

**Development of a methodology of computational  
intelligence for robust composite biomarker  
discovery: Targeting breakthrough in the therapeutic  
management of melanoma**



A dissertation submitted to the Department of Digital Systems,  
University of Piraeus in complete fulfilment of the requirements  
for the degree of Doctor of Philosophy

by

**Georgia A. Kontogianni**

Piraeus, June 2019



**«Ανάπτυξη μιας μεθοδολογίας υπολογιστικής νοημοσύνης για την ανακάλυψη σύνθετων εύρωστων βιοδεικτών: Στοχεύοντας στη θεραπευτική αντιμετώπιση του μελανώματος»**

**“Development of a methodology of computational intelligence for robust composite biomarker discovery: Targeting breakthrough in the therapeutic management of melanoma”**

Written by:

Georgia Kontogianni



I. Maglogiannis  
Associate Professor,  
University of Piraeus

Approved by:



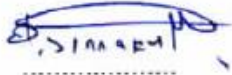
A. Prentza  
Professor,  
University of Piraeus



A. Chatziioannou  
Researcher B,  
NHRF



D. Kyriazis  
Assistant Professor,  
University of Piraeus



M. Philippakis  
Associate Professor,  
University of Piraeus



Q. Telelis  
Assistant Professor,  
University of Piraeus



H. Karanikas  
Lecturer,  
University of Thessaly



# Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Dr. Ilias Maglogiannis and Dr. Aristotelis Chatziioannou for the continuous support during my PhD thesis and research all these years, for their patience, motivation, and immense knowledge. Their guidance helped me throughout these years, and I could not have imagined having better advisors and mentors for my PhD study. They gave me the opportunity to join their team and have access to the laboratory and research facilities. Without their precious support it would not be possible to conduct this research. I would also like to thank Dr. Andriana Prentza for taking part in my committee.

My earnest appreciation also goes to Dr. Olga Papadodima for her insightful comments, encouragement and for the inspiring discussions that we had over the course of this and other projects. Without her valuable help this study would not have been completed.

Special thanks to Georgia Piroti, for assisting with the DNA extraction procedure, and all the members of NHRF's bioinformatics group, e-NIOS, HBD and Datamed, who generously shared their time and experiences for the purposes of this study.

I am also grateful to the people who saw me through this thesis; to all those friends who provided support, talked things over, read and offered comments. Special thanks are due to Maria Socratous for her irreplaceable help these years and all the vital comments, concerning this PhD and more. Also, thanks to Maria Ntzamili and Georgia Kourti for listening to the difficulties that I experienced throughout my PhD (and not only).

Last but not least, I would like to thank my family, Andreas, Koula, Dimitra and Isidoros, whose unconditional love, encouragement and support made this project possible. I dedicate this thesis to you.



# Summary

Cancer is a complex and intricate disease, and the scientific community has been struggling for decades to identify any febleness or rudimentary characteristics to discover effective treatments. Next generation sequencing technologies have eased the way for the systematic discovery of diagnostic biomarkers for cancers and other pathologies. Melanoma continues to be a rare form of skin cancer but causes the majority of skin cancer related deaths. For many years research has focused on the investigation of the pathogenesis leading to melanoma, with the aim of better understanding its complexity and the potential advancement of therapeutic strategies.

In this PhD thesis a computational model for the integrated analysis of multi-source cancer datasets is proposed, using cutaneous melanoma as disease-model, in order to identify robust composite biomarkers that allow the classification between healthy and disease state. Along this road, for the first time primary cutaneous melanoma biopsies from Greek patients were subjected to whole exome sequencing and were analysed in order to derive their mutational profile landscape. Moreover, in the context of big data analytical methodologies, integration of exome sequencing and transcriptomic data was performed, in an attempt to achieve a multi-layered analysis and infer a tentative disease network for primary melanoma pathogenesis, offering deeper insight in the underlying mechanisms affected by melanoma and potentially contributing to the valuable effective epidemiological characterisation of this disease.

This study exhibits a modular and distributed workflow that can integrate heterogeneous, multidimensional (omics, histological images and clinical) data for the multi-angled portrayal and classification of melanoma patients. All the proposed methodologies achieve satisfying performance through the proposed framework. The specific architecture aspires to lower the barrier for the introduction of personalised therapeutic approaches, towards precision medicine.





# Περίληψη

Ο καρκίνος αποτελεί μια πολύπλοκη ασθένεια που έχει απασχολήσει την επιστημονική κοινότητα παγκοσμίως για δεκαετίες, στοχεύοντας στον εντοπισμό τυχόν αδυναμιών ή στοιχειωδών χαρακτηριστικών που θα οδηγήσουν σε αποτελεσματικές θεραπείες. Οι τεχνολογίες αλληλούχισης νέας γενιάς έχουν διευκολύνει τη συστηματική εξεύρεση διαγνωστικών βιοδεικτών για καρκίνους και άλλες παθολογίες. Το μελάνωμα αποτελεί μια σπάνια μορφή καρκίνου του δέρματος, αλλά ευθύνεται για την πλειοψηφία των θανάτων που σχετίζονται με τον συγκεκριμένο τύπο καρκίνου. Για πολλά χρόνια οι αιτίες και η παθογένεια που οδηγούν στο μελάνωμα αποτελούν αντικείμενο έρευνας, με στόχο την καλύτερη κατανόηση της πολυπλοκότητάς του και της πιθανής εξέλιξης των θεραπευτικών στρατηγικών που ακολουθούνται.

Στην παρούσα διδακτορική διατριβή προτείνεται ένα υπολογιστικό μοντέλο για την ολιστική ανάλυση πολλαπλών πηγών καρκινικών δεδομένων, χρησιμοποιώντας το δερματικό μελάνωμα ως ασθένεια-μοντέλο, με στόχο να εντοπιστούν εύρωστοι σύνθετοι βιοδείκτες που επιτρέπουν την ταξινόμηση μεταξύ υγιούς και νοσηρής κατάστασης. Κατά τη διάρκεια αυτής της ανάλυσης, για πρώτη φορά οι πρωτογενείς βιοψίες δερματικού μελανώματος από Έλληνες ασθενείς υποβλήθηκαν σε αλληλούχιση εξωνίων (WES) και αναλύθηκαν προκειμένου να αντληθεί το προφίλ μεταλλάξεών τους. Επιπλέον, στο πλαίσιο ανάλυσης μεθοδολογιών μεγάλων δεδομένων (big data), ολοκληρώθηκε η ενσωμάτωση και σύντηξη των εξωνικών με μεταγραφικά δεδομένα, σε μια προσπάθεια επίτευξης μιας πολύ-επίπεδης ανάλυσης και εξαγωγής ενός γενικευμένου δικτύου ασθένειας για το μελάνωμα, επιτρέποντας την εμβάθυνση της γνώσης των υποκείμενων βιολογικών μηχανισμών που επηρεάζονται και συμβάλλοντας ενδεχομένως στον πολύτιμο αποτελεσματικό επιδημιολογικό χαρακτηρισμό αυτής της ασθένειας.

Αυτή η μελέτη παρέχει μια αρθρωτή και κατανεμημένη ροή εργασίας που μπορεί να ενσωματώνει ετερογενή, πολυδιάστατα δεδομένα (omics, ιστολογικές εικόνες και κλινικά) για τη διεξοδική ταξινόμηση ασθενών με μελάνωμα. Όλες οι προτεινόμενες μεθοδολογίες επιτυγχάνουν ικανοποιητικές επιδόσεις μέσω του προτεινόμενου πλαισίου.

Η συγκεκριμένη αρχιτεκτονική φιλοδοξεί να μειώσει τα εμπόδια για την εισαγωγή εξατομικευμένων θεραπευτικών προσεγγίσεων, οδηγώντας στην ιατρική ακρίβειας.

# Contents

Acknowledgements.....	5
Summary .....	7
Περίληψη .....	9
List of Figures .....	15
List of Tables .....	17
List of Abbreviations .....	18
Chapter 1. Introduction .....	21
1.1 Purpose & Research questions .....	21
1.2 Prologue .....	22
1.3 Biomarkers .....	24
1.4 Machine learning .....	26
1.5 Genomics & Big Data era .....	27
1.6 Skin cancer & Melanoma .....	29
1.7 Biological terminology .....	31
Chapter 2. Related work .....	34
2.1 Skin Imaging & Dermoscopy background .....	34
2.1.1 Basic methods for melanoma detection using a computer-based approach.....	35
2.1.1.1 Evaluation of the basic methods using a computer-based approach.....	40
2.1.2 Diverse techniques for melanoma detection.....	42
2.1.3 Molecular techniques for melanoma detection .....	43
2.1.4 Data Integration .....	43
2.1.5 Discussion.....	46
2.2 Biological background .....	51

2.2.1	Germline susceptibility .....	51
2.2.2	Somatic alterations .....	52
2.2.2.1	Mutation burden and specific signatures in melanoma .....	53
2.2.2.2	Genes bearing causative somatic mutations in melanoma .....	54
2.2.2.3	Affected pathways & gene expression .....	59
2.2.3	FFPE samples.....	60
2.2.4	Discussion.....	61
2.3	Bioinformatics & Cancer genomic data .....	62
2.3.1	Genomic Data Resources .....	62
2.3.2	Bioinformatic cancer genomic data analysis exploiting NGS.....	65
Chapter 3.	Materials & Methods .....	68
3.1	Biological material and molecular analysis.....	71
3.1.1	Melanoma Samples.....	71
3.1.2	DNA Extraction and Exome Sequencing .....	72
3.2	Variant calling & mutational biomarker discovery .....	72
3.2.1	NGS complexity and information.....	82
3.3	Transcriptomics microarray data analysis .....	82
3.4	Machine learning techniques for data integration & classification.....	84
3.4.1	Mutational data .....	84
3.4.2	Imaging data & Integration.....	86
Chapter 4.	Analysis & Results .....	88
4.1	Initial molecular analysis and integration.....	88
4.1.1	Initial molecular analysis.....	89
4.1.2	Pathway analysis.....	90
4.1.3	Discussion.....	97
4.2	Extended molecular analysis.....	97

4.2.1	Sequencing Data Analysis .....	98
4.2.2	Identification of Germline Variation .....	99
4.2.3	Identification of Somatic Coding Mutations .....	101
4.2.3.1	Characterisation of Mutated Genes and Copy Number Variations .....	103
4.2.3.2	Enhanced Pathway Analysis.....	108
4.2.4	Discussion.....	111
4.3	Data integration & Classification .....	112
4.3.1	Molecular data classification system .....	114
4.3.2	Imaging data classification system.....	117
4.3.3	Integrated data classification system .....	119
4.3.4	Discussion.....	122
Chapter 5.	Conclusions .....	123
	General Discussion.....	123
	Challenges .....	125
	Algorithm complexity & Contribution.....	125
	Future work.....	126
	Closing remarks.....	127
	Bibliography .....	129
	Conferences & Publications .....	175
	Presentations & Posters on International Conferences .....	175
	Publications on Conference Papers .....	177
	Publications on International Journals .....	177
	Appendix .....	179
	Tables A1: Patient and melanoma lesion characteristics (extended).....	179
	Tables A2: Quality control for whole-exome sequencing of 18 samples .....	181
	Table A3: Binary tables of patients carrying mutations for MAPK pathway and cell cycle .....	184



# List of Figures

Figure 1: Processing pipeline of the presented research .....	24
Figure 2: (upper) Number of publications per year on Pubmed (until 4th December 2018) using terms 'melanoma' and 'cutaneous melanoma', (lower) Major landmarks concerning the study of melanoma .	53
Figure 3: Comprehensive processing pipeline of the presented research .....	70
Figure 4: Commands to run BWA, specific parameters used are given in brackets .....	74
Figure 5: Commands to run Picard .....	75
Figure 6: Commands to run GATK RealignerTargetCreator, IndelRealigner, Base Recalibrator and PrintReads in consecutive steps .....	77
Figure 7: Commands to run GATK HaplotypeCaller and MuTect2 .....	78
Figure 8: Commands to run MutSigCV.....	79
Figure 9: BioInfoMiner online tool.....	80
Figure 10: Commands to run CNVkit in consecutive steps .....	81
Figure 11: Workflow of analysis for the identification of variance and somatic mutations.....	82
Figure 12: Microarray analysis commands using GEOquery and limma .....	83
Figure 13: Commands to run SMOTE in R.....	85
Figure 14: Commands used to build the classifiers in R .....	85
Figure 15: Commands to create Random Forests classifier and test its accuracy in R.....	86
Figure 16: Hierarchical clustering of the patients, based on the mutational profiles, red-present mutation, blue-absent .....	90
Figure 17: Venn diagram for the significant gene lists from the two analyses.....	94
Figure 18: Bar plot of significant terms with the number of associated genes (x-axis). Terms are ranked using the corrected p-value. The colours of the genes specify their expression fold change, green -on the left- for under-expressed genes and red -on the right- for over-expressed genes, neutral indicating somatic mutation .....	95
Figure 19: Venn diagram for the significant pathway lists from the two distinct analyses, as well as their integration .....	96
Figure 20: Mutation spectrum for each patient. C > T transitions account for 85.6% of the mutations (median rate). .....	103

Figure 21: Common genes between the 73-genes list, the 1571 Network of Cancer Genes (NCG) genes and the >5% mutated genes in melanoma from COSMIC .....	104
Figure 22: Characterisation of genes carrying non-synonymous mutations based on COSMIC data. M-melanoma-associated genes (>20% mutated in COSMIC) and C-cancer-census genes (>5% in melanoma samples) *—denotes genes highlighted by MutSigCV. ....	106
Figure 23: Top Census genes in melanoma from COSMIC database and the type of mutation found in all patients .....	108
Figure 24: Statistically significant biological processes with the corresponding number of genes found as mutated in at least one patient .....	109
Figure 25: 30 top-prioritised mutated genes, D-probably damaging and P-possibly damaging mutation, as predicted by PolyPhen2. ....	111
Figure 26: Data assortment for classification .....	115
Figure 27: Results for the Molecular Classifier. ROC curve for rf-Random Forests, glmnet-Gaussian linear model, gbm-Stochastic Gradient Boosting, c50-Decision Trees C5.0, lda-Linear Discriminant Analysis, svm-Support Vector Machines, knn-k Nearest Neighbours, glm- Logistic Regression.....	117
Figure 28: Results for the Imaging Classifier. ROC, Sensitivity and Specificity for rf-Random Forests, glmnet-Gaussian linear model, gbm-Stochastic Gradient Boosting, c50-Decision Trees C5.0, svm-Support Vector Machines, logistic- Logistic Regression, knn-k Nearest Neighbours.....	119
Figure 29: ROC curves for the 3 Random Forests classifiers, immo-integrated features (82) classifier, molecular features (51) classifier, im-imaging features (31) classifier. The integrated feature classifier performs best, with a mean AUC of 0.9432.....	120
Figure 30: Results for the Integrated RF Classifiers, ROC, Sensitivity and Specificity for immo-82 features, gr10-gain ratio top-10, gr20-gain ratio top-20, ig10-information gain top-10 and ig20- information gain top-20 features .....	122



# List of Tables

Table 1: Comparative table for the surveyed classification systems.....	46
Table 2: Somatic variance calling tools.....	66
Table 3: Driver mutation calling tools.....	67
Table 4: Patient and melanoma lesion characteristics, Border R-regular, I-irregular .....	71
Table 5: Number of somatic mutations, missense/nonsense mutations, and unique genes affected per patient.....	89
Table 6: Table of the significant biological processes influenced by the mutated genes. Enrichment represents the ratio of the number of genes in the input list annotated with a GO term to the total number of genes annotated to this specific term, Hypergeometric and Corrected p-values represent the statistic score used for ranking the terms, given by BioInfoMiner .....	91
Table 7: Table of the significant biological processes influenced by the differentially expressed genes. Enrichment represents the ratio of the number of genes in the input list annotated with a GO term to the total number of genes annotated to this specific term, Hypergeometric and Corrected p-values represent the statistic score used for ranking the terms, given by BioInfoMiner .....	92
Table 8: Sequencing characteristics for all the samples. ....	99
Table 9: Germline single nucleotide polymorphisms (SNPs) putatively associated with melanoma, based on genome-wide association studies (GWAS) and MelGene databases. ....	100
Table 10: Somatic mutation characteristics for each patient.....	101
Table 11: Characterisation of the somatic mutations for each patient.....	102
Table 12: Recurrent mutations, based on COSMIC (v84) (n>15).....	106
Table 13: Top 20 mutated genes, given by MutSigCV .....	107
Table 14: Genes used as features for the molecular classifier .....	114
Table 15: Metrics used as features for the imaging classifier.....	118
Table 16: Top 20 features selected using information gain and information gain ratio, gene features in bold .....	121

# List of Abbreviations

ABCDE- Asymmetry, Border, Colour, Diameter, Evolution,

AUC- area under the curve,

BAM- binary alignment format,

BED- browser extensible data,

BMR- background mutation rate,

BWA- Burrows-Wheeler aligner,

CBIR- content-based image retrieval,

CM- cutaneous melanoma,

CMY- Cyan, Magenta, Yellow,

CNV- copy number variation,

COSMIC- Catalogue of Somatic Mutations in Cancer,

dbNSFP- database of Non-synonymous SNPs' Functional Predictions

DDBJ- DNA Databank of Japan

EHR- Electronic health record

ELM- epiluminescence microscopy or dermoscopy,

EMBL- European Molecular Biology Laboratory

FFPE- formalin fixed paraffin embedded,

FISH- fluorescence in situ hybridisation,

GATK- Genome Analysis Toolkit,

GDC- Genomic Data Commons

GDP/GTP- guanosine diphosphate/ guanosine-5'-triphosphate,

GEO- Gene Expression Omnibus,

GO- Gene Ontology,  
GWAS- Genome Wide Association Study,  
HIS- Hue, Intensity, Saturation,  
HLS- Hue, Lightness, Saturation,  
HSV- Hue, Saturation, Value,  
Indel- insertion/ deletion,  
kNN- k nearest neighbours  
MAF- Mutation Annotation Format  
MAPK- mitogen activating protein kinase  
Mb- mega-base,  
MB- mega-byte,  
ML- machine learning  
MRI- magnetic resonance imaging,  
NCBI- National Centre for Biotechnology Information,  
NCG- Network of Cancer Genes,  
NCI- National Cancer Institute  
NGS- next generation sequencing,  
PI3K- phosphoinositide 3 kinase,  
RGB- red, green, blue,  
RF- random forests,  
ROC- receiver operating characteristic,  
SAM- sequence alignment map  
SBFS- sequential backward floating selection,  
SFFS- sequential forward floating selection,

SNP/SNV- single nucleotide polymorphism/variant,

SVM- support vector machine,

TCGA- The Cancer Genome Atlas,

TPM- Transcripts per million,

UTR- un-translated region,

UV/UVR- ultra-violet /radiation,

vcf- variant call format,

WGS- whole genome sequencing,

WES-whole exome sequencing,

YIQ- Y component for luma information In phase Quadrature,

Nucleotides: A-adenine, C-cytosine, G-guanine, T-thymine

Amino-Acids: E- glutamic acid, V- valine

# Chapter 1. Introduction

## 1.1 Purpose & Research questions

The purpose of this research is to find the best combination of molecular, histological and clinical features to create an automated integrative processing system for the detection of cutaneous malignant melanoma.

The questions that will be addressed are the following:

1. Can the integration of diverse molecular data offer new knowledge on melanoma?
2. Which composite biomarkers differentiate between malignant melanoma and benign nevus?
3. Which feature combination presents better performance on an automated processing system for the detection of melanoma?
4. Which classification method performs better on a given feature combination?

The motivation is to attain a breakthrough in the translational biomedical research, focusing in the pervasive study of such an important disease as melanoma. The goal is to design and implement a layered analytical framework, which can integrate high-volume molecular omic data with imaging data of skin lesions (dermoscopy).

The output of this study is a novel method which combines molecular components with image processing for the detection of melanoma with high accuracy. Another goal is the reduction of the number of features that are used for classification. An important aspect is that the presented analysis scheme can be utilised for diverse classification problems, i.e. cutaneous melanoma subtype classification, given that enough training input data becomes available. Finally, the ultimate goal is the production of an approach that is based on high throughput technology that is cost-effective at the same time.

## 1.2 Prologue

The following thesis is a multi-disciplinary analysis concerning cutaneous melanoma (CM), or melanoma of the skin. Various biological and computational aspects are explored to understand the deeper mechanisms leading to the disease's manifestation and offer new insights in the distinction between healthy and disease state.

A concept that needs to be discussed is the difference between data integration and fusion. The two terms are often used interchangeably, still in many applications a difference can be observed. Fusion refers to the process of combining input data into a common representational arrangement, probably transforming the original input, while integration refers to the use of multi-source information to assist in a particular task (Mitchell, 2012). In this thesis, we mostly use the term data integration, to avoid any misinterpretation.

Several chapters in this dissertation contain material that has been published previously. These do not essentially represent the final published form and might have been edited slightly. At the end, the list of publications related to this study, finalised or under preparation, is presented. In some cases, supplementary data has not been included but can be found online in the corresponding publication.

As declared, this is a multi-disciplinary research analysis and the various biological and computational aspects need to be explored. Partly this research focuses on unearthing important associations between molecular, histological and clinical features concerning the characterisation of melanoma. This, requested separate analysis of the different levels of data and feature selection, followed by statistical analysis or machine learning techniques. Then, emphasis was given on examining which combination of features improves classification performance upon various classifiers. The structure of this thesis is as follows:

- The remains of chapter 1 discuss several basic concepts used throughout this analysis and a general introduction on the subject is given.

- In chapter 2, the related work of the field is presented. In the first subsection of the chapter, the literature concerning the imaging field is presented, while in the second subsection the biological background is provided.
- Chapter 3 includes the data acquisition and methodology used for analysis. More specifically, the bioinformatic and machine learning framework is given, along with specific commands used for the analysis. The software and hardware information are also denoted.
- In chapter 4, a number of case studies that utilise the proposed parts of the framework, along with their results, is presented. The first subsection presents the initial molecular analysis and integration of mutational and transcriptomic data, the second subsection gives an extended description of the molecular analysis and the resulting molecular biomarkers, while the third and final subsection describes the classification systems built and the integration of imaging and molecular data.
- Chapter 5 includes a final discussion of the outputs of this dissertation and the conclusion.

Answers to the research questions are given throughout this dissertation, predominantly in chapter 4, where the analysis is performed, and the results are presented. Firstly, a broad molecular network implicated in CM is given at section 4.1, where integration of diverse molecular data takes place, elucidating the important mechanisms involved in this type of cancer; some formerly concealed by the statistical cut-offs. Secondly, the composite biomarkers that can discriminate melanoma from healthy nevus are given as lists (Tables Table 14 and Table 15), after molecular and imaging data integration and successful classification at section 4.3. For the third question, the composite features are statistically evaluated, again in section 4.3. For the fourth question, multiple algorithms are evaluated to achieve the highest accuracy possible. Figure 1 presents the basic outline and processing pipeline of this research. The basis of this study is the experimental and bioinformatic analysis of exome sequencing data, deriving from new patients. Through integration with transcriptomic data (top rectangle of Figure 1, pink shade), a broad molecular network implicated in CM is given,

elucidating the important mechanisms involved in this type of cancer. Through integration with skin imaging characteristics and machine learning modalities, the creation of classification systems (see bottom rectangle of Figure 1, green and grey shades) for the layered scheme is acquired.

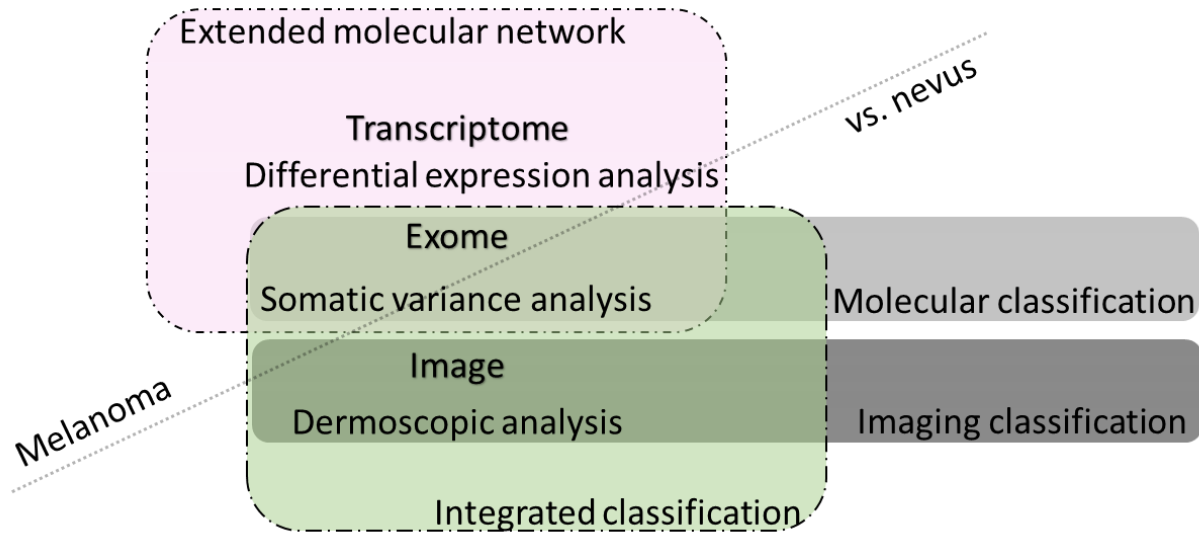


Figure 1: Processing pipeline of the presented research

### 1.3 Biomarkers

The term “biomarker” refers to a wide subgroup of signs indicating a medical state that can be measured accurately in a reproducible manner. These signs differ from medical symptoms, which are limited to the warnings of disease perceived by patients themselves (Strimbu and Tavel, 2010). A biomarker is “any substance, structure or process that can be measured in the body or its products and influence or predict the incidence or outcome of disease” (Organization and Safety, 2001). A more updated definition for biomarker is “a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types of biomarkers” (FDA-NIH Biomarker Working Group, 2016).



Medical imaging can be a source of diagnostic, predictive, prognostic, and monitoring biomarkers, and has already been used extensively in many applications (Weaver and Leung, 2017). In clinical oncology, imaging can act as an operative cancer diagnostic tool, offering non-invasive and low-cost handling. By accumulating innovative and auspicious means to precision medicine, the customisation of cancer care can be achieved (Parmar et al., 2015).

Biomarkers' role in the drug development process and the general biomedical research initiative is crucial. The association of an easily measurable biological aspect with a clinical result is vibrant to escalating the medicinal actions against disease, deepening the knowledge of human physiology (Strimbu and Tavel, 2010). An important aspect of biomarker discovery is procedure stability, in terms of sample variation or robustness of selection processes. Biomarker robustness can influence any succeeding biological validation and reinforce the confidence of a selection scheme.

Currently in biomedicine, next generation sequencing (NGS) technologies are invaluable for biomarker discovery through applications of computational biology. "NGS, massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionised genomic research. Using NGS an entire human genome can be sequenced within a single day. In contrast, the previous Sanger sequencing technology, used to decipher the human genome, required over a decade to deliver the final draft" (Behjati and Tarpey, 2013). NGS applications include whole-genome sequencing (WGS), whole-exome sequencing (WES) and RNA sequencing (RNA-seq), among others. This study mostly focuses on WES, though other methodologies are discussed, as well. Specially in the case of cancer, high-dimensional technologies have become state-of-the-art for the comparison between healthy and disease state (Abeel et al., 2010).

Combining different levels of information concerning a problem can improve the total knowledge on that problem and assist in the effort of finding a solution (Lanckriet et al., 2004). The combination of high-dimensional, multimodal, multivariate models and data sets can provide a more comprehensive view of the response to disease. The ascending products of such combination can be described with the term "composite

biomarkers”. Through coalescing different levels of information and utilising diverse approaches, researchers can generate cancer-specific or patient-specific tactics dedicated to Precision Medicine (Lambin et al., 2013).

## 1.4 Machine learning

Machine learning (ML) is the field of artificial intelligence that relies on statistical and mathematical algorithms to detect patterns in complex data sets, with the prospect of inferring knowledge on these and additional similar data sets. Computers can “learn” from former instances and decipher the patterns to allow classification of new data. This ability is highly compatible with biomedical applications, especially in the case of complex NGS technologies. To this end, ML has been frequently utilised in cancer research (Cruz and Wishart, 2006). Various ML techniques have been used in cancer detection and diagnosis for over 30 years (Cicchetti, 1992; Maclin et al., 1991; Simes, 1985).

From the ML perspective, selection of biomarkers can be viewed as a feature selection task for classification, where the objective is to identify the set of features that correctly differentiate the samples. This set of features can serve as a signature for the disease under investigation, given that the results allow reproducibility and are biologically validated (Abeel et al., 2010).

Current advances in high-performance computing in the fields of artificial intelligence and deep learning, have increased the accessibility of bulky annotated data sets, moving forward to the development of innovative frameworks, ensuing the unparalleled acceleration of these fields. Potentially any level of data -molecular, imaging, clinical- can be viewed through the artificial intelligence standpoint, offering critical insights and aiding therapeutics and diagnostics (Zhavoronkov, 2018).

## 1.5 Genomics & Big Data era

As general terminology "omics" sciences include genomics for DNA variants, transcriptomics for RNA analysis, proteomics for proteins, and metabolomics for metabolic products (Romero et al., 2006). Genomics refers to the study of genes and their functions, and associated techniques. Genomics addresses all genes aiming to investigate their relationship to categorise shared impact on the development of an organism ("WHO | WHO definitions of genetics and genomics", n.d.). Hence, genomics comprises of studies that are conducted at the level of the genome ("genomics | Learn Science at Scitable", n.d.).

Utilising high-throughput sequencing technologies, genomics has shed light to new prospects for the characterisation of diseases and drug discovery. The potential of genomics revolution is endless, allowing powerful openings with social, economic, and cultural impact, apart from the obvious effect on science. The research community has and will continue to embrace genomics technologies, gaining extraordinary power for novelties and modernisations, to confront any challenge menacing the human kind (Jimenez-Sanchez, 2015).

"Precision medicine" aims to accurately categorise patients, sharing a mutual biological root for a disease, into specific subgroups to improve treatment and outcome. To succeed, it requires data efficacy from various bases and in vast volumes, including data gathering, supervision and analytics. A required step is including omic information for each patient into the electronic health record (EHR) (Wu et al., 2017). The widespread implementation of EHR for the entire population delivers a foundation for healthcare efficiency and wellbeing. The arrival of high-throughput omic analyses, like NGS, has led to the fast accretion of omic data (Hillestad et al., 2005). Omic analysis regularly aims at discovering biomarkers using high-throughput technologies, through the extraction of molecular and disease profiles, identifying the substantial genes and networks altered in a given state by statistical models and validation. EHR holds patients' disease information with the prospect of predicting imminent outcome, based on individual and population characteristics. The incorporation of omic knowledge on EHR can enable efficient

therapeutics and empower disease diagnosis, prognosis, and treatment. Up until this day, disease variability, specially in the case of cancer, is the highest challenge, so accurate classification into subtypes necessitates methodical big data models for validation and reliable group allocation (Wu et al., 2017). To keep up with the continuous fruition of technologies implemented in EHR, decision support systems must be offered in the same pace (Andreu Perez et al., 2015).

The decreasing cost of data generation has led science to the Big Data era. The ever-growing technological developments permit the detailed profiling of biological systems, at the same time demanding the constant improvement of biological/clinical data mining and analysis tools (Greene et al., 2016).

The uncovering of hidden outlines, correlations, and other perceptions has been made possible through the exploratory investigation of large-scale data sets. Implementation of additional technological advancements in EHR can allow patient-centred precision medicine in clinical practice (He et al., 2017). The cumulative amount of medical imaging data increased the challenge of systematising, mining, and information extraction from large-scale medical imaging datasets. Novel modalities are constantly emerging, while others are attaining universal acceptance as state-of-the-art (Andreu Perez et al., 2015).

Big Data structures offer the incorporation and manipulation of genomic information in wide-ranging EHRs, delivering a viable prospect of developing operative tactics for variant discovery, personalised medicine and patient stratification (He et al., 2017). Still, variant discovery for an individual genome remains a complex procedure and requires high computational power, for population and medical genomics (Stephens et al., 2015).

Big Data refers to data sets possessing the five Vs, i.e. Volume, Velocity, Variety, Verification/Veracity, and Value (Huang et al., 2015). Big data in medicine describes datasets that are generated quickly, in high levels and are difficult to interpret due to variability and heterogeneity (Andreu Perez et al., 2015). This describes data sets of

immense volume and high complexity that require special processing methodologies and are subjected to technological developments (He et al., 2017).

Among the Big Data domains, genomics is the most demanding, bearing in mind the computational strains required for data acquisition, long-time storage, data distribution, and, of course, analysis. Any novel approach needs to take these under consideration, though it is highly unlikely that a sole development or technology can solve genomics' complexity (Stephens et al., 2015).

Big Data developments on various fields, like medical informatics, bioinformatics, imaging and sensors, will have an excessive effect on forthcoming clinical research (Andreu Perez et al., 2015).

## 1.6 Skin cancer & Melanoma

Predominantly in Dermatology, skin cancers are the world's most common cancer. Currently one in every three cancers diagnosed is a skin cancer, and according to the Skin Cancer Foundation Statistics ("The Skin Cancer Foundation", n.d.), three out of ten Caucasians will develop skin cancer during their lifetime. The most common skin cancer is basal cell carcinoma, which is rarely deadly since it generally does not metastasise. In contrast, melanoma is a rare type of skin cancer but is considered among the most lethal forms of cancers. Non-melanoma skin cancers rarely spread to other parts of the body, but melanomas are considered a metastatic type of cancer. A characteristic of all skin cancers is that their incidence has increased in the last decades. Specifically, in the case of cutaneous melanoma, its incidence rates in Caucasian populations have risen faster than those of any other malignant entity over the last 30 years. Melanoma's incidence has been increasing since the mid-60s in most fair skinned populations and is predicted to continue increasing for at least two more decades ("The Skin Cancer Foundation", n.d.).

Melanoma is a malignant tumour that originates from melanocytes; the cells specialised to produce melanin pigment. Melanocytes derive from the neural crest, a

transient embryonic structure, consisting of highly migratory pluripotent cells, which give rise to a number of different cell types (Uong and Zon, 2010). During development, melanocyte progenitors migrate, differentiate and colonise the skin- epidermis and hair follicles, the uvea of the eye and mucous membranes throughout the body. Consequently, melanoma can arise at all these sites, leading to phenotypically, histologically, clinically and genetically diverse types of disease. In Caucasian populations, the most common type of melanoma is Cutaneous Melanoma (CM), originating from the epidermal melanocytes of non-glabrous skin. Among skin cancers, melanoma is the most aggressive, and although it accounts for less than 5% of skin cancer incidence, it is responsible for the majority of related deaths (Nikolaou and Stratigos, 2014). In this dissertation, discussion primarily focuses on CM, as there are many differences in the genetic background implicated in different types of melanoma, such as mucosal or uveal melanoma. A distinct melanoma subtype, often referred to as a subtype of CM, is acral melanoma, occurring on glabrous (nonhair-bearing) acral skin of palms, soles and nail beds, which is not further discussed, as it is not part of the analysis of this thesis.

CM development is a complex multi-factorial process, arising through multiple etiologic pathways and involving the interplay of genetic and environmental risk factors. Among them, the most well-established risk factors are exposure to ultraviolet (UV) radiation, family history, and phenotypic traits carrying a strong genetic component- including hair and eye colour, and the number of common and atypical melanocytic nevi on the body (Nikolaou and Stratigos, 2014).

During the last decades, a continuous increase of CM frequency rates has been observed in Caucasian populations worldwide, making CM the cancer with the most rapidly increasing occurrence. CM incidence varies significantly between populations from different geographic regions, with Australia and New Zealand presenting the highest incidence rates worldwide. In Europe, rates are lower, but still have shown a three-fold to five-fold increase during this time period (Garbe and Leiter, 2009). CM occurrence differs substantially between European countries, with Switzerland showing the highest rate and Greece belonging to the group of low-incidence countries (“EUCAN”, n.d.; Forsea et al., 2012).

CM diagnosis requires visual detection through dermoscopy, followed by lesion excision, biopsy and histopathological confirmation of malignancy. Regarding the classification of melanoma, clinical morphologists have traditionally divided the disease into several subgroups, including superficial spreading melanoma, nodular melanoma, acral lentiginous melanoma, and lentigo maligna melanoma, plus other uncommon variants such as desmoplastic melanoma and nevoid melanoma (De Vries et al., 2005). Concerning the molecular characterisation of melanoma, the arrival of NGS technologies in the marketplace has changed our understanding regarding the complexity of its genomic profile, characterised by heterogeneity and a notably high mutation rate; in fact, any significant progress towards the characterisation of the somatic mutational landscape of melanoma, can be mainly attributed to the rapid evolution of sequencing technologies during the last fifteen years. An increasing number of studies employing NGS have suggested that characterising the patient's mutation profile could be the first step towards the administration of tailored drugs. Current medical informatics standards are being adapted to manage efficiently extremely large medical images. In order to achieve medical imaging and non-imaging integration, there should be multidisciplinary work, including collaboration between informaticians, engineers, pathologists, technicians, clinicians, primary care professionals, and administrators. Digital images can play a significant role in disease early detection and prevention (e.g. cervical cancer screening) (De Vries et al., 2005).

## 1.7 Biological terminology

In this subsection, important biological terms that are used throughout this study are clarified to the reader. Melanoma is caused by genetic mutations in oncogenes or tumour suppressor genes that lead to the unrestrained proliferation of melanocytes (Schramm and Mann, 2011; Walia et al., 2012). Gene expression and cellular mechanisms are also affected, so understanding every molecular component implicated is imperative.

Nucleotides are organic molecules that create long consecutive chains with other nucleotides, tied with special bonds, building long genetic sequences and constituting DNA and RNA molecules, containing the genetic information of an organism. There are four main nucleotides on DNA with diverse nucleobases, adenine (A), cytosine (C), guanine (G) and thymine (T). A mutation – switch of any nucleotide to another on a given spot – is a permanent alteration in a genetic sequence. If only a single nucleotide is affected, then we have a point mutation. This can be insertion/ deletion of one nucleotide or substitution, classified as transition (purine for a purine-A/G- or pyrimidine for a pyrimidine-C/T) or transversion (purine for a pyrimidine, or opposite). Large-scale mutations affect chromosomal structure, and include amplifications (or gene duplications, leading to multiple copies, increasing the dosage of the genes located within them) and deletions of large chromosomal regions (leading to loss of the genes within those regions). Genes and chromosomes can mutate in either somatic or germinal tissue, and these changes are called somatic and germline mutations, respectively. If a somatic mutation occurs in a single cell from developing somatic tissue, then this cell is the progenitor of a population of identical mutant cells, all of which are carrying the mutation (Griffiths et al., 2000). A single nucleotide polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals. A particular SNP may not cause a disorder, but some SNPs are associated with certain diseases. A single-nucleotide variant (SNV), or single-nucleotide alteration, is a variation in a single nucleotide without any limitations of frequency and may arise in somatic cells (“Scitable | Learn Science at Nature”, n.d.).

Gene expression is the process through which information from a gene is used in the synthesis of a functional gene product. Misexpression of wild-type gene products, can cause instabilities in the cellular processes, leading to mutant phenotypes (Prelich, 2012). Differential gene expression refers to the difference in the expression of a gene between two states, i.e. change in the amount of the output product, so that a gene can be up-regulated or down-regulated, or similarly over-expressed or under-expressed, in one state as opposed to another state. For example, a gene may be over-expressed in a disease state, as opposed to the healthy state, changing the balance and effect on cellular processes, implicating this misexpression to the disease phenotype.



A biological pathway is a series of interactions among molecules in a cell that leads to a certain product or a change in a cell. A pathway can trigger the assembly of new molecules, turn genes on and off, or spur a cell to move. Mutations and gene expression have direct effects on pathways and any other molecule in that pathway.

Cell cycle is a mechanism where a cell duplicates its DNA sequence and divides to produce two new cells. It consists of phases: G0 (resting), G1 phase, S phase (synthesis), G2 phase and M phase (mitosis and cytokinesis).

In terms of cancer, “a driver mutation is causally implicated in oncogenesis. It has conferred growth advantage on the cancer cell and has been positively selected in the microenvironment of the tissue in which the cancer arises. A driver mutation need not be required for maintenance of the final cancer, but it must have been selected at some point along the lineage of cancer development”. “A passenger mutation has not been selected, has not conferred clonal growth advantage and has therefore not contributed to cancer development” (Stratton et al., 2009).

The following chapter delivers the related research performed on CM to the reader. In the first subsection of the chapter, the literature concerning the skin imaging field is presented, while in the second subsection the biological background is provided.

# Chapter 2. Related work

## 2.1 Skin Imaging & Dermoscopy background

Melanocytic neoplasms vary from benign lesions, labelled as nevi, to malignant lesions, labelled as melanomas (Shain and Bastian, 2016). Melanoma diagnosis can be challenging and relies on the experience of the dermatologist or physician (“Melanoma Research Foundation”, n.d.). The most common technique for detection of melanoma is dermoscopy (or dermatoscopy or epiluminescence microscopy ELM), which performs the examination through an optical system (magnifying glass) with a light source (polarised light), allowing an in depth visualisation of features used for the diagnosis (Arroyo and Zapirain, 2014).

Over the past decades efforts have been made to create computer-based systems that will improve detection of skin cancer and will also allow repeatability of results (Arroyo and Zapirain, 2014; Maglogiannis and Kosmopoulos, 2006; Mishra and Celebi, 2016). Concerning the digital dermoscopy analysis various approaches based on image analysis exist for the diagnosis of melanoma lesions. The Menzies scale, the Seven-point scale, the Total Dermoscopy Score based on the ABCD rule, and the ABCDE rule (Asymmetry, Border, Colour, Diameter, Evolution) comprise some examples (Argenziano et al., 1998, 2003; Betta et al., 2005). As human interpretation of image content can be subjective, advanced computerised techniques may aid clinicians in the diagnostic process (Manousaki et al., 2006; Ogorzałek et al., 2011). In this context, expert computer systems have been proposed as alternatives to the naked-eye expert prediction. The majority of the existing systems focus on the detection of malignant melanoma and its discrimination from dysplastic or common nevus. However there exist systems aiming at the detection of different modalities. The most common installation type seems to be the video camera, obviously due to the control features that it provides (Tomatis et al., 1998; Umbaugh et al., 1991, 1997). The still camera is of use in some installations, e.g. (Herbin et al., 1993), while infrared or ultraviolet illumination (in situ or in vivo) using appropriate cameras is a popular choice as well, e.g., (Bono et al., 1996; Chwirot et al., 1998; Lohmann and Paul,

1988) correspondingly. Microscopy (or epiluminescence microscopy) installations are applied in the works of (Ganster et al., 2001; Sanders et al., 1999) and digital video microscopy in (Grana et al., 2003).

Dermoscopic images offer a better view of diagnostic features compared to normally magnified images (Rahman and Bhattacharya, 2010), thus improve the diagnostic accuracy. This technique enables the use of methods such as pattern analysis (Pehamberger et al., 1987), the ABCD rule (Stolz and Landthaler, 1994), the Menzies method (Menzies et al., 1996) and the 7-point checklist (Argenziano et al., 1998) which are the most common detectors for melanoma (Arroyo and Zapirain, 2014). Pattern analysis focuses on diagnosing pigmented skin lesions. The ABCD rule describes the asymmetry of shape, border irregularity, colour variety and shape diameter (“Dermoscopy”, n.d.) or differential structures (Maglogiannis et al., 2005). The Menzies scoring method is based on specific features that are either present or not. The 7-point checklist detects melanoma based on the existence of specific patterns. When melanoma is confirmed a biopsy and excision of the malignant region is carried out (Arroyo and Zapirain, 2014).

One major issue of dermoscopy is the inability to detect early melanoma or cases that lack optical features (Goodson and Grossman, 2009). To deal with that issue researchers have focused on molecular techniques.

Aim of this part is to present the state-of-the-art concerning the detection methods of malignant melanoma and describe the contributions made in this area of research.

### 2.1.1 Basic methods for melanoma detection using a computer-based approach

An automated system for the detection/ diagnosis of melanoma consists of five rudimentary steps (Arroyo and Zapirain, 2014; Mishra and Celebi, 2016):

- 1) image acquisition,

- 2) pre-processing,
- 3) segmentation of lesion area,
- 4) detection and classification using important features,
- 5) diagnosis

The first step is to acquire the image of the lesion that is under investigation. A major concern at this point is to ensure the reproducibility of results by standardising camera positioning, distance from the lesion and lighting parameters. These can ensure an accurate comparison for follow-up studies (Maglogiannis et al., 2001). Calibration of an XYZ system on particular RGB (Red Green Blue) colour space can be applied, to describe each instrument, so as to easily compare images among different studies and to achieve better results (Grana et al., 2004). Alternatively, methods for error detection upon acquisition time, utilising spatial domain frequency methods in combination with morphological methods are used (Gutenev et al., 2001).

The pre-processing step includes colour correction, resizing, masking, cropping and hair removal (Abbas, Garcia, Celebi and Ahmad, 2013; Yuan et al., 2006; Zouridakis et al., 2004). Various software correction algorithms can be implemented, such as calibration to black and white, shading correction (Maglogiannis et al., 2001), median filtering (Celebi et al., 2009; Chiem et al., 2007; Maglogiannis et al., 2001; Messadi et al., 2009; Motoyama et al., 2004), Gretag-McBeth colour calibration chart (Maglogiannis et al., 2001), gaussian filtering, anisotropic diffusion filters (Celebi et al., 2009), contrast enhancement (Abbas, Garcia, Celebi, Ahmad, et al., 2013; Chiem et al., 2007), Dull-Razor algorithm (Lee et al., 1997) for thick hair removal, Karhunen-Loeve transform (Messadi et al., 2009).

The segmentation of lesion area in an image is the most important step of the process. Several segmentation algorithms can be utilised to ensure correct segmentation. Most groups use thresholding (Celebi et al., 2009, 2013; Chiem et al., 2007; Jain et al., 2015; Maglogiannis et al., 2001, 2005; Taouil et al., 2006), weighted functions (Maglogiannis et al., 2001, 2005), region growing (Celebi et al., 2009; Maglogiannis et al.,

2001, 2005), principal components transform (PCT) (Maglogiannis et al., 2001, 2005; Zouridakis et al., 2004), CIELAB colour space transform (Maglogiannis et al., 2001, 2005), sigmoid, fuzzy c-Means (Zouridakis et al., 2004), dynamic thresholding (Ganster et al., 2001), clustering algorithms (Metz et al., 2011), model-based, morphological, active contours (Celebi et al., 2009; Nasir et al., 2018; Zhou et al., 2013). Segmentation based on declarative knowledge, implementing multilayer perceptron neural network algorithm and following specific segmentation rules has also been implemented (Kwasnicka and Paradowski, 2005). Common methodology is the application of different algorithms or combinations of them; Taouil et al. (Taouil et al., 2006) use thresholding based on Otsu method (Otsu, 1975), a combination of thresholding and active contours (snakes) and a combination of morphology functions and snakes method for segmentation, and Jain et al. (Jain et al., 2015) utilise the Otsu method on RGB colour planes. Several groups have tested the efficiency of the segmentation algorithms by comparing computer-calculated border with the border manually defined by a dermatologist (Maglogiannis et al., 2001).

The fourth step depends on feature extraction based on pixel-calculation on the segmented image of the lesion. Several categories of features can be examined at this point. The categories include border-based features, such as border irregularity or asymmetry, colour features, which include colour plane measurements, such as RGB, HIS (Hue, Intensity, Saturation), HSV (Hue, Saturation, Value), YIQ (Y component for luma information In phase Quadrature), HLS (Hue, Lightness, Saturation), CMY (Cyan, Magenta, Yellow) or colour variegation (Arroyo et al., 2011; Kaur et al., 2015; Maglogiannis et al., 2001), and size and shape features (Ganster et al., 2001; Ruela et al., 2017; Sadeghi et al., 2013). Another very basic feature for melanoma detection is the pigment network, also known as reticular pattern. Many of the proposed techniques are solely based on the presence or not of pigment network for classification of melanoma (Arroyo and Zapirain, 2014; Dreiseitl et al., 2001; Grana et al., 2004; Sadeghi et al., 2011). Ganster et al. (Ganster et al., 2001) utilised two algorithms for feature selection, sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS), with the SFFS algorithm outperforming SBFS by selecting the smaller number of features. Tanaka et al. (Tanaka et al., 2004) examined over 100 features and evaluated them statistically so as to narrow down the number to the top 10, showing that only a

number of features can prove useful for diagnosis. Chiem et al. (Chiem et al., 2007) used wavelet packet transform (WPT) for feature extraction and continued with principal components analysis (PCA) for feature number reduction.

The extracted dermoscopy image features that are used for automated lesion characterisation are usually the ones that are associated with colour in various colour spaces (RGB, HIS, CIELab), e.g, colour values in (Maglogiannis et al., 2005; Umbaugh et al., 1991, 1997; Zhang et al., 2003) and Colour bin (i.e., the percentage of the lesion coloured foreground pixels) (Zhang et al., 2003). Some of the systems combine features in more than one colour spaces for better results, e.g., HIS and RGB in (Hansen et al., 1997; Herbin et al., 1993; Sanders et al., 1999; Tomatis et al., 1998), or RGB and colours peculiar to malignant melanomas (Motoyama et al., 2004). The intensity characteristics are also used in works like (Chwirot et al., 1998). Asymmetry and border features are also quite common e.g., (Ganster et al., 2001; Grana et al., 2003; Zhang et al., 2003), while features based on differential structures are very rare. Some works (Boldrick et al., 2007; Stanley et al., 2003), rely also on the whole ABCD rule for lesion characterisation, while others combine these with geometrical features (Jain et al., 2015; Maglogiannis and Doukas, 2009). Shape and colour features, like Area and Elevation, calculated manually by dermatologists, have also been used (Zhang et al., 2003).

For the classification, data is divided into groups. Several algorithms have been implemented. Blanzieri et al. (Blanzieri et al., 2000) suggested a multi-classifier system utilising discriminant analysis, decision tree and k-nearest neighbour (kNN) that would improve the performance compared to a single classifier. Lefevre et al. (Lefevre et al., 2000) compared two methods of basic belief assignment (Denoeux's method (Denoeux, 1995) and Appriou's method (Appriou, 1999)) with a novel method for the classification of lesions. Dreiseitl et al. (Dreiseitl et al., 2001) compare several algorithms (kNN, logistic regression, artificial neural networks, decision tree and support vector machine) for the detection of pigment network and classification into either 3 (common nevi vs. dysplastic nevi vs. melanoma) or 2 (common/ dysplastic nevi vs melanoma and common nevi vs. dysplastic nevi/ melanoma) classes. Maglogiannis et al. (Maglogiannis et al., 2001) used linear discriminant analysis and neural networks to compare a combination of groups,

first, the entire region of melanoma with dysplastic nevus, and second, the dark region of melanoma against dysplastic nevus. In 2005, Maglogiannis et al. (Maglogiannis et al., 2005) followed the same techniques using features for each comparison. In 2004, Maglogiannis & Zafiropoulos (Maglogiannis and Zafiropoulos, 2004) compared those previously used classification algorithms with support vector machine (SVM) algorithm. Zhang et al. and Messadi et al. (Messadi et al., 2009; Zhang et al., 2003) implement neural network algorithm for their diagnostic system. Ensemble learners have been employed by (Abedini et al., 2015; Schaefer et al., 2014). Grana et al. (Grana et al., 2003) use discriminant analysis to propose an algorithm that improves border description for classification. Similarly, Tanaka et al. (Tanaka et al., 2004) used discriminant analysis on a low number of features. d'Amico et al. (d'Amico et al., 2004) implement size functions and SVM in their classifier. Motoyama et al. (Motoyama et al., 2004) classify groups by using only colour information. Betta et al. (Betta et al., 2005) propose a simple algorithm that detects important criteria from the 7-point checklist, focusing mostly on 'irregular streaks' and 'atypical pigment network'. Yuan et al. (Yuan et al., 2006) apply SVM algorithm for texture classification. Chiem et al. (Chiem et al., 2007) compared back-propagation neural networks (BNN) to SVM for the classification. Serrano & Acha (Serrano and Acha, 2009) used supervised classification of Markov random fields (MRF) for pattern analysis resulting to melanoma detection. Rahman & Bhattacharya (Rahman and Bhattacharya, 2010) incorporate a classifier combination (SVM, Gaussian maximum likelihood, kNN) for melanoma distinction, where each classifier is given different input features. Sadeghi et al. (Sadeghi et al., 2011) present a graph-based approach for the detection of pigment network, proposing a feature extraction part of a system for melanoma detection. Garcia Arroyo & Garcia Zapirain use a decision tree classifier (Arroyo and Zapirain, 2014) and fuzzy classification of pixels (Garcia-Arroyo and Garcia-Zapirain, 2018) to detect reticular pattern. Deep learning approaches have also been applied (Codella et al., 2015, 2017).

Maglogiannis et al. (Maglogiannis et al., 2001) add an extra step after segmentation denoted as registration, needed for monitoring the progress or possible changes that may occur on skin lesions. An image registration algorithm exports four parameters: magnification, rotation, horizontal and vertical shifting (Venot et al., 1988).

The measurements of two pictures can be compared using statistical correlation. An optimisation algorithm is used to maximise the similarity of the measurements. In particular, the group implemented a deterministic algorithm that utilises cross-correlation of the log-polar Fourier spectrum (Cideciyan, 1995), so as to compare magnification and rotation measurements (Maglogiannis et al., 2001, 2005).

### *2.1.1.1 Evaluation of the basic methods using a computer-based approach*

This section focuses on the performance and contribution of the methods described above. An evaluation and comparison of the methods cannot be absolute because of the different datasets used for each study. As mentioned before, repeatability and reproducibility of results, and of course standardisation of image acquisition and processing, are crucial.

Blanzieri et al. (Blanzieri et al., 2000) achieved low numbers of sensitivity and specificity when using single classifiers but improved most of their results using a combination of the classifiers (over 80% sensitivity and over 70% specificity). Apart from that, their system performs equally, when compared to diagnosis from dermatologists. Lefevre et al. (Lefevre et al., 2000) achieve 98.8% of good classification with their method as opposed to Appriou's 97.5% and Denoeux's 91.35%. Dreiseitl et al. (Dreiseitl et al., 2001) showed that logistic regression, artificial neural networks and SVM perform equally good in classification accuracy and that kNN and decision tree, although they do not perform as good, produce results comparable to the accuracy of expert diagnosis. Maglogiannis et al. (Maglogiannis et al., 2001) achieved accuracy of 96.2% and 96% using discriminant analysis, whereas neural networks had 100% in both cases (4 principal components), but reduced accuracy when 2 principal components were used (84.6% and 96%). Similarly, in 2005 the same group showed accuracy of 97% in both cases with discriminant analysis, 97% and 100% for neural network model of 4 components and 85% and 94% for neural network using 2 components (Maglogiannis et al., 2005). Maglogiannis & Zafiropoulos (Maglogiannis and Zafiropoulos, 2004) achieved 94.1%



accuracy using SVM as opposed to 88% for discriminant analysis and 94.1% for 4 component neural network. They suggest 'Using Area' and 'Thinness Ratio' as important features for the diagnosis of melanoma. Ganster et al. (Ganster et al., 2001) showed that melanoma detection can be more difficult when comparing 3 groups (benign nevi, dysplastic nevi and melanoma) both for automated classification and clinical diagnosis achieving 73% and 72% sensitivity respectively. When combining dysplastic nevi and melanoma in one group they showed 87% sensitivity and 92% specificity. When combining benign with dysplastic nevi, sensitivity was 77% and specificity 84%. Overall, they showed that results improve when dealing with two categories for classification. (Zhang et al., 2003)'s system achieves 92% of correctly diagnosed results using neural networks with an automatic border detector, as opposed to 81% of average correct diagnostic rate performed by 16 dermatologists. Grana et al. (Grana et al., 2003) manage 80% diagnostic accuracy with 85.88% sensitivity and 74.12% specificity. As in (Zhang et al., 2003), they propose automatic assignment of border parameters. Tanaka et al. (Tanaka et al., 2004) achieve 90% classification ratio for melanoma (sensitivity), with 98.3% specificity (96% accuracy), on the basis of 10 features. d'Amico et al. (d'Amico et al., 2004) showed over 84% sensitivity and over 83% specificity in their tests, managing 100% sensitivity at 63.65% specificity. (Motoyama et al., 2004)'s results may not present great accuracy (only 26%) but showed the importance of using several features (they used only one colour space) in the detection of melanoma. Yuan et al. (Yuan et al., 2006) achieve about 70% accuracy for their SVM algorithm suggesting optimal values for window size and degree. Chiem et al. (Chiem et al., 2007) have compared BNN to SVM showing 95% and 85% accuracy of classification, respectively. Messadi et al. (Messadi, Bessaid, and Taleb-Ahmed 2009) reach 74.5% accurate classification rate with 67.5% sensitivity and 80.5% specificity. Serrano & Acha (Serrano and Acha, 2009) achieve a correct classification rate of 86%, not focusing on feature extraction, but using pattern analysis instead. Rahman & Bhattacharya (Rahman and Bhattacharya, 2010) classify between 3 groups achieving a 83.75% accuracy for the malignant category using a classifier combination as opposed to 72.45% accuracy when using only one classifier (SVM had the best performance). Sadeghi et al. (Sadeghi et al., 2011) present 94.3% accuracy in pigment network detection, suggesting their method to be used as part of a

diagnostic system for melanoma. Garcia Arroyo & Garcia Zapirain (Arroyo and Zapirain, 2014) achieve 86% sensitivity and 81.67% specificity in detection of pigment network using a decision trees approach, similarly to fuzzy sets (Garcia-Arroyo and Garcia-Zapirain, 2018). Kasmi & Mokrani (Kasmi and Mokrani, 2016) attain 91.25% sensitivity and 95.83% specificity using an automatic ABCD scoring methodology. More recently, 97.5% accuracy was achieved using SVM on entropy-based selected features (Nasir et al., 2018).

### 2.1.2 Diverse techniques for melanoma detection

Apart from the declared image analysis techniques, several alternatives have been suggested the last two decades. Claridge et al. (Claridge et al., 2003) suggested that image colouration represents specific histological measurements. The classification, in this case, is based on physical maps created for each histological parameter, such as melanin and blood concentration and collagen thickness. This combination achieved 80.1% sensitivity and 82.7% specificity for their classification results, as well as 96.2% sensitivity using the presence of melanin as a diagnostic feature. A procedure which also differs from the 'classic' steps was introduced by Buzug et al. (Buzug et al., 2006). The group used infrared imaging to distinguish between malignant melanoma and healthy skin, based on the fact that cancerous tissue shows increased metabolism and blood flow, thus assuming higher temperature and higher absorbance. Similarly, side-transillumination, which displays both subsurface and surface features, paired with cross polarisation, which only displays surface features, can help highlight a malignant specimen, demonstrating a much larger area due to increased blood vessels (Zouridakis et al., 2004). These led to 75% accurate classification on a small dataset. Total body photography to monitor high-risk patients (carrying many dysplastic nevi) and facilitate their skin examination has been eagerly used (Banky et al., 2005; Edmondson et al., 1999; Feit et al., 2004), and specific algorithms for mole identification, counting and segmentation, have proven very valuable (Lee et al., 2005). This kind of methodology

could help in the identification of new nevi and comparison with existing nevi, which could prove very useful in cancer detection (Goodson and Grossman, 2009).

### 2.1.3 Molecular techniques for melanoma detection

Molecular techniques for the diagnosis of malignant melanoma include non-invasive methods (tape stripping) (Wong et al., 2006), tissue microarrays (Rothberg et al., 2009) or fluorescence in situ hybridisation (FISH) (Gerami et al., 2009). In (Wachsman et al., 2007, 2011) genomic microarray analysis is utilised to identify genetic biomarkers that can accurately detect melanoma. The group achieved 100% accuracy, using a 5-gene discrimination method, in 2007, and showed 100% sensitivity and 88% specificity, this time using a 12-gene classifier, in 2011. Chandler et al. (Chandler et al., 2012) presented single nucleotide polymorphism genomic microarray (SNP-GMA), achieving high degrees of accuracy in detection of malignancy - reached 89% sensitivity and 100% specificity. Liu et al. (Liu, Peng, et al., 2013) proposed a model that evaluated the significance of existent datasets from (Hoek et al., 2004; Riker et al., 2008; Rose et al., 2011; Scatolini et al., 2010; Smith et al., 2005) to produce a 12-gene biomarker. For their classification model they utilised support vector machine and leave-one-out-method, achieving an average of 99.1% accuracy. Another aspect that should be mentioned is the potential benefits emerging from the development of content-based image retrieval (CBIR) systems. A database of already diagnosed cases of melanoma has been proposed by Rahman et al. (Rahman and Bhattacharya, 2010). Iakovidis et al. (Iakovidis et al., 2009) presented an approach for CBIR based on pattern similarity measures, associated with feature clustering; PANDA framework (PAtterns for Next generation DAtabase systems).

### 2.1.4 Data Integration

Combining different levels of information concerning a problem can improve the total knowledge on that problem and assist in the effort of finding a solution (Lanckriet et

al., 2004). This leads to the conclusion that diagnosis should be based on the correct integration of molecular, histological as well as clinical features, so as to be more accurate. The association of the phenotypical characteristics of nevi with genetic variants has been shown and reinforces the use of multi-source data integration (Cuéllar et al., 2009; Vallone et al., 2018; Zalaudek et al., 2007). Combining those complementary pieces of information can be expected to enhance not only diagnosis, but biomarker detection as well. It has been well accepted that research engagement will be most productive when illuminating the intersection made up of genetics, clinical data, and imaging features (Jaffe, 2012). The integration of multi-level biological data, including prognostic quantitative imaging biomarkers and signatures, and clinical variables, can generate more robust, detailed descriptors that can aid diagnosis and supplementary personalised patient plans (Katrib et al., 2016).

Lanckriet et al. proposed a statistical model that combines heterogeneous data in order to achieve better performance on a given statistical task, based on kernel functions. Each kernel function offers specific information for the data. To combine the kernels, they utilised the semidefinite programming method. For classification they used support vector machine, which is kernel-based. This method can find use in any biological problem with heterogeneous features (Lanckriet et al., 2004). Feature level image fusion for multimodal medical images using wavelet transform has been applied, showing that the fused image offers more valuable information, evaluated by standard deviation, entropy, cross entropy and gradient parameters (Kor and Tiwary, 2004). Classification can be upgraded by accumulation of lesion depth and structure information, obtained from the multivariate images on the surface representative information obtained from the dermoscopic images (Patwardhan et al., 2005). Winnepenninckx et al. (Winnepenninckx et al., 2006) suggested that there is a correlation of genomic profiling and clinical outcome and used statistical analysis (Pearson's correlation values) to study the correlation of gene expression and histological features. A similar approach to (Lanckriet et al., 2004) was utilised by Ye et al. (Ye et al., 2008) for the study of Alzheimer's disease (AD). The group integrated imaging with genetic measurements to achieve accurate prediction for AD, based on kernel methods. They further extended their research to identifying biomarkers from the heterogeneous dataset. Again, support vector machine was used for

classification. Similarly, Metsis et al. (Metsis et al., 2012) presented a computational and machine learning based framework for integrating heterogeneous gene expression and metabolomic data to classify the different types of brain tumours and extract biomarkers, using naïve Bayes and support vector machines for classification. Kashani-Sabet et al. (Kashani-Sabet et al., 2009) presented a multi-marker model for the prediction of melanoma, based on immunohistochemical features derived from genomic profiling. To assess the correlation of the two datasets Cox regression and Kaplan-Meier analyses were utilised. Rothberg et al. (Rothberg et al., 2009) proposed a multi-marker molecular model for melanoma classification, derived from microarray, immunohistochemical and image association, performed by statistical tests. Li & Patra (Li and Patra, 2010) proposed a 'random walk with restart on heterogeneous network' (RWRH) algorithm for better understanding of the gene-phenotype relationship. This method can optimise data fusion and offer a ranking for biomarker extraction. Mann et al. (Mann et al., 2013) presented a prediction model based both on clinical features and genomic data, that showed improved accuracy over unimodal datasets. Concerning glioblastoma, Zinn et al. (Zinn et al., 2011) combined molecular with magnetic resonance imaging (MRI) volumetrics, accessing public datasets to invent a new diagnostic imaging technique, like Jamshidi et al. (Jamshidi et al., 2013) who used expression and MRI data. Regarding non-small cell lung cancer, Gevaert et al. (Gevaert et al., 2012) explored the clinical prognostic value of radio-genomic imaging, concluding that it is highly possible to leverage expression data to determine prognosis and therapeutic response as a function of image features.

