Developing Advanced Computing Techniques in Bioinformatics and Biomedical Engineering.

Application in Diagnosis and Assessment of Skin Cancer

PhD Thesis Dissertation

Author:

Juan Manuel Gálvez Gómez (PhD Student)

Directors:

Ignacio Rojas Ruiz (Director)

Francisco Manuel Ortuño Guzmán (Co-Director)

UNIVERSITY OF GRANADA

Department of Computer Architecture and

Computer Technology

PhD Thesis Dissertation:

Developing Advanced Computing Techniques in Bioinformatics and Biomedical Engineering.Application in Diagnosis and Assessment of Skin Cancer

by

Juan Manuel Gálvez Gómez (PhD Student)

Directors

Ignacio Rojas Ruiz (Director)

Francisco Manuel Ortuño Guzmán (Co-Director)

Doctoral Program in Information and Communication Technologies of the University of Granada

Granada, May 24, 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan Manuel Gálvez Gómez
ISBN: 978-84-1306-251-8
URI: http://hdl.handle.net/10481/56469

Agradecimientos

Han sido casi 3 años y medio de lucha, entrega y constancia. Muchas horas de angustia, de desazón cuando cuesta sacar adelante aquello que tienes en marcha. Pero este día llegó. Y llegó, como muchos otros, porque pasaron cientos de experiencias de por medio. Experiencias, ni amargas ni dulces, experiencias. Vivencias que llevaré siempre conmigo, en mi memoria mientras me dure. Porque las viví durante estos 3 años rodeado de personas a las que estimo, y a las que hasta puedo llegar a querer.

Quería en primer lugar manifestar mi más profundo agradecimiento a mis 2 Directores de Tesis. Ignacio, gracias por facilitarme la vida siempre que ha sido posible ante cualquier trámite, por introducirme y guiarme en el mundo de la Bioinformática, por tener la oportunidad de conocer tan cercanamente a personas de tanto prestigio en esta disciplina. Curro, me has ayudado más de lo que nadie podría esperar y te has entregado con mi labor, creyendo en mí en todo momento y mirando por mí más que yo mismo. Sólo tengo gratitud hacia vosotros. Quería añadir aquí, si me lo permiten, a Luis Javier. Porque no has ejercido como Director, si no como algo mucho más: un verdadero compañero y amigo. Quiero que sepas que agradeceré por siempre todos y cada uno de los consejos, iniciativas y propuestas que tuviste para que mi trabajo apuntase a lo más alto y tratar de llegar a ser brillante.

Quiero también dedicarle unas palabras al Departamento de Arquitectura

y Tecnología de los Computadores. Ha sido un honor pertenecer a este Departamento durante esta etapa pre-doctoral. He llegado hasta a disfrutar de la Docencia impartiendo una materia que no se me da especialmente bien, pero en la que me esforcé y traté de dar lo mejor de mí para aprendizaje de los alumnos. Saludos para todos ellos también. No quiero olvidarme de Mari Carmen, José, Bea, Pacos, y personas que trabajan en la gestión y administración del CITIC. Siempre dispuestos a ayudar y a ahorrarme algún quebradero de cabeza. Gracias por siempre.

To the people at Institute of Bioinformatics WWU Muenster. Thank you for welcoming me with such affection and for treating me from day one as one more. Thank you for all the help I received from you. In particular, Eberhard, thank you for your time dedicated to me, for your supervision. Thank you all very much: Wojtek, Vika, Norbert, Reza, Nikhil, Wolfgang, and the rest of your wonderful IoB team. Vielen Dank für alles!

A mis compañeros, de fatigas, pero también de alegrías y buenos momentos. Porque detrás de horas tratando de darle sentido a lo que hacíamos, nos hemos reído. Y he desayunado. E incluso trabajado: Gonzalo, Antares, Fergu, Antonio, Paloma, Juanjo, Luis, Ramón, Iván. Antares... qué te voy a decir que no te haya dicho: gracias a tu amplia sabiduría informática y trabajo, he podido presentar unas gráficas estupendas. ¡Mil gracias compañero! ¿Me dejé a alguien del Despacho? Ya lo sé Dani, imposible no acordarme de tí. Si hemos hablado, discutido y sacado trabajo adelante juntos como campeones. ¡Un placer!

A mis amigos más cercanos, ahora un poco lejanos en distancia, pero nunca en pensamiento y corazón. Siempre deseándoos a todos lo mejor y que nos podamos juntar pronto. Manolo, Miri, Oresti, Claudia, y un largo etcétera. Gracias por vuestro más sincero aliento y ánimos hacia mí, no en esta etapa, si no siempre. ¡Os llevo siempre conmigo amigos míos!

A mi familia. A mi abuelo Francisco, al que siempre recordaré con orgullo y

amor. A mis padres, por darme todo lo que tengo, por hacer de mí un hombre con valores, por darme vuestro amor cada día. ¡Os quiero! Y a mi hermana, la rubia más inteligente y divertida que conozco. Decir suerte es poco para referirme a ella. Todo corazón y hermana de hermano. Aquí no tengo suficiente para esplayarme todo lo que quisiera, pero eres sencillamente maravillosa. Tu hermano, ¡que te quiere y siempre te querrá mucho!

A mi Eva, porque ella sí que es una auténtica heroína. Comprender, aguantar, esperar... casi cada día de mi Doctorado. No hay palabras para expresar tan ampliamente lo que me haces sentir. Por animarme y no dejar que desista. Por simple y llanamente, quererme. Muy feliz de haber pasado esta etapa de mi vida junto a tí y lo que nos queda. Te amo.

A todas aquellas personas que me mostraron su afecto e interés en mí y en el desarrollo de esta tesis, ¡muchas gracias!

Y esto va también dedicado a todas aquellas personas que lucharon y luchan cada día. Deseo que mi trabajo sea útil y se pueda seguir mejorando, acercándonos cada vez más a su solución, hasta que lo venzamos juntos como raza. A los que conocí y quise, cada día dedicado a mi labor ha sido impulsado por vuestra lucha.

Para mis padres, para mi hermana,

para Eva mi amor ...

La ciencia más útil es aquella
cuyo fruto es el más comunicable.

Leonardo Da Vinci.

Contents

Abstract

The sequencing of the Human Genome has opened a new era of opportunities in the field of Bioinformatics. Now more than ever, the biological knowledge about the human being continues to widen thanks to immersion and research dedication in multiple interdisciplinary fields at different scales: Transcriptomics, Genomics, Metabolomics, Proteomics, etc. Both governmental organisations and different international institutions have made strong economic investments in search of providing their research centers and laboratories with the best possible equipment. The explosion of the number of experiments carried that have been carried out in these last 2 decades on the different technologies of sequencing at transcriptomical level (mainly microarrays and RNA-seq) has meant the collection of an enormous amount of information that does not stop growing. Over time, many of these isolated experiments have been shared with the scientific community both publicly and under controlled access. In this sense, the potential of the information stored in such repositories is extremely high and the biological knowledge to be derived may still be an unknown to be revealed. This is due in large part to the fact that the experiments carried out usually have a very reduced number of samples, which implies the extraction of specific conclusions dependent on the characteristics of the cohort of samples analysed. Bringing together all the multilevel biological information on the same disease, one could collect a much broader and more robust set of data from which to extract more significant results at the biological level and widely supported at the statistical level. In

addition to glimpsing general conclusions about the most prominent biomarkers of a disease, there is the possibility of immersing oneself in the search for more specific biomarkers, taking into account clinical data that offer a much closer approach to the patient and to that what is increasingly demanded in healthcare: the pursued dream of personalised medicine. In this sense, advanced strategies for the efficient integration of information and the selection of reliable biomarkers are increasingly valued and necessary in order to advance the understanding, knowledge and treatment of diseases. Although the methodological approaches proposed in this thesis can be extrapolated and applied to any type of disease for which there is a relevant number of samples, the research carried out has focused on improving the diagnosis of skin cancer. This cancerous disease is biologically very heterogeneous and its incidence is increasing worldwide, so there is great alarm and social concern. Since cancer is essentially considered a disease on genetic level, all efforts have been made to extract knowledge from two main sources of information at the transcriptomic and genomic levels: gene expression levels and copy number variations. Besides providing some insights into the most informative biomarkers for knowing skin cancer predisposition, this dissertation opens new opportunities to develop innovative methodological approaches that consider highly heterogeneous data, quantified in multiple omic viewpoints and leading to the establishment of greater awareness and knowledge about the analysed diseases.

Resumen

La secuenciación del Genoma Humano ha abierto una nueva era para una disciplina como la Bioinformática. Ahora más que nunca, el conocimiento biológico sobre el ser humano sigue ensanchándose cada vez más gracias a la inmersión y dedicación investigadora en múltiples campos interdisciplinares a diferentes escalas: Transcriptómica, Genómica, Metabolómica, Proteómica, etc. Tanto organizaciones gubernamentales como diferentes instituciones internacionales han realizado fuertes inversiones económicas en búsqueda de dotar sus centros de investigación y laboratorios con los mejores equipamientos. La explosión del número de experimentos realizados sobre las diferentes tecnologías de secuenciación a nivel transcriptómico que se han ido sucediendo en estas 2 últimas décadas (principalmente microarrays y RNA-seq) ha supuesto la recolección de una cantidad ingente de información que no para de crecer. Poco a poco, muchos de estos experimentos aislados han sido compartidos con la comunidad científica tanto de forma pública como bajo acceso controlado. En este sentido, el potencial de la información alojada en dichos repositorios es extremadamente alto y el conocimiento biológico a derivar puede seguir siendo aún una incógnita. Esto es debido en gran parte a que los experimentos llevados a cabo usualmente cuentan con un número muy reducido de muestras, lo que supone la extracción de conclusiones muy concretas y adaptadas al cohorte de muestras analizado. Aunando toda la información biológica posible sobre una misma enfermedad, se podría extraer un conjunto de datos mucho más amplio

y robusto del que poder extraer resultados más significativos a nivel biológico y ampliamente sustentados a nivel estadístico. Además de llegar a obtener conclusiones más generales sobre aquellos biomarcadores más prominentes a informar mejor sobre el padecimiento de una enfermedad, es posible sumergirse en la búsqueda de biomarcadores más específicos teniendo en cuenta datos clínicos que ofrezcan un acercamiento mucho mayor al paciente y a aquello que se reclama cada vez más en la atención sanitaria: la medicina personalizada. En este sentido, estrategias avanzadas para la integración eficiente de la información y la selección de biomarcadores fiables son cada vez más valoradas y necesarias en pos de avanzar en el entendimiento, conocimiento y tratamiento de las enfermedades. Aunque las aproximaciones metodológicas planteadas en esta tesis pueden ser extrapoladas y aplicadas a cualquier tipo de enfermedad sobre la que exista un número relevante de muestras, la investigación realizada se ha centrado en la mejora del diagnóstico del cáncer de piel. Se trata de una enfermedad cancerosa altamente heterogénea y cuya incidencia es cada vez mayor a nivel mundial, por lo que existe una gran alarma y preocupación social. Dado que el cáncer se considera fundamentalmente una enfermedad de los genes, todos los esfuerzos han sido destinados a extraer conocimiento desde principalmente 2 fuentes de información a nivel transcriptómico y genómico: la expresión de gen y el número de copias de gen. Finalmente, además de proporcionar ciertas averiguaciones sobre aquellos biomarcadores más informativos para conocer la predisposición a padecer cáncer de piel, esta tesis abre nuevas oportunidades para desarrollar innovadoras aproximaciones metodológicas que consideren datos altamente heterogéneos, cuantificados en múltiples puntos de vista ómicos y liderando al establecimiento de una conciencia y un conocimiento mayor sobre las enfermedades analizadas.

1. Introduction

1.1. Thesis Goal

The main goal of this thesis aims to contribute in the development of advanced computing techniques for the processing of biological data in the field of Bioinformatics and Biomedical Engineering. In particular, this thesis focuses on the application of efficient strategies by means of the integration of heterogeneous information sources for the determination of reliable biomarkers which help in improving the diagnosis of complex and cancerous diseases. All different methodological approaches proposed in this dissertation have been designed and assessed on a specific cancerous disease: the skin cancer.

1.2. Problem & Motivation

The decryption of the Human Genome has not only marked a before and after in the biological analysis of the human being. Its achievement has meant the opening to an endless number of human biological studies, offering the possibility of collecting and analysing a myriad of patient samples presenting evident signs of suffering from certain pathologies. Different laboratories and research groups around the world have made significant efforts to elucidate those more prominent biomarkers to better inform about the disease or predisposition to suffer from certain diseases. However, these initiatives are strongly limited to the availability

of small repertoires of samples that make it difficult to generalise and state their biological implication as well as their significance and statistical robustness to interpret those obtained results. Over the years, both publicly and protectively accessible databases have appeared, sharing and making those data available to the scientific community for experimental analysis. Thus, the main challenge now lies in obtaining the most reliable, robust and meaningful results possible based on the consideration of multiple samples from different experiments achieved for the same disease. As an immediate consequence, multiple bioinformatics issues appear to be treated and considered in order to effectively analyse an integrated dataset made up of all those isolated experiments. In order to show the potential of the different strategies of information integration and biomarker selection proposed in this thesis, skin cancer was chosen as a study case. And, because the cancer is considered a genetic disease, the experimental and research efforts carried out throughout this thesis focused on using information directly related to this fact: gene expression levels and gene copy number variations. As it will be seen below, skin cancer is a highly heterogeneous and complex cancerous disease. In this sense, with the purpose of facing this problem, the improvement of the reliability of the diagnostic process is postulated as fundamental, being able to be supported by automated computational tools that complement the subjective judgment of medical experts with an objective mathematical point of view. This experimental part has been extensively assessed by applying multiple machine learning techniques.

1.3. Thesis Aims and Contributions

In this section, the main objectives and goals together with the most remarked contributions of this thesis are highlighted.

1.3.1. Thesis Aim and Objectives

Although there are more and more studies that show promising results pointing to specific biomarkers better informing about multiple complex and cancerous diseases, it is true that there is a global concern about both their reliability at the level of generalisation and global occurrence and their particular and differential appearance in different cohorts of patients.

The main motivation behind this dissertation is to make available a range of advanced computing strategies to the scientific community for integrating heterogeneous biological data. This fact is thought to facilitate the discovery and emergence of more representative and discerning biomarkers of the analysed disease. By having paid all the attention and focus on the analysis of skin cancer, the main objectives have been specialised in order to improve the detection and diagnosis of this cancerous disease.

Some of the issues addressed in this thesis take into account the following:

- Is it possible to obtain a widespread diagnosis of skin cancer on the basis of the biological data available and accessible to the scientific community? And if so, is there a repertoire of biomarker candidates reliable and contrasted enough to describe and help diagnose this heterogeneous disease?

- What technological, scientific and design considerations are essential and required to establish a stable pipeline that allows such biomarkers to emerge? And in that case, what are the real possibilities of effectively integrating biological data from such diverse sources? What feature extraction methods can provide the most informative and discerning biomarkers that facilitate the classification process?

- In this last point, how is it possible to efficiently train a classifier where the generalisation model has to take into consideration multiple pathological states of skin cancer?

The research work carried out during the development of this thesis confronts these questions broadly in search of being resolved by presenting advanced computational strategies that bring light to these uncertainties raised. The proposed strategies address several essential and standardised parts of automated data processing such as preprocessing, dimensionality reduction, feature selection and/or classification.

1.3.2. Contributions of the Thesis

The main contribution of the thesis is the development of efficient strategies for the integration of heterogeneous data and the selection of reliable biomarkers that facilitate the research work of the scientific community when dealing with large studies. It should be noted that the methodological approaches proposed here can be applied to a wide range of diseases by using machine learning techniques. Extensively, under the assessment of a specific study case such as skin cancer, some insights are provided about biomarkers that emerged from our analysis taking into consideration gene expression and copy number variation. All these points will be discussed in more detail in the related chapters.

1.3.2.1. Contributions to Skin Cancer Diagnosis

1. Proposing a panel of informative biomarkers on the main skin cancer pathologies at epidermal level (**Chapter 3**).

2. Proposing a panel of informative biomarkers that additionally discern some skin diseases considered pre-cancerous of skin cancer (**Chapter 4**).

3. Proposing a panel of highly informative biomarkers that may be responsible for the progression of cutaneous melanoma (**Chapter 5**).

1.3.2.2. More General Contributions

1. Proposing different effective integration pipelines dealing with heterogeneous information sources based exclusively on gene expression from multiple microarray platforms (**Chapter 3**) or considering both microarrays and RNA-seq platforms (**Chapter 4**).

2. Proposing a methodological approach for the efficient selection of informative biomarkers for multiple pathological states of skin cancer (**Chapter 4**).

3. Proposing an integration pipeline where heterogeneous information is quantified in gene expression and copy number variation (**Chapter 5**).

1.4. Outline

In this **Chapter 1**, the main objectives and motivations revolving around the achievement of this thesis have been presented, also motivating the accomplishment of various milestones translated into various contributions at the methodological level and on findings about the diagnosis of skin cancer.

In **Chapter 2**, a brief exposition on some fundamental concepts for the efficient treatment of the biological information used in this research is carried out. In addition, information is included on the repositories of inspected data, the types of biological data used and the pathological states of skin considered for the development of the different studies approached during this thesis.

In **Chapter 3**, a study on the integration of different microarray platforms is presented. In this first work, the informative potential stored in the microarrays for the extraction of knowledge about the analysed disease is highlighted and an innovative biomarker selection strategy is established, considering genes robust to the influence of diverse and potential present batch effects. Particularised to

the problem of skin cancer analysed in this dissertation, the 17 most informative biomarkers are indicated to discern up to 7 pathological states of skin considering 5 tumor states (basal cell carcinoma, squamous cell carcinoma, Merkel cell carcinoma, cutaneous melanoma in primary state and cutaneous melanoma in metastatic state).

In **Chapter 4**, an extension of the work carried out and presented in chapter 3 is presented, which is reflected in several points: the selection of data from additional repositories, the additional consideration of RNA-seq data to co-integrate with microarray data, the introduction of skin diseases considered precancerous of the skin (psoriasis and actinic keratosis) and the implementation of an algorithm of selection of informative biomarkers to discern different pathological states simultaneously.

In **Chapter 5**, a new methodological extension of the previous works is made. In the first place, samples of healthy states (normal skin and moles) and melanomas (primary and metastatic) already used previously are considered. Next, the integrated dataset is reinforced with a cohort of 73 patients suffering from cutaneous melanoma, for which quantified biological data are available at the transcriptomic level (gene expression from RNA-seq data) and genomic level (copy number variation from whole exome data). The methodological approach presented allows biomarkers to be selected based on the informative correlation between the level of expression and copy number alterations of these genes.

In **Chapter 6**, based on the findings revealed, some conclusions and a series of suggestions are indicated to be carried out in the future.

1.5. Summary

This chapter was intended for providing an introduction of the main aims and motivations behind this thesis, thus remarking an overview of the research

contributions as well as a brief exposition of the content included in the next chapters.

2. Methological Review and Fundamentals

2.1. Designing Customised Bioinformatics Pipelines

In the pursuit of extracting underlying knowledge from biological data, it is absolutely essential to become aware that a series of bioinformatics tasks are necessary for their correct determination. Traditionally, bioinformatics pipelines have been guided to simply understand which patterns or biomarkers best define or distinguish the object of analysis (human, mouse, plant, etc.), considering 3 standardised steps: preprocessing, experimental analysis and results. However, when it is intended to bring together as much information as possible to be integrated and extract additional knowledge such as the most informative biomarkers for diagnosis, these steps have to be extended on both sides of the simplest classical pipeline. A standard pipeline to meet these requirements involves the following steps: identification and acquisition, pre-processing, post-processing, feature selection and classification. Each of these pipeline phases can be custom-designed according to the scope and purpose of the study approached by considering different strategies and tools for implementation. For this reason, the studies presented in **Chapters 3, 4 and 5** require different considerations for their implementation. Although some key concepts will be used later, here they are briefly introduced and specific information for the

accomplishment of the studies is included.

2.1.1. Identification & Acquisition

This first step requires an exhaustive search for those data that are intended to be analysed. This implies taking into account of which biological nature is (sample type) and how this is acquired (sequencing technology), where these data are accessible (webdata repositories) and how much can be acquired from each type of samples that are stored there (skin pathological states).

2.1.1.1. Sample Type

In general, and knowing that there may be other forms of biological origin, the choice of the type of sample for the experimental analyses was one of the most important decisions: tissue or cell line. After inspecting the current possibilities, it was decided to analyse those samples that are extracted from tissue specimen, following the rules established by ICD-10 [1]. Specifically, this assumes that those samples are usually extracted by means of punch biopsies or as slice sections. Special care was taken in order to not select samples on which drugs were applied or viruses were evaluated. Similarly, those skin tissues corresponding to trunk, upper limb (including shoulder) and lower limb (including hip) were selected.

2.1.1.2. Sequencing Technology

As previously introduced, this thesis has focused on the analysis of gene expression and alterations in gene copy numbers for the presentation of the results. In this sense, and based on the platforms of available data for the analysis of this type of biological variables, we considered the 2 technologies with the greatest number of experiments and co-existing at present: hybridisation-based microarrays (those experiments based on array) and high-throughput sequencing (those experiments based mainly on Next-Generation Sequencing (NGS) such as

RiboNucleic Acid sequencing (RNA-seq) and Whole Exome Sequencing (WXS)). For the experimental analysis of the studies carried out for the presentation of this dissertation, gene expression values and Copy Number Variations (CNVs) were specifically considered and used. The different sequential processes for obtaining gene expression (from microarrays and RNA-seq) and somatic copy number variation (from WXS) are briefly introduced here:

- **Microarrays:** The main foundation of this technology is based on DNA hybridisation process, taking into account the 4 different nucleotide types: Adenine (A) binding to Thymine (T) and Cytosine binding to Guanine (G). In order to create the microarray for further analysis, different sequential events have to take place. First of all, the oligonucleotides probes are adhered to the array surface. Following, each patient sample is subjected to fluorescent lighting and added to the array. As a consequence, that non-hybridised material to each probe is removed. Later, the hybridised material is subjected to a laser whose reflected light is detected by a scanner. At this point, the surface of microarray can be scanned in order to obtain a microarray image. The proportion of hybridised sample can be quantified by means of a process analysis of this image and the results are stored in a CEL file. This raw files contain values quantified in gene expression values. The main manufacturers of this technology are Affymetrix [2] and Illumina [3].

- **RNA-seq:** Based on the use of NGS, this technology reveals the presence and quantity of RNA in a biological sample at a specific temporal moment. In this sense, the use of this current alternative is thought to monitorise the continuous changes within the cellular transcriptome at gene expression level, among others. The RNA sequencing process mainly considers three general steps for the obtaining of reads: RNA isolation,

RNA selection/depletion and cDNA synthesis. Following, based on guided genome, these raw sequence reads can be aligned by means of a reference genome. At this point, RNA-seq read counts can be obtained by counting the number of reads mapping to each locus in the transcriptome assembly step [4]. Finally, correspondence and conversion to gene expression values is achieved under application of conditional quantile normalisation [5, 6]. Nowadays, Illumina [3] has already monopolised the RNA-seq market.

- **WXS:** This technology is essentially a genomic technique for sequencing the exome, which implies the analysis of the protein-coding region of genes in a genome. These regions are known as exons and only constitute the 1% of the human genome. The identification of genetic variants altering protein sequences can be achieved by means of its analysis. This is the main reason to consider this technology for determining somatic CNVs and will be briefly justified in Section 2.1.2.2.

2.1.1.3. Webdata Repositories

Up to 3 web repositories were consulted for the collection of biological samples:

- **National Center for Biotechnology Information - Gene Expression Omnibus (NCBI GEO):** This international public repository archives and freely distributes microarray, next-generation sequencing (NGS) and other forms of high-throughput functional genomic data. The resource supports archiving of raw data, processed data and metadata which are indexed, cross-linked and searchable [7].

- **ArrayExpress (AE):** One of the major international public repositories of functional genomics data which includes biological data generated by sequencing or array-based technologies. This repository maintains

an exchange agreement with NCBI GEO in order to import directly experiments for both technologies [8].

- **National Cancer Institute - Genomic Data Commons (NCI GDC):** Conceived as an information system, this repository contains multiple biological raw data as well as harmonised data by means of standardised pipelines. Besides storing diagnostic, histologic and clinical outcome of patient samples, this database offers patient cohorts quantified in multiple omics point of view such as transcriptomic or genomic [9].

2.1.1.4. Skin Pathological States

After a long research work at biological level, 10 pathological skin states were finally considered for experimental analyses. They can be taxonomically classified within 4 groups: healthy states (with regard to healthy normal skin and nevus/moles), Non-Melanoma Skin Cancer (NMSC) (with regard to biological alterations in keratinocytes), Melanoma Skin Cancer (MSC) (with regard to biological alterations in melanocytes) and precancerous states (concerning skin diseases with a possible predisposition to tumorally evolve). It should be noted that only those samples showing the lesion at epidermal level were taken into account. This decision led to discard those referring to lymph nodes or metastases in other parts of the body other than the metastasis itself cutaneous. The pathological states are the following:

- **Normal Skin (NSK):** Taken as a reference sample to observe alterations, this healthy state can be collected either from patients with no apparent signs of suffering from any skin disease or from patients with skin lesions but extracting from an area without skin lesion.

- **Nevus (NEV):** Also considered a priori as a healthy state, NEV are considered as a birthmark or a mole on the skin, especially a birthmark

in the form of a raised red patch. It is fundamental to take into account this state, since it is proven that many of them can tumorally degenerate into melanoma. Their morphological and histological aspects become very similar and they can represent a high risk of melanoma [10].

- **Basal Cell Carcinoma (BCC):** Included among NMSC, this skin carcinoma is considered to be the most common pathology of skin cancer [11]. Although BCC is the one with the least risk of spreading and becoming deadly [12], it can be disfiguring if it is not treated promptly. It is a highly complex cancerous manifestation histopathologically, which historically has an incidence of >80% among NMSC [13].

- **Squamous Cell Carcinoma (SCC):** This is the second most common NMSC [14], although its incidence is dramatically increasing even with respect to BCC [15].

- **Primary Merkel Cell Carcinoma (PMCC):** As a global manifestation, it is a highly aggressive and rare cancer with neuroendocrine characteristics [16]. It would be the third NMSC in order of incidence after BCC y SCC. Merkel Cell Carcinoma (MCC) development is linked to exposure to UV radiation as with other skin cancers, and PMCC lesions typically appear on sun-exposed skin [17].

- **Metastatic Merkel Cell Carcinoma (MMCC):** When MCC gets to metastasise, it is really out of precise medical control and it becomes very complicated to be able to treat it. It does not even seem clear that chemotherapy can always be effective. Its diagnosis and treatment still require a broader consensus to establish clearer and more precise guidelines [18].

- **Primary Melanoma (PRIMEL):** Cutaneous melanoma is undoubtedly

the most deadly manifestation of skin cancer, included in MSC. However, detected at an early stage and thanks to appropriate biopsy methods, it can be easily treated and reduce the risk of death [19]. Primary melanoma is usually diagnosed following the ABCDE signs [20]: asimmetry, border, color, diamater and evolving.

- **Metastatic Melanoma (METMEL):** Despite great advances in treatment, the long-term prognosis of this MSC disease in advance state remains poor [21]. Probably, its highly mutable and heterogeneous character precludes setting standards for generalised treatment.

- **Psoriasis (PS):** For those patients suffering from this immune skin disease, the risk of developing some form of cancer (lung, gastrointestinal tract, urinary tract, etc.) have already been alerted [22]. However, in recent years, there have also been indications that there is a risk of deriving in some tumor manifestation of skin [23].

- **Actinic Keratosis (AK):** Considered to be a cancerous precursor to SCC [24], its early diagnosis and treatment could prevent the development of more dangerous skin cancer later on.

2.1.2. Pre-processing

After having identified and acquired all relevant and existing information, it is crucial to properly process it so that the subsequent downstream analysis is as reliable as possible on the basis of the underlying biological knowledge. Different tasks are usually carried out at this stage:

2.1.2.1. Key Considerations for Processing Microarray and RNA-seq Platfoms

A. Data Adequacy All datasets must be properly preprocessed under the application of different procedures, regardless of the technological platform and the sequencing technology used. In this sense, each dataset must be processed individually in order to convert the raw data into expression values. Traditionally, the processing of microarrays has involved the use of Robust Multi-array Average (RMA) algorithm [25]. This algorithm performs background correction, normalisation and summarisation in a modular way. Regarding the RNA-seq data, this part is mainly covered by the consideration of conditional quantile normalisation to correct GC-content [5, 26].

B. Heterogeneous Sources Integration When all datasets have been individually adequated, checking the depth of scale for each dataset has to be performed in order to homogenise the analysis of all samples considered and thinking that they will be subsequently integrated. This supposes to apply logarithmic transformation on those series that were not previously pre-processed as well as homogenisation of the bit depth (more widely known as *dynamic range*) to equalise the expression ranges for all datasets. Following, it is completely necessary to check the gene annotation of each series and map to common gene symbols for all datasets. For this purpose, the use of standardised gene symbols is highly recommended for gene annotation: HUman Genome Organization (HUGO) [27], Entrez [28], Ensembl [29], etc. For example, HUGO is the official gene symbol approved by the HUGO Gene Nomenclature Committee (HGNC) [30]. This committee approves those symbols and sets the standards in accordance with the guidelines for Human Gene Nomenclature (HGN). In this sense, the HGNC approves a unique and meaningful name for every known human gene, based on a query of experts. Extensively, HGNC is responsible for approving unique symbols and names for human loci, including protein coding genes, RNA genes and pseudogenes. This fact also helps in avoiding ambiguous

scientific communication. Additionally, the use of standardised gene symbols makes possible to know truthfully how many genes are common throughout all the datasets considered and avoids integration errors when applying some tool for this purpose.

C. Batch Effects Removal After having jointly corrected all the considered datasets, a joint evaluation of possible deviations between different datasets is essential. Obtaining each dataset involves at least one different experiment that can insert variations in gene expression due to biological, technical and even atmospheric agents [31]. This concept is widely known as batch effects. For the experimental analysis of each study, several batch effects correction algorithms have been considered. They are briefly introduced here:

- **Quantile Discretization (QD):** Although the main purpose of this method consisted in data normalisating at the probe level [32], batch effect removal has been also benefited from being taken into consideration [25, 33]. Conceived as a discrete method, this algorithm assigns the same value to all genes that fall into the same bin across all studies. These bin values can be continuous or cardinal numbers obtained by calculating mean gene expression values.

- **Mean Rank Scores (MRS):** This method considers one batch reference and all genes are ranked based on their median expression. Following, all genes contained in each sample in the non-reference batch are also ranked and their value replaced by the corresponding ranked median from the reference [33].

- **Gene Quantile (GQ):** This discrete method applies a quantile normalisation for each individual gene among different considered datasets. More specifically, it is considered as an extension of MRS that enforces an

extra transformation of gene expression values such that the median values for each gene are equal in all batches (in our case, different datasets) [33].

- **Empirical Bayes (EB):** Widely-known as *ComBat* and based on empirical bayesian techniques, this method assumes that the expression of each gene within a performed analysis is directly affected by a known designed batch factor [33, 34]. In order to remove the influence of these batch effects, estimations considering additive and multiplicative factors are calculated for each gene in each batch. These estimations are usually well-known and simple parameters such as mean and variance. Those parameters are calculated by means of gathering information from multiple genes with similar expression characteristics in each considered batch or analysed dataset. One of the major strengths of this approach focuses on avoiding over-correcting which is critical for use with small batches.

- **Normal Discretization (NORDI):** Based on discrete normalisation as well, this method fits a normal distribution to each expression profile and following detected outlier genes are removed. To achieve it, genes are categorised and separated in three different groups: under-expressed, over-expressed and unexpressed. Finally, each gene expression value is replaced by -1, +1 or 0 in accordance with the previously assigned group. This method was thought to reinforce the extraction of relevant association rules [35].

- **Mean Centering (MC):** Considering the occurrence of systematic multiplicative biases within batches, this simple method transforms the data by subtracting the mean of each gene over all samples (per batch) from its observed expression value, such that the mean for each gene becomes sero [36].

In general terms, an attempt has been made to apply batch effect correction

algorithms that preserve the biological information contained in the datasets. However, dealing with batch effects removal is becoming currently becoming challenging because there is no absolute certainty about removing it even after applying specific correction algorithms [37].

D. Normalisation Although the deletion of batch effects becomes successful, the expression levels between samples do not remain completely homogeneous. In order to avoid possible subsequent errors in the classification phase (see Section 2.1.5), a homogeneous range is established by means of an inter-array normalisation [38]. More consistency is achieved among all samples put together, forcing an identical empirical distribution on each of them based on quantile normalisation. Finally, a data matrix is available in gene expression where the rows correspond to the patient samples and the columns correspond to the genes, or vice versa. In this sense, the broad set of potential biomarkers candidates to discern between different pathologies is ready to be analysed.

2.1.2.2. Key Considerations for Detecting Copy Number Variation

When considering the analysis of genomic data, adequate chromosomal segmentation of the genome is crucial in order to efficiently detect the length and position of copy number variations. Extensively, for effective partitioning, the availability of as many control samples as possible will allow a reliable comparison to tumor samples, helping in emerging copy number variations between them. After inspecting the different possibilites for this purpose, cn.MOPS [39] was selected against other alternative methods: MOFDOC [40], EWT [41], JointSLM [42], CNV-Seq [43] and FREEC [44]. In addition to significantly improving performance with respect to its predecessor alternatives, this tool enables the application of various configurations for the determination of CNVs. Among them, there is opportunity to determine somatic CNVs by subtracting the part

of germline CNVs. This can be achieved thanks to the availability of control and tumor samples from the same cohort of patients. Technically, while tumor sample will contain together somatic and germline CNVs part, control sample will contain only germline CNVs part.

2.1.3. Post-processing

Various tasks can be included here when assessing the validity of biological information that is available after being pre-processed.

2.1.3.1. Dimensionality Reduction

The integration of multiple data sources can lead to a severe problem of computational analysis. When the size of the integrated dataset grows significantly, gathering not only thousands of genes but also hundreds or thousands of samples, finding those most informative biomarkers can be computationally expensive. When analysing gene expression data, a first step is to significantly reduce the search space by determining a set of Differentially Expressed Genes (DEGs). This implies that only those genes that are more informative between each pair of analysed pathologies could be selected. To achieve this, it is necessary to restrict the selection of biomarkers based on some statistical parameters. *limma* package has been postulated as a powerful tool to extract this type of information from both microarrays and RNA-seq [38] by applying several statistical restrictions. Among them, Log2-Fold-Change (LFC) and P-Value (PV) can be highlighted by their widespread use. On the one hand, LFC requires a minimum absolute threshold of gene expression level change between each pair of pathological states. On the other hand, PV establishes a cutoff value for adjuste p-values, only allowing those genes with lower p-values to be considered.

2.1.3.2. Statistical Assessment Application

The consideration of statistical values to delimit the set of biomarkers candidates is commonly applied in this type of research studies. Traditionally, it is possible to impose a selection of genes based on the statistical significance of parameters such as p-value or logarithmic fold change. On the other hand, the influence of various factors on those biomarkers candidates must be thoroughly evaluated by applying some statistical test. This will determine whether the variations in gene expression between the different considered pathologies are due to the biological nature itself or may be due to some of the analysed factors. Any complementary information to the selected biological samples may be useful at this point (clinical or other data). ANalysis Of VAriance (ANOVA) statistical test [45] was used for this purpose and extended documentation was consulted in order to perform the statistical analyses [45–48]. This well-standardised and widely-used test is useful for comparing more than two factor means for statistical significance. Additionally, the consideration of correlation tests can help to corroborate the informative correlation between evaluated characteristics. In this thesis, this is applied to see the informational correlation between gene expression and the number of gene copies: Kendall [49], Pearson [50] and Spearman [51]. These non-parametric statistical tests allow to validate the significance of the method without the need to check the normality of the analysed distribution. Finally, other statistical parameters were used to make comparisons of multiple corrections based on confidence intervals: Fisher [52], Bonferroni [53] and Benjamini-Hochberg [54].

2.1.3.3. Functional Enrichment Analysis

In many cases, obtaining DEGs is not enough to know the relevance of the outstanding biomarkers discerning between pathological states. In this sense, it is usually necessary to take a further step forward by seeking to prove the

associations of gene sets with disease phenotypes. For this purpose, there are methods that also use statistical approximations in order to identify those significantly enriched or depleted groups of biomarkers. From the retrieved result, there is an opportunity to better understand the underlying biological processes thanks to the determination of the functional profile of that gene set. Taking into account this type of approaches has allowed associating different gene set, grouped together by their involvement in the same biological pathway, or by proximal location on a chromosome. In this dissertation, both database-specific and programmatic queries under the use of several tools have been carried out. On the one hand, DAVID Bioinformatics Database offers a wide range of results associated with functional properties of the submitted genes [55]. Among them, Gene Ontology (GO) terms divided into Biological Processes (BP), Cellular Components (CC) and Molecular Functions (MF) can be easily retrieved [56]. Reactome [57] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [58] web browsers were also inspected in search of checking involved pathways for gene sets.

2.1.4. Feature Selection

Although the reduction of dimensionality allows to dramatically decrease the set of biomarkers to be evaluated, usually therapeutic diagnoses require only some target genes. With personalised and patient-oriented approaches in mind, it is difficult to determine very specific biomarkers. However, with respect to generalised studies, it may be interesting to determine what diagnostic potency is possible to achieve with small sets of genes. In order to do this, it is necessary to apply some informative ordering criterion that evaluates which genes are more informative than others. For example, there are feature selection algorithms based on mutual information which aim to find the largest dependency between a subset of features and the output variable. In this sense, this dissertation took into

account the Minimum Redundancy Maximum Relevance (mRMR) algorithm [59]. The basis criterion consists in considering mutual information among variables (in our case, genes) in search of assesing variables relevance. In this sense, the algorithm will rank in first position that gene containing the maximum relevance information, followed by those genes providing minimum redundant information. Also, well-known correlation tests have been applied in order to establish another way to order them based on informative correlation, mentioned above: Kendall [49], Pearson [50] and Spearman [51].

2.1.5. Classification

This last step helps in assessing the informative power of the those selected genes to provide an intelligent diagnosis of a new unseen sample. In this thesis, well-known state-of-the-art Machine Learning (ML) techniques have been trained and tested for this purpose:

- **Support Vector Machine (SVM):** These classification models are discriminative classifiers formally defined by a separating hyperplane. This implies that the algorithm outputs an optimal hyperplane that maximises the distance between different classes (in our case, different pathological states). Thanks to this, new unseen samples will be assigned to categories, even when overlapping data is happening. These models have the capability to define a higher dimensional space from a reduced space by means of kernel functions. Extensively, fault tolerance is also managed by this algorithm by controlling γ hyperparameter. This fact improves the generalisation capability of the model [60].

- **K-Nearest Neighbors (KNN):** This type of instance-based learning model assigns a category to a new unseen sample under majority voting decision. This consensus is achieved by inducing the predominant class

among the k nearest neighbors. This technique provides an outstanding performance, even being one of the simplest machine learning techniques [61].

- **Naive Bayes (NB):** This classifier is based on a conditional probability model which applies Bayes theorema with strong (naive) independence assumptions between the features. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers [62].

- **Tree Bagging (TB):** This algorithm is considered as a special case of the model averaging approach. In other words, a very simple and powerful ensemble method. The basis for operation lies in combining the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Traditionally, this algorithm is implemented by means of multiple decision trees [63].

- **Ensemble Learning (ENS):** This approach is very similar to TB, but it is implemented by means of multiple learning algorithms (for example, those classification models previously commented and put all together). This type of implementation is thought to obtain better predictive performance than could be obtained from any of the constituent machine learning algorithms alone. Additionally, each of the considered classifiers can be weighted, allowing to tune the influence of each one of them in the performance [64].

In order to assess the classifier performance, several cross-validation tecniques can be taken into account: Leave-One-Out Cross-Validation (LOO-CV) [65] and K-Fold Cross-Validation (KFOLD-CV) [66]. These techniques are applied over the training dataset to obtain the optimal hyperparameters for the previous methodologies: σ (kernel width) and γ for SVM, and k for KNN. Finally, different metrics for recognition assessment are usually considered, mainly highlighting

Accuracy (ACC) and Overall F1-score (OF1) among others. Specifically, ACC may suffer from some limitations under presence of data imbalancement [67]. In this case, OF1 is the most recommendable metric by tackling better that issue.

2.2. Summary

This chapter presented the main key concepts revolve around the development of this thesis. The content is intended to provide a useful guide for understanding the following 3 chapters focused on presenting the research studies conducted.

3. Offering New Opportunities to Microarrays

3.1. Introduction

This chapter addresses the possibility of offering a comprehensive skin cancer diagnosis based exclusively on the integration of multiple microarray platforms. For this purpose, a novel methodological approach is proposed involving the integration of several heterogeneous skin cancer datasets, and a later multiclass classifier design. This approach is thus a way to provide the clinicians with an intelligent diagnosis support tool based on the use of a robust set of selected biomarkers, which simultaneously distinguishes among different cancer-related skin states. The study has made use of the following resources:

- **Number of datasets:** 24
- **Number of samples:** 678
- **Skin Pathological States:** 7 (NSK, NEV, BCC, SCC, MCC, PRIMEL and METMEL)
- **Web Data Repositories:** NCBI GEO
- **Sequencing Technology:** Microarrays

For more general information on the different resources consulted and used in this dissertation, see Section 2.1.1. The content presented here is largely based on the journal publication entitled *"Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series"* (PloS one, Volume 13, Number 5, `https://doi.org/10.1371/journal.pone.0196836`

3.2. Background

The analysis of microarray data has become a common practice of research groups for the determination of biomarkers of interest for many years. The ability to simultaneously measure the expression levels of thousands of genes has not gone unnoticed by researchers. This fact has promoted its widespread use for determining biomarkers discerning among very specific pathological states (usually, control versus tumor). As a consequence of the above, multiple isolated experiments have been performed and, little by little, collected by means of different webdata repositories. However, despite successful results that have helped to develop diverse target therapies for the treatment of specific diseases, the definitive cure for extremely worrying diseases such as cancer has not been elucidated. In fact, the trend of new cases is gradually increasing for cancerous pathologies such as skin cancer. Almost two decades ago, this cancerous disease was predicted to account for more than a third of all cancers [68], and that prediction is already a crude reality. Although the occurrence of skin cancer is becoming alarming, the registration standards of NMSC are incomprehensibly precarious almost worldwide [69]. This is largely due to an insufficient data collection in cancer registries on BCC cases which prevents its actual incidence from being known [70]. Even so, skin cancer is considered the major public health problem in Australia [71, 72] and the most commonly diagnosed cancer in United States [70, 73]. Therefore, several consciousnees campaigns and

programs have been promoted in relation to the prevention of skin cancer in both countries [70, 74]. With respect to its incidence in Europe, NMSC has been categorised as one of the most worrying malignancies in Germany [75] as well as a systematic review reflected the worrying current situation in Spain [76]. Focusing on examining what research efforts have been perfomed for revealing clues about how to treat this cancerous disease, a wide range of research studies have been performed by using microarray technology [77–80]. Additionally, in search of broadening the knowledge about this disease, a wide range of ML and computer science approaches have been also proposed: neural networks [81], image preprocessing and classification [82–84], prediction models [85, 86], pattern recognition [87], optical techniques [88, 89], etc.

3.3. Motivation

Despite great efforts have been done for bringing light to effective diagnosis of multiple diseases, there are suspicions that a much more widespread and robust studies could be carried out. In the light of what has been investigated so far, there is a high probability that most of the research studies developed applying microarray technology to the characterisation of different pathological states of any disease may fail in reaching statistically significant results. This is largely due to the small repertoire of analysed samples, and to the limitation in the number of states or pathologies usually addressed. Focusing again on the diagnosis of skin cancer, this seems to be the general trend of previous microarray studies such as those presented in Section 3.2 where the researches are often conducted on a limited sample set. This fact leads to obtain different DEGs sets by using traditionally binary classifiers for each isolated experiment. As a solution to these limitations, collecting different datasets including skin cancer samples of diverse pathological states from various experiments, may considerably

increase the robustness of the study and help in identifying biomarkers for the differentiation of a wider range of pathological states. This initiative would entail the challenge of analysing a multiclass scenario. Although multiclass classification has been approached for a wide range of cancerous diseases in several previous works (breast [90], colorectal [91], ovarian [92], prostate [93], etc.), the truth is that the consideration of this approach on skin cancer analysis remains practically unprecedented. Merely, hierarchical clustering has been used in order to compare gene expression signatures from different skin pathological states [94]. Also, a number of skin cancer studies have used this strategy from the analysis of histopathological [95, 96] or dermoscopic images [97–100]. Therefore, it is sighted that an excellent opportunity presents itself for performing a comprehensive analysis of the gene expression, eventually becoming able to extract revealing genes which could be responsible for a number of manifestations of this disease of the genes [101]. However, the joint consideration of cancer datasets with different technical characteristics usually involves dealing with the removal of batch effects. The influence of potential deviations on the gene expression quantification is wrongly and usually disregarded, so it should be always taken into account for an effective integration [37]. Extensively, although the imposition of RNA-seq is a matter of time with regard to gene expression analysis, microarrays still have many factors in their favor. Above all, microarrays have been used so far, and are still in use, because they are cheaper. Additionally, the existing availability of a vast amount of gene expression microarray datasets encourages to take them into consideration and should still be properly exploited. With all these premises, a multi-platform combination of microarray datasets from Affymetrix and Illumina manufacturers [2, 3] was carried out. This integration is expected to strengthen the statistical robustness of the study as well as the finding of highly-reliable skin cancer biomarkers. For this end, ML techniques efficiently help to select those genes with the highest informative power for the diagnosis. Under this general

idea and based on the use of highly-discriminant DEGs, any new patient skin sample could be assessed and correctly classified by distinguishing among several skin pathological states in a single analysis [102]. Since the cancer prognosis is much more encouraging when a patient diagnosis is available at an early stage, clinicians can take advantage of relying the final diagnosis on its assessment [103]. Consequently, at the dawn of the personalised medicine, predisposition to certain skin cancer manifestations could be properly detected [104], and unnecessary medical treatments such as radiation therapies, excision surgeries or medications supply could be prevented [105].

3.4. Methodologies and Experiments

3.4.1. Samples

All analysed RNA samples were obtained from NCBI GEO web platform [106, 107]. An exhaustive search was carried out covering the main skin cancerous manifestations for which registers were found in this public database. The two most well-known microarray technologies (Illumina [3] and Affymetrix [2]) were considered for this purpose. Thus NMSC, MSC and healthy skin categories were finally chosen. The first category is comprised by the NMSC variants already mentioned in Section 2.1.1.4: BCC, SCC and MCC samples. The next category collects melanoma samples, distinguishing between two types: PRIMEL and METMEL. The last category includes those samples from healthy skin (NEV and NSK). Other important cancer manifestations such as Langerhans cells, among others, were not considered as no registers were found in the database. This fact led to not including it among the considered skin pathological states introduced in Section 2.1.1.4 for this dissertation.

Under the specified operation framework, 24 datasets from Affymetrix and Illumina platforms were selected. Specifically, 770 microarray samples were

Table 3.1: Taxonomic classifications for the three skin cancer scenarios: 2, 3 & 7 classes.

	Carcinoma (NMSC)			Melanoma (MSC)		Healthy Skin	
	BCC	SCC	MCC	PRIMEL	METMEL	NSK	NEV
7 classes	43	84	33	118	118	250	32
3 classes	160			236		282	
2 classes	396						
TOTAL	678						

contained in these 24 datasets and were collected in first instance. However, only 678 of them finally passed the quality control and were subjected to the pre-processing phase: 554 samples from Affymetrix platforms and 124 samples from Illumina platforms. In order to obtain them, these datasets are publicly available and accessible at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=S.NAME` where S.NAME is the name of each series at NCBI GEO web platform. From the collection of the all selected RNA samples, the following taxonomies were proposed (see Table 3.1 for complete information including number of samples for each category):

- tumor and healthy samples as the most general taxonomy (2 classes taxonomy).
- carcinoma, melanoma and healthy samples (3 classes taxonomy).
- BCC, SCC, MCC, PRIMEL, METMEL, NSK and NEV samples (7 classes taxonomy).

Extensively, Table 3.2 offers a summarisation of the information about the series before and after the quality control phase. Finally, Table 3.3 details the distribution of the skin samples for each microarray series used for this study.

Table 3.2: NCBI GEO series selected for this study. Criteria for series selection was getting a relative balancing of the different categories, including all possible samples from the least frequent diseases. Technology and total number of samples/outliers are included.

Series	**Technology**	**Samples origin**	**Skin states ordered by frequency (*)**	**# High quality samples**	**# Excluded outliers**
GSE2503	Affymetrix	Berlin (Deutschland)	SCC, NSK	10	1
GSE3189	Affymetrix	San Diego (USA)	PRIMEL, NEV, NSK	66	4
GSE6710	Affymetrix	Berlin (Deutschland)	NSK	12	1
GSE7553	Affymetrix	Tampa (USA)	BCC, PRIMEL, SCC, NSK	44	2
GSE13355	Affymetrix	Ann Arbor (USA)	NSK	57	7
GSE14905	Affymetrix	Gaithersburg (USA)	NSK	20	1
GSE15605	Affymetrix	Nashville (USA)	PRIMEL, NSK, METMEL	46	18
GSE29359	Illumina	New Lambton Heights (Australia)	METMEL	75	7
GSE30999	Affymetrix	Spring House (USA)	NSK	74	11
GSE32407	Affymetrix	New York (USA)	NSK	10	0
GSE32628	Illumina	Leiden (Netherlands)	SCC	14	1
GSE32924	Affymetrix	New York (USA)	NSK	7	1
GSE36150	Affymetrix	Royal Oak (USA)	MCC	10	5
GSE39612	Affymetrix	Ann Arbor (USA)	MCC, SCC, BCC	28	12
GSE42109	Affymetrix	New York (USA)	BCC	10	1
GSE42677	Affymetrix	New York (USA)	SCC	10	0
GSE45216	Affymetrix	London (United Kingdom)	SCC	28	2
GSE46517	Affymetrix	Houston (USA)	METMEL, PRIMEL, NSK, NEV	78	10
GSE52471	Affymetrix	New York (USA)	NSK	10	3
GSE53223	Affymetrix	New York (USA)	NEV, NSK	14	4
GSE53462	Illumina	Suwon (South Korea)	BCC, SCC, NSK	25	1
GSE55664	Illumina	Philadelphia (USA)	NSK	10	0
GSE66359	Affymetrix	Turku (Finland)	SCC	8	0
GSE82105	Affymetrix	New York (USA)	METMEL, NSK	12	0
TOTAL	Integrated			678	92

(*) Skin states of each series are ordered from most frequent to the lowest frequent one

Table 3.3: RNA skin samples selected after the quality control analysis.

Series	Technology	Most frequent state	Carcinoma (NMSC)			Melanoma (MSC)		Healthy skin		Total
			BCC	SCC	MCC	PRIMEL	METMEL	NSK	NEV	
GSE2503	Affymetrix	SCC		5				5		10
GSE3189	Affymetrix	PRIMEL				44		6	16	66
GSE6710	Affymetrix	NSK						12		12
GSE7553	Affymetrix	BCC	15	11		14		4		44
GSE13355	Affymetrix	NSK						57		57
GSE14905	Affymetrix	NSK						20		20
GSE15605	Affymetrix	PRIMEL				31	2	13		46
GSE29359	Illumina	METMEL					75			75
GSE30999	Affymetrix	NSK						74		74
GSE32407	Affymetrix	NSK						10		10
GSE32628	Illumina	SCC		14						14
GSE32924	Affymetrix	NSK						7		7
GSE36150	Affymetrix	MCC			10					10
GSE39612	Affymetrix	MCC	2	3	23					28
GSE42109	Affymetrix	BCC	10							10
GSE42677	Affymetrix	SCC		10						10
GSE45216	Affymetrix	SCC		28						28
GSE46517	Affymetrix	METMEL				29	35	7	7	78
GSE52471	Affymetrix	NSK						10		10
GSE53223	Affymetrix	NEV						5	9	14
GSE53462	Illumina	BCC	16	5				4		25
GSE55664	Illumina	NSK						10		10
GSE66359	Affymetrix	SCC		8						8
GSE82105	Affymetrix	METMEL					6	6		12
TOTAL	Integrated		43	84	33	118	118	250	32	**678**

3.4.2. Tools

R [108] and MATLAB [109] programming languages were used for performing this study. Most of the used R packages derived from Bioconductor platform [110]. This platform is an open-source and open-development software built in the R statistical programming environment for the analysis and comprehension of genomic data. The tools contained in the Bioconductor project represent many state-of-the-art methods for the analysis of microarray and genomic data. Other R packages come from CRAN [111], a network of ftp and web servers around the world storing identical, up-to-date versions of code and documentation for R.

3.4.3. Pipeline

Our work has been based on the steps specified in Figure 3.1, dealing with a part of the key concepts introduced in Chapter 2. Each one of the phases carried out is detailed in the next subsections.

3.4.3.1. Raw Data Acquisition and Preparation

Acquiring raw data is the very first step in any analysis. Each vendor quantifies its raw data in a different format, even with different platforms. Therefore, a particular procedure has to be applied for each series. In this study, several R packages have been used to download the microarray datasets in a programmatic manner. The Bioconductor *affy* package was used to read and process Affymetrix CEL files for their later preprocessing [112]. *GEOquery* package [113] was necessary in order to obtain already preprocessed RNA samples (when RNA samples CEL files are not available). For the newer Affymetrix microarrays, the Bioconductor *oligo* package [114] was employed. For the Illumina microarrays, the *lumi* package [115] has been used.

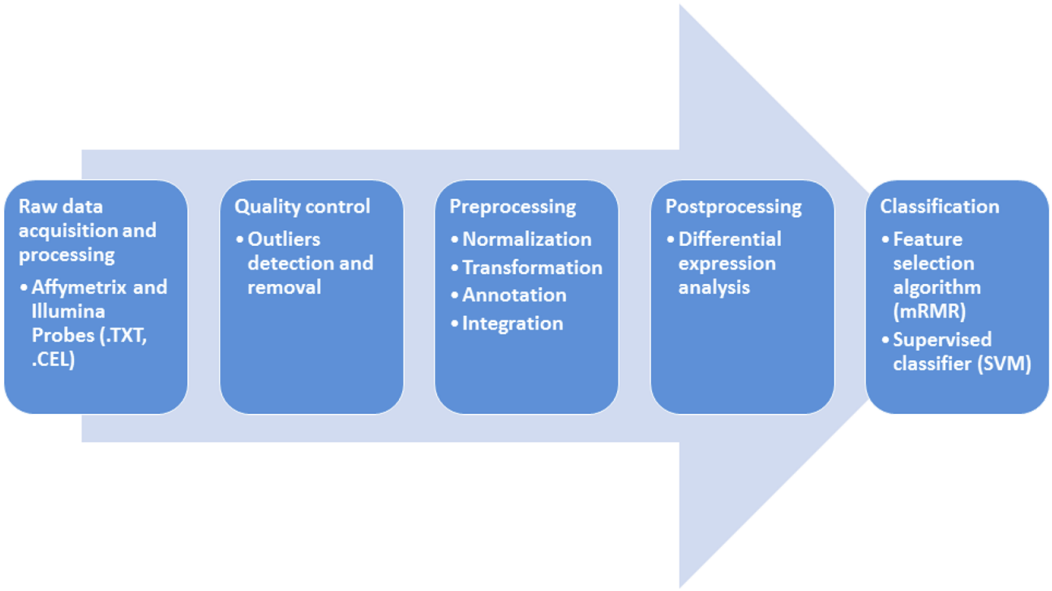


Figure 3.1: Microarray gene expression analysis pipeline. The process has been developed sequentially in different phases. This pipeline summarises the decisions made throughout the study.

3.4.3.2. Quality Control

Assessing the quality of the experiments is an essential step in microarray analysis as array-based technologies present inherent biases. Bioconductor *arrayQualityMetrics* package [116] is widely used for chip analysis and its use is not limited to one technology. It provides tests that consider quality metrics over the series samples as comparisons, intensity distributions, variance mean dependence and individual quality, for the detection of samples with insufficient quality (outliers). These tests include: distance among samples, principal component analysis (PCA), Kolmogorov-Smirnov test based on the K_a parameter, density distribution plots, standard deviation of the samples intensities and Hoeffding's D-statistic (normally executed with $D < 0.15$). All of them are iteratively applied over a given series until outliers are no longer detected or considered. Final number of excluded outliers from the considered series is shown in the last column of Table 3.2.

3.4.3.3. Preprocessing

Applying a preprocessing step on microarray data is crucial, especially when different platforms and technologies are integrated. More specifically, microarray technologies usually require normalisation, which involves a platform-dependent process necessary for converting raw data probe intensities into expression values. In this study, the Robust Multi-array Average (RMA) algorithm [25] was applied on the collected microarray data. RMA performs background correction, normalisation, and summarisation in a modular way. For Affymetrix microarrays, it can be achieved by means of the *rma* function from *affy* and *oligo* packages. In the case of Illumina microarrays analysis, the homologous *lumiExpresso* function from *lumi* package was used, allowing to do all processing steps simultaneously.

After microarray normalisation, other factors have to be taken into account for a correct microarray integration. On one hand, the logarithmic transformation must be done on the different series as well as the bit depth homogenisation. Both processes are necessary in order to avoid scale errors in further analysis. In particular, all series required logarithmic transformation in base 2. However, only 4 series had to be changed to 16-bit depth: GSE2503, GSE3189, GSE29359 and GSE55664. This type of transformations should be applied to any new sample before it can be classified correctly through the pipeline proposed in this study.

A final verification of correct series annotation was made by checking annotation data for different chips from Bioconductor AnnotationData Packages website. The main reason lies in avoiding further integration errors. They can likely come from either a missing annotation in the raw data taken from NCBI GEO web platform or after the application of the previous pre-processing R routines. Table 3.4 summarises different R packages of annotation data chips included in this work. Finally, the sample integration is possible by means of packages as virtualArray [117], readbulk [118] or inSilicoMerging [119] in association with inSilicoDb [120].

Table 3.4: Bioconductor R AnnotationData packages and available symbols for the selected series integration.

Annotation data chip	**# Platform**	**# Possible symbols**	**# Symbols with annotation**	**# Symbols NA's**	**# Symbols integrated by virtualArray**	**# Series**
hgu133a	GPL96	24549	23392	1157	12442	4
hgu133a2	GPL571	24543	23390	1153	12441	4
hgu133plus2	GPL570	58616	48709	9907	20545	11
huex10sttranscriptcluster	GPL5175	26387	21988	4399	15016	1
illuminaHumanv2	GPL6104	22916	21965	951	17296	1
illuminaHumanv4	GPL10558	50613	39181	11432	21035	1
lumiHumanAll	GPL6102	47323	31403	15920	20787	2
TOTAL (Integrated Symbols)					9978	24

(*) Different number of symbols were achieved depending on the platform and technology employed. NA ≡ Not Available.

These tools have in common that allow combining multiple microarray samples with different strategies, but not all have the necessary characteristics for this study. While readbulk can only collect heterogeneous datasets, inSilicoMerging only works with Affymetrix platforms. This last package can also normalise and remove batch effect over multiple datasets of Affymetrix technology, but virtualArray package allows merging additional datasets from other technologies as Illumina. For this reason, the package virtualArray was chosen for this approach.

Additionally, the impact of two factors on the quantification of genes can be evaluated with this tool: batch effect and union method. The first one takes into account the variations in gene expression due to biological, technical and even atmospheric agents [31]. Taking into account the hypothetical influence of this factor is considered as a compulsory step in any study of high-throughput data [121]. Currently, dealing with it is becoming challenging because there is no absolute certainty about removing the batch effects even after applying correction algorithms. An effective removal may be essential for effective integration of different datasets [37]. In this sense, the virtualArray package allows evaluating up to 6 different batch effects without losing biological information on the quantification of the gene expression: GQ [33], EB [34], NORDI [35], QD [33], MRS [33] and MC [36].The second one allows summarising in a single value all the values of expression of genes that transcribe the same gene identifier. All transcripts can be gathered into a single expression value in order to be consistent in evaluating the impact of each gene selected in the study. To evaluate its effect, this tool allows 2 union methods: mean and median. Therefore, and in search of independence in the process, a total of 12 configurations from the combination between the 6 batch effects and the 2 union methods have been tested. Consequently, only those genes that are also robust to these factors are obtained.

3.4.3.4. Post-processing

The next step in the microarray analysis methodology is calculating and obtaining DEGs. In this study, a seven-classes taxonomy was considered for DEGs identification. Then, those results were translated to the three-classes and the two-classes taxonomies for assessment.

The *limma* package [38] is commonly used since it includes interesting supplementary features: in addition to calculating DEGs, it allows making heatmaps and Venn diagrams. Although there are several statistical parameters that are taken into account in this type of studies as moderated t-statistic (T) or B-statistic (B), special attention was paid to other two parameters: log-fold change (LFC) and p-value (PV). Restrictive values for those two parameters were considered in order to guarantee statistically highly differentiated candidate genes.

This decision is motivated by the fact that certain variations can be expected among the quantification values of the genes since data are being taken from different platforms. Because of this, they could influence the selection of the genes that define the considered skin states. To avoid the potential influence of these factors, it is important to impose severe statistical restriction values on these parameters with the aim of taking those genes that are as representative as possible.

With these premises, each configuration was subjected to evaluation from the imposition of the finally chosen values for LFC and PV of 4 and 0.001, respectively. Then, a joint result was obtained, by selecting as definitive candidates based on the matches among those configurations returning candidates.

Once a set of genes has been selected, it is very important to know the robustness of the expression of these DEGs when processing microarrays from different technologies. From this perspective, the main goal is to analyse whether

the variation in the expression of these DEGs is mainly due to the different cancer-related skin states considered in this study or there are also other relevant factors involved in the processing (such as the batch effect, the country of origin of the samples or the union methods considered). In order to perform a statistical analysis that can encompass the information of all DEGs simultaneously, a dependent variable has been designed based on the Least Squares concept [122]. This algorithm takes into account the difference between the expression value of each of the candidate genes and their mean over all experiments and preprocessing variants. An ANOVA statistical test [45] was performed in order to verify the robustness of the selected genes with respect to a number of factors: "country", "type" (7 cancer-related skin states), "batch effect" and "union method". This test allowed us to confirm the study feasibility and robustness, in the selection of the identified skin cancer biomarkers.

Finally, after all the post-processing tasks were performed, the DEGs identified by the proposed methodology were consulted in different databases in order to assess their hypothetical relationship with skin cancer. DisGeNET [123], WikiGenes [124], DISEASES [125] and Open Targets [126] databases were employed for this purpose. Additionally, a text mining tool, "Gene Set to Disease" (GS2D) [127], was applied to extract the relation among the DEGs with skin diseases or disorders.

3.4.3.5. Classification

The traditional microarray data processing typically ends with the determination of DEGs. The experts can usually check these highlighted genes with laboratory experimentation or contrast them with past works. However, a great interest is aroused in relation to which DEGs are more relevant according to the analysed data groups.

This work moves one step ahead by applying ML techniques in order to gain

knowledge on the relevance of the selected genes. Similarly, a classification model is designed to automatically classify new data samples.

With the objective of discerning among the involved seven cancer-related skin states, a ranking of the most significant DEGs was obtained by using the well-known and effective mRMR algorithm [59]. This algorithm takes into account the redundancy contained among the considered genes, identifying the genes that add complementary information. This leads to attaining simpler classifiers with lower number of genes. The mRMR algorithm made use of the Kraskov Mutual Information estimator [128].

The classification technique considered in this study is the SVMs [60]. Then, two cross-validation techniques were applied to assess the classifier performance: K-Fold cross-validation (KFOLD-CV, where K = 10) [129] and Leave-One-Out cross-validation (LOO-CV) [65].

3.5. Results

3.5.1. Biological Samples Integration

24 series from Affymetrix and Illumina platforms were selected. Table 3.2 summarised the series selection process and the samples relevant to the study. 92 RNA samples were considered outliers and discarded after a strict quality control from the initial selection of 770 RNA samples. The joint representation of individual series normalisation reflected several expression value ranges (Figure 3.2). An additional preprocessing was carried out by using virtualArray tool in order to remove the samples dynamic variability, so that a homogeneous expression range was obtained for 678 high quality RNA samples (Figure 3.3).

12 different configurations, coming from six batch effects using two different union methods, were applied through this tool on all 678 RNA samples. This was made in order to invalidate the influence of intrinsic anomalies on the

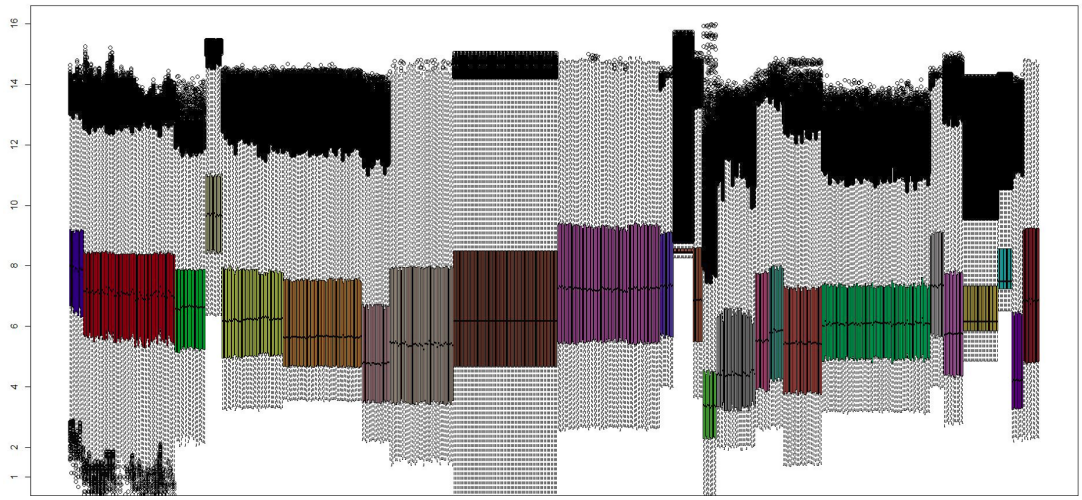


Figure 3.2: Expression values of each series after independent normalisation. The aggregation of the high quality samples shows dynamic variability among different datasets.

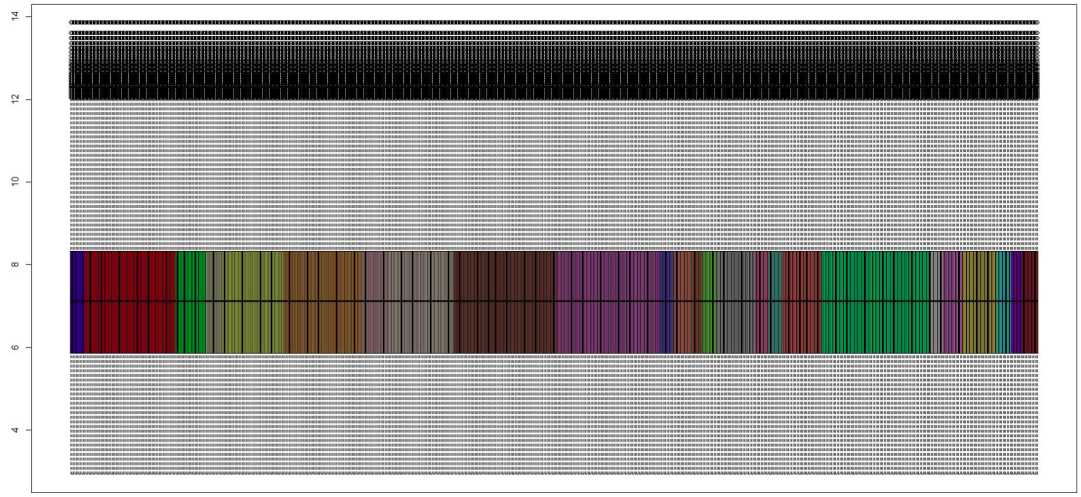


Figure 3.3: Expression values of each series after joint platforms normalisation. The integration tool used on the high quality samples reflects a homogeneous expression range.

quantification of the genes. A cross-platform normalisation and batch effect removal was simultaneously applied. Regardless of the configuration applied, 9978 genes were coerced through correct annotation (Table 3.4).

Table 3.5: Total number of obtained DEGs depending on several restrictions imposed by different evaluated configurations of virtualArray tool. The batch effect removal and union method factors were considered. The statistical parameters LFC ≥ 4 and PV ≤ 0.001 were selected.

Batch Effect / Union	GQ	QD	EB	NORDI	MRS	MC
Mean	0	25	0	0	39	39
Median	0	23	0	0	41	41

3.5.2. Expressed Genes Selection

As several heterogeneous data series were put together, and with the aim of attaining statistical robustness in the selection of DEGs, all possible batch effect validations provided by virtualArray package were tested. Similarly, strong conditions were imposed to the statistical parameters involved. Values of LFC ≥ 4, PV ≤ 0.001 were finally selected. Table 3.5 summarises the number of expressed genes after evaluating each of the 12 configurations.

DEGs appearing in several of the configuration outcomes were expected to perform robustly as potential biomarkers of skin cancer. Therefore, the intersection of candidate DEGs for configurations QD and MRS by using both union methods (configuration MC got the same results as MRS) was carried out. This guarantees that possible anomalies, coming from the heterogeneous union of datasets, would have no effect on the discriminative gene selection. The Venn diagram in Figure 3.4 shows the common DEGs among the 4 considered configurations. Resulting DEGs selected from this intersection are shown in Table 3.6; it includes the main statistical parameters presented by limma package in a summary way. Average and standard deviation values for LFC, T and B parameters were included considering the cases in where required statistical restrictions were fulfilled. Also, minimum and maximum PV were specified for these cases. Additionally, in the DEG cases column, the number of times

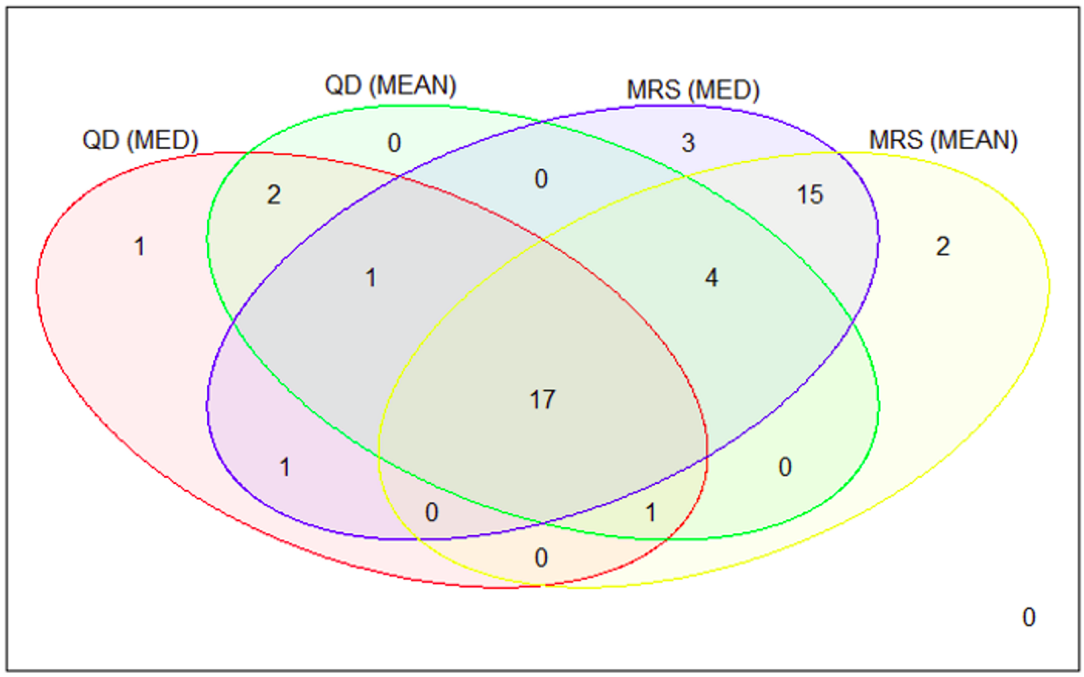


Figure 3.4: Final common DEGs obtained by considering common genes from QD and MRS results intersection. 17 common DEGs were obtained between QD and MRS effect batch removal in addition to apply union methods intersection.

each gene is present as a DEG is given. This number is calculated doing pair comparisons between two classes which results in a total of 21 pair comparisons taking into account the 7 cancer-related skin states.

3.5.2.1. ANOVA Statistical Test

Our aim is to accurately determine the influence on the DEGs when various factors or ways of treating the microarray are used. In addition, factors related to the skin pathological state analysed in this study will be included in this statistical analysis, in order to compare the statistical significance of the disease. A well-known technique such as ANOVA was used for performing this analysis. First of all, those factors are identified and distinguished (Table 3.7).

Table 3.6: List of 17 DEGs which are independent to the union method, batch removal method and multiclass problem. One of the virtualArray configurations (union method by mean, MRS batch effect and 7 classes taxonomy) was selected for showing those DEGs. All of them were listed and ordered by μ_{LFC}.

Gene Symbol	# DEG cases	$\mu_{LFC} \pm \sigma_{LFC}$	$\mu_T \pm \sigma_T$	$[PV_{min}, PV_{max}]$	$\mu_B \pm \sigma_B$	Related to skin cancer
PCP4	5	6,4649 ± 0,7160	29,8327 ± 5,3161	[1,37E-147 , 2,34E-75]	275,6113 ± 68,3340	No
TYRP1	5	5,8649 ± 0,9581	13,3323 ± 2,3796	[4,06E-47 , 4,29E-23]	72,3469 ± 24,6074	Yes (DisGeNET)
ISL1	6	5,7691 ± 0,4387	24,3430 ± 1,9227	[5,16E-106 , 1,78E-76]	204,9844 ± 24,8309	Yes (TargetValidation)
POU4F1	6	5,6337 ± 0,5265	30,5412 ± 3,4649	[2,94E-164 , 2,93E-112]	284,6693 ± 43,9275	Yes (DISEASES)
DSC3	8	5,3559 ± 0,4448	22,3521 ± 8,3026	[6,89E-179 , 1,65E-41]	180,4227 ± 104,3494	Yes (WikiGenes)
DSC1	9	5,3304 ± 0,5153	16,3742 ± 5,0402	[3,20E-111 , 7,55E-27]	108,0049 ± 60,6726	Yes (DisGeNET)
MLANA	8	5,3174 ± 1,1594	19,0411 ± 4,0277	[3,02E-94 , 1,48E-30]	138,9094 ± 48,6644	Yes (DISEASES)
SOSTDC1	3	5,0053 ± 0,4610	18,9934 ± 1,1818	[8,34E-70 , 1,05E-57]	136,8830 ± 14,5191	No (*)
TGM3	2	4,9463 ± 0,5166	19,9389 ± 1,6022	[3,32E-76 , 8,20E-64]	148,6677 ± 20,0540	Yes (TargetValidation)
CLDN1	6	4,8115 ± 0,5509	17,3556 ± 1,1748	[1,16E-63 , 3,65E-48]	116,9768 ± 14,0670	Yes (TargetValidation)
MYO15A	6	4,7396 ± 0,4262	33,1131 ± 5,4205	[3,90E-178 , 1,70E-101]	316,6801 ± 68,1730	No
BNC2	2	4,7061 ± 0,0239	24,5090 ± 3,5099	[1,56E-109 , 1,60E-81]	206,8020 ± 44,9266	Yes (DISEASES)
SCGB2A1	5	4,6701 ± 0,3740	16,6978 ± 2,7768	[6,47E-68 , 4,84E-32]	110,0158 ± 31,7911	No (*)
CRYBA2	5	4,6490 ± 0,2739	33,2294 ± 4,2398	[9,09E-172 , 3,93E-117]	318,5401 ± 53,3786	No
ANXA3	2	4,5959 ± 0,1186	17,7576 ± 1,0363	[4,05E-62 , 2,21E-54]	121,7490 ± 12,5127	No
KRT20	6	4,5916 ± 0,2552	27,1811 ± 3,9432	[7,86E-142 , 3,50E-77]	241,6544 ± 50,8763	Yes (TargetValidation)
LGR5	1	4,3562 ± 0,0000	21,5629 ± 0,0000	6,22E-79	169,1057 ± 0,0000	Yes (DisGeNET)

(*) Related to epithelial tissues

Table 3.7: Variables used in the statistical study. All the possible configurations of factors levels.

Factors	**Levels of the Factors**						
Country	GERMANY	NETHERLANDS	SOUTH KOREA	USA	UNITED KINGDOM	AUSTRALIA	FINLAND
Type	NEV	NSK	PRIMEL	SCC	METMEL	BCC	MCC
Batch	MRS	QD					
Method	MED	MEAN					

Due to the existence of multiple genes that are significant once the pipeline of genes selection is carried out (17 genes have been selected, which are presented in Table 3.6), in order to perform a statistical analysis that can encompass all the information of all those genes simultaneously, a dependent variable has been designed based on the concept of Least Squares (regression analysis method) [122]. This variable is defined as:

$$\overline{d}_i = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \sum_{j=1}^{M} (g_{i,j} - \widehat{g}_i)^2 \tag{3.1}$$

where N is the number of genes used in this study (a total of 17), M is the number of measures which have been performed in the various experiments and/or pre-processing variants with these genes (a total of 2712), $g_{i,j}$ is the value of gene i in the experiment j, and $\widehat{g}_i$ is therefore the average of the gene i in all the experiments and/or pre-processing variants:

$$\widehat{g}_i = \frac{1}{\mathrm{M}} \sum_{j=1}^{M} g_{i,j}^2 \tag{3.2}$$

Thus, because of having different data from several microarray, the influence of the gene expression (over a selected set of genes) is analysed using $\overline{d}_i$ as the dependent variable. Following, Table 3.8 gives the four-way variance analysis for the whole set of processing examples of the microarray analysed in this study. The ANOVA table containing the sum of squares, degrees of freedom, mean square, test statistics, etc., representing the experimental analysis in a compact form. This kind of tabular representation is customarily used to set out the results of

Table 3.8: Results of the ANOVA test. The statistical analysis includes the main factors assessed, such as relevant statistics parameters among which highlights associated PV.

Source (Main Factors)	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A: TYPE (*)	6,28026	6	1,04671	1521,63	*0,0000*
B: BATCH (*)	1,51534	1	1,51534	2202,88	*0,0000*
C: METHOD	0,00112746	1	0,00112746	1,64	**0,2005**
D: COUNTRY (*)	0,332589	6	0,0554316	80,58	*0,0000*
RESIDUAL	1,85524	2697	0,000687889		
TOTAL (CORRECTED)	10,1783	2711			

the ANOVA calculations.

Therefore, an ANOVA test allowed determining the influence of different factors considered on the 17 expressed common genes quantified and extracted from different microarrays. From its assessment, the analysed cancer-related skin state has been showed to be the factor with greater repercussion on the variation in the expression of such genes. Therefore, these 17 expressed common genes were cataloged as hopeful candidates for skin cancer biomarkers. Also, these genes are able to discern as much as possible among the seven skin states considered in this study. In accordance with this, Table 3.8 summarised the main statistics parameters of this analysis and supported the independent selection of any configuration for the subsequent analysis of the 17 DEGs quantification values.

3.5.3. Gene Set Assessment & Hierarchical Clustering

With the aim of illustrating the joint discriminatory power of the 17 DEGs analysed in this study, a hierarchical clustering of a selection of samples from each skin state is presented in Figure 3.5. A suitable cluster separation and a inter-cluster grouping among similar cancer-related skin states were achieved thanks to the dendrogram reorder performed by using the Ward's method [130]. On the top, both skin carcinomas (BCC and SCC) were put together. Next, both healthy skin states (NSK and NEV) and both skin melanoma states (PRIMEL

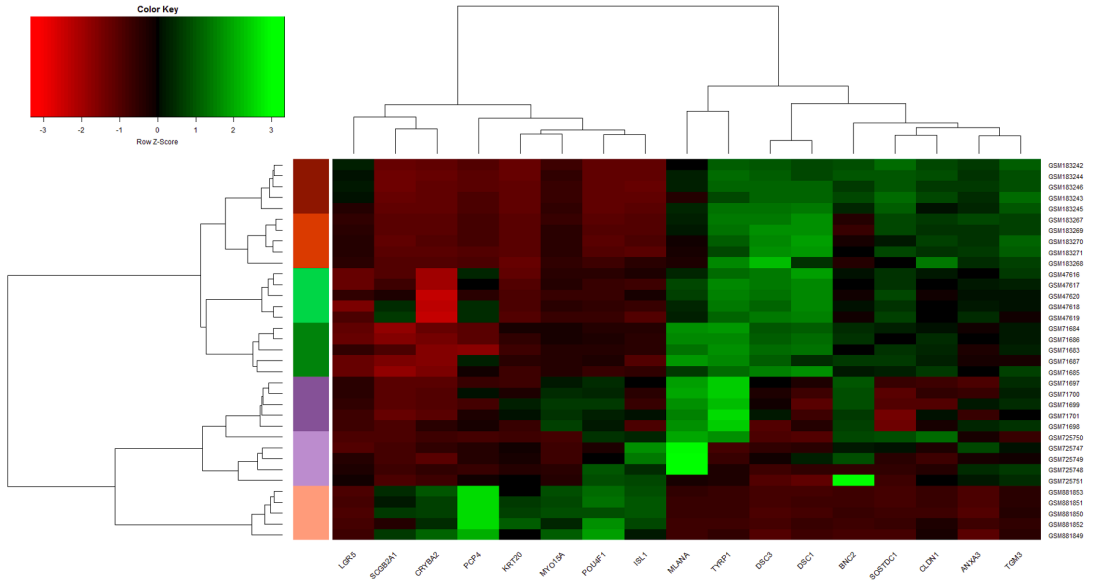


Figure 3.5: Hierarchical clustering of healthy and skin cancer samples by using the 17 DEGs. A perfect differentiation among the 7 cancer-related skin states was obtained after applying clustering and dendrogram reorder. Five samples from each skin state were used. Different colors are used for each skin sample type: NSK (light green), NEV (dark green), PRIMEL (dark purple), METMEL (light purple), BCC (chocolate), SCC (orange) and MCC (salmon).

and METMEL) were sequentially listed. At the bottom, MCC was separated from the other skin carcinomas as it practically exhibits opposite expression values for almost all the selected genes. In the light of all this, the different selected genes show to have an expectable remarkable discriminative power to differentiate among the different cancer-related skin states as well as to obtain a reliable skin cancer diagnosis.

3.5.4. Gene Relevance Identification & Classification Process

An assessment of the quality of the information provided by the 17 validated DEGs is necessary in order to reduce the complexity of the study. It also allows to limit the effective diagnostic potential of skin cancer to only a small set of genes.

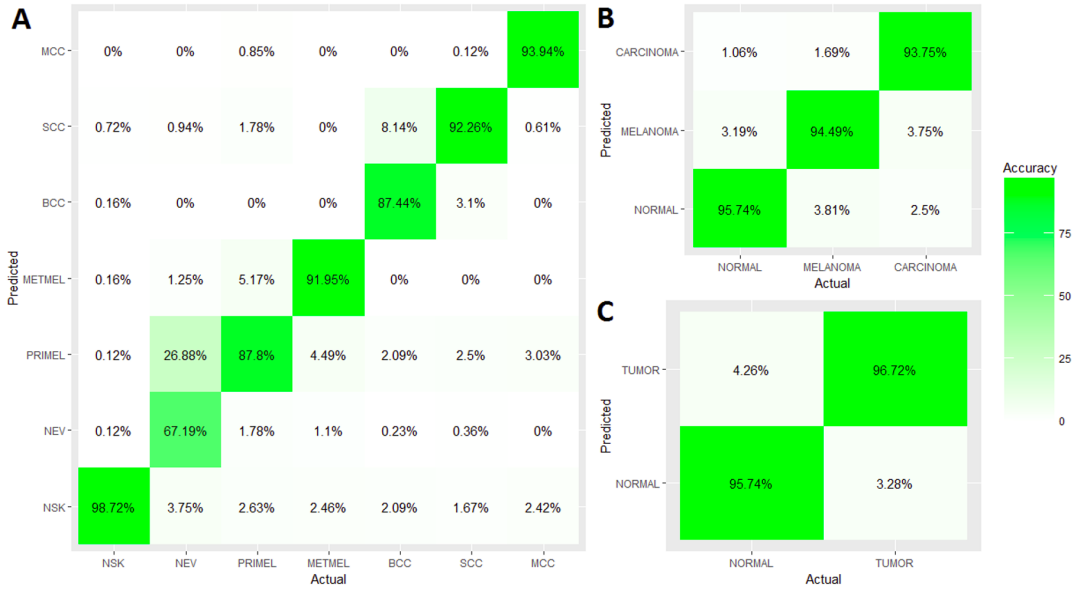


Figure 3.6: Classification accuracy achieved for each of the considered taxonomies: (A) 7 classes, (B) 3 classes and (C) 2 classes. The confusion matrix for taxonomy A was constructed with 10-CV and 17 DEGs. The other confusion matrices were constructed from the previous, by summing the respective sub-matrices associated with each skin super-state.

Different databases were consulted with the aim of checking the relationship between these genes and skin cancer. Table 3.6 points out if the identified DEGs were previously reported as related to the cancer-related skin states, according to the consulted databases. Full and exhaustive information about the biological relationship of these genes with skin cancer and other cancers can be consulted in S3 Appendix [131]. Main insights and findings about the involvement of these genes in skin cancer are discussed below in Section 3.6.2.

In order to assess a hypothetical classification procedure, special precaution must be taken regarding the information provided by the selected set of genes in a new skin sample. For this reason, a classification model based on SVM multiclass was designed together with two cross-validation processes (LOO and 10-FOLD) for its assessment. The results reflect an overall accuracy recognition for the 7 cancer-related skin states considered up to 92% for both

cross-validation processes. Translating this percentage into the 2 additional taxonomies considered of 3 classes (melanoma, carcinoma and healthy skin) and 2 classes (tumoral and healthy skin), this percentage increased to 95% and 96%, respectively. The associated confusion matrices can be seen in the Figure 3.6.

This previous result does not allow appreciating objectively the informative contribution of each gene to the skin state recognition. For this reason, the mRMR algorithm was employed in order to obtain a ranking of these genes according to their potential in the seven skin states discernment. The genes ranking returned by the algorithm is as follows: DSC3, SCGB2A1, BNC2, TYRP1, ISL1, DSC1, MLANA, CRYBA2, ANXA3, PCP4, LGR5, CLDN1, POU4F1, SOSTDC1, KRT20, TGM3 and MYO15A. The expression value distribution of each selected gene sorted by this ranking over each of the cancer-related skin states can be seen in the Figure 3.7.

Next, distinct SVM models were designed and retested by cross-validation processes in order to assess the classification capacity of different subgroups of genes returned by this ranking. The gene ranking classification results on the three considered taxonomies can be seen in the Figure 3.8. Finally, an evaluation of the designed classifiers behaviour was carried out for each of the cancer-related skin states. The accuracy results for each skin state and for each gene subset are showed in the Figure 3.9.

3.6. Discussion

3.6.1. Heterogeneous Dataset Integration & Expressed Gene Selection

Two main reasons motivate the integration of multiple gene expression datasets. Firstly, an extensive quantity of high quality samples from different platforms and

Figure 3.7: Expression level of the selected genes ordered by the ranking returned by mRMR algorithm. Different colors are used for each cancer-related skin state: NSK (Normal Skin), NEV (Nevus), PRIMEL (Primary Melanoma), METMEL (Metastatic Melanoma), BCC (Basal Cell Carcinoma), SCC (Squamous Cell Carcinoma) and MCC (Merkel Cell Carcinoma).

technologies must be put together. This decision enriches the heterogeneity of the study, thus reinforcing its reliability and statistical robustness as well. Secondly, resulting from the previous reason, the independence of the results obtained can be guaranteed by analysing a wider heterogeneous dataset. The collection of a large repertoire of samples increases significantly the dimensionality, the diversity and the complexity of the experimental analysis, more so when it comes to addressing a multiclass problem. This ambitious challenge is driven by jointly analysing multiple batches where each of them collects only a part

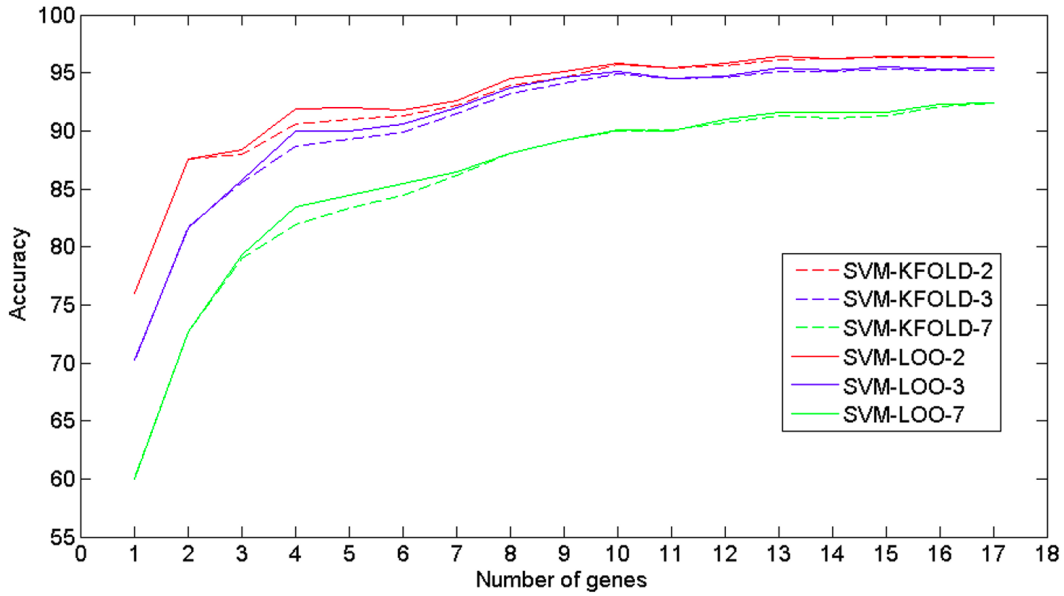


Figure 3.8: Evolution of the classification accuracy for each subset of genes considered, and for each taxonomy. Similar trends can be observed for both LOO-CV and 10-CV.

of the classes involved in the final approach design. Table 3.2 reflects how the heterogeneity can be achieved by taking into account samples that have been experimentally processed at different time points, from different technologies and different platforms. Moreover, a large racial diversity can be expected given the origin of the samples. As a result of the foregoing, Table 3.1 includes the 678 RNA samples that were finally considered after a strict quality control phase. These samples represent 7 different cancer-related skin states from which was aimed to extract genes that may be truly representative of their manifestation. By considering several series with different number of skin states, the emergence of batch effects may become inevitable and could be seen as a possible limitation because of the partial association between series and skin states. However, in spite of the great heterogeneity that can be observed from the expression values of the 24 unprocessed series (Figure 3.2), a simultaneous preprocessing step across all the samples attains an homogeneous expression range (Figure 3.3).

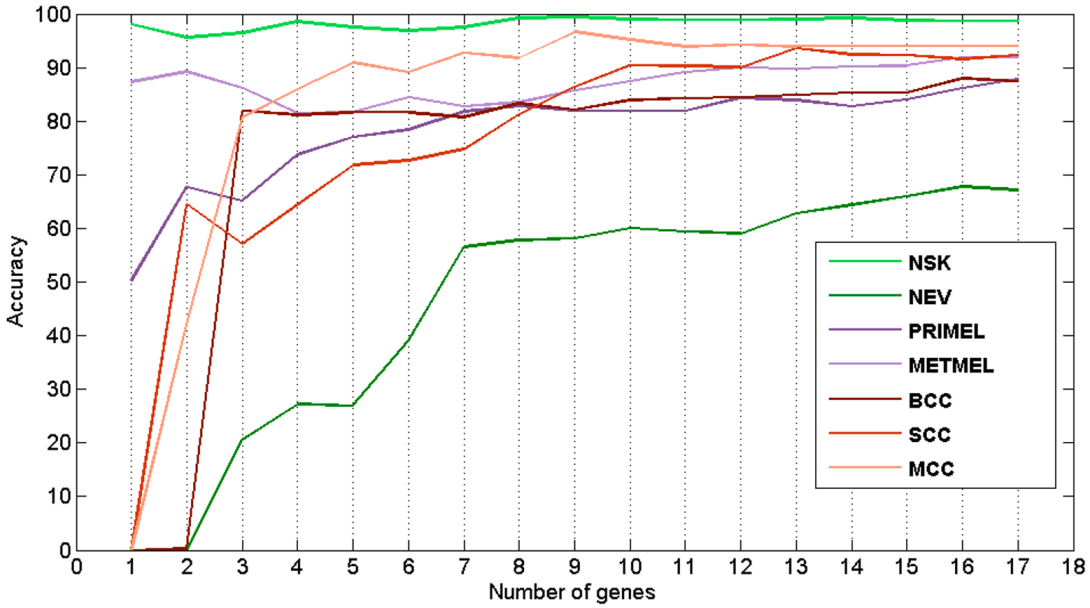


Figure 3.9: Evolution of the classification accuracy for each cancer-related skin state according to the number of genes from the mRMR ranking considered in the classifier. Different colors are used for each skin sample type: NSK (Normal Skin), NEV (Nevus), PRIMEL (Primary Melanoma), METMEL (Metastatic Melanoma), BCC (Basal Cell Carcinoma), SCC (Squamous Cell Carcinoma) and MCC (Merkel Cell Carcinoma). SVM with 10-CV was used.

In the translation from samples to genes, only those common genes that have the same coded symbol for any considered microarray platform, will appear after heterogeneous sample integration. The lack of uncommon gene symbols from different platforms is an assumed trade-off since the main purpose of this study is to integrate as many samples as possible that significantly represent each cancer-related skin state. Table 3.4 showed how the series from GPL96 and GPL571 Affymetrix platforms could integrate a little more than 12400 genes. This imposes a maximum number of potential genes that may eventually appear as common after the microarray integration. However, those series contain more than half of the PRIMEL samples (specifically, 73) and almost three quarters of the total NEV samples (in this case, 23). Not including those series would have

had direct repercussions on the balance of classes and their representativeness in the study.

In view of this decision, a total of 9978 genes with common symbols appeared after integration and were exposed to the statistical significance process. In order to obtain genes that can become robust and reliable, very restrictive values were imposed for the statistical parameters LFC and PV. At this point, ensuring the statistical significance of the selected genes is thought to be primordial. This imposition can restrict the finding of skin cancer biomarkers that are strongly invariant against different anomalies or deviations. Under these restrictions, a small set of genes were highlighted by the tested configurations as presented in Table 3.5. Those genes were obtained from the intersection of configurations returning candidate biomarkers as shown in Figure 3.4. The final validity of the selected 17 gene set has been supported through the application of a statistical test. That test confirms the relevance of those genes to classify the 7 different cancer-related skin states versus other intrinsic factors of the heterogeneous datasets integration.

3.6.2. Biological Relevance of the DEGs

The relevance of these DEGs in the diagnosis of cancerous manifestations on the skin was investigated from an exhaustive search in the literature. Table 3.6 summarised how 11 of the 17 highlighted genes have already been strongly related to skin cancer in previous studies (ISL1, POU4F1, CLDN1, TYRP1, DSC1, TGM3, DSC3, BNC2, KRT20, LGR5 and MLANA). Regarding the 6 remaining genes, 2 of them have been linked to epithelial tissues (SOSTDC1 and SCGB2A1). The other 4 genes have not been previously highlighted as reliable biomarkers of the disease (PCP4, MYO15A, ANXA3 and CRYBA2). Additionally, Table 3.9 reflects the outcome of the "Gene Set to Disease" (GS2D) text mining tool for the identified DEGs.

Table 3.9: Relation between DEGs in this study and different skin diseases or disorders. A minimum number of two disease-related citations for each gene was selected, as wll as one gene significantly associated at least with a disease. A maximum False Discovery Rate (FDR) equal to 0.05 was imposed.

Disease	# Genes involved	Genes (%)	LFC	PV	FDR	Gene symbols
Hutchinson's Melanotic Freckle	1	0.06	9,4778	2,80E-03	2,10E-01	MLANA
Dermatitis Herpetiformis	1	0.06	8,4778	2,80E-03	1,05E-01	TGM3
Lentigo	1	0.06	7,4778	5,60E-03	7,00E-02	MLANA
Unknown Primary Neoplasms	1	0.06	6,4777	1,12E-02	9,31E-02	ISL1
Oculocutaneous Albinism	1	0.06	5,7778	1,81E-02	1,36E-01	TYRP1
Merkel Cell Carcinoma	1	0.06	5,3903	2,36E-02	1,47E-01	KRT20
Pemphigus	1	0.06	4,6702	3,86E-02	1,93E-01	DSC3
Vitiligo	2	0.12	4,6450	2,81E-03	7,04E-02	MLANA, TYRP1
Ichthyosis	1	0.06	4,6200	3,99E-02	1,87E-01	CLDN1
Nevus	1	0.06	4,3903	4,67E-02	1,94E-01	MLANA
Circulating Neoplastic Cells	2	0.12	4,2683	4,70E-03	7,05E-02	LGR5, KRT20
Experimental Melanoma	1	0.06	3,6702	7,58E-02	2,37E-01	TYRP1
Neuroendocrine Tumors	1	0.06	3,5472	8.23E-02	2.47E-01	ISL1
Skin Diseases	1	0.06	3,1027	1,10E-01	2,76E-01	DSC3
Basal Cell Carcinoma	1	0.06	2,6826	1,45E-01	3,11E-01	TYRP1
Atopic Dermatitis	1	0.06	2,2780	1,88E-01	3,52E-01	TGM3
Lymphatic Metastasis	4	0.25	1,6276	3,54E-02	1,90E-01	ANXA3, LGR5, CLDN1, KRT20
Skin Neoplasms	2	0.12	1,1440	2,28E-01	3,97E-01	MLANA, TYRP1
Adenoma	1	0.06	0,9561	4,08E-01	5,78E-01	LGR5
Carcinogenesis	1	0.06	0,9184	4,16E-01	5,78E-01	LGR5
Neoplasm Metastasis	1	0.06	-0,8365	8,50E-01	9,11E-01	CLDN1
Melanoma	2	0.12	0,6871	3,56E-01	5,34E-01	TYRP1, MLANA
Squamous Cell Carcinoma	3	0.19	0,6781	2,87E-01	4,79E-01	DSC3, CLDN1, TGM3
Genetic Predisposition to Disease	6	0.38	-0,6666	9,80E-01	9,93E-01	DSC3, BNC2, TYRP1, ISL1, LGR5, MYO15A
Carcinoma	1	0.06	-0,2515	7,11E-01	7,73E-01	LGR5
Neoplasm Invasiveness	3	0.19	0,2016	4,99E-01	6,24E-01	LGR5, CLDN1, KRT20

From these results, 13 of the 17 genes in this study have been related to some pathology, disorder or disease of the skin, including the cancer-related skin states studied in this work. However, in addition to the possible relationship of the expressed genes with different cancerous manifestations and skin diseases, it is important to emphasise that 4 genes (ANXA3, LGR5, CLDN1 and KRT20) have been related to lymphatic metastasis. Similarly, 6 genes (DSC3, ISL1, TYRP1, LGR5, MYO15A and BNC2) are related to genetic predisposition to disease. In this sense, the potential relevance of these genes surpasses the scope of this study: the potential biomarkers not only reflect their relationship with different cancerous manifestations of the skin but they also seem to have some relationship with the predisposition to metastasise and to become ill.

3.6.3. Gene Ranking Assessment

Although the potential of all the identified DEGs as skin cancer biomarkers became evident, an additional evaluation of the actual information provided in a possible diagnosis test was carried out. Two additional objectives were aimed: on one hand, to further reduce the final repertoire of DEGs in order to decrease the test complexity; on the other hand, to check the potential relevance of each of the considered DEGs, especially those that were not previously related to skin cancer. The procedure followed in this study was the evaluation of different classifiers taking into account the gradual insertion of the highlighted genes according to their maximisation of the discernment capability among the cancer-related skin states. From the mRMR algorithm point of view, this is translated to a gradual increase of mutual information between the identified DEGs and the skin states, avoiding as much as possible the redundancy among them.

3.6.3.1. Gene Relevance Analysis

DSC3 gene was chosen from the selected gene set as the most discriminating gen by mRMR algorithm to differentiate among the 7 cancer-related skin states. This gene, which has already been previously cataloged as skin oncogene, tends to present low gene expression levels on patients who suffer from skin melanoma (see S3 Appendix in [131]). This can be seen in Figure 3.7, where gene expression levels for each gene in each skin state are observed. In this gene, only its PRIMEL gene expression wide range prevents separating this skin state from the rest. Even so, DSC3 allows separating those skin states that present a greater probability of provoking malignant tumor formations and spreading (PRIMEL, METMEL and MCC) from those less aggressive or simply healthy skin states (BCC, SCC, NSK and NEV).

Following, the mRMR algorithm selected the SCGB2A1 gene as the next with more information to discern among the 7 cancer-related skin states. In this gene, at least 2 groups can be easily differentiated: 1) NSK together with MCC, and 2) the rest. It is noteworthy that this gene, that had never been related to skin cancer before, appeared in second position. However, this gene has certainly been linked to epithelial tissues and other cancers (ovary, prostate, uterus, primary and occult breast, liver, colorectal, etc.), and what it is more important, with up-expression in almost all of them. From Figure 3.7, we observe that in skin cancer, gene expression levels appeared down-expressed with respect to NSK for the remaining cancer-related skin states, except for MCC. In this sense, there is evidence that its gene expression level is lower for cancer-related skin states than for the other healthy skin state (NEV). For all of this, this gene could be a novel and valid biomarker that provides clues about the predisposition to suffer from some type of skin cancer. Extended information can be consulted in S3 Appendix in [131].

BNC2 gene was ranked in third position by the feature selection algorithm.

Already previously accepted as skin oncogene, this biomarker allows clearly differentiating among the 2 most diagnosed skin carcinomas (BCC and SCC). Additionally, its expression adds complementary information to what it is already provided by DSC3 and SCGB2A1, providing a better discernment among the 7 cancer-related skin states.

The gene expression differences for each of the next selected genes in the ranking can be also observed in Figure 3.7. It should be noted at this point that, although the mRMR ranking proposes genes with greater ability to discriminate among cancer-related skin states than others, all of them present relevant information for the specific skin states diagnosis. For example, several genes from the final part of the ranking, present specific clear information on MCC against the rest skin states as LGR5, POU4F1, SOSTDC1, KRT20, TGM3 and MYO15A genes, as their gene expression levels are opposite against to the other cancer-related skin states. Among all of them, up-expression of POU4F1 and KRT20 genes was previously related to MCC. Surprisingly, although LGR5 and TGM3 have been linked before to BCC risk, they showed here down-expressed values in MCC (Figure 3.7). Even going beyond, SOSTDC1 and MYO15A have not been previously reported as skin cancer biomarkers. However, they show down-expression and up-expression in MCC, respectively. On the other hand, PCP4 gene appeared as down-expressed in several skin states with respect to NSK as well as so did SCGB2A1. More biological details about these genes and their relationship with skin cancer can be seen in S3 Appendix in [131].

3.6.3.2. Accuracy-Complexity Trade-Off

Although only 17 genes fulfilled all the statistical constraints and a high overall recognition rate was obtained, there are chances that not all of them have a direct influence on improving the classifier performance. In this regard, a detailed analysis of the influence of each DEG on the classifier improvement can be made.

Multiple interpretations could be drawn from the gene relevance analysis. On the one hand, it could be achieved from the interlaced analysis of their distribution on each cancer-related skin state. On the other hand, together with the previous one, it could be analysed from their influence on the classifying power of the classifier model both in the global recognition and in the specific recognition of each skin state. Thus, in search of informative power for the genes to be finally selected for the diagnosis tool, a classification accuracy improvement was assessed, by gradually adding genes from the ranking into the classifier.

The actual contribution of each gene to the classifier can be more clearly verified from the overall and specific trends in the evolution analysis seen in Figure 3.8 and Figure 3.9. If the 17 DEGs are used, an overall accuracy above 92%, 95% and 96% can be attained when the 7, 3 and 2 classes taxonomy are used. The curves associated with each taxonomy evolve similarly for both cross-validation processes. This fact indicates that a great robustness was reached in this study from the large sample integration, which leads to the convergence of both validation processes. With respect to the 7 classes taxonomy curve trend, an ascending order is clearly observed as the genes are introduced into the classifier. Therefore, it shows that there is a gradual real information input.

Since the 3 and 2 classes taxonomies results were obtained from the 7 classes confusion matrix summary, there are certain local convergence zones in their accuracy evolutions. These events occur among the fourth and sixth genes, and from the tenth gene, from which the accuracy practically reaches its maximum value. Therefore, this quantity of genes can be considered as a suboptimal gene subset, allowing to establish a trade-off between the number of genes considered for the diagnosis model and its accuracy. Precision rounded 95.5% for 2 classes, 95% for 3 classes and 90% for 7 classes for the 10 genes model. This implies a decrease of around 2% of accuracy in the classifier performance for the main 7 classes taxonomy, at the expense of reducing in 40% the number of genes needed

for diagnosing. Thus a simpler diagnosis model is possible, with the resulting economical and time reduction. To sum up, different accuracy-complexity trade-offs can be raised depending on the benefits that intend to be optimised:

(**a**) **Minimum number of genes:** 4 DEGs, accuracies around 92% (2 classes), 90% (3 classes) and 83% (7 classes).

(**b**) **Maximum accuracy:** All 17 DEGs, accuracies around 96% (2 classes), 95% (3 classes) and 92.5% (7 classes) (see Figure 3.8).

(**c**) **Accuracy-genes trade-off approach:** 10 DEGs, accuracies around 95.5% (2 classes), 95% (3 classes) and 90% (7 classes).

Figure 3.9 showed how different accuracy evolutions were reached by each cancer-related skin state as the genes were gradually aggregated into the classifier model. For example, with only the first 3 genes (DSC3, SCGB2A1 and BNC2), an accuracy above 80% is insured for 4 skin states (NSK, METMEL, BCC and MCC). By selecting 10 genes as trade-off, high classification rates are reached for most cancer-related skin states: NSK (99%), PRIMEL (82%), METMEL (90%), BCC (84%), SCC (90%) and MCC (96%).

These observations suggest that different gene rankings could be returned when pursuing an optimal classification of a specific cancer-related skin state. For example, although MCC shows expression values contrary to the rest of skin states in the identified DEGs, there are genes like LGR5, POU4F1, SOSTDC1, KRT20 and MYO15A which are clearly postulated as differentiating genes in MCC diagnosing in comparison to other cancer-related skin states. However, their contribution on the MCC diagnosis improvement can not be appreciated because these genes were ranked after eleventh position and the diagnosis of this skin carcinoma does not improve after the ninth gene as can be seen in Figure 3.9. From the same figure, a similar conclusion can be drawn from the PCP4 gene that was ranked in tenth position and its potential informative power for diagnosing

some skin state seems to be irrelevant despite having a distribution similar to SCGB2A1.

3.7. Conclusions

Through a restrictive pipeline process, 17 DEGs were obtained for discriminating up to seven cancer-related skin states from the integration of multiple skin cancer datasets. In the light of all results and discussions presented in this study, these genes have been seen as reliable skin cancer biomarkers. Consequently, they are expected to serve as a guide to improve the early diagnosis of skin cancer because these indicate the potential predisposition to suffer from it. Many of these genes have been linked even to other pathologies or disorders of the skin that are considered as precancerous skin states.

The vast heterogeneity of the sample collection with respect to diverse factors like platforms, origin, parts of the body, etc. positively influenced in the finding of 6 genes that had not previously related to skin cancer: SCGB2A1, CRYBA2, ANXA3, PCP4, SOSTDC1 and MYO15A. In this sense, beyond the importance of each DEG in the overall recognition, the relevance analysis of each DEG showed the differentiating role of the SCGB2A1 gene. This is greatly due to the fact that the massive heterogeneous sample integration has allowed extracting extremely useful underlying information from the joint study of up to 7 different cancer-related skin states. SCGB2A1 appeared as down-expressed for all the cancer-related skin states, but MCC. The same gene was also down-expressed for the NEV state, in comparison with NSK gene expression levels. In terms of accuracy recognition, an overall recognition around 92.5% of accuracy has been achieved to distinguish among 7 cancer-related skin states. More briefly, an accuracy of 96% is guaranteed to discriminate between healthy and tumor samples from the 17 DEGs.

Our next objectives include the idea of using this pipeline in other types of cancers or diseases with a good number of existing samples from public repositories, available private data or even from further generation sequencing techniques, having data quantified in gene expression values. Specifically, the proposed integration scheme is expected to allow the co-integration with more innovative state-of-the-art technologies such as RNA-seq. Additionally, modifications of the general pipeline are aimed to be used in the improvement of the diagnosis of those cancer-related skin states with lowest diagnostic accuracies.

4. Integrating Transcriptomic Technologies

4.1. Introduction

This chapter presents a new methodological approach that integrates skin cancer datasets at the gene expression level, and whose information comes from the 2 co-existing sequencing technologies: microarray and RNA-seq. The study aims to take a step forward, reinforcing the methodology presented in Chapter 3 from several fronts: the consideration of precancerous diseases, the implementation of an algorithm for selection of biomarkers and an improvement in the classification process.

The study has made use of the following resources:

- **Number of datasets:** 27
- **Number of samples:** 968
- **Skin Pathological States:** 10 (NSK, NEV, BCC, SCC, PMCC, MMCC, PRIMEL, METMEL, AK and PS)
- **Web Data Repositories:** NCBI GEO and AE
- **Sequencing Technology:** Microarrays and RNA-seq

As in the previous Chapter 3, it is suggested to review the concepts and resources previously explained in Section 2.1.1 for a better understanding of the procedure presented here. The content included in this chapter is a part of the submitted journal article entitled *"Towards Improving Skin Cancer Diagnosis by Integrating Microarray and RNA-seq Datasets"* (revision process).

4.2. Background

Skin cancer is a worrying complex disease taking a wide range of skin pathological states (SPSs). The complex heterogeneity of its occurrence is determined by the abnormal and out of control proliferation of specific cells (squamous, basal, Merkel, melanocyte, keratinocyte, etc.) that incur the development of multiple skin cancerous pathologies. Among them, the most frequent in order of incidence are related to non-melanoma skin cancer (NMSC) which is led by basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and Merkel cell carcinoma (MCC) [69]. With regard to melanoma skin cancer (MSC), the main pathologies can be summarised in primary melanoma (PRIMEL) and metastatic melanoma (METMEL) whose mortality rate is higher [132]. The concerning current global trend is reflected in epidemiological studies that show how the incidence and occurrence of both MSC and NMSC cases have already become the most common types of cancer in white populations [133]. This is supported by the statistical analyses of cohorts of MSC rates on United States whites, United Kingdom, Norway and Sweden which increased up to 3% annually during the last 3 decades [134]. With respect to NMSC cases, its incidence is around 20 times higher than MSC cases [135] despite being widely understudied. As a result of the fateful combination of both factors, an extensive global alarm is being increased together with the possibility of suffering from any skin cancer type by two main drivers: on the one hand, because of tumor evolution of other

skin diseases previously considered precancerous states such as psoriasis (PS) [23, 136, 137] or actinic keratosis (AK) [24, 138], and on the other hand, because of tumor degeneration and mutation from healthy states such as normal skin (NSK) and nevus (NEV). The narrow biological relationship among several SPSs may complicate the successful diagnosis of skin cancer. Certain researches have pointed out the difficulty in discerning among specific SPSs from the clinical, histological and molecular points of view: AK vs SCC [139], AK and SCC vs PRIMEL [140], SCC vs BCC and MSC [141], primary MCC (PMCC) vs metastatic MCC (MMCC) [142], etc. Different editions of the American Joint Committee on Cancer (AJCC) have gradually introduced the most outstanding clinical parameters for the diagnosis (tumor mitotic rate, TNM classification, Breslow thickness, Clark levels, etc.). Consequently, the AJCC Cancer Staging Manual has been considered the gold standard by clinicians when making their diagnoses [143]. However, the way to diagnose this cancerous disease continues to be limited and each AJCC edition implies controversies and corrections on which are the best criteria to efficiently diagnose each SPS. Conversely, other studies insist on the possibility of differentiating them from the identification of gene expression patterns such as AK vs SCC [144]. Although discerning among SPSs by using DEGs has been revealed, the biological complexity of the skin cancer may put its validity into question.

4.3. Motivation

The opportunity to efficiently improving the discernment among multiple SPSs related to cancer from biological data involves taking into account a set of requirements. Firstly, different technological alternatives which allow to quantify in terms of gene expression have to be inspected. Although microarray technology has been vastly used, RNA-seq technology is definitely ending up replacing

it thanks to various notorious advantages [145]: i) RNA-seq allows detecting the variation of a single nucleotide; ii) it does not need genomic sequence knowledge; iii) it provides quantitative expression levels and isoform-level expression measurements; and finally, iv) it offers a broader dynamic range than microarrays. Nonetheless, the absence of open access datasets from experiments of the newest technologies still invites to consider analysing microarrays. In addition to its low cost, it may not have been properly exploited yet because of being analysed for isolated experiments. By combining diverse skin cancer datasets containing samples of different SPSs, there is the chance to reinforce the statistical robustness of the study as well as to obtain highly DEGs from a wider range of SPSs. This fact adds the challenge of adequately integrating data from both technologies in order to increase as much as possible the repertoire of samples of each identified SPS for the study. Previous studies have proven the consistency of applying multi-platform integration among both microarray platforms and technologies at gene expression level [146–149], encouraging to continue carrying it out. However, the researchers have traditionally kept in mind the mandatory correction of eventual batch effects with the purpose of achieving an effective integration of multiple experiments over different microarray platforms [121], mainly coming from two manufacturers: Affymetrix [2] and Illumina [3]. By additionally taking into account experiments conducted on RNA-seq technology, the hypothetical influence of this factor may be modified in an unpredictable way. Although despite the efforts to remove them completely, there is not even certainty that a complete elimination of these effects will take place [37], the treatment and the attempt of correction should never be disregarded. Among the multiple batch effect correction algorithms, ComBat [34] has been proven to show the highest effectiveness when integrating microarrays [31] and, recently, has been strongly validated by integrating RNA-seq datasets from different sources: GTEx and TCGA projects [150]. In the case of favorably

dealing with all these limitations, discerning multiple SPSs by using changes in gene expression implies a new experimental challenge. Although hierarchical clustering highly helps in graphically showing such changes [94], methodological approaches based on multiclass classification are postulated as an innovative alternative when assessing the validity of DEGs for simultaneously diagnosing multiple SPSs [131]. Finally, the use of feature selection algorithms must be explored with the objective of selecting only informative DEGs, thus dramatically reducing the search space. Under the fulfillment of the previous premises, the integration of microarray and RNA-seq technologies at gene expression level [151] opens new possibilities for skin cancer analysis. Concretely, this advance could improve the understanding about the hypothetical biological relationships and differences among SPSs that may be discerned in a simple simultaneous analysis. Clinicians could directly benefit from its validity in multiple ways. Firstly, the suspicions about the patient tumor evolution from healthy skin states to cancerous states, even through precancerous skin diseases, could be eventually assessed by presenting certain genetic susceptibility to change [77]. A personalised and patient-oriented medical service could be derived from the above by knowing the genetic signs. Consequently, unnecessary medical treatments such as radiation therapies, excision surgeries or medications supply could be prevented [105]. Definitely, their diagnosis decisions could be supported thanks to the use of an intelligent diagnosis tool that offers another complementary point of view [103]. In view of the benefits and clinical coverage that its use could offer, this study presents a novel methodological approach that addresses all the requirements presented to advance in the improvement of the skin cancer diagnosis. The integration of different skin cancer datasets from microarray and RNA-seq technologies based on gene expression analysis has not been widely explored by the scientific community. First of all, an exhaustive sample search of multiple SPSs was carried out from public data repositories. Thus, 22 microarray

and 5 RNA-seq series containing 1090 samples in total were finally collected. However, after applying a strict quality control phase, only 968 samples passed and were subjected to the preprocessing phase: 666 samples from Affymetrix and Illumina microarray platforms and 302 samples from Illumina RNA-seq platforms. Subsequently, the sample integration consisted in considering only those genes sharing a common annotation for all the series selected for this study. After merging multiple batches and applying batch effect correction on them, the challenge was to efficiently find valid genes to simultaneously discern up to 10 SPSs: from a priori healthy states (NSK and NEV) to cutaneous carcinomas (BCC, ISCC, PMCC and MMCC) or melanomas (PRIMEL and METMEL), including skin diseases with a higher risk of tumor degeneration that have already been cataloged as precancerous states (AK and PS). From the assessment of a highly heterogeneous multiclass dataset of 968 samples and almost 7700 genes, a subset of DEGs was identified by applying a simple one-vs-one (OVO) multiclass gene selection algorithm. This was achieved by means of consciously tuning critical and highly selective parameters. Concretely, log2 fold change (LFC) and maximum number of selected DEGs (NMAX) among each pair of SPSs were considered. By relying on a widely used feature selection algorithm and assessing different subgroups of multiclass candidate DEGs, an ANOVA statistical test [152] assessed the influence of these critical parameters together with the use of different classification models and performance metrics. Finally, the biological relationship of these DEGs with skin cancer was consulted by examining their functional properties and inspecting specific literature.

4.4. Methodologies and Experiments

Under the operation of a specific designed pipeline, an overall flowchart of our approach is presented (Figure 4.1). Each of the experimental steps of this

proposed pipeline will be sequentially addressed in the following subsections.

4.4.1. Transcriptomics Technologies Integration

To obtain the integration of skin cancer datasets coming from different platforms and technologies, three steps have to be carried out (see left part in Figure 4.1).

4.4.1.1. Raw Data Acquisition

One of the first steps involves carrying out an in-depth information search about skin cancerous pathologies and, subsequently, finding out the current availability of datasets. On the one hand, AK and PS have been previously cataloged as precancerous skin diseases. On the other hand, a wide range of SPSs related to cancer have been specified: from carcinomas (BCC, SCC or MCC) to melanomas (PRIMEL and METMEL), to even lymphomas or sarcomas. Next, the identification of transcriptomics webdata resources implied inspecting the availability of the above SPSs together with healthy states (such as NSK or NEV) in public repositories such as NCBI GEO [106] and ArrayExpress [8] web platforms. Initially, guidelines indicated in Section 2.1.1.1 for sample selection were followed. Moreover, only those SPSs containing a representative number of samples were considered in order to increase the possibilities of characterising their manifestation [153]. Under these considerations, Bowen's disease samples (also known as SCC in situ) were not finally considered (only two datasets containing data samples from this SPS were found, summing up to only 12 samples which was considered too low for the study). Extensively, no representative number of lymphoma and sarcoma samples were available to be considered in this study. By dealing with different microarray technologies and platforms, several R packages were considered from Bioconductor web platform [154] in order to acquire the RNA samples: from GEOquery [113], affy [112] and oligo [114] for different Affymetrix platforms to lumi [115] for Illumina platforms.

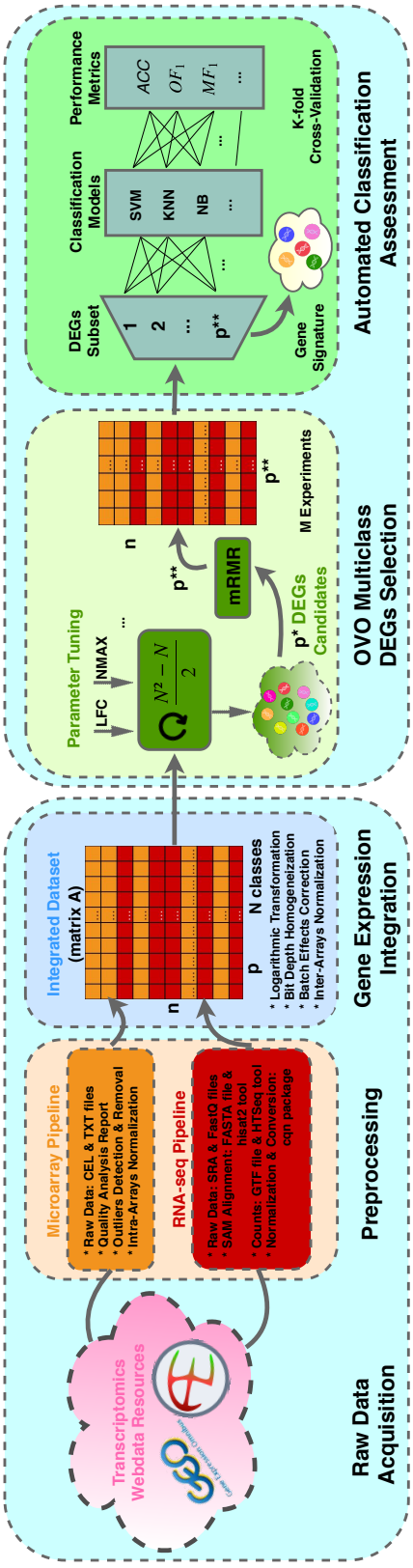


Figure 4.1: Overall flowchart of the designed gene expression analysis pipeline. Two main bioinformatic tasks are addressed based on gene expression analysis: transcriptomics technologies integration and machine learning techniques application.

In the case of RNA-seq series, Sequence Read Archive (SRA) and FASTQ files containing raw information were directly downloaded in a programmatic manner before being preprocessed. Only those series whose samples were aligned to the GRCh37 reference genome, were considered for this study due to its greater public availability. Specifically, the extensive RNA sample collection from 27 series used in this work led to the analysis of up to 10 SPSs (Table 4.1). Each of the series can be identified under accession ID, highlighting most of them being submitted from United States and other countries where their population is predominantly white: Deutschland, Netherlands, Great Britain and Australia (Table 4.2).

4.4.1.2. Preprocessing

This phase begins by checking the quality of the samples under a restrictive evaluation procedure. To achieve it, the arrayQualityMetrics R package was iteratively applied on every microarray series by assessing up to 6 quality tests [116]: distance among samples, principal Principal Component Analysis (PCA), Kolmogorov-Smirnov test based on the Ka parameter, density distribution plots, standard deviation of the samples intensities and Hoeffding's D-statistic (normally executed with $D < 0.15$). In order to discard all samples presenting low quality (outliers), all of these tests were iteratively applied over each series. With respect to RNA-seq series, 5 samples were excluded by avoiding sample duplication. The total number of excluded samples from each series is specified in the last column of Table 4.2. Subsequently, each of the sequencing technologies requires a wide range of intra-array processing steps which have to be carefully performed when both are going to be integrated at gene expression level. Because of being processed from different platforms, a normalisation procedure has to be applied on each microarray series. RMA algorithm [25] was applied in this work by modularly performing background correction, normalisation and summarisation on the microarray data. For its application, rma function from

Table 4.1: Taxonomic classification of skin pathological states for the 968 collected RNA samples.

Super-state	**SPS**	**Microarray**	**RNA-seq**	**Integrated**
Healthy state	NSK	151	84	235
	NEV	30	27	57
Non-melanoma skin cancer (NMSC)	BCC	43	0	43
	ISCC	69	14	83
	PMCC	26	0	26
	MMCC	23	0	23
Melanoma skin cancer (MSC)	PRIMEL	69	51	120
	METMEL	39	0	39
Precancerous skin disease	AK	29	23	52
	PS	187	103	290
Total		666	302	968

SPS = Skin pathological state, NSK = Normal skin, NEV = Nevus, BCC = Basal cell carcinoma, ISCC = Invasive squamous cell carcinoma, PMCC = Primary Merkel cell carcinoma, MMCC = Metastatic Merkel cell carcinoma, PRIMEL = Primary melanoma, METMEL = Metastatic melanoma, AK = Actinic keratosis, PS = Psoriasis.

Table 4.2: Series information selected for this study from NCBI GEO and ArrayExpress web platforms.

Technology	Manufacturer	Series	Samples origin	Skin pathological states (Selected samples)	High quality samples	Excluded outliers
Microarray	Affymetrix	GSE2503	Berlin (DEU)	ISCC (5), NSK (4), AK (3)	12	2
		GSE3189	San Diego (USA)	NEV (16), NSK (6)	22	3
		GSE6710	Berlin (DEU)	PS (13)	13	0
		GSE7553	Tampa (USA)	BCC (15), PRIMEL (14), ISCC (11), NSK (4)	44	2
		GSE13355	Ann Arbor (USA)	NSK (61), PS (56)	117	5
		GSE14905	Gaithersburg (USA)	PS (31), NSK (16)	47	7
		GSE15605	Nashville (USA)	PRIMEL (30), NSK (13), METMEL (2)	45	18
		GSE30999	Spring House (USA)	PS (73)	73	12
		GSE32407	New York (USA)	NSK (10)	10	0
		GSE32924	New York (USA)	NSK (7)	7	1
		GSE36150	Royal Oak (USA)	PMCC (5), MMCC (5)	10	5
		GSE39612	Ann Arbor (USA)	PMCC (12), MMCC (5), ISCC (2), BCC (2)	21	15
		GSE42109	New York (USA)	BCC (10)	10	1
		GSE42677	New York (USA)	NSK (9), ISCC (5), AK (5)	19	1
		GSE45216	London (GBR)	ISCC (27), AK (8)	35	5
		GSE46517	Houston (USA)	METMEL (31), PRIMEL (25), NSK (6), NEV (6)	68	20
		GSE50451	Bethesda (USA)	MMCC (13), PMCC (9)	22	1
		GSE52471	New York (USA)	PS (14), NSK (10)	24	7
		GSE53223	New York (USA)	NEV (8), NSK (5)	13	5
		GSE82105	New York (USA)	METMEL (6)	6	0
	Illumina	GSE32628	Leiden (NLD)	ISCC (14), AK (13)	27	2
		GSE53462	Suwon (PRK)	BCC (16), ISCC (5)	21	5
RNA-seq		GSE54456	Ann Arbor (USA)	PS (89), NSK (80)	169	0
		GSE67785	Ann Arbor (USA)	PS (14)	14	0
		GSE84293	Houston (USA)	AK (10), ISCC (9)	19	0
		GSE98394	New York (USA)	PRIMEL (51), NEV (27)	78	0
		E-MTAB-5678	Brisbane (AUS)	AK (13), ISCC (5), NSK (4)	22	0
Integrated					**968**	122

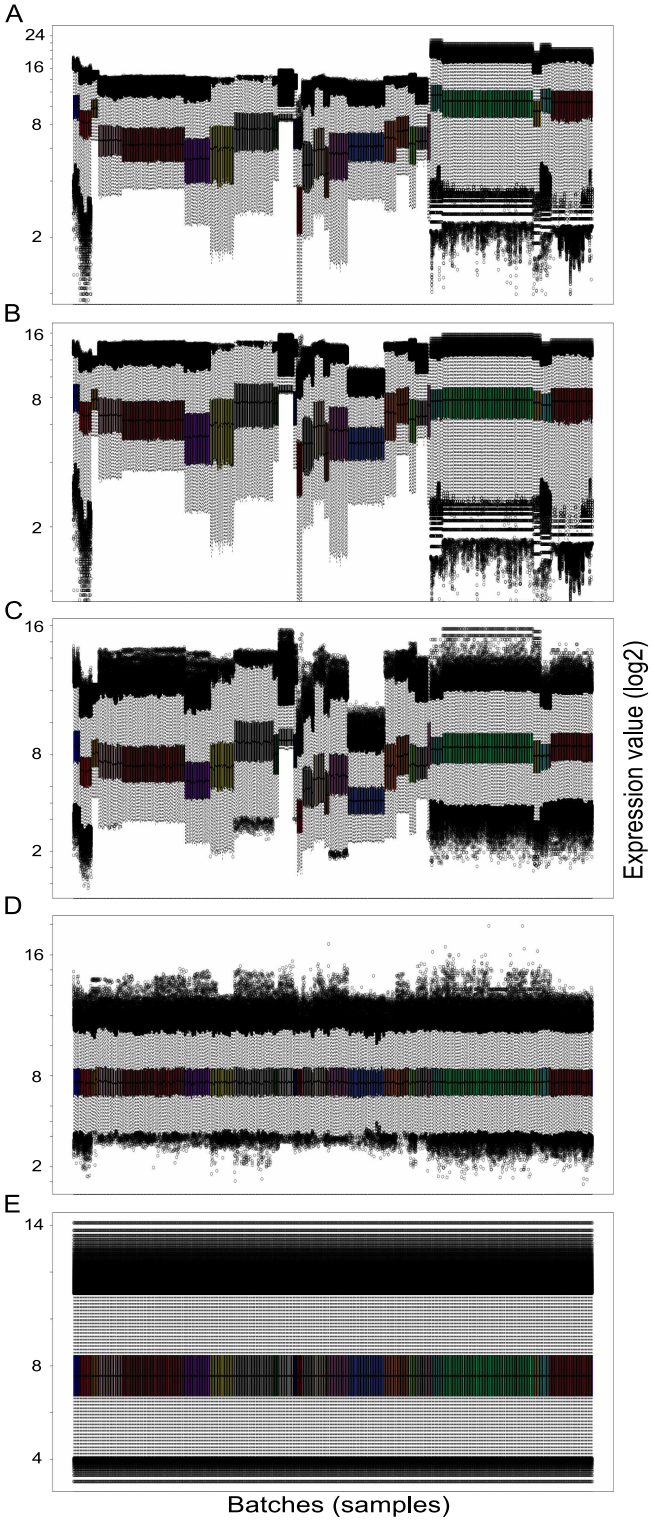
Samples purity for each skin pathological state was critically required and inspected. Manufacturer, technology and total number of samples/outliers are included.

affy and oligo R packages was used for Affymetrix microarrays as well as lumiExpresso from lumi R package was used for Illumina microarrays. Gene annotation of each series was provided by the annotate R package, which eases the mapping from the manufacturer chip identifiers to standardised symbols by using a wide range of annotation packages from Bioconductor website. With respect to RNA-seq series processing, the proposed pipeline by Anders et al. [155] was partially followed but changing certain tools. Once a large number of FASTQ and SRA files are available, several tools such as sra-toolkit [156], hisat2 [157], bowtie2 [158], samtools [159] and htseq [4] were used until getting read count files containing the located genes in each sample. Before obtaining these files, gene annotation was retrieved by means of biomaRt R package [160], a data-mining tool which allows to connect with Ensembl database [161]. After all these steps, other R packages such as cqn [5] helped in correcting and normalising GC content bias, and NOISeq [6] allowed to calculate the gene expression values.

4.4.1.3. Gene Expression Integration

After preprocessing each of the microarray and RNA-seq series individually, additional requirements have to be considered before inter-array normalising and correctly integrating them all [162]. On the one hand, each of the expression values of the genes transcribing the same gene identifier have to be summarised in a single value. In order to be consistent in assessing the impact of each gene selected, all transcripts were gathered by applying the mean of them on each series separately. On the other hand, several simultaneous steps were carried out on the 27 series (Figure 4.2). 28 batches were established because different samples from GSE42677 series were processed by two different platforms. Firstly, logarithmic transformation was performed on 2 series (GSE2503 and GSE3189) in order to adequate the scale representing gene expression values, establishing base 2 for all the batches (Figure 4.2A). Following, 16-bit depth homogenisation

Figure 4.2: Series processing procedure for gene expression integration: (A) logarithmic transformation, (B) 16-bit depth homogenisation, (C) complete cases selection along the batches, (D) batch effect correction with ComBat and (E) inter-array normalisation with normalizeBetweenArrays.



was applied (Figure 4.2B) after previously analysing the maximum value of gene expression for each series in function of the platform, establishing different consensus values in the bit depth: 20-bit depth for Human Genome U133A Array platform (GSE2503, GSE3189, GSE6710 and GSE46517), 16-bit depth for Human Genome U133 Plus 2.0 Array platform (GSE7553, GSE13355, GSE14905, GSE15605, GSE30999, GSE32924, GSE39612, GSE42677, GSE45216, GSE50451, GSE53223 and GSE82105), 16-bit depth for Human Genome U133A 2.0 Array platform (GSE32407, GSE42109, GSE42677 and GSE52471), 12-bit depth for Human Exon 1.0 ST Array platform (GSE36150), 16-bit depth for HumanAll platform (GSE32628 and GSE53462), 22-bit depth for Genome Analyzer platform (GSE54456), 20-bit depth for Genome Analyzer IIx platform (GSE67785), 24-bit depth for HiSeq 2000 platform (GSE84293 and E-MTAB-5678) and 22-bit depth for HiSeq 2500 platform (GSE98394). Thereupon, by having previously established a common gene annotation for all the considered series, only common genes were identified and selected for all the samples coming from the series / batches. At this point, batch effect correction was thought to be applied because hypothetical batch effects could be appearing among all 28 batches considered (Figure 4.2C). By dealing with this issue, ComBat method [34] from sva R package [163] was considered, correcting and establishing a harmonised sample distribution along all the samples from all batches (Figure 4.2D). Finally, an inter-array normalisation was applied by means of normalizeBetweenArrays function from limma R package [38]. This achieves consistency among all the samples put together and forces an identical empirical distribution on each of them based on quantile normalisation (Figure 4.2E). Before any new sample is properly assessed by this procedure, all these transformations are completely necessary and have to be applied in the same way. At the end of this procedure, the whole integrated dataset formed by p common genes and all n quality samples selected including N classes is achieved (matrix A in Figure 4.1).

4.4.2. Machine Learning and Soft Computing

Bioinformatics researches have been successfully benefited from the use of machine learning and soft computing techniques [164] in a wide range of problems such as expression profiling identification, feature selection and classification [59]. As the number of biological experiments and applications using high-throughput technologies continue to increase, new approaches using this type of techniques for knowledge discovery have to be proposed [165, 166].

4.4.2.1. OVO Multiclass DEGs Selection

Traditionally, the gene selection from expression profiles analysis deals with the curse of dimensionality problem (np-hard) because of pitting few n samples against thousands of p genes [62]. By reducing such dimensionality to highly discriminatory DEGs, this issue becomes even more challenging when increasing the number of SPSs (in our work, N) (see nomenclature in Figure 4.1). With the purpose of handling such challenge, this work presents a simple and intuitive one-vs-one (OVO) multiclass DEGs selection approach based on the assessment of all possible pair comparisons of SPSs. This concept of comparing two SPSs has been defined in this work as class pair comparison (CPC). Each CPC is analysed under the criterion of selecting those DEGs with higher LFC by having a higher discernment power at the gene expression level. For this purpose, this process was carried out by means of tuning the two parameters LFC and NMAX. On the one hand, LFC establishes a minimum threshold value to be genes considered as DEGs throughout all CPCs. On the other hand, NMAX indicates the maximum number of DEGs selected for each CPC. An additional threshold can be established by means of p-value (PV), but a constant value of 0.001 was established to present our approach. By extending to a problem of N SPSs, the total number of CPCs amounts to $(N2 - N)/2$. This is particularised in 45 CPCs in this work in which 10 SPSs are simultaneously analysed. The expected maximum number of

DEGs would attain NMAX * $(N2 - N)/2$ value after applying this methodology. This step forward with respect to the classical gene selection process, which is exclusively controlled by PV and LFC, may eventually avoid the lack of capacity of the selected DEGs discerning among specific CPCs or different SPSs subsets by easily tuning NMAX parameter. In order to finish this process, the union of all the DEGs sets after considering each CPC has to be performed. This consideration allows identifying repeated DEGs because of having higher difference of gene expression for several CPCs. Such DEGs coincidence helps in reducing even further up to p* the final candidate multiclass DEGs, where p* $\leq$ NMAX*$(N2 - N)/2 \ll$ p. In search of strengthening the selection of DEGs as much as possible, up to M different experiments were performed, splitting the whole integrated dataset into two datasets: 90% for training and validation and the remaining 10% for testing. Similar representativeness of each SPS was ensured within both datasets. The feature selection and parameter tuning processes were initially applied on the 90% similarly to a cross-validation procedure for each of these M experiments, thus returning different DEGs sets for each LFC and NMAX combination. With the aim of improving the reliability and the interpretability of the subsequent results, only those p* common genes matching all the M experiments for each parameter combination were selected. This fact discards spurious DEGs only emerging in specific experiments and preserves from subsequent classification biases. Before evaluating the different p* common genes sets within each of the M experiments, an additional assessment of their informative power was performed by means of minimum-Redundancy Maximum-Relevance (mRMR) feature selection algorithm [59]. This algorithm returns a ranking according to the criterion of placing those DEGs with the most relevant and the lowest redundant information among themselves with respect to the class variable. After applying it, different DEGs rankings were established by assessing the different p* candidates sets on the whole integrated dataset for each

LFC and NMAX combination. To sum up, twofold DEGs selections were carried out: firstly, reducing the computational complexity from p thousands of genes to the p* most reliable candidate DEGs of the disease; secondly, after considering the previous reason, exclusively selecting those p** DEGs with higher informative capability for the intelligent diagnosis (see right part in Figure 4.1).

4.4.2.2. Automated Classification Assessment

Three classification techniques assessed the informative power of different DEGs subsets from the ranking returned by mRMR: Support Vector Machines (SVM) [60], K-Nearest Neighbour (KNN) [61] and Naive Bayes (NB) [62]. K-fold cross validation technique (K-fold CV, where K = 10) [66], the most considered accurate approach for model selection, was used on the training set of each M experiment with the purpose of providing a realistic performance of the DEGs on new unseen data. Once again, samples from each SPS were equally distributed among K-folds in search of improving the possibilities of correctly classifying any new sample. The 10-fold CV classification assessment was repeated 10 times by randomly shuffling the dataset, thus achieving statistical robustness by procuring asymptotic convergence to a reliable estimation of the classifier performance [167]. Finally, three metrics were used in order to measure the recognition rate by combining each classifier in association with different DEGs set sizes: accuracy (ACC), overall F_1-score (OF_1) and mean multiclass F_1-score (MF_1).

These are calculated by using the Equation 1, 2 and 3, respectively. Each of these metrics can be expressed in function of certain parameters (precision (P) and recall (R)) or different rates (T_p, T_n, F_p and F_n) which can be identified from a confusion matrix of N classes:

$$ACC = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{4.1}$$

$$OF_1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot T_p}{2 \cdot T_p + F_p + F_n} \tag{4.2}$$

$$MF_1 = \frac{\sum_{i=1}^{N} F_1^{classs}(i)}{N} \tag{4.3}$$

The metrics related to F1-score [168] were considered particularly suited and robust for the multiclass study tackled, as they provide a better measurement of the recognition rate of each of the classes under unbalanced data. After our model has been validated by the K-fold CV, these metrics were also calculated for the remaining 10% testing dataset for each experiment. Following, in order to assess the influence of the multiple factors considered for identifying multiclass DEGs, an ANOVA statistical test was performed over the entire dataset. Although factors such as assessed dataset type (TYPE), analysed K-fold cross validation (KFOLD) or M experiment performed (EXPERIMENT) were also evaluated by this test, 4 factors were specifically highlighted because of their further relevance in the subsequent analysis. On the one hand, LFC and NMAX parameters were subjected to evaluation by tuning the proposed algorithm. On the other hand, the hypothetical differences of applying different classifiers in combination with a number of DEGs set sizes (GenMax) were also inspected by means of this test. By checking the validity of each factor (LFC, NMAX, classifier and GenMax), the different performance metrics (ACC, OF1 and MF_1) were measured for both training and test sets. Finally, a functional enrichment analysis was performed by means of DAVID 6.8 [169] in order to functionally annotate and interrelate the obtained DEGs using Gene Ontology (GO) terms.

4.5. Results and Discussion

By taking into account the integration at gene expression level from 22 microarrays and 5 RNA-seq series containing multiple SPSs related to cancer, the opportunity to determine a skin cancer gene signature of up to N = 10 SPSs, formed by highly reliable multiclass DEGs, has been addressed in this work. The experimental analysis of this study have been conducted under the proposal of an OVO multiclass DEGs selection algorithm which has been thoroughly tested by means of an ANOVA statistical test. The interpretation of the results obtained from this analysis have been used in order to select suitable setting parameters. By tuning our proposed algorithm, this study was focused on assessing the informative power of the p* identified multiclass DEGs. After selecting p** multiclass DEGs from the previous one, their biological relationship to skin cancer was finally consulted. This discussion has been guided on presenting all the results derived from the procedure above.

4.5.1. Impact of Tuning Algorithm Parameters

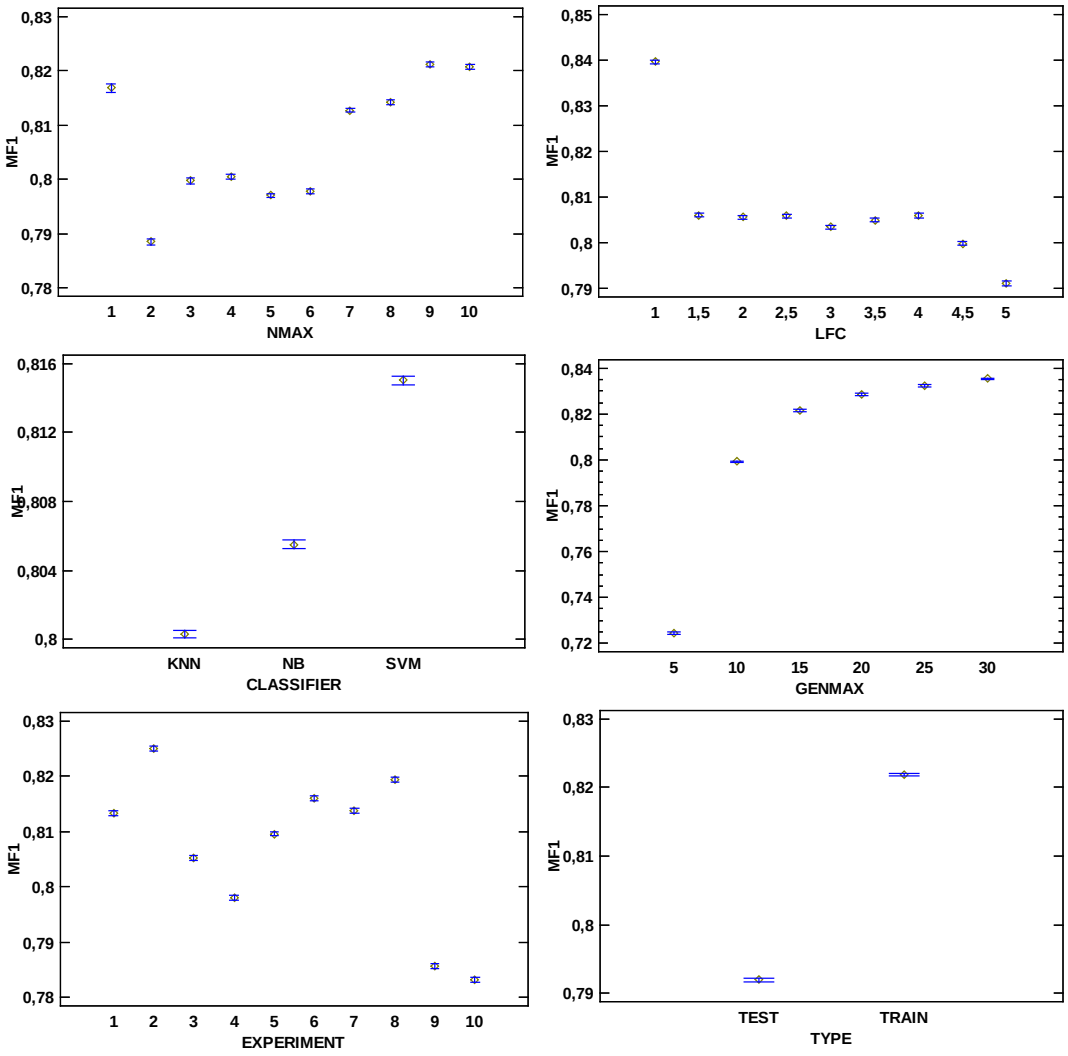
The statistical significance of each considered and highlighted factor (NMAX, LFC, GenMax, Classifier) was confirmed by means of the ANOVA statistical test, showing the influence of each of them on the classification performance (Table 4.3). Type III sums of squares was chosen and the contribution of each factor was measured having removed the effects of all other factors. P-values tested the statistical significance of each of the factors. Since 6 P-values are less than 0,05, these factors have a statistically significant effect on MF_1 at the 95,0% confidence level (highlighted in bold). All F-ratios are based on the residual mean square error. The most significant differences were exclusively appreciated by checking the scale depth when using MF_1 (Figure 4.3). While the lowest NMAX parameter value reflected one of the highest classification performances

Table 4.3: ANOVA statistical test for MF_1 performance metric

Source (Main Effects)	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A: TYPE	61.4457	1	61.4457	22681.85	**0.0000**
B: EXPERIMENT	48.0765	9	5.34183	1971.87	**0.0000**
C: KFOLD	0.00597	9	0.00066	0.24	0.9878
D: CLASSIFIER	10.1689	2	5.08447	1876.86	**0.0000**
E: LFC	42.0707	8	5.25884	1941.23	**0.0000**
F: NMAX	30.6972	9	3.4108	1259.05	**0.0000**
G: GENMAX	419897	5	83.9794	30999.84	**0.0000**
RESIDUAL	739442	272955	0.00270		
TOTAL (CORRECTED)	1398.59	272998			

and discarded the consideration of a wide range of DEGs for each of the 45 CPCs, the impact of tuning LFC helped to elucidate the disadvantage of selecting high threshold values because of dropping value more than 3%. Classification models results ranged from 80% to 82%, establishing these performances around 10 genes (see Classifier and GenMax factors in Figure 4.3). Similar statistical results and distribution for each factor were achieved for ACC and OF1, and these can be facilitated under petition. Following, in order to present and illustrate the utility of the proposed algorithm in this work, a choice of parameters was required. The decision was motivated under the criterion of restrictively selecting DEGs while being preserved the information of all considered SPSs for this study. For this purpose, NMAX = 1 was established by presenting one of the highest recognition rate for each performance metric assessed (as clearly showed and supported the results of ANOVA statistical test), leading to drastically reduce the computational complexity to a maximum of $(N^2 - N)/2 = 45$ highly discerning DEGs. This fact prevents of arbitrarily tuning LFC and relying decision power on it in search of an enough threshold for discerning among multiple SPSs. Extensively, this decision may avoid the removal of DEGs to discern those hardly distinguishable when applying highly restrictive LFC values. Hereafter, these setting parameters were used to identify the candidate multiclass DEGs and present a potential gene signature of the skin cancer.

Figure 4.3: ANOVA statistical test results for MF_1 in function of different factors: Type, Experiment, NMAX, LFC, Classifier and GenMax. All these factors were determined as significant statistically.



4.5.2. Selection of Informative DEGs

Although up to 45 genes could have potentially been returned by our proposed algorithm under the selected configuration, exclusively p* = 10 candidate multiclass DEGs appeared as common genes from the intersection of DEGs

for each of the M = 10 experiments performed, as many of these genes were highly discriminating among several CPCs. However, in order to reduce the repertoire of candidate DEGs set for intelligent diagnosis, the informative capability of different subgroups of up to p* DEGs ordered by means of mRMR, was subjected to an automated classification assessment. This algorithm then established the following DEGs ranking: MLANA, LTF, MMP1, ADAMTS3, LY6D, SCGB2A2, KRT14, PI3, PMEL and S100A7. As a result, the classification results are presented when increasing the size of DEGs set following the previously established ranking, showing asymptotic convergence for the different performance assessments (Figure 4.4). Our classification procedure even demonstrates how the recognition rate for unseen data does not drastically drop, reinforcing the overall reliability of these DEGs for skin cancer diagnosis. By reducing the complexity of the study, the subsequent experimental analysis was limited to consider the first p** = 8 DEGs given that the average improvement of MF_1 per gene is lower than 0.6%. The results associated with this size of DEGs set even improved those showed by GenMax parameter for ANOVA test, outperforming recognition rates of 94% OF1 and 80% MF_1 when considering any classifier. Afterwards, with the purpose of knowing the overall discernment capabilities of the 8 multiclass candidate DEGs, the number of SPSs and CPC cases being covered by each one of them when being appeared with the highest |LFC| for any CPC was summarised (Table 4.4).

4.5.3. Recognition of SPSs

Despite establishing setting parameters which help in emerging DEGs to discern from each CPC, this fact does not prevent from having difficulties in distinguishing among certain SPSs. Most CPCs can be properly discerned from any of the 8 DEGs by presenting significant LFC values (Figure 4.5). However, there is a small set of CPCs which are harder to distinguish when examining

changes at gene expression level such as ISCC vs AK (LFC < 2) or PMCC vs MMCC (LFC < 1). This incurs in observing the informative limits of gene expression when intending to offer a reliable diagnosis among a lot of SPSs which are close at the biological level.

Figure 4.4: Evolution of the recognition rate for training and test datasets. Three classification models (SVM, KNN and NB) were assessed by means of several performance metrics (ACC, OF_1 and MF_1) when considering different subgroups of DEGs ranked by mRMR algorithm.

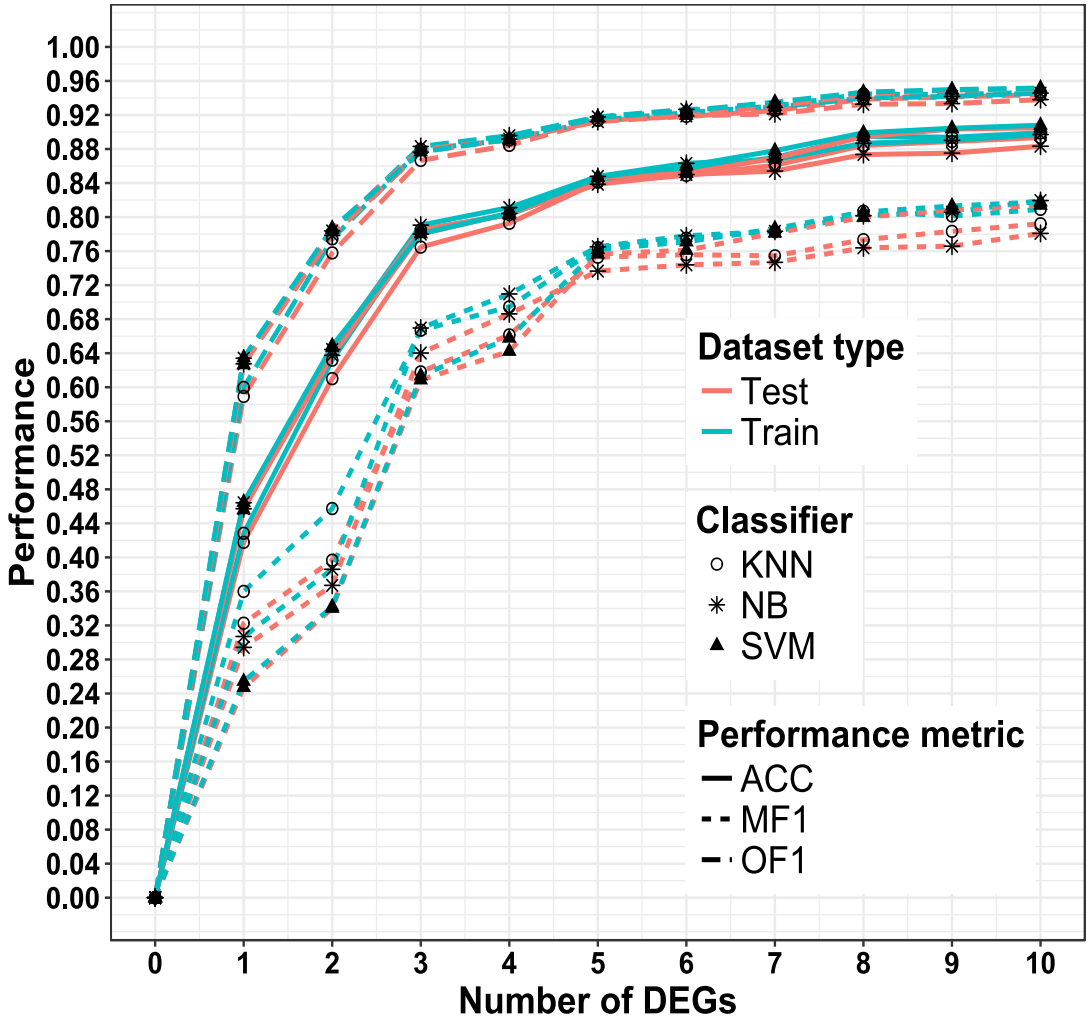


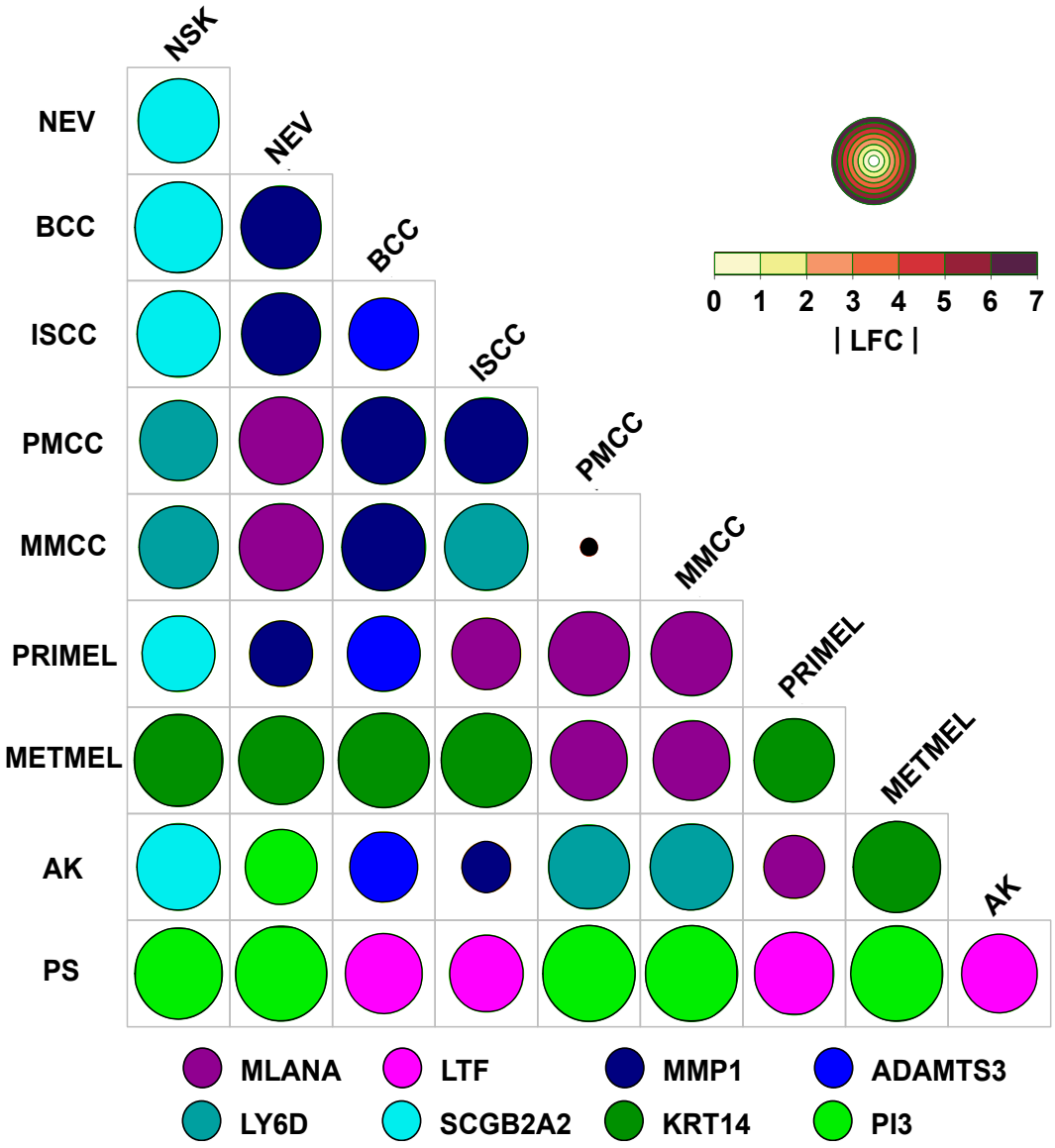
Table 4.4: ANOVA statistical test for MF_1 performance metric

Gene Symbol	**SPSs**	**CPCs (%)**	$\mu_{\mathbf{LFC}} \pm \sigma_{\mathbf{LFC}}$	$[\mathbf{PV_{MIN}}, \mathbf{PV_{MAX}}]$
MLANA	7	8 (17.8)	4.86 ± 0.97	[1.30E-157, 8.52E-86]
LTF	5	4 (8.9)	4.81 ± 0.29	[3.87E-186, 2.70E-94]
MMP1	7	7 (15.6)	4.67 ± 1.48	[1.34E-77, 8.30E-14]
ADAMTS3	4	3 (6.7)	4.02 ± 0.27	[3.91E-230, 1.67E-160]
LY6D	5	5 (11.1)	5.35 ± 0.33	[4.55E-135, 5.71E-121]
SCGB2A2	6	5 (11.1)	5.48 ± 0.69	[7.58E-121, 4.41E-89]
KRT14	7	6 (13.3)	6.24 ± 0.56	[2.42E-264, 1.96E-189]
PI3	7	6 (13.3)	6.39 ± 1.07	[9.99E-220, 7.23E-73]

By extensively checking how a new unseen sample could be classified, the different classification models assessed the 8 highlighted DEGs set (Figure 4.6). The recognition rates confirm the real challenge of properly discerning the CPC cases previously highlighted, although presenting accuracy differences among models when classifying certain SPSs (for example, ISCC achieves 72% for NB, 76% for KNN and 77% for SVM). On the one hand, 3 SPSs achieved high recognition rates for SVM classification model: NSK (97%), BCC (100%) and PS (98%). On the other hand, recognition rates dropped for the 7 remaining SPSs mainly to be confused with another SPS as predecessor studies had already advanced [9–12]: NEV (83% and confused with NSK above 4%), ISCC (77% and confused with AK above 20%), PMCC (58% and confused with MMCC above 37%), MMCC (45% and confused with PMCC above 54%), PRIMEL (91% and confused with METMEL above 2%), METMEL (90% and confused with PRIMEL above 7%) and AK (65% and confused with ISCC above 31%). This fact remarks the difficulty of achieving reliable DEGs between precancerous and invasive states because they present molecular similarities. By considering the fusion of certain CPCs (for example, MCC formed by PMCC and MMCC, MSC

formed by PRIMEL and METMEL or combining ISCC and AK), the recognition rates would have practically outperformed percentages ranged from 87% to 99% for these skin super-states in a much more generalised study.

Figure 4.5: Distribution map of the 8 multiclass DEGs set. Highest |LFC| value for each CPC by considering NMAX = 1 and applying mRMR algorithm. Circle sise and color are correlated with |LFC| value and multiclass DEG with highest |LFC|, respectively. CPC, class pair comparison.



4.5.4. Determination of Potential Target Genes

One of the main reasons to separate in specific SPSs lies in finding relevant biomarkers of their occurrence from gene expression analysis. The determination of potential target genes could help clinicians when making their diagnoses, eventually avoiding the application of inappropriate therapies to combat certain SPSs. For example, the search of therapeutic alternatives for the treatment of MMCC [170] has been necessary due to the ineffectiveness of chemotherapy by failing to ensure successful outcomes when applying on long-term MMCC [171]. This fact has driven the pursuit of personalised therapies to deal with diverse MCC stages such as PMCC or MMCC [172]. Before making any medical decision on any new skin sample, clinicians could rely on intelligent diagnosis based on assessing reliable multiclass DEGs. In this case, our approach highlighted the informative capacity of these 8 candidate multiclass DEGs for an overall diagnosis of suffering from skin cancer (Figure 4.7).

In view of these results, certain multiclass DEGs such as MLANA, MMP1, LY6D or PI3 appeared down-expressed for both SPSs and, among others, may discern better PMCC and MMCC with respect to other SPSs (Figure 4.5). All these genes have previously proven to be of great importance for expression pattern characterisation and skin cancer diagnosis: from inhibition in SCC (MLANA), positive dysregulation in BCC and AK (MMP1) to correlated overexpression in SCC and PS (PI3) [173–175]. Therefore, a preventive clinical analysis of these genes could help to avoid erroneous therapies by examining their hypothetical involvement in other SPSs addressed by this study.

4.5.5. Biological Interpretation of the Multiclass DEGs

In order to understand the functional properties of the 8 highlighted DEGs, a functional enrichment analysis based on GO terms was performed from DAVID

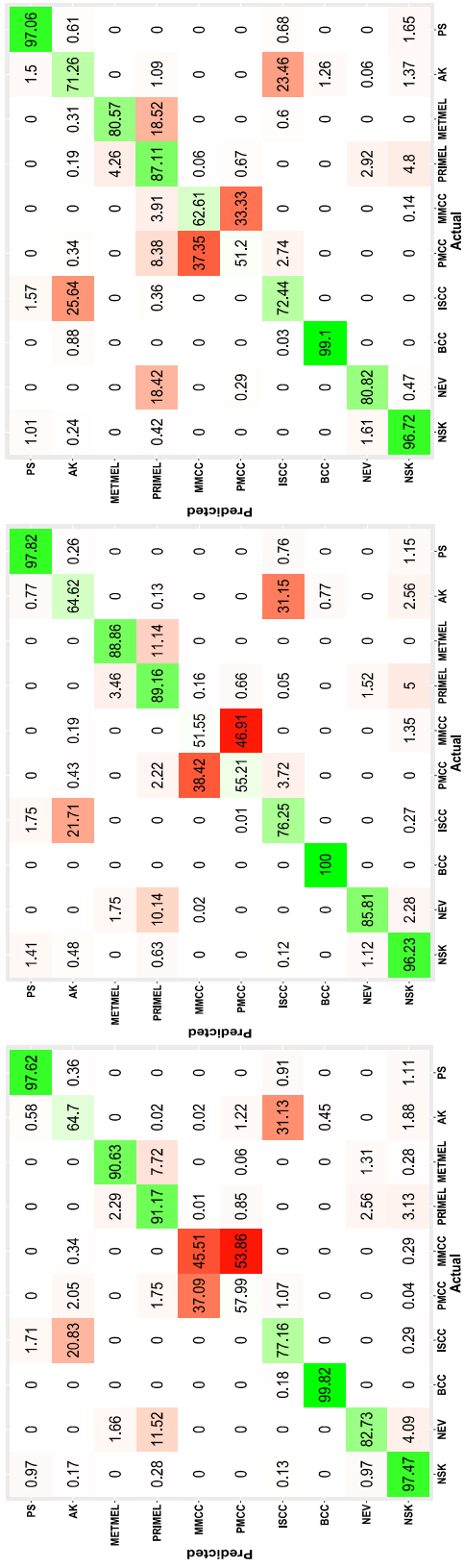


Figure 4.6: Different classification models were assessed: (A) SVM, (B) KNN and (C) NB. For each designed model, the 8 highlighted multiclass DEGs set were selected and assessed by 10-fold CV for discerning 10 SPSs. SVM = Support Vector Machines, KNN = K-Nearest Neighbors, NB = Naive Bayes, CV = Cross-Validation, SPS = Skin pathological state.

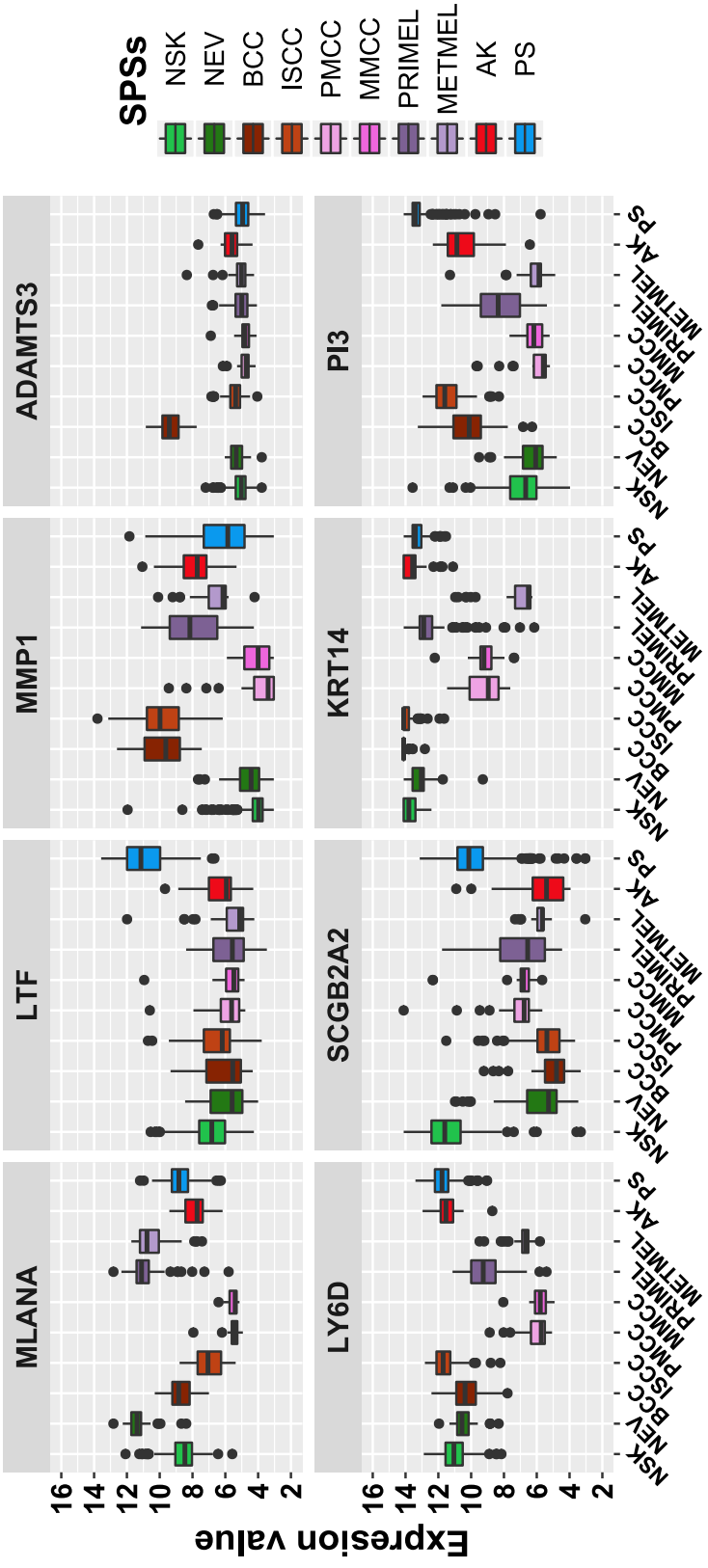


Figure 4.7: Expression level of the 8 multiclass DEGs set. Highlighted DEGs by our approach are ordered from left to right and from top to bottom by the ranking returned by mRMR. SPS = Skin pathological state.

Bionformatics Database [169]. The three GO ontologies for biological processes (BP), cellular components (CC) and molecular functions (MF) were considered for our analysis. A total of 6 BPs, 6 CCs and 4 MFs were determined to be significant throughout these genes (Table 4.5). As shown, MMP1, ADAMTS3 and LTF genes are highly related in terms of their proteolysis process and endo- and metalloendo-peptidase activity. According to the activity of proteolytic enzymes, this fact has been associated with angiogenesis and tumor progression of skin cancer [176, 177]. Following, by exhaustively inspecting specific literature, the biological relationship of the 8 highlighted DEGs with skin cancer was consulted. On the one hand, the most remarkable inquiries underlined the dysregulation of up to 6 DEGs in MSC cases [178, 179] and development risk [180] (excepting ADAMTS3 and PI3) and the implication of up to 5 DEGs in PS development or inflammatory processes [175, 181] (excepting MLANA, ADAMTS3 and KRT14). On the other hand, the differentiating role of specific DEGs in NMSC cases was highlighted: the overexpression of ADAMTS3 in BCC [182] or the hypothetical implication of KRT14 in the malignant transformation of potential stem cells as origin of MCC [183]. Based on all these precedent evidences and the results showed (Figure 4.7), the 8 multiclass DEGs highlighted by this approach should be particularly taken into account by being related to tumorigenesis and pathogenesis of skin cancer. Concretely, MLANA has been remarkably demonstrated to be upregulated in NEV [178], inhibited in SCC [173] and differentiated between MCC and PRIMEL by highlighting absence and overexpression by means of immunohistochemical analysis [184]. Further, multiple genetic dysregulations of DEGs have been reported in several studies: from the downregulation of LY6D, SCGB2A2 and KRT14 in METMEL with respect to PRIMEL [178] to the dysregulation in SCC versus NSK by showing inhibition of MLANA and SCGB2A2 or overexpression of MMP1 and PI3 [173, 185, 186].

Table 4.5: Functional Enrichment Analysis for the 8 multiclass DEGs using GO terms.

Ontology	GO ID	GO Term	# Genes (%)	PV	Gene Symbol
BP	GO:0006508	Proteolysis	4 (50.0)	8.75E-3	LTF, MMP1, ADAMTS3, PI3
	GO:0030574	Collagen catabolic process	2 (25.0)	1.92E-2	ADAMTS3, MMP1
	GO:0032963	Collagen metabolic process	2 (25.0)	3.23E-2	ADAMTS3, MMP1
	GO:0044236	Multicellular organism metabolic process	2 (25.0)	3.90E-2	ADAMTS3, MMP1
	GO:0044243	Multicellular organism catabolic process	2 (25.0)	2.13E-2	ADAMTS3, MMP1
	GO:0044259	Multicellular organismal macromolecule metabolic process	2 (25.0)	3.38E-2	ADAMTS3, MMP1
CC	GO:0005576	Extracellular region part	5 (62.5)	3.95E-2	LTF, MMP1, ADAMTS3, KRT14, PI3
	GO:0005578	Proteinaceous extracellular matrix	3 (37.5)	5.57E-3	MMP1, ADAMTS3, PI3
	GO:0031012	Extracellular matrix	3 (37.5)	1.17E-2	MMP1, ADAMTS3, PI3
	GO:0031982	Vesicle	5 (62.5)	1.89E-2	MLANA, LTF, ADAMTS3, KRT14, PI3
	GO:0031988	Membrane-bounded vesicle	5 (62.5)	1.64E-2	MLANA, LTF, ADAMTS3, KRT14, PI3
	GO:0044421	Extracellular region part	5 (62.5)	2.12E-2	LTF, MMP1, ADAMTS3, KRT14, PI3
MF	GO:0004175	Endopeptidase activity	3 (37.5)	1.19E-2	LTF, MMP1, ADAMTS3
	GO:0004222	Metalloendopeptidase activity	2 (25.0)	3.95E-2	MMP1, ADAMTS3
	GO:0008233	Peptidase activity, acting on L-amino acid peptides	3 (37.5)	2.51E-2	LTF, MMP1, ADAMTS3
	GO:0070011	Peptidase activity, acting on L-amino acid peptides	3 (37.5)	2.35E-2	LTF, MMP1, ADAMTS3

Fisher's exact statistical test was performed to determine their significance (PV < 5E-2). GO = Gene Ontology, PV = P-Value, BP = biological process, CC = cellular component, MF = molecular function.

Finally, the dysregulation of certain DEGs has been interestingly reflected in both SCC and PS in a similar way: from inhibition of SCGB2A2 together with overexpression of MMP1 and PI3 [175] to slight and strong upregulation of LTF in SCC and PS, respectively [175, 181]. Because of being a chronic inflammatory skin disease, special attention should be paid to the psoriasis evolution because the cancer development also generates inflammatory reactions around surrounding tissue [23]. From the preventive point of view, clinicians should remain attentive to the high gene expression variability of these specific DEGs by showing changes between NSK, PS and diverse SPSs related to cancer (see gene expression changes for all these DEGs in Figure 4.7). In accordance with our results, this subset of multiclass DEGs could represent a genetic signature offering clues about the overall state of the disease.

4.6. Conclusions

Throughout this study, the validity of integrating transcriptomic data from the main technologies for quantifying gene expression has been underlined. Specifically, an even more generalised study on skin cancer has been approached, extending the methodological approach presented in Chapter 3. Some new insights on biomarkers that might be offering clues on skin tumor degeneration have been shown. Despite specific skin pathological states are hardly distinguishable due to high intrinsic biological similarities, obtaining an intelligent skin cancer diagnosis for 10 pathological states with an overall classification rate higher than 94% with only 8 genes is quite promising. This result took on more value when inspecting the biological involvement of these 8 biomarkers on skin cancer. New clinical evaluations will determine the diagnostic potential of these biomarkers, thus encouraging to develop innovative target therapies for combating the skin cancer.

5. Considering Somatic CNVs to Improve Intelligent Diagnosis

5.1. Introduction

This chapter is intended to motivate the potential of genomic information to help narrow further the candidate gene set that may show greater biological involvement and hypothetical responsibility promoting the development of the analysed cancer. Specifically, taking advantage of the knowledge acquired from the studies presented in Chapters 3 and 4, gene expression has been integrated with somatic CNVs, even analysing the informative correlation between both biological information sources in order to select those biomarkers with greater informative power in an intelligent diagnosis.

The study has been made from the use of the following resources:

- **Number of datasets:** 18
- **Number of samples (patients):** 605 (532)
- **Skin Pathological States:** 4 (NSK, NEV, PRIMEL and METMEL)
- **Web Data Repositories:** NCBI GEO, AE and NCI GDC
- **Sequencing Technology:** Microarrays, RNA-seq and CNVs

All these resources were defined in Chapter 2 and can be consulted there. The content included in this chapter is a part of the submitted journal article entitled *"Supporting Clinical Decisions - Determining Biomarkers Driven by Somatic Copy Number Variations Being Responsible for the Progression of Cutaneous Melanoma"* (under review).

5.2. Background

Cutaneous melanoma is unquestionably the deadliest form of all skin cancers. Despite historically being a rare cancer, its incidence in recent decades has increased faster than any other cancer [187]. This growing projection represents a significant health burden in worldwide [134]], being remarkably alarming in United States where it is estimated >96000 new cases and >7000 deaths during 2019 [73]. Nowadays, the personalised treatment of this disease considerably depends on the diagnosed clinical stage which is usually based on the AJCC guidelines [143]. However, the melanoma heterogeneity prevents of establishing long-lived therapeutic solutions [188]. On the one hand, surgical excision is clearly considered the main recommendation for treating the primary cutaneous melanoma. On the other hand, multiple controversies appear about which is the most proper therapy to treat metastatic malignant melanoma. Depending on whether the malignancy is unresectable or not, those acquired mutations may potentially influence in the progression of the disease [189]. In this sense, despite having been an important therapeutic strategy for palliation, chemotherapy has been shifted to secondary choice. Among the new therapeutic trends, the design of dysregulated pathway inhibitors (for example, for RAS/RAF/MEK/ERK MAPK pathway), the application of targeted therapies and the consideration of different immunotherapy strategies are currently prevailing [190]. More importantly, it should be noted that the ongoing research focused on the development of effective

resistance for treating cutaneous melanoma is highly challenging because of being a highly mutated cancer.

5.3. Motivation

With the demand for accurate and customised solutions for the patient, new targeted therapies have to be increasingly oriented towards extremely personalised medicine. Moreover, the success of these novelty therapeutic strategies may come with early diagnoses which anticipate the progression of the cancer. As it has been showed in the previous Chapter 3 and 4, DEGs usually help in offering clues about what genetic biomarkers discern better among different pathological states. However, the predisposition of certain biomarkers to present multiple mutations could hinder the determination of reliable DEGsDEGs, thus influencing in the variation of their gene expression levels. Despite having widely studied Single Nucleotide Polymorphisms (SNPs) using Genome-Wide Association Studies (GWAS), the inter-individual genetic variation provided by CNVs has been mostly ignored [191]. In this sense, the best findings about the occurrence of CNVs in cancer have been determined from the use of array Comparative Genomic Hybridization (aCGH) [192]. Nowadays, with the arrival of NGS, copy number variation extracted from WXS could help in elucidating which of those DEGs are likely dosage-sensitive by changing their expression due to alterations promoting loss or gain of gene copies. This fact could bring light in determining the susceptibility of genes affected by CNVs to tumorigenesis and angiogenesis. In search of offering new insights on the diagnosis of cutaneous melanoma, this chapter presents a methodological approach which determines reliable biomarkers related to the progression of the cutaneous melanoma from primary to metastatic state. For this purpose, an integrative analysis of gene expression and somatic copy number variation has

been designed and included within our clinical support approach (Figure 5.1). In search of offering new insights on the diagnosis of cutaneous melanoma, this work presents a methodological approach which determines reliable biomarkers related to the progression of the cutaneous melanoma from primary to metastatic state. For this purpose, an integrative analysis of gene expression and somatic copy number variation has been designed and included within our clinical support approach (Figure 5.1). On the basis of the motivations presented, the following experimental procedure has been performed. Firstly, by taking advantage of the previously demonstrated consistency when integrating microarray and RNA-seq datasets at gene expression level [148, 149] (also showed in the previous studies presented in Chapter 3 and 4), up to 18 different skin cancer datasets coming from 3 webdata repositories were considered. Concretely, 13 microarray and 5 RNA-seq datasets containing 596 samples in total were finally collected. For the preprocessing phase, only 532 samples were subjected after passing the quality control phase: 289 samples from Affymetrix and Illumina microarray platforms and 243 samples from Illumina RNA-seq platforms. Secondly, among the all previously selected samples, 73 of them corresponded to patients also having DNA-seq information: blood derived normal and primary or metastatic tumor samples. This allowed to perform an integrated analysis by using a 73 patients cohort containing 42 primary and 31 metastatic tumor samples. As a result of the whole integrative analysis, 26 DEGs showing remarkable somatic copy number variations were highlighted. Besides checking the functional properties of these biomarkers, their discernment capability for an eventual intelligent diagnosis was subjected by means of a robust ML process considering different classification techniques of the state-of-the-art. This study brings certain findings along about the existent informative correlation between gene expression and somatic copy number variation which helps in explaining the progression of cutaneous melanoma.

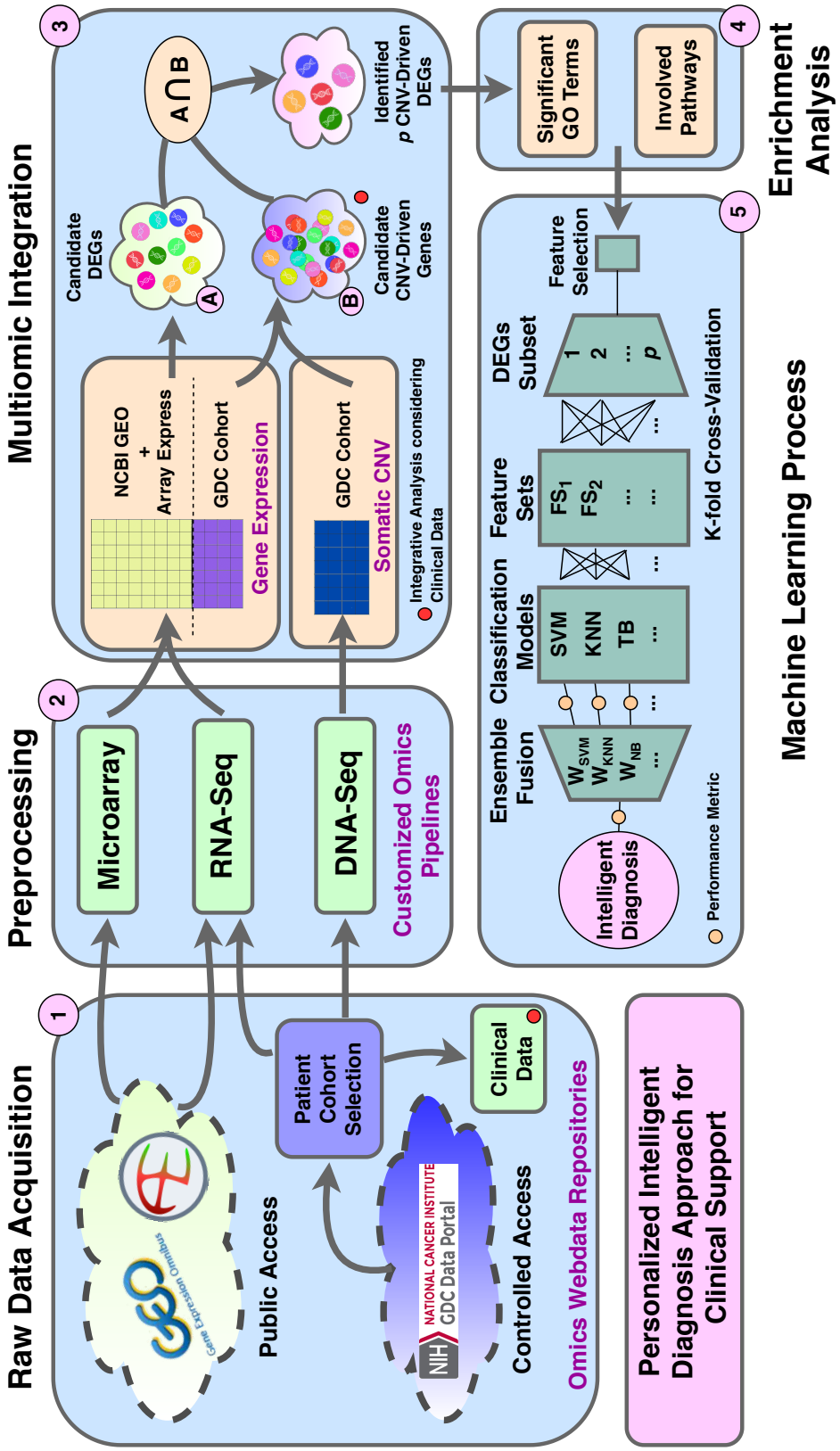


Figure 5.1: Workflow scheme for the design of our clinical support approach. Several challenging bioinformatic tasks have to be widely addressed: (1) raw data acquisition, (2) data preprocessing, (3) multiomic integration, (4) enrichment analysis and (5) machine learning process.

5.4. Methodologies and Experiments

5.4.1. Raw Data Acquisition

Three webdata repositories were inspected for collecting the biological samples: NCBI GEO [106], ArrayExpress [8] and NCI Genomic Data Commons (GDC) [9] (See Section 2.1.1.3 for details). In order to implement our multi-omic integration approach, several heterogeneous information sources were considered: from transcriptomic datasets (microarray and RNA-seq) to genomic datasets (DNA-seq). Both technologies have been widely introduced in Section 2.1.1.2. Finally, Up to 4 different sample types were considered for our study: healthy skin, nevus, primary melanoma and metastatic melanoma (see a wide description of samples types for skin cancer in Section 2.1.1.1). Tissue specimens were exclusively considered based on the guidelines of International Classification of Diseases (ICD-10) [1], selecting those skin tissues corresponding to trunk, upper limb (including shoulder) and lower limb (including hip). Total number of collected samples for each dataset and skin pathological state is specified in Table A.3A.

Concretely, a total of 13 different microarray series were selected from NCBI GEO and identified by means of their accession ID: GSE2503, GSE3189, GSE7553, GSE14905, GSE15605, GSE32407, GSE32924, GSE42677, GSE46517, GSE52471, GSE53223 and GSE8210. A range of R packages from Bioconductor web platform [154] were used to acquire these datasets collected on different Affymetrix platforms [2]: *GEOquery* [113], *affy* [112] and *oligo* [114]. RNA-seq sample sets under GSE54456, GSE58375 and GSE98394 accession ID from NCBI GEO and E-MTAB-5678 accession ID from ArrayExpress were also considered. Finally, RNA-seq and WXS samples for a cohort of 73 patients were downloaded under controlled access from GDC Portal. These data are part of TCGA-SKCM project under dbGaP study accession phs000178.v10.p8. They were downloaded

by means of authorised data access request under the following project title: "Multi-Omic integration using different sources of information. Advancing in Personalized Precision Medicine". Sample IDs and clinical data for the identification of this cohort within the webdata repository are included in Table A.

5.4.2. Preprocessing

5.4.2.1. Microarray Pipeline

As in the previous chapters, several quality metrics were assessed in order to discard low quality samples by means of *arrayQualityMetrics* R package [116]: density distribution plots, standard deviation of the samples intensities, Kolgomorov-Smirnov test based on the Ka parameter, distance among samples, principal component analysis (PCA) and Hoeffdings D-statistic (executed with $D < 0.15$). Total number of selected samples for each dataset can be consulted in Supplemental Table A.3B. Each dataset was separately preprocessed by considering Robust Multiarray Average (RMA) algorithm [25] whose implementation is included in *rma* function of *affy* and *oligo* R packages. Gene annotation standardisation of different datasets was established by translating chip identifiers to official gene symbols from the use of annotate R package [193].

5.4.2.2. RNA-Seq Pipeline

Standardised and well-established pipeline [155] was modified in order to preprocess all collected FASTQ and SRA files from Illumina manufacturer [3]. Count files were achieved after applying sra-toolkit [194], Hisat2 [157], bowtie2 [158], samtools [159] and HTSeq [4] tools. Gene annotation was retrieved by using *biomaRt* R package [160] and connecting to Ensembl database [195]. Correction and normalisation of GC content bias was performed by means of cqn [5]. Gene expression values were calculated by using NOISeq [6].

5.4.2.3. DNA-seq Pipeline

cn.MOPS [39] was selected for CNV detection by clearly improving the performance versus alternative methods: MOFDOC [40], EWT [41], JointSLM [42], CNV-Seq [43] and FREEC [44]. Segmentation window size selection and algorithm were carefully designed to setting CNV detection. Following the suggestions from cn.MOPS authors, an average number or reads in a window of 50-100 base pairs (bp) was established to guarantee a good performance. Window length (WL) can be calculated from the next simple relation:

$$W_L = \frac{m \cdot L}{C} \tag{5.1}$$

where m is the average number of reads per bin, L is the sequence read length and C represents the number of unique reads including a single nucleotide in the reconstructed sequence (redundancy of coverage or depth). Coverage was calculated by averaging all samples from the cohort of patients (see Table A) and subsequently applying the Lander-Waterman equation [196]:

$$C = \frac{N \cdot L}{G} \tag{5.2}$$

where N corresponds to the number of sequence reads, L is the sequence read length and G is the haploid genome size (in our study, diploid genome for human species). Sequence read length was equal to 76 bp/read for each DNA-seq sample by applying *samtools* and *idxstats* command. Number of sequence reads was averaged by considering 22 chromosomes together with X and Y chromosomes (around 84477952 reads). Genome size was retrieved as 3088269832 bp, so:

$$W_L = \frac{m \cdot G}{N} \tag{5.3}$$

where m is ranged from 50 to 100 base pairs. This implies window lengths

ranging from 1827 to 3655 bp. Window length was selected to 2000 bp. Circular binary segmentation algorithm was applied on our study [197]. Somatic CNVs were determined by individually applying referencecn.mops function on paired patient samples: blood derived normal versus tumor sample (control versus case setting). Copy number regions were mapped to loss value (-1 for CN0 and CN1) and gain value (1 for CN3 to CN128). Genes appearing in several chromosome regions were detected, cataloged as transition genes by changing the segmentation window and finally discarded of our downstream analysis.

5.4.3. Multiomic Integration

Logarithmic transformation (base 2) and 16-bit depth homogenisation (also referred to *dynamic range*) were required before applying gene expression integration. As mentioned above, batch effect correction was applied by means of ComBat method [34] from *sva* R package [121], showing the highest effectiveness when integrating microarrays [31] and RNA-seq datasets coming from different projects [150]. Inter-array normalisation was also considered, establishing an identical empirical distribution by applying normalizeBetweenArrays function from *limma* R package [38]. LFC and PV were used for selecting DEGs among both cutaneous melanoma states: LFC > 1 and PV < 0.001. Integrative analysis of gene expression and copy number variation was performed by using *iGC* R package [198]. False discovery rate (FDR) and PV were used as statistical parameters to restrict the selection of CNV-driven DEGs: FDR < 0.05 and PV < 0.001. Up to 9 different patient cohort analyses were carried out in function of the skin pathological state and gender:

1. Full patient cohort.
2. Only samples from men.
3. Only samples from women.

4. Only primary melanoma samples.
5. Only metastatic melanoma samples.
6. Only primary melanoma samples for men.
7. Only primary melanoma for women.
8. Only metastatic melanoma for men.
9. Only metastatic melanoma for women.

5.4.4. Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was performed by using DAVID Bioinformatics Database 6.8 [55] and *topGO* R package [199]. Highlighted GO terms divided into biological processes (BP), cellular components (CC) and molecular functions (MF) were retrieved. As introduced in Section 2.1.3.3, Reactome [57] and KEGG [58] pathway web browsers were inspected by using pathway identifiers retrieved from DAVID. Fisher's exact statistical test [52] was carried out to determine their significance (PV < 0.05) together with associated FDR. Our analysis was accompanied by widely-accepted standard statistics for multiple comparison corrections: Bonferroni [53] and Benjamini-Hochberg [54].

5.4.5. Machine Learning Process

As previously mentioned, feature selection was carried out by weighting three different rank correlation coefficients: Kendall [49], Pearson [50] and Spearman [51]. Weighted ranking was used to assess different gene subsets. Two feature sets were assessed: FS1 (considering only gene expression) and FS2 (considering gene expression together with CNV). Three individual classification models were trained and tested to compare their performance: SVM [60], KNN [61] and TB [63]. Additionally, ensemble fusion of these classifiers ENS was particularly designed and assessed [64], assigning different weights for each classifier (W_{SVM},

W_{KNN} and W_{TB}) in function of the individual recognition rates: W = 0.25 for the worst, W = 0.5 for the intermediate and W = 1 for the best. KFOLD-CV (where K = 10) [66] and overall F1-score [168] performance metric were considered to assess and measure the informative capability of the biomarkers.

5.5. Results

5.5.1. Determination of Somatic CNV-Driven DEGs Candidates

79 genes were cataloged as DEGs between primary and metastatic melanoma from the integration of microarray and RNA-seq datasets. This first gene set was obtained by intersecting three DEGs lists coming from assessing diverse sample subsets: (1) DEGs appearing from NCBI GEO and ArrayExpress samples (microarray + RNA-seq), (2) DEGs appearing from GDC Portal (RNA-Seq), and (3) DEGs coming from the integrated dataset (microarray + RNA-seq). Following, an integrative analysis considering gene expression and somatic CNVs determined which of those genes could be changing their gene expression value in function of alterations in gene copy number. By exclusively analysing the cohort of 73 patients from GDC, up to 26 of them showed simultaneous changes of both magnitudes, being statistically significant within any of the 9 cohort integrative analyses. Those were designed based on 3 discriminant search criteria of CNVs: disease (full patient cohort), state (primary or metastatic melanoma) and gender (men or women). No candidate gene appeared when primary melanoma in men or women, or metastatic melanoma in general were analysed. However, the 6 remaining criteria combinations presented candidate somatic CNV-driven DEGs: full patient cohort (DSG3), primary melanoma (SERPINB4), metastatic melanoma in men (DEFB1), metastatic melanoma in

women (CST6), cutaneous melanoma in men (CLCA2, CST6, IVL, KLK11, KRT5, KRT6A, KRT6B, KRT10, KRT14, KRT16, LCE3D, LOR, LYPD3, PKP3, S100A2, S100A7, S100A7A, SPRR1A, SPRR1B, SPRR2G, SFN and TRIM29) and cutaneous melanoma in women (KLK7 and SFN). Behind these results, the most remarkable interpretations highlight the involvement and alteration of a range of keratins and members of S100 family mainly affecting men, together with the generalised loss of gene copy number. No less important is to emphasise the behavior of SFN by losing copies in men and gaining them in women. In both cases, this eventuality seems to be more prominent for patients suffering from metastasised cutaneous melanoma (above 16% of cases for men and 23% for women). All the information about the determination of the 26 candidate somatic CNV-driven DEGs can be consulted in Table A.1.

5.5.2. Functional Characteristics Related to Highlighted Biomarkers

Only those 26 candidate somatic CNV-driven DEGs were subjected to an enrichment analysis. The functional profiles of these biomarkers were retrieved based on involved gene ontology terms and affected pathways. High statistical significance was imposed in order to only highlight those functional properties with the highest opportunities to have association with cutaneous melanoma. Diverse functional properties were classified within the 3 categories (Table 5.1), highlighting in bold 10 terms annotated coinciding from the use of two functional annotation tools (see Section 5.4 for details). Beyond showing apparent relationship with biological processes of the skin (keratinisation, keratinocyte differentiation, etc.), a wide range of those 26 candidate genes were associated with diverse cellular components involving the extracellular space. Among our highlighted biomarkers set, the structural integrity of the cytoskeleton and the complex assembly within or outside the cell were related.

Table 5.1: Functional enrichment analysis of the 26 candidate somatic CNV-driven DEGs using DAVID 6.8 and topGO R package. (Abbreviations: GO ID: Gene Ontology Identifer; GO: Gene Ontology; PV: P-Value; FDR: False Discovery Rate; BP: Biological Process; CC: Cellular Component; MF: Molecular Function)

Ontology	GO ID	GO Term	# Genes (%)	PV	FDR
BP	GO:0008544	epidermis development	13 (50,0)	9,49E-16	1,48E-12
	GO:0009888	tissue development	15 (57,7)	5,84E-09	8,66E-06
	GO:0009913	epidermal cell differentiation	9 (34,6)	5,83E-11	8,64E-08
	GO:0030216	keratinocyte differentiation	9 (34,6)	2,25E-12	3,33E-09
	GO:0030855	epithelial cell differentiation	10 (38,5)	2,70E-08	4,00E-05
	GO:0031424	keratinisation	7 (26,9)	4,96E-11	7,36E-08
	GO:0043588	skin development	9 (34,6)	5,53E-10	8,19E-07
	GO:0060429	epithelium development	11 (42,3)	5,17E-07	7,66E-04
CC	GO:0001533	cornified envelope	6 (23,1)	2,66E-09	2,88E-06
	GO:0005576	extracellular region	18 (69,2)	1,44E-06	1,55E-03
	GO:0005882	intermediate filament	6 (23,1)	5,01E-06	5,42E-03
	GO:0031982	vesicle	16 (61,5)	4,45E-06	4,81E-03
	GO:0031988	membrane-bounded vesicle	16 (61,5)	2,67E-06	2,88E-03
	GO:0043230	extracellular organelle	15 (57,7)	9,15E-07	9,89E-04
	GO:0044421	extracellular region part	17 (65,4)	8,94E-07	9,66E-04
	GO:0045111	intermediate filament cytoskeleton	7 (26,9)	5,09E-07	5,50E-04
	GO:0070062	extracellular exosome	15 (57,7)	8,56E-07	9,25E-04
	GO:1903561	extracellular vesicle	15 (57,7)	9,11E-07	9,85E-04
MF	**GO:0005198**	structural molecule activity	11 (42,3)	4,67E-08	5,09E-05
	GO:0005200	structural constituent of cytoskeleton	6 (23,1)	4,13E-07	4,50E-04

Table 5.2: Pathway analysis using DAVID 6.8 and linking to Reactome and KEGG web browsers. (Abbreviations: SNP: Single Nucleotide Polymorphism; MMP: matrix metalloproteinase; TP53: tumor protein p53)

Gene Symbol	Source	Pathway ID	Related topic
DEFB1	Reactome	R-HSA-1461957	SNPs
	Reactome	R-HSA-1461973	Extracellular region part
DSG3	Reactome	R-HSA-351906	Apoptosis
KLK7	Reactome	R-HSA-1474228	MMPs and extracellular matrix degradation
KRT14	Reactome	R-HSA-446107	Hemidesmosomes
KRT5	Reactome	R-HSA-446107	Hemidesmosomes
SFN	Reactome	R-HSA-5628897	TP53 regulation
	KEGG	hsa04110	Cell growth and death
	KEGG	hsa04115	TP53 signaling pathway

Those genes involved in each annotated term together with additional statistical results are specified in Table A.2. Additionally, pathway analysis elucidated the biological involvement of several highlighted biomarkers by our approach. By directly interpreting the retrieved information, those genes are related to biological processes involving the interaction with the extracellular matrix, playing a critical role in the maintenance and integrity of tissue structure. Different biological events of the cell cycle are critically related to specific biomarkers as well (cell development, apoptosis, etc.). Highlighted information about these findings is specified in Table 5.2.

5.5.3. Development and Progression of Cutaneous Melanoma

Gene expression levels of the 26 candidate somatic CNV-driven DEGs were compared by considering both cutaneous melanoma states (primary and metastatic) together with healthy skin and nevus samples. Our straightforward purpose consisted in inspecting whether gene expression levels could be changing among the different skin pathological states. Despite the identified genes were not selected to simultaneously discern among all the considered states

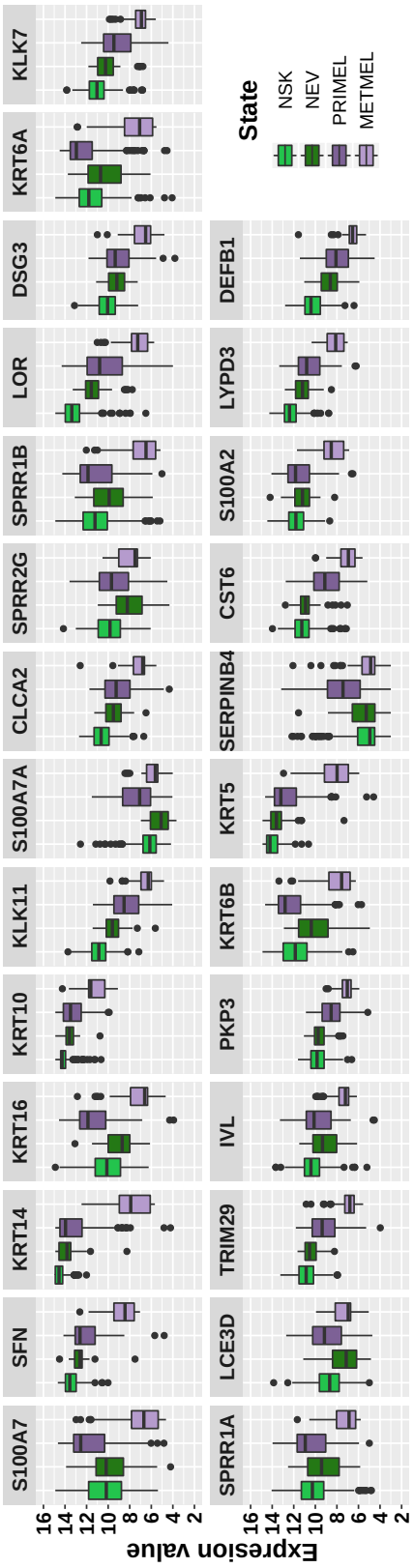


Figure 5.2: Expression level of the 26 candidate somatic CNV-driven DEGs, comparing up to 4 skin pathological states considered in our approach: NSK (normal skin), NEV (nevus), PRIMEL (primary melanoma) and METMEL (metastatic melanoma).

(healthy states were not initially considered for the gene selection purpose), our highlighted genes show a clearly revealing result: gene expression levels usually decrease when metastasising cutaneous melanoma. This trend is not exclusively appreciated when melanoma progresses from primary to metastatic state, but tumor degeneration could be taking place from healthy skin states to tumor states (Figure 5.2). For example, different gene expression levels can be distinguished for KLK7, KLK11 or LOR genes, ranging from 11-13 to 7 for gene expression values. In this sense, a wide range of these candidate biomarkers highlighted could offer clues about a more general progression of the cutaneous melanoma.

5.5.4. Intelligent Diagnosis for Clinical Support

A machine learning process was designed in search of assessing the informative capability of those biomarkers based on gene expression level and copy number variation. Our feature selection procedure ranked those 26 somatic CNV-driven DEGs candidate in the order shown in Figure 5.2 from left to right and up-bottom. Different subgroups of p genes, ranging p from the most informative gene (S100A7) to the whole set of 26 highlighted genes, were assessed by means of several well-trained and tested classification models. This experimental analysis was focus on assessing two feature sets (FS1 and FS2, defined in Section 5.4) and clearing two concerns up: 1) how much piece of information the somatic copy number variation provides to the gene expression for recognition purposes, and 2) how many genes are enough to offer a reliable intelligent diagnosis of cutaneous melanoma by considering both sources of information. Both concerns can be elucidated by interpreting the results showed in Figure 5.3. In general terms, the recognition rate increases from 0.5% to 2% for every additional gene including somatic copy number variation information (for example, LOR, ranked as 11th, showed improvements for KNN and TB models).

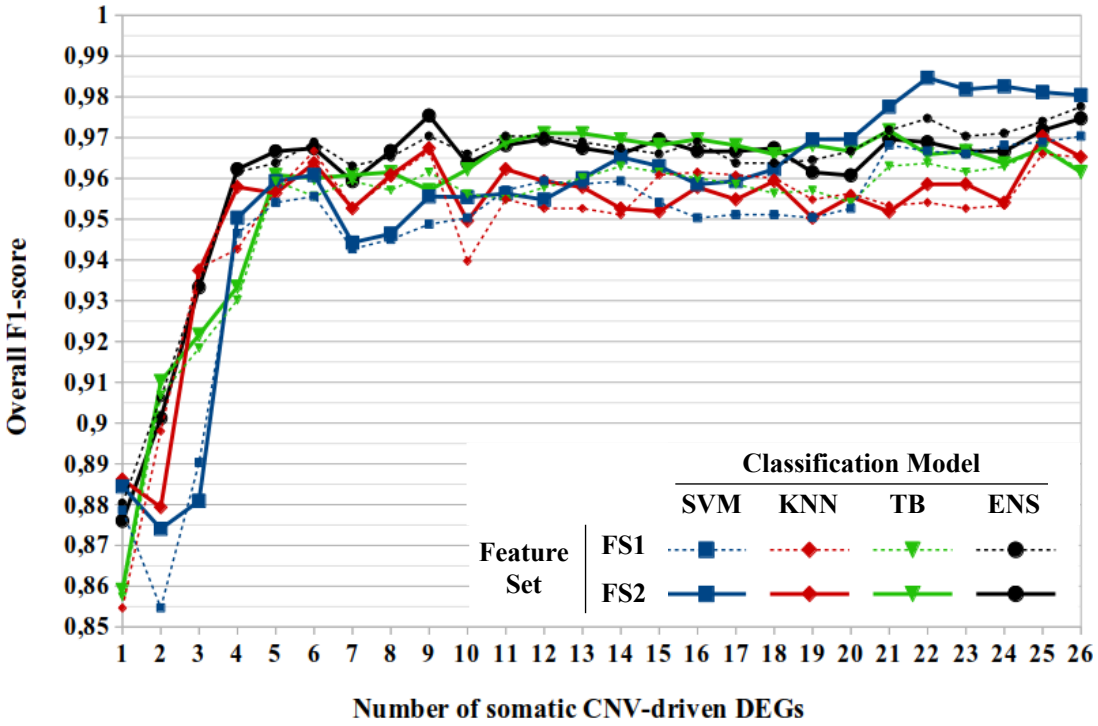


Figure 5.3: Comparison of the informative capability of the 26 candidate somatic CNV-driven DEGs by using two different feature sets. Three classification models (SVM, KNN and TB) and an ensemble fusion of all of them (ENS) was assessed by means of overall F1-score.

However, there are also specific genes with scarce recognition improvements when they are inserted in the classification process (for example, S100A7, ranked as 7th, for any classification model). Consequently, in order to reduce the complexity of our study, the classification using only the first 6 ranked genes was finally considered. This decision was taken given that the average improvement of the overall recognition rate per gene was lower than 0.05% after those 6 genes. By using exclusively this last reduced gene set, our classification procedure indicates that an intelligent diagnosis with a success probability above 95% and close to 97% could be attained for different classification models and for our ensemble solution, respectively.

5.5.5. Correlation between Gene Expression and Copy Number Variation

Gene expression levels and somatic copy number variation information were represented in order to see the influence of copy number alterations on gene expression. The generalised loss of gene copy number for the highlighted biomarkers by our methodological approach was previously mentioned. Figure 5.4 shows how gene expression level and specific alterations in somatic copy number variation (divided into loss and gain) seem to be correlated when patients are suffering from the disease. Specifically, those 6 most informative genes underlined by our approach reflected gene expression levels above median value for primary melanoma and around median value for metastatic melanoma when losing copy number. This result was significantly remarkable for men (see Table A.1). However, this fact may not be distinguished when gaining gene copy number. Furthermore, separation of samples from different webdata repositories was performed in order to check the distribution of gene expression levels. In spite of performing multiple preprocessing steps before integrating, the biological information was preserved by presenting similar distribution ranges within each dataset (for example, KRT14 was ranged from 8 to 15 log2 expression values in primary melanoma for both datasets). All these genes show unequivocally loss of gene expression when cutaneous melanoma metastasises.

5.6. Discussion

In this study, the informative correlation between gene expression and somatic copy number variation of significant genes being responsible for the progression of cutaneous melanoma was elucidated. This relationship has been previously noted under different experimental conditions: ulcerated versus non-ulcerated tumor subgroups [200], genetic subtypes defined by the presence of BRAF, NRAS or

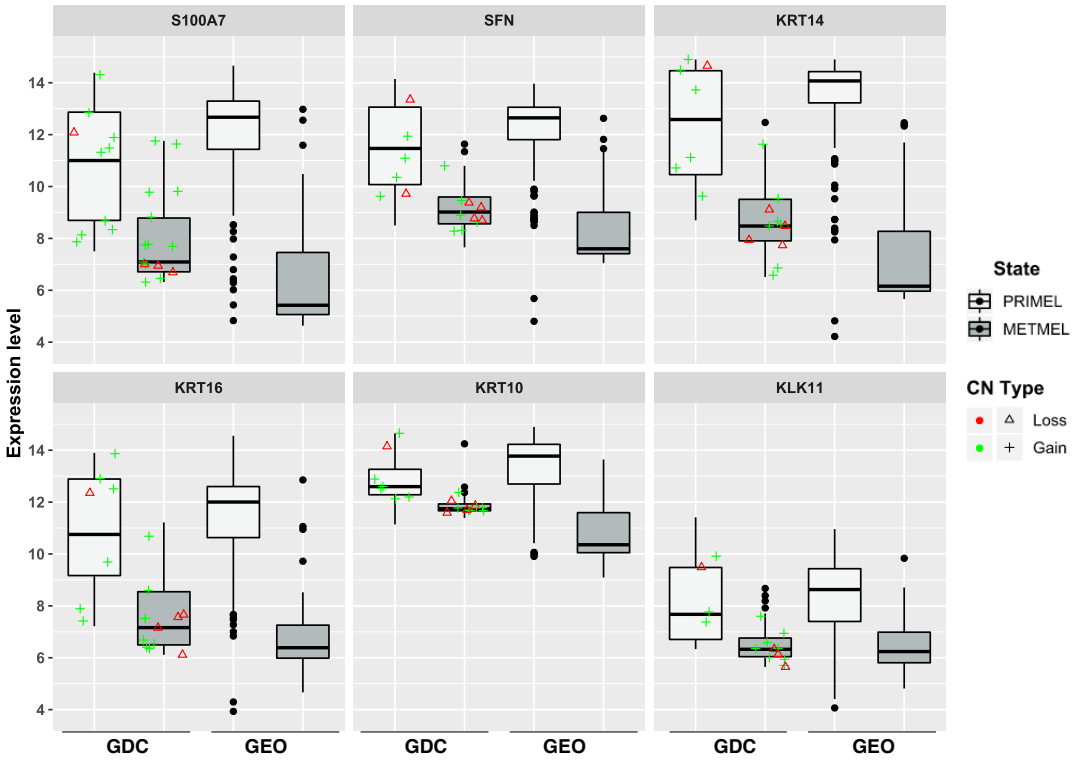


Figure 5.4: Gene expression values and somatic copy number variation for the 6 most informative genes highlighted by our approach. Distribution of two datasets was compared for both skin disease states: PRIMEL (primary melanoma) and METMEL (metastatic melanoma). CN: Copy Number.

NF1 mutations [201], etc. After carefully preprocessing and integrating multiple heterogeneous information sources, up to 26 candidate somatic CNV-driven DEGs emerged from our integrative analysis. Interestingly, all of them showed decreased expression when progressing cutaneous melanoma to advanced stages (Figure 5.2), which agrees with other findings in previously reported studies [202–204]. With respect to the detection of CNVs, it has been demonstrated that primary and metastatic states are highly similar in terms of their gene copy number alterations, loss of heterosygosity and single nucleotide variation (SNVs), thus making their differentiation extremely challenging [205]. In this sense, several previous works have performed great efforts in search of CNV differences among skin states related to cancer. On the one hand, genomic

differences between cutaneous melanoma and nevus have been showed by means of aCGH experiments [206, 207]. On the other hand, somatic CNVs have been detected by examining the influence of dividing into different sample thicknesses for distinguishing multiple primary cutaneous melanomas [208].

The functional enrichment analysis performed by analysing those 26 candidate somatic CNV-driven DEGs determined their significance and potential relationship to both gene expression and copy number variation (Table 5.1 and Table 5.2). Firstly, although different biological processes were associated with cellular and tissue development at epidermal and epithelial level, both keratinocyte differentiation and keratinisation were specially remarked. These findings are in accordance with recent studies of cutaneous melanoma [209], where the differentiation process is marked by contrasted markers such as loricrin (LOR) or involucrin (IVL) [210] together with family members of keratins. In fact, the differentiator role of this protein family has been also investigated in breast cancer [211–213]. Additionally, cutaneous melanoma may be delineated from epidermal stem cells marked accordingly such as KRT5 and KRT14, which are attached to the basement membrane through hemidesmosomes [214]. These structures keep adhered those epidermal keratinocytes to the extracellular matrix. In order to correctly preserve the skin homeostasis, their disruption may not take place. This fact supports the invasive behavior of tumor cells in absence of those keratins [215] and could explain the similar downregulation of other highlighted genes in our work. Despite being an approach that identifies potential relevant biomarkers from a reduced patient cohort, the biological involvement of all of them is consistent with findings of the current panorama dealing with melanoma diagnosis since the broadest analytic context. The importance of cellular components related to extracellular matrix such as vesicle or exosome has been recently demonstrated by analysing liquid biopsy samples [216], being also evidenced within our enrichment analysis (Table 5.1). Extensively, the integration of

heterogeneous datasets is becoming popular to extract robust knowledge by additionally considering different omic points of view. Our approach considered a patient cohort containing RNA-seq and WXS samples, but other previous studies extracted relevant knowledge by additionally considering whole genome sequencing (WGS) data, microRNA (miRNA) and methylation, among others [200, 201].

After identifying a set of outstanding biomarkers related to the progression of cutaneous melanoma, our approach determined which of them have the most powerful informative capability for an intelligent diagnosis. Under a robust machine learning classification process, the joint use of the first 6 ranked candidate somatic CNV-driven DEGs (S100A7, SFN, KRT14, KRT16, KRT10 and KLK11) guarantees to outperform recognition rates of 96% overall F_1-score (Figure 5.3). Beyond the role of keratins which was already discussed, the remaining biomarkers were also interestingly highlighted. S100A7 has been widely studied for several skin lesions, suggesting that the expression could be altered in association with early stages of skin tumorigenesis with highest levels and, conversely, downregulated in invasive state [217]. These findings correlate with our results, where expression level was upregulated in primary melanoma and downregulated in metastatic melanoma with respect to healthy skin states (Figure 5.2). Among both melanoma states, the downregulation of this gene was already associated with metastasis [204]. SFN has been simultaneously showed hypermethylated and dramatically downregulated in cutaneous melanoma, suggesting the pathogenic role of inhibiting angiogenesis [218, 219]. Finally, KLK11 has been associated with spreading of metastatic melanoma and overall survival of patients with primary melanoma [220]. Other member of kallikrin family, KLK7, was highlighted by our approach, presenting 31% women cases with copy number alterations in metastatic state (see Table A.1). Based on their diagnosis power and involvement in cutaneous

melanoma, our results confirmed previous findings of those biomarkers by using an independent dataset of samples. Furthermore, our approach is expected to contribute with an innovative point of view on how to integrate heterogeneous and multi-omic information in order to identify reliable biomarkers which improve the effectiveness of the clinical diagnosis of cutaneous melanoma. This work underlines the biological involvement of outstanding biomarkers being responsible for the progression of this malignancy. Extensively, the validity of our integrative pipeline is aimed to be applied on a wide range of complex and cancerous diseases, offering support for clinical diagnosis decisions. Among our next objectives, new improvements of our methodological approach are thought to reinforce the diagnosis power by considering miRNA [221] and methylation data [222] together with clinical data within the machine learning process.

5.7. Conclusions

Although diverse therapeutics strategies continue to appear for the effective treatment of cutaneous melanoma, the metastasis process remains uncontrollable and misunderstood due in large part to its biological heterogeneity. In search of biomarkers being responsible of this tumor degeneration, our methodological approach identified 26 somatic CNV-driven DEGs which reflected generalised loss of expression level and copy number losses ranging from 4% to 31% of total cases, being statistically significant within our assessed patient cohort. Besides supporting clinical decisions by means of an intelligent diagnosis, these findings encourage to deeply study the role of these biomarkers and dedicate new efforts in determining their influence on the metastatic process of cutaneous melanoma.

6. Conclusions & Future Work

6.1. Conclusions

This section gathers and underlines the general and most relevant conclusions of this doctoral thesis, as well as going into detail on those more specific conclusions that appeared from the carrying out of the different studies that make up the development of this thesis, previously noted in each corresponding chapter.

6.1.1. Conclusions about exclusively integrating microarrays

Based on the results obtained in Chapter 3, it is clear that the informational potential of this widely used sequencing technology is extremely high. As has been introduced and motivated in the Section 3.3, the arrangement of a large number of experiments on the same disease can contribute to the development of much more robust and statistically reliable studies than those carried out in isolation. It is precisely now, when data repositories are full of heterogeneous information, that a scientific effort must be made to try to exploit this information. This involves irrevocably developing advanced computing strategies that take into account multiple factors to achieve reliable integration of data sources. It is precisely now time to extend the knowledge about the appearance and development of multiple diseases, taking advantage of the existence of multiple algorithms widely-used and well-standardised for microarray analysis. However, only by addressing the following detected weak points will it be possible to extract more benefit from

the integration strategies:

1. **Microarray platforms reannotation:** It is postulated that it is essential that the scientific community make an additional effort in trying to implement tools that apply re-annotation to the different microarray platforms. This is due to the fact that the specifications and annotations between them vary significantly. This fact directly influences the extraction of relevant potential biomarkers because many of them will not appear after the integration phase. For example, by examining the Table 3.4, it can be glimpsed that the potential number of identified official gene symbols (under HUGO standard) varies between 12441 and 21035. It could be expected that if those symbols are common to each other and specific to each platform, the number of common symbols could become 12441. However, after integration, only 9978 common genes were obtained. This shows that there is at least a divergence of about 20% between platforms that will prevent subsequent analysis and possible loss of biomarkers potentially important for improving the diagnosis of the disease analysed. To a large extent, it is also influenced by the timing at which the experiments are performed, as the quality of genome assembly and the amount of annotation increases, and new experiments may improve in this respect significantly over predecessors.

2. **Effective batch effect removal:** Although the procedure approached for the minimisation of eventual batch effects by the integration proposed in Chapter 3 is considered highly robust, going into more detail on this issue is extremely necessary. As already motivated in the Section 3.3, dealing with this technical factor is highly challenging. In addition to applying methodological approaches emerging robust genes to the existence of hypothetical batch effects such as the one proposed in this thesis, it

must be asserted that the correction is taking place. Having as much information as possible regarding the conduct of experiments, conditions, assumptions, etc. can help to improve the performance of the several batch effect correction methods included in the literature. Extensively, iterative batch effect corrections could help to understand and determine the potential emerging deviations by microarray platform and technology. The biological relevance of subsequent analysis is profoundly influenced by this technical challenge. Based on the results obtained regarding the biological involvement of the proposed biomarkers and the classification rate (above 92% to discern 7 pathologies by means of 17 genes), the proposed integration pipeline is intended to partly satisfy this problem and is presumed to be extensively applied by the scientific community under the idea of selecting robust genes versus batch effects. This contribution is also considered achieved.

3. **Integration with a greater number of platforms:** This thesis has been based on the use of data from the main firms in the microarray sector: Affymetrix and Illumina. However, the repositories have microarray experiments carried out by Agilent, Taqman, Exon, etc. In the case of applying adequate procedures in order to have gene expression from these providers, additional challenges would come into play. On the one hand, challenging factors such as dealing with those mentioned above on a larger scale given the intrinsic heterogeneity of considering new data sources; on the other hand, computational complexity would increase to evaluate those emerging biomarkers in a classification process. On the contrary, in favour would be 2 issues: on the one hand, the reduction of the existing margin due to the course of dimensionality: more samples (n) versus the same number of genes (p) or lower due to integration without improvement by re-annotation; on the other hand, statistical robustness and reliability

should be further reinforced. The opposite option would suppose to consider for the integration a select group of samples coming from specific platforms where the number of genes annotated is higher and avoids losses in the integration. This implies irremediably sacrificing robustness in recognising certain pathological states at the expense of this being compensated by improving the efficiency of diagnosis in classification.

6.1.2. Conclusions about simultaneously co-integrating microarrays and RNA-seq

The proven consistency of integrating both transcriptomic data sources has extended the possibilities of research studies. By applying a rigorous integration process, a simple and intuitive way of determining relevant biomarkers for diagnosis has been devised: to check the optimal number of genes by discerning between each pair of pathologies that are necessary to improve intelligent diagnosis. Our methodological approach demonstrated that by simply selecting the DEG with the highest LFC among each pair of pathological states, it is possible to dramatically reduce the set of candidate genes to obtain an intelligent diagnosis of the analysed disease. Therefore, this study, which was conceived as an extension of the study presented in the Chapter 3, also manages to deal with and solve in a simple way the problem of the course of dimensionality: those "p" thousands of genes become tens and practically units of genes. This experimental contribution is intended to be applied in approaches based on inaccurate but highly efficient solutions such as those proposed under the paradigm of "soft computing". In addition, it is possible to invalidate something completely usual in an experimental process: the subjective criterion of the researcher, avoiding having to establish statistical thresholds, simply has to choose the best gene for each pair of pathologies (in this sense, this assessment must be taken with caution). Therefore, the second general contribution is thought to have

been widely satisfied by the satisfactory result offered by the methodological implementation presented here. As an extension to the conclusions pointed out previously in the Section 6.1.1, it can be insisted that considering now more potential transcripomic data can be integrated, the challenges posed must be even more carefully addressed. From this analytical perspective of promoter and to improve the previous, in the Chapter 4 it was decided to consider the correction of potential batch effects under the use of ComBat method. This method is highly recommended and effective against the integration of microarray and RNA-seq platforms as it was introduced, justified and motivated in the Section 4.3. As a learning from the study presented in Chapter 3, the sacrifice between leaving out of the integration some series of data in favor of not losing potentially relevant genes was taken into account. The introduction of gene enrichment analysis was also an improvement over the previous methodological approach.

6.1.3. Conclusions about co-integrating microarrays, RNA-seq and CNV

Taking a step forward, the informative correlation between gene expression and somatic variations in gene copies is the cornerstone of the study presented in Chapter 5. On this occasion, since gene expression is commonly used to discern between pathological states, we tried to see to what extent those genes appearing as DEGs could be really reliable. The consideration of information for the same cohort of patients at both genomic and transcriptomic levels was decisive in this study. The somatic variations helped to establish a filter for the selection of DEGs due to the additional presence of gene alterations. This approach is highly innovative and according to the results presented in the Chapter 5, the achievement of another objective can be highlighted with this contribution. By evaluating the different previous influencing factors that can be determinant in the subsequent analysis, it is necessary to highlight the selection of outstanding

biomarkers informing about the hypothetical progression of cutaneous melanoma. These insights are further discussed and concluded in the Section 6.1.4.

6.1.4. Biological level conclusions about panels of biomarkers

These observations can be specifically commented for each study carried out, with subtle differences in the omic viewpoints considered and the pathologies involved in the analysis. However, one of the most positive aspects of the application of the integration techniques proposed by this thesis lies in the fact that they were sequentially tested on the same cancerous disease: skin cancer. In this way, in spite of the small divergences at the experimental level to which one is subject, it is possible to extract a series of enriching conclusions in this respect:

1. **Panel of biomarcadores from microarray analysis:** Among the most important experimental findings, we must highlight the joint informative power of the genes DSC3, SCGB2A1 and BNC2 in a potential intelligent diagnosis evaluating the main epidermal cancerous pathologies of the skin (see Figure 3.9 where 4 pathological states surpassed 80%). More important has been the fact of verifying the biological involvement of these biomarkers (see S3 Appendix in the publication associated with this study [131] with more detailed information): on the one hand, DSC3 has shown a contradictory character and its simple deregulation in melanoma patients seems to imply predisposition to suffer tumorigenic processes in the skin. In addition, it is a protein-coding gene of the family of desmosomes that is responsible for maintaining the adhesion of cells. BNC2, associated with the risk of developing SCC, is considered a tumor suppressor gene during cancer development. A special mention deserves SCGB2A1. Although it not was directly related to skin cancer, the presented methodological approach in this dissertation, integrating a wide amount of heterogeneous data, brought

it to light, showing low levels expressed for practically all skin pathologies except MCC (see Figure 3.7 for more detail). Even so, it is glimpsed that this gene could have some biological implication in the development of skin cancer because it has already been related to the development of other epithelial cancers, whose cells are in charge of covering the inside and outside surfaces of the body.

2. **Panel of biomarkers from co-integrating transcriptomics technologies:** In addition to bringing together transcriptomic data from both co-existing technologies, this study focused on further testing the predisposition of those diseases already considered pre-cancerous of the skin to become skin cancer such as actinic keratosis and psoriasis. Highly discriminatory genes such as ADAMTS3 and LTF (discerning BCC and PS versus the rest, respectively) can offer valuable scientific knowledge to detect the early development of skin cancer. Besides highlighting the role of genes such as MMP1 thanks to perform an enrichment analysis, it is again curious and interesting to see how another gene from the secretoglobin family, SCGB2A2, has been highlighted as highly relevant. Future research will have to determine the validity of this gene as a differential biomarker in the diagnosis of skin cancer. Also, the appearance as a candidate of a keratin, KRT14, highly related to the skin.

3. **Panel of biomarkers considering simultaneously transcriptomic and genomic data:** Up to 18 of the 26 genes highlighted as DEGs and presenting somatic CNVs are related to the extracellular region according to the term GO. Interestingly, KRT14 was again highlighted and is related to hemidesmosomes that keep the epidermal keratinocytes attached to the extracellular matrix. Interactions with this region are critical since alterations can cause disruption. As a potential biological

contribution, it has been hypothesised that the generalised low-expression of the outstanding biomarkers can clearly indicate the invasive and progressive character of cutaneous melanoma. Other insights were justified based on the use of clinical data for the determination of these biomarkers, which gives the study a further point of experimental rigor.

On the basis of all the biological findings derived from the three research studies presented in this dissertation, the intersection of similar results between the different approaches is valuable. Among them, the emergence of a priori unexpected genes (such as SCGB2A1 and SCGB2A2 in the studies exclusively using gene expression) and the influence on biological processes of angiogenesis, tumorigenesis and melanogenesis of genes such as DSC3 or KRT14 together with other biomarkers such as MLANA, S100A7 or MMP1 can be highlighted. It is strongly thought that future studies further reinforcing the research initiatives carried out for the development of this thesis will contribute to the confirmation of some of these biomarkers as crucial in the diagnosis of skin cancer. Thanks to the knowledge acquired during this time, it is also possible to indicate some possible future projects that confirm the arguments presented throughout these conclusions.

6.2. Future Work

The main research efforts carried out during the course of this thesis were focused on the development of strategies for the integration of heterogeneous data with a global character. This was especially highlighted by the transcriptomic studies in which multiple pathological states of the skin were considered (7 for the study of Chapter 3 and 10 for the study of Chapter 4). However, this type of strategy must be irremediably complemented by other approaches that insist on determining what is exclusively and specifically related to the pathogenic

emergency. With more and more popular demand and translated into widespread economic and governmental implication, the marked guidelines for the treatment of today's and future diseases converge on the same point: the application of therapy and personalised and patient-oriented medicine. Evidently, Biomedical Engineering takes full advantage of existing advances in biological knowledge to be implemented in the health field and that can help in patient care. However, it is still necessary to find personalised solutions that satisfy the patient in a univocal improvement of his state of health under suffering from any disease.

Since throughout this thesis all experimental analyses have been carried out on skin cancer diagnosis, it must be said that promising and fascinating advances are expected for its early detection and effective treatment. The advanced computational techniques developed here have to be accompanied not only by other biological views translated into biological quantification of the patient's vital state, but also by image analysis algorithms of the disease area itself. It is expected the immediate irruption of strategies of integration of diverse sources of omic information together with own characteristics of the analysis of histopathological and dermoscopic images. The process of selection of relevant features and classification under operation of powerful machine learning algorithms and other approaches will be clearly enriched. Extensively, the personalised and strict follow-up of the patient must be translated into highly reliable and reliable clinical data. This fact would help to promote the discovery of biological knowledge that until now was probably unpredictable, avoiding the introduction of errors (deviations) in experimental analyses.

Although Science is in a vertiginous wave of scientific and technical advances that help us to delegate in automatic processes in favor of our wellness, the tools of intelligent diagnosis here exposed are not thought to eliminate the work of the specialist. On the contrary, they are directly designed to help and facilitate medical diagnosis. For example, the diagnosis of cutaneous melanoma under the

signs ABCDE will be accompanied by automatic analysis of histopathological and dermoscopic images together with transcriptomic and genomic analysis of the patient's sample. In the search for personalised attention, it is increasingly valuable to determine mechanisms for detecting biomarkers that promote the disease. In this sense, under a biomarker diagnostic kit offering a generalised background of the disease determined by methodological approaches such as those presented in Chapters 3, 4 and 5, it will be necessary to contrast with clinical data of the patient and look for which biomarkers can be specifically the triggers of the appearance, development and mutation of the disease. This will be the case of diseases such as skin cancer, which throughout this thesis has been described as a highly complex, heterogeneous and mutable disease. In addition to advancing in the improvement of the findings made in the Section 6.1 (re-annotation, effective deletion of batches, integration with more platforms, etc.), it would be desirable to enhance the realignment of samples that were sequenced in a genome prior to the current one. This would help to apply an update of the resources available for the discovery of new knowledge as well as correction and refinement for having improved biological precision (for example, new pseudogenes).

Specifically, and under the views previously argued and the proposals made in this thesis, it is intended to continue contributing in the future with new methodological approaches considering:

1. **Integration with other omic points of view:** Preferably using data from the same cohort of patients (such as those used for the study of the Chapter 5), it would be highly fascinating and challenging to process and integrate together proteomic, metabolomic, epigenomic data, etc.

2. **Integration with associated image characteristics:** Extensively, to have disease data in this format to extract new characteristics that help improve the classification process.

3. **Extension to cell line analysis:** It was based on Section 2.1.1.1 that only tissue samples were analysed. It is thought that valuable results could also be added by considering cell lines, serving at least as a reference or contrast in the analyses.

4. **Application of meta-analysis techniques:** It is just as important to offer generalised diagnoses in order to have a broad background as it is to compare with possible different genetic signatures coming from the different isolated and integrated studies.

5. **Immersion in the analysis of biomarker networks:** Considering different evaluations of enrichment analysis, one could take advantage of diagnostics based on evaluating groups of genes operating under predictably known biological mechanisms.

Conclusiones y Trabajo Futuro

Conclusiones

Esta sección reúne y subraya las conclusiones generales y más relevantes de esta tesis doctoral, así como ahonda en aquellas conclusiones más específicas que surgieron de la realización de los diferentes estudios que conforman el desarrollo de esta tesis, previamente apercibidas en cada capítulo correspondiente.

Conclusiones acerca de integrar exclusivamente microarrays

En base a los resultados obtenidos en el Capítulo 3, queda de manifiesto que el potencial informativo de esta ampliamente usada tecnología de secuenciación es extremadamente alto. Como ha sido introducido y motivado en la Sección 3.3, la disposición de un gran número de experimentos sobre una misma enfermedad puede contribuir al desarrollo de estudios mucho más robustos y estadísticamente fiables que aquellos llevados a cabo aisladamente. Es precisamente ahora, cuando los repositorios de datos se encuentran repletos de información heterogénea, cuando hay que hacer un esfuerzo científico para tratar de explotar dicha información. Esto pasa irrevocablemente por desarrollar estrategias de cómputo avanzadas que tengan en cuenta múltiples factores para la consecución de una integración fiable de fuentes de datos. Es ahora el momento de extender el conocimiento sobre la aparición y desarrollo de múltiples enfermedades, aprovechando que existen múltiples algoritmos ampliamente

usados y estandarizados para procesamiento de microarrays. Sin embargo, sólamente se conseguirá extraer mayor provecho de las estrategias de integración si se abordan los siguientes puntos débiles detectados:

1. **Reanotación de las plataformas de microarrays:** Se postula como fundamental que la comunidad científica haga un esfuerzo adicional en tratar de implementar herramientas que reanoten las diferentes plataformas de microarrays. Esto es debido a que las especificaciones y anotaciones entre ellas varían sensiblemente. Este hecho influye directamente en la extracción de potenciales biomarcadores relevantes debido a que muchos de ellos no aparecerán tras la fase de integración. Por ejemplo, examinando la Tabla 3.4, puede ser vislumbrado que el número potencial de símbolos de gen oficial identificados (bajo estándar HUGO) varía entre 12441 y 21035. Podría ser esperado que si aquellos símbolos son comunes entre ellos y específicos para cada plataforma, el número de símbolos comunes aspiraría a alcanzar 12441. Sin embargo, tras la integración, únicamente se obtuvieron 9978 genes comunes. Esto demuestra que existe al menos una divergencia del 20% aproximadamente entre plataformas que impedirá el análisis subsecuente y la posible pérdida de biomarcadores potencialmente importantes para la mejora del diagnóstico de la enfermedad analizada. En gran medida, también viene influenciado por el momento temporal en que los experimentos son realizados, ya que la calidad de ensamblado del genoma y la cantidad de anotación incrementa, y los nuevos experimentos pueden mejorar en este aspecto sensiblemente con respecto a los predecesores.

2. **Borrado efectivo de efectos de batch:** Aunque el procedimiento abordado para la minimización de eventuales efectos de batch por parte de la integración propuesta en el Capítulo 3 es considerado altamente robusto, profundizar en esta problemática es extremadamente necesario.

Como ya fue motivado en la Sección 3.3, lidiar con este factor técnico es altamente retante. Además de aplicar aproximaciones metodológicas emergiendo genes robustos a la existencia de hipotéticos efectos de batch como la propuesta en esta tesis, hay que aseverarse de que la correción está teniendo lugar. Disponer de la mayor cantidad de información con respecto a la realización de los experimentos, condiciones, suposiciones, etc. pueden ayudar a mejorar las prestaciones de los diversos métodos de corrección de efectos de batch existentes en la literatura. Extensiblemente, correcciones de efectos de batch de manera iterativa podría ayudar a comprender y a determinar cuáles son las potenciales desviaciones emergentes por plataforma y tecnología de microarray. La relevancia biológica del análisis subsequente se ve profundamente influenciada por este reto técnico. A tenor de los resultados obtenidos en lo que respecta a implicación biológica de los biomarcadores propuestos y la tasa de clasificación (superior al 92% para discernir 7 patologías con 17 genes), es pensado que el pipeline de integración propuesto puede satisfacer en parte esta problemática y se presume que pueda ser extensiblemente aplicado por la comunidad científica bajo la idea de seleccionar genes robustos a efectos de batch. Esta contribución se considera alcanzada.

3. **Integración con un mayor número de plataformas:** Esta tesis ha sido cimentada sobre la utilización de datos procedentes de las principales firmas en el sector de microarrays: Affymetrix e Illumina. Sin embargo, los repositorios cuentan con experimentos de microarrays llevados a cabo por Agilent, Taqman, Exon, etc. En el caso de aplicar procedimientos adecuados en pos de disponer de expresión de gen desde estos proveedores, retos adicionales entrarían en juego. Por un lado, factores retantes como lidiar con los mencionados anteriormente a mayor escala dada la heterogeneidad intrínseca de considerar nuevas fuentes de datos; por

otro lado, aumentaría la complejidad computacional para evaluar aquellos biomarcadores emergentes en un proceso de clasificación. Por el contrario, a favor se encontrarían 2 cuestiones: por un lado, la reducción del margen existente por curso de la dimensionalidad: más muestras (n) frente a mismo número de genes (p) o inferior debido a la integración sin mejora por la reanotación; por otro lado, la robustez y fiabilidad estadística debería verse aún más reforzada. La opción contraría supondría considerar para la integración un selecto grupo de muestras provenientes de específicas plataformas donde el número de genes anotados es más alto y evita pérdidas en la integración. Esto implica irremediablemente sacrificar robustez en reconocer ciertos estados patológicos siempre y cuando ese sacrificio merezca la pena en términos de eficiencia de diagnóstico en clasificación superior.

Conclusiones acerca de co-integrar microarrays y RNA-seq

La probada consistencia de integrar ambas fuentes de datos transcriptómicos ha extendido las posibilidades de análisis. Aplicando un proceso riguroso de integración, se ha pensado una forma sencilla e intuitiva de determinar biomarcadores relevantes para el diagnóstico: comprobar el número óptimo de genes discerniendo entre cada par de patologías que son necesarios para mejorar el diagnóstico inteligente. Nuestra aproximación metodológica demostró que simplemente seleccionando el DEG con mayor LFC entre cada par de estados patológicos, es posible reducir dramáticamente el conjunto de genes candidatos para obtener un diagnóstico inteligente de la enfermedad analizada. Por tanto, este estudio que fue pensado como una extensión del estudio presentado en el Capítulo 3, además consigue lidiar y solventar de forma sencilla el problema del curso de la dimensionalidad: aquellos ”p” miles de genes pasan a ser decenas y prácticamente unidades de genes. Esta contribución experimental es pensada a ser de gran valor para el abordaje de aproximaciones cimentadas en soluciones

inexactas pero altamente eficientes como las planteadas bajo el paradigma de "computación flexible". Además, se consigue invalidar algo completamente habitual en un proceso experimental: el criterio subjetivo del investigador, que no tiene que establecer ningún tipo de umbral estadístico, simplemente tiene que escoger el mejor gen para cada par de patologías (eso sí, esta apreciación hay que tomarla con matices). En este sentido, la segunda contribución general es pensada a haber sido ampliamente satisfecha por el resultado satisfactorio ofrecido por la implementación metodológica presentada. Como extensión a las conclusiones puntualizadas anteriormente en la Sección 6.2, se puede insistir en que considerando ahora más potenciales datos transcriptómicos pudiendo ser integrados, los retos planteados deben ser aún más cuidadosamente abordados. Desde esta perspectiva analítica y promotora de mejorar lo previo, en el Capítulo 4 se optó por considerar la corrección de potenciales efectos de batch bajo el método ComBat, que es altamente recomendado y eficaz frente a la integración de plataformas de microarrays y de RNA-seq como fue introducido, justificado y motivado en la Sección 4.3. Como aprendizaje del estudio presentado en el Capítulo 3, se evaluó el sacrificio entre dejar fuera de la integración algunas series de datos en favor de no perder genes potencialmente relevantes. La introducción de análisis de enriquecimiento de los genes fue también una mejora con respecto a la aproximación metodológica anterior.

Conclusiones acerca de co-integrar microarrays, RNA-seq y variación de números de copias

Como una vuelta de tuerca más, la correlación informativa entre la expresión de gen y las variaciones somáticas en copias de gen es el pilar fundamental del estudio presentado en el Capítulo 5. En esta ocasión, dado que la expresión de gen es usada habitualmente para discernir entre estados patológicos, se trato de ver hasta qué punto aquellos genes apareciendo como DEGs podrían ser

realmente fiables. La consideración de información genómica desde el mismo cohorte de pacientes que para información transcriptómica fue determinante en este estudio. Las variaciones somáticas ayudaron a establecer un filtro de selección de DEGs por presentar adicionalmente alteraciones génicas. Esta aproximación es altamente innovativa y a tenor de los resultados presentados en el Capítulo 5, se puede remarcar la consecución de otro objetivo con esta contribución. Evaluando los diferentes factores influyentes previos que pueden ser determinantes en el análisis subsecuente, hay que destacar la selección de excelentes biomarcadores informando sobre la hipotética progresión del melanoma cutáneo. Estas apreciaciones son más ampliamente comentadas y concluidas en la Sección 6.2.

Conclusiones a nivel biológico sobre los paneles de biomarcadores

Estas observaciones pueden ser específicamente comentadas para cada estudio llevado a cabo, con diferencias sutiles en lo que respecta a los puntos de vista ómicos considerados y las patologías envueltas en el análisis. Sin embargo, uno de los aspectos más positivos de la aplicación de las técnicas de integración propuestas por esta tesis reside en que fueron secuencialmente testeadas sobre una misma enfermedad cancerosa: el cáncer de piel. De esta manera, a pesar de las pequeñas divergencias a nivel experimental a las que se encuentra uno sujeto, es posible extraer una serie de conclusiones enriquecedoras al respecto:

1. **Panel de biomarcadores desde análisis de microarrays:** Entre las averiguaciones experimentales más importantes, hay que destacar el poder informativo conjunto de los genes DSC3, SCGB2A1 y BNC2 en un potencial diagnóstico inteligente evaluando las principales patologías cancerosas epidermales (ver Figura 3.9 donde 4 estados patológicos ya

superan el 80%). Más importante ha sido comprobar la implicación biológica de estos biomarcadores (ver S3 Appendix en la publicación asociada a este estudio [131] con información más detallada): por un lado, DSC3 ha mostrado un carácter contradictorio y su simple desregulación en pacientes con melanoma parece implicar predisposición a padecer procesos tumorigénicos en la piel. Se trata de un gen codificando proteína de la familiar de desmosomas que se encarga de mantener la adhesión de las células. BNC2, asociado a riesgo de desarrollar SCC, es considerado un gen supresor de tumor durante desarrollo del cáncer. Una mención especial merece SCGB2A1, que aún no habiendo sido relacionado directamente a cáncer de piel, la aproximación metodológica considerando una amplia integración de datos heterogéneos presentada en esta tesis lo sacó a la luz, mostrando niveles bajo expresados para prácticamente todas las patologías de piel excepto MCC (ver Figura 3.7 para más detalle). Aún así, se vislumbra que este gen podría tener alguna implicación biológica en el desarrollo de cáncer de piel por haber sido ya relacionado al desarrollo de otros cánceres epiteliales, cuyas células se encargan de cubrir las superficies internas y externas del cuerpo.

2. **Panel de biomarcadores co-integrando las tecnologías transcriptómicas:** Además de aunar datos transcriptómicos de ambas tecnologías co-existentes, este estudio fue enfocado a comprobar adicionalmente la predisposición a degenerar en cáncer de piel aquellas enfermedades ya consideradas pre-cancerosas de la piel como son la keratosis actínica y la psoriasis. Genes altamente discriminatorios respecto al resto como ADAMTS3 y LTF (discerniendo BCC y PS frente al resto, respectivamente) pueden ofrecer valiosos conocimientos científicos para detectar precozmente el desarrollo de cáncer de piel. Además de destacar el rol de genes como MMP1 gracias al análisis de enriquecimiento,

resulta nuevamente curioso e interesante ver cómo otro gen de la familia secretoglobin, SCGB2A2, ha sido destacado como altamente relevante. Futuras investigaciones tendrán que determinar la validez de este gen como biomarcador diferencial en el diagnóstico de cáncer de piel. También, la aparición como candidato de una keratina, KRT14, altamente relacionada a la piel.

3. **Panel de biomarcadores considerando transcriptómica y genómica simultáneamente:** Hasta 18 de los 26 genes destacados como DEGs y presentando CNVs somáticos se encuentran relacionados con la región extracelular según el término GO. Interesantemente, KRT14 fue nuevamente destacado y se encuentra relacionado con los hemidesmosomas que guardan adheridos los keratinocitos epidermales a la matriz extracelular. Las interacciones con esta región son críticas ya que alteraciones pueden provocar disrupción. Como potencial contribución biológica, ha sido hipotetizado que la bajo-expresión generalizada de los biomarcadores destacados pueden denotar claramente el carácter invasivo y progresivo del melanoma cutáneo. Otras percepciones fueron justificadas en base al uso de datos clínicos para la determinación de dichos biomarcadores, lo que le otorga un punto más de rigurosidad experimental al estudio llevado a cabo.

En base a todas las apreciaciones biológicas derivadas de la realización de los tres estudios de investigación presentados en esta disertación, es valorable la intersección de resultados semejantes entre las diferentes aproximaciones. Entre ellas, destacan la aparición de genes a priori inesperados (como SCGB2A1 y SCGB2A2 en los estudios exclusivamente usando expresión de gen) y la influencia en procesos biológicos de angiogenesis, tumorigenesis y melanogenesis de genes como DSC3 o KRT14 junto con otros biomarcadores como MLANA, S100A7

o MMP1. Es conscientemente pensado que estudios futuros reforzando aún más las iniciativas de investigación llevadas a cabo para el desarrollo de esta tesis contribuirán a la confirmación de algunos de estos biomarcadores como determinantes en el diagnóstico de cáncer de piel. Gracias al conocimiento adquirido durante esta etapa, es posible indicar algunos posibles proyectos futuros que confirmen los argumentos presentados a lo largo de estas conclusiones.

Trabajo Futuro

Los principales esfuerzos de investigación llevados a cabo durante el transcurso de esta tesis fueron enfocados en el desarrollo de estrategias de integración de datos heterogéneos con carácter global. Esto fue especialmente remarcado por los estudios transcriptómicos en los que se consideraron múltiples estados patológicos de la piel (7 para el estudio del Capítulo 3 y 10 para el estudio del Capítulo 4). Sin embargo, este tipo de estrategia debe ser complementada irremediablemente por otras que ahonden en la búsqueda de lo exclusivo y específico de la emergencia patogénica. Cada vez con más reclamo popular y traducido en extendida implicación económica y gubernamental, las líneas marcadas para el tratamiento de las enfermedades del hoy y del futuro convergen en un único punto: la aplicación de la terapia y la medicina personalizada y orientada al paciente. Evidentemente, la Ingeniería Biomédica aprovecha por completo los avances existentes en conocimiento biológico para ser implementados en el ámbito sanitario y que puedan ayudar en la atención del paciente. Sin embargo, sigue siendo necesario encontrar soluciones personalizadas que satisfagan al paciente en una mejora unívoca de su estado de salud bajo padecimiento de cualquier enfermedad.

Puesto que a lo largo de esta tesis se han llevado a cabo todos los análisis experimentales sobre diagnóstico del cáncer de piel, cabe decir que

prometedores y fascinantes avances son esperados para su detección precoz y efectivo tratamiento. Las técnicas de cómputo avanzadas desarrolladas aquí tienen que ser acompañadas no sólamente por otros puntos de vista biológicos traducidos en cuantificación biológica del estado vital del paciente, si no por algortimos de análisis de imágenes de la propia zona de la enfermedad. Es esperada la inmediata irrupción de estrategias de integración de diversas fuentes de información ómicas junto con características propias del análisis de imágenes histopatológicas y dermoscópicas. El proceso de selección de características relevantes y clasificación bajo operación de algoritmos poderosos de aprendizaje máquina y otras aproximaciones se verá francamente enriquecido. Extensiblemente, el seguimiento personalizado y estricto del paciente debe traducirse en unos datos clínicos altamente fehacientes y fiables. Este hecho ayudaría a potenciar el descubrimiento de conocimiento biológico hasta ahora probablemente impredecible, evitando la introducción de errores (desviaciones) en los análisis experimentales.

Aunque estamos en una vertiginosa ola de avances científicos y técnicos que nos ayudan a delegar en procesos automáticos en favor de nuestro bienestar, las herramientas de diagnóstico inteligente aquí expuestas no son pensadas para eliminar la labor del especialista. Al contrario, son directamente pensadas para ayudar y facilitar el diagnóstico médico. Por ejemplo, el diagnóstico del melanoma cutáneo bajo los signos ABCDE será acompañado por análisis automáticos de imágenes histopatológicas y dermoscópicas junto con análisis transcriptómicos y genómicos de la muestra del paciente. En la búsqueda de una atención personalizada, resulta cada vez más valorable la determinación de mecanismos de detección de biomarcadores promotores de la enfermedad. En este sentido, bajo un kit de diagnóstico de biomarcadores ofreciendo un background generalizado de la enfermedad determinado por aproximaciones metodológicas como las presentadas en los Capítulos 3, 4 y 5, será necesario contrastar con

datos clínicos del paciente y buscar qué biomarcadores pueden ser específicamente los detonantes de la aparición, el desarrollo y la mutación de la enfermedad. Éste será el caso de enfermedades como el cáncer de piel, que a lo largo de esta tesis se ha descrito como una enfermedad altamente compleja, heterogénea y mutable. Además de avanzar en la mejora de las apreciaciones realizadas en la Sección 6.2 (reanotación, borrado efectivo de batches, integración con más plataformas, etc.), sería deseable potenciar el realineamiento de muestras que fueron secuenciadas en un genoma anterior al actual. Esto ayudaría a aplicar una actualización de los recursos disponibles tanto para descubrimiento de nuevo conocimiento como corrección y refinamiento por haber mejorado en la precisión biológica (por ejemplo, nuevos pseudogenes).

Específicamente, y bajo las visiones previamente argumentadas y las propuestas realizadas en esta tesis, se pretende continuar aportando en lo venidero con nuevas aproximaciones metodológicas considerando:

1. **Integración con otros puntos de vista ómicos:** Preferiblemente bajo la utilización de datos provenientes del mismo cohort de pacientes (como los usados para el estudio del Capítulo 5), sería altamente fascinante y retante procesar e integrar conjuntamente datos proteómicos, metabolómicos, epigenómicos, etc.

2. **Integración con características de imágenes asociadas:** Extensiblemente, disponer de datos referentes a la enfermedad en este formato para extraer nuevas características que ayuden a mejorar el proceso de clasificación.

3. **Extensión al análisis de líneas celulares:** Fue fundamentado en Sección 2.1.1.1 que únicamente se analizaron muestras de tejido. Es pensado que resultados valuables podrían ser también añadidos por considerar líneas celulares, sirviendo al menos como referencia o contraste en el análisis.

4. **Aplicación de técnicas de meta-análisis:** Tan importante es ofrecer diagnósticos generalizados para tener un background amplio como cotejar con eventuales firmas genéticas diferentes provenientes de los diferentes estudios aislados e integrados.

5. **Inmersión en el ánalisis de redes de biomarcadores:** En torno a la consideración de diferentes evaluaciones del análisis de enriquecimiento, se podría tomar ventaja de diagnósticos basados en evaluar grupos de genes operando bajo mecanismos biológicos previsiblemente conocidos.

Acronyms

ACC	Accuracy.
aCGH	array Comparative Genomic Hybridization.
AE	ArrayExpress.
AJCC	American Joint Committee on Cancer.
AK	Actinic Keratosis.
ANOVA	ANalysis Of VAriance.
BCC	Basal Cell Carcinoma.
BP	Biological Processes.
CC	Cellular Components.
CNAs	Copy Number Alterations.
CNVs	Copy Number Variations.
DEGs	Differentially Expressed Genes.
EB	Empirical Bayes.
ENS	Ensemble Learning.
GO	Gene Ontology.
GQ	Gene Quantile.
GSEA	Gene Set Enrichment Analysis.
GWAS	Genome-Wide Association Studies.
HGN	Human Gene Nomenclature.

HGNC	HUGO Gene Nomenclature Committee.
HUGO	HUman Genome Organization.
KEGG	Kyoto Encyclopedia of Genes and Genomes.
KFOLD-CV	K-Fold Cross-Validation.
KNN	K-Nearest Neighbors.
LFC	Log2-Fold-Change.
LOO-CV	Leave-One-Out Cross-Validation.
MC	Mean Centering.
MCC	Merkel Cell Carcinoma.
METMEL	Metastatic Melanoma.
MF	Molecular Functions.
MF1	Mean F1-score.
ML	Machine Learning.
MMCC	Metastatic Merkel Cell Carcinoma.
mRMR	Minimum Redundancy Maximum Relevance.
MRS	Mean Rank Scores.
MSC	Melanoma Skin Cancer.
NB	Naive Bayes.
NCBI GEO	National Center for Biotechnology Information - Gene Expression Omnibus.
NCI GDC	National Cancer Institute - Genomic Data Commons.
NEV	Nevus.
NGS	Next-Generation Sequencing.
NMSC	Non-Melanoma Skin Cancer.
NORDI	Normal Discretization.

NSK Normal Skin.

OF1 Overall F1-score.

PCA Principal Component Analysis.

PMCC Primary Merkel Cell Carcinoma.

PRIMEL Primary Melanoma.

PS Psoriasis.

PV P-Value.

QD Quantile Discretization.

RMA Robust Multi-array Average.

RNA-seq RiboNucleic Acid sequencing.

SCC Squamous Cell Carcinoma.

SNPs Single Nucleotide Polymorphisms.

SRA Sequence Read Archive.

SVM Support Vector Machine.

TB Tree Bagging.

WXS Whole Exome Sequencing.

List of Figures

List of Tables

Bibliography

[1] Organization WH. International statistical classification of diseases and related health problems. vol. 1. World Health Organization; 2004.

[2] Gohlmann H, Talloen W. Gene expression studies using Affymetrix microarrays. CRC Press; 2009.

[3] Illumina, Inc. Illumina: Illumina Gene Expression arrays.; 2009. Available from: `http://www.illumina.com/techniques/microarrays/gene-expression-arrays.html`.

[4] Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–169.

[5] Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2):204–216.

[6] Tarazona S, Furió-Tar\'\i P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43(21):e140—-e140.

[7] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012;41(D1):D991–D995.

[8] Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic acids research. 2010;39(suppl_1):D1002–D1004.

[9] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. New England Journal of Medicine. 2016;375(12):1109–1112.

[10] Goldstein AM, Tucker MA. Dysplastic nevi and melanoma. AACR; 2013.

[11] Kauvar AN, Cronin Jr T, Roenigk R, Hruza G, Bennett R. Consensus for nonmelanoma skin cancer treatment: basal cell carcinoma, including a cost analysis of treatment methods. Dermatologic Surgery. 2015;41(5):550–571.

[12] Gandhi SA, Kampp J. Skin cancer epidemiology, detection, and management. Medical Clinics. 2015;99(6):1323–1335.

[13] Asgari MM, Moffet HH, Ray GT, Quesenberry CP. Trends in basal cell carcinoma incidence and identification of high-risk subgroups, 1998-2012. JAMA dermatology. 2015;151(9):976–981.

[14] Stratigos A, Garbe C, Lebbe C, Malvehy J, Del Marmol V, Pehamberger H, et al. Diagnosis and treatment of invasive squamous cell carcinoma of the skin: European consensus-based interdisciplinary guideline. European journal of cancer. 2015;51(14):1989–2007.

[15] Muzic JG, Schmitt AR, Wright AC, Alniemi DT, Zubair AS, Lourido JMO, et al. Incidence and trends of basal cell carcinoma and cutaneous squamous cell carcinoma: a population-based study in Olmsted County, Minnesota, 2000 to 2010. In: Mayo Clinic Proceedings. vol. 92. Elsevier; 2017. p. 890–898.

[16] Becker JC, Stang A, DeCaprio JA, Cerroni L, Lebbé C, Veness M, et al. Merkel cell carcinoma. Nature Reviews Disease Primers. 2017;3:17077.

[17] Schadendorf D, Lebbé C, zur Hausen A, Avril MF, Hariharan S, Bharmal M, et al. Merkel cell carcinoma: epidemiology, prognosis, therapy and unmet medical needs. European Journal of Cancer. 2017;71:53–69.

[18] Lebbe C, Becker JC, Grob JJ, Malvehy J, Del Marmol V, Pehamberger H, et al. Diagnosis and treatment of Merkel cell carcinoma. European consensus-based interdisciplinary guideline. European Journal of Cancer. 2015;51(16):2396–2403.

[19] Shenenberger DW. Cutaneous malignant melanoma: a primary care perspective. American family physician. 2012;85(2).

[20] Rigel DS, Friedman RJ, Kopf AW, Polsky D. ABCDE—an evolving concept in the early detection of melanoma. Archives of dermatology. 2005;141(8):1032–1034.

[21] Eggermont AM, Spatz A, Robert C. Cutaneous melanoma. The Lancet. 2014;383(9919):816–827.

[22] Richard MA, Barnetche T, Horreau C, Brenaut E, Pouplard C, Aractingi S, et al. Psoriasis, cardiovascular events, cancer risk and alcohol use: evidence-based recommendations based on systematic review and expert opinion. Journal of the European Academy of Dermatology and Venereology. 2013;27:2–11.

[23] Egeberg A, Thyssen J, Gislason G, Skov L. Skin cancer in patients with psoriasis. Journal of the European Academy of Dermatology and Venereology. 2016;30(8):1349–1353.

[24] Fuchs A, Marmur E. The kinetics of skin cancer: progression of actinic keratosis to squamous cell carcinoma. Dermatologic Surgery. 2007;33(9):1099–1101.

[25] Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–193.

[26] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC bioinformatics. 2011;12(1):480.

[27] Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO gene nomenclature database, 2006 updates. Nucleic acids research. 2006;34(suppl_1):D319–D321.

[28] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic acids research. 2005;33(suppl_1):D54–D58.

[29] Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic acids research. 2018;47(D1):D745–D751.

[30] Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO gene nomenclature committee (HGNC). Human genetics. 2001;109(6):678–680.

[31] Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PloS one. 2011;6(2):e17238.

[32] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–264.

[33] Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: A survey. Briefings in Bioinformatics. 2013;14(4):469–490. Cited By 52.

[34] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–127.

[35] Martinez R, Pasquier N, Pasquier C. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. Bioinformatics. 2008;24(22):2643–2644.

[36] Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. BMC medical genomics. 2008;1(1):42.

[37] Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends in biotechnology. 2017;35(6):498–507.

[38] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47–e47.

[39] Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, et al. cn. MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic acids research. 2012;40(9):e69–e69.

[40] Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics. 2009;41(10):1061.

[41] Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome research. 2009;19(9):1586–1592.

[42] Magi A, Benelli M, Yoon S, Roviello F, Torricelli F. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. Nucleic acids research. 2011;39(10):e65–e65.

[43] Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC bioinformatics. 2009;10(1):80.

[44] Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics. 2010;27(2):268–269.

[45] Rutherford A. Introducing ANOVA and ANCOVA: a GLM approach. Sage; 2001.

[46] Fisher RA. Contributions to mathematical statistics. 1950;.

[47] Turner JR, Thayer J. Introduction to analysis of variance: design, analyis & interpretation. Sage Publications; 2001.

[48] Montgomery DC. Design and analysis of experiments. John wiley & sons; 2017.

[49] Abdi H. The Kendall rank correlation coefficient. Encyclopedia of Measurement and Statistics Sage, Thousand Oaks, CA. 2007;p. 508–510.

[50] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Springer; 2009. p. 1–4.

[51] Zar JH. Spearman rank correlation. Encyclopedia of Biostatistics. 2005;7.

[52] Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society. 1922;85(1):87–94.

[53] BONFERRONI C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze. 1936;8:3–62. Available from: `https://ci.nii.ac.jp/naid/20001561442/en/`.

[54] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289–300.

[55] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2009;4(1):44.

[56] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25.

[57] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic acids research. 2017;46(D1):D649–D655.

[58] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic acids research. 2018;47(D1):D590–D595.

[59] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology. 2005;3(02):185–205.

[60] Noble WS. What is a support vector machine? Nature biotechnology. 2006;24(12):1565.

[61] Parry R, Jones W, Stokes T, Phan J, Moffitt R, Fang H, et al. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. The pharmacogenomics journal. 2010;10(4):292.

[62] Kohavi R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Kdd. vol. 96. Citeseer; 1996. p. 202–207.

[63] Breiman L. Bagging predictors. Machine learning. 1996;24(2):123–140.

[64] Dietterich TG. Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer; 2000. p. 1–15.

[65] Shao J. Linear model selection by cross-validation. Journal of the American statistical Association. 1993;88(422):486–494.

[66] Arlot S, Celisse A, et al. A survey of cross-validation procedures for model selection. Statistics surveys. 2010;4:40–79.

[67] He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering. 2008;(9):1263–1284.

[68] DePinho RA. The age of cancer. Nature. 2000;408(6809):248–254. Cited By 635.

[69] Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. British Journal of Dermatology. 2012;166(5):1069–1080.

[70] Watson M, Thomas CC, Massetti GM, McKenna S, Gershenwald JE, Laird S, et al. CDC Grand Rounds: Prevention and Control of Skin Cancer. American Journal of Transplantation. 2016;16(2):717–720. Cited By 1.

[71] Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. 2013; Lyon, France: International Agency for Research on Cancer. globocan iarc fr/Default aspx. 2014;.

[72] Staples MP, Elwood M, Burton RC, Williams JL, Marks R, Giles GG. Non-melanoma skin cancer in Australia: The 2002 national survey and trends since 1985. Medical Journal of Australia. 2006;184(1):6–10. Cited By 348.

[73] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: a cancer journal for clinicians. 2018;68(1):7–30.

[74] Volkov A, Dobbinson S, Wakefield M, Slevin T. Seven-year trends in sun protection and sunburn among Australian adolescents and adults. Australian and New Zealand Journal of Public Health. 2013;37(1):63–69. Cited By 44.

[75] Katalinic A, Kunze U, Schäfer T. Epidemiology of cutaneous melanoma and non-melanoma skin cancer in Schleswig-Holstein, Germany: Incidence, clinical subtypes, tumour stages and localization (epidemiology of skin cancer). British Journal of Dermatology. 2003;149(6):1200–1206. Cited By 152.

[76] Tejera-Vaquerizo A, Descalzo-Gallego MA, Otero-Rivas MM, Posada-García C, Rodríguez-Pazos L, Pastushenko I, et al. Cancer Incidence and Mortality in Spain: A Systematic Review and Meta-Analysis [Incidencia y mortalidad del cáncer cutáneo en España: revisión sistemática y metaanálisis Skin]. Actas Dermo-Sifiliograficas. 2016;107(4):318–328. Cited By 14.

[77] Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. Genome Biology. 2013;14(7):R75. Cited By 78.

[78] Mitsui H, Suárez-Fariñas M, Gulati N, Shah KR, Cannizzaro MV, Coats I, et al. Gene expression profiling of the leading edge of cutaneous squamous cell carcinoma: IL-24-driven MMP-7. Journal of Investigative Dermatology. 2014;134(5):1418–1427. Cited By 15.

[79] Sand M, Skrygan M, Georgas D, Sand D, Hahn SA, Gambichler T, et al. Microarray analysis of microRNA expression in cutaneous squamous cell carcinoma. Journal of Dermatological Science. 2012;68(3):119–126. Cited By 53.

[80] Harms PW, Patel RM, Verhaegen ME, Giordano TJ, Nash KT, Johnson CN, et al. Distinct gene expression profiles of viral- and nonviral-associated merkel cell carcinoma revealed by transcriptome analysis. Journal of Investigative Dermatology. 2013;133(4):936–945. Cited By 43.

[81] Salah B, Alshraideh M, Beidas R, Hayajneh F. Skin cancer recognition by using a neuro-fuzzy system. Cancer Informatics. 2011;10:1–11. Cited By 11.

[82] Hoshyar AN, Al-Jumaily A, Hoshyar AN. The beneficial techniques in preprocessing step of skin cancer detection system comparing. vol. 42; 2014. p. 25–31. Cited By 5.

[83] Ray PJ, Priya S, Kumar TA. Nuclear segmentation for skin cancer diagnosis from histopathological images; 2015. p. 397–401. Cited By 0.

[84] Jaworek-Korjakowska J, Tadeusiewicz R. Determination of border irregularity in dermoscopic color images of pigmented skin lesions; 2014. p. 6459–6462. Cited By 11.

[85] Van Der Geer S, Kleingeld PAM, Snijders CCP, Rinkens FJCH, Jansen GAE, Neumann HAM, et al. Development of a non-melanoma skin cancer detection model. Dermatology. 2015;230(2):161–169. Cited By 0.

[86] Vuong K, McGeechan K, Armstrong BK, Cust AE. Risk prediction models for incident primary cutaneous melanoma: A systematic review. JAMA Dermatology. 2014;150(4):434–444. Cited By 15.

[87] Li Y, Esteva A, Kuprel B, Novoa R, Ko J, Thrun S. Skin Cancer Detection and Tracking using Data Synthesis and Deep Learning. arXiv preprint arXiv:161201074. 2016;.

[88] Calin MA, Parasca SV, Savastru R, Calin MR, Dontu S. Optical techniques for the noninvasive diagnosis of skin cancer. Journal of Cancer Research and Clinical Oncology. 2013;139(7):1083–1104. Cited By 47.

[89] Sattlecker M, Stone N, Bessant C. Current trends in machine-learning methods applied to spectroscopic cancer diagnosis. TrAC - Trends in Analytical Chemistry. 2014;59:17–25. Cited By 6.

[90] Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Scientific reports. 2017;7(1):4172.

[91] Kather JN, Weis CA, Bianconi F, Melchers SM, Schad LR, Gaiser T, et al. Multi-class texture analysis in colorectal cancer histology. Scientific reports. 2016;6:27988.

[92] Misganaw B, Vidyasagar M. Exploiting Ordinal Class Structure in Multiclass Classification: Application to Ovarian Cancer. IEEE life sciences letters. 2015;1(1):15–18.

[93] Doyle S, Feldman M, Tomaszewski J, Shih N, Madabhushi A. Cascaded multi-class pairwise classifier (CASCAMPA) for normal, cancerous, and cancer confounder classes in prostate histology. In: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE; 2011. p. 715–718.

[94] Haqq C, Nosrati M, Sudilovsky D, Crothers J, Khodabakhsh D, Pulliam BL, et al. The gene expression signatures of melanoma progression. Proceedings of the National Academy of Sciences. 2005;102(17):6092–6097.

[95] Romo-Bucheli D, Moncayo R, Cruz-Roa A, Romero E. Identifying histological concepts on basal cell carcinoma images using nuclei based sampling and multi-scale descriptors. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. IEEE; 2015. p. 1008–1011.

[96] Maryam, Setiawan NA, Wahyunggoro O. A hybrid feature selection method using multiclass SVM for diagnosis of erythemato-squamous disease. In: AIP Conference Proceedings. vol. 1867. AIP Publishing; 2017. p. 020048.

[97] Maurya R, Singh SK, Maurya AK, Kumar A. GLCM and Multi Class Support vector machine based automated skin cancer classification. In: Computing for Sustainable Global Development (INDIACom), 2014 International Conference on. IEEE; 2014. p. 444–447.

[98] Choudhury D, Naug A, Ghosh S. Texture and color feature based WLS framework aided skin cancer classification using MSVM and ELM. In: India Conference (INDICON), 2015 Annual IEEE. IEEE; 2015. p. 1–6.

[99] Pérez-Ortiz M, Sáez A, Sánchez-Monedero J, Gutiérrez PA, Hervás-Martínez C. Tackling the ordinal and imbalance nature of a melanoma image classification problem. In: Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE; 2016. p. 2156–2163.

[100] Sundar RS, Vadivel M. Performance analysis of melanoma early detection using skin lession classification system. In: Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on. IEEE; 2016. p. 1–5.

[101] Bishop JM. Molecular themes in oncogenesis. Cell. 1991;64(2):235–248. Cited By 1273.

[102] Yang J, Zhou J, Zhu Z, Ma X, Ji Z. Iterative ensemble feature selection for multiclass classification of imbalanced microarray data. Journal of Biological Research-Thessaloniki. 2016;23(1):13.

[103] Lê Cao KA, McLachlan GJ. Statistical analysis on microarray data: selection of gene prognosis signatures. In: Computational Biology. Springer; 2009. p. 55–76.

[104] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences. 2001;98(26):15149–15154.

[105] Jenefer BM, Cyrilraj V. An innovative hybrid mathematical hierarchical regression model for breast cancer diseases analysis. Cluster Computing;p. 1–14.

[106] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: Mining tens of millions of expression profiles - Database and tools update. Nucleic Acids Research. 2007;35(SUPPL. 1):D760–D765. Cited By 792.

[107] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002;30(1):207–210. Cited By 4581.

[108] Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014; 2014.

[109] Sharma G, Martin J. MATLAB®: A language for parallel computing. International Journal of Parallel Programming. 2009;37(1):3–36. Cited By 56.

[110] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):R80. Cited By 7132.

[111] Hornik K. The Comprehensive R Archive Network. Wiley Interdisciplinary Reviews: Computational Statistics. 2012;4(4):394–398. Cited By 9.

[112] Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–315. Cited By 2104.

[113] Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846–1847.

[114] Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010;26(19):2363–2367. Cited By 315.

[115] Du P, Kibbe WA, Lin SM. lumi: A pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547–1548. Cited By 1106.

[116] Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics - A bioconductor package for quality assessment of microarray data. Bioinformatics. 2009;25(3):415–416.

[117] Heider A, Alt R. VirtualArray: A R/bioconductor package to merge raw data from different microarray platforms. BMC Bioinformatics. 2013;14(1). Cited By 21.

[118] Kieslich PJ, Henninger F. Package 'readbulk'; 2016. Available from: `https://github.com/pascalkieslich/readbulk`.

[119] Taminau J, Taminau MJ, Meganck S, BiocGenerics S. Package 'inSilicoMerging'. March; 2013.

[120] Taminau J, Steenhoff D, Coletta A, Meganck S, Lazar C, De schaetzen V, et al. inSilicoDb: An R/bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. Bioinformatics. 2011;27(22):3204–3205. Cited By 23.

[121] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010;11(10):733.

[122] Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry. 1964;36(8):1627–1639. Cited By 9197.

[123] Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015. Cited By 131.

[124] Hoffmann R. A wiki for the life sciences where authorship matters. Nature Genetics. 2008;40(9):1047–1051. Cited By 117.

[125] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ.

DISEASES: Text mining and data integration of disease-gene associations. Methods. 2015;74:83–89. Cited By 65.

[126] Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: A platform for therapeutic target identification and Validation. Nucleic Acids Research. 2017;45(D1):D985–D994. Cited By 13.

[127] Fontaine J, Andrade-Navarro M. Gene Set to Diseases (GS2D): Disease Enrichment Analysis on Human Gene Sets with Literature Data. Genomics and Computational Biology. 2016;2(1):33. Available from: `https://www.genomicscomputbiol.org/ojs/index.php/GCB/article/view/33`.

[128] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics. 2004;69(6 2):066138–1–066138–16. Cited By 886.

[129] Refaeilzadeh P, Tang L, Liu H. Cross-validation. In: Encyclopedia of database systems. Springer; 2009. p. 532–538.

[130] Strauss T, Von Maltitz MJ. Generalising ward's method for use with manhattan distances. PLoS ONE. 2017;12(1). Cited By 1.

[131] Gálvez JM, Castillo D, Herrera LJ, San Román B, Valenzuela O, Ortuño FM, et al. Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series. PloS one. 2018;13(5):e0196836.

[132] Giblin AV, Thomas JM. Incidence, mortality and survival in cutaneous melanoma. J Plast Reconstr Aesthet Surg. 2007;60(1):32–40.

[133] Apalla Z, Lallas A, Sotiriou E, Lazaridou E, Ioannides D. Epidemiological trends in skin cancer. Dermatol Pract Concept. 2017;7(2):1–6.

[134] Whiteman DC, Green AC, Olsen CM. The growing burden of invasive melanoma: projections of incidence rates and numbers of new cases in six susceptible populations through 2031. J Invest Dermatol. 2016;136(6):1161–1171.

[135] Eide MJ, Krajenta R, Johnson D, Long JJ, Jacobsen G, Asgari MM, et al. Identification of patients with nonmelanoma skin cancer using health maintenance organization claims data. Am J Epidemiol. 2009;171(1):123–128.

[136] Pouplard C, Brenaut E, Horreau C, Barnetche T, Misery L, Richard MA, et al. Risk of cancer in psoriasis: a systematic review and meta-analysis of epidemiological studies. J Eur Acad Dermatol. 2013;27:36–46.

[137] Dai H, Li WQ, Qureshi AA, Han J. Personal history of psoriasis and risk of nonmelanoma skin cancer (NMSC) among women in the United States: A population-based cohort study. Journal of the American Academy of Dermatology. 2016;75(4):731–735. Cited By 2.

[138] Schmitz L, Gambichler T, Gupta G, Stücker M, Dirschka T. Actinic keratosis area and severity index (AKASI) is associated with the incidence of squamous cell carcinoma. J Eur Acad Dermatol. 2018;32(5):752–756.

[139] Lober BA, Lober CW. Actinic keratosis is squamous cell carcinoma. South Med J. 2000;93(7):650–655.

[140] Fix WC, Yun SJ, Groft MacFarlane CM, Jambusaria A, Elenitsas R, Chu E, et al. MART-1-labeled melanocyte density and distribution in actinic keratosis and squamous cell cancer in situ: Pagetoid melanocytes are a potential source of misdiagnosis as melanoma in situ. J Cutan Pathol. 2018;45(10):734–742.

[141] Tan KB, Tan SH, Aw DCW, Jaffar H, Lim TC, Lee SJ, et al. Simulators of squamous cell carcinoma of the skin: diagnostic challenges on small biopsies and clinicopathological correlation. J Skin Cancer. 2013;2013.

[142] Bichakjian CK, Lowe L, Lao CD, Sandler HM, Bradford CR, Johnson TM, et al. Merkel cell carcinoma: critical review with guidelines for multidisciplinary management. Cancer. 2007;110(1):1–12.

[143] Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin. 2017;67(2):93–99.

[144] Padilla RS, Sebastian S, Jiang Z, Nindl I, Larson R. Gene expression patterns of normal human skin, actinic keratosis, and squamous cell carcinoma: a spectrum of disease progression. Arch Dermatol. 2010;146(3):288–293.

[145] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57–63. Cited By 4695.

[146] Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res. 2005;33(18):5914–5923.

[147] Irigoyen A, Jimenez-Luna C, Benavides M, Caba O, Gallego J, Ortuño FM, et al. Integrative multi-platform meta-analysis of gene expression profiles in pancreatic ductal adenocarcinoma patients for identifying novel diagnostic biomarkers. PLoS One. 2018;13(4):e0194844.

[148] Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in Saccharomyces cerevisiae. Nucleic Acids Research. 2012;40(20):10084–10097. Cited By 135.

[149] Castillo D, Gálvez JM, Herrera LJ, San Román B, Rojas F, Rojas I. Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling. BMC bioinformatics. 2017;18(1):506.

[150] Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, Zhang L, et al. Unifying cancer and normal RNA sequencing data from different sources. Sci Data. 2018;5:180061.

[151] Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457.

[152] Fisher RA. The statistical utilization of multiple measurements. Ann Eugen. 1938;8(4):376–386.

[153] Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics. 2004;20(15):2429–2437.

[154] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115–121.

[155] Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8(9):1765.

[156] Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(Suppl. 1):D19–D21.

[157] Kim D, Langmead B, Salzberg S. HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes; 2017.

[158] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–359.

[159] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079.

[160] Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):W589—-W598.

[161] Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2017;46(D1):D754—-D761.

[162] Önskog J, Freyhult E, Landfors M, Rydén P, Hvidsten TR. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC Bioinformatics. 2011;12(1):390.

[163] Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD, et al.. sva: Surrogate Variable Analysis; 2018.

[164] Mitra S, Hayashi Y. Bioinformatics with soft computing. IEEE Trans Syst Man Cybern C Appl Rev. 2006;36(5):616–635.

[165] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70–79.

[166] Turki T, Wei Z. Boosting support vector machines for cancer discrimination tasks. Comput Biol Med. 2018;101:236–249.

[167] Stone M. Asymptotics for and against cross-validation. Biometrika. 1977;p. 29–35.

[168] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Comm Com Inf Sc. 2009;45(4):427–437.

[169] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4(1):44.

[170] Kaufman HL, Russell J, Hamid O, Bhatia S, Terheyden P, D'Angelo SP, et al. Avelumab in patients with chemotherapy-refractory metastatic Merkel cell carcinoma: a multicentre, single-group, open-label, phase 2 trial. Lancet Oncol. 2016;17(10):1374–1385.

[171] Asgari MM, Sokil MM, Warton EM, Iyer J, Paulson KG, Nghiem P. Effect of host, tumor, diagnostic, and treatment variables on outcomes in a large cohort with Merkel cell carcinoma. JAMA Dermatol. 2014;150(7):716–723.

[172] Amaral T, Leiter U, Garbe C. Merkel cell carcinoma: Epidemiology, pathogenesis, diagnosis and therapy. Rev Endocr Metab Disord. 2017;18(4):517–532.

[173] Shen L, Liu L, Yang Z, Jiang N. Identification of genes and signaling pathways associated with squamous cell carcinoma by bioinformatics analysis. Oncol Lett. 2016;11(2):1382–1390.

[174] Kuznetsova EV, Snarskaya ES, Zavalishina LE, Tkachenko SB. Immunohistochemical study of the specific features of expression of matrix

metalloproteinases 1, 9 in the photoaged skin, the foci of actinic keratosis and basal cell carcinoma. Arkh Patol. 2016;78(6):17–22.

[175] Swindell WR, Sarkar MK, Liang Y, Xing X, Gudjonsson JE. Cross-disease transcriptomics: unique IL-17A signaling in psoriasis lesions and an autoimmune PBMC signature. J Invest Dermatol. 2016;136(9):1820–1830.

[176] Sun Y, Huang J, Yang Z. The roles of ADAMTS in angiogenesis and cancer. Tumour Biol. 2015;36(6):4039–4051.

[177] Kerkelä E, Saarialho-Kere U. Matrix metalloproteinases in tumor progression: focus on basal and squamous cell skin cancer. Exp Dermatol. 2003;12(2):109–125.

[178] Ren S, Liu S, Howell Jr P, Xi Y, Enkemann SA, Ju J, et al. The impact of genomics in understanding human melanoma progression and metastasis. Cancer Control. 2008;15(3):202–215.

[179] Degesys CA, Powell HB, Hsia LB, Merritt BG. Outcomes for Invasive Melanomas Treated With Mohs Micrographic Surgery: A Retrospective Cohort Study. Dermatol Surg. 2018;.

[180] Gerami P, Cook RW, Wilkinson J, Russell MC, Dhillon N, Amaria RN, et al. Development of a prognostic genetic signature to predict the metastatic risk associated with cutaneous melanoma. Clinical Cancer Research. 2015;21(1):175–183.

[181] Melero JL, Andrades S, Arola L, Romeu A. Deciphering psoriasis. A bioinformatic approach. J Dermatol Sci. 2018;89(2):120–126.

[182] Dai J, Lin K, Huang Y, Lu Y, Chen WQ, Zhang XR, et al. Identification of critically carcinogenesis-related genes in basal cell carcinoma. Onco Targets Ther. 2018;11:6957.

[183] Sauer CM, Haugg AM, Chteinberg E, Rennspiess D, Winnepenninckx V, Speel EJ, et al. Reviewing the current evidence supporting early B-cells as the cellular origin of Merkel cell carcinoma. Crit Rev Oncol Hematol. 2017;116:99–105.

[184] Kontochristopoulos GJ, Stavropoulos PG, Krasagakis K, Goerdt S, Zouboulis CC. Differentiation between Merkel cell carcinoma and malignant melanoma: an immunohistochemical study. Dermatology. 2000;201(2):123–126.

[185] Wei W, Chen Y, Xu J, Zhou Y, Bai X, Yang M, et al. Identification of Biomarker for Cutaneous Squamous Cell Carcinoma Using Microarray Data Analysis. Journal of Cancer. 2018;9(2):400.

[186] Zheng L, Wang R, Chi S, Li C. Matrix metalloproteinase 1: a better biomarker for squamous cell carcinoma by multiple microarray analyses. Giornale italiano di dermatologia e venereologia: organo ufficiale, Societa italiana di dermatologia e sifilografia. 2017;.

[187] Guy Jr GP, Thomas CC, Thompson T, Watson M, Massetti GM, Richardson LC. Vital signs: melanoma incidence and mortality trends and projections—United States, 1982–2030. MMWR Morb Mortal Wkly Rep. 2015;64(21):591.

[188] Ahmed F, Haass NK. Microenvironment-driven dynamic Heterogeneity and phenotypic plasticity as a mechanism of Melanoma therapy resistance. Frontiers in oncology. 2018;8.

[189] Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017;545(7653):175.

[190] O'Sullivan J, O'Connor D. The modern approach to targeting melanoma. In: Human Skin Cancers-Pathways, Mechanisms, Targets and Treatments. IntechOpen; 2018. .

[191] Shlien A, Malkin D. Copy number variations and cancer. Genome medicine. 2009;1(6):62.

[192] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nature genetics. 2005;37(6s):S11.

[193] Gentleman R. annotate: Annotation for microarrays; 2019.

[194] Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. Nucleic acids research. 2011;40(D1):D54—-D56.

[195] Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic acids research. 2018;47(D1):D745—-D751.

[196] Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988;2(3):231–239.

[197] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557–572.

[198] Lai YP, Wang LB, Wang WA, Lai LC, Tsai MH, Lu TP, et al. iGC—an integrated analysis package of gene expression and copy number alteration. BMC bioinformatics. 2017;18(1):35.

[199] Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology; 2018.

[200] Rakosy Z, Ecsedi S, Toth R, Vizkeleti L, Herandez-Vargas H, Lazar V, et al. Integrative genomics identifies gene signature associated with melanoma ulceration. PLoS One. 2013;8(1):e54958.

[201] Akbani R, Akdemir KC, Aksoy BA, Albert M, Ally A, Amin SB, et al. Genomic classification of cutaneous melanoma. Cell. 2015;161(7):1681–1696.

[202] Riker AI, Enkemann SA, Fodstad O, Liu S, Ren S, Morris C, et al. The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. BMC Med Genomics. 2008;1(1):13.

[203] Jaeger J, Koczan D, Thiesen HJ, Ibrahim SM, Gross G, Spang R, et al. Gene expression signatures for tumor progression, tumor subtype, and tumor thickness in laser-microdissected melanoma tissues. Clinical cancer research. 2007;13(3):806–815.

[204] Koh SS, Wei JPJ, Li X, Huang RR, Doan NB, Scolyer RA, et al. Differential gene expression profiling of primary cutaneous melanoma and sentinel lymph node metastases. Modern Pathology. 2012;25(6):828.

[205] Turajlic S, Furney SJ, Lambros MB, Mitsopoulos C, Kozarewa I, Geyer FC, et al. Whole genome sequencing of matched primary and metastatic acral melanomas. Genome research. 2012;22(2):196–207.

[206] Bastian BC, Olshen AB, LeBoit PE, Pinkel D. Classifying melanocytic tumors based on DNA copy number changes. The American journal of pathology. 2003;163(5):1765–1770.

[207] Raskin L, Ludgate M, Iyer RK, Ackley TE, Bradford CR, Johnson TM, et al. Copy number variations and clinical outcome in atypical spitz tumors. The American journal of surgical pathology. 2011;35(2):243–252.

[208] Montagnani V, Benelli M, Apollo A, Pescucci C, Licastro D, Urso C, et al. Thin and thick primary cutaneous melanomas reveal distinct patterns of somatic copy number alterations. Oncotarget. 2016;7(21):30365.

[209] Wang B, Qu XL, Chen Y. Identification of the potential prognostic genes of human melanoma. Journal of cellular physiology. 2019;234(6):9810–9815.

[210] Nithya S, Radhika T, Jeddy N. Loricrin–an overview. Journal of oral and maxillofacial pathology: JOMFP. 2015;19(1):64.

[211] Korsching E, Packeisen J, Agelopoulos K, Eisenacher M, Voss R, Isola J, et al. Cytogenetic alterations and cytokeratin expression patterns in breast cancer: integrating a new model of breast differentiation into cytogenetic pathways of breast carcinogenesis. Laboratory investigation. 2002;82(11):1525.

[212] Korsching E, Jeffrey SS, Meinerz W, Decker T, Boecker W, Buerger H. Basal carcinoma of the breast revisited: an old entity with new interpretations. Journal of clinical pathology. 2008;61(5):553–560.

[213] Boecker W, van Horn L, Stenman G, Stürken C, Schumacher U, Loening T, et al. Spatially correlated phenotyping reveals K5-positive luminal progenitor cells and p63-K5/14-positive stem cell-like cells in human breast epithelium. Laboratory Investigation. 2018;98(8):1065.

[214] Jones JCR, Green KJ. Intermediate filament plasma membrane interactions. Current opinion in cell biology. 1991;3(1):127–132.

[215] Seltmann K, Roth W, Kröger C, Loschke F, Lederer M, Hüttelmaier S, et al. Keratins mediate localization of hemidesmosomes and repress cell motility. Journal of Investigative Dermatology. 2013;133(1):181–190.

[216] García-Silva S, Benito-Martín A, Sánchez-Redondo S, Hernández-Barranco A, Ximénez-Embún P, Nogués L, et al. Use of extracellular vesicles from lymphatic drainage as surrogate markers of melanoma progression and BRAFV600E mutation. Journal of Experimental Medicine. 2019;p. jem—-20181522.

[217] Alowami S, Qing G, Emberley E, Snell L, Watson PH. Psoriasin (S100A7) expression is altered during skin tumorigenesis. BMC dermatology. 2003;3(1):1.

[218] Liu S, Howell P, Ren S, Fodstad O, Riker AI. The 14-3-3σ gene promoter is methylated in both human melanocytes and melanoma. BMC cancer. 2009;9(1):162.

[219] Schultz J, Ibrahim SM, Vera J, Kunz M. 14-3-3σ gene silencing during melanoma progression and its role in cell cycle control and cellular senescence. Molecular cancer. 2009;8(1):53.

[220] Winnepenninckx V, Lazar V, Michiels S, Dessen P, Stas M, Alonso SR, et al. Gene expression profiling of primary cutaneous melanoma and clinical outcome. Journal of the National Cancer Institute. 2006;98(7):472–482.

[221] Gajos-Michniewicz A, Czyz M. Role of miRNAs in Melanoma Metastasis. Cancers. 2019;11(3):326.

[222] Tang M, Rai K. Computational Analysis of Epigenetic Modifications in Melanoma. In: Computational Epigenetics and Diseases. Elsevier; 2019. p. 327–342.

A. Appendix

Table A.1: Highlighted information of the 26 somatic CNV-driven DEGs. Different statistical parameters are specified for each gene: LFC to be considered DEG and being remarked within each iGC substudy, PV and FDR. Total number and percentage of cases are included for each substudy, divided into loss, normal and gain. Statistically significant cases are highlighted for loss (red) and gain (green). Additionally, those significant cases being included in another substudies are also remarked for loss (brown) and gain (purple).

Gene Symbol	Highlighted iGC Substudy	LFC DEG	PV	FDR	LFC iGC	CNV Type	MEN PRIMEL			MEN METMEL			WOMEN PRIMEL			WOMEN METMEL			MEN			WOMEN			PRIMEL			METMEL			ALL		
							L (%)	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*	*L (%)*	*N (%)*	*G (%)*
S100A7	*MEN*	*-4,54*	*5,31E-07*	*1,83E-04*	-2,74	L	0 (0)	18 (75)	6 (25)	2 (11)	9 (50)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	2 (5)	27 (64)	13 (31)	2 (6)	21 (68)	8 (26)	1 (2)	32 (76)	9 (21)	3 (10)	16 (52)	12 (39)	4 (5)	48 (66)	21 (29)
SFN	*MEN – WOMEN*	*-3,26*	*3,69E-04 // 2,07E-05*	*2,66E-02 // 8,9E-03*	-1,54 // -2,21	L / G	1 (4)	19 (79)	4 (17)	3 (17)	12 (67)	3 (17)	1 (6)	17 (94)	0 (0)	1 (8)	9 (69)	3 (23)	4 (10)	31 (74)	7 (17)	2 (6)	26 (84)	3 (10)	2 (5)	36 (86)	4 (10)	4 (13)	21 (68)	6 (19)	6 (8)	57 (78)	10 (14)
KRT14	*MEN*	*-5,15*	*6,58E-04*	*3,97E-02*	-2,53	L	0 (0)	19 (79)	5 (21)	3 (17)	10 (56)	5 (28)	1 (6)	16 (89)	1 (6)	1 (8)	10 (77)	2 (15)	3 (7)	29 (69)	10 (24)	2 (6)	26 (84)	3 (10)	1 (2)	35 (83)	6 (14)	4 (13)	20 (65)	7 (23)	5 (7)	55 (75)	13 (18)
KRT16	*MEN*	*-3,98*	*6,58E-06*	*1,23E-03*	-2,19	L	0 (0)	19 (79)	5 (21)	3 (17)	10 (56)	5 (28)	1 (6)	16 (89)	1 (6)	1 (8)	10 (77)	2 (15)	3 (7)	29 (69)	10 (24)	2 (6)	26 (84)	3 (10)	1 (2)	35 (83)	6 (14)	4 (13)	20 (65)	7 (23)	5 (7)	55 (75)	13 (18)
KRT10	*MEN*	*-2,04*	*6,77E-05*	*7,52E-03*	-0,78	L	0 (0)	19 (79)	5 (21)	3 (17)	10 (56)	5 (28)	1 (6)	16 (89)	1 (6)	1 (8)	10 (77)	2 (15)	3 (7)	29 (69)	10 (24)	2 (6)	26 (84)	3 (10)	1 (2)	35 (83)	6 (14)	4 (13)	20 (65)	7 (23)	5 (7)	55 (75)	13 (18)
KLK11	*MEN*	*-1,8*	*6,04E-05*	*6,76E-03*	-1,24	L	0 (0)	21 (88)	3 (12)	2 (11)	12 (67)	4 (22)	1 (6)	17 (94)	0 (0)	1 (8)	8 (62)	4 (31)	2 (5)	33 (79)	7 (17)	2 (6)	25 (81)	4 (13)	1 (2)	38 (90)	3 (7)	3 (10)	20 (65)	8 (26)	4 (5)	58 (79)	11 (15)
S100A7A	*MEN*	*-1,43*	*3,12E-06*	*7,27E-04*	-1,35	L	0 (0)	18 (75)	6 (25)	2 (11)	9 (50)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	2 (5)	27 (64)	13 (31)	2 (6)	21 (68)	8 (26)	1 (2)	32 (76)	9 (21)	3 (10)	16 (52)	12 (39)	4 (5)	48 (66)	21 (29)
CLCA2	*MEN*	*-1,87*	*2,96E-05*	*3,94E-03*	-1,3	L	1 (4)	20 (83)	3 (12)	4 (22)	9 (50)	5 (28)	1 (6)	16 (89)	1 (6)	1 (8)	12 (92)	0 (0)	5 (12)	29 (69)	8 (19)	2 (6)	28 (90)	1 (3)	2 (5)	36 (86)	4 (10)	5 (16)	21 (68)	5 (16)	7 (10)	57 (78)	9 (12)
SPRR2G	*MEN*	*-1,48*	*2,44E-05*	*3,41E-03*	-1,7	L	0 (0)	19 (79)	5 (21)	3 (17)	8 (44)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	3 (7)	27 (64)	12 (29)	2 (6)	21 (68)	8 (26)	1 (2)	33 (79)	8 (79)	4 (13)	15 (48)	12 (39)	5 (7)	48 (66)	20 (27)
SPRR1B	*MEN*	*-4,12*	*6,08E-04*	*3,78E-02*	-2,18	L	0 (0)	19 (79)	5 (21)	3 (17)	8 (44)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	3 (7)	27 (64)	12 (29)	2 (6)	21 (68)	8 (26)	1 (2)	33 (79)	8 (79)	4 (13)	15 (48)	12 (39)	5 (7)	48 (66)	20 (27)
LOR	*MEN*	*-3,03*	*5,49E-04*	*3,58E-02*	-1,43	L	0 (0)	19 (79)	5 (21)	3 (17)	8 (44)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	3 (7)	27 (64)	12 (29)	2 (6)	21 (68)	8 (26)	1 (2)	33 (79)	8 (79)	4 (13)	15 (48)	12 (39)	5 (7)	48 (66)	20 (27)
DSG3	*ALL*	*-2,18*	*1,11E-04*	*3,06E-02*	-1,65	L	1 (4)	16 (67)	7 (29)	3 (17)	7 (39)	8 (44)	0 (0)	16 (89)	2 (11)	1 (8)	11 (85)	1 (8)	4 (10)	23 (55)	15 (36)	1 (3)	27 (87)	3 (10)	1 (2)	32 (76)	9 (21)	4 (13)	18 (58)	9 (29)	5 (7)	50 (68)	18 (25)
KRT6A	*MEN*	*-4,53*	*1,59E-04*	*1,49E-02*	-1,66	L	0 (0)	18 (75)	6 (25)	2 (11)	9 (50)	7 (39)	1 (6)	15 (83)	2 (11)	1 (8)	7 (54)	5 (38)	2 (5)	27 (64)	13 (31)	2 (6)	22 (71)	7 (23)	1 (2)	33 (79)	8 (79)	3 (10)	16 (52)	12 (39)	4 (5)	49 (67)	20 (27)
KLK7	*WOMEN*	*-2,09*	*1,16E-05*	*5,48E-03*	-2,05	G	0 (0)	21 (88)	3 (12)	2 (11)	12 (67)	4 (22)	1 (6)	17 (94)	0 (0)	1 (8)	8 (62)	4 (31)	2 (5)	33 (79)	7 (17)	2 (6)	25 (81)	4 (13)	1 (2)	38 (90)	3 (7)	3 (10)	20 (65)	8 (26)	4 (5)	58 (79)	11 (15)
SPRR1A	*MEN*	*-3,04*	*7,12E-07*	*2,41E-04*	-1,85	L	0 (0)	19 (79)	5 (21)	3 (17)	8 (44)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	3 (7)	27 (64)	12 (29)	2 (6)	21 (68)	8 (26)	1 (2)	33 (79)	8 (79)	4 (13)	15 (48)	12 (39)	5 (7)	48 (66)	20 (27)
LCE3D	*MEN*	*-1,7*	*2,19E-04*	*1,87E-02*	-1,09	L	0 (0)	19 (79)	5 (21)	2 (11)	9 (50)	7 (39)	1 (6)	14 (78)	3 (17)	0 (0)	8 (62)	5 (38)	2 (5)	28 (67)	12 (29)	1 (3)	22 (71)	8 (26)	1 (2)	33 (79)	8 (79)	2 (6)	17 (55)	12 (39)	3 (4)	50 (68)	20 (27)
TRIM29	*MEN*	*-2,12*	*4,94E-05*	*5,74E-03*	-1,65	L	1 (4)	18 (75)	5 (21)	4 (22)	8 (44)	6 (33)	1 (6)	15 (83)	2 (11)	2 (15)	8 (62)	3 (23)	5 (12)	26 (62)	11(26)	3 (10)	23 (74)	5 (16)	2 (5)	33 (79)	7 (17)	6 (19)	16 (52)	9 (29)	8 (11)	49 (67)	16 (22)
IVL	*MEN*	*-2,26*	*1,73E-07*	*7,83E-05*	-1,87	L	0 (0)	19 (79)	5 (21)	3 (17)	8 (44)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	3 (7)	27 (64)	12 (29)	2 (6)	21 (68)	8 (26)	1 (2)	33 (79)	8 (79)	4 (13)	15 (48)	12 (39)	5 (7)	48 (66)	20 (27)
PKP3	*MEN*	*-1,41*	*1,75E-04*	*1,63E-02*	-1,36	L	0 (0)	19 (79)	5 (21)	2 (11)	9 (50)	7 (39)	1 (6)	16 (89)	1 (6)	1 (8)	8 (62)	4 (31)	2 (5)	28 (67)	12 (29)	2 (6)	24 (77)	5 (16)	1 (2)	35 (83)	6 (14)	3 (10)	17 (55)	11 (35)	4 (5)	52 (71)	17 (23)
KRT6B	*MEN*	*-4,12*	*1,64E-08*	*1,00E-05*	-2,55	L	0 (0)	18 (75)	6 (25)	2 (11)	9 (50)	7 (39)	1 (6)	15 (83)	2 (11)	1 (8)	7 (54)	5 (38)	2 (5)	27 (64)	13 (31)	2 (6)	22 (71)	7 (23)	1 (2)	33 (79)	8 (79)	3 (10)	16 (52)	12 (39)	4 (5)	49 (67)	20 (27)
KRT5	*MEN*	*-4,12*	*4,51E-06*	*9,40E-04*	-1,89	L	0 (0)	18 (75)	6 (25)	2 (11)	9 (50)	7 (39)	1 (6)	15 (83)	2 (11)	1 (8)	7 (54)	5 (38)	2 (5)	27 (64)	13 (31)	2 (6)	22 (71)	7 (23)	1 (2)	33 (79)	8 (79)	3 (10)	16 (52)	12 (39)	4 (5)	49 (67)	20 (27)
SERPINB4	*PRIMEL*	*-2,07*	*6,75E-05*	*4,11E-03*	-1,57	L	1 (4)	15 (63)	8 (33)	3 (17)	6 (33)	9 (50)	1 (6)	15 (83)	2 (11)	1 (8)	11 (85)	1 (8)	4 (10)	21 (50)	17 (40)	2 (6)	26 (84)	3 (10)	2 (5)	30 (71)	10 (24)	4 (13)	17 (55)	10 (32)	6 (8)	47 (64)	20 (27)
CST6	*MEN – WOMEN METMEL*	*-1,89*	*3,42E-04 // 3,55E-04*	*2,52E-02 // 4,62E-02*	-0,88 // 1,19	L / L	1 (4)	19 (79)	4 (17)	1 (6)	9 (50)	8 (44)	0 (0)	17 (94)	1 (6)	2 (15)	8 (62)	3 (23)	2 (5)	28 (67)	12 (29)	2 (6)	25 (81)	4 (13)	1 (2)	36 (86)	5 (12)	3 (10)	17 (55)	11 (35)	4 (5)	53 (73)	16 (22)
S100A2	*MEN*	*-3,01*	*4,19E-05*	*5,06E-03*	-1,51	L	0 (0)	18 (75)	6 (25)	2 (11)	9 (50)	7 (39)	1 (6)	14 (78)	3 (17)	1 (8)	7 (54)	5 (38)	2 (5)	27 (64)	13 (31)	2 (6)	21 (68)	8 (26)	1 (2)	32 (76)	9 (21)	3 (10)	16 (52)	12 (39)	4 (5)	48 (66)	21 (29)
LYPD3	*MEN*	*-2,31*	*3,23E-04*	*2,45E-02*	-0,95	L	1 (4)	20 (83)	3 (12)	2 (11)	13 (72)	3 (17)	1 (6)	16 (89)	1 (6)	1 (8)	9 (69)	3 (23)	3 (7)	33 (79)	6 (14)	2 (6)	25 (81)	4 (13)	2 (5)	36 (86)	4 (10)	3 (10)	22 (71)	6 (19)	5 (7)	58 (79)	10 (14)
DEFB1	*MEN METMEL*	*-1,4*	*8,52E-05*	*2,22E-02*	-1,22	L	1 (4)	19 (79)	4 (17)	2 (11)	10 (56)	6 (33)	1 (6)	16 (89)	1 (6)	1 (8)	11 (85)	1 (8)	3 (7)	29 (69)	10 (24)	2 (6)	27 (87)	2 (6)	2 (5)	35 (83)	5 (12)	3 (10)	21 (68)	7 (23)	5 (7)	56 (77)	12 (16)

LEGEND					
	ABBREVIATIONS	LFC (log2 fold change), PV (p-value), FDR (false discovery rate), CNV (copy number variation), L (loss), N (normal), G (gain), PRIMEL (primary melanoma), METMEL (metastatic melanoma), ALL (full patient cohort)			
	COLORS		Statistically significant for LOSS	Statistically significant for GAIN	Substudy including cases remarked for LOSS	Substudy including cases remarked for GAIN

Table A.2: Extended information of the functional enrichment analysis of the 26 somatic CNV-driven DEGs. Gene set included in each GO term is specified.

Ontology	GO ID	GO Term	# Genes (%)	PV	BONF	BH	FDR	Genes
BP	GO:0008544	epidermis development	13 (50,0)	9,49E-16	6,16E-13	6,16E-13	1,48E-12	LOR, KLK7, CST6, KRT5, S100A7, SPRR1A, KRT16, SPRR1B, KRT14, LCE3D, KRT10, SFN, IVL
	GO:0009888	tissue development	15 (57,7)	5,84E-09	3,60E-06	5,99E-07	8,66E-06	LOR, KLK7, KRT6A, KRT6B, S100A7, LCE3D, KRT10, SFN, KRT5, CST6, SPRR1A, KRT16, SPRR1B, KRT14, IVL
	GO:0009913	epidermal cell differentiation	9 (34,6)	5,83E-11	3,59E-08	8,98E-09	8,64E-08	LOR, S100A7, SPRR1A, KRT16, SPRR1B, LCE3D, KRT10, SFN, IVL
	GO:0030216	keratinocyte differentiation	9 (34,6)	2,25E-12	1,39E-09	6,93E-10	3,33E-09	LOR, S100A7, SPRR1A, KRT16, SPRR1B, LCE3D, KRT10, SFN, IVL
	GO:0030855	epithelial cell differentiation	10 (38,5)	2,70E-08	1,66E-05	2,37E-06	4,00E-05	LOR, S100A7, SPRR1A, KRT16, SPRR1B, KRT14, LCE3D, KRT10, SFN, IVL
	GO:0031424	keratinization	7 (26,9)	4,96E-11	3,06E-08	1,02E-08	7,36E-08	LOR, SPRR1A, KRT16, SPRR1B, LCE3D, SFN, IVL
	GO:0043588	skin development	9 (34,6)	5,53E-10	3,40E-07	6,81E-08	8,19E-07	LOR, S100A7, SPRR1A, KRT16, SPRR1B, LCE3D, KRT10, SFN, IVL
	GO:0060429	epithelium development	11 (42,3)	5,17E-07	3,18E-04	3,98E-05	7,66E-04	LOR, KRT6A, S100A7, SPRR1A, KRT16, SPRR1B, KRT14, LCE3D, KRT10, SFN, IVL
CC	GO:0001533	cornified envelope	6 (23,1)	2,66E-09	2,32E-07	2,32E-07	2,88E-06	LOR, CST6, SPRR1A, SPRR1B, LCE3D, IVL
	GO:0005576	extracellular region	18 (69,2)	1,44E-06	1,25E-04	1,79E-05	1,55E-03	KLK7, KRT6A, CLCA2, KRT6B, LYPD3, S100A7, KRT10, SFN, KRT5, CST6, DSG3, KRT16, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
	GO:0005882	intermediate filament	6 (23,1)	5,01E-06	4,36E-04	4,36E-05	5,42E-03	KRT6A, KRT6B, KRT5, KRT16, KRT14, KRT10
	GO:0031982	vesicle	16 (61,5)	4,45E-06	3,87E-04	4,30E-05	4,81E-03	KLK7, KRT6A, KRT6B, S100A7, KRT10, SFN, KRT5, CST6, KRT16, DSG3, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
	GO:0031988	membrane-bounded vesicle	16 (61,5)	2,67E-06	2,32E-04	2,90E-05	2,88E-03	KLK7, KRT6A, KRT6B, S100A7, KRT10, SFN, KRT5, CST6, KRT16, DSG3, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
	GO:0043230	extracellular organelle	15 (57,7)	9,15E-07	7,96E-05	1,33E-05	9,89E-04	KRT6A, KRT6B, S100A7, KRT10, SFN, KRT5, CST6, KRT16, DSG3, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
	GO:0044421	extracellular region part	17 (65,4)	8,94E-07	7,78E-05	1,94E-05	9,66E-04	KLK7, KRT6A, KRT6B, LYPD3, S100A7, KRT10, SFN, KRT5, CST6, DSG3, KRT16, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
	GO:0045111	intermediate filament cytoskeleton	7 (26,9)	5,09E-07	4,43E-05	2,21E-05	5,50E-04	KRT6A, KRT6B, KRT5, KRT16, TRIM29, KRT14, KRT10
	GO:0070062	extracellular exosome	15 (57,7)	8,56E-07	7,45E-05	2,48E-05	9,25E-04	KRT6A, KRT6B, S100A7, KRT10, SFN, KRT5, CST6, KRT16, DSG3, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
	GO:1903561	extracellular vesicle	15 (57,7)	9,11E-07	7,93E-05	1,59E-05	9,85E-04	KRT6A, KRT6B, S100A7, KRT10, SFN, KRT5, CST6, KRT16, DSG3, SPRR1B, KRT14, KLK11, SERPINB4, DEFB1, IVL
MF	**GO:0005198**	structural molecule activity	11 (42,3)	4,67E-08	4,25E-06	4,25E-06	5,09E-05	LOR, KRT6A, KRT6B, KRT5, SPRR1A, KRT16, SPRR1B, KRT14, LCE3D, KRT10, IVL
	GO:0005200	structural constituent of cytoskeleton	6 (23,1)	4,13E-07	3,76E-05	1,88E-05	4,50E-04	LOR, KRT6A, KRT6B, KRT5, KRT16, KRT14

LEGEND	GO (Gene Ontology), BP (Biological Process), CC (Cellular Component), MF (Molecular Function), PV (P-Value), BONF (Bonferroni Score), BH (Benjamini-Hochberg Score), FDR (False Discovery Rate)

Table A.3: Selected transcriptomic sample summary for this study. Webdata repository, dataset accession ID and sample type are specified for each dataset: (A) pre-selection (all collected samples) and (B) final selection (all considered samples).

(A) PRE-SELECTION

Technology	Repository	Accesion ID	NSK	NEV	PRIMEL	METMEL	Total samples – Collected		
Microarray	GEO	2503	5				5	351	596
	GEO	3189	7	18			25		
	GEO	7553	4		16		20		
	GEO	13355	64				64		
	GEO	14905	21				21		
	GEO	15605	16		45	2	63		
	GEO	32407	10				10		
	GEO	32924	8				8		
	GEO	42677	10				10		
	GEO	46517	7	9	31	41	88		
	GEO	52471	13				13		
	GEO	53223	6	12			18		
	GEO	82105				6	6		
RNA-seq	AE	5678	4				4	245	
	GEO	54456	82				82		
	GEO	58375	8				8		
	GEO	98394		27	51		78		
	GDC	CA			42	31	73		
			265	66	185	80			
			596						

(B) FINAL SELECTION

Technology	Repository	Accesion ID	NSK	NEV	PRIMEL	METMEL	Total samples – Selected		
MICRO	GEO	2503	4				4	289	532
	GEO	3189	6	16			22		
	GEO	7553	4		14		18		
	GEO	13355	61				61		
	GEO	14905	16				16		
	GEO	15605	13		30	2	45		
	GEO	32407	10				10		
	GEO	32924	7				7		
	GEO	42677	9				9		
	GEO	46517	6	6	25	31	68		
	GEO	52471	10				10		
	GEO	53223	5	8			13		
	GEO	82105				6	6		
RNA-seq	AE	5678	4				4	243	
	GEO	54456	80				80		
	GEO	58375	8				8		
	GEO	98394		27	51		78		
	GDC	CA			42	31	73		
			243	57	162	70			
			532						

LEGEND	GEO (NCBI GEO webdata repository), AE (ArrayExpress webdata repository), GDC (Genomic Data Commons webdata repository), CA (Controlled Access), NSK (normal skin), NEV (nevus), PRIMEL (primary melanoma), METMEL (metastatic melanoma)

Table A.4: Clinical data related to GDC patient cohort selected for this study.

Case	Sample Type	Gender	A	B	C	D	E	F	G	H	I	J	K
A17Z	MET	M	1951	2008	C44.5	iib	20970	D	8720/3	263.0	C44.5	C44.5	–
A180	MET	M	1932	2008	C44.6	iii	25523	D	8720/3	2889.0	C44.6	C44.6	–
A184	MET	M	1931	2008	C44.6	ib	26447	D	8720/3	2073.0	C44.6	C44.6	–
A185	MET	F	1953	2008	C44.5	iiic	20235	D	8720/3	151.0	C44.5	C44.5	–
A196	PT	F	1943	–	C44.7	iic	23533	A	8744/3	–	C44.7	C44.7	1785.0
A19E	MET	F	1972	2009	C44.5	ib	13297	D	8720/3	396.0	C44.5	C44.5	–
A19F	MET	M	1921	2005	C44.6	NR	30176	D	8720/3	802.0	C44.6	C44.6	–
A19G	MET	F	1939	–	C44.5	NR	17820	A	8720/3	–	C44.5	C44.5	9188.0
A19M	MET	M	1970	2011	C44.5	ib	13300	D	8720/3	1857.0	C44.5	C44.5	–
A19O	MET	M	1950	–	C44.6	iiib	20740	D	8720/3	–	C44.6	C44.6	–
A19Q	MET	F	1964	2005	C44.5	NR	13648	D	8720/3	1548.0	C44.5	C44.5	–
A19S	MET	F	1927	–	C44.5	NR	29603	A	8720/3	–	C44.5	C44.5	1505.0
A19T	PT	M	1955	2006	C44.7	iv	18732	D	8744/3	270.0	C44.7	C44.7	–
A19T	MET	M	1955	2006	C44.7	iv	18732	D	8744/3	270.0	C44.7	C44.7	–
A20B	MET	F	1936	–	C44.7	ii	24445	A	8720/3	–	C44.7	C44.7	4070.0
A20C	MET	M	1938	2009	C44.5	0	21660	D	8720/3	4601.0	C44.5	C44.5	–
A20H	MET	M	1939	2009	C44.5	i	20476	D	8720/3	5118.0	C44.5	C44.5	–
A20I	MET	M	1929	2009	C44.7	iv	28883	D	8720/3	412.0	C44.7	C44.7	–
A24C	PT	M	1955	–	C44.5	NR	20539	A	8720/3	–	C44.5	C44.5	632.0
A24D	PT	M	1939	–	C44.6	iiib	26439	A	8720/3	–	C44.6	C44.6	645.0
A262	MET	F	1954	–	C44.5	NR	17239	A	8720/3	–	C44.5	C44.5	4255.0
A263	PT	M	1980	2005	C44.5	iv	8952	D	8720/3	467.0	C44.5	C44.5	–
A264	MET	M	1937	2006	C44.5	NR	22254	D	8720/3	3587.0	C44.5	C44.5	–

A26A	MET	F	1943	2008	C44.5	iiia	23288	D	8720/3	988.0	C44.5	C44.5	–
A26D	MET	F	1936	–	C44.5	iic	26494	D	8720/3	1460.0	C44.5	C44.5	1204.0
A29B	MET	M	1937	–	C44.7	iib	24618	D	8720/3	2588.0	C44.7	C44.7	2452.0
A29C	MET	M	1987	–	C44.5	ib	7510	D	8720/3	2402.0	C44.5	C44.5	1455.0
A29V	MET	M	1923	2010	C44.5	iiic	31092	D	8720/3	787.0	C44.5	C44.5	–
A2NB	PT	M	1953	–	C44.5	iiib	20938	D	8720/3	857.0	C44.5	C44.5	486.0
A2ND	MET	F	1947	2005	C44.5	iiic	20867	D	8720/3	710.0	C44.5	C44.5	–
A3J5	MET	M	1929	2003	C44.5	iii	26226	D	8720/3	1124.0	C44.5	C44.5	–
A3R1	PT	M	1942	–	C44.6	iic	25271	A	8721/3	–	C44.6	C44.6	685.0
A3Z4	PT	M	1958	–	C44.6	iiic	19749	D	8720/3	519.0	C44.6	C44.6	119.0
A41B	PT	F	1936	–	C44.6	iic	27998	A	8721/3	–	C44.6	C44.6	291.0
A430	PT	M	1929	–	C44.5	iic	30344	A	8721/3	–	C44.5	C44.5	-2.0
A44O	PT	M	1943	–	C44.5	iib	25537	A	8720/3	–	C44.5	C44.5	81.0
A44P	PT	F	1954	–	C44.5	iic	21519	A	8720/3	–	C44.5	C44.5	741.0
A4IS	PT	M	1935	–	C44.5	iib	28315	A	8720/3	–	C44.5	C44.5	774.0
A4OY	PT	F	1947	–	C44.7	iiib	23773	A	8720/3	–	C44.7	C44.7	977.0
A4OZ	PT	F	1971	–	C44.5	iiic	15337	A	8720/3	–	C44.5	C44.5	620.0
A4P0	PT	M	1930	–	C44.5	iic	30225	D	8730/3	326.0	C44.5	C44.5	-2.0
A4Z2	PT	M	1962	–	C44.5	iiic	18462	D	8721/3	190.0	C44.5	C44.5	93.0
A4Z3	PT	F	1939	–	C44.6	iiic	26812	A	8721/3	–	C44.6	C44.6	505.0
A51B	PT	M	1959	–	C44.6	iic	19645	A	8720/3	–	C44.6	C44.6	931.0
A550	PT	F	1937	–	C44.5	iic	27556	D	8720/3	264.0	C44.5	C44.5	6.0
A551	PT	F	1934	–	C44.7	iiic	28671	A	8730/3	–	C44.7	C44.7	590.0
A553	PT	M	1950	–	C44.5	iic	22972	A	8720/3	–	C44.5	C44.5	226.0
A5DX	PT	M	1941	–	C44.5	iiic	26056	A	8720/3	–	C44.5	C44.5	640.0
A5EO	PT	M	1947	–	C44.5	iic	23940	A	8720/3	–	C44.5	C44.5	703.0
A5EP	PT	F	1937	–	C44.5	iiic	27622	A	8720/3	–	C44.5	C44.5	335.0

A5EQ	PT	M	1949	–	C44.5	iic	23353	A	8720/3	–	C44.5	C44.5	323.0
A5ER	PT	M	1949	–	C44.5	iic	23199	A	8720/3	–	C44.5	C44.5	327.0
A5ES	PT	F	1936	–	C44.5	iic	28092	A	8720/3	–	C44.5	C44.5	490.0
A5GO	MET	M	1940	–	C44.6	ii	22287	A	8720/3	–	C44.6	C44.6	4195.0
A5SE	PT	M	1939	–	C44.5	iib	26965	D	8721/3	401.0	C44.5	C44.5	0.0
A5UM	PT	F	1964	–	C44.5	iic	17867	A	8720/3	–	C44.5	C44.5	779.0
A728	PT	F	1959	–	C44.6	iiib	19936	A	8743/3	–	C44.6	C44.6	583.0
A769	PT	M	1971	2012	C44.7	iic	14304	D	8721/3	1070.0	C44.7	C44.7	–
A85I	PT	M	1947	–	C44.7	iic	24408	A	8720/3	–	C44.7	C44.7	362.0
A85J	PT	F	1947	–	C44.7	iib	24263	A	8720/3	–	C44.7	C44.7	360.0
A8K4	PT	M	1928	–	C44.6	iib	31308	A	8721/3	–	C44.6	C44.6	614.0
A8RT	MET	F	1970	–	C44.6	iib	15342	A	8720/3	–	C44.6	C44.6	839.0
A8ZY	MET	M	1948	–	C44.5	iia	22676	D	8720/3	1506.0	C44.5	C44.5	1330.0
A97M	PT	M	1947	–	C44.5	iic	24228	A	8720/3	–	C44.5	C44.5	414.0
A9QB	MET	F	1960	–	C44.5	NR	9058	A	8720/3	–	C44.5	C44.5	11217.0
A9VZ	PT	F	1923	–	C44.7	ii	32872	A	8720/3	–	C44.7	C44.7	11.0
A9W2	PT	M	1932	–	C44.6	i	29836	A	8720/3	–	C44.6	C44.6	417.0
A9WB	MET	M	1941	2013	C44.5	ia	26176	D	8720/3	518.0	C44.5	C44.5	–
AAA4	MET	F	1956	–	C44.7	iiic	20708	A	8720/3	–	C44.7	C44.7	760.0
AAOU	PT	F	1940	–	C44.5	iic	26750	A	8730/3	–	C44.5	C44.5	476.0
AAZV	PT	F	1957	–	C44.5	ii	20545	A	8720/3	–	C44.5	C44.5	412.0
AAZW	PT	F	1951	–	C44.7	ii	22778	D	8720/3	393.0	C44.7	C44.7	18.0
AAZY	PT	F	1937	–	C44.5	iiic	27938	A	8720/3	–	C44.5	C44.5	405.0

Terms: PT (Primary tumor), MET (Metastatic), M (Male), F (Female), NR (Not reported), A (Alive), D (Dead). Meaning of columns: A (Year of birth), B (Year of death), C (Primary diagnosis), D (Tumor stage), E (Age at diagnosis), F (Vital status), G (Morphology), H (Days)to death, I (Tissue or organ of origin), J (Site of resection or biopsy), K (Days to last follow up).

Table A.5: Segmentation window size analysis on the selected full patient cohort.

Internal ID	BP/R Global	C	BP 24(G)	Mapped	Unmapped Reads	24(N)	BP/R 24	BP/R 22
A17Z_DNA_BDN	36,037	2,109	3088269832	85537708	160509	85698217	47,958	38,040
A17Z_DNA_MET	40,670	1,869	3088269832	75783061	152590	75935651	58,625	45,317
A180_DNA_BDN	28,645	2,653	3088269832	107619532	193268	107812800	36,917	30,316
A180_DNA_MET	21,627	3,514	3088269832	142680867	114791	142795658	28,881	23,938
A184_DNA_BDN	31,842	2,387	3088269832	96760034	226354	96986388	43,766	34,428
A184_DNA_MET	38,742	1,962	3088269832	79583950	129612	79713562	53,417	41,410
A185_DNA_BDN	39,474	1,925	3088269832	78112309	123236	78235545	255,725	43,266
A185_DNA_MET	52,867	1,438	3088269832	58239286	176608	58415894	374,847	59,929
A196_DNA_BDN	38,434	1,977	3088269832	80169988	182129	80352117	225,962	42,456
A196_DNA_PT	38,307	1,984	3088269832	80482501	137452	80619953	233,981	41,977
A19E_DNA_BDN	42,901	1,772	3088269832	71804178	182276	71986454	264,323	47,604
A19E_DNA_MET	53,198	1,429	3088269832	57930401	122030	58052431	441,749	63,332
A19F_DNA_BDN	51,417	1,478	3088269832	59979780	83350	60063130	69,782	55,455
A19F_DNA_MET	38,513	1,973	3088269832	80086732	100400	80187132	114,300	42,759
A19G_DNA_BDN	50,586	1,502	3088269832	60860375	189906	61050281	277,372	55,894
A19G_DNA_MET	43,889	1,732	3088269832	70278875	87323	70366198	300,489	49,654
A19M_DNA_BDN	28,996	2,621	3088269832	106393730	114815	106508545	36,942	30,868
A19M_DNA_MET	35,924	2,116	3088269832	85834660	132930	85967590	58,593	38,022
A19O_DNA_BDN	45,524	1,669	3088269832	67719966	117873	67837839	59,491	48,192
A19O_DNA_MET	50,887	1,494	3088269832	60574432	114361	60688793	68,620	55,378
A19Q_DNA_BDN	37,721	2,015	3088269832	81716305	154614	81870919	212,448	41,802
A19Q_DNA_MET	50,357	1,509	3088269832	61195427	131984	61327411	331,469	57,331
A19S_DNA_BDN	38,790	1,959	3088269832	79409931	204410	79614341	249,502	43,005
A19S_DNA_MET	51,436	1,478	3088269832	59942105	99297	60041402	384,148	57,101
A19T_DNA_BDN	32,385	2,347	3088269832	95294865	66941	95361806	42,368	34,260
A19T_DNA_MET	35,444	2,144	3088269832	87042902	88178	87131080	79,752	38,679
A19T_DNA_PT	32,721	2,323	3088269832	94228482	152514	94380996	71,140	41,953

A20B_DNA_BDN	38,656	1,966	3088269832	79661821	228573	79890394	198,977	42,135
A20B_DNA_MET	39,209	1,938	3088269832	78472433	291406	78763839	228,332	46,817
A20C_DNA_BDN	36,585	2,077	3088269832	84266399	146030	84412429	52,830	38,603
A20C_DNA_MET	50,363	1,509	3088269832	61120173	199945	61320118	70,772	56,111
A20H_DNA_BDN	38,201	1,989	3088269832	80620539	221875	80842414	50,101	40,508
A20H_DNA_MET	32,470	2,341	3088269832	94990364	120212	95110576	129,058	36,875
A20I_DNA_BDN	39,605	1,919	3088269832	77807807	169729	77977536	52,478	42,013
A20I_DNA_MET	56,215	1,352	3088269832	54811712	125383	54937095	84,264	66,835
A24C_DNA_BDN	54,788	1,387	3088269832	56222654	145191	56367845	75,287	59,255
A24C_DNA_PT	36,136	2,103	3088269832	85294171	167890	85462061	46,173	37,357
A24D_DNA_BDN	51,392	1,479	3088269832	59934970	157529	60092499	68,277	54,754
A24D_DNA_PT	34,003	2,235	3088269832	90712350	112072	90824422	86,457	37,165
A262_DNA_BDN	38,695	1,964	3088269832	79633360	177197	79810557	260,548	43,016
A262_DNA_MET	38,069	1,996	3088269832	81011301	111381	81122682	223,328	43,925
A263_DNA_BDN	47,133	1,612	3088269832	65381009	141840	65522849	64,834	51,425
A263_DNA_PT	33,081	2,297	3088269832	93260513	94978	93355491	44,915	35,863
A264_DNA_BDN	45,725	1,662	3088269832	67393346	147032	67540378	59,682	48,425
A264_DNA_MET	38,001	2,000	3088269832	81023029	244119	81267148	50,473	42,853
A26A_DNA_BDN	32,039	2,372	3088269832	96214145	175665	96389810	186,617	34,801
A26A_DNA_MET	57,970	1,311	3088269832	53073941	199899	53273840	287,033	65,035
A26D_DNA_BDN	46,361	1,639	3088269832	66497779	115985	66613764	308,242	51,101
A26D_DNA_MET	38,926	1,952	3088269832	79177134	159162	79336296	253,248	46,522
A29B_DNA_BDN	38,324	1,983	3088269832	80360684	223330	80584014	51,231	40,880
A29B_DNA_MET	53,856	1,411	3088269832	57201290	141753	57343043	219,433	59,960
A29C_DNA_BDN	17,915	4,242	3088269832	172073519	313808	172387327	23,118	18,903
A29C_DNA_MET	37,462	2,029	3088269832	82150829	286031	82436860	51,729	42,466
A29V_DNA_BDN	40,313	1,885	3088269832	76375001	231806	76606807	58,113	42,864
A29V_DNA_MET	54,645	1,391	3088269832	56324191	190915	56515106	76,063	60,659
A2NB_DNA_BDN	57,006	1,333	3088269832	54012810	161734	54174544	78,692	61,955
A2NB_DNA_PT	36,431	2,086	3088269832	84662585	108665	84771250	55,988	41,520
A2ND_DNA_BDN	33,624	2,260	3088269832	91753686	93448	91847134	217,283	36,345

A2ND_DNA_MET	30,076	2,527	3088269832	102523629	159742	102683371	204,444	33,374
A3J5_DNA_BDN	30,682	2,477	3088269832	100501214	153728	100654942	37,204	31,585
A3J5_DNA_MET	31,515	2,412	3088269832	97819042	173663	97992705	47,814	35,402
A3R1_DNA_BDN	27,509	2,763	3088269832	112128388	136068	112264456	35,138	29,053
A3R1_DNA_PT	13,253	5,734	3088269832	232819178	197145	233016323	17,904	14,537
A3Z4_DNA_BDN	23,438	3,243	3088269832	131589807	175324	131765131	31,650	25,372
A3Z4_DNA_PT	29,827	2,548	3088269832	103422832	116196	103539028	39,550	32,690
A41B_DNA_BDN	40,690	1,868	3088269832	75720435	176676	75897111	245,703	43,959
A41B_DNA_PT	41,487	1,832	3088269832	74308815	131502	74440317	280,108	47,082
A430_DNA_BDN	35,929	2,115	3088269832	85800963	154613	85955576	49,527	38,158
A430_DNA_PT	34,069	2,231	3088269832	90491632	156913	90648545	69,220	37,488
A44O_DNA_BDN	50,009	1,520	3088269832	61692180	61816	61753996	67,505	53,770
A44O_DNA_PT	59,932	1,268	3088269832	51464381	65042	51529423	79,080	64,333
A44P_DNA_BDN	14,368	5,290	3088269832	214810999	132630	214943629	107,652	15,850
A44P_DNA_PT	55,716	1,364	3088269832	55373928	54654	55428582	424,630	61,692
A4IS_DNA_BDN	56,560	1,344	3088269832	54538251	63625	54601876	75,540	60,865
A4IS_DNA_PT	49,484	1,536	3088269832	62311712	98080	62409792	77,514	53,541
A4OY_DNA_BDN	60,035	1,266	3088269832	51370891	70144	51441035	429,359	65,629
A4OY_DNA_PT	50,349	1,509	3088269832	61254093	83075	61337168	364,314	55,792
A4OZ_DNA_BDN	58,902	1,290	3088269832	52372654	57622	52430276	409,527	64,233
A4OZ_DNA_PT	88,129	0,862	3088269832	35003219	39525	35042744	592,376	96,592
A4P0_DNA_BDN	61,779	1,230	3088269832	49941872	46984	49988856	83,733	66,295
A4P0_DNA_PT	50,251	1,512	3088269832	61379355	77820	61457175	64,797	54,371
A4Z2_DNA_BDN	47,491	1,600	3088269832	64933805	94734	65028539	61,946	50,440
A4Z2_DNA_PT	39,820	1,909	3088269832	77434272	120895	77555167	51,971	43,254
A4Z3_DNA_BDN	60,036	1,266	3088269832	51386771	53272	51440043	461,110	65,905
A4Z3_DNA_PT	59,149	1,285	3088269832	52094082	117986	52212068	470,515	64,898
A51B_DNA_BDN	35,361	2,149	3088269832	87226970	109345	87336315	46,318	37,911
A51B_DNA_PT	37,325	2,036	3088269832	82647579	92356	82739935	90,015	40,744
A550_DNA_BDN	36,465	2,084	3088269832	84604051	87217	84691268	262,188	40,001
A550_DNA_PT	30,383	2,501	3088269832	101481975	161519	101643494	220,552	33,715

A551_DNA_BDN	46,626	1,630	3088269832	66162454	71974	66234428	334,379	51,155
A551_DNA_PT	45,355	1,676	3088269832	68011562	79556	68091118	334,042	49,801
A553_DNA_BDN	34,883	2,179	3088269832	88437674	95305	88532979	45,799	37,513
A553_DNA_PT	32,693	2,325	3088269832	94357544	106036	94463580	58,998	37,446
A5DX_DNA_BDN	39,978	1,901	3088269832	77151885	98020	77249905	52,914	43,060
A5DX_DNA_PT	35,098	2,165	3088269832	87874449	114781	87989230	56,898	38,373
A5EO_DNA_BDN	39,029	1,947	3088269832	79016102	111604	79127706	52,903	42,334
A5EO_DNA_PT	30,641	2,480	3088269832	100608471	179100	100787571	41,296	33,633
A5EP_DNA_BDN	32,398	2,346	3088269832	95177160	144367	95321527	220,288	36,110
A5EP_DNA_PT	28,476	2,669	3088269832	108307814	145499	108453313	206,618	33,828
A5EQ_DNA_BDN	40,550	1,874	3088269832	76076828	83320	76160148	54,804	43,861
A5EQ_DNA_PT	45,070	1,686	3088269832	68413448	107978	68521426	59,639	48,838
A5ER_DNA_BDN	33,318	2,281	3088269832	92607244	84166	92691410	43,614	35,751
A5ER_DNA_PT	40,782	1,864	3088269832	75628726	98161	75726887	87,990	44,340
A5ES_DNA_BDN	34,609	2,196	3088269832	89142350	91699	89234049	244,922	37,950
A5ES_DNA_PT	46,082	1,649	3088269832	66936946	79489	67016435	286,209	50,698
A5GO_DNA_BDN	34,205	2,222	3088269832	90196248	91932	90288180	45,251	36,902
A5GO_DNA_MET	43,285	1,756	3088269832	71292882	54653	71347535	63,557	47,682
A5SE_DNA_BDN	30,857	2,463	3088269832	98411251	1671109	100082360	40,564	32,562
A5SE_DNA_PT	36,621	2,075	3088269832	82954566	1375892	84330458	50,071	40,731
A5UM_DNA_BDN	33,403	2,275	3088269832	90980949	1473334	92454283	77,727	36,130
A5UM_DNA_PT	35,632	2,133	3088269832	85252482	1419440	86671922	78,746	38,159
A728_DNA_BDN	16,711	4,548	3088269832	184317900	486045	184803945	32,429	17,051
A728_DNA_PT	16,826	4,517	3088269832	183099259	447379	183546638	32,440	17,595
A769_DNA_BDN	17,526	4,336	3088269832	175727193	480965	176208158	19,842	17,618
A769_DNA_PT	15,712	4,837	3088269832	195967323	581689	196549012	18,518	16,317
A85I_DNA_BDN	39,887	1,905	3088269832	77310629	113987	77424616	65,494	41,979
A85I_DNA_PT	37,646	2,019	3088269832	81927281	107033	82034314	52,800	39,855
A85J_DNA_BDN	42,179	1,802	3088269832	73147616	71150	73218766	258,369	45,668
A85J_DNA_PT	34,623	2,195	3088269832	89136649	60191	89196840	244,546	37,512
A8K4_DNA_BDN	36,324	2,092	3088269832	84909042	111315	85020357	46,332	38,425

A8K4_DNA_PT	49,334	1,541	3088269832	62526408	72556	62598964	65,255	54,991
A8RT_DNA_BDN	33,348	2,279	3088269832	92460173	147995	92608168	190,354	36,024
A8RT_DNA_MET	28,212	2,694	3088269832	109299259	166820	109466079	167,128	31,661
A8ZY_DNA_BDN	30,173	2,519	3088269832	102196243	155198	102351441	38,510	31,905
A8ZY_DNA_MET	44,596	1,704	3088269832	69093170	156374	69249544	107,306	54,401
A97M_DNA_BDN	46,414	1,637	3088269832	57559372	8977648	66537020	57,922	48,720
A97M_DNA_PT	29,749	2,555	3088269832	89650260	14159062	103809322	38,871	32,655
A9QB_DNA_BDN	28,188	2,696	3088269832	94607312	14954037	109561349	54,516	29,267
A9QB_DNA_MET	30,697	2,476	3088269832	86909536	13694212	100603748	70,597	33,874
A9VZ_DNA_BDN	26,594	2,858	3088269832	100106613	16018554	116125167	50,650	27,318
A9VZ_DNA_PT	43,457	1,749	3088269832	61414394	9650545	71064939	172,060	46,673
A9W2_DNA_BDN	48,586	1,564	3088269832	63430249	132586	63562835	64,027	51,481
A9W2_DNA_PT	47,187	1,611	3088269832	65312286	135089	65447375	75,539	49,797
A9WB_DNA_BDN	42,901	1,772	3088269832	62186787	9799322	71986109	53,325	44,974
A9WB_DNA_MET	32,438	2,343	3088269832	82096198	13109651	95205849	41,542	35,473
AAA4_DNA_BDN	33,359	2,278	3088269832	79873342	12703597	92576939	68,368	34,460
AAA4_DNA_MET	33,324	2,281	3088269832	80005824	12667983	92673807	167,621	36,576
AAOU_DNA_BDN	45,484	1,671	3088269832	67807445	90367	67897812	281,418	49,352
AAOU_DNA_PT	45,279	1,678	3088269832	68082810	121909	68204719	271,375	49,573
AAZV_DNA_BDN	41,708	1,822	3088269832	73904342	140307	74044649	255,104	45,128
AAZV_DNA_PT	48,396	1,570	3088269832	63730117	82836	63812953	321,844	52,942
AAZW_DNA_BDN	44,720	1,699	3088269832	68967495	90247	69057742	292,673	48,554
AAZW_DNA_PT	48,028	1,582	3088269832	64155241	146276	64301517	305,372	54,412
AAZY_DNA_BDN	55,641	1,366	3088269832	55371325	131980	55503305	347,804	60,111
AAZY_DNA_PT	42,596	1,784	3088269832	72236413	265068	72501481	248,397	46,294
SUM (MEAN)	39,986	**2,079**	3088269832	83439527,2	1038425,1	84477952,2	143,1	43,7
SUM (STD)	10,947	0,753	0,000	30432756,0	3214363,6	30614138,7	123,7	12,3

B. Curriculum Vitae

Personal Information

Juan Manuel Gálvez Gómez
March 20, 1986
Huétor - Tájar, Granada (Spain)

Education

2016 - 2019	PhD studies at CASIP group, Department of Computer Architecture and Computer Technology, University of Granada, Spain
2015 - 2016	Master's Degree in Data Science (Postgraduate Course), University of Granada, Spain
2014 - 2015	Collaborator Researcher at Department of Computer Architecture and Computer Technology, University of Granada, Spain
2010 - 2013	M.Sc. Electrical Engineering at Faculty of Sciences, University of Granada, Spain
2004 - 2010	M.Sc. Telecommunication Engineering at High Technical School of Informatics and Telecommunication Engineerings, University of Granada, Spain
2000 - 2004	Secondary and High School at La Salle, Córdoba, Spain
1998 - 2000	Secondary School at La Salle, Astorga, León, Spain
1994 - 1998	Primary School at CEIP Los Canos, Guadiaro, Cádiz, Spain
1991 - 1994	Primary School at CEIP Gloria Fuertes, Guadiaro, Cádiz, Spain

Research/Work Experience

2016 - 2109	PhD Position (Project of Excellence, P12-TIC-2082) at Department of Computer Architecture and Computer Technology, University of Granada, Spain
2018 - 2018	Visiting Research at Complex and Cancer Diseases Group, Institute of Bioinformatics, Muenster, Germany
2017 - 2017	Visiting Research at Complex and Cancer Diseases Group, Institute of Bioinformatics, Muenster, Germany

C. List of Publications

International Journals (SCI-Indexed)

Gálvez JM, Castillo D, Herrera LJ, Korsching E, Makalowski W, Ortuño FM, Rojas I. **Supporting clinical decisions - Determining biomarkers driven by somatic copy number variations bein responsible for the progression of cutaneous melanoma**. *Application of Bioinformatics in Cancers* (under review)

Gálvez JM, Castillo D, Herrera LJ, Valenzuela O, Caba O, Prados JC, Ortuño FM, Rojas I. **Towards improving skin cancer diagnosis by integrating microarray and RNA-seq datasets**. *Journal of Biomedical and Health Informatics* (revision process)

Castillo D, Gálvez JM, Herrera LJ, Rojas F, Valenzuela O, Caba O, Prados JC, Rojas I. **Leukemia multiclass assessment and classification from microarray and RNA-seq technologies integration at gene expression level**. *PLoS one* 14(2):e0212127 (2019).

Gálvez JM, Castillo D, Herrera LJ, San Román B, Valenzuela O, Ortuño FM, Rojas I. **Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series**. *PLoS one* 13(5):e0196836 (2018).

Castillo D, Gálvez JM, Herrera LJ, San Román B, Rojas F, Rojas I. **Integration of RNA-seq data with heterogeneous microarray data for breast cancer profiling**. *BMC Bioinformatics*. 18(1):506 (2017).

Baños O, Gálvez JM, Damas M, Pomares H, Rojas I. **Window size impact in human activity recognition**. *Sensors*. 14(4):6474-99 (2014).

International Conferences

González S, Castillo D, Gálvez JM, Rojas I, Herrera LJ. **Feature selection and assessment of lung cancer sub-types by applying predictive models**. *International Work-Conference on Artificial Neural Networks (IWANN)* (2019) (Accepted).

Baños O, Gálvez JM, Damas M, Guillén A, Herrera LJ, Pomares H, Rojas I, Villalonga C. **Improving wearable activity recognition via fusion of multiple equally-sized data subwindows**. *International Work-Conference on Artificial Neural Networks (IWANN)* (2019) (Accepted).

Castillo D, Gálvez JM, Herrera LJ, Rojas I. **Breast cancer microarray and RNASeq data integration applied to classification**. *International Work-Conference on Artificial Neural Networks* (pp. 123-131). Springer, Cham (2017).

Valenzuela O, Castillo D, Gálvez JM, Rojas I. **Development of intelligent systems for the classification and automatic diagnosis of fluoroscopy images corresponding to different skin pathologies**. *International Work-Conference on Bioinformatics and Biomedical Engineering* (2017).

Baños O, Gálvez JM, Damas M, Guillén A, Herrera LJ, Pomares H, Rojas I, Villalonga C, Hong CS, Lee S. **Multiwindow fusion for wearable activity recognition**. *International Work-Conference on Artificial Neural Networks* (pp. 290-297). Springer, Cham (2015).

Baños O, Gálvez JM, Damas M, Guillén A, Herrera LJ, Pomares H, Rojas I.

Evaluating the effects of signal segmentation on activity recognition. *International Work-Conference on Bioinformatics and Biomedical Engineering* (pp. 759-765) (2014).

Proceedings Editions

Editors: Valenzuela O, Castillo D, Gálvez JM, Delgado E, Sáez MJ, Rojas F, Rojas I. **Proceedings of Extended Abstracts**. *International Work-Conference on Bioinformatics and Biomedical Engineering*. Editorial Godel S.L. I.S.B.N.: 978-84-17293-94-9 (2019).

Editors: Castillo D, Gálvez JM, Sáez MJ, Rojas F, Herrera LJ, Rojas I. **Proceedings of Abstracts**. *International Work-Conference on Bioinformatics and Biomedical Engineering*. Editorial Godel S.L. I.S.B.N.: 978-84-17293-38-3 (2018).

Editors: Castillo D, Gálvez JM, Sáez MJ, Rojas F, Herrera LJ, Rojas I. **Proceedings of Extended Abstracts**. *International Work-Conference on Bioinformatics and Biomedical Engineering*. Editorial Godel S.L. I.S.B.N.: 978-84-17293-36-9 (2018).