clinical and sequence data which is accessible via the data portal. Samples are available for over 20 cancer types, including melanoma. All samples are anonymised by the clinical centre at which they are initially processed, prior to receipt by the BCR and risks of re-identification are minimized. However, because of greater concerns over possible re-identification related to some kinds of sequence data, TCGA operates a two-tiered access policy. For individual level germ-line variations, whole exome or whole genome sequence data, certification must be provided to researchers by TCGA prior to access being granted. For other types of data, access is free and no certification from TCGA is required prior to download and analysis; this includes the RNA-Seq expression level data we used in our study. All patients included provided written informed consent prior to sample submission.[41]

1.8 R packages Used

All statistical analysis for this studied was carried out using open source R software. For the more advanced analyses specific Bioconductor packages were required:

1. DESeq2 for analyses of differential expression between various groups within the sample. DESeq2 models read counts as following a negative binomial (Poisson) distribution. General linear models with log link are fitted for each gene, an empirical Bayes approach is used for shrinkage and false discovery rate is reduced by eliminating genes with mean normalized counts lower than a particular threshold prior to testing for significance using a Wald test.[42]

2. topGO was used for gene ontology enrichment analysis on lists of differentially expressed genes used. In our topGO analyses the Fisher exact test was used to assess for significant enrichment, but this can be modified.[43]

3. stringr was used in some of the data cleaning tasks.

Part 2: Methods

2.1 Sequence Data Used in Study

The Cancer Genome Atlas (TCGA) provides RNA-Seq data for 472 melanoma tumour biopsy specimens via the Data Portal of its Biospecimen Core Resource (BCR). Each patient is assigned a unique alphanumeric identifier, referred to as BCR barcode. RNA-Seq expression data is among the bioinformatic data regarded by BCR as having a low risk of patient re-identification (see Introduction, Section 1.7) and as such is freely available for download from the data portal without need for investigators to seek approval from TCGA's steering committee.

2.2 Clinical Data Used in Study

468 of the 472 melanoma specimens for which RNA-Seq expression data is available via the BCR data portal have matched clinical data also available. BCR barcode enables matching of clinical and sequence data in BCR dataset while maintaining subject anonymity. A total of 32 variables in the data set describe: 1) demographic information on the patient, 2) details of the type of specimen acquired from each patient, 3) information on disease features at diagnosis and 4) information on patient follow-up, treatment and outcome.

2.2.1: Demographic features of the Patients

Table 2.1 demonstrates the seven variables used to describe baseline demographic features of the patients in the TCGA data set and describes their data type. Ethnicity was a categorical variable with three levels; white, black/African-American and Asian. Ethnicity had two levels; Hispanic or non-Hispanic. All patients of black or Asian race were of non-Hispanic ethnicity. Patients whose race was recorded as White or where race was unknown, could be recorded as having Hispanic ethnicity.

Variable	**Type of Data**
Days to Birth	Numeric
Gender	Categorical, binary
Height in cm	Numeric
Weight in kg	Numeric
Prior Cancer Diagnosis	Categorical, binary
Race	Categorical, three levels
Ethnicity	Categorical, binary

Table 2.1: Variables describing patient demographics

2.2.2: Variables describing aspects of Specimen

Table 2.2 describes the two variables which provide information regarding the circumstances of sample collection. Days to sample submission indicates the amount of time that elapsed between melanoma diagnosis and specimen submission. Ideally, all specimens would be submitted at diagnosis, i.e. days to sample submission equal to zero. This was not the case for many of the specimens, with years having elapsed between diagnosis and specimen submission in many cases. In some cases the sample had only been acquired when disease previously designated Stage I - III, i.e. without distant metastases, advanced to involve organs outside the skin. As a result, some patients were described as having stage I - III disease but specimen was obtained from a distant metastasis.

Site of sampling had three levels, primary tumour (PT), Regional Lymph Node (RLN) and Distant Metastasis (DM). Again, this was a non-ideal situation from the point of view of gene expression analysis. A preferable arrangement would have seen all tissue collected from primary tumours, allowing analysis of differential gene expression in groups defined by disease state or outcome measures without differences in specimen site introducing confounding factors. Furthermore, gene expression in metastases is known to be highly chaotic and of little use in predicting outcomes. However, the availability of a mixture of sites of tissue origin did allow comparisons to be made between samples from primary tumour, regional nodes and distant metastases to attempt to identify changes in gene expression associated with blood-borne metastasis to distant organs and spread to regional lymph nodes.

Variable	Type of Data
Days to sample submission	Numeric
Site of sampling	Categorical, three levels

Table 2.2: Variables detailing circumstances of tissue sampling

2.2.3: Disease Characteristics at diagnosis

Table 2.3 demonstrates the twelve variables used to describe disease features at time of diagnosis. 'Multiple primary present at diagnosis' was yes if the patient had two different primary tumours diagnosed at the same time, with primary tumour count being 1 for all patients with 'Multiple primary present at diagnosis assigned 'No'

Breslow depth and Clark level both describe the depth of invasion of the primary lesion into the skin and provide similar information. Breslow depth is the depth of invasion below the surface of the skin in millimetres and as such is a numeric variable. Clark level is a categorical value, I to V, with higher values representing greater depth of invasion through the skin.

Ulceration of primary tumour is assessed visually and presence of ulceration is an independent predictor of poor prognosis in melanoma. 'Mitoses per high power field' is assessed visually on light microscopy with higher values associated with worse prognosis.

The details of TNM staging and its three constituent components are discussed in the introduction to this thesis.

Lactate Dehydrogenase (LDH) is an enzyme measurable in the blood. Higher levels are associated with worse prognosis in metastatic melanoma.

Variable	Type of Data
Multiple primary present at diagnosis	Categorical, binary
Primary tumour count	Numeric
Breslow depth in mm	Numeric
Clark level	Ordinal
Ulceration of primary tumour	Categorical, binary
Mitoses per high power microscopic field	Numeric
Pathologic T-stage	Ordinal
Pathologic N-stage	Ordinal
Pathologic M-stage	Ordinal
Overall TNM stage	Ordinal

Serum LDH level (IU/L)	Numeric
Year of diagnosis	Date

Table 2.3: Variables describing features of disease at diagnosis

2.2.4: Outcome, Follow-up and Treatment Variables

Table 2.4 describes the variables describing patient follow-up, treatment and outcomes in TCGA's dataset.

Not all patients were followed up for the same amount of time and days to last follow up gives an indication of when they were last contacted relative to their initial diagnosis. Vital status indicates whether the patient was dead or alive at last follow up and tumour indicates whether they still had evidence of cancer or were considered cured when last followed up. Days from diagnosis to death was recorded for all patients who died during follow-up (it was N/A for those still alive at the end of follow-up).

The variable length of time to last follow-up meant that the information provided by vital status, tumour status and days to death had to be treated with caution. It was possible that some patients with a short follow-up still alive at the end of that time in fact died sooner than some who did so during the course of a long follow up. Similarly it is possible that patients presumed tumour free after a short follow relapsed thereafter.

Interferon treatment and radiotherapy treatment indicate whether or not a patient received either of these two modes of treatment after diagnosis. Interferon is a biologic, immune-modulatory agent which was until recently the only licensed treatment to extend life in advanced melanoma. However, as discussed in the introduction, recent years have seen the development of a number of newer small molecule and monoclonal antibody agents for melanoma treatment. It is likely that few of the patients in this database received these newer agents but the failure to record where this was the case is a weakness of the data.

Variable	**Type of Data**
Days to Last Follow up	Numeric
Vital status	Categorical, binary
Tumour status	Categorical, binary
Days to death	Numeric
Interferon treatment	Categorical, binary
Radiotherapy treatment	Categorical, binary

Table 2.4: Variables describing patient follow-up, outcome and treatment received

2.3 Download and Cleaning of Data

A perl script (see Appendix) was written to extract RNA Seq data and matched clinical information for all melanoma specimens from the TCGA's Data Portal. Clinical data was saved in Microsoft Excel CSV form and RNA Seq reads in Microsoft Notepad form. Clinical data was anonymous but each patient was assigned a TCGA barcode number by which clinical data could be matched to RNA-Seq data.

Clinical data was read into R, version 3.1.2 as a data frame identified as 'clinical' and RNA-Seq data as a data frame identified as 'counts'. A function was written to convert variables to numeric type where appropriate (see tables 2.1-4). Additionally, as '-' was used as a separator in patient barcode in the clinical data while '.' was used in the RNA Seq read dataframe, the str_replace_all() function in the stringr library was used to replace '-' for '.' in the patient barcode variable in the 'clinical' data frame.

As mentioned section 2.1, above, time from diagnosis to tumour sampling was variable among patients included in the database; years in some cases. The variables Days to Last Follow up and Days to Death described time from disease diagnosis to these events. For the purposes of our study, the time patients were followed up for and/or survived following tissue sampling was of more interest; RNA-Seq data obviously describes gene expression at the time of tissue sampling rather than diagnosis, if these were not the same, so time of sampling is a more appropriate t = 0. This was easily dealt with by redefining Days to Last Follow-Up as:

Days_To_Last_Follow_Up = (Days_To_Last_Follow_Up) - (Days_To_Specimen_Collection)

And similarly days to death was redefined as:

Days_To_Death = (Days_To_Death) - (Days_To_Specimen_Collection)

TNM disease stage was recorded major stage I to IV, with stage II and III further subdivided into substage A to C. To reduce the number of levels for the disease stage categorical variable, substages were ignored with all stage II disease comprising a single level and similar for all stage III. We justified this on the basis that, to our knowledge, no previous expression studies have subdivided mid-stage disease into A to C substages and doing so here would have dramatically increased the number of factor levels for the disease stage variable while reducing the number of samples included in each level. Similarly, the individual components of TNM stage, i.e. T-stage, N-stage and M-stage are normally divided into integer 'major stage' and these are subdivided into substage A - C, but in the interest of simplicity and limiting factor levels, we again regarded all substages of a major stage as a single factor level.

No clinical data was available for five patients and these individuals were deleted from the counts data frame.

2.4 Selection of Clinical Variables for Study

For variables included in the clinical data, version 3.1.2 of R was used to generate summary tables if variables were categorical or ordinal and histograms if variables were numeric.

Our primary goal in conducting this study was to investigate differences in expression profile in groups of patients who differed according to the behaviour of their disease. As discussed in the previous section, clinical variables describing patients fell mostly into four categories; variables describing patient demographic features, variables describing the specimen obtained from the patient, variables describing patient disease features at diagnosis and variables describing the long term outcomes of the patients. Prior knowledge of the biology of melanoma was combined with the summaries of variable distribution generated in R in order to select variables on which the sample could be split in order to compare gene expression in different groups likely to exhibit meaningful differences in tumour behaviour.

Of the four types of clinical variable discussed in section 2.2, those in the 'patient demographic' groups describe only features of the patients themselves and are unlikely to have any relevance to the disease process or the genomic features of the tumour.

Differences in the specimen type used to generate RNA-seq data was, however, considered a potential source of variation in gene expression and the second group of variables discussed previously did provide one key piece of information in this regard. The source of the specimen was recorded as being primary tumour, tumour in regional lymph node or distant metastasis (see introduction for a discussion of the significance of disease at these three different sites). It would be expected, both from first principles and previous microarray studies that cancer in distant metastases would have a markedly different gene expression profile from primary tumour, while regional lymph node disease might be expected to have an expression profile more similar to (but still distinct from primary) tumour. Moreover, site of specimen origin was a relatively attractive option as a variable to study in this data set as it was recorded for the majority of specimens and was directly relevant to the specimen from which sequence data was generated.

Analysing differences in expression profile between groups of patients with differences in disease characteristics at diagnosis, as described by the third group of clinical variables discussed in the previous section, seemed a promising means of identifying expression signatures associated with more or less advanced disease. Several different variables in the third group described related to features of melanoma which have been shown to correlate with prognosis.

As was discussed in the introduction, TNM stage of the cancer is the most important variable in individualising prognosis and treatment in melanoma. Increase in overall TNM stage, or an increased value for any of it three components, has been shown to correlate with worse patient prognosis (as discussed in the Introduction). TNM stage at diagnosis was recorded for the majority of patients in TCGA's melanoma dataset. Each of T, N and M stage were also recorded independently as ordinal variables. Primary tumour size, the most important factor in determining T stage, was also recorded in terms of both the numeric Breslow depth (depth of primary tumour invasion in millimetres) and Clark depth (a five-level categorical variable also reflecting depth of invasion). Also, mitotic rate (a numeric variable) and presence of ulceration of primary (binary yes/no) which also contribute to T staging were also documented. Finally, a numeric level serum LDH level was also among those listed as available in the dataset. Serum LDH is a relevant prognostic indicator in metastatic disease only, with high blood levels of this enzyme correlated with worse prognosis.

Unfortunately, there were significant practical difficulties with analyzing gene expression on the basis of disease features at diagnosis. The most important of these was that specimens were often not collected and sequenced until a substantial length of time had elapsed since diagnosis. In such cases, it is likely that the values which accurately described the disease at diagnosis were no longer applicable at the time the specimen was taken. Furthermore, groups which included samples from primary tumour, regional lymph node disease and distant metastasis had the potential to introduce confounding factors, owing to differences in gene expression between these different sites. As a consequence, for analysis seeking to correlate variables describing disease at diagnosis with gene expression, only patients who had primary tumour specimens obtained at the time of diagnosis (i.e. 'Days to Sample Submission' = 0 and 'Site of Sampling' = Primary Tumour) could be usefully included. This subgroup consisted of 82 patients. Among these patients the vast majority had either stage II (n = 56) or stage III (n = 20) disease at time of presentation. Only one stage IV and three stage I cases were present in this group, with stage unknown in two cases. Because of the lack of data for other stages, our analysis for this group of patients consisted of comparison of gene expression in stage II versus stage III disease only. As TNM stage incorporates data on all clinical features which have shown independent value as predictors of outcome, analysis on the basis of other disease features at diagnosis was not deemed likely to be of use.

The final group of variables described patient outcomes at the end of follow-up. For patients followed up for a fixed period of time, differential gene expression analysis based on vital status at the end of follow-up (dead or alive) could be of great interest as it has the potential to identify specific gene expression profiles associated with better or worse outcomes. For the data used in this study however this approach was difficult as the length of follow up varied widely between patients and the length of follow-up was, not unexpectedly, a key indicator as to whether patients would die during that period. Median time to end of follow up for patients alive at the end of follow up was 688 days, while median time to death among those dead at the end of follow up was 1110. Thus, it is likely that many of the patients who were described as alive at the end follow up in fact died at an earlier point than some of those dead at the end of follow up, but these deaths were not recorded due to earlier end of follow up. Thus analysis on the basis of vital status at the end of follow up was deemed unlikely to generate meaningful data.

An alternative approach was to study patients who had been followed until death and analyse for differences in gene expression in patients with longer as opposed to shorter time to death. One of the issues here was the relatively small number of patients who had been followed until death in many groups. For example, in those patients with sequence data from primary tumour at time of diagnosis available, the subgroup on which we attempted analysis on the basis of stage at diagnosis, only 6 out of 82 had been followed until death. Analysis of such limited data was not deemed likely to be useful. However, owing to shorter survival times in patients with metastatic disease we reasoned that among the group of patients with tissue sampled from distant metastases were more likely to be followed up until death. Indeed, when the subgroup of patients whose sample came from a distant metastasis, over half (36 out of 68) had been followed until death, allowing us to divide these 36 into those who survived longer than the median survival time and those who did not. Rather than raw survival time from diagnosis, the new variable of 'Survival post tissue sampling' was used, as discussed previously under the heading of download and cleaning of data.

2.4 Selection of Genes for Study

Interrogation of the counts data frame with the nrow() function in R demonstrated that it contained expression data for 20531 genes. For our initial analyses, we studied a subset of 10% of these genes with the highest standard deviations in order to reduce time required for computation. We reasoned that the subset of genes with highest standard deviation was likely to be enriched for genes with significant variation between subpopulations of specimens. This approach has been applied before most notably by the The Cancer Genome Atlas Network in their analysis of genomic features of melanoma samples from 331 patients, not included in the database used in our study (Cell 2015 paper).

For the comparison of gene expression profiles in distant metastases compared to primary tumour, several microarray studies had been conducted previously and we were interested to assess the replicability of the findings of these studies in our own data. Two previous studies, one by Mauerer et al and one by Jaeger et al, had compared expression profiles in primary versus metastatic melanoma and for both of these lists of genes found to be differentially expressed between the two groups were available (see Introduction). For each study, a vector was constructed in R, listing the differentially expressed genes. This was used to select columns from the counts data frame describing expression of these genes in our data set in order to assess for differential expression of these in cases included in TCGA's data.

5. DESeq2 Analysis

DESeq2 is a bioconductor package for analysis of RNA-Seq based gene expression data, discussed in the Introduction. We used DESeq2 to analyze for genes above the tenth percentile in terms of variability of expression (as assessed by the standard deviation) which were differentially expressed between different subsets of cases, defined by clinical variables. DESeq2 analysis conducted sought differentially expressed genes in different groups defined as follows:

1. Distant metastases versus primary tumour

2. Regional lymph node disease versus primary tumour

3. Among patients with primary tumour specimens from time of diagnosis available, stage III disease versus stage II disease

4. Among patients with sample obtained from distant metastasis who were followed up until death, comparison was made of those surviving longer than the median time post-biopsy with others.

6. Gene Ontology Enrichment (GOE) Analysis

The Bioconductor package topGO was used to assess for enrichment of gene ontology categories in the genes found to be differentially expressed in distant metastases versus primary tumour and lymph node disease versus primary tumour. A named vector was created in R, with each member of the vector representing one of the 2053 genes analysed and the value of the vector member indicating whether the gene was considered 'of interest' , i.e. differentially expressed, or not. We then conducted analysis to determine if certain functional categories of genes according to Entrez gene annotations were over-represented in the differentially expressed genes as compared to the set of all genes analyzed. Enrichment analysis was not carried out for genes differentially expressed in lymph node disease versus primary as the relatively small number of genes included (24) was not thought likely to generate meaningful results.

Part 3: Results

1. Demographics of Patients Included

468 patients had both RNA-Seq2 sequence data and clinical information. Of these 180 were female. Median age at diagnosis of patients included was 58 years, with an interquartile range of 47 to 71 years. Height and weight were not analyzed as the clinical implications of this data were not clear. 22 out 468 had had a prior diagnosis of cancer prior to emergence of melanoma. 445 out of 468 were of white race, 12 were Asian, just one African American and race was unknown for 10 patients. 11 of the white patients were of Hispanic ethnicity.

With regard to time of tissue sampling, just 116 specimens had been obtained at the same time as melanoma was first diagnosed, i.e. days to sample submission = 0. The median delay from diagnosis to submission of specimen to TCGA was 352 days, just under a year. For 81 patients the delay was five years or more. Site of tissue sampling was from primary tumour in 105 patients, from involved regional lymph nodes in 200 and from distant metastases in 68. This data was not recorded in the remainder.

Only 12 patients had more than one confirmed primary tumour present at diagnosis. 384 were confirmed as having one primary only. For 72 the information was not recorded but it is reasonable to assume that a large majority of these would have just one primary at diagnosis. Of the 12 who had more than one primary tumour confirmed, 11 had two primaries identified and one had three.

The overall TNM stage at diagnosis was recorded as per table 3.1 (see below). It should be noted that although only stage IV disease features distant metastases at diagnosis, the number of samples taken from distant metastases (n = 68) exceeded the number of patients with stage IV disease confirmed at diagnosis (n = 23). Indeed, only 6 of the samples from distant metastases came from patients with stage IV disease at diagnosis. Samples from distant metastases in patients classified as having lower stage disease at presentation was mostly due to significant delay between diagnosis of melanoma and receipt of tissue by TCGA, with presumed progression of disease during the intervening time. The median time to specimen submission in patients with samples obtained from distant metastases that were not stage IV at presentation was 1111 days, over three years. However, two of these patients did have specimen obtained at time of diagnosis and one just four days thereafter, suggesting that in these three patients at least stage of disease at diagnosis may have been incorrectly recorded in the data set.

Other potentially useful variables describing disease features at diagnosis with prognostic implication provided limited information due to the high number of NA values. Breslow depth was

Stage at initial diagnosis	**Number of samples**
Not available	38
Stage 0	7
Stage I	76
Stage II	138
Stage I or Stage II NOS	14
Stage III	173
Stage IV	23

Table 3.1 Pathological stage of disease at diagnosis in patients in the sample studied

unknown for 111 out of 468 patients, ulceration status of the primary tumour was unknown or unavailable for 157 patients and mitotic rate per high power field was unavailable for 297. Although serum LDH level was listed as one of the variables provided, this in fact turned out not to be known for any of the patients included in the version of the dataset which we accessed.

Follow up of the patients in this data set was also somewhat unreliable. Of 468 patients, 307 were still alive when clinical follow up ended while 161 were dead. Time to death was known for all but two of the deceased. The time to last follow up was unclear for 7 patients whose follow up was described as 'Completed' in the clinical data and who were alive at the end of follow up. Of the remaining 300, the first quartile for length of follow up was just 28 days, indicating that clinical follow up data was available for less than one month post diagnosis in at least one quarter of the patients described as being alive at the end of follow up.

Of 161 patients followed to death, just 9 were tumour free by the time they died (presumably of unrelated causes). 146 were confirmed as not tumour free at time of death. For 6, their tumour status at time of death was unknown. Of those still alive at the end of follow up, 209 were tumour free, 77 confirmed alive but with tumour and tumour status was unavailable in 21. However, as mentioned above, the brief follow up times for many patients meant that interpretation of these figures was difficult

3.2 DESeq2 Analysis of Primary Tumour vs. Regional Lymph Node vs. Distant Metastases

The simplest comparison which allowed the largest groups was gene expression in primary tumour (PT) compared to distant metastasis (DM) or regional lymph nodes (RLN). As previously mentioned there were 105 samples from primary tumour, 200 from regional lymph nodes and 68 from distant metastases.

As discussed in methods, for our initial analysis we focussed on genes with a standard deviation among the highest 10% of those included. DESeq2 was used to compare primary tumour versus distant metastasis, primary tumour versus regional lymph nodes and regional lymph nodes versus distant metasastasis for significant differences in gene expression.

41 out of 2053 genes analysed were found to be significantly differentially expressed in distant metastases as compared to primary tumour, with a p-value adjusted for multiple analyses of less than 0.05 taken to indicate significance. These are summarised in table 3.2. The 24 differentially expressed in tumour-involved regional lymph nodes as compared to primary are shown in table 3.3.. This finding of fewer genes differentially expressed to a significant degree in regional lymph nodes versus primary as compared to distant metastasis versus primary is in line with expectations, given that disease is generally considered to advance from primary tumour to more advanced disease with distant metastases with regional lymph node involvement and intermediate stage.

In order to confirm that genes being detected as differentially expressed in distant metastases compare to primary tumour were mostly due to genuine differences in tumour biology between the two sites, rather than random false positives due to the large number of genes being tested, we ran twenty analyses comparing expression of groups in the same size as the distant metastases group (n = 68) and the primary tumour group (n = 105) but with group member chosen randomly. For these twenty analyses, median number of genes detected as differentially expressed was 7, with a range from 2 to 14. Thus it is likely that the detection of 41 out of 2053 genes as significantly differentially expressed in the 68 distant metastases compared to the 105 primary tumours did indeed reflect differences in important differences in the behaviour of cancer cells with regard to transcription.

Of these genes only one, SYNM was significantly differentially expressed in both distant metastases and regional lymph nodes compared to primary tumour. (SYNM encodes the protein synemin which is believed to play a role in the adhesion of cells to the protein network in their local tissue environment). Such a small intersection of differentially expressed genes is surprising and suggests that the mechanisms by which melanoma metastasises to distant organs may differ from those by which it spreads to regional lymph nodes. Still more surprising is the fact that the one gene differentially expressed in both groups compared to primary had it expression altered in different directions. Table 3.2 demonstrates that the log2fold change of SYNM in distant metastases is -0.892, indicating an expression level just over half as great as that seen in primary tumours. This concords with previous studies of the effect of SYNM in view of evidence that SYNM is downregulated in other human cancers and this reduced expression has been correlated with more aggressive clinical behaviour in breast cancer. On the other hand SYNM was modestly but significantly upregulated in lymph node disease compared to primary tumour. The relationship between cell adhesiveness and spread of malignancy beyond its site of origin is likely to be a complex one; metastatic cells must first cleave away from the main body of tumour to enter blood or lymph but subsequently successfully adhere to microscopic structures in the tissue which they invade. The different directions of SYNM regulation in lymph node disease versus distant metastases may reflect such complexity.

Gene name	Entrez ID	Base Mean	Log2-Fold-Change	lfcSE	stat	Pvalue	padj
PNMA5	114824	26.91026	1.6534905	0.2067387	7.997973	1.264837e-15	2.596710e-12
VAT1L	57687	255.68206	-1.1930841	0.2086274	-5.718732	1.073222e-08	1.101662e-05
ADAM19	8728	2284.73864	0.9942959	0.1832004	5.427368	5.719114e-08	3.913780e-05
EGR1	1958	9503.04824	-0.7874400	0.1703695	-4.621953	3.801446e-06	1.254722e-03
FAM84B	157638	2455.11531	-0.8917475	0.1929048	-4.622734	3.787149e-06	1.254722e-03
SYNM	23336	8205.16597	-0.8920456	0.1909669	-4.671204	2.994393e-06	1.254722e-03
THBS1	7057	9584.99859	0.8517716	0.1852728	4.597390	4.278155e-06	1.254722e-03
IGF1R	3480	8836.73797	-0.6092997	0.1348962	-4.516802	6.278042e-06	1.611103e-03
GRIK3	2899	790.63694	-0.9278101	0.2085600	-4.448648	8.641257e-06	1.857463e-03
LOC151162	151162	10106.52306	-0.5621730	0.1266507	-4.438767	9.047553e-06	1.857463e-03
TSPAN7	7102	4786.427	-0.8810663	0.2010215	-4.382945	1.170859e-05	0.002185249
HSPA7	3311	1262.694	-0.8486585	0.1968648	-4.310869	1.626140e-05	0.002782055
ZFP106	64397	20839.610	-0.6572983	0.1534042	-4.284748	1.829466e-05	0.002889149
NRP1	8829	2916.065	0.6733424	0.1686313	3.992986	6.524644e-05	0.009567925
GAB2	9846	5601.541	-0.5482851	0.1381387	-3.969093	7.214685e-05	0.009874499
GNS	2799	23151.992	-0.5196117	0.1315125	-3.951043	7.781133e-05	0.009957765
ISG15	9636	3177.173	0.6983543	0.1773757	3.937148	8.245592e-05	0.009957765
NGFR	4804	6228.044	0.7978342	0.2047715	3.896217	9.770672e-05	0.011143994
NDRG1	10397	27581.443	-0.7160878	0.1846127	-3.878866	1.049447e-04	0.011339549
GPX3	2878	5486.639	-0.7269875	0.1907704	-3.810798	1.385188e-04	0.014218956
CYP17A1	1586	7.248821	-0.7030209	0.18575864	-3.784593	0.0001539601	0.01505143
ETV5	2119	14872.383444	-0.4319270	0.11468185	-3.766307	0.0001656802	0.01546098
ATP1B1	481	4420.099181	-0.6731271	0.18010279	-3.737461	0.0001858882	0.01659255
ATXN7L3B	552889	7433.299153	-0.3294595	0.08845479	-3.724609	0.0001956182	0.01673351
SFTPB	6439	8.040516	0.7530116	0.20491637	3.674727	0.0002381042	0.01955312
TRIB1	10221	7795.180427	-0.4843786	0.13479741	-3.593382	0.0003264139	0.02577414
NOV	4856	5848.728472	0.7162553	0.20021533	3.577425	0.0003469959	0.02638454
EDN3	1908	660.186984	-0.6436706	0.18076573	-3.560800	0.0003697268	0.02710890
ALDH1A3	220	14987.625313	-0.7224305	0.20362829	-3.547790	0.0003884772	0.02750151
SPP1	6696	38258.210927	-0.7044106	0.19912982	-3.537444	0.0004040201	0.02764844
CHRDL1	91851	683.2553	-0.7307687	0.20788608	-3.515236	0.0004393631	0.02909718
F2R	2149	9031.3786	-0.6369625	0.18188151	-3.502074	0.0004616514	0.02961782
INPP5F	22876	6203.0570	-0.5014949	0.14467990	-3.466238	0.0005277966	0.03283535
APOD	347	79018.3040	-0.6889502	0.19964691	-3.450843	0.0005588379	0.03374395
COL15A1	1306	5920.4954	0.5690201	0.16614992	3.424739	0.0006153906	0.03509436
TRIB3	57761	4004.7684	-0.4852155	0.14162770	-3.425993	0.0006125566	0.03509436
TGFBI	7045	17081.8893	0.6456996	0.18910080	3.414579	0.0006388073	0.03544517
ZMIZ1	57178	7512.7483	-0.3185038	0.09418671	-3.381622	0.0007205928	0.03893097
ANXA2	302	38154.5323	0.4024951	0.12018324	3.349012	0.0008110036	0.04269206
ITGA10	8515	3923.9422	-0.6940738	0.20807927	-3.335622	0.0008510882	0.04368210
SPRY4	81848	15949.68	-0.5948785	0.1795975	-3.312288	0.0009253627	0.04633584

Table 3.2: *List of genes in top 10% in terms of variability of expression (standard deviation) significantly differentially expressed in distant metastases as compared to primary tumour. Significance taken as p-value adjusted for multiple testing less than 0.05. Genes ordered from lowest to highest p-value. Log2-Fold-Change > 0 indicates higher expression in distant metastases compared to primary tumour. Log2-Fold-Change < 0 indicates lower expression in distant metastases compared to primary tumour.*

Gene name	Entrez ID	Base Mean	Log2-Fold-Change	lfcSE	stat	Pvalue	padj
SCIN	85477	2389.153571	-0.7667814	0.13535778	-5.664849	1.471541e-08	3.021074e-05
SDC2	6383	4778.360379	0.6750001	0.12805495	5.271175	1.355530e-07	1.391452e-04
CYP11B1	1584	1.255849	-0.4894565	0.09502461	-5.150840	2.593224e-07	1.774630e-04
AQP1	358	8115.037407	0.6116671	0.12592823	4.857267	1.190170e-06	6.108545e-04
FGL2	10875	3023.991259	0.5945656	0.13350606	4.453473	8.449230e-06	3.095918e-03
STATH	6779	18.426018	-0.4055083	0.09135627	-4.438757	9.047982e-06	0.003095918
SYNM	23336	7910.566065	0.5560766	0.12915717	4.305426	1.666650e-05	0.004888047
IGFBP2	3485	8268.625234	0.5595811	0.13345319	4.193089	2.751818e-05	0.007061852
PTN	5764	605.363232	0.5515058	0.13509251	4.082430	4.456715e-05	0.010166263
CYP17A1	1586	5.685141	-0.4548180	0.11445229	-3.973865	7.071555e-05	0.014517901
ZNF697	90874	4230.909996	-0.4107765	0.1039887	-3.950204	7.808465e-05	0.01457344
CFI	3426	2086.481064	0.5223998	0.1335295	3.912243	9.144295e-05	0.01564437
RGS5	8490	4878.023949	0.5086354	0.1321700	3.848342	1.189199e-04	0.01878020
ECE1	1889	6951.713570	-0.4081931	0.1073419	-3.802739	1.431052e-04	0.02098536
C4A	720	5080.437405	0.4856630	0.1288175	3.770162	1.631415e-04	0.02232863
CPXM2	119587	935.322272	0.5057705	0.1348446	3.750765	1.762958e-04	0.02262096
MRGPRX3	117195	233.451896	0.4892490	0.1312362	3.728003	1.930032e-04	0.02330798
PRH2	5555	33.759974	-0.5009912	0.1350618	-3.709348	2.077939e-04	0.02370005
TNS1	7145	8497.398684	-0.4210604	0.1139535	-3.695020	2.198698e-04	0.02375751
KRT9	3857	6.740968	-0.3789670	0.1034013	-3.665012	2.473272e-04	0.02538813
LAMC2	3918	359.6092	-0.4818238	0.1354688	-3.556713	0.0003755236	0.03671191
ITIH5	80760	5034.3191	0.4688389	0.1341156	3.495781	0.0004726758	0.04410925
DCBLD2	131566	7813.3371	-0.3900584	0.1128447	-3.456597	0.0005470431	0.04766108
TBX2	6909	5464.4830	0.3927680	0.1137913	3.451651	0.0005571680	0.04766108

Table 3.3: *List of genes in top 10% in terms of variability of expression (standard deviation) significantly differentially expressed in regional lymph nodes as compared to primary tumour. Significance taken as p-value adjusted for multiple testing less than 0.05. Genes ordered from lowest to highest p-value. Log2-Fold-Change > 0 indicates higher expression in regional lymph nodes compared to primary tumour. Log2-Fold-Change < 0 indicates lower expression in distant metastases compared to primary tumour.*

3.3 Analysis of Genes Previously Reported in Metastases versus Primary Tumour

Mauerer et al had previously reported the results of microarray analysis comparing RNA expression levels in 20 melanoma primaries to 20 distant metastases. Of the 189 genes they found to be differentially expressed between the two groups, 178 had expression values available in our data. Of these, 17 were found to be differentially expressed in distant metastases compare to primary tumour and 9 were differentially expressed in regional lymph node disease compared to primary tumour. Table 3.4 is a complete list of these genes. As can be seen in that table, two genes, GPX3 and ZNF185 showed altered expression in both distant metastases and involved regional lymph nodes in our dataset as well as showing altered expression between the two groups in Mauerer's study

It is noteworthy that, of 178 genes analysed in this analysis, 17 (9.5%) showed altered expression levels in distant metastases compared to primary tumour. In contrast, analysis of the

2053 genes with the highest standard deviation demonstrated altered expression levels in 41 genes (2%). Similarly, for involved regional lymph nodes, 9 of 178 genes reported by the previous study (5.1%) showed evidence of altered expression levels compared with 24 out of 2053 (1.2%). Thus, selection of genes previously reported as significantly differentially expressed in primary versus metastatic tumour (even though the previous report had used a microarray-based methodology rather than direct RNA sequencing) appeared slightly more effective for enriching the set of genes analysed for those differentially expressed between the two groups than did selection of more variable genes according to the standard deviation of count reads.

Distant Metastases	**Regional Lymph Nodes**	**Both**
SPP1|6696	CFI|3426	GPX3|2878
NELL1|4745	STC1|6781	ZNF185|7739
IGFBP2|3485	GPX3|2878	
C7|730	BNC1|646	
GPX3|2878	ANK3|288	
SELE|6401	SERPINB2|5055	
ANK3|288	ZNF185|7739	
ALDH3B2|222	AKR1B10|57016	
CA12|771	KRT2|3849	
SCUBE2|57758		
CA2|760		
ZNF185|7739		
GJA1|2697		
CAPG|822		
CXADR|1525		
MMP1|4312		
SCEL|8796		

Table 3.3: *List of genes reported as differentially expressed in distant metastases compared to primary tumour in Mauerer et al's microarray analysis also found to be differentially expressed between those groups in TCGA's RNAseq data. The gene name and Entrez gene ID of all genes are given, separated by '|'.*

Jaeger et al reported a similar microarray analysis comparing RNA expression in 19 primary tumours to that of 22 metastases. Their group reported 269 genes differentially expressed in primary compared to metastatic melanoma. 99 of these genes had also been reported as differentially expressed in primary versus metastatic melanoma by Jaeger's group.246 of the genes reported as differentially expressed by Jaeger's group had count data available in our dataset and 24 of these were significantly differentially expressed in distant metastases versus primary tumours. In contrast only 4 out of 246 were differentially expressed in lymph node disease compared to primary tumour in our dataset. Table 3.4 summarizes the genes which were differentially expressed in distant metastases as compared to primary tumour in both our own data and in Jaeger et al's study. It is interesting to note, as table 3.4 demonstrates, that while the number of genes that were differentially expressed in lymph nodes as compared to primary tumour in our data which had been reported in Jaeger's study was only four, two of these (THBD and ETS2) were also reported as differentially expressed in distant metastases. Also of note is the fact that approximately 9.75% of genes reported as differentially expressed by Jaeger et al were differentially expressed in distant metastases versus primary tumour in our study, similar to the rate of 'hits' for genes reported by

Mauerer et al and greater than the 2% of genes in the 10% most variable found to be differentially expressed.

Distant Metastases	Regional Lymph nodes	Both
CXADR|1525	KRT10|3858	THBD|7056
GJA1|2697	THBD|7056	ETS2|2114
NEBL|10529	ETS2|2114	
MAST4|375449	LTF|4057	
FHOD3|80206		
ALDH3B2|222		
PYCARD|29108		
SCUBE2|57758		
NTRK2|4915		
THBD|7056		
ETS2|2114		
HLF|3131		
F2RL1|2150		
PALMD|54873		
FXYD3|5349		
FCGBP|8857		
CFH|3075		
KRT19|3880		
LSR|51599		
ZNF185|7739		
ABCC3|8714		
MID2|11043		
JAG1|182		
NFIB|4781		

Table 3.4: *List of genes reported as differentially expressed in distant metastases compared to primary tumour in Jaeger et al's microarray analysis also found to be differentially expressed between those groups in TCGA's RNAseq data. The gene name and Entrez gene ID of all genes are given, separated by '|'.*

Of the 99 genes differentially expressed in both the Jaeger and Mauerer studies, only 5 were also differentially expressed in our studies. This 'hit' rate of less than 5% was lower than for genes found differentially expressed in any one study alone. This somewhat surprising result may have been due to the relatively small number of genes expressed in both studies. The five genes found to be differentially expressed in distant metastases versus primary tumour in both our study and the two older microarray experiments were as follows:

1. CXADR (ID = 1525) - NCBI Genbank describes this as a human membrane receptor protein, known to bind viral particles of Coxsackie and adenovirus. Genbank does not report any previous evidence of a role for this gene product in human cancer. It may have a role in cell-cell adhesion.[45]

2. GJA1 (ID = 2697) - Genbank describes this as a member of the connexin family, also known as connexin 43 and important in the formation of gap junctions via which electrolytes can be trafficked between cells and in adhesion between cells. There is limited evidence to suggest a role for the gene product in cancer; Ghosh et al reported that its downregulation in laboratory cultures of malignant cells lead to increased killing of these cells by low dose radiation treatment. The gene was

upregulated in distant metastases compared to primaries in our data (a log2fold change of 0.67) suggesting it may promote aggressive behaviour and survival of cancer cells.[46]

3. ALDH3B2 (ID = 222) - An enzyme, also known as ALDH8, of the aldehyde dehydrogenase family with a role in the metabolism of toxic aldehydes. Although no definite evidence of a role in cancer is reported in Genbank, expression in melanocytes is certainly abnormal, as expression in normal tissue has only been reported in the salivary gland.[47]

4. SCUBE2 (ID = 57758) - A signal transduction and tumour suppressor gene, downregulation of which is believed to play a role in breast cancer pathogenesis, according to Genbank. Of note, its log2fold change in our data in distant metastases compared to primary tumour was -0.9135227, indicating that expression levels in distant metastases were reduced by almost one half.

5. ZNF185 (ID = 7739) - Genbank reports that the product of this gene is a DNA-binding zinc finger protein with a role in cell proliferation and apoptosis, both processes affected during carcinogenesis. Its silencing has been associated with prostate cancer progression. In our data its log2fold change in metastases compared to primary tumour was 0.5234868, meaning that in contrast to prostate cancer upregulation of this gene may promote melanoma progression.

Figure 3.1 demonstrates graphically the relationship between the two microarray studies and the degree to which the findings of each was replicated in our data.

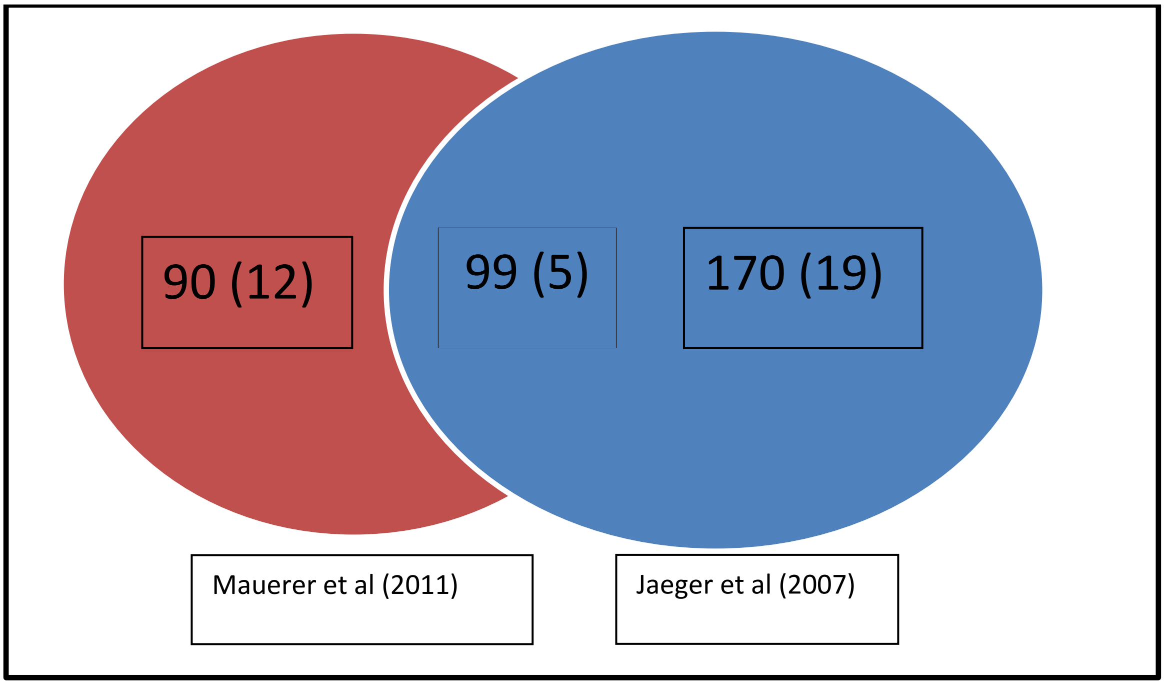


Figure 3.1: *Comparison of findings of previous microarray studies comparing expression profile in distant melanoma metastases to primary tumours. The figures represent the number of genes found to be differentially expressed between the two groups in the two studies. Figures in brackets represent the number of genes in each category which were differentially expressed according to the RNA-Seq data used in our study. Note that total figures in this figure represent all genes reported in the earlier studies, while figures quoted in the text mainly refer to those which were reported as differentially expressed in the previous studies and for which counts were available in our own data.*

3.4 Gene Ontology Enrichment Analysis of Genes Differentially Expressed in Distant Metastases

41 genes of the 2053 most variably expressed were found to be differentially expressed in distant metastases compared to primary melanoma tumours. We were interested to discover whether genes were more likely to be differentially expressed in distant metastases compared to primary tumour based on biological processes in which they participated. Accordingly we carried gene ontology enrichment analysis using the topGO R package to assess if the list of 41 differentially expressed genes was significantly. Analysis was based on annotations by the Entrez gene database. Table 3.5 demonstrates the 12 categories of biological process which were significantly ($p < 0.05$) enriched in the set of genes differentially expressed in distant metastases.

Process ID	Process	Total genes with in category	Genes differentially expressed DM vs PT	Expected number in differentially expressed set	Classic Fischer p-value
GO:0007162	Negative regulation of cell adhesion	10	2	0.16	0.0095
GO:0002683	Negative regulation of immune system	14	2	0.22	0.0186
GO:0009653	Anatomical structure morphogenesis	121	5	1.94	0.0230
GO:0045596	Negative regulation of cell differentiation	17	2	0.27	0.0271
GO:0050900	Leukocyte migration	17	2	0.27	0.0271
GO:0009887	Organ morphogenesis	46	3	0.74	0.0296
GO:0007417	Central nervous system development	47	3	0.75	0.0313
GO:0048646	Anatomical structure formation involved in morphogenesis	47	3	0.75	0.0313
GO:0008283	Cell proliferation	87	4	1.39	0.034
GO:0048585	Negative stimulus response	52	3	0.83	0.0410
GO:0030030	Cell projection organisation	53	3	0.85	0.0432
GO:0030155	Regulation of cell adhesion	22	2	0.35	0.0443

Table 3.5: *Gene ontology categories significantly enriched among genes differentially expressed in distant metastases compared to primary tumour compared to the data set in general*

3.5 Analysis of Primary Tumour Samples from Time of Diagnosis by Stage

82 samples were available from primary tumour which been obtained at time of diagnosis, i.e. day from diagnosis to specimen submission = 0. This subset was of particular interest as such circumstances of biopsy represent the circumstances under which melanoma tissue is most

commonly biopsied in clinical practice. Further, features of disease at diagnosis are the features of disease at specimen sampling in this group, with no need to consider the possibility that disease has advanced or regressed due to treatment between diagnosis and sampling.

Of the 82 available samples of this subgroup, 56 were stage II and 20 were stage III. There were insufficient samples present representing other stages to make their analysis worthwhile. Nonetheless, we proceeded to compare gene expression in stage II and stage III disease in this population using DESeq2. We found just one gene differentially expressed between disease of the two stages; HSD3B2. This finding is of interest as it suggests that there is relatively little overall change in genomic features of melanoma primary tumour between stage II and stage III and may point towards the importance of other factors in driving local lymph node invasion, the key process causing upstaging from stage II to stage III.

3.6 Analysis of Gene Expression Profile in Distant Metastases by Survival Post-Biopsy

We divided 36 samples from distant metastases in patients who had been followed up until death into a group of 18 who lived longer than the median survival post biopsy and 18 who survived for less time. We then performed DESeq2 analysis to compare gene expression patterns between the two groups. Over 33% of genes were reported as significantly differentially expressed between the two groups, with 823 our 2053 genes in the top 10% most variably expressed genes showing differential expression. This is an enormously high number which likely reflects the chaotic and random transcription behaviour in metastases. While these random variations may average out over a large number of distant metastasis samples, allowing informative comparisons to be made between metastases as a group and less advanced tumours, further attempts to compare two relatively small groups both composed entirely of samples from metastases was likely to prove futile. While it is possible that increased or decreased expression of a small subset of the 823 genes reported as differentially expressed in the good versus bad survival groups do have a role in driving survival differences, it is likely that these meaningful 'signals' would be impossible to extract from such excessively noisy data.

Part 4: Discussion

4.1 Clinical Data Quality Considerations

Our study represents an attempt to correlate differences in gene expression profile of melanoma as assessed by RNA-Seq analysis of patient samples with clinical features of the disease with relevance to prognosis and treatment of the condition. We made use of sequence data with matched clinical data which is freely available via the data portal of the Cancer Genome Atlas (TCGA) consortium and which is matched to clinical data from the same patients

Unfortunately our analysis was somewhat limited by the quality of the clinical data made available. Many of the variables in the clinical data frame describe the disease status of the patient at the time of diagnosis, while tissue sampling sequence data generation did not take place until years later in many cases. The median time from diagnosis to obtaining the specimen was 352 days, just under a year. Such a long time lapse would give ample opportunity for clinical status of the patient to change dramatically, due to factors such as the administration of treatment or metastasis of an initially lower stage tumour to distant sites. Therefore features of the disease at diagnosis were not necessarily relevant to the clinical situation at the time sequence data was generated, meaning any attempt to draw conclusions from the correlation of these clinical data with sequence reads highly unreliable, except in carefully selected subpopulations of the patients represented in the database.

Missing values were also a problem in much of the clinical data (for example, the serum LDH level was recorded as unknown in all cases, while mitotic count rate of the primary tumour was unknown in 297 out 468 patients included). Also the length of time for which patients were followed up by TCGA varied widely, with patients who died during follow up having a much longer median time to last follow up (1110 versus 688 days) meaning that death during follow up was as likely to reflect greater duration of follow up as any meaningful biological difference in the behaviour of the tumour. Finally, information on other clinical data beyond those provided in the dataset could have been of value In particular data on treatment were lacking. Only receipt of radiotherapy and receipt of interferon (an older mode of chemotherapy) were documented. Patients who received the newer agents discussed in the Introduction - including the small molecule agents vermurafenib and dabrafenib and the monoclonal antibodies ipilimumab and nivolumab - would have been an interesting subgroup for further study but there was no means of ascertaining which, if any, patients were treated with these.

There are ethical and pragmatic reasons why clinical data collection for this dataset may fall short of the ideal. Given that the data is freely available, concerns about reidentification of patients may preclude provision of detailed information on treatment received. Uniform long term follow up of patients from several different sites is often achievable in the context of a well-resourced clinical trial but can be extremely difficult for an observational study run by a research consortium without significant industry support. Perhaps then the most useful approach is to collect data pragmatically and rely on the size of the sample (468 melanoma cases in total) and careful data interpretation by researchers to enable useful conclusions to be drawn despite incomplete and sometimes difficult to interpret clinical data. Certainly, while more complete data would have allowed us to carry out several further interesting analyses, we were able to draw some conclusions based on differential gene expression and gene ontology enrichment analyses of TCGA's RNA-seq data in its current form.

4.2 Differential gene expression in primary tumour versus distant metastases and regional lymph nodes

Owing to the issues with many of the clinical variables discussed, we first sought to analyse specimens on the basis of site from which sample was obtained (primary tumour versus regional lymph node versus distant metastasis). This was an appealing starting point for the analysis as site of sampling is clearly and unambiguously related to the sequence data generated in a way that features of disease at diagnosis (which may have occurred years prior to sequence data generation) is not. Current understanding of cancer biology suggests that, while diseased tissue at all of these sites should show behaviour different to normal tissue, cancer tissue from distant metastatic sites ought to show relatively more aberrant gene expression than lymph node disease, with primary tumour tissue most similar to normal tissue. While normal tissue samples were not available for comparison, our analysis of the 2053 genes with the most variable expression in the dataset did suggest that more genes were significantly differentially expressed in distant metastases compared to primary tumour than in lymph node disease compared to primary tumour (41 versus 24). This finding offers reassurance that the sequence data provided in the counts data frame is reliable. Furthermore, twenty runs of DESeq2 analysis using groups of similar size to those in the distant metastasis versus lymph node category but with randomly selected samples in each group produced a median of 7 differentially expressed genes per run, less than one sixth of the number found when distant metastases were compared to primary tumour, thereby indicating our results were highly unlikely to be the result of random variation in two groups substantively similar in terms of expression profile.

Gene ontology enrichment analysis was carried out to identify whether genes related to any specific biological processes were over-represented in set of 41 genes differentially expressed in distant metastases compared to primary tumour. This analysis identified 12 functional gene categories over-represented in the differentially expressed set, summarised in Table 3.5 of the results section. There is a clear link between some of these functions and the hallmarks of cancer discussed in the introduction to this thesis. Most notably, two of the categories enriched in the differentially expressed list - 'negative regulation of immune system' and 'leucocyte migration' - could affect interaction of the tumour cells with the immune system. Evasion of immune mechanisms for killing of abnormal, potentially harmful cells is one of the eight hallmarks of cancer discussed in the introduction to this thesis. Targeting the mechanisms by which melanoma cells achieve this outcome has lead to the development of clinically effective anti-tumour therapy using agents such as ipilimumab which has proven efficacy in extending life in metastatic melanoma. It is likely that changes in transcriptions of gene products regulating interaction with the immune system confer a greater capacity to evade the immune system upon aggressive melanoma which has metastasised to distant organs than on comparatively less lethal early stage disease confined to the site of primary tumour.

The functional categories 'cell proliferation' and 'negative stimulus response' were also over-represented among genes differentially expressed in primary tumours versus distant metastases. Changes in expression of genes in these categories may also be related acquisition or enhancement of the hallmarks of cancer, by promoting uncontrolled cell growth and causing insensitivity to negative growth factors, respectively.

Although the link is less direct, it is likely that a further two of the enriched categories- 'negative regulation of cell adhesion' and 'regulation of cell adhesion' - have a roll in another of the hallmarks of cancer, namely metastasis and invasion. There is considerable preclinical data to suggest that at least some of the molecules involved in cellular adhesion are important in allowing metastatic deposits to colonize tissue distant from the primary site of the tumour in several cancers.[50] Our findings here suggest a possible role for these mechanisms in melanoma metastasis in the clinical setting of human disease also. Nonetheless the relationship between expression of cell adhesion molecules is unlikely to be straightforward. As noted in the Results section, metastatic cells must firstly break away from the tissue of the primary malignancy before becoming adherent again at a distant site. This complexity is perhaps reflected in the opposite direction of expression change in distant metastases as compared regional lymph nodes of the only gene found differentially expressed compared to primary tumour in both groups, namely the cell adhesion regulator SYNM.

The physical conformation of cells is also important in determining how they interact with their surroundings, thus genes in the 'cell projection organisation' group may interact with genes involved in cellular adhesion to cause changes in cell adhesiveness.

Of the remaining categories of gene function significantly enriched, 'negative regulation of cellular differentiation' indicates that genes which cause cell to lose specialization of their function tend to be differentially expressed between primary tumour and distant metastasis. In the case of melanocytes, the main specialised function of the cells is production of melanin in response to sunlight exposure. Although not one of the hallmarks of cancer described by Hanahan and Weinberg, loss of specialization or dedifferentiation is widely recognized as a common occurrence in malignant cells. Indeed, extreme dedifferentiation of some subpopulations of tumour cells leads to the formation of 'cancer stem cells', malignant cells which exhibit pluripotency, the ability to take on the features of almost any other cell type in the body when exposed to the correct environment.[51, 52]

It is possible that loss of specialization is an important factor in allowing metastatic cells to survive in various organs other than those in which they occur in state of health, allowing cancer cells to take on characteristics more like those of the cell population in the tissue to which they spread. Dedifferentiation of cells usually implies regression to a more immature form closer to that seen in foetal tissue, thus this process could explain enrichment for genes with functions described as 'anatomical structure morphogenesis', 'organ morphogenesis', and 'anatomical structure formation formed in morphogenesis'. Morphogenesis is the process by which immature cells proliferate to form rudimentary tissues and organs, thus the process would be more active in more foetal-like cells. Enrichment of genes involved in central nervous system development is similarly not a surprise, as melanocytes develop from the same 'neural crest' embryological tissue as most neurons, thus advanced dedifferentiation of melanocytes might be expected to give expression profiles somewhat similar to the developing nervous system.[52]

Only one gene was common to both the set differentially expressed in distant metastases versus primary and the set differentially expressed in involved regional nodes versus primary. This gene, SYNM, was in fact downregulated in distant metastases and upregulated in lymph node disease. This surprising set of results may indicate that, in spite of a tendency to regard lymph node invasion as a step on the path to acquiring widely metastatic disease, the mechanisms by which these two adverse events occur may in fact be quite different. Decreased SYNM expression has previously been shown to correlate with increased aggressiveness of breast cancer, and so it reduced expression in distant metastases in melanoma is not in itself altogether surprising.[48] The product of this gene is involved in cell adhesion and the apparently contradictory results here may reflect the complexity of the relationship of this process to development of metastases, with malignant cells initially needing to reduce the strength of binding to the tissue of primary tumour in order to enter blood or lymph, but subsequently needing to adhere to tissue at other sites to establish a metastatic deposit.

4.3 Comparison to previous microarray expression studies

Two earlier studies had compared gene expression in primary tumour to distant metastases in clinical melanoma samples using microarray analysis. Jaeger et al's study[34] reported 269 differentially expressed genes between the two groups, while Mauerer et al's study[33] reported 189. Of these, count data was available for 246 and 179 in our sequence data, respectively. 99 genes were recorded by both groups. For each of the older studies we sought to determine how many of the genes differentially expressed according to their data could be demonstrated to have similar changes in expression in distant metastases in our dataset.

The numbers were perhaps surprisingly low. For both the Jaeger et al study and the Mauerer et al study, only just under 10% of genes reported by the microarray experiments were replicated in our data. Part of the explanation for this may be intrinsic discrepency between microarray and RNA-seq approaches for transcriptome study. Low rates of concordance have been reported, especially in genes with low median expression values. However, even in these cases, concordance rates of 40% are reported.[53] We propose that two other factors may have contributed to lack of agreement between our studies and the microarray data.

Firstly, the sample population in our data was much larger, expression data from 68 distant metastases and 105 primary tumours, as compared to 20 and 20 for the Mauerer study and 19 versus 22 for the Jaeger et al study. Thus there was less chance of random variation leading to false positives in TCGA's data owing to improved sample size. This hypothesis is supported by the fact that a relatively large number of genes in each of the microarray studies were not supported by the other; 170 out of 269 in Jaeger's study and 90 out of 189 in Mauerer et al. It should be noted that a previous study by Shyr and Li investigated sample size requirements for RNA Seq based experiments using a simulation-based method to calculate sample size requirements for adequate power in RNA-Seq experiments. For TCGA datasets for lung, colorectal and breast cancer, sample sizes of 18, 20 and 25 were required, respectively, for 80% power to detect changes.[54] It should be noted that for our main analysis of primary tumour, versus distant metastasis versus regional lymph node metastasis all group sizes greatly exceeded these values. However, it is unfortunate that Shyr and Li's work did not assess sample sizes that would be required for powerful analysis in the melanoma dataset and so, although our groups sizes were likely sufficient, this cannot be assumed with absolute confidence.

Secondly, closer inspection of TCGA's data revealed that samples from primary tumours tended to come from larger, more advanced cancers. The T score, which mainly depends on the size of the tumour, was T4 (the most advanced) in 87 out 105 primaries included in the analysis. Neither of the microarray studies give an indication the distribution of T scores for the cases included in their analysis. However, the proportion of T4 primaries in TCGA's data set is sufficiently high that it is speculate that the proportion in the other studies was lower. This is relevant as larger, more advanced primary tumour would be expected to have features more similar to distant metastases than smaller cancers. Thus the advanced stage of the majority of TCGA's primaries may have contributed to failure to detect some of the genes reported in previous studies as differentially expressed in primary tumour compared to distant metastases.

Although the replicability of findings from both earlier microarray studies was just under 10% using TCGA's RNA-Seq data, this was nonetheless higher than the approximately 2% rate of detection of differentially expressed genes in distant metastases versus primary tumour when the 10% most variable genes were analysed. This does suggest a modest increase in enrichment for differentially expressed genes when genes for study are selected on the basis of previously published data, as compared to when they are selected using a relatively crude statistical measure such as standard deviation. It should be noted that the TCGA consortium's recent clustering analysis of RNA seq data from melanoma specimens was based on the 1500 genes with highest standard deviation without reference to previous literature on melanoma expression profile.[36] We would suggest that further work of this kind should consider selecting genes for analysis on the basis of previous work indicating alterations in their expression levels in cancer of different stages of advancement in order to include more transcripts whose varying levels may correspond to differences in aggressiveness of tumour. It should be noted however, that genes reported as differentially expressed by both of the previous two microarray studies identified did not appear to be more likely to be differentially expressed here than those reported by one previous study. It should be noted, however, that only 99 genes were reported by both of the earlier studies, so small sample size and random variation may have explained this.

Five genes were differentially expressed in distant metastases compared primary tumour in both of the two microarray studies and in our analysis of TCGA's sequence data. For three of these GJA1, SCUBE2 and ZNF185, plausible explanations could be offered as to why their increased expression might lead to more aggressive tumour behaviour on the basis of previous studies of their effects in other tumour types. What mechanism, if any, links the remaining two genes - ALDH3B2 and CXADR - to aggression of tumour behaviour is unclear. Further research into the role of all these genes and their products in the context of malignant melanoma may be justified in future work

4.4 Other analyses

As mentioned previously, limited and heterogeneous clinical data limited our ability to conduct further gene expression analyses. However, the comparison of stage II to stage III primary tumours with sample available at diagnosis was interesting in that it revealed only one gene out of the 2053 most variable differentially expressed between these two groups; HSD3B2. The protein product of this gene is involved in the metabolism of steroids and related hormones such as testosterone. While mutations in this gene have been associated with increased risk of developing prostate cancer, and increased aggression of tumours of this type, this is likely due to the role of testosterone in promoting growth of prostate cancers. No such role has been described for the hormone in melanoma and no mechanism could be proposed to explain why HSD3B2 should be a driver of advancing stage in melanoma. It is likely therefore that the detection of differential expression between these two groups for this gene is a spurious finding due largely to random variation. Our previous analysis, in which we carried out 20 runs of DESeq2 analyses on samples randomly chosen the entire dataset did indicate that a low level of signal detection can occur by chance (see Methods and Discussion part 4.2).

In general, the main difference between stage II and stage III disease is lymph node involvement in stage III. The lack of significant differences in gene expression between primary tumours in these two stages may suggest factors other than expression profile differences account for differences in whether melanoma involves lymph nodes, for example proximity of the primary tumour to the nearest regional node. On the other hand, 23 out of the 2053 most variable genes were found to be differentially expressed in nodal disease compared to all primary tumours. It is possible however, that these differences in expression developed after spread to the node, with alterations in transcription profile necessary for melanoma cells to survive in the environment of a lymph node.

References:

1. LeBoit P.E., Burg G., Weedon D, Sarasain A. (Eds.): World Health Organization Classification of Tumours. Pathology and Genetics of Skin Tumours. IARC Press: Lyon 2006.

2. R. Shashanka and B. R. Smitha. Head and Neck Melanoma. ISRN Surgery, vol. 2012, Article ID 948302, 7 pages, 2012.

3. Belmar-Lopez C, Mancheno-Corvo P, Saornil MA, Baril P, Vassaux G, Quintanilla M, Martin-Duque P. Uveal vs. Cutaneous Melanoma: Origins and Causes of Difference. Clin Transl Oncol. 2008 Mar;10(3):137-42.

4. Khalil DN, Carvajal RD. Treatments for Non-Cutaneous Melanoma. Hematol Oncol Clin North Am. 2014 Jun;28(3):507-21

5. http://www.seer.cancer.gov/statfacts

6. Tas F. Metastatic Pattern in Melanoma: Timing, Pattern, Survival and Influencing Factors. J of Oncology. 2012, Article ID 647684, 9 pages.

7. NCCN Clinical Practice Guidelines in Oncology: Melanoma. V.2..2009

8. Melanoma of the skin. In: Edge, Byrd & Compton eds. AJCC Cancer Staging Manual, 7th ed. New York, NY: Springer, 2010

9. Balch CM, Soong SJ, Gershenwald JE, Thompson JF, Reintgen DS, Cascinelli N, Urist M, McMasters KM, Ross MI, Kirkwood JM, Atkins MB, Thompson JA, Coit DG, Byrd D, Desmond R, Zhang Y, Liu PY, Lyman GH, Morabito A. Prognostic factors analysis of 17,600 melanoma patients: validation of the American Joint Committee on Cancer melanoma staging system. J Clin Oncol 2001 Aug 15;19(16):3622-34.

10. John F. Thompson, Seng-Jaw Soong, Charles M. Balch, Jeffrey E. Gershenwald, Shouluan Ding, Daniel G. Coit, Keith T. Flaherty, Phyllis A. Gimotty, Timothy Johnson, Marcella M. Johnson, Stanley P. Leong, Merrick I. Ross, David R. Byrd, Natale Cascinelli, Alistair J. Cochran, Alexander M. Eggermont, Kelly M. McMasters, Martin C. Mihm Jr, Donald L. Morton, and Vernon K. Sondak. Prognostic Significance of Mitotic Rate in Localized Primary Cutaneous Melanoma: An Analysis of Patients in the Multi-Institutional American Joint Committee on Cancer Melanoma Staging Database. J Clin Oncol. 2011 Jun 1;29(16):2199-205.

11. Meral R, Duranyildiz D, Tas F, *et al*. Prognostic significance of melanoma inhibiting activity levels in malignant melanoma. *Melanoma Res*2001;11:627–632.

12. Hanahan D, Weinberg RA (2000). The Hallmarks of Cancer. *Cell* 100 (1): 57–70.

13. Hanahan, D.; Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell* 144 (5): 646–674.

14. Ultraviolet radiation and cutaneous malignant melanoma. Jhappan C, Noonan FP,, Merlino G. *Oncogene* (2003) 22, 3099–3112.

15. Ahmedin Jemal, Susan S. Devesa, Patricia Hartge, Margaret A. Tucker. Recent Trends in Cutaneous Melanoma Incidence Among Whites in the United States. JNCI J Natl Cancer Inst (2001) 93 (9):678-683.

16. Armstrong BK, Kricker A, English DR. Sun exposure and skin cancer. Australas J Dermatol. 1997 Jun; 38 Suppl 1:S1-6

17. Amaro-Ortiz A, Yan B, D'Orazio JA. Ultraviolet radiation, aging and the skin: prevention of damage by topical cAMP manipulation. Molecules. 2014 May 15;19(5):6202-19

18. Hill VK, Gartner JJ, Samuels Y, Goldstein AM. The genetics of melanoma: recent advances. Annu Rev Genom Human Genet 2013; 14:247-79

19. Bradford PT, Goldstein AM, Tamura D, Khan SG, Ueda T, Boyle J, Oh KS, Imoto K, Inui H, Moriwaki S, Emmert S, Pike KM, Raziuddin A, Plona TM, DiGiovanna JJ, Tucker MA, Kraemer KH. Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair. J Med Genet, 2011 Mar; 48(3):168-76

20. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, DiCara D, Ramos AH, Lawrence MS, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio RC, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton DL, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury JE, Davies MA, Gershenwald JE,Wagner SN, Hoon DSB, Schadendorf D, Lander ES, Gabriel SB, Getz G, Garraway LA, Chin L. A landscape of driver mutations in melanoma. Cell 2012; 150(2): 251 - 263.

21. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature. 2002 27;417(6892):949-54.

22. Eggermont AM, Spitz A, Robert C. Cutaneous melanoma. Lancet 2014; 383: 1816-27

23. Nickolaev SD, Rimoldi D, Iseli C. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet 2011 44(2):133-9.

24. Curtin JA, Fridlyand J, Kageshita T et al. Distinct sets of genetic alterations in melanoma. N Eng J Med. 2005; 353(20):2135-47.

25. Ball NJ, Yohn JJ, Morelli JG, Norris DA, Golitz LE, Hoeffler JP. Ras mutation in human melanoma: a marker of malignant progression. J Invest Dermatol. 1994; 102(3):285-90

26. Scortegagna M, Ruller C, Feng Y. Genetic inactivation or pharmacological inhibition of Pdk1 delays development and inhibits metastasis of Braf(V600E)::Pten(-/-) melanoma. Oncogene 2014; 33(34):4330-9

27. Pardoll DM. The blockade of immune checkpoints in cancer chemotherapy. Nature Reviews Cancer. 2012; 12:252-64.

28. Sarkar D, Leung EY, Baguley BC, Finlay GJ, Askarian-Amiri. Epigenetic regulation in human melanoma: past and future. Epigentics 2015 2015;10(2):103-21

29. Kunz M, Ibrahim SM, Koczan D, *et al*. DNA microarray technology and its applications in dermatology. *Exp Dermatol* 2004;13:593–606

30. Koh SS, Opel ML, Jia-Perng JW. Molecular classification of melanomas and nevi using gene expression microarray signatures and formalin-fixed and paraffin-embedded tissue. Modern Pathology (2009) 22, 538–546

31. Bittner M, Meltzner P, Chen Y, et al. A. molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406:536–540.

32. Becker B, Roesch A, Hafner C, et al. Discrimination of melanocytic tumors by cDNA array hybridization of tissues prepared by laser pressure catapulting. J Invest Dermatol 2004;122:361–368

33. Mauerer A, Roesch A, Hafner C. Identification of new genes associated with melanoma. Experimental Dermatology 2011(20); 502–507.

34. Jaeger J, Koczan D, Thiesen HJ. Gene Expression Signaturesfor Tumor Progression,Tumor Subtype, and TumorThicknessin Laser-Microdissected MelanomaTissues. Clin Cancer Res 2007;13(3)

35. Marioni JC, Mason CE, Mane SM. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 2008; 18:1508-1517

36. The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. Cell 2015; 161, 1681–1696

37. Targeted therapies for cutaneous melanoma. Kee and McArthur. Hematol Oncol Clin North Am. 2014

38. Chapman et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. New England Journal of Medicine 2011

39. Combined BRAF and MEK Inhibition versus BRAF Inhibition alone in Melanoma. Long et al. New England Journal of Medicine. 2014

40. Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. Larkin et al. New England Journal of Medicine. 2015

41. The Cancer Genome Atlas Program: Human Subjects Protection and Data Access Policies. Revision 01-16-14.

42. Love MI, Huber W and Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, 15, pp. 550.

43. Alexa A and Rahnenfuhrer J (2010). *topGO: topGO: Enrichment analysis for Gene Ontology*. R package version 2.20.0.

44. Noetzel E, Rose M, Sevinc E. Intermediate filament dynamics and breast cancer: aberrant promoter methylation of the Synemin gene is associated with early tumor relapse. Oncogene 2010 Aug 26;29(34):4814-25.

45. Krivega M, Geens M, van de Velde H. CAR expression in human embryos and hESC illustrates its role in pluripotency and tight junctions. Reproduction 2014. Nov;148(5):531-44

46. Ghosh S, Kumar A, Chandna S. Connexin-4 3 downregulation in G2/M phase enriched tumour cells causes extensive low-dose hyper-radiosensitivity (HRS) associated with mitochondrial apoptotic events. Cancer lett 2015 Jul 10;363(1):46-59.

47. Hsu LC, *et al*. Sequencing and expression of the human ALDH8 encoding a new member of the aldehyde dehydrogenase family. Gene, 1996 Oct 3

48. Lin YC et al. Tumor suppressor SCUBE2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial-mesenchymal transition. J Cell Science 2014 Jan 1;127(Pt 1):85-100

49. Zhang JS et al. ZNF185, an actin-cytoskeleton-associated growth inhibitory LIM protein in prostate cancer. Oncogene 2007 Jan 4;26(1):111-22

50. Bendas G, Borsig L. Cancer Cell Adhesion and Metastasis: Selectins, Integrins, and the Inhibitory Potential of Heparins. Int J of Cell Biology 2013 doi:10.1155/2012/676731.

51. Friedmann-Morvinski D, Verma IM. Dedifferentiation and reprogramming: origins of cancer stem cells. EMBO Rep. 2014 Mar; 15(3): 244–253.

52. Grichnik J. Melanoma, Nevogenesis, and Stem Cell Biology. J Investigative Dermatology (2008) 128, 2365–2380

53. Wang C et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnology. 2014 Sep;32(9):926-32

54. Shyr D, Li CI. Sample Size Calculation of RNA-sequencing Experiment-A Simulation-Based Approach of TCGA Data. Biometrics and Biostatistics 2014.