More recently, Moutselos et al. (Moutselos et al., 2014) and Valavanis et al. (Valavanis, Maglogiannis, and Chatziioannou 2015) presented a methodology that integrates fusion of imaging and microarray data for the study of melanoma. They use a number of feature extraction algorithms (principal components analysis, linear discriminant analysis and random forests) for the discovery of composite biomarkers that would achieve higher performance upon classification. Lazova et al. (Lazova et al., 2017) integrated molecular with mass spectrometry imaging to assist in the classification of diagnostically challenging atypical Spitzoid melanomas, as well as diagnosis and prediction of outcome. Clustering analysis can be used to determine likely clusters formed

when integrating datasets from different platforms, as shown by (Hoadley et al., 2014), who extracted different biomarkers to characterise each cluster deriving from twelve types of cancer. In 2016, Kavakiotis et al. presented a methodology for integrating multiple immunogenetic and clinico-biological data sources, for the study chronic lymphocytic leukaemia. Their method is based on ranking aggregation approach and formalisation of voting systems (Kavakiotis et al., 2016). Another approach is using several fusion methods simultaneously in a dispersed system to reach a decision. This was done by (Przybyła-Kasperek et al., 2017) using two separate datasets, utilising basic statistical fusion methodologies, and improving the final outcome as compared to the discrete analyses.

## 2.1.5 Discussion

Due to high mortality rate and the difficulty of treatment, in case of late diagnosis of malignant melanoma, the sensitivity of a method is more important. Success of classification depends mainly on feature selection. A summarisation of the aforementioned surveyed works concerning skin imaging classification systems is presented in Table 1.

*Table 1: Comparative table for the surveyed classification systems*

Reference	Classification System	Information	Groups for comparison	Results *
(Blanzieri et al., 2000)	Discriminant analysis, decision trees, kNN	Single vs. multi-classifier methodology	Melanoma vs. nevus	Multi-class. Sens>80%, Spec>70%, Acc=98.8%
(Lefevre et al., 2000)	Basic relief assignment	vs. similar methods	2	Acc=98.8%
(Dreiseitl et al., 2001)	Logistic regression, ANN, SVM, kNN, decision trees	Detection of pigmentation network	3 or 2 groups, Nevus vs. dysplastic vs. melanoma	Logistic regression ~ANN~SVM >kNN, decision trees

(Ganster et al., 2001)	Automated classification	3 group comparison vs. clinical diagnosis	3 or 2 groups, dysplastic & melanoma vs. nevus / melanoma vs. dysplastic & nevus	3 class. Auto. Sens=73%, Clin. Sens=72%, dys&mel Sens=87%, Spec=92%, dys&nev Sens=77%, Spec=84%
(Maglogiannis et al., 2001)	LDA, neural networks	PCA, focus on regions	2 Melanoma vs. dysplastic	LDA 4PCs Acc=96.2% / 2PCs Acc=96% NN 4PCs Acc=100% / 2PCs Acc=84.6%
(Grana et al., 2003)	Discriminant analysis	Automatic assignment for border parameters	2	Acc=80%, Sens=85.88%, Spec=74.12
(Zhang et al., 2003)	Neural networks	Automatic assignment for border parameters	2	Acc=92%
(d'Amico et al., 2004)	Size function on SVM	ABCDE rule	2	Sens>84%, Spec>83%, →Sens=100% - Spec=63.65%
(Maglogiannis and Zafiroopoulos, 2004)	SVM, LDA	-	2	SVM Acc=94.1%, LDA Acc=88%
(Tanaka et al., 2004)	Discriminant analysis	10 features	2	Acc=96%, Sens=90%, Spec=98.3%
(Betta et al., 2005)	Criteria detection	7-point checklist	2	-
(Yuan et al., 2006)	SVM	Texture classification	2	Acc=70%
(Chiem et al., 2007)	Back-propagation neural networks, SVM	BNN to SVM	2	BNN Acc=95%, SVM Acc=85%

(Messadi et al., 2009)	Neural networks	-	2	Acc=74.5%, Sens=67.5%, Spec=80.5%
(Serrano and Acha, 2009)	Supervised classification of Markov fields	Pattern analysis	2	Acc=86%
(Rahman and Bhattacharya, 2010)	SVM, Gaussian maximum likelihood, kNN	Class. combination	3	Acc=83.75%, SVM-only Acc=72.45%
(Sadeghi et al., 2011)	Graph-based	Pigmentation network detection	2	Acc=94.3%
(Arroyo and Zapirain, 2014)	Decision tree for reticular pattern	Pigmentation network detection	2	Sens=86%, Spec=81.67%
(Schaefer et al., 2014)	Ensemble classifiers	SVM fusion	2	Sens=93.76%, Spec=93.84%
(Abedini et al., 2015)	Ensemble classifiers	Discriminant analysis, ANN, kNN, SVM, decision trees	2	Acc=91%, Sens=97%, Spec=65%
(Codella et al., 2015)	Deep learning	Sparse coding, SVM	2, melanoma vs. non-melanoma, melanoma vs. atypical	Mel vs. non Acc=93.1%, Sens=94.9%, Spec=92.8% Mel vs. atyp Acc=73.9%, Sens=73.8%, Spec=74.3%
(Kasmi and Mokrani, 2016)	Automatic scoring	ABCD rule	2	Sens=91.25%, Spec=95.83%
(Codella et al., 2017)	Deep learning	-	2	Acc=76%
(Garcia-Arroyo and Garcia-Zapirain, 2018)	Fuzzy class. of image pixels	Based on decision trees	2	Acc=88%, Sens=90.71%, Spec=83.44%
(Nasir et al., 2018)	SVM	Entropy-based features	2	Acc=97.5%

\* Acc: Accuracy, Sens: Sensitivity, Spec: Specificity



The techniques mentioned here offer very promising results for the detection of melanoma. This part of the thesis was mostly focused on comparing classification methods and investigating which machine learning method delivers the best results. A number of different classifiers have been utilised, with neural networks, support vector machines and discriminant analysis achieving the highest accuracy (d'Amico et al., 2004; Arroyo and Zapirain, 2014; Blanzieri et al., 2000; Celebi et al., 2009; Dreiseitl et al., 2001; Garcia-Arroyo and Garcia-Zapirain, 2018; Grana et al., 2003; Maglogiannis et al., 2001, 2005; Maglogiannis and Kosmopoulos, 2006; Maglogiannis and Zafiroopoulos, 2004; Messadi et al., 2009; Nasir et al., 2018; Rahman and Bhattacharya, 2010; Sadeghi et al., 2011; Serrano and Acha, 2009; Tanaka et al., 2004; Yuan et al., 2006; Zhang et al., 2003). Other groups were concerned mostly about the feature extraction step, using statistical methods and supervised learning to lower the feature level, to achieve better performance on classification. The number of features that was used in each case varies, as does their nature. It is clear, though, that the success of classification depends mainly on feature selection (Chiem et al., 2007; Ganster et al., 2001; Maglogiannis et al., 2005; Nasir et al., 2018; Tanaka et al., 2004).

The most common classification methods are rule-based, e.g., (Bono et al., 1996; Chwirot et al., 1998; Dreiseitl et al., 2001; Grana et al., 2003; Sanders et al., 1999; Stanley et al., 2003; Tomatis et al., 1998). More advanced techniques such as neural networks and support vector machines are presented in works like (Boldrick et al., 2007; Ercal et al., 1994; Maglogiannis and Zafiroopoulos, 2004; Nasir et al., 2018; Rubegni et al., 2002; Umbaugh et al., 1991, 1997; Zhang et al., 2003), while the k-nearest neighbourhood classification scheme is applied in (Ballerini et al., 2013; Ganster et al., 2001). Evidence Theory (upper and lower probabilities induced by multivalued mapping) based on the concept of lower and upper bounds for a set of compatible probability distributions is used in (Lefevre et al., 2000) for melanoma detection. The success rates for the methods presented in the literature indicate that the work towards automated classification of lesions and melanoma may provide good results. Detailed descriptions and results regarding the methods used in existing dermoscopy analysis systems are presented in (Korotkov and Garcia, 2012; Mishra and Celebi, 2016).

Molecular techniques have led to the discovery of a number of biomarkers for the detection of the disease. The biological function of these biomarkers is known, proving that linkage to histology and imaging is possible. Still only few efforts have been made to understand the molecular or histological level of association with the clinical view (image), to base the prediction on this association. Research by (Buzug et al., 2006; Claridge et al., 2003; Zouridakis et al., 2004) showed that there is a clear connection between histological and image features, suggesting a combination of parameters to be used as features for detection. Suggestions that the phenotypical characteristics of nevi are correlated to genetic variants have been given (Cuéllar et al., 2009; Vallone et al., 2018; Zalaudek et al., 2007). Some groups have used data fusion to improve predictions concerning melanoma (Kashani-Sabet et al., 2009; Mann et al., 2013; Moutselos et al., 2014; Rothberg et al., 2009; Valavanis et al., 2015; Winnepenninckx et al., 2006). Research on the latter issue is at least scanty, therefore further investigation is needed.

## 2.2 Biological background

Melanoma is caused by genetic mutations in oncogenes or tumour suppressor genes that lead to the unrestrained proliferation of melanocytes (Schramm and Mann, 2011; Walia et al., 2012). In this part, we summarise the progress towards the genomic characterisation of CM.

### 2.2.1 Germline susceptibility

Regarding the genetic background predisposing to melanoma, several susceptibility loci acting as high, moderate or low penetration genes, have been identified. *CDKN2A* (Cyclin-dependent kinase inhibitor 2A), the first familial melanoma gene identified (Hussussian et al., 1994; Kamb et al., 1994), is found mutated in approximately 40% of melanoma high-density families. *CDKN2A* encodes for two distinct proteins, p16INK4A (p16) and p14ARF (p14), both involved in the regulation of the cell cycle (de Snoo and Hayward, 2005). The p16 and p14 mRNAs are transcribed from alternative first exons, so the related proteins have no similarity in their amino acid sequence, since they are translated in alternative reading frames. Mutations in p16 are predominantly loss of-function missense mutations, distributed throughout the protein, while in p14 inactivating mutations like whole gene deletions, insertions or splice-site mutations are mainly observed. Germline mutations in *CDK4* are much less frequent and were initially identified by screening for p16 interacting partners. A mutational hotspot in codon 24, leading to an arginine substitution, abrogates the capacity of p16 to inactivate the kinase, thus promoting the G1-S phase transition of the cell cycle. Other mutations have been identified in genes of more moderate penetrance, including *BAP1*, *TERT*, *POT1*, *ACD*, *TERF2IP* and *MITF* (Aoude et al., 2015). Genome-wide association studies (GWASs) have also revealed numerous recurring single nucleotide polymorphisms (SNPs) associated with melanoma risk (Antonopoulou et al., 2015; Athanasiadis et al., 2014; Chatzinasiou et al., 2011; Law et al., 2015).

## 2.2.2 Somatic alterations

Identifying somatic mutations in the genome of melanoma is of great importance in order to understand the molecular basis of the disease's genesis and progression. A number of oncogenes and tumour suppressor genes have been found to carry causative mutations. The first oncogene identified in melanoma was *NRAS* (Padua et al., 1984), which is also found mutated in other cancers. In 2002, the *BRAF*V600E somatic mutation was identified (Davies et al., 2002), which is the most frequent mutation found in CM patients. Since then, the advances in sequencing technology have enabled the application of massively parallel sequencing, thus dramatically changing our understanding of the somatic mutation landscape of melanoma. The first catalogue of somatic mutations of a cancer genome, at the whole-genome level, concerned a melanoma cell line (Plesance et al., 2010), indicating the presence of a great number of mutations per Mb and suggesting a mutational signature related to UV exposure. Whole-exome sequencing studies exploiting clinical samples demonstrated that *NF1*, *ARID2*, *PPP6C*, *RAC1*, *SNX31*, *TACC1*, and *STK19* are genes significantly mutated in melanoma (Hodis et al., 2012; Krauthammer et al., 2012). The Cancer Genome Atlas Skin Cutaneous Melanoma (SKCM-TCGA) project confirmed, through exome sequencing, previously reported melanoma oncogenes and tumour suppressors (*BRAF*, *NRAS*, *CDKN2A*, *TP53*, and *PTEN*) and identified several additional significantly mutated melanoma genes, namely, *MAP2K1*, *IDH1*, *RB1*, and *DDX3X* ("The Cancer Genome Atlas", n.d.). The study proposed the classification of CM into four major genomic subtypes, related to the presence of specific mutations in established driver genes. In particular, the proposed genetic subtypes are the *BRAF* mutant, *RAS* mutant, *NF1* mutant, and the triple wild-type (no mutations in the aforementioned genes). Low-frequency mutations were identified in the triple wild-type subtype in *KIT*, *CTNNB1*, *GNA11*, and *GNAQ*. More recently, the first large-scale study exploiting whole-genome sequencing supported the involvement of the non-coding genome in melanoma pathogenesis and revealed diverse carcinogenic processes across the different melanoma subtypes. Figure 2 summarises the research on melanoma during the last decades, pinpointing key milestones in understanding its complexity.

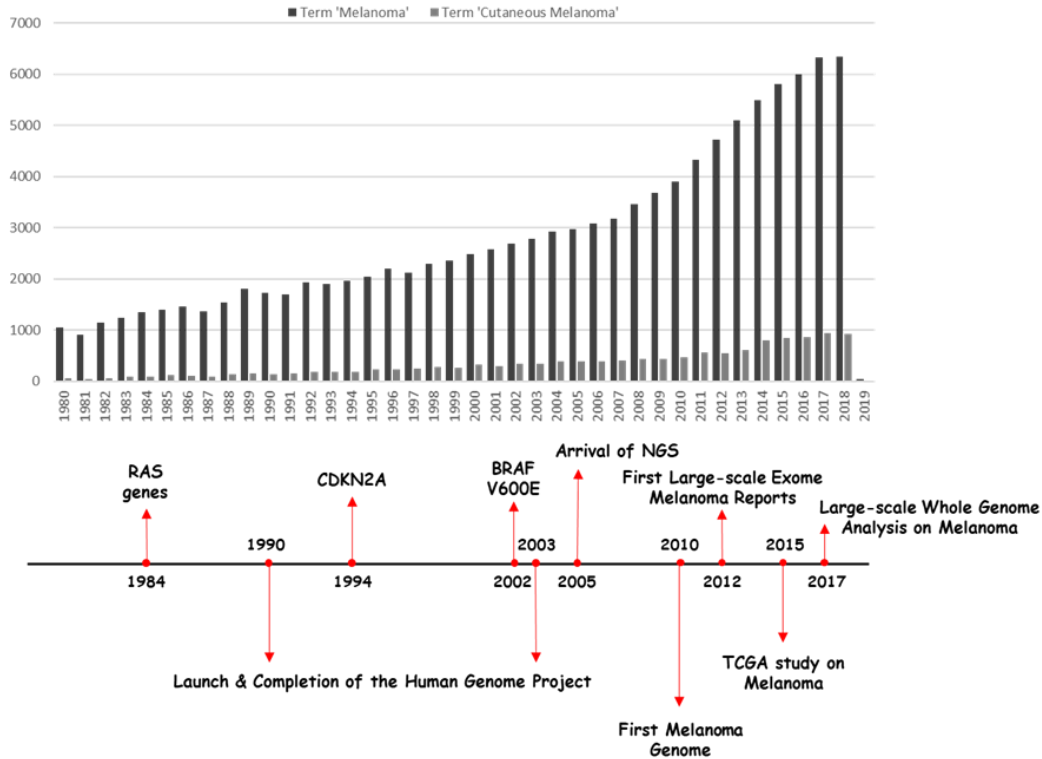


Figure 2: (upper) Number of publications per year on Pubmed (until 4th December 2018) using terms 'melanoma' and 'cutaneous melanoma', (lower) Major landmarks concerning the study of melanoma

### 2.2.2.1 Mutation burden and specific signatures in melanoma

Sequencing of different cancers has revealed that the melanoma genome shows a substantial prevalence of somatic mutations (Alexandrov et al., 2013; Greenman et al., 2007; Pleasance et al., 2010). Particularly, in CM an increased abundance of cytidine to thymidine (C > T) transitions is observed. This specific alteration is considered characteristic of a UV-light-induced mutational signature. A recent study, exploiting whole-genome sequencing of cutaneous, acral and mucosal melanomas, revealed distinct mutation profiles among these melanoma subtypes. The number of base substitutions and short insertions and/or deletions in CM was generally much higher than of those observed in acral and mucosal melanomas. In addition, the UV-related C > T transition was not observed in the latter melanoma subtypes. In contrast, somatic

structural rearrangements were more frequent in acral and mucosal subtype (Hayward et al., 2017). These data suggest that different etiologic pathways are involved in the manifestation of diverse melanoma subtypes.

#### 2.2.2.2 *Genes bearing causative somatic mutations in melanoma*

One of the most well-established pathways commonly affected in melanoma is the mitogen activating protein kinase (MAPK) signalling cascade, governing cell growth and survival. *BRAF*, *NRAS* and *NF1* are the most frequently mutated genes of this pathway. Other pathways found significantly altered in CM, include the phosphoinositide 3-kinase (PI3K) pathway, Tumour Protein 53 (TP53) signalling, cell cycle regulation and the telomere length maintenance pathway. In the next sections, the most significant genes involved in such key processes, harbouring driver mutations, are summarised.

##### *BRAF*

The *BRAF* gene encodes a serine/threonine protein kinase, belonging to the RAF family. This protein acts as a downstream effector of *RAS*-signalling in the MAPK cascade, affecting cell proliferation and survival. Mutations in this gene have been identified in various cancers. According to the COSMIC database (“COSMIC database”, n.d.; Forbes et al., 2017), 44% of melanomas arising from skin tissue bear mutations in *BRAF*. In non-acral CM, the *BRAF* mutation of the kinase-activation domain at amino acid position 600, is the most common somatic mutation. Interestingly, *BRAF* V600E mutation results from a T > A transversion and not a C > T substitution, which is characteristic of UV light induced mutagenesis. Nevertheless, epidemiological and genomic evidence implies that UV radiation contributes to the formation of *BRAF* V600E. Soon after the characterisation of *BRAF* V600E mutation in melanomas, it became apparent that its distribution greatly differs among different melanoma subtypes (Maldonado et al., 2003). In particular, *BRAF* V600E mutations are more common in younger CM patients, whose melanomas arise on intermittently sun-exposed skin, on anatomical sites, such as the

trunk and proximal extremities. In contrast, melanomas arising on chronically sun-damaged skin, usually on anatomical sites like head, neck and the distal extremities of older individuals, have infrequent *BRAF* mutations, with *BRAF* V600K being more frequent than *BRAF* V600E (Bastian, 2014). Acral melanomas bear *BRAF* mutations much less frequently. Targeting the *BRAF*-V600E mutant protein with specific inhibitors exposed new therapeutic aspects for the management of such an aggressive disease. The oncogenic activation of *BRAF* mutations is considered a necessary but not sufficient condition to transform melanocytes to melanoma cells, a suggestion which is also supported by the frequent occurrence of such mutations in benign nevi (Bastian, 2014).

## *RAS*

*RAS* proteins are small GTPases (enzymes that control signal transduction) functioning as GDP–GTP-regulated binary switches that control many fundamental cellular processes. *RAS* proteins connect a great variety of upstream signals from activated membrane receptors to downstream pathways controlling cell cycle, growth, apoptosis, and senescence (Simanshu et al., 2017). The *HRAS*, *KRAS*, and *NRAS* oncogenes were the first human oncogenes to be discovered (Cox and Der, 2010). In the case of CM, *NRAS* mutations are found in 17% of the cases, according to COSMIC database (“COSMIC database”, n.d.; Forbes et al., 2017). *NRAS* hot-spot mutations are mutually exclusive of *BRAF* hot-spot mutations. *HRAS* and *KRAS* mutations are much less frequent in CM. Regarding *NRAS*, the most common mutations cause a change of the amino acid at position 61, lying at the GTP-binding domain. These substitutions disrupt the GTPase activity of the protein, locking it in its active conformation (Fedorenko et al., 2013).

## *NF1*

*NF1* is a tumour suppressor gene encoding for a direct negative regulator of *RAS* signalling (Vigil et al., 2010). In particular, *NF1* is a GTPase-activating protein known to

downregulate RAS activity by stimulating the hydrolysis of GTP and returning the protein to its inactive form. A significant enrichment of *NF1* mutations was found in *BRAF* and *NRAS* wild-type melanomas (Hodis et al., 2012; Krauthammer et al., 2012). In the TCGA study, *NF1* was found as the third most frequently observed significantly mutated gene of the MAPK pathway (“Genomic Classification of Cutaneous Melanoma”, 2015). Mutations in the *NF1* gene are loss of function mutations, mainly nonsense point mutations (Cirenajwis et al., 2017), which can be considered as an alternative way to activate the MAPK signalling pathway.

### *TERT*

The *TERT* gene encodes for the telomerase reverse transcriptase, the catalytic subunit of the telomerase ribonucleoprotein, essential for the maintenance of telomeres and chromosomal stability. Recurrent somatic mutations in the *TERT* promoter have been characterised in CM, with high frequency in sporadic melanoma. Specifically, the two hot-spot mutations, located at -124 and -146bp relative to the transcriptional start site, are C > T transitions, consistent with a UV signature mutational profile (Horn et al., 2013; Huang et al., 2013). In a recent study exploiting whole genome sequencing, 86% of CM cases were found mutated at one or more out of four positions upstream of the transcriptional start site (Hayward et al., 2017). All these mutations are mutually exclusive and create new binding sites for the E26 transformation-specific (ETS) family transcription factor GA-binding protein (GABP). Recent evidence suggests that *TERT* promoter mutations result in *TERT* over-expression (Shain et al., 2018; Zhao et al., 2016). They are established after MAPK-pathway activating mutations, but still during the early stage of melanoma progression (Shain et al., 2018).

### *CDKN2A*

*CDKN2A* is a well characterised tumour-suppressor gene, found to harbour somatic alterations in a wide variety of different tumour types (Zhao et al., 2016).



Regarding CM, in addition to its association with familial melanoma, somatic alterations resulting in *CDKN2A* inactivation are also frequently observed in sporadic melanoma. The most frequent alteration is the deletion of the *CDKN2A* gene, reported in 41% of CM (Palmieri et al., 2018). *CDKN2A* expression is additionally regulated at the epigenetic level, mainly by methylation of its promoter and subsequent gene silencing. The two proteins encoded by *CDKN2A*, p16 and p14, have distinct roles in the regulation of cell cycle. p16 modulates G1 to S phase transition by inhibiting the kinase activity of cyclin dependent kinases 4 and 6 (CDK4 and CDK6), while p14 acts through TP53 stabilisation. Biallelic loss of *CDKN2A* and subsequent disruption of the G1/S checkpoint, is believed to be a crucial step in melanoma progression towards transition to the invasive phenotype (Shain et al., 2018).

### *TP53*

*TP53* is a well-known tumour suppressor gene, involved in the transcriptional regulation of several target genes. *TP53* is mutated in 27 different types of cancer (Bailey et al., 2018). Regarding melanoma, 15% of cases harbour mutations in *TP53* (“COSMIC database”, n.d.; Forbes et al., 2017). Based on mutational studies, comparing primary melanomas and metastases, *TP53* was found to be more frequently mutated in melanoma metastases, indicating that *TP53* mutations may arise later during melanoma progression (Bastian, 2014).

### *PTEN*

*PTEN* is a tumour-suppressor gene, coding for the phosphatidylinositol-3,4,5-triphosphate 3-phosphatase. PTEN phosphatase is a fundamental regulator of the PI3K/AKT pathway, exerting its inhibitory effects on AKT signalling, by dephosphorylating PIP3. PIP3 acts as a second messenger, triggering a number of signalling cascades—among them AKT— which play a key role in processes like cell survival and proliferation, apoptosis, and cellular metabolism (Ortega-Molina and Serrano, 2013).

Somatic mutations in *PTEN*, primarily deletions but also loss-of-function SNVs- 18 % and 8%, respectively in CM (Palmieri et al., 2018), result in *PTEN* inactivation and promote cell survival through sustained activation of the PI3K signalling pathway (Helgadottir et al., 2018).

### *MITF*

Microphthalmia-associated transcription factor (MITF) is a basic helix–loop–helix/leucine zipper transcription factor required for melanocyte development. MITF is essential for establishing the melanocytic lineage during differentiation of neural crest cells (Mort et al., 2015). Transcriptional targets of MITF, include genes encoding for components of melanosomes, enzymes of the melanin synthesis pathway, as well as genes involved in cell cycle regulation and cell survival. Somatic amplification of *MITF* has been identified in melanomas, but MITF activity is mainly altered by its upstream activators and suppressors acting on the transcriptional, post-transcriptional and post-translational levels.

### *Other genes*

Other genes causatively implicated in melanomagenesis and progression include *KIT*, *RAC1* and *ARID2*. *KIT* encodes for a tyrosine kinase, which is the receptor of the Stem Cell Factor. Upon ligand binding, multiple signalling pathways affecting cell growth, proliferation, survival, and migration are activated. In CM, mutations in *KIT* occur most commonly in melanomas originating from chronically sun damaged skin and in the acral subtype (Bastian, 2014). The *RAC1* gene encodes for a GTPase of the Ras superfamily with important roles in cell motility. A hot spot mutation at P29S, is the result of a C > T transition, consistent with the molecular signature associated with UV damage (Krauthammer et al., 2012). The *ARID2* gene encodes for a subunit of the switch/sucrose non-fermentable (SWI/SNF) chromatin remodelling complex, a multiprotein complex that alters chromatin structure to regulate gene expression (Mehrotra et al., 2014). Recent

evidence suggests that components of the SWI/SNF complex, function as tumour suppressors in several types of cancer. In the case of CM, loss-of-function mutations in the *ARID2* gene are the most frequent among SWI/SNF enzymes.

### 2.2.2.3 *Affected pathways & gene expression*

As have been mentioned, among the major affected pathways during the development and progression of melanoma are the MAPK pathway, cell cycle, DNA damage response and cell death pathways, and PI3K/Akt pathway (“Genomic Classification of Cutaneous Melanoma”, 2015). Additional pathways that have been shown to be implicated in the process of melanoma-genesis include Notch, Wnt, TGF- $\beta$ , NF- $\kappa$ B, PKC, JNK/c-JUN, BCL-2, and APAF-1 (Kong et al., 2010). Basically, each gene taking part in those mechanisms may have altered expression, leading to the advancement of melanoma. Shain et al. (Shain et al., 2018) revealed the sequential events that lead to the evolution from pre-malignant lesions to melanoma. MAPK activation is preceded by telomerase up-regulation, then chromatin modulation, cell cycle checkpoint and p53 pathway disruption, and PI3K path activation. *TERT* expression is significantly elevated. *EZH2* is up-regulated in the presence of immune infiltration, and its expression in melanoma cells silences the immune response (Zingg et al., 2017). Extracellular signal-regulated kinase (ERK) activity plays a role in immune evasion by melanoma cells, since targeting of *BRAF* and MAPK decreases production of the immunosuppressive factors IL-10, VEGF (vascular endothelial growth factor), or IL-6. Therefore, constitutive activation of the MAPK pathway not only promotes increased proliferation of melanoma cells but also is important for immune evasion of this disease (Kong et al., 2010). The PI3K/Akt pathway is often activated in melanoma because of mutations in the tumour suppressor gene *PTEN* or activation of *AKT*. Loss of functional *PTEN* in tumour cells causes *AKT* phosphorylation and activation, leading to reduced apoptosis or amplified mitogenic signalling (Kong et al., 2010). Low expression level of *PTEN* has been observed in melanoma samples (Zhou et al., 2000), perhaps as a consequence of inactivation by epigenetic silencing, altered subcellular localisation, or

ubiquitination. In metastatic melanoma, a higher percentage of *PTEN* methylation is observed (Mirmohammadsadegh et al., 2006), suggesting that this gene plays a role during melanoma progression. *AKT3* is also responsible for progression (Dai et al., 2005; “Genomic Classification of Cutaneous Melanoma”, 2015; Kong et al., 2010; Madhunapantula and Robertson, 2009). High frequency of amplifications and over-expression of *AKT3* in *RAS*-mutant, *NF1*-mutant, and triple wild-type melanoma subtypes has been observed (“Genomic Classification of Cutaneous Melanoma”, 2015). *AKT*'s activity has been implicated with up-regulation of cell adhesion protein Mel-CAM, as well as high expression of *MMP-2* and *MMP-9*, through activation of nuclear factor  $\kappa$ B (NF- $\kappa$ B) (Kong et al., 2010). *MITF* is amplified in primary and metastatic melanomas, as opposed to nevi (Garraway et al., 2005). Over-expression of both *MITF* and mutant *BRAF* may induce transformation of normal melanocytes to cancerous, implicating *MITF* as an oncogene (Kong et al., 2010). *TP53* is often over-expressed in melanomas whereas expression is absent in nevi (Ragnarsson-Olding et al., 2004). Expression of *BRAF* V600E in cells induces up-regulation of *IGFBP7*, which inhibits BRAF-MEK-ERK signalling and induces senescence or apoptosis (Wajapeyee et al., 2008). Down-regulation of pro-apoptotic genes occurs promptly in the development of invasive melanoma (Jensen et al., 2007). These include tumour suppressor genes (*TPL73L* and *P53AIP1*, involved in cellular apoptosis), tumour necrosis factors and receptors (*TNFSF10*, *TNFRSF25*, *Apo-2*, *Apo-3*, *DR3*, *DR4*, *LARD*, involved in cell death) and the caspase family of proteases (fundamental role in the apoptotic pathway).

### 2.2.3 FFPE samples

The majority of studies utilising NGS for the characterisation of melanoma genome are using fresh-frozen tissue samples (Zhang et al., 2016). Nevertheless, formalin-fixed, paraffin-embedded (FFPE) tissue is the most common specimen available for molecular assays on tissue after diagnostic histopathological examination, and a number of archived samples are available for retrospective studies. The major restraint for using FFPE samples in molecular biology analyses is that nucleic acids isolated from FFPE tissue

often suffer from degradation and chemical modifications. Particularly in the case of primary melanomas, which generally have a small size at the time of diagnosis, most of the tissue is used for diagnostic evaluation, rendering an additional issue of tissue-availability. However, several protocols have been developed to improve the isolation of DNA/RNA from FFPE specimens (Pikor et al., 2011; Sengüven et al., 2014) as well as specialised library construction methods allowing NGS-based analyses starting from nucleic acids of limited quantity and poor quality (De Paoli-Iseppi et al., 2016). Studies evaluating the quality of genomic variant calling and/or gene expression quantification by NGS, based on nucleic acids isolated from FFPE specimens as compared to fresh–frozen tissue, revealed that this approach, although challenging, can produce accurate data (De Paoli-Iseppi et al., 2016; Menon et al., 2012; Spencer et al., 2013; Van Allen et al., 2014; Zhang et al., 2017).

#### 2.2.4 Discussion

In this part, the main genetic features contributing to the development of CM were presented. Marked advances in dealing with this complex disease have been achieved over the last years, due to the diligent efforts of researchers to shed light on the biological mechanisms involved in melanoma manifestation, assisted by the advent of NGS technologies. Elucidating the mechanisms underlying melanoma biology and progression can enable the development of targeted and immune-related therapeutic approaches. Still, melanoma remains one of the most lethal types of cancer. Additional understanding of the resistance to targeted therapies is crucial, and ought to remain a central aspect of cancer research.

## 2.3 Bioinformatics & Cancer genomic data

Hi-tech developments have urged the research community to consider advanced methodologies to deal with the large-scale genomic data, with the prospect of aiding clinical outcomes. Great efforts are requested to generate a global database, with all available information, aiding scientific research in the fight against disease (Bean and Hegde, 2016). This section focuses on the genomic resources and bioinformatics tools developed for the analysis of large-scale data, mainly NGS technologies.

### 2.3.1 Genomic Data Resources

From the computational biology point of view, genomics embodies the assembly and storage of massive amounts of data, involving any essential information for an organism's life processes, in digital format. This information can be available to public and general scientific community through genomic databases. Few of the numerous and widespread databases include NCBI (National Centre for Biotechnology Information) database [[www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/)], EMBL (European Molecular Biology Laboratory) database [[www.ebi.ac.uk/embl/](http://www.ebi.ac.uk/embl/)], and DDBJ (DNA Database of Japan) database [[www.ddbj.nig.ac.jp/](http://www.ddbj.nig.ac.jp/)]. Most of the databases incorporate scientific results, storing nucleic/amino acid sequences, allowing universal access to all public; essentially, the matching data incorporated is reachable and reusable by the means of genomic databases (Akhtar et al., 2017).

The NGS revolution has directed the formation of many databases that not only store the information but can relate to additional effects of a given variant. This addition can be based on previous evidence from literature search, or built-in database tools and prediction models (Bean and Hegde, 2016).

The following paragraphs present briefly several of the major public databases and repositories that can be consulted for general functional analyses, or explicit cancer research.

The Gene Expression Omnibus (GEO) [<https://www.ncbi.nlm.nih.gov/geo/>] is a public repository of NCBI, which files and dispenses full datasets of high-throughput genomic data, like microarray or NGS, succumbed by the research community globally. Apart from storage, a variety of web-based applications are available in GEO, allowing users to download and analyse the data, and extract the gene expression profiles offered (Barrett et al., 2013; Edgar et al., 2002).

The Cancer Genome Atlas (TCGA) [<https://gdc.cancer.gov/>] represents a revolutionary multidisciplinary cancer genomics programme, where over 20,000 primary malignancies from 33 distinct cancers were molecularly characterised, along with matched normal samples. TCGA programme belongs to the National Cancer Institute's (NCI) Genomic Data Commons (GDC) next generation cancer knowledge network, whose major goals include the importation and tuning of genomic and clinical data, the coordination of sequence data (genome or transcriptome), and the standardisation of state-of-the-art approaches for manifold data, like variant calling or gene expression. Ultimate aim is to deliver a unified data repository for cancer research, through enabling driver identification, therapy outcome, and various resources for storage, quality regulation, data integration, and redeployment of cancer genomic sets (Grossman et al., 2016). Amongst these resources lies cBioPortal for Cancer Genomics (Cerami et al., 2012; Gao et al., 2013), a free-access, open-source repository for the exploration and visualisation of most cancer sets included in GDC, aiming to abridge the complexity of genomic data by offering high-quality admission to molecular profiles and translational clinical applications.

The Gene Ontology (GO) [<http://geneontology.org/>] resource offers a computational depiction of the up-to-date scientific information regarding gene function from various organisms; humans to bacteria. GO permits genes' functional annotation through the incorporation of biomedical knowledge on the molecular and cellular level of an organism, or closely related phylogenetic families. GO is allied to many other similar ontologies, underpinning computer science applications in biology and medicine (Ashburner et al., 2000; The Gene Ontology Consortium, 2019).

Reactome [<https://reactome.org/>] is a free access, manually curated, peer-reviewed pathway database, aiming to deliver bioinformatics approaches for visualising, clarifying and investigating molecular mechanism information to clinical research and systems biology. Emphasis is given on signalling and metabolic reactions, and molecule inter-relations, extracting biological pathways and processes, through evidence supported by literature (Fabregat et al., 2018).

dbNSFP (database for Nonsynonymous SNPs' Functional Predictions) [<https://sites.google.com/site/jpopgen/dbNSFP>] is an established resource specializing on functional prediction and annotation of all potential non-synonymous SNVs located on the human genome. Currently, dbNSFP includes more than 84 million non-synonymous SNVs and splice site SNVs (splicing-site SNVs). Prediction scores are compiled by various prediction algorithms and databases (Liu, Jian, et al., 2013; Liu, Wu, et al., 2016).

The Catalogue of Somatic Mutations in Cancer (COSMIC) [<https://cancer.sanger.ac.uk/cosmic>] is the largest database for somatic mutation evidence concerning human cancers and is manually curated by proficient researchers. COSMIC comprises of millions of mutations, incorporating thousands of cancer types and subtypes. It contains the Cancer Gene Census list, with detailed disseminations and effects of driver mutations, and curated annotations for cancer genomes in the direction of target discovery. COSMIC updates occur every three months ("COSMIC database", n.d.; Forbes et al., 2017).

The Network of Cancer Genes (NCG) [<http://ncg.kcl.ac.uk/>] is a manually curated cancer database, containing information based on the literature. This information is evaluated, providing evidence on the experimental validation backing up the addition of cancer genes; including annotations of gene functionality. Currently, more than 2300 genes are included in NCG (An et al., 2016; Repana et al., 2019).

Genomic data resources can assist the scientific community on the integration and analysis of varied large-scale projects. Their contribution on functional annotations and relationship disclosure of the diverse features included is irreplaceable for combating



disease and general clinical research (Akhtar et al., 2017). At this stage, focus was given on the databases used in this dissertation.

### 2.3.2 Bioinformatic cancer genomic data analysis exploiting NGS

As mentioned, nowadays, NGS has become the state-of-the-art tool in cancer research and is the most common and advanced technology for *de novo* somatic mutation detection. NGS technologies are in continuous development and improvement, both at the level of the applied protocols for library preparation and sequencing chemistry, but also at the bioinformatics level. A large number of bioinformatics tools have been developed for general pre-processing and basic analysis of NGS (WES/WGS) data with the aim of revealing altered variants for the cases under investigation. This part of the thesis focuses on tools developed for somatic mutation calling, which can be described as the first level of analysis, bypassing those needed to reach this step of the analysis. Furthermore, we present most of the available tools for driver-mutation identification, including the approaches that are used to achieve this step. Discriminating driver from passenger mutation remains a challenge from the experimental as well as the bioinformatics points of view (Gonzalez-Perez et al., 2013; Gonzalez-Perez and Lopez-Bigas, 2012; Hodis et al., 2012; Lawrence et al., 2013; Raphael et al., 2014). In the case of melanoma, which is one of the cancers with the highest mutation burdens and heterogeneity, this problem is even more difficult to address, due to the confounding impact of melanoma's high mutation rate. More detailed evaluations and comparisons of the tools are available by (Raphael et al., 2014; Xu, 2018; Zhang et al., 2014).

The basic approach for somatic variance identification is to compare paired samples, i.e. analyse matched tumour-normal samples collected from the same patient. Most callers are structured after this notion and use different approaches to extract the desired list of variants, meeting certain criteria. Among the strategies utilised are heuristic approaches combined with statistical tests, analysis and evaluation of a joint genotype likelihood, allele frequency or haplotype-based analyses, or exploitation of machine learning methods for variant classification. Apart from these, there are specialised tools

that offer single-sample somatic mutation calling (lack of normal samples), through association with databases like COSMIC (“COSMIC database”, n.d.; Forbes et al., 2017) and application of machine learning and statistical algorithms. Table 2 lists most somatic mutation callers based on their aforementioned strategic approaches.

*Table 2: Somatic variance calling tools*

<b>Analysis Tactic</b>	<b>Variant Callers</b>
Heuristic approaches	qSNP (Kassahn et al., 2013), RADIA (Radenbaugh et al., 2014), Shimmer (Hansen et al., 2013), SOAPsnv (“SOAP :: Short Oligonucleotide Analysis Package”, n.d.), VarDict (Lai et al., 2016), VarScan2 (Koboldt et al., 2012)
Joint genotype analysis	CaVEMan (Jones et al., 2016), FaSD-somatic (Wang, Wang, et al., 2014), JointSNVMix2 (Roth et al., 2012), SAMtools (Li, 2011), Seurat* (Christoforides et al., 2013), SNVSniffer (Liu, Loewer, et al., 2016), SomaticSniper (Larson et al., 2012), Virmid (Kim et al., 2013)
Allele frequency	deepSNV (Gerstung et al., 2012), EBCall (Shiraishi et al., 2013), LoFreq (Wilm et al., 2012), LoLoPicker (Carrot-Zhang and Majewski, 2017), MuTect (Cibulskis et al., 2013), Strelka (Saunders et al., 2012)
Haplotype analysis	FreeBayes (Garrison and Marth, 2012), HapMuC (Usuyama et al., 2014), LocHap (Sengupta et al., 2016), MuTect2 (Cibulskis et al., 2013), Platypus (Rimmer et al., 2014)
Machine Learning	BAYSIC (Cantarel et al., 2014), MutationSeq (Ding et al., 2012), SNooPer (Spinella et al., 2016), SomaticSeq (Fang et al., 2015)
Single-sample analysis	GATKcan (Hsu et al., 2017), ISOWN (Kalatskaya et al., 2017), OutLyzer (Muller et al., 2016), Pisces (Dunn et al., 2018), SiNVICT (Kockan et al., 2017), SomVarIUS (Smith et al., 2016)
Structural or Copy Number variation calling	APOLOH (Yang et al., 2013), BIC-Seq (Xi et al., 2010), BreakDancer (Chen et al., 2009), Break-Pointer (Drier et al., 2013), CNVkit (Talevich et al., 2016), CoNIFER (Krumm et al., 2012), Delly (Rausch et al., 2012), HYDRA (Malhotra et al., 2013), GASV (Sindi et al., 2009), GASVPro (Sindi et al., 2012), Meerkat (Yang et al., 2013), PeSV-Fisher (Escaramís et al., 2013), VariationHunter-CommonLaw (Hormozdiari et al., 2011)
RNA-seq variant calling	eSNVdetect (Tang et al., 2014), SNPiR (Piskol et al., 2013), VarDict (Lai et al., 2016), VarScan2 (Koboldt et al., 2012)

As a latter step, after obtaining a list of somatic mutations, it is important to distinguish the driver mutations which actively contribute to carcinogenesis (Stratton et al., 2009). This step can be described as the second level of analysis. Driver-mutation discrimination can be accomplished through mutation frequency analysis, functional impact investigation or machine learning algorithms based on known sets of

driver/passenger genes. Another approach followed is enrichment analysis on known pathways or networks. Table 3 summarises several tools which focus on driver mutation identification, classified by the strategic approach used. It is important to mention that distinction of driver/passenger genes faces many challenges mostly due to lack of annotation, additive effects of passenger mutations or a possible change in roles during cancer progression and the development of tumour heterogeneity (Zhang et al., 2014).

*Table 3: Driver mutation calling tools*

<b>Analysis Tactic</b>	<b>Driver Callers</b>
Functional impact	CanPredict (Kaminker, Zhang, Watanabe, et al., 2007), Condell (González-Pérez and López-Bigas, 2011), FATHMM (Shihab et al., 2013), GERP++ (Davydov et al., 2010), GOSS (Kaminker, Zhang, Waugh, et al., 2007), MutationAssessor (Reva et al., 2007, 2011), MutationTaster (Schwarz et al., 2014), Oncodrive-fm (Gonzalez-Perez and Lopez-Bigas, 2012), PMUT (Ferrer-Costa et al., 2005), PolyPhen-2 (Adzhubei et al., 2010), PROVEAN (Choi et al., 2012), SIFT (Ng and Henikoff, 2001), SNPs3D (Yue et al., 2006), TransFIC (Gonzalez-Perez et al., 2012)
Mutation frequency	DrGaP (Hua et al., 2013), MuSiC (Dees et al., 2012), MutSig /MutSigCV (Lawrence et al., 2013), Youn <i>et al.</i> (Youn and Simon, 2011)
Machine Learning	CHASM (Carter et al., 2009; Wong et al., 2011), DMI (Tan et al., 2012)
Structural or Copy Number focus	ADMIRE (van Dyk et al., 2013), CMD5 (Zhang et al., 2010), GISTIC2 (Mermel et al., 2011), JISTIC (Sanchez-Garcia et al., 2010)
Positional/Structural clustering	iPAC (Ryslik et al., 2013), NMC (Ye et al., 2010)
Pathway/Network analysis	BioInfoMiner (Koutsandreas et al., 2016), Dendrix (Vandin et al., 2011), GSEA (Subramanian et al., 2005), HotNet (Vandin et al., 2011), MEMo (Ciriello et al., 2012), Multi-Dendrix (Leiserson et al., 2013), NetBox (Cerami et al., 2010), PathScan (Wendl et al., 2011), Patient-oriented gene sets (Boca et al., 2010), RME (Miller et al., 2011)

# Chapter 3. Materials & Methods

The techniques and methodologies mentioned in the previous chapter offer auspicious results for the detection of melanoma. Dermoscopy images are low-cost and extensively available, permitting a feasible option for early diagnosis. The success rates given by literature show that the work towards automated classification of lesions and melanoma may provide decent results; still there is always room for improvement. Dermoscopy images are frequently accompanied by numerous irregularities and huge deviations between specimens. For this reason, it is critical to find the appropriate means to overcome these oddities and attain a truthful finding (Mishra and Celebi, 2016). Image processing constitutes an important aspect of disease confrontation. Coupled with the biological facet, together they can contribute to an exemplary outcome, enabling the simultaneous co-optation of therapeutic approaches. Towards this idea, this thesis offers a multi-layered approach, bridging the aforementioned aspects, a concept lacking in existing research.

A number of methods were utilised at the different stages of this study:

- Molecular techniques for the extraction of biological material
- Bioinformatics techniques for NGS analysis and biomarker detection
- Statistical analysis techniques to examine the correlation of different groups of data features
- Supervised machine learning techniques for the classification

A more comprehensive outline of the analysis performed is given by Figure 3. The basis of this study is the experimental and bioinformatic analysis of WES data, deriving from new patients. Through exome and transcriptomic data integration (top rectangle of

Figure 3, pink shade), a broad molecular network implicated in CM is given, elucidating the important mechanisms involved in this type of cancer; lists of significant genes and mechanisms is the output. Specifically, for the newly analysed WES data, the list of extensively analysed mutated genes acts as input for classification and CM detection – genes as potential biomarkers. Through integration with imaging features (list acquired from data previously analysed by collaborator) and creation of classification systems (see bottom rectangle of Figure 3, green and grey shades) for the layered scheme, another output is the list of potential composite biomarkers.

The following sections of this chapter present the methodology and approaches used for this study, describing the algorithms and tools used in detail, with the corresponding hyper-parameter settings.

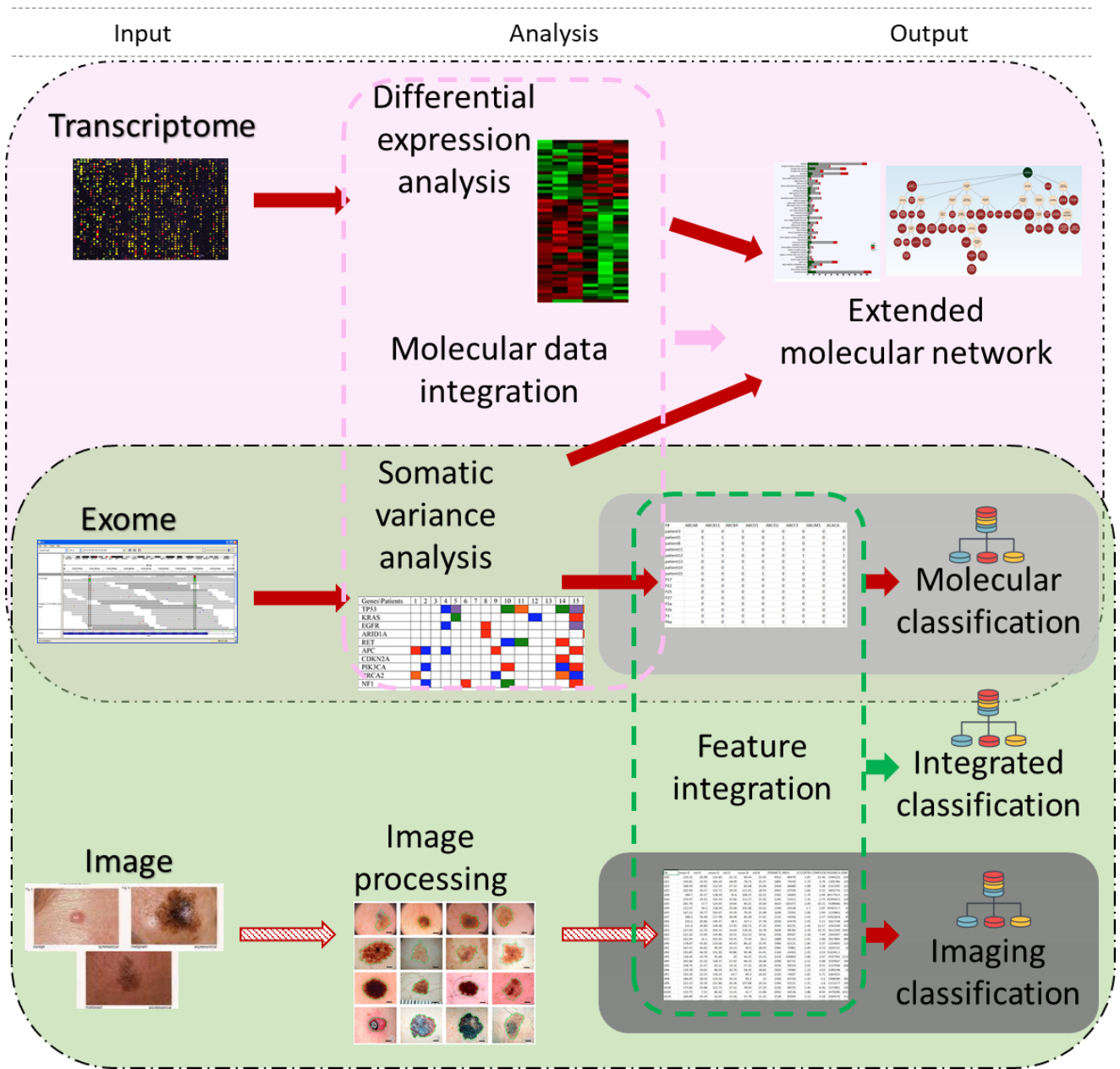


Figure 3: Comprehensive processing pipeline of the presented research

## 3.1 Biological material and molecular analysis

### 3.1.1 Melanoma Samples

All samples were acquired in the context of the 12CHN-204 PROMISE project [12CHN 204 Entrepreneurial Program Competitiveness & Entrepreneurship & Districts in Transition, (Action Bilateral Research and Technology Cooperation between Greece and China) “Personalisation of melanoma therapeutic management through the fusion of systems biology and intelligent data mining methodologies-PROMISE”], under the strict conformity to the rules of the call. The samples derived from FFPE tissue blocks from excisional biopsies histo-pathologically confirmed as melanomas. Areas from tumour and adjacent healthy nevus tissue were assessed and separated by a pathologist. Paired tissue samples, tumour and normal individually, from nine patients, both male and female, with cutaneous melanoma were collected. All patients had no reported family history of melanoma and all examined melanoma tissues were from the primary lesion. More information concerning the patients and excised lesion can be found at Table 4 and Appendix Tables A1.

Table 4: Patient and melanoma lesion characteristics, Border R-regular, I-irregular

Patient No.	AGE	SEX	PRE-X NEVUS	ASYMMETRY	BORDER	SIZE INCREASED	DIAMETER	COLOR CHANGE	SITE
3	52	♀	yes	yes	I	yes	>5mm	yes	waste
5	82	♂	yes	no	R	yes	>5mm	yes	back
8	80	♂	no	yes	I	yes	>5mm	yes	head
10	77	♀	no	yes	I	yes	>5mm	yes	subungual (foot)
11	72	♂	yes	yes	I	yes	>5mm	yes	back
12	56	♀	no	yes	I	yes	>5mm	yes	head

13	69	♀	no	yes	R	yes	>5mm	no	back
14	73	♀	yes	yes	I	yes	>5mm	NA	tibia
15	38	♂	no	no	R	yes	>5mm	no	abdomen

---

### 3.1.2 DNA Extraction and Exome Sequencing

DNA was isolated from the samples using QIAamp DNA FFPE Tissue protocol from QIAGEN (Hilden, Germany), with several modifications to deparaffinisation, washing and proteinase K digestion steps, to ensure better quality and higher quantity of the extracted DNA. More specifically, deparaffinisation of FFPE samples was performed with xylene (×2 times) at 56 °C for 3 min and the precipitate was sequentially washed with 100%, 70% and 50% ethanol (Sengüven et al., 2014). Proteinase K digestion was performed at 56 °C while stirring the samples and the incubation time was increased to 3 days with daily re-addition of proteinase K. The quantity and purity of the samples were checked using Nanophotometer (IMPLEN, Munich, Germany). The extracted DNA was prepared and captured with the Agilent SureSelect Human All Exon 50 Mb kit (Agilent SureSelect v5, Santa Clara, CA, USA) and whole exome sequencing was performed on an Illumina HiSeq 4000 sequencer (San Diego, CA, USA), as paired-end (PE) reads.

## 3.2 Variant calling & mutational biomarker discovery

The following analysis was performed utilising various state-of-the-art tools. Unless stated otherwise, all bioinformatics analyses were performed using shell interface command line, on a Linux-based 64GB RAM/ 12 processor cluster server. Whole exome sequencing was performed on 18 DNA samples (paired normal/tumour from each patient) with an average of 14911,30 Mb of raw bases given by the Illumina sequencer. After removing low-quality reads, we obtained on average 93500006 clean reads (12953,84 Mb). The clean reads of each sample had high Q20 and Q30, which showed high



sequencing quality. The average GC content was 47.88%. The complete tables of quality control for the whole exome sequencing procedure can be found at Appendix Tables A2. The clean reads were received in fastq format, 2 files per sample (PE reads), 4 per patient (paired normal/tumour analysis).

The first step of the analysis was to align the reads from the fastq files to the reference genome (hg19, version b37), using BWA (Burrows-Wheeler Aligner) (Li and Durbin, 2010) for DNA reads, version 0.7.5, adjusted for paired-end sequencing and ran in consecutive steps for finding the correct coordinates and generate the final alignment per sample, in SAM format. Figure 4 includes the 3 commands given to BWA so as to perform the alignment for each sample.

**index**      `bwa index [-p prefix] <in.db.fasta>`  
 Index database sequences in the FASTA format.

**-p**      Prefix of the output database [same as db filename]

**aln**      `bwa aln [-n maxDiff] [-o maxGapO] [-e maxGapE] [-d nDelTail] [-i nIndelEnd] [-k maxSeedDiff] [-t nThrds] [-cRN] [-M misMsc] [-O gapOsc] [-E gapEsc] <'@RG\tID:group1\tSM:sample \tPL:illumina\tLB:lib1\tPU:unit1'> <in.db.fasta> <in.query.fq> > <out.sai>`  
 Find the SA coordinates of the input reads. Maximum *maxSeedDiff* differences are allowed in the first *seedLen* subsequence and maximum *maxDiff* differences are allowed in the whole sequence.

**-n**      Maximum edit distance given, or the fraction of missing alignments given 2% uniform base error rate [0.04]

**-o**      Maximum number of gap opens [1]

**-e**      Maximum number of gap extensions, -1 for k-difference mode (disallowing long gaps) [-1]

**-d**      Disallow a long deletion within INT bp towards the 3'-end [16]

**-i**      Disallow an indel within INT bp towards the ends [5]

**-k**      Maximum edit distance in the seed [2]

**-t**      Number of threads (multi-threading mode) [6]

**-M**      Mismatch penalty [3]

**-O**      Gap open penalty [11]

**-E**      Gap extension penalty [4]

**-R**      Proceed with suboptimal alignments for paired-end mapping.

**sampe**      `bwa sampe [-a maxInsSize] [-o maxOcc] [-n maxHitPaired] [-N maxHitDis] [-P] <in.db.fasta> <in1.sai> <in2.sai> <in1.fq> <in2.fq> > <out.sam>`  
 Generate alignments in the SAM format given paired-end reads. Repetitive read pairs will be placed randomly.

**-a**      Maximum insert size for a read pair to be considered being mapped properly [500]

**-o**      Maximum occurrences of a read for pairing [100000]

**-n**      Maximum number of alignments to output in the XA tag for reads paired properly [3]

**-N**      Maximum number of alignments to output in the XA tag for dis-concordant read pairs (excluding singletons) [10]

*Figure 4: Commands to run BWA, specific parameters used are given in brackets*

The second step was performed using Picard (“Picard Tools - By Broad Institute”, n.d.), version 1.98. This step was required for the pre-processing of the aligned reads, to ensure the reads were in the correct format for further analysis. This included sorting sequences based on the reference sequence, marking duplicate reads and building an index for the output sample file, which allows fast look-up of data, essential for supplementary steps. Figure 5 includes the 3 commands given to Picard for the pre-processing of each sample, with the output acquiring BAM format.

```
java -jar picard.jar SortSam INPUT=input.sam OUTPUT=sorted.bam SORT_ORDER=coordinate
java -jar picard.jar MarkDuplicates INPUT=sorted.bam OUTPUT=marked_duplicates.bam
METRICS_FILE=marked_dup_metrics.txt
java -jar picard.jar BuildBamIndex INPUT=sorted.bam
```

*Figure 5: Commands to run Picard*

Third step was processing the reads with GATK (Genome Analysis Toolkit) (McKenna et al., 2010), to certify the good quality of reads (all reads are given quality scores and can be dismissed if necessary) and perform realignments and recalibrations based on the scores and references (commands given on Figure 6), to optimise the output reads and permit the succeeding variance and somatic mutation investigation. For this study, version 3.6 of GATK was used, which incorporates somatic SNP calling with somatic indel (insertions & deletions) calling, as carried out by MuTect2, based on the original MuTect (Cibulskis et al., 2013) and Indelocator (“Indelocator”, n.d.), comparing the tumour-normal pairs in order to characterise somatic mutations. MuTect2 permits varying allelic fraction for each variant, as is often seen in tumours with lower purity, multiple subclones, and/or copy number variations. It also incorporates information from COSMIC database (“COSMIC database”, n.d.; Forbes et al., 2017), annotating previously described somatic mutations from preceding studies. Germline variants were identified using the HaplotypeCaller tool, by comparing the normal samples with the reference sequence. This tool can call SNPs and indels simultaneously, through performing local *de novo* assembly of haplotypes at a given active region. Specific coding SNPs were investigated from a known panel of germline variants associated with melanoma based on GWAS studies and established databases (Antonopoulou et al., 2015; Kypreou et al., 2016; MacArthur et al., 2016), focusing on those found on coding regions. Basic coding to perform HaplotypeCaller and MuTect2 analyses is given in Figure 7. Both tools output lists of sites describing the altered alleles, along with specific coordinates on the DNA, various quality scores and precise quantifications for each sample, in vcf format. Strand-specific artefacts, i.e. SNPs that the alternate allele was not supported by both forward and reverse orientation of the DNA, were considered as false positives- possibly due to DNA damage resulting from formalin fixation and storage time- and were excluded

manually from the results. The aforementioned analysis was restricted on the exome region, where possible, utilising the target coordinates of the exome that were used during the sequencing procedure for capturing DNA (Agilent SureSelect v5, S04380110\_v5 in bed format file), to ensure more accurate results and faster realisation of the workflow. Most of the GATK tools that were used require assistance to distinguish true variants from false positives or known sites of variation, which is given by the synchronised use of additional resources like the 1000 Genomes Project ('1000 Genomes | A Deep Catalog of Human Genetic Variation' n.d.) and dbSNP database ("Home - SNP - NCBI", n.d.), build\_132.b37.

```
java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R reference.fasta -I input.bam --known gold_standard.1000G_phase1.indels.b37.vcf
-o forIndelRealigner.intervals.vcf -L [restrict processing to specific genomic intervals, exome regions only] S04380110_v5.bed
```

<u>--maxIntervalSize</u>	500	maximum interval size; any intervals larger than this value will be dropped
<u>--minReadsAtLocus</u>	4	minimum reads at a locus to enable using the entropy calculation
<u>--mismatchFraction</u>	0.0	fraction of base qualities needing to mismatch for a position to have high entropy
<u>--windowSize</u>	10	window size for calculating entropy or SNP clusters

```
java -jar GenomeAnalysisTK.jar -T IndelRealigner -R reference.fasta -I input.bam -known gold_standard.1000G_phase1.indels.b37.vcf
-targetIntervals intervalListFromRTC.intervals -o realignedBam.bam -L [restrict processing to exome regions only] S04380110_v5.bed
```

<u>--entropyThreshold</u>	0.15	Percentage of mismatches at a locus to be considered having high entropy (0.0 < entropy <= 1.0)
<u>--maxConsensuses</u>	30	Max alternate consensuses to try (necessary to improve performance in deep coverage)
<u>--maxSizeForMovement</u>	3000	maximum insert size of read pairs that we attempt to realign
<u>--maxPositionalMoveAllowed</u>	200	Maximum positional move in basepairs that a read can be adjusted during realignment
<u>--maxReadsForConsensuses</u>	120	Max reads used for finding the alternate consensuses (necessary to improve performance in deep coverage)
<u>--maxReadsForRealignment</u>	20000	Max reads allowed at an interval for realignment
<u>--maxReadsInMemory</u>	150000	max reads allowed to be kept in memory at a time by the SAMFileWriter

```
java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R reference.fasta -I my_reads.bam -known gold_standard.1000G_phase1.indels.b37.vcf
-knownSites dbsnp_build132.b37.vcf -o recal_data.table -L [restrict processing to exome regions only] S04380110_v5.bed
```

<u>--indels_context_size</u>	3	Size of the k-mer context to be used for base insertions and deletions
<u>--maximum_cycle_value</u>	500	The maximum cycle value permitted for the Cycle covariate
<u>--mismatches_context_size</u>	2	Size of the k-mer context to be used for base mismatches

```
java -jar GenomeAnalysisTK.jar -T PrintReads -R reference.fasta -I input.bam -BQSR [apply the recalibration using] recal_data.table -o output.bam
```

*Figure 6: Commands to run GATK RealignerTargetCreator, IndelRealigner, Base Recalibrator and PrintReads in consecutive steps*

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R reference.fasta -l sample.bam [--dbsnp dbsnp_build132.b37.vcf] [-stand_call_conf 30]
[-stand_emit_conf 10] -L S04380110_v5.bed -o output.raw.snps.indels.vcf
```

<u>--heterozygosity</u>	0.001	Heterozygosity value used to compute prior likelihoods for any locus
<u>--heterozygosity_stdev</u>	0.01	Standard deviation of heterozygosity for SNP and indel calling
<u>--indel_heterozygosity</u>	0.00125	Heterozygosity for indel calling
<u>--maxReadsInRegionPerSample</u>	10000	Maximum reads in an active region
<u>--min_base_quality_score</u>	10	Minimum base quality required to consider a base for calling
<u>--minReadsPerAlignmentStart</u>	10	Minimum number of reads sharing the same alignment start for each genomic location in an active region
<u>--sample_ploidy</u>	2	Ploidy per sample
<u>--stand_call_conf</u>	10.0	The minimum phred-scaled confidence threshold at which variants should be called

```
java -jar GenomeAnalysisTK.jar -T MuTect2 -R reference.fasta -l:tumor tumour.bam -l:normal normal.bam [--dbsnp dbsnp_build132.b37.vcf]
[--cosmic COSMIC.vcf] -L S04380110_v5.bed -o output.vcf
```

<u>--dbsnp_normal_lod</u>	5.5	LOD threshold for calling normal non-variant at dbsnp sites
<u>--heterozygosity</u>	0.001	Heterozygosity value used to compute prior likelihoods for any locus
<u>--heterozygosity_stdev</u>	0.01	Standard deviation of heterozygosity for SNP and indel calling.
<u>--indel_heterozygosity</u>	0.00125	Heterozygosity for indel calling
<u>--initial_normal_lod</u>	0.5	Initial LOD threshold for calling normal variant
<u>--initial_tumor_lod</u>	4.0	Initial LOD threshold for calling tumor variant
<u>--max_alt_allele_in_normal_fraction</u>	0.03	Threshold for maximum alternate allele fraction in normal
<u>--max_alt_alleles_in_normal_count</u>	1	Threshold for maximum alternate allele counts in normal
<u>--max_alt_alleles_in_normal_qscore_sum</u>	20	Threshold for maximum alternate allele quality score sum in normal
<u>--maxReadsInRegionPerSample</u>	1000	Maximum reads in an active region
<u>--min_base_quality_score</u>	10	Minimum base quality required to consider a base for calling
<u>--minReadsPerAlignmentStart</u>	5	Minimum number of reads sharing the same alignment start for each genomic location in an active region
<u>--normal_lod</u>	2.2	LOD threshold for calling normal non-germline
<u>--pir_mad_threshold</u>	3.0	threshold for clustered read position artifact MAD
<u>--pir_median_threshold</u>	10.0	threshold for clustered read position artifact median
<u>--power_constant_qscore</u>	30	Phred scale quality score constant to use in power calculations
<u>--sample_ploidy</u>	2	Ploidy per sample. For pooled data, set to (Number of samples in each pool * Sample Ploidy).
<u>-stand_call_conf</u>	10.0	The minimum phred-scaled confidence threshold at which variants should be called
<u>--tumor_lod</u>	6.3	LOD threshold for calling tumor variant

*Figure 7: Commands to run GATK HaplotypeCaller and MuTect2*

The next stage of the analysis was to annotate the resulting sites, SNPs or somatic mutations, using Oncotator (Ramos et al., 2015), which utilises several databases to link the sites to specific genes. Oncotator is a web application, as well as stand-alone tool, for annotating human genomic point mutations and indels with data relevant to cancer. For this analysis, we used stand-alone Oncotator version 1.5.1.0, which requires the download of multiple database annotations in a specific folder, to perform genomic,

protein, cancer and non-cancer variant annotations, by specifying the input file, its format, the genomic build, and gives the result in MAF text format.

Significantly mutated genes among the patients were identified using MutSigCV (version 1.41) (Lawrence et al., 2013), which ranks the genes by estimating a background mutation rate (BMR) through the number of silent versus non-coding mutations in the gene and the surrounding regions. This BMR model is not constant and changes due to patient- and genomic position-based factors. Figure 8 shows the command for MutSigCV.

```
run_MutSigCV.sh [my_mutations.maf] [exome_full192.coverage.txt] [gene.covariates.txt] my_results
[mutation_type_dictionary_file.txt] [chr_files_hg19]
```

mutation table file	Mutation list in <u>Mutation Annotation Format (MAF)</u> .
coverage table file	the number of sequenced bases in each patient, per gene per mutation category.
covariates table file	Covariates table in a tab-delimited text file.
output filename base	Base name for the output files.
mutation type dictionary	The mutation type dictionary to use for automatic category and effect discovery.
genome build	Genome build to use for automatic category and effect discovery.

*Figure 8: Commands to run MutSigCV*

BioInfoMiner (Koutsandreas et al., 2016) was used for the functional analysis of the mutated genes, so as to identify the molecular pathways influenced by these mutations, and to isolate the genes with central role, implicated in diverse and major mechanisms from various vocabularies; for this analysis GO (Ashburner et al., 2000; The Gene Ontology Consortium, 2019) and Reactome (Fabregat et al., 2018) were utilised. BioInfoMiner combines the StRAnGER2 (Pilalis and Chatziioannou, 2013) and GOrevenge (Moutselos et al., 2011) algorithms and is an online tool. Figure 9 shows a draft experiment on BioInfoMiner, requiring a gene list, species and database specifications, and statistical cut-off. The output can be lists and illustrations of pathways and genes with a significant role emerging from the original list.

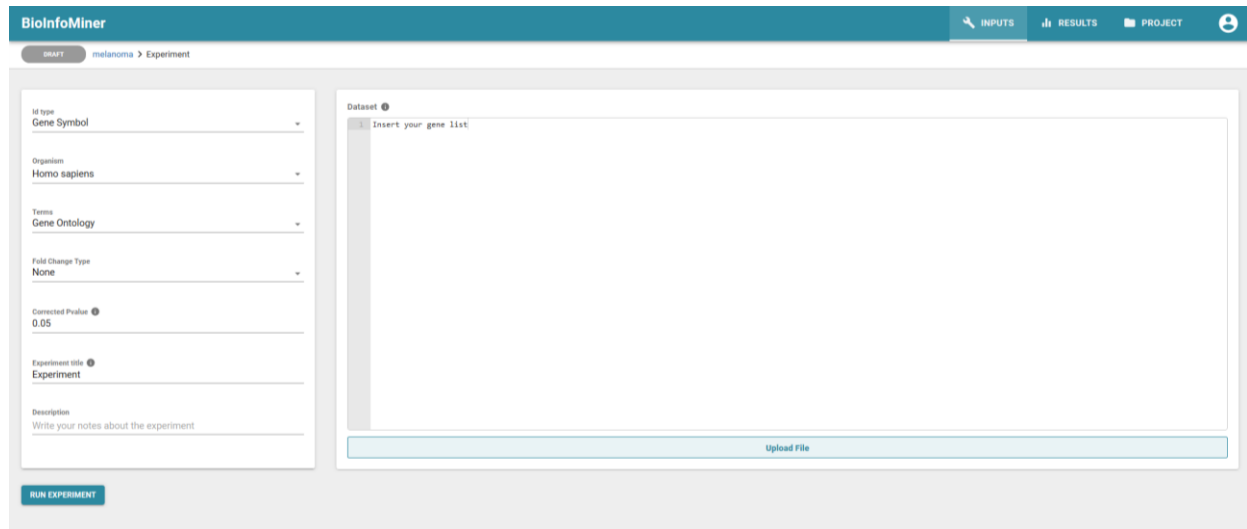


Figure 9: BioInfoMiner online tool

For the extended molecular analysis (chapter 4.2), we performed additional investigations and database searches. For the functional prediction of the somatic mutations we utilised dbNSFP (Liu, Wu, et al., 2016) through Oncotator, a database that provides information for functional predictions and annotations for human non-synonymous variants. Copy number variation (CNV) analysis was performed using CNVkit (Talevich et al., 2016), version 0.8.5, which specialises on CNV detection on targeted DNA sequencing (including WES). CNVkit also runs on the exome region, utilising the target coordinates of the exome (Agilent SureSelect v5, S04380110\_v5 in bed format file) and anti-targets to avoid, to ensure more accurate results. Figure 10 presents the commands used to run CNVkit in consecutive steps, first creating target and anti-target regions, then merging the normal samples in one 'global' reference for copy numbers and finally comparing each pair of patient samples, incorporating this global reference.



```
cnvkit.py target S04380110_Regions.bed --annotate refFlat.txt --split -o S04380110_Regions.bed

cnvkit.py antitarget my_targets.bed -g data/access-5kb-mappable.hg19.bed -o my_antitargets.bed

cnvkit.py coverage normal.bam S04380110_Regions.bed -o normal.targetcoverage.cnn

cnvkit.py reference *coverage.cnn -f reference.fa -o Reference.cnn

cnvkit.py batch tumour*T.bam --normal normal*N.bam --targets S04380110_Regions.bed --fasta reference.fa --output-reference my_reference.cnn --output-dir results_cnvkit/ --diagram --scatter
```

*Figure 10: Commands to run CNVkit in consecutive steps*

Expression of mutated genes was evaluated through TCGA's cBioPortal datasets (Cerami et al., 2012; Gao et al., 2013). Genes with TPM (transcripts per million) between 0.5 and 10 were considered to have low expression, between 11 and 1000 medium, and if TPM was over 1000, the genes were considered as highly expressed in a given case. Overall, genes were considered as expressed when encompassing at least low expression in over 30% of the cases.

Figure 11 presents the workflow of analysis starting from raw whole exome sequencing data for the identification of variance and somatic mutations concerning the disease under investigation. Final output is a list of genes -potential biomarkers- that are affected and play a crucial role in the manifestation of melanoma.

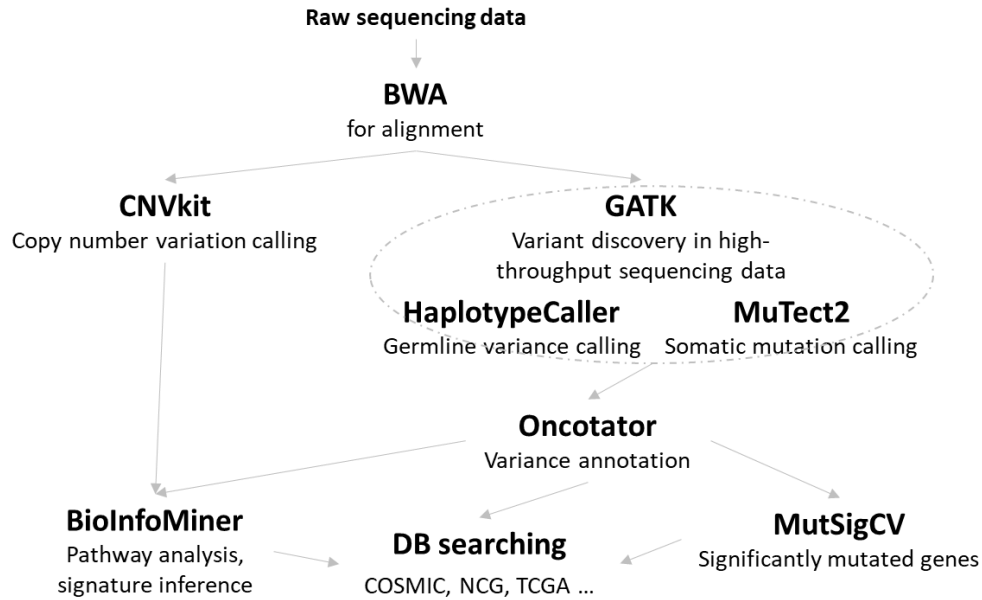


Figure 11: Workflow of analysis for the identification of variance and somatic mutations

### 3.2.1 NGS complexity and information

The complexity of NGS data is high, due to the high amount of information contained in each separate sample (compressed ~10GB per sample/ 20GB per patient/ ~150GB for all) and the fact that several distinct parameters need to be adjusted at each step, so as to optimise the performance and the quality of the results (i.e., BWA needs to be adjusted for paired-end sequencing and run in consecutive steps to find the correct coordinates and generate the final alignment in proper format). It is worth noting that the complete workflow for a pair of samples (tumour and normal samples from one patient) needs approximately 35 hours running time, summarising the results in ~10MB.

## 3.3 Transcriptomics microarray data analysis

The microarray dataset was downloaded from NCBI's Gene Expression Omnibus (GEO) database (Barrett et al., 2013). Transcriptomic differential expression analysis was performed on R programming environment (Development, n.d.) version 3.5.1, using



## 3.4 Machine learning techniques for data integration & classification

Machine learning analysis was performed utilising various algorithms, i.e. classification was performed using random forests, Gaussian linear modelling, stochastic gradient boosting, decision trees, linear discriminant analysis, support vector machines, k-nearest neighbours and logistic regression. For feature selection the topological prioritisation of Bioinforminer was used, as well as entropy-based (information gain and gain ratio) criteria. Unless stated otherwise, all further analyses were performed on R programming environment (Development, n.d.) version 3.5.1, using RStudio version 1.0.136, on a Windows-based 8GB RAM/ 4 processor/64-bit personal computer.

### 3.4.1 Mutational data

The concept is to build a classifier exploiting the data that were produced. Since the number of patients analysed was limited, melanoma samples from TCGA database were added, through cBioPortal (Cerami et al., 2012; Gao et al., 2013), to expand the lists of samples. As healthy state (non-melanoma) mutational data from dysplastic nevus that were acquired through similar experimental procedure (Melamed et al., 2017) were used. For feature selection, we reduced the list of mutated genes by prioritising them according to their centrality (genes taking part in numerous distinct mechanisms are ranked higher), using BioInfoMiner (Koutsandreas et al., 2016). The samples (samples of dysplastic nevus and melanoma) were separated under two labels, dysplastic nevus (*represented by DNS*) and melanoma (*represented by MEL*) and each sample is attributed a multi-dimensional binary vector, showing if the corresponding gene contains a mutation or not. To deal with unbalanced classes, the SMOTE (Chawla et al., 2002) algorithm was utilised to generate synthetic data for the DNS label. Several classification algorithms (random forests, Gaussian linear modelling, stochastic gradient boosting, decision trees, linear discriminant analysis, support vector machines, k-nearest neighbours and logistic regression) were examined, to find the one with the best outcome. An exhaustive grid

search for fine-tuning of classification parameters was performed. Figures Figure 13- Figure 15 present the commands used for this part of the study.

```
#Let's create observations using SMOTE. We set perc.over = 100 to double the quantity of positive cases, and set
perc.under=200 to keep half of what was created as negative cases.

library(DMwR)
input$class <- as.factor(input$class)
input_SM <- SMOTE(class ~ ., input, perc.over = 100, perc.under=200)

print(table(input$class))
#DNS MEL
#121 121
```

*Figure 13: Commands to run SMOTE in R*

```
library(caret); library(e1071); library(glmnet); library(klaR); library(C50); library(gbm)
set.seed(seed) #seed=7
##Algorithms
# Linear Discriminant Analysis
fit.lda <- train(class~, data=input, method="lda", metric=metric, trControl=control)
# Logistic Regression
fit.glm <- train(class~, data=input, method="glm", metric=metric, trControl=control)
# GLMNET
fit.glmnet <- train(class~, data=input, method="glmnet", metric=metric, trControl=control)
# SVM Radial
fit.svmRadial <- train(class~, data=input, method="svmRadial", metric=metric, trControl=control, fit=FALSE)
# kNN
fit.knn <- train(class~, data=input, method="knn", metric=metric, trControl=control)
# CART
fit.cart <- train(class~, data=input, method="rpart", metric=metric, trControl=control)
# C5.0
fit.c50 <- train(class~, data=input, method="C5.0", metric=metric, trControl=control)
# Bagged CART
fit.treebag <- train(class~, data=input, method="treebag", metric=metric, trControl=control)
# Random Forest
fit.rf <- train(class~, data=input, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalised Boosted Modeling)
fit.gbm <- train(class~, data=input, method="gbm", metric=metric, trControl=control, verbose=FALSE)

results <- resamples(list(lda=fit.lda, logistic=fit.glm, glmnet=fit.glmnet, svm=fit.svmRadial, knn=fit.knn, c50=fit.c50, rf=fit.rf,
gbm=fit.gbm))
```

*Figure 14: Commands used to build the classifiers in R*

```

library(caret)
set.seed(7)
## split dataset in train-test, with equal representation of classes
splitIndex <- createDataPartition(input$class, p = .50,
                                  list = FALSE,
                                  times = 1)
trainSplit_input <- input[ splitIndex,]
testSplit_input <- input[-splitIndex,]

prop.table(table(trainSplit_input$class))
#DNS MEL
#0.5 0.5

prop.table(table(testSplit_input$class))
#DNS MEL
#0.5 0.5

rfmodel <- train(class ~ ., data = trainSplit_input, method = "rf", metric=metric, trControl = ctrl)

predictors <- names(trainSplit_input)[names(trainSplit_input) != 'class']
pred <- predict(rfmodel,type = "prob",testSplit_input)

library(pROC)
auc <- roc(testSplit_input$class, pred[,2])
print(auc)

#Call:
# roc.default(response = testSplit_input$class, predictor = pred_mo[, 2])
#Data: pred[, 2] in 60 controls (testSplit_input$class DNS) < 60 cases (testSplit_input$class MEL).
#Area under the curve: 0.9289 ## output for the molecular dataset

```

*Figure 15: Commands to create Random Forests classifier and test its accuracy in R*

The packages used to run the aforesaid analysis were caret (Kuhn 2015), DMwR (Torgo 2016), pROC (Robin et al. , e1071 (Meyer et al., 2019), glmnet (Friedman et al., 2018), klaR (Roever et al., 2018), C50 (Kuhn et al., 2018), gbm (Greenwell et al., 2019), ROCR (Sing et al., 2015), cvAUC (LeDell et al., 2014).

### 3.4.2 Imaging data & Integration

Concerning skin imaging, data from Moutselos et al. (Moutselos et al., 2014) were acquired, where 1041 dermoscopy samples of 32 features are characterised. The feature list included is based on the ABCD-rule of dermatology (Border, Colour, Textural features)

after analysing the 1041 (69 MEL/ 972 DNS) images. To deal with unbalanced classes, as before, the SMOTE algorithm was utilised, this time to generate synthetic data for the MEL label. Classification algorithms were explored as described above.

In order to validate the proof of concept of the proposed design a final integrative dataset, containing molecular and imaging data for the melanoma case, was explored. More specifically, a synthetic dataset was constructed to incorporate images from different nevi (dysplastic or melanomas) together with molecular measurements, integrating the two datasets from before. Again, a classification system was explored. Several classification algorithms were examined (random forests, Gaussian linear modelling, stochastic gradient boosting, decision trees, linear discriminant analysis, support vector machines, k-nearest neighbours and logistic regression), to find the one with the best outcome. An exhaustive grid search for fine-tuning of classification parameters was performed and comparisons were made between this and the ones described before.

A challenge that may arise when dealing with such an analysis is overfitting, i.e., data fitting the training set well, but achieving poor performance in the validation set. This might be the case when building a complex risk prediction model with the inclusion of many biomarkers. Information gain and gain ratio were used to prioritise the integrated biomarkers and comparisons were made between the classification systems. Those are entropy-based metrics that express the amount of information contained in a given attribute, which can characterise one class from another.

# Chapter 4. Analysis & Results

## 4.1 Initial molecular analysis and integration

The complexities of cellular metabolism and regulatory pathways involved have, until recently, obstructed the formulation of a unified description for melanoma (Dummer and Hoek, 2004). Thus, despite the descent of gene signatures for various cancers, e.g. breast or colon cancer, a similar progress remains elusive for malignant melanoma. This could be attributed to the intricate nature of the molecular basis of cutaneous melanoma, which needs neatly stratified epidemiological cohorts to effectively address the issue of the high heterogeneity of this disease. Anyhow, melanoma genomic studies are limited by the availability and quality of the biological material and therefore are relatively sparse. In any case, genomic studies are limited by the shortage of similar melanoma cohorts, collecting and maintaining frozen tumour tissue, therefore rendering gene expression profiling studies of melanoma relatively scarce (Winnepeninckx et al., 2006). Still, efforts have been made to overcome any issues and shed some light on the underlying mechanisms associated with melanoma pathogenesis and metastases (Raskin et al., 2013; Winnepeninckx et al., 2006). A number of important emerging biological pathways and gene targets recently identified in melanoma are reported in (Dutton-Regester and Hayward, 2012). Key biological pathways, where several significant genes (e.g. *CDKN2A*, *CDK4*, *RB1*) are involved, include proliferation, transcriptional control, extracellular matrix remodelling, glutamate signalling, and apoptosis.

In this part of the study, focus was given on integrating different levels of molecular data through functional analysis to improve our understanding of the underlying mechanisms involved in melanoma. Established microarray datasets were incorporated with next generation sequencing mutational data creating a potential disease network for melanoma. This stage was presented in (Kontogianni et al., 2016).

The following sections describe the results deriving from the analysis of next generation sequencing and transcriptomic data (methodology parts 3.2, 3.3). Particularly,



focus was given on the NGS analysis results and the integration with the transcriptomic dataset.

#### 4.1.1 Initial molecular analysis

WES data derived from tumour and normal samples were aligned to the human genome, with an average sequence coverage of > 100x (number of reads aligning to known reference bases), ideal for achieving the mutational profile required. Overall, the individual samples have depth of coverage > 90, with only one sample achieving a lower score. Still this lower score is found only in normal sample, which does not affect further analysis, since high coverage is necessary mainly for the tumour samples, to overcome endogenous heterogeneity. Table 5 shows the number of putative sites of somatic mutations, after the MuTect analysis, as well as the count of missense and nonsense mutations for each patient. These mutations affect gene products, by amino acid substitutions or protein truncation, and require further analysis as candidate genetic biomarkers. Figure 16 shows how the patients are grouped based on their mutational profiles, using Ward's criterion (Ward, 1963) for hierarchical clustering.

*Table 5: Number of somatic mutations, missense/nonsense mutations, and unique genes affected per patient*

<b>patient</b>	<b>Sites of somatic mutations</b>	<b>Missense/Nonsense Mutations</b>	<b>Unique genes affected</b>
<b>3</b>	855	224	214
<b>5</b>	1134	309	295
<b>8</b>	826	281	265
<b>10</b>	73	10	10
<b>11</b>	944	275	265

<b>12</b>	5985	1811	1474
<b>13</b>	812	226	200
<b>14</b>	922	224	219
<b>15</b>	1111	224	214

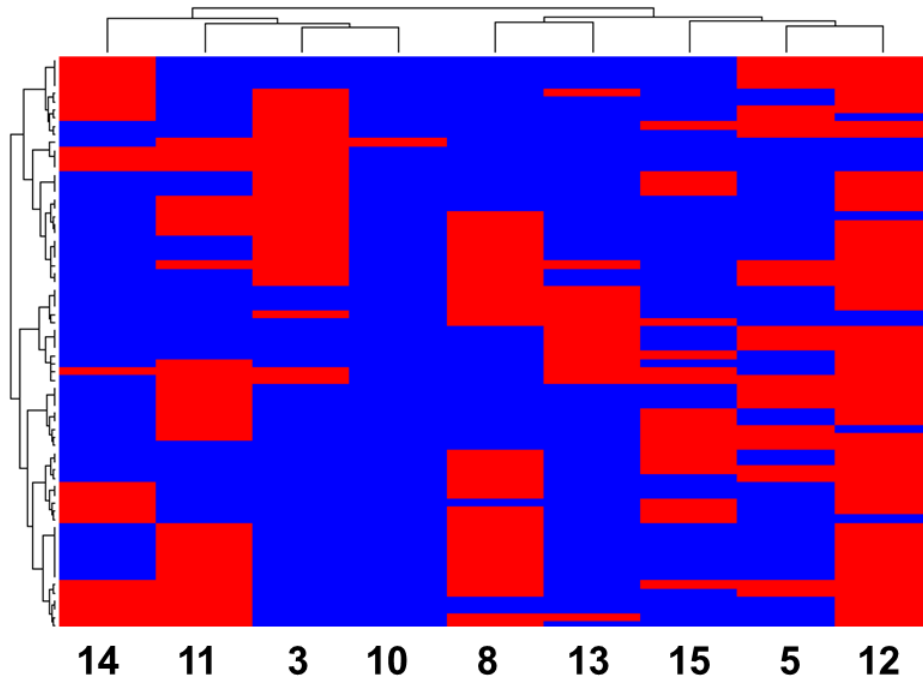


Figure 16: Hierarchical clustering of the patients, based on the mutational profiles, red-present mutation, blue-absent

#### 4.1.2 Pathway analysis

In order to discover the molecular pathways affected by the specific mutations, after annotating the mutations to specific genes, we performed functional analysis of the union of affected genes from all the patients (on 2685 unique genes), which revealed 40 statistically significant biological processes (p-value < 0.05), shown in Table 6.

*Table 6: Table of the significant biological processes influenced by the mutated genes. Enrichment represents the ratio of the number of genes in the input list annotated with a GO term to the total number of genes annotated to this specific term, Hypergeometric and Corrected p-values represent the statistic score used for ranking the terms, given by BioInfoMiner*

<b>Term id</b>	<b>Term Definition</b>	<b>Enrichment</b>	<b>Hypergeometric p-value</b>	<b>Corrected p-value</b>
<b>GO:0007156</b>	homophilic cell adhesion via plasma membrane adhesion molecules	69/150	4.33E-20	0.0014
<b>GO:0007155</b>	cell adhesion	148/531	2.17E-15	0.0027
<b>GO:0050911</b>	detection of chemical stimulus involved in sensory perception of smell	105/389	1.68E-10	0.0037
<b>GO:0030198</b>	extracellular matrix organisation	84/313	1.38E-08	0.0048
<b>GO:0086010</b>	membrane depolarization during action potential	17/30	1.10E-07	0.0063
<b>GO:0007411</b>	axon guidance	95/375	3.64E-08	0.0068
<b>GO:0006811</b>	ion transport	82/319	1.53E-07	0.0101
<b>GO:0022617</b>	extracellular matrix disassembly	39/117	2.98E-07	0.0108
<b>GO:0006814</b>	sodium ion transport	37/106	1.59E-07	0.0115
<b>GO:0055085</b>	transmembrane transport	162/767	7.19E-07	0.012
<b>GO:0007608</b>	sensory perception of smell	61/224	7.06E-07	0.0125
<b>GO:0019228</b>	neuronal action potential	16/31	1.42E-06	0.0144
<b>GO:0035725</b>	sodium ion transmembrane transport	30/89	5.11E-06	0.0145
<b>GO:0007268</b>	synaptic transmission	97/428	5.75E-06	0.0178
<b>GO:0042391</b>	regulation of membrane potential	36/117	6.79E-06	0.0195
<b>GO:0007186</b>	G-protein coupled receptor signalling pathway	192/976	9.61E-06	0.0198
<b>GO:0030574</b>	collagen catabolic process	26/74	9.12E-06	0.0203
<b>GO:0007605</b>	sensory perception of sound	39/133	1.03E-05	0.0223
<b>GO:0034765</b>	regulation of ion transmembrane transport	35/118	2.18E-05	0.0257
<b>GO:0060080</b>	inhibitory postsynaptic potential	8/11	2.26E-05	0.0257
<b>GO:0070588</b>	calcium ion transmembrane transport	38/129	1.19E-05	0.0258
<b>GO:0018108</b>	peptidyl-tyrosine phosphorylation	37/130	3.53E-05	0.0287
<b>GO:0016339</b>	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	13/27	3.64E-05	0.0306
<b>GO:0070509</b>	calcium ion import	13/28	5.89E-05	0.0323
<b>GO:0007018</b>	microtubule-based movement	24/74	8.72E-05	0.0331

<b>GO:0001539</b>	cilium or flagellum-dependent cell motility	6/7	5.96E-05	0.034
<b>GO:0032228</b>	regulation of synaptic transmission, GABAergic	7/10	0.0001	0.0353
<b>GO:0007399</b>	nervous system development	72/322	0.0001	0.0376
<b>GO:0007169</b>	transmembrane receptor protein tyrosine kinase signalling pathway	33/119	0.0002	0.0382
<b>GO:0034220</b>	ion transmembrane transport	65/286	0.0002	0.0395
<b>GO:0001964</b>	startle response	10/20	0.0002	0.0399
<b>GO:0050907</b>	detection of chemical stimulus involved in sensory perception	28/96	0.0002	0.0405
<b>GO:0007416</b>	synapse assembly	17/47	0.0002	0.0445
<b>GO:0071625</b>	vocalization behaviour	7/12	0.0006	0.0447
<b>GO:2000821</b>	regulation of grooming behaviour	4/4	0.0005	0.0455
<b>GO:0016337</b>	single organismal cell-cell adhesion	30/109	0.0003	0.0465
<b>GO:0030534</b>	adult behaviour	12/29	0.0004	0.0468
<b>GO:0034332</b>	adherent junction organisation	14/38	0.0006	0.0476
<b>GO:0034329</b>	cell junction assembly	22/76	0.001	0.0492
<b>GO:0015721</b>	bile acid and bile salt transport	11/27	0.0009	0.0493

The transcriptomic analysis from the microarray dataset revealed 1425 unique differentially expressed genes. Enrichment analysis showed 36 statistically significant biological processes (p-value < 0.05), which are presented in Table 7.

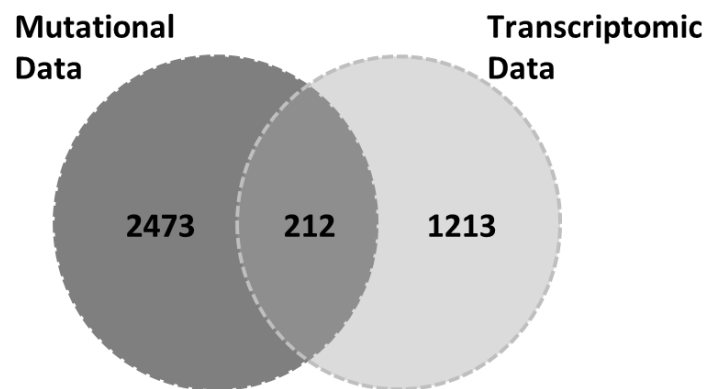
*Table 7: Table of the significant biological processes influenced by the differentially expressed genes. Enrichment represents the ratio of the number of genes in the input list annotated with a GO term to the total number of genes annotated to this specific term, Hypergeometric and Corrected p-values represent the statistic score used for ranking the terms, given by BioInfoMiner*

<b>Term id</b>	<b>Term Definition</b>	<b>Enrichment</b>	<b>Hypergeometric p-value</b>	<b>Corrected p-value</b>
<b>GO:0030198</b>	extracellular matrix organisation	66/313	0.00000676	0.0014
<b>GO:0008544</b>	epidermis development	31/109	0.00000027	0.0033
<b>GO:0030216</b>	keratinocyte differentiation	19/56	0.000003067	0.0043
<b>GO:0006094</b>	gluconeogenesis	16/48	0.00002341	0.0053

<b>GO:0048013</b>	ephrin receptor signalling pathway	21/91	0.0005	0.0078
<b>GO:0060512</b>	prostate gland morphogenesis	4/4	0.0001	0.0079
<b>GO:0033599</b>	regulation of mammary gland epithelial cell proliferation	4/5	0.0006	0.0094
<b>GO:0045861</b>	negative regulation of proteolysis	9/26	0.0011	0.0114
<b>GO:0061436</b>	establishment of skin barrier	7/17	0.0012	0.0116
<b>GO:0060326</b>	cell chemotaxis	15/57	0.0008	0.0132
<b>GO:0071230</b>	cellular response to amino acid stimulus	13/48	0.0013	0.0155
<b>GO:0051591</b>	response to cAMP	14/54	0.0013	0.0157
<b>GO:0048538</b>	thymus development	12/45	0.0022	0.0182
<b>GO:0045669</b>	positive regulation of osteoblast differentiation	14/57	0.0023	0.0199
<b>GO:0001954</b>	positive regulation of cell-matrix adhesion	8/23	0.0019	0.021
<b>GO:0042060</b>	wound healing	20/95	0.0024	0.022
<b>GO:0007155</b>	cell adhesion	78/531	0.0028	0.0235
<b>GO:0061036</b>	positive regulation of cartilage development	6/15	0.0032	0.0236
<b>GO:0022617</b>	extracellular matrix disassembly	23/117	0.003	0.025
<b>GO:0045765</b>	regulation of angiogenesis	9/30	0.0033	0.027
<b>GO:0071526</b>	semaphorin-plexin signalling pathway	7/20	0.0036	0.0292
<b>GO:0048661</b>	positive regulation of smooth muscle cell proliferation	13/54	0.004	0.0298
<b>GO:0050773</b>	regulation of dendrite development	5/11	0.0038	0.0313
<b>GO:0048678</b>	response to axon injury	9/32	0.0053	0.0337
<b>GO:0010951</b>	negative regulation of endopeptidase activity	26/144	0.0056	0.0343
<b>GO:0061621</b>	canonical glycolysis	8/27	0.0059	0.0346
<b>GO:0070373</b>	negative regulation of ERK1 and ERK2 cascade	12/50	0.0057	0.0374
<b>GO:0055086</b>	nucleobase-containing small molecule metabolic process	16/78	0.008	0.0402
<b>GO:0007160</b>	cell-matrix adhesion	18/92	0.0084	0.0402
<b>GO:0060441</b>	epithelial tube branching involved in lung morphogenesis	6/17	0.0066	0.0405
<b>GO:0030032</b>	lamellipodium assembly	9/33	0.0066	0.0407

<b>GO:0030324</b>	lung development	20/106	0.0086	0.0435
<b>GO:0002009</b>	morphogenesis of an epithelium	7/23	0.0084	0.045
<b>GO:0043153</b>	entrainment of circadian clock by photoperiod	6/18	0.009	0.0454
<b>GO:0007266</b>	Rho protein signal transduction	13/59	0.0087	0.0465
<b>GO:0030855</b>	epithelial cell differentiation	16/79	0.009	0.0483

To facilitate a deeper examination of the datasets, we compared the gene lists from the mutational and transcriptomic analyses. Figure 17 illustrates the total unique and common genes, from the two types of datasets. Only 5% of the total genes were common between the two sets. Nevertheless, among the highly ranked processes, presented in tables Table 6 Table 7, cell adhesion, extracellular matrix organisation and extracellular matrix disassembly, all containing large numbers of genes, are found as significantly affected in both cases.



*Figure 17: Venn diagram for the significant gene lists from the two analyses*

In order to create a feasible disease network for melanoma, the previous results were merged together, and additional functional analysis was carried out. This enrichment analysis revealed 45 statistically significant biological processes ( $p$ -value < 0.05), presented in Figure 18, ranked according to their corrected  $p$ -values.

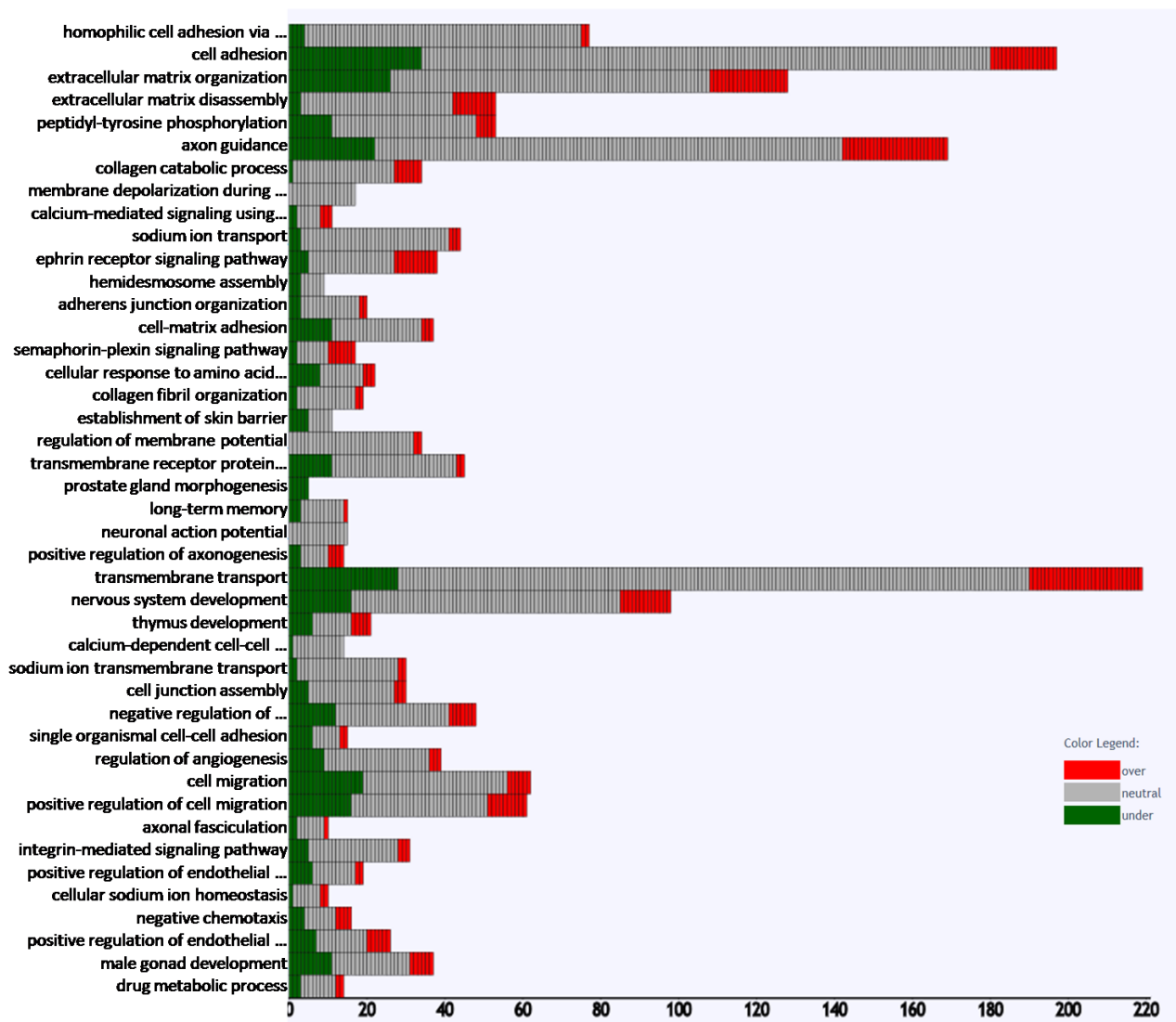
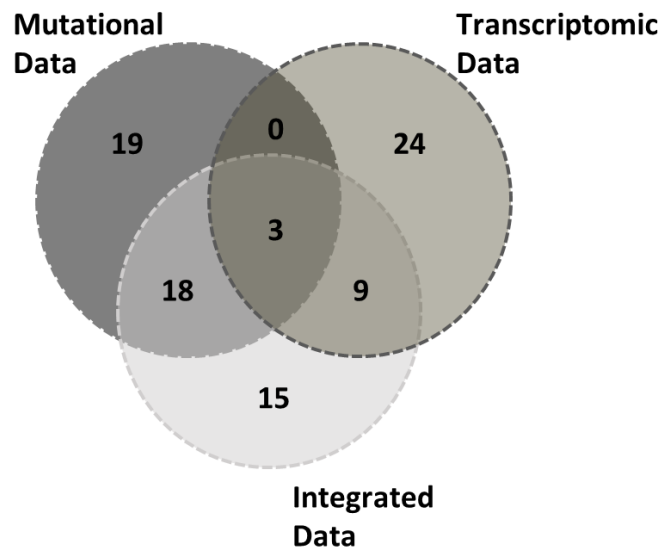


Figure 18: Bar plot of significant terms with the number of associated genes (x-axis). Terms are ranked using the corrected p-value. The colours of the genes specify their expression fold change, green -on the left- for under-expressed genes and red -on the right- for over-expressed genes, neutral indicating somatic mutation

This potential Disease Network revealed several mechanisms with known significance, consistent with melanoma. Enrichment of GO terms, such as epithelial tube branching involved in lung morphogenesis, morphogenesis of an epithelium, epithelial cell differentiation, and regulation of mammary gland epithelial cell proliferation reflects the topological origin of cutaneous melanoma (Jogi et al., 2012; Martin-Belmonte and

Perez-Moreno, 2011). Furthermore, cell-matrix procedures (organisation, adhesion) have been previously reported as significantly altered in tumours (Hart et al., 1991; Saladi et al., 2010), as well as lamellipodium assembly, an essential structure for cell migration, which plays an important part in cell invasion and metastasis of cancer (Kato et al., 2014; Machesky, 2008). In relation to the ephrin receptor and Rho protein signalling pathways, the Eph receptor tyrosine kinases and their ephrin ligands have specific expression patterns in cancer cells (Pasquale, 2010), while Rho-like GTPase have been identified as key regulators of epithelial architecture and cell migration, both correlated to cancer development (Ridley, 2004; Sander and Collard, 1999).

As expected, the previously discussed significant pathways from tables Table 6 Table 7 are complemented by the additional data, incorporating an increased number of genes, with considerable implication in melanoma manifestation and progression. Among the significant processes, there are several previously highlighted by the distinct datasets, but also a number of newly generated, after data integration. Figure 19 indicates the unique and common pathways in each case.



*Figure 19: Venn diagram for the significant pathway lists from the two distinct analyses, as well as their integration*



### 4.1.3 Discussion

In this stage, we sought to export the broader molecular network implicated in cutaneous melanoma. We integrated molecular data of different levels in order to identify the important mechanisms that are involved in this type of cancer. This integration advanced our understanding about the mechanisms involved in melanoma, by observing the correlation between different sets and levels of data. More importantly, it allowed the manifestation of additional mechanisms previously concealed by the statistical cut-offs, thus enhancing the disease network and our general understanding of the phenomenon.

## 4.2 Extended molecular analysis

CM development is a complex, multi-factorial process involving the interplay of genetic and environmental risk factors. The most well-established environmental risk factor is the exposure to ultraviolet radiation (UVR) (Nikolaou and Stratigos, 2014). Regarding the genetic background, several susceptibility genes have been identified, including highly penetrative genes such as *CDKN2A*, the first familial melanoma gene identified (Hussussian et al., 1994; Kamb et al., 1994) which is found mutated in approximately 40% of melanoma high-density families. Other less frequent mutations have been identified in genes of high or more moderate penetration, including *CDK4* and the more recently described *BAP1*, *TERT*, *POT1*, *ACD*, *TERF2IP* and *MITF* (Aoude et al., 2015). Genome-wide association studies (GWASs) have also revealed numerous recurring single nucleotide polymorphisms (SNPs) associated with melanoma risk (Antonopoulou et al., 2015; Athanasiadis et al., 2014; Chatzinasiou et al., 2011; Law et al., 2015).

In the last decade, important steps towards characterising the somatic mutational landscape of melanoma have been achieved (Dutton-Regester and Hayward, 2012; Walia et al., 2012). Identifying causative melanoma mutations is of great importance in order to understand the molecular basis of melanoma genesis and progression. Towards this end, next generation sequencing (NGS) technologies are a valuable tool and have

been exploited in a number of recent studies, comparing sequencing data from melanoma tissue and a matched normal control in order to identify somatic mutations (Dutton-Regester and Hayward, 2012; Krauthammer et al., 2012; Nikolaev et al., 2011; Stark et al., 2012, 2015; Wei et al., 2011). In such approaches, discriminating driver mutations from passenger ones remains a challenge from both the experimental as well as the bioinformatics point of view (Gonzalez-Perez et al., 2013; Gonzalez-Perez and Lopez-Bigas, 2012; Hodis et al., 2012; Lawrence et al., 2013; Raphael et al., 2014). Especially in the case of melanoma, which is one of the cancers with the highest mutation burden and heterogeneity, this problem is even more difficult to address, due to the confounding impact of melanoma's high mutation rate.

In this part, the aim was to exploit FFPE samples for the identification of somatic mutations and germline variants in patients with primary melanomas by exome sequencing and perform a more thorough search than the one described in the previous section. Towards this, analysis was repeated, and additional steps were included (e.g. CNV identification), all presented in detail in the following subsection. Regarding the genetic factors involved in melanoma susceptibility, a number of studies concerning melanoma patients from the Greek population have been reported (Kypreou et al., 2016; Law et al., 2015). Still, the characterisation of melanoma somatic mutations in Greek patients has been limited and, to the best of my knowledge, this was the first attempt to characterise the somatic mutational profile at the exome level of primary melanoma patients in Greece. This thorough analysis is presented in (Kontogianni et al., 2018).

#### 4.2.1 Sequencing Data Analysis

For this analysis, the paired tumour and surrounding normal skin FFPE tissue from nine patients with cutaneous melanoma after excisional biopsy were used. The whole exome sequencing (WES) data were aligned to the human genome, with an average alignment rate of >91%, an average sequence coverage of >100× and over 96% of targets with at least 20× coverage, enabling the achievement of the intended mutational profile (Cibulskis et al., 2013; Majewski et al., 2011; Sims et al., 2014; Van Allen et al.,

2014). Only one sample attained a lower score in terms of coverage (normal sample for patient 8), but this did not affect further analysis, since high coverage is necessary in the tumour samples to overcome endogenous heterogeneity. Table 8 summarises the calculated alignment scores and sequencing depths for all the samples.

*Table 8: Sequencing characteristics for all the samples.*

<b>Patients</b>	<b>3</b>	<b>5</b>	<b>8</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>
<b>Normal</b>									
Alignment rate (%)	96	84.2	65.6	93.8	96.8	96.8	97.4	87.9	95.6
Average sequencing depth on target	129.8	93.9	72.2	103.9	101.5	123.3	117.4	102.5	128.4
Fraction of target covered by >20× (%)	97.37	96.37	94.41	96.44	95.03	97.53	97	97.38	97.96
<b>Tumour</b>									
Alignment rate (%)	95.4	89.5	93.6	92.6	96.7	96.2	96.8	88.8	92
Average sequencing depth on target	118.9	100.8	102.2	111.5	111.4	104.8	111.9	102.2	120.7
Fraction of target covered by >20× (%)	97.4	95.86	96.48	96.75	96.74	96.62	96.16	96.3	96.66
Average sequencing depth on target					<b>108.7</b>				

## 4.2.2 Identification of Germline Variation

Aiming to examine whether the patients had germline variations on possible melanoma susceptibility loci, focus was given on a panel of SNPs previously reported to be associated with CM risk. In particular, a list of SNPs from the GWAS catalogue

database (MacArthur et al., 2016) enriched by putative melanoma risk SNPs based on the MelGene database (Antonopoulou et al., 2015; Kypreou et al., 2016) was assessed. The analysis was restricted to 22 SNPs located in exon regions, since all SNPs on intronic or intragenomic regions were excluded because the data was derived from exome sequencing. Table 9 demonstrates the germline SNPs associated to melanoma risk found in the analysed patients. The relevant genes include pigmentation associated genes (*SLC45A2*, *OCA2*, *TYR*), as well as cell cycle and DNA repair genes (*ATM*, *CDKN2A*, *ERCC5*). Specific melanoma susceptibility alleles (Aitken et al., 1999; Barrett et al., 2011; Gerstenblith et al., 2010; Nan et al., 2009; Schrama et al., 2011; Sturm et al., 2008) were found in a number of patients.

*Table 9: Germline single nucleotide polymorphisms (SNPs) putatively associated with melanoma, based on genome-wide association studies (GWAS) and MelGene databases.*

<b>dbSNP_ID</b>	<b>Gene</b>	<b>Chr</b>	<b>Variant Classification</b>	<b>Ref. Allele</b>	<b>MA Allele</b>	<b># Hom . Ref.</b>	<b># Hom . MA</b>	<b># Heter.</b>
<b>rs1801516</b>	<b>ATM</b>	11	Missense Mutation	G	A	6	0	3
<b>rs11515</b>	<b>CDKN2A</b>	9	3'UTR	C	G	0	7	2
<b>rs16891982</b>	<b>SLC45A2</b>	5	Missense Mutation	C	G	0	9	0
<b>rs17655</b>	<b>ERCC5</b>	13	Missense Mutation	G	C	7	0	2
<b>rs1800407</b>	<b>OCA2</b>	15	Missense Mutation	G	A	8	0	1
<b>rs1042602</b>	<b>TYR</b>	11	Missense Mutation	C	A	2	2	5

\* The corresponding gene, chromosome position, classification type, reference allele, Melanoma-associated allele and the number of patients in Homozygous/Heterozygous state, are shown in the relevant columns.

### 4.2.3 Identification of Somatic Coding Mutations

In order to identify somatic mutations, the MuTect algorithm (Cibulskis et al., 2013) was used, which detects somatic variations at low allelic fraction with high sensitivity and low false positive rate, based on the paired analysis of tumour and matched-normal sequencing data (Xu et al., 2014). A total of 10,030 somatic mutations in all patients were identified (can be found at the online version of (Kontogianni et al., 2018) as Table S1). The majority of patients had comparable numbers of somatic mutations (median: 589), with the exception of two patients; patient 12, who had a total of 5324 somatic mutations, affecting 3752 genes, and patient 10 who had 27 somatic mutations, affecting 23 genes. Table 10 displays the number of somatic mutations for all the patients, the mutation frequency per Mb, along with the number of non-synonymous mutations and the relevant number of affected genes; Table 11 reports the types of the mutations per patient. In particular, 3955 protein-altering somatic mutations were identified. Excluding patient 10, the median mutation frequency was 12.75 mutations/Mb (ranging from 10.1 to 105.7), which is in agreement with the previously reported mutation burden of melanoma genome, which is considered one of the highest among cancer genomes (Alexandrov et al., 2013). Regarding the sample from patient 10, it was the only case of acral melanoma, which has been reported to have markedly less somatic mutations. Next, the distribution of somatic substitutions per base change was analysed and all patients, except patient 10, showed a UVR characteristic mutational spectrum with a high ratio of C > T transitions (median rate 85.6%), which has been reported to characterise sun-exposed melanomas (Brash, 2015; Zhang et al., 2016) (Figure 20).

*Table 10: Somatic mutation characteristics for each patient.*

<b>Patients</b>	<b>3</b>	<b>5</b>	<b>8</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>
Number of Mutations	522	693	901	27	935	5324	511	589	528
Frequency per Mb	10.4	13.8	17.9	0.5	18.6	105.7	10.1	11.7	10.5
Non-synonymous	226	284	387	18	364	2036	222	224	193

Genes with non-synonymous 216 274 360 16 338 1634 201 218 185

Table 11: Characterisation of the somatic mutations for each patient

<b>Patients:</b>	<b>3</b>	<b>5</b>	<b>8</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>
<b>3'UTR</b>	9	9	13	14	110	9	13	12
<b>5'UTR</b>	6	12	15	22	87	10	4	11
<b>Frame Shifting Deletion</b>	1	3	2	0	8	0	1	1
<b>Frame Shifting Insertion</b>	1	1	0	3	0	1	0	4
<b>IGR</b>	13	20	14	19	110	9	22	19
<b>Intron</b>	152	209	237	277	1523	132	197	184
<b>lincRNA</b>	5	7	7	5	29	0	5	7
<b>Missense</b>	198	251	356	332	1890	200	196	168
<b>Nonsense</b>	17	10	15	18	84	12	15	8
<b>RNA</b>	10	17	21	20	94	12	10	10
<b>Silent</b>	100	135	206	214	1335	117	114	92
<b>Splice site</b>	10	14	13	7	48	6	8	10
<b>In Frame Insertion</b>	0	1	1	2	1	0	1	1
<b>In Frame Deletion</b>	0	2	0	1	1	1	1	0
<b>De Novo Start In Frame</b>	0	1	0	0	1	1	0	0
<b>De Novo Start Out of Frame</b>	0	0	0	0	0	0	0	1
<b>Start Codon SNP</b>	0	0	0	1	3	1	2	0
<b>5'Flank</b>	0	0	1	0	0	0	0	0
<b>Nonstop</b>	0	1	0	0	0	0	0	0
<b>Total:</b>	522	693	901	935	5324	511	589	528

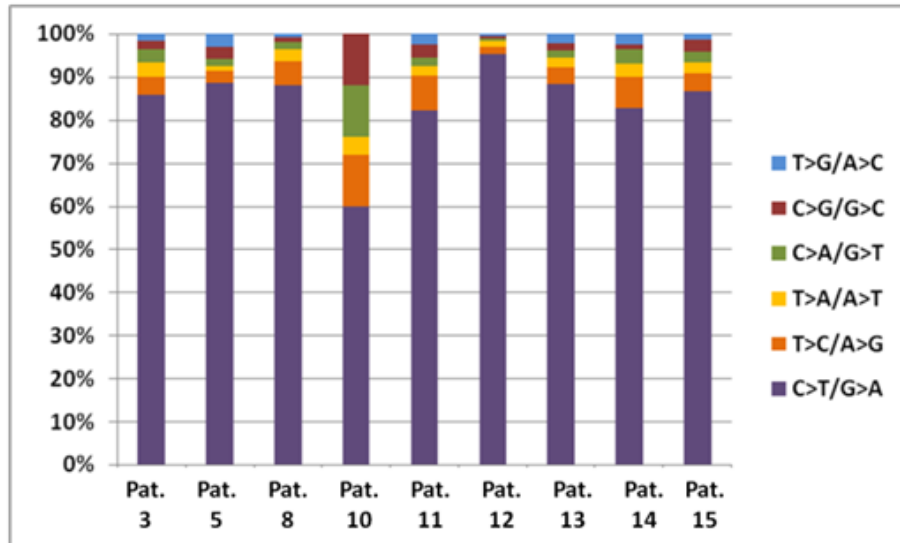


Figure 20: Mutation spectrum for each patient. C > T transitions account for 85.6% of the mutations (median rate).

#### 4.2.3.1 Characterisation of Mutated Genes and Copy Number Variations

The 10,030 mutations corresponded to 6030 unique genes, of which 2890 harboured non-synonymous mutations, most likely affecting protein functionality. Among them, 421 genes were found in at least two and 73 genes were common in at least three patients. In order to gain insights about the functional role of these common genes, the Network of Cancer Genes (NCG) (An et al., 2016) was accessed so as to identify all cancer-related genes from this 73-gene top common list. NCG is a manually curated literature-based repository containing 1571 cancer genes with either known involvement in cancer or high probability of association due to statistical analysis from numerous NGS studies. Out of the 73 genes, 33 were referred to as cancer genes according to NCG, namely *DNAH7*, *PCLO*, *TTN*, *CSMD1*, *GPR98*, *MUC16*, *PKHD1L1*, *MYOM2*, *NEB*, *RELN*, *SPHKAP*, *UNC13C*, *ADCY8*, *ANK3*, *BAI3*, *CD163L1*, *CNTN5*, *COL22A1*, *DNAH14*, *EYS*, *FAT1*, *FAT3*, *FLT1*, *GRIN2A*, *KMT2D*, *PCDH18*, *PKHD1*, *SHROOM3*, *THSD7B*, *TNC*, *BRAF*, *LRP1B* and *RYR1*. In addition, the COSMIC database (“COSMIC database”, n.d.; Forbes et al., 2017) was accessed to identify genes previously reported

in melanoma. 65 of the 73 top common genes were previously identified in melanoma cases, with a frequency over 5%. Figure 21 demonstrates that the majority of the 73 genes, found mutated in at least three patients, either belonged to the candidate cancer gene list of NCG (containing 1571 genes) or were previously reported in COSMIC with a mutation frequency of >5% in melanoma samples (1181 genes). Among the seven genes identified only in this study, WD repeat domain 87 (*WDR87*), was found mutated in seven patients (87.5%). *WDR87* is a protein coding gene, but little is known about its function and its implication in melanoma. In COSMIC, it is found mutated in 4.2% of the samples. Additionally, TCGA's cBioPortal database (Cerami et al., 2012; Gao et al., 2013) was accessed to investigate preceding discoveries for *WDR87*. This search exposed two melanoma studies, with *WDR87* mutated in 55% and 16% of the samples examined (Berger et al., 2012; Shain et al., 2015). Further analysis is needed to clarify the potential significance of the high mutation frequency observed for *WDR87* gene in the specific subjects.

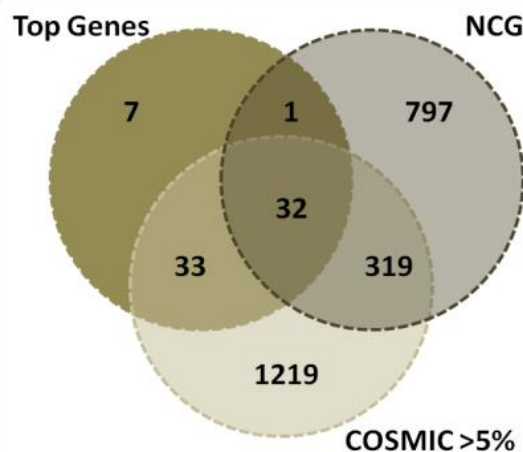


Figure 21: Common genes between the 73-genes list, the 1571 Network of Cancer Genes (NCG) genes and the >5% mutated genes in melanoma from COSMIC

Somatic copy number variation (CNV) was assessed using differences in sequence coverage between each tumour specimen and all same-sex adjacent skin



samples utilising CNVkit, a methodology which uses both on-target and off-target reads to infer copy number consistently across the genome. This analysis revealed several CNV events in genes implicated in melanoma and reported to harbour amplifications or deletions (Bastian, 2014; Krauthammer et al., 2012). Specifically, *CDKN2A* (9p21) presented a deletion signal on three of the patients, and *PTEN* (10q23) on two. In addition, *CCND1* (11q13.3) and *MITF* (3p13) were amplified in two patients and *BRAF* (7q34) was amplified in one patient (Figure 22).

Next, we searched the whole list of 2890 genes found to contain non-synonymous mutations in at least one patient, exploring the COSMIC database which contains data of somatic mutations for specific cancer types but also data for genes causally implicated in cancer. The notable melanoma-associated mutation *BRAF V600E* was detected in three patients, while *RAS* mutations were not detected. Among the mutated genes, only *BRAF*, *CTNNB1*, *NF1* and *TP53* carried specific mutations that have been previously reported in COSMIC (in more than 15 cases), as shown in Table 12. Two criteria to characterise the genes carrying non-synonymous mutations in this study were used; the frequency of a gene found mutated in melanoma and the characterisation of a gene as cancer census. Specifically, we searched for genes mutated in melanomas with a frequency >20% and in addition Cancer Census genes reported as mutated in melanoma with a frequency >5%, both based on COSMIC (Figure 22). Furthermore, the MutSigCV algorithm was used to identify significantly mutated genes, incorporating patient-specific mutation frequency with gene expression and replication time data. The small sample size prevents statistical significance in the results; still, the algorithm offers valuable information, by prioritising genes with putative significant mutations (denoted with \* in Figure 22, top-20 genes in Table 13), mainly after correcting for gene-specific mutation rates. It should be noted that among the most frequently mutated genes in these results, there were several constantly found mutated in cancer (e.g., *PCLO*, *TTM*) that are considered non-oncogenic (Lawrence et al., 2013). Still, focusing only on the top melanoma census genes from COSMIC, the majority of them are also mutated in the analysed cases (Figure 23).

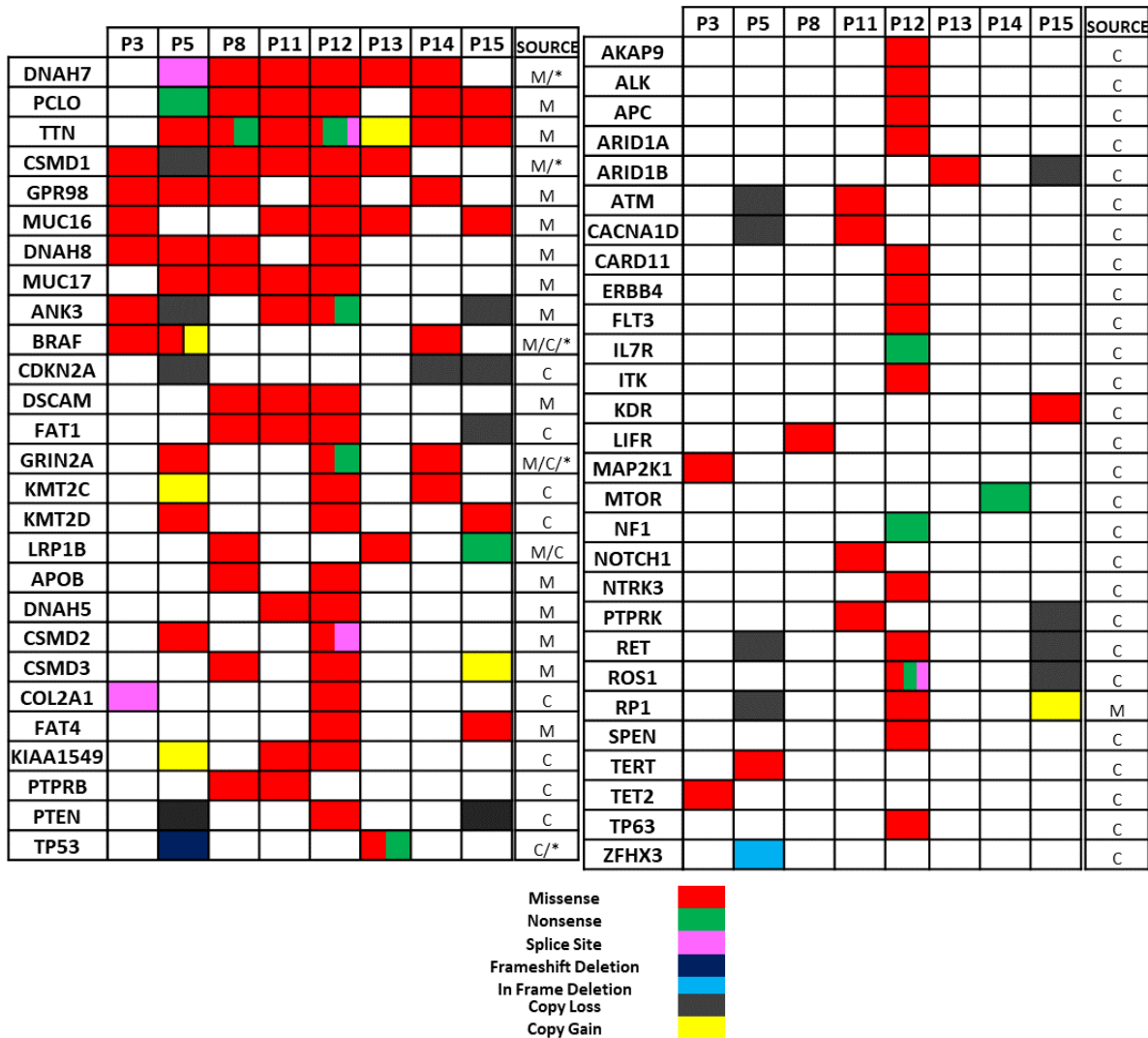


Figure 22: Characterisation of genes carrying non-synonymous mutations based on COSMIC data. M-melanoma-associated genes (>20% mutated in COSMIC) and C-cancer-census genes (>5% in melanoma samples) \*—denotes genes highlighted by MutSigCV.

Table 12: Recurrent mutations, based on COSMIC (v84) (n>15)

Hugo_Symbol	Chr	Variant_Classification	Patient	cDNA_Change	Protein_Change	COSMIC count (v84)
BRAF	7	Missense_Mutation	14	c.1799T>A	p.V600E	27133
BRAF	7	Missense_Mutation	3	c.1799T>A	p.V600E	27133
BRAF	7	Missense_Mutation	5	c.1799T>A	p.V600E	27133
BRAF	7	Missense_Mutation	5	c.1798G>A	p.V600M	31
CTNNB1	3	Missense_Mutation	15	c.109T>G	p.S37A	76

NF1	17	Nonsense_Mutation	12	c.4084C>T	p.R1362*	16
TP53	17	Nonsense_Mutation	13	c.438G>A	p.W146*	49

Table 13: Top 20 mutated genes, given by MutSigCV

rank	gene	Pat. 3	Pat. 5	Pat. 8	Pat. 11	Pat. 12	Pat. 13	Pat. 14	Pat. 15	sum	p-value
1	CYP4X1	0	0	0	0	1	0	1	0	2	0.00010
2	TP53	0	1	0	0	0	1	0	0	2	0.00018
3	GPC5	0	1	0	0	1	0	0	0	2	0.00065
4	BRAF	1	1	0	0	0	0	1	0	3	0.00072
5	CHRM3	1	0	1	1	1	0	0	0	4	0.00077
6	SYNGR1	0	0	1	1	0	0	0	0	2	0.00083
7	IQSEC3	0	1	0	0	0	1	0	0	2	0.00100
8	ZDHHC11	1	0	0	1	0	0	0	0	2	0.00114
9	UGT2A3	0	1	0	0	0	1	0	0	2	0.00115
10	DSG4	0	0	1	0	1	0	0	0	2	0.00119
11	DENND2C	0	0	0	1	1	0	1	0	3	0.00176
12	SAMD7	0	0	1	0	1	0	0	0	2	0.00196
13	MYOM2	1	0	1	0	1	0	0	1	4	0.00235
14	C12orf36	0	1	0	0	1	0	0	0	2	0.00386
15	SPAM1	0	0	1	0	1	0	0	0	2	0.00419
16	KEL	0	1	0	0	1	0	0	0	2	0.00429
17	GALNT9	0	1	0	0	1	0	0	0	2	0.00469
18	FANCB	0	0	0	1	1	0	0	0	2	0.00474
19	C6orf120	0	0	0	0	0	1	0	0	1	0.00501
20	FLRT2	0	1	0	0	1	0	0	0	2	0.00530

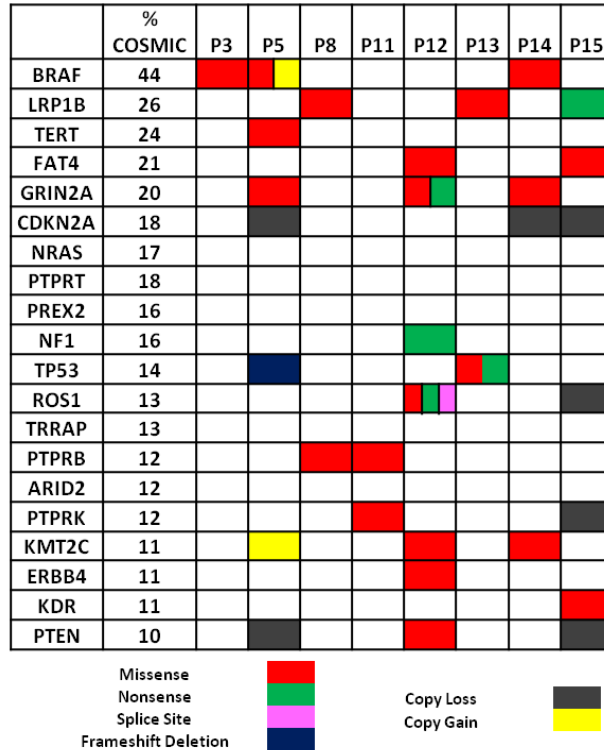
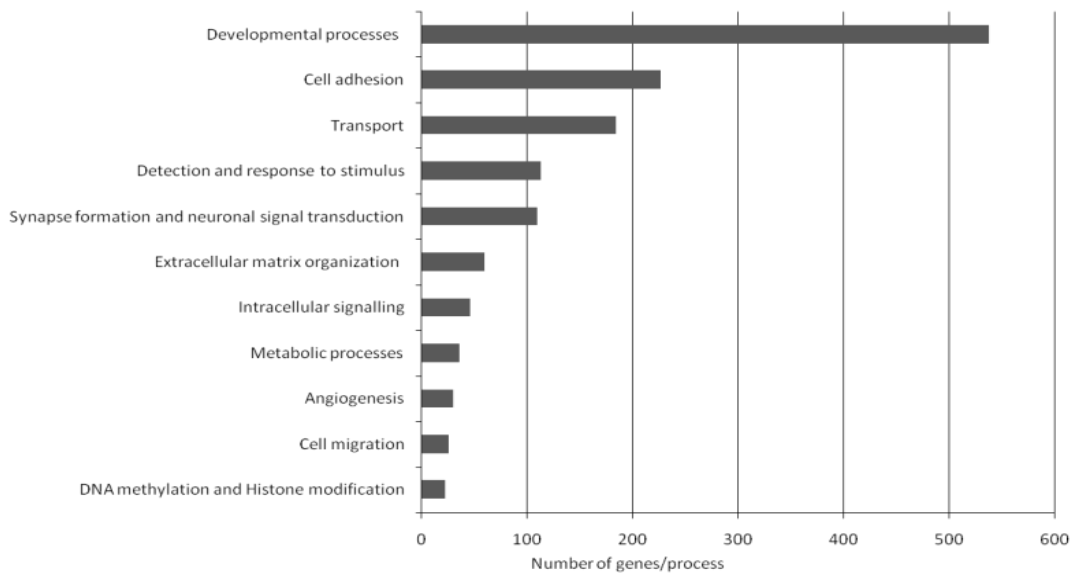


Figure 23: Top Census genes in melanoma from COSMIC database and the type of mutation found in all patients

#### 4.2.3.2 Enhanced Pathway Analysis

In order to examine whether the genes found to carry somatic mutations were related to specific biological mechanisms, we performed enrichment analysis on the union of non-synonymous mutations for all the patients, particularly missense, nonsense, frameshifting, splice site and non-stop mutations. Excluding the genes that were solely mutated in patient 12 (1303 genes) to avoid patient-specific bias, as well patient 10, who harboured very few mutated genes, a starting list of 1587 genes was obtained. Aiming to focus on genes putatively contributing to melanoma pathophysiology and filter out those carrying non-significant mutations, we applied two filtering steps at the 1587 gene list. Firstly, taking into account the predicted impact of each mutation on protein functionality, as predicted by PolyPhen2 tool (Adzhubei et al., 2010), which excluded all genes carrying neutral mutations. Moreover, we explored whether these genes are expressed in melanoma through TCGA's cBioPortal datasets and retained for pathway analysis only

those appearing to have at least low expression in over 30% of the cases. These filtering steps reduced the list to 769 genes, which were used as input for enrichment analysis based on GO (Ashburner et al., 2000; The Gene Ontology Consortium, 2019) and Reactome (Fabregat et al., 2018). The BioInfoMiner tool was used and statistically significant enriched terms were revealed, which were grouped according to their biological relevance in Figure 24.



*Figure 24: Statistically significant biological processes with the corresponding number of genes found as mutated in at least one patient*

A great number of genes fall in the categories of developmental processes (295 genes) and cell adhesion (138 genes). Interestingly, 67 genes were related to neural system characteristic mechanisms, as indicated by GO terms such as ‘neuronal action potential’, ‘synapse organisation’, ‘regulation of myelination’ and ‘neuron projection guidance’, grouped under the label ‘synapse formation and neuronal signal transduction’. With the scope of distinguishing putatively causative genes, focus was given on those with implication in diverse cross-talking biological processes, reflecting genes with a central role in cellular physiology. For this reason, we performed topological analysis

using BioInfoMiner, which exploits semantic information to detect and rank genes based on their centrality, as described in different databases (e.g., GO and Reactome). This analysis resulted in a short list of genes (Figure 25) with possible causal implications in melanoma. Interestingly, in the proposed list there are several genes with a well-established connection to cancer, like *BRAF*, *ATM* and *TP53*, but also others like *PDPK1* and *DMD* which could represent intriguing, yet poorly explored targets for further evaluation and possibly cancer treatment. Particularly, *PDPK1* (3-phosphoinositide dependent kinase 1) was found altered in three patients, two of them carrying a gene amplification and one carrying a possibly damaging point mutation. Regarding *DMD* (Duchenne muscular dystrophy), it is a long gene of 2.5 Mb, located on chromosome X. In two patients, *DMD* was found containing protein-altering point mutations. It is worth mentioning that significant pathways are enriched by different genes in each patient, suggesting that the great diversity observed in genes affected by somatic mutations could reflect deregulation of common molecular mechanisms. Indeed, regarding processes with an established role in melanoma genesis and progression, such as the MAPK pathway and cell cycle (Appendix Tables A3), all patients are found to have at least one mutated gene annotated by GO to the aforementioned biological processes. The fact that all these genes are expressed and bear damaging mutations supports their potential implication in a malfunctioning mechanism contributing to melanoma.

	P3	P5	P8	P11	P12	P13	P14	P15
CTNNB1								D
PTK2B								D
NOTCH1				D				
LRRK2			P					
BRAF	D	D						D
DMD	P				D			
RELN					D			
ATM				D				
EPHA2		D						
PDPK1							P	
ZC3H12A	P							
TP53						P		
ANGPT1			D					
NR1H4		P			D			
TLR4		P						
KDR								D
CLU			D		Start			
CASP8								P
HSF1		D						
CDKN1B						D		
ROCK1			D					
ANK3				D	D			
GSN				D				
TERT		P						
HNF1A	P							
PPP1R9A			P		D			
DCN	P				P			
AKAP6						P		
ROBO2								D
PKP2					D			D







Missense		Copy Loss	
Nonsense		Copy Gain	
Start Codon SNP			
Frameshift Deletion			

Figure 25: 30 top-prioritised mutated genes, D-probably damaging and P-possibly damaging mutation, as predicted by PolyPhen2.

#### 4.2.4 Discussion

In this part, the characterisation of somatic mutations and germline variants in patients with primary melanomas from Greece by exome sequencing analysis was reported. This was the first analysis, to our knowledge, where primary CM tissue from a low-incidence, southern European country is analysed at the exome level. In particular, FFPE tissue paired samples were used, which represent a valuable source of knowledge that needs to be exploited, especially in the case of CM, where clinical practice renders fresh-frozen primary tissue availability limited. Towards this end, the present part

consisted of an investigation aiming to overcome technical difficulties and establish bioinformatics workflows for the exploitation of NGS approaches on FFPE clinical samples. Here, the consequences of the fixation procedure were minimised, ensuring the validity of the presented results, at the cost of the inevitable loss in sensitivity. A multi-level analysis was performed, exploiting vastly established databases and state-of-the-art tools to incorporate information aiming at a better understanding of the underlying mechanisms involved with melanoma. A short list of candidate genes with probable causative role in CM was obtained, which contains both well-known melanoma-associated genes, but also potential new players, such as *PDPK1* and *DMD*. *PDPK1* was originally characterised as a serine-threonine kinase, phosphorylating and activating *AKT* (Alessi et al., 1997). *PDPK1* is a key element at the crossroad of signal transduction pathways such as Ras/MAPK pathway and Myc-cascade, in addition to PI3K/AKT (Gagliardi et al., 2017). Furthermore, *PDPK1* is frequently amplified at the gene level or over-expressed in several tumour types (Choucair et al., 2012; Maurer et al., 2009), including melanoma (Scortegagna et al., 2014). As far as *DMD* is concerned, it was recently reported as a tumour suppressor in cancers, featuring myogenic programmes (Wang, Marino-Enriquez, et al., 2014). In melanoma cell lines, the *DMD* gene was found bearing deletions while the protein was frequently absent or down-regulated (Körner et al., 2007). In addition, a recent study based on genomic data from public repositories of diverse cancer types, showed that *DMD* expression was decreased in the majority of the analysed tumours. Specifically in the case of melanoma, *DMD* was down-regulated as compared to benign nevi that already showed a reduced expression compared to normal skin (Luce et al., 2017).

## 4.3 Data integration & Classification

Melanoma diagnosis can be challenging and relies solely on the experience of the dermatologist or physician (“Melanoma Research Foundation”, n.d.). One major issue of dermoscopy is the inability to detect early melanoma or cases that lack optical features (Goodson and Grossman, 2009). To deal with that issue one can turn to molecular



techniques. The onset and constant advancement of molecular technologies has enabled the parallel, high-throughput process of millions of sequence reads, thus ushering a new era with numerous, novel applications in basic, applied and clinical research. Molecular technologies allow the extrapolation of profiling patterns of genomic sequences (whole genome, whole exome (WES), or targeted sequencing of a gene panel) with classification ability for the different phenotypic classes of a disease/pathology. On these grounds, the aim was to integrate the different levels of molecular and imaging data, so as to produce a robust diagnostic signature for the classification of melanoma.

Coalescing diverse levels of information improves the total knowledge on a problem and promotes its resolution (Lanckriet et al., 2004). Based on this, diagnosis should be based on the correct integration of molecular, histological and clinical features, so as to become more accurate. Previous analyses were able to achieve better performance on given tasks, through combination of heterogeneous data (Lanckriet et al., 2004; Ye et al., 2008), or by building multi-marker models for accurate classification of melanoma (Kashani-Sabet et al., 2009; Mann et al., 2013; Rothberg et al., 2009). Better understanding of the etiological aspects and mechanisms of cancer development are vital to improve survival rate and prevention. Given this perspective, recent studies have shown an improved performance, when combining transcriptomics with gene regulatory data in ovarian cancer (Xu et al., 2016). Efficient predictive biomarkers from multiple approaches or different levels of analyses support optimal characterisation of the tumour under investigation. Gene signature strategies are tested extensively for their potential to transform clinical practice, i.e. to support immunotherapy-based, management of cancer-patients (Gibney et al., 2016).

Previous work by our group (Moutselos et al., 2014; Valavanis et al., 2015) has shown that data integration offers key information on melanoma manifestation. Here, the intension was to extend this knowledge to mutational data. Ultimate aim is to produce a robust diagnostic gene signature that allows the classification of the patients and at the same time aid in the context of personalised medicine.

### 4.3.1 Molecular data classification system

The molecular data derived from exome sequencing of melanoma tissue and matched healthy control, from the previously described analysis. At this point, we sought to build a classifier exploiting the mutational data that were produced. Since the number of patients analysed in this work was limited, samples were added from TCGA database through cBioPortal (Cerami et al., 2012; Gao et al., 2013). As healthy state (non-melanoma) mutational data from dysplastic nevi that were acquired through similar experimental procedure (Melamed et al., 2017) were used. On the molecular level, this state holds a considerably lower mutational load compared to melanoma, and few mutated genes in total, 232 genes, as opposed to the 1587 genes, found in this study (see chapter 4.2.3). For feature selection, we reduced the list of mutated genes, to a total of 51 genes, by distinguishing the ‘driver’ mutations, i.e. mutations with high impact on the product, using PolyPhen2 (Adzhubei et al., 2010), and then prioritising them according to their centrality (genes taking part in numerous distinct mechanisms are ranked higher), using BioInfoMiner (Koutsandreas et al., 2016). Table 14 presents the 51 genes used for the molecular signature.

Table 14: Genes used as features for the molecular classifier

PTK2B	KDR	MYOC	BRAF	FZD4
RELN	ROCK1	FLT1	SCN5A	NR1H4
CTNNA1	DMD	DCN	NOTCH1	SEMA3C
COL3A1	PKP2	NPTN	CDH1	ROBO2
TEK	CSF1R	PPP1R9A	CACNA1C	NRXN1
LRRK2	PDGFD	TRPV4	ANK3	GRIN2A
ANGPT1	ERBB4	EPHB2	LAMA2	PTPRO
KALRN	ROBO1	POSTN	ITGB4	CFTR
DOCK1	PDPK1	FLRT2	BVES	DCHS1
EPHA2	EPHA7	SEMA3E	CELSR1	KMT2A
CARMIL1				

The samples (samples of dysplastic nevi and melanomas) were separated under two labels, dysplastic nevus (*represented by DNS*) and melanoma (*represented by MEL*)

and each sample is attributed a 51-dimensional binary vector showing whether the corresponding gene contains a mutation or not. To deal with unbalanced classes, the SMOTE (Chawla et al., 2002) algorithm was utilised to generate synthetic data for the DNS label. This data assortment is presented in Figure 26. Due to the binary type of the classification problem, the *Random Forests (RF)* algorithm (Chen and Ishwaran, 2012) was selected, as an appropriate and effective methodology. Additional classification algorithms were examined, generally showing equivalent outcome, due to the evident discrepancy of the two classes (see Figure 27). RF implementations are often more parametrizable than similar tree-based algorithms (like *Decision Trees*) and this permitted an exhaustive grid search for fine-tuning of classification parameters. Also, RF is a recursive algorithm, an asset that prevents being trapped in a subset of solutions and so, all contingencies are included, with the appropriate statistical weight.

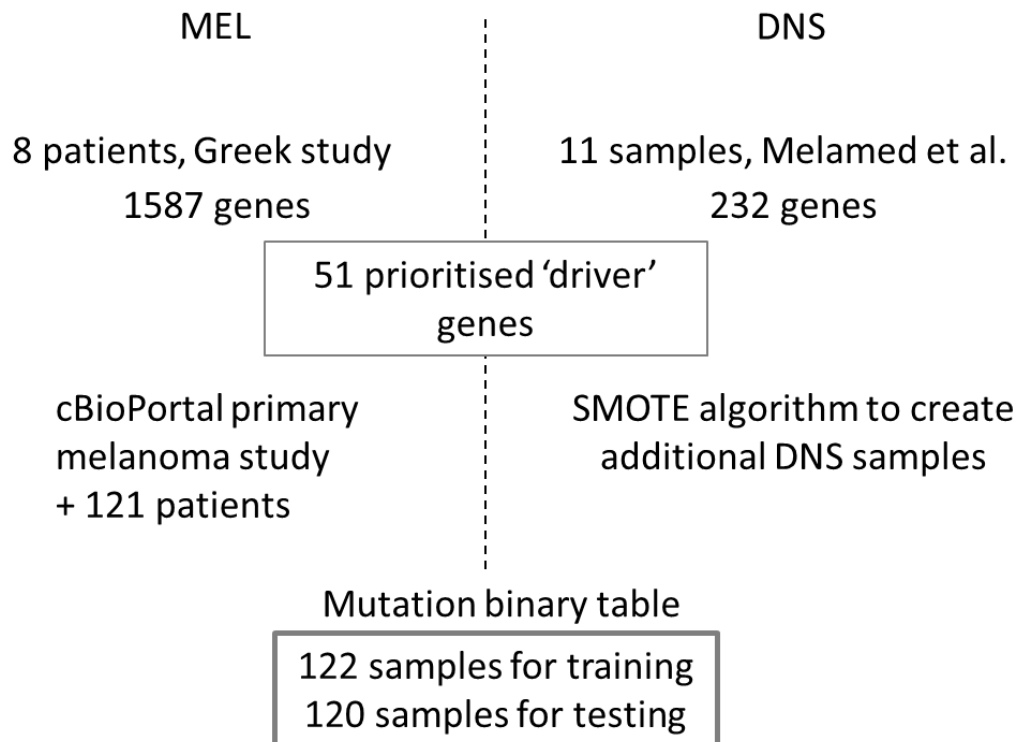


Figure 26: Data assortment for classification

The best performance was reported for the RF classifier with the following parameters:

- 122 samples
- 51 predictors
- 2 classes: 'DNS', 'MEL'
- No pre-processing
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- mtry = 26

As a criterion for the cross validation performance, the receiver operating characteristic (ROC) curve was used, which controls the sensitivity with respect to the specificity (Hajian-Tilaki, 2013). The area under the curve (AUC) of the plot gives an unbiased estimation of the classifier's performance at each round. The classifier performed very well, reaching a mean accuracy of 0.93. This result justifies the utilisation of this classifier as a model for class prediction (melanoma vs. dysplastic nevus) of unknown samples of mutation data.

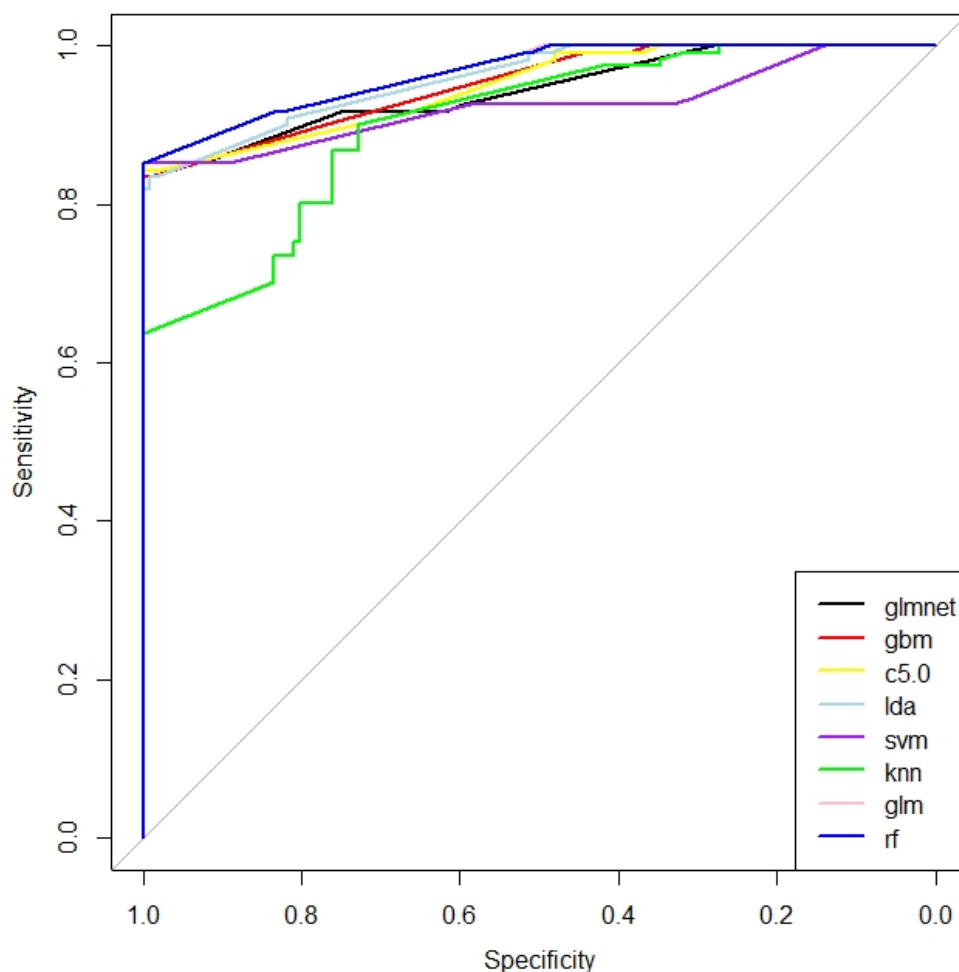


Figure 27: Results for the Molecular Classifier. ROC curve for rf-Random Forests, glmnet-Gaussian linear model, gbm-Stochastic Gradient Boosting, c50-Decision Trees C5.0, lda-Linear Discriminant Analysis, svm-Support Vector Machines, knn-k Nearest Neighbours, glm- Logistic Regression

### 4.3.2 Imaging data classification system

The imaging data derived from our group's previous study (Moutselos et al., 2014), analysing 1041 (69 melanomas / 972 dysplastic nevi) images. To deal with unbalanced classes, as before, the SMOTE algorithm was utilised, this time to generate synthetic data for the MEL label. The samples (samples of dysplastic nevi and melanomas) were separated under two labels, dysplastic nevus (*represented by DNS*) and melanoma

(represented by MEL), and each sample is attributed a 31-dimensional vector showing the measurements for each feature, since one feature was removed due to having zero variance in both classes. Table 15 presents the metrics used as features for the imaging classification system. Only 122 samples (balanced classes) were chosen, randomly, for analysis to make sure this set can be used for integration in a following step.

Table 15: Metrics used as features for the imaging classifier

mean-R	DISSIMILARITY	L-mean
std-R	ASM	L-std
mean-G	GMSM-mean	A-mean
std-G	GLSM-std	A-std
mean-B	I-mean	B-mean
std-B	I-std	B-std
PERIMETER	S-mean	Grad-mean
AREA	S-std	Grad-std
ECCENTRICITY	H-mean	Grad-max
COMPLEXITY	H-std	Distance-std
Asymmetry		Grad-min (zero variance)

The best performance was reported for the RF classifier with the following parameters:

- 122 samples
- 31 predictors
- 2 classes: 'DNS', 'MEL'
- No pre-processing
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- mtry = 2

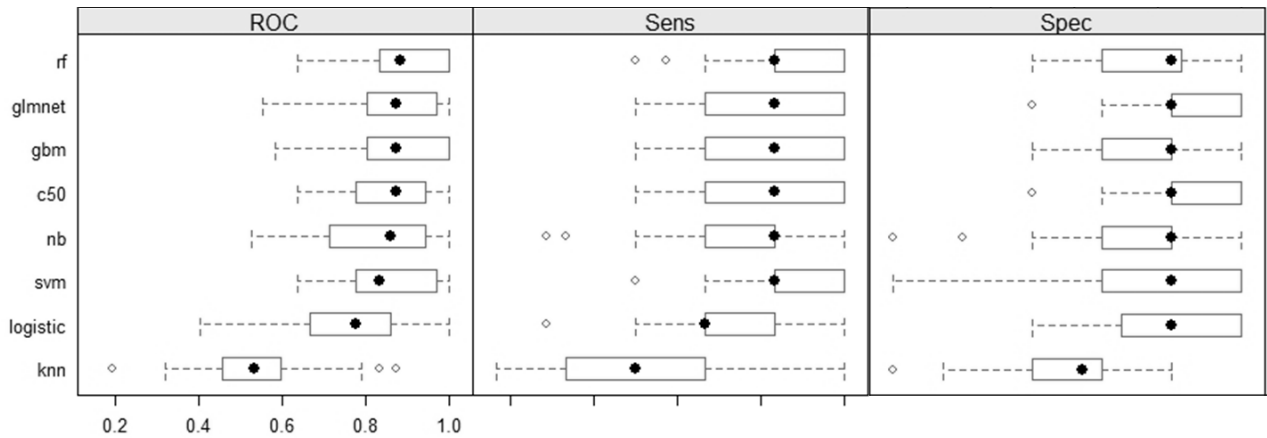


Figure 28: Results for the Imaging Classifier. ROC, Sensitivity and Specificity for *rf*-Random Forests, *glmnet*-Gaussian linear model, *gbm*-Stochastic Gradient Boosting, *c50*-Decision Trees C5.0, *svm*-Support Vector Machines, *logistic*- Logistic Regression, *knn*-k Nearest Neighbours

In order to evaluate the performance of the imaging classifiers the ROC analysis was used (presented in Figure 28). The RF classifier performed very well, reaching a mean accuracy of 0.898. This result justifies the utilisation of this classifier as a model for class prediction (melanoma vs. dysplastic nevus) of unknown samples of imaging data.

### 4.3.3 Integrated data classification system

In order to validate the proof of concept of the proposed design we created a fused dataset containing molecular and clinical data for the melanoma case. More specifically, a synthetic dataset was constructed to incorporate images from different nevi (dysplastic or melanomas) together with molecular measurements which are encountered in the same stages, using the imaging and WES data that were available, described previously. The nature of the molecular features allows for this ‘random’ integration, due to the small number of mutations, especially in the dysplastic nevus class. In total, tests were performed on three (3) different datasets i) the molecular dataset of 51 features, ii) the imaging dataset of 31 features and iii) the integrated dataset of 51+31 features. The parameters used for each RF classifier are similar (section 4.3.1), apart from the number

of predictors used, that equals the number of features (=82). Performance metrics obtained by classification modules (DNS vs. MEL) support that integrated features perform best, regarding the discrimination between malignant and benign sample classes, and constitute to an improved classifier, compared to the molecular and imaging classifiers. From the statistical perspective, the use of synthetic data is more conservative when the number of replicates is large. Essentially, it is the closest and more plausible approach to be adopted for simulation purposes. The corresponding results in the form of ROC curves are illustrated in Figure 29.

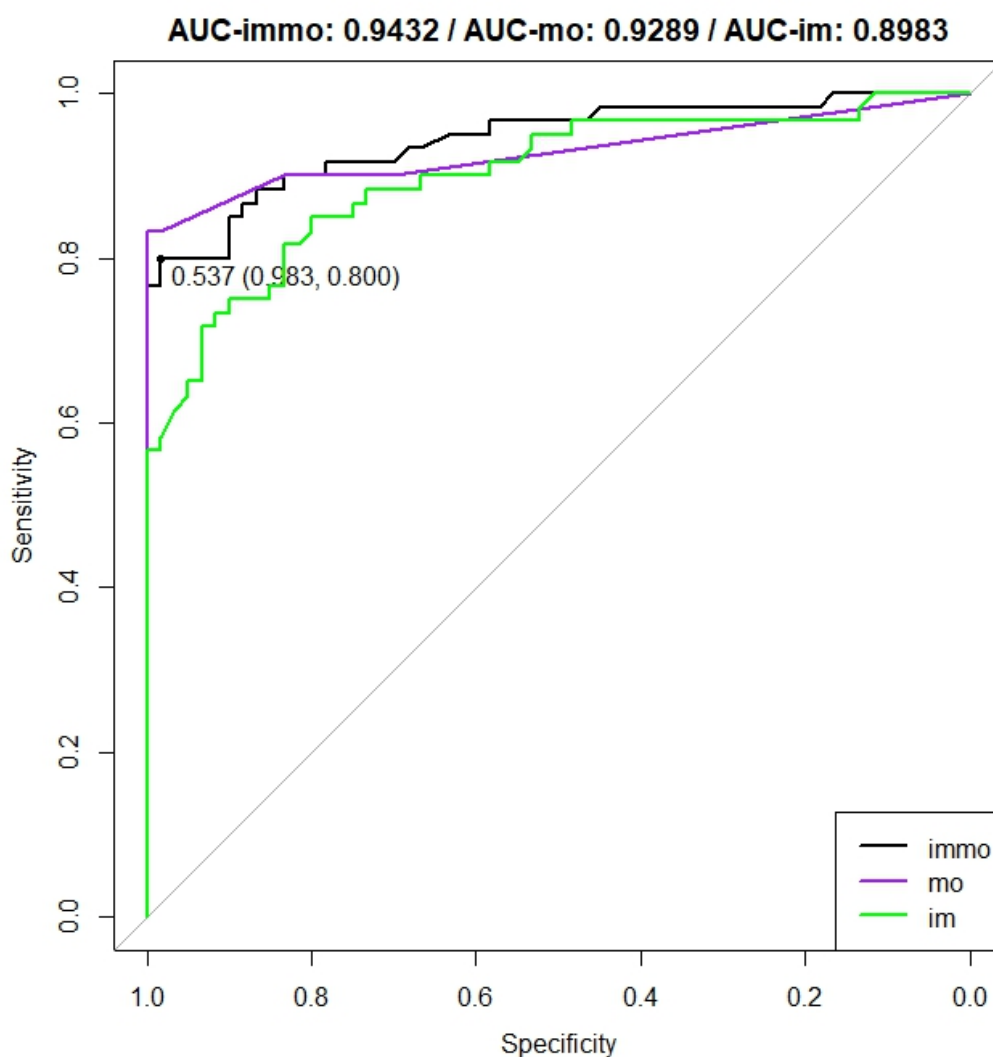


Figure 29: ROC curves for the 3 Random Forests classifiers, immo-integrated features (82) classifier, mo-molecular features (51) classifier, im-imaging features (31) classifier. The integrated feature classifier performs best, with a mean AUC of 0.9432



The entropy-based information gain (IG) and the information gain ratio (GR) were measured for the set of features from the integrated dataset. Four extra datasets were created containing the top 10 and top 20 features per measurement, to evaluate the classification accuracy. Table 16 contains the top 20 features selected using information gain and information gain ratio. It is worth mentioning that both lists contain mostly gene features, deriving from the molecular dataset.

*Table 16: Top 20 features selected using information gain and information gain ratio, gene features in bold*

Information Gain	Gain Ratio
L-mean	<b>ANK3</b>
mean-R	<b>RELN</b>
<b>ANK3</b>	mean-R
I-mean	<b>GRIN2A</b>
<b>RELN</b>	<b>SCN5A</b>
<b>GRIN2A</b>	<b>FLT1</b>
mean-G	<b>COL3A1</b>
<b>SCN5A</b>	<b>KALRN</b>
<b>FLT1</b>	<b>CFTR</b>
H-std	<b>ROBO2</b>
<b>COL3A1</b>	<b>LAMA2</b>
<b>KALRN</b>	<b>NRXN1</b>
std-B	<b>DMD</b>
<b>CFTR</b>	<b>EPHA7</b>
<b>ROBO2</b>	<b>CELSR1</b>
<b>LAMA2</b>	<b>ANGPT1</b>
<b>NRXN1</b>	<b>CACNA1C</b>
mean-B	I-mean
<b>DMD</b>	<b>PTPRO</b>
<b>EPHA7</b>	L-mean

Performance metrics obtained by classification modules (DNS vs. MEL) support that all integrated feature datasets perform equally good (Figure 30).

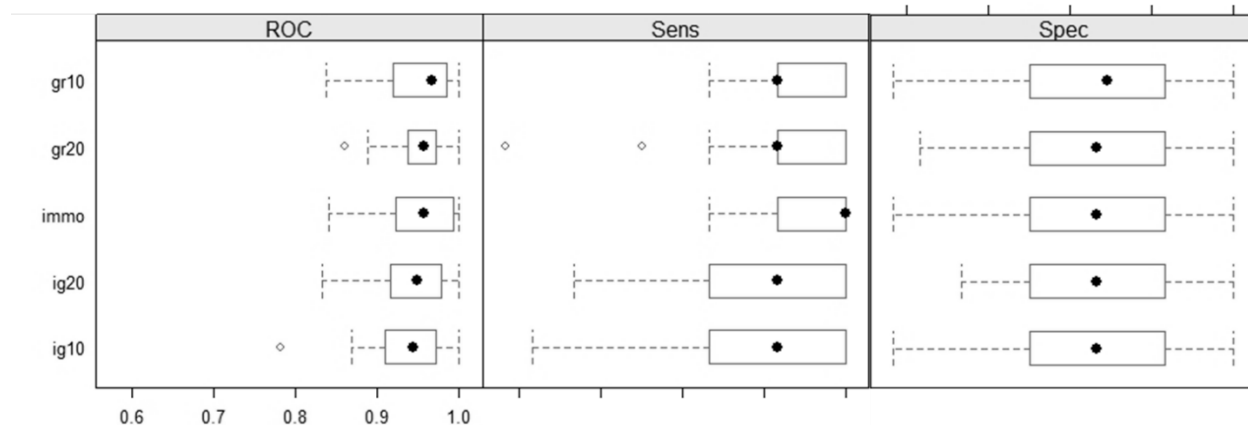


Figure 30: Results for the Integrated RF Classifiers, ROC, Sensitivity and Specificity for immo-82 features, gr10-gain ratio top-10, gr20-gain ratio top-20, ig10-information gain top-10 and ig20-information gain top-20 features

#### 4.3.4 Discussion

Performance metrics obtained by classification modules (DNS vs. MEL) support that integrated features perform best, regarding the discrimination between malignant and benign sample classes, and constitute to an improved classifier, compared to the molecular and imaging classifiers. Results are similar and equally plausible when selecting the top features for classification, by statistical means. The performance metrics obtained for the classifiers support the basic notion of this study.

# Chapter 5. Conclusions

## General Discussion

Since the start of this thesis there have been exhilarating advances in the field of cancer research. New technologies have allowed investigators to screen cancer in new ways, explore tumours at higher resolution and in greater numbers than ever before and unearth novel aspects of the molecular characteristics of cancer development, leading to an explosion of new data sources and a corresponding need for new methodologies to analyse such big data. Computer science has met the challenges presented by this oncoming overflow of data, and upon continuing to do so in the next decade we will learn more about cancer biology than was dreamed possible at the time the field began.

Marked advances in dealing with melanoma have been achieved over the last years, due to the diligent efforts of researchers to shed light on the biological mechanisms involved in melanoma manifestation, as well as the classification methodologies allowing skin image recognition with high accuracy. Regarding this thesis, as stated in the very beginning, the analysis performed is multi-disciplinary, consequently having multi-level outcomes.

One of the main questions was whether integrated diverse molecular data can advance our knowledge concerning melanoma. Exportation of the broader molecular network implicated in cutaneous melanoma was established through integration of different molecular levels. Additional mechanisms previously concealed by the statistical cut-offs were revealed, thus enhancing our general understanding of the phenomenon.

Another main goal of this project was to infer the composite biomarkers with robust discriminative ability between melanoma and healthy nevus. My analysis started from raw whole exome sequencing data for the identification of variance and somatic mutations concerning the disease under investigation. For the first time the characterisation of somatic mutations and germline variants in patients with primary melanomas from Greece by exome sequencing analysis was reported. The output was a list of genes – potential

biomarkers – that are affected and play a crucial role in the manifestation of melanoma. This list signifies the molecular component, and after integration with the imaging component, together they represent the composite biomarkers under investigation. Further characterisation of the genetic risk factors in different patient populations could help develop more efficient prevention strategies and improve tactics for early diagnosis.

The ultimate purpose of this research was to find the best combination of features in order to create an automated integrative processing system for the detection of cutaneous melanoma. Success of classification depends mainly on feature selection. Based on this, the total list of features (82) achieves high accuracy, improved only by the top features taken through gain ratio. Still, cutting the number of features does not necessarily improve this system, since the performance was equal to begin with, plus extra information is considered for the therapeutic approach, when required. The total list offers a solution that not only is very successful in the discrimination of the classes but can actually assist with treatment and personalised medicine at the next stage, in a cost-effective manner. The most important aspect is that the integration of the mutational with the imaging data improved the classification system, as was hypothesised. Multiple algorithms were tested, with Random Forests achieving the highest accuracy for the integrated dataset.

The proposed design offers a tiered analytical framework as an expansion of current EHR systems, which can integrate high-volume molecular omics data, imaging data, as well as relevant clinical observations. It enables a “molecular-enabled computational approach” that can be incorporated in clinical practice, revolutionising therapeutic strategies, not only for melanoma, but analogous complex diseases. The overall idea advances the importance of building a sturdy alliance between researchers and clinicians in the context of translational cancer research. An imperative aid towards this coalition is the constant decrease of NGS-based testing cost, recently shown to be lower than that of conventional methods (de Unamuno Bustos et al., 2017).

Elucidating the mechanisms underlying melanoma biology and progression aids the development of targeted and immune-related therapeutic approaches. A potential scenario for the application of this analysis would be, upon request from the physician,

targeted sequencing of the specific highlighted genes, screening for mutations, integration with the dermoscopic features, successful classification, and in case of melanoma diagnosis, patient-specific therapy incorporating the mutational pattern.

## Challenges

One important challenge during this analysis was using FFPE samples for NGS analyses. FFPE samples represent a valuable source of knowledge that needs to be exploited, especially in the case of cutaneous melanoma, where clinical practice renders fresh–frozen primary tissue availability limited. Towards this end, the presented analysis consisted of an investigation aiming to overcome technical difficulties and establish bioinformatics workflows for the exploitation of NGS approaches on FFPE clinical samples. The consequences of the fixation procedure were minimised, ensuring the validity of the presented results, at the cost of losing few “true” mutations.

Another restraint was the limited number of patients analysed. To deal with this a multi-level analysis was performed, exploiting vastly established databases and state-of-the-art tools to incorporate information aiming at a better understanding of the underlying mechanisms involved with melanoma and finally, building a synthetic dataset for classification. The methodology used combined functional impact analysis with pathway enrichment, to deal with the limited dataset, in order to distinguish important genes and possible drivers, and based on these, inclusion of additional samples from public databases. Still on the same restraint, the datasets also suffered from unbalanced abundance of benign vs malignant samples, which was dealt with using specific algorithms.

## Algorithm complexity & Contribution

The offered analysis presented a workflow bridging together miscellaneous tools and methodologies, overcoming technical and experimental hitches, like the limited

sample set, to establish a strong foundation supporting the initial hypotheses. An important facet of the presented scheme is that it can be utilised for varied analyses, exploring different types of cancer with analogous characteristics.

As has been stated, the complexity of NGS data is high, due to the high volume of information contained in each sample and several distinct parameters needed specific adjustments, so as to optimise the performance and the quality of the results. Aside from that, most of the analysis carried out required specialised hardware and processing power. It is worth noting that exome analysis workflow for a pair of samples (tumour and normal samples from one patient) needed approximately 35 hours running time, starting from ~20GB and summarising the results in ~10MB.

The work in this thesis required advanced knowledge of bioinformatics programming and machine learning, as well as solid understanding of the biological background. Nevertheless, the multi-level outcomes can be used directly, lists of molecular/ composite biomarkers for example.

## Future work

The future goal is to expand this analysis to a greater number of patients, aiming to study any possible associations between the various levels of data in melanoma, i.e. germline and somatic alterations or phenotype/ imaging features. In order to prove the diagnostic value of the presented integration scheme and further validate the design of the implemented system, more experiments need to be carried out, incorporating “real” paired data. Assimilation of sufficient numbers of actual paired molecular and imaging data can offer new insights and paths for exploration.

Additionally, integration of supplementary layers of data in the presented framework, like RNA sequencing data, cannot only improve our understanding of disease manifestation, but potentially improve the classification scheme and overall accuracy, by introducing new biomarkers with better discrimination ability.

An important aspect is that the offered analysis scheme can be utilised for diverse classification problems. This means that the same analysis can be expanded to specific CM subtype classification, rather than melanoma or healthy distinction; like discrimination between superficial spreading melanoma, nodular melanoma, lentigo maligna melanoma, desmoplastic melanoma or amelanotic melanoma, given of course that enough training data becomes available.

Furthermore, patterns or clusters of differential morphology, presented on dermoscopy images, can be linked to specific molecular traits – germline or somatic alterations – perhaps allowing the derivation of personalised treatment, through a non-invasive aspect, directly from phenotypic features.

## Closing remarks

The presented multi-level analysis offers a flexible and distributed workflow, which integrates heterogeneous, multidimensional data for the multi-angled portrayal and classification of melanoma patients. It highlights a short list of candidate genes with a probable causative role in melanoma that were used as promising targets, allowing the successful classification of melanoma versus dysplastic nevus, achieving satisfactory accuracy.

The analysis framework may have required advanced knowledge of bioinformatics programming and machine learning, but the output can be used directly as a list. This list includes candidate genes with probable causative role in CM, containing both well-known melanoma-associated genes, but also potential new players.

Melanoma is one of the most lethal types of cancer. Additional understanding of the resistance to targeted therapies is crucial, and ought to remain a central aspect of cancer research. The intervention schemes based on combination approaches are the most promising therapeutic ways, in the context of personalised treatment, strengthening knowledge discovery and computer-aided intelligent diagnosis in the direction of precision

medicine. These schemes can be introduced as an extension to current EHR systems, aiding in clinical decisions.

Digital dermoscopy has been established in clinical practice for melanocytic lesion monitoring. The last few years NGS methodologies, mainly targeted approaches focusing on up to few hundreds of genes are being introduced for the characterization of several cancers. The presented analysis proposes a paradigm, which through the massive integration of multi-layered, heterogeneous data, depicting phenotypic aspects of the disease manifestation and the parallel processing of those streams, independently or together, will produce appropriate sets of composite biomarkers that ultimately assist and accelerate medical diagnosis and patient therapeutic management.



# Bibliography

- Abbas, Q., Garcia, I.F., Celebi, M.E. and Ahmad, W. (2013), "A Feature-Preserving Hair Removal Algorithm for Dermoscopy Images", *Skin Research and Technology*, Vol. 19 No. 1, pp. e27–e36.
- Abbas, Q., Garcia, I.F., Celebi, M.E., Ahmad, W. and Mushtaq, Q. (2013), "A perceptually oriented method for contrast enhancement and segmentation of dermoscopy images", *Skin Research and Technology*, Vol. 19 No. 1, pp. e490–e497.
- Abedini, M., Chen, Q., Codella, N.C.F., Garnavi, R. and Sun, X. (2015), "Accurate and Scalable System for Automatic Detection of Malignant Melanoma", *Dermoscopy Image Analysis*, M. E. Celebi, T. Mendonça and J. S. Marques, CRC Press, pp. 293–343.
- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. and Saeys, Y. (2010), "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods", *Bioinformatics*, Vol. 26 No. 3, pp. 392–398.
- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., et al. (2010), "A method and server for predicting damaging missense mutations", *Nat Methods*, Vol. 7, United States, pp. 248–9.
- Aitken, J., Welch, J., Duffy, D., Milligan, A., Green, A., Martin, N. and Hayward, N. (1999), "CDKN2A variants in a population-based sample of Queensland families with melanoma", *Journal of the National Cancer Institute*, Vol. 91 No. 5, pp. 446–452.
- Akhtar, M.S., Swamy, M.K., Alaraidh, I.A. and Panwar, J. (2017), "Genomic Data Resources and Data Mining", in Hakeem, K.R., Malik, A., Vardar-Sukan, F. and Ozturk, M. (Eds.), *Plant Bioinformatics: Decoding the Phyta*, Springer International Publishing, Cham, pp. 267–278.

- Alessi, D.R., James, S.R., Downes, C.P., Holmes, A.B., Gaffney, P.R., Reese, C.B. and Cohen, P. (1997), "Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase B $\alpha$ ", *Current Biology*, Vol. 7 No. 4, pp. 261–269.
- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., et al. (2013), "Signatures of mutational processes in human cancer", *Nature*, Vol. 500 No. 7463, pp. 415–421.
- d'Amico, M., Ferri, M. and Stanganelli, I. (2004), "Qualitative asymmetry measure for melanoma detection", presented at the Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on, IEEE, pp. 1155–1158.
- An, O., Dall'Olio, G.M., Mourikis, T.P. and Ciccarelli, F.D. (2016), "NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings", *Nucleic Acids Res*, Vol. 44 No. D1, pp. D992-9.
- Andreu Perez, J., C. Y. Poon, C., Merrifield, R. and Guang-Zhong, Y. (2015), "Big Data for Health", *1208*, available at:<https://dx.doi.org/10.1109/JBHI.2015.2450362>.
- Antonopoulou, K., Stefanaki, I., Lill, C.M., Chatzinasiou, F., Kypreou, K.P., Karagianni, F., Athanasiadis, E., et al. (2015), "Updated field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma: the MelGene database", *J Invest Dermatol*, Vol. 135 No. 4, pp. 1074–9.
- Aoude, L.G., Wadt, K.A., Pritchard, A.L. and Hayward, N.K. (2015), "Genetics of familial melanoma: 20 years after CDKN2A", *Pigment Cell & Melanoma Research*, Vol. 28 No. 2, pp. 148–160.
- Appriou, A. (1999), "Multisensor signal processing in the framework of the theory of evidence", *Tiré à Part- Office National d'études et de Recherches Aérospatiales*.
- Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E. and Delfino, M. (1998), "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison

- of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis”, *Arch Dermatol*, Vol. 134 No. 12, pp. 1563–70.
- Argenziano, G., Soyer, H.P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., et al. (2003), “Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet”, *Journal of the American Academy of Dermatology*, Vol. 48 No. 5, pp. 679–693.
- Arroyo, J.L.G. and Zapirain, B.G. (2014), “Detection of pigment network in dermoscopy images using supervised machine learning and structural analysis”, *Computers in Biology and Medicine*, Vol. 44, pp. 144–157.
- Arroyo, J.L.G., Zapirain, B.G. and Zorrilla, A.M. (2011), “Blue-white veil and dark-red patch of pigment pattern recognition in dermoscopic images using machine-learning techniques”, *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, presented at the 2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 196–201.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., et al. (2000), “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”, *Nat Genet*, Vol. 25 No. 1, pp. 25–9.
- Athanasiadis, E.I., Antonopoulou, K., Chatzinasiou, F., Lill, C.M., Bourdakou, M.M., Sakellariou, A., Kypreou, K., et al. (2014), “A Web-based database of genetic association studies in cutaneous melanoma enhanced with network-driven data exploration tools”, *Database (Oxford)*, Vol. 2014, available at:<https://doi.org/10.1093/database/bau101>.
- Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., et al. (2018), “Comprehensive Characterization of Cancer Driver Genes and Mutations”, *Cell*, Vol. 173 No. 2, pp. 371–385.e18.

- Ballerini, L., Fisher, R.B., Aldridge, B. and Rees, J. (2013), "A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions", in Celebi, M.E. and Schaefer, G. (Eds.), *Color Medical Image Analysis*, Springer Netherlands, Dordrecht, pp. 63–86.
- Banky, J.P., Kelly, J.W., English, D.R., Yeatman, J.M. and Dowling, J.P. (2005), "Incidence of new and changed nevi and melanomas detected using baseline images and dermoscopy in patients at high risk for melanoma", *Archives of Dermatology*, Vol. 141 No. 8, pp. 998–1006.
- Barrett, J.H., Iles, M.M., Harland, M., Taylor, J.C., Aitken, J.F., Andresen, P.A., Akslén, L.A., et al. (2011), "Genome-wide association study identifies three new melanoma susceptibility loci", *Nat Genet*, Vol. 43 No. 11, pp. 1108–13.
- Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., et al. (2013), "NCBI GEO: archive for functional genomics data sets—update", *Nucleic Acids Research*, Vol. 41 No. Database issue, pp. D991–D995.
- Bastian, B.C. (2014), "The molecular pathology of melanoma: an integrated taxonomy of melanocytic neoplasia", *Annual Review of Pathology: Mechanisms of Disease*, Vol. 9, pp. 239–271.
- Bean, L.J.H. and Hegde, M.R. (2016), "Gene Variant Databases and Sharing: Creating a Global Genomic Variant Database for Personalized Medicine", *Human Mutation*, Vol. 37 No. 6, pp. 559–563.
- Behjati, S. and Tarpey, P.S. (2013), "What is next generation sequencing?", *Archives of Disease in Childhood. Education and Practice Edition*, Vol. 98 No. 6, pp. 236–238.
- Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., et al. (2012), "Melanoma genome sequencing reveals frequent PREX2 mutations", *Nature*, Vol. 485 No. 7399, pp. 502–6.
- Betta, G., Di Leo, G., Fabbrocini, G., Paolillo, A. and Scalvenzi, M. (2005), "Automated Application of the '7-point checklist' Diagnosis Method for Skin Lesions: Estimation of Chromatic and Shape

- Parameters”, Vol. 3, presented at the Instrumentation and Measurement Technology Conference, 2005. IMTC 2005. Proceedings of the IEEE, IEEE, pp. 1818–1822.
- Blanzieri, E., Eccher, C., Forti, S. and Sboner, A. (2000), “Exploiting classifier combination for early melanoma diagnosis support”, presented at the European Conference on Machine Learning, Springer, pp. 55–62.
- Boca, S.M., Kinzler, K.W., Velculescu, V.E., Vogelstein, B. and Parmigiani, G. (2010), “Patient-oriented gene set analysis for cancer mutation data”, *Genome Biology*, Vol. 11 No. 11, p. R112.
- Boldrick, J.C., Layton, C.J., Nguyen, J. and Swetter, S.M. (2007), “Evaluation of digital dermoscopy in a pigmented lesion clinic: clinician versus computer assessment of malignancy risk”, *Journal of the American Academy of Dermatology*, Vol. 56 No. 3, pp. 417–421.
- Bono, A., Tomatis, S., Bartoli, C., Cascinelli, N., Clemente, C., Cupeta, C. and Marchesini, R. (1996), “The invisible colours of melanoma. A telespectrophotometric diagnostic approach on pigmented skin lesions”, *European Journal of Cancer*, Vol. 32 No. 4, pp. 727–729.
- Brash, D.E. (2015), “UV signature mutations”, *Photochem Photobiol*, Vol. 91 No. 1, pp. 15–26.
- Buzug, T.M., Schumann, S., Pfaffmann, L., Reinhold, U. and Ruhlmann, J. (2006), “Functional infrared imaging for skin-cancer screening”, presented at the Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE, IEEE, pp. 2766–2769.
- Cantarel, B.L., Weaver, D., McNeill, N., Zhang, J., Mackey, A.J. and Reese, J. (2014), “BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity”, *BMC Bioinformatics*, Vol. 15, p. 104.
- Carrot-Zhang, J. and Majewski, J. (2017), “LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples”, *Oncotarget*, Vol. 8 No. 23, pp. 37032–37040.

- Carter, H., Chen, S., Isik, L., Tyekucheveva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., et al. (2009), "Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations", *Cancer Research*, Vol. 69 No. 16, pp. 6660–6667.
- Celebi, M.E., Iyatomi, H., Schaefer, G. and Stoecker, W.V. (2009), "Lesion border detection in dermoscopy images", *Computerized Medical Imaging and Graphics*, Vol. 33 No. 2, pp. 148–153.
- Celebi, M.E., Wen, Q., Hwang, S., Iyatomi, H. and Schaefer, G. (2013), "Lesion Border Detection in Dermoscopy Images Using Ensembles of Thresholding Methods", *Skin Research and Technology*, Vol. 19 No. 1, pp. e252–e258.
- Cerami, E., Demir, E., Schultz, N., Taylor, B.S. and Sander, C. (2010), "Automated network analysis identifies core pathways in glioblastoma", *PLoS One*, Vol. 5 No. 2, p. e8918.
- Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., et al. (2012), "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data".
- Chandler, W.M., Rowe, L.R., Florell, S.R., Jahromi, M.S., Schiffman, J.D. and South, S.T. (2012), "Differentiation of malignant melanoma from benign nevus using a novel genomic microarray with low specimen requirements", *Archives of Pathology & Laboratory Medicine*, Vol. 136 No. 8, pp. 947–955.
- Chatzinasiou, F., Lill, C.M., Kypreou, K., Stefanaki, I., Nicolaou, V., Spyrou, G., Evangelou, E., et al. (2011), "Comprehensive field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma", *Journal of the National Cancer Institute*, Vol. 103 No. 16, pp. 1227–1235.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002), "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357.

- Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., et al. (2009), "BreakDancer: An algorithm for high resolution mapping of genomic structural variation", *Nature Methods*, Vol. 6 No. 9, pp. 677–681.
- Chen, X. and Ishwaran, H. (2012), "Random forests for genomic data analysis", *Genomics*, Vol. 99 No. 6, pp. 323–9.
- Chiem, A., Al-Jumaily, A. and Khushaba, R.N. (2007), "A novel hybrid system for skin lesion detection", presented at the Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on, IEEE, pp. 567–572.
- Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. (2012), "Predicting the functional effect of amino acid substitutions and indels", *PLoS One*, Vol. 7 No. 10, p. e46688.
- Choucair, K.A., Guérard, K.-P., Ejdelman, J., Chevalier, S., Yoshimoto, M., Scarlata, E., Fazli, L., et al. (2012), "The 16p13.3 (PDPK1) genomic gain in prostate cancer: a potential role in disease progression", *Translational Oncology*, Vol. 5 No. 6, pp. 453–460.
- Christoforides, A., Carpten, J.D., Weiss, G.J., Demeure, M.J., Von Hoff, D.D. and Craig, D.W. (2013), "Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs", *BMC Genomics*, Vol. 14, p. 302.
- Chwirot, B., Chwirot, S., Redziński, J. and Michniewicz, Z. (1998), "Detection of melanomas by digital imaging of spectrally resolved ultraviolet light-induced autofluorescence of human skin", *European Journal of Cancer*, Vol. 34 No. 11, pp. 1730–1734.
- Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., et al. (2013), "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples", *Nat Biotechnol*, Vol. 31 No. 3, pp. 213–9.
- Cicchetti, D.V. (1992), "Neural networks and diagnosis in the clinical laboratory: state of the art.", *Clinical Chemistry*, Vol. 38 No. 1, pp. 9–10.

- Cideciyan, A.V. (1995), "Registration of ocular fundus images: an algorithm using cross-correlation of triple invariant image descriptors", *IEEE Engineering in Medicine and Biology Magazine*, Vol. 14 No. 1, pp. 52–58.
- Cirenajwis, H., Lauss, M., Ekedahl, H., Törngren, T., Kvist, A., Saal, L.H., Olsson, H., et al. (2017), "NF1-mutated melanoma tumors harbor distinct clinical and biological characteristics", *Molecular Oncology*, Vol. 11 No. 4, pp. 438–451.
- Ciriello, G., Cerami, E., Sander, C. and Schultz, N. (2012), "Mutual exclusivity analysis identifies oncogenic network modules", *Genome Research*, Vol. 22 No. 2, pp. 398–406.
- Claridge, E., Cotton, S., Hall, P. and Moncrieff, M. (2003), "From colour to tissue histology: physics-based interpretation of images of pigmented skin lesions", *Medical Image Analysis*, Vol. 7 No. 4, pp. 489–502.
- Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A. and Smith, J.R. (2015), "Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images", in Zhou, L., Wang, L., Wang, Q. and Shi, Y. (Eds.), *Machine Learning in Medical Imaging*, Springer International Publishing, pp. 118–126.
- Codella, N.C.F., Nguyen, Q.-, Pankanti, S., Gutman, D.A., Helba, B., Halpern, A.C. and Smith, J.R. (2017), "Deep learning ensembles for melanoma recognition in dermoscopy images", *IBM Journal of Research and Development*, Vol. 61 No. 4/5, pp. 5:1-5:15.
- "COSMIC database". (n.d.) , available at: <http://cancer.sanger.ac.uk> (accessed 1 January 2017).
- Cox, A.D. and Der, C.J. (2010), "Ras history: The saga continues", *Small GTPases*, Vol. 1 No. 1, pp. 2–27.
- Cruz, J.A. and Wishart, D.S. (2006), "Applications of Machine Learning in Cancer Prediction and Prognosis", *Cancer Informatics*, Vol. 2, p. 117693510600200030.



- Cuéllar, F., Puig, S., Kolm, I., Puig-Butille, J., Zaballos, P., Martí-Laborda, R., Badenas, C., et al. (2009), “Dermoscopic features of melanomas associated with MC1R variants in Spanish CDKN2A mutation carriers”, *The British Journal of Dermatology*, Vol. 160 No. 1, pp. 48–53.
- Dai, D.L., Martinka, M. and Li, G. (2005), “Prognostic significance of activated Akt expression in melanoma: a clinicopathologic study of 292 cases”, *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, Vol. 23 No. 7, pp. 1473–1482.
- Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., et al. (2002), “Mutations of the BRAF gene in human cancer”, *Nature*, Vol. 417 No. 6892, pp. 949–954.
- Davis, S. and Meltzer, P.S. (2007), “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”, *Bioinformatics (Oxford, England)*, Vol. 23 No. 14, pp. 1846–1847.
- Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010), “Identifying a high fraction of the human genome to be under selective constraint using GERP++”, *PLoS Computational Biology*, Vol. 6 No. 12, p. e1001025.
- De Paoli-Iseppi, R., Johansson, P.A., Menzies, A.M., Dias, K.R., Pupo, G.M., Kakavand, H., Wilmott, J.S., et al. (2016), “Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: implications for clinical decision making”, *Pathology*, Vol. 48 No. 3, pp. 261–6.
- De Vries, E., De Poll-Franse, V., Louwman, W., De Gruijl, F. and Coebergh, J. (2005), “Predictions of skin cancer incidence in the Netherlands up to 2015”, *British Journal of Dermatology*, Vol. 152 No. 3, pp. 481–488.
- Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., et al. (2012), “MuSiC: identifying mutational significance in cancer genomes”, *Genome Research*, Vol. 22 No. 8, pp. 1589–1598.

- Denoeux, T. (1995), "A k-nearest neighbor classification rule based on Dempster-Shafer theory", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 25 No. 5, pp. 804–813.
- "Dermoscopy". (n.d.). , available at: <http://www.dermoscopy.org/> (accessed 11 February 2019).
- Development, R. (n.d.). *Core Team (2011) R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0. Available: <http://www.R-project.org>.
- Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., et al. (2012), "Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data", *Bioinformatics (Oxford, England)*, Vol. 28 No. 2, pp. 167–175.
- Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H. and Binder, M. (2001), "A comparison of machine learning methods for the diagnosis of pigmented skin lesions", *Journal of Biomedical Informatics*, Vol. 34 No. 1, pp. 28–36.
- Drier, Y., Lawrence, M.S., Carter, S.L., Stewart, C., Gabriel, S.B., Lander, E.S., Meyerson, M., et al. (2013), "Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability", *Genome Research*, Vol. 23 No. 2, pp. 228–235.
- Dummer, R. and Hoek, K. (2004), "Human Melanoma: From Transcriptome to Tumor Biology", *Forschungsdatenbank Der Universität Zürich*, Vol. 2008.
- Dunn, T., Berry, G., Emig-Agius, D., Jiang, Y., Lei, S., Iyer, A., Udar, N., et al. (2018), "Pisces: An Accurate and Versatile Variant Caller for Somatic and Germline Next-Generation Sequencing Data", *Bioinformatics (Oxford, England)*, available at: <https://doi.org/10.1093/bioinformatics/bty849>.
- Dutton-Regester, K. and Hayward, N.K. (2012), "Reviewing the somatic genetics of melanoma: from current to future analytical approaches", *Pigment Cell Melanoma Res*, Vol. 25 No. 2, pp. 144–54.

- van Dyk, E., Reinders, M.J.T. and Wessels, L.F.A. (2013), "A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control", *Nucleic Acids Research*, Vol. 41 No. 9, p. e100.
- "ECIS - European Cancer Information System". (n.d.). *European Cancer Information System*, available at: <https://ecis.jrc.ec.europa.eu> (accessed 10 October 2018).
- Edgar, R., Domrachev, M. and Lash, A.E. (2002), "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Research*, Vol. 30 No. 1, pp. 207–210.
- Edmondson, P.C., Curley, R.K., Marsden, R.A., Robinson, D., Allaway, S.L. and Willson, C.D. (1999), "Screening for malignant melanoma using instant photography", *Journal of Medical Screening*, Vol. 6 No. 1, pp. 42–46.
- Ercal, F., Chawla, A., Stoecker, W.V., Lee, H.-C. and Moss, R.H. (1994), "Neural network diagnosis of malignant melanoma from color images", *IEEE Transactions on Biomedical Engineering*, Vol. 41 No. 9, pp. 837–845.
- Escaramís, G., Tornador, C., Bassaganyas, L., Rabionet, R., Tubio, J.M.C., Martínez-Fundichely, A., Cáceres, M., et al. (2013), "PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data", *PLoS One*, Vol. 8 No. 5, p. e63377.
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., et al. (2018), "The Reactome Pathway Knowledgebase", *Nucleic Acids Research*, Vol. 46 No. D1, pp. D649–D655.
- Fang, L.T., Afshar, P.T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J.C., Gibeling, G., et al. (2015), "An ensemble approach to accurately detect somatic mutations using SomaticSeq", *Genome Biol*, Vol. 16, p. 197.

- FDA-NIH Biomarker Working Group. (2016), *BEST (Biomarkers, EndpointS, and Other Tools) Resource*, Food and Drug Administration (US), Silver Spring (MD), available at:  
<http://www.ncbi.nlm.nih.gov/books/NBK326791/> (accessed 23 February 2019).
- Fedorenko, I.V., Gibney, G.T. and Smalley, K.S.M. (2013), “NRAS mutant melanoma: biological behavior and future strategies for therapeutic management”, *Oncogene*, Vol. 32 No. 25, pp. 3009–3018.
- Feit, N.E., Dusza, S.W. and Marghoob, A.A. (2004), “Melanomas detected with the aid of total cutaneous photography”, *The British Journal of Dermatology*, Vol. 150 No. 4, pp. 706–714.
- Ferrer-Costa, C., Gelpí, J.L., Zamakola, L., Parraga, I., de la Cruz, X. and Orozco, M. (2005), “PMUT: a web-based tool for the annotation of pathological mutations on proteins”, *Bioinformatics (Oxford, England)*, Vol. 21 No. 14, pp. 3176–3178.
- Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., et al. (2017), “COSMIC: somatic cancer genetics at high-resolution”, *Nucleic Acids Research*, Vol. 45 No. D1, pp. D777–D783.
- Forsea, A.-M., Del Marmol, V., De Vries, E., Bailey, E. and Geller, A. (2012), “Melanoma incidence and mortality in Europe: new estimates, persistent disparities”, *British Journal of Dermatology*, Vol. 167 No. 5, pp. 1124–1130.
- Friedman, J., Hastie, T., Tibshirani, R., Simon, N., Narasimhan, B. and Qian, J. (2018), *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, available at: <https://CRAN.R-project.org/package=glmnet> (accessed 15 February 2019).
- Gagliardi, P.A., Puliafito, A. and Primo, L. (2017), “PDK1: At the crossroad of cancer signaling pathways”, presented at the Seminars in Cancer Biology, Elsevier.
- Ganster, H., Pinz, P., Rohrer, R., Wildling, E., Binder, M. and Kittler, H. (2001), “Automated melanoma recognition”, *IEEE Transactions on Medical Imaging*, Vol. 20 No. 3, pp. 233–239.

- Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., et al. (2013), “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”, *Sci Signal*, Vol. 6 No. 269, p. p11.
- Garbe, C. and Leiter, U. (2009), “Melanoma epidemiology and trends”, *Clinics in Dermatology*, Vol. 27 No. 1, pp. 3–9.
- Garcia-Arroyo, J.L. and Garcia-Zapirain, B. (2018), “Recognition of pigment network pattern in dermoscopy images based on fuzzy classification of pixels”, *Computer Methods and Programs in Biomedicine*, Vol. 153, pp. 61–69.
- Garraway, L.A., Widlund, H.R., Rubin, M.A., Getz, G., Berger, A.J., Ramaswamy, S., Beroukhim, R., et al. (2005), “Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma”, *Nature*, Vol. 436 No. 7047, pp. 117–122.
- Garrison, E. and Marth, G. (2012), “Haplotype-based variant detection from short-read sequencing”, *ArXiv Preprint ArXiv:1207.3907*.
- “Genomic Classification of Cutaneous Melanoma”. (2015), *Cell*, Vol. 161 No. 7, pp. 1681–96.
- “genomics | Learn Science at Scitable”. (n.d.). , available at:  
<https://www.nature.com/scitable/definition/genomics-126> (accessed 10 April 2019).
- Gerami, P., Jewell, S.S., Morrison, L.E., Blondin, B., Schulz, J., Ruffalo, T., Matushek IV, P., et al. (2009), “Fluorescence in situ hybridization (FISH) as an ancillary diagnostic tool in the diagnosis of melanoma”, *The American Journal of Surgical Pathology*, Vol. 33 No. 8, pp. 1146–1156.
- Gerstenblith, M.R., Shi, J. and Landi, M.T. (2010), “Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis”, *Pigment Cell & Melanoma Research*, Vol. 23 No. 5, pp. 587–606.

- Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H. and Beerenwinkel, N. (2012), “Reliable detection of subclonal single-nucleotide variants in tumour cell populations”, *Nature Communications*, Vol. 3, p. 811.
- Gevaert, O., Xu, J., Hoang, C.D., Leung, A.N., Xu, Y., Quon, A., Rubin, D.L., et al. (2012), “Non–Small Cell Lung Cancer: Identifying Prognostic Imaging Biomarkers by Leveraging Public Gene Expression Microarray Data—Methods and Preliminary Results”, *Radiology*, Vol. 264 No. 2, pp. 387–396.
- Gibney, G.T., Weiner, L.M. and Atkins, M.B. (2016), “Predictive biomarkers for checkpoint inhibitor-based immunotherapy”, *Lancet Oncol*, Vol. 17 No. 12, pp. e542–e551.
- Gonzalez-Perez, A., Deu-Pons, J. and Lopez-Bigas, N. (2012), “Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation”, *Genome Medicine*, Vol. 4 No. 11, p. 89.
- González-Pérez, A. and López-Bigas, N. (2011), “Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel”, *American Journal of Human Genetics*, Vol. 88 No. 4, pp. 440–449.
- Gonzalez-Perez, A. and Lopez-Bigas, N. (2012), “Functional impact bias reveals cancer drivers”, *Nucleic Acids Research*, p. gks743.
- Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., et al. (2013), “IntOGen-mutations identifies cancer drivers across tumor types”, *Nature Methods*, Vol. 10 No. 11, pp. 1081–1082.
- Goodson, A.G. and Grossman, D. (2009), “Strategies for early melanoma detection: approaches to the patient with nevi”, *Journal of the American Academy of Dermatology*, Vol. 60 No. 5, pp. 719–735.

- Grana, C., Pellacani, G., Cucchiara, R. and Seidenari, S. (2003), "A new algorithm for border description of polarized light surface microscopic images of pigmented skin lesions", *IEEE Transactions on Medical Imaging*, Vol. 22 No. 8, pp. 959–964.
- Grana, C., Pellacani, G., Seidenari, S. and Cucchiara, R. (2004), "Color calibration for a dermatological video camera system", Vol. 3, presented at the Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, IEEE, pp. 798–801.
- Greene, C.S., Tan, J., Ung, M., Moore, J.H. and Cheng, C. (2016), "Big Data Bioinformatics", *Methods (San Diego, Calif.)*, Vol. 111, pp. 1–2.
- Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., et al. (2007), "Patterns of somatic mutation in human cancer genomes", *Nature*, Vol. 446 No. 7132, pp. 153–158.
- Greenwell, B., Boehmke, B., Cunningham, J. and Developers (<https://github.com/gbm-developers>), G.B.M. (2019), *Gbm: Generalized Boosted Regression Models*, available at: <https://CRAN.R-project.org/package=gbm> (accessed 15 February 2019).
- Griffiths, A.J., Miller, J.H., Suzuki, D.T., Lewontin, R.C. and Gelbart, W.M. (2000), "Somatic versus germinal mutation", *An Introduction to Genetic Analysis. 7th Edition*, available at: <https://www.ncbi.nlm.nih.gov/books/NBK21894/> (accessed 27 February 2019).
- Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A. and Staudt, L.M. (2016), "Toward a Shared Vision for Cancer Genomic Data", *The New England Journal of Medicine*, Vol. 375 No. 12, pp. 1109–1112.
- Gutenev, A., Skladnev, V. and Varvel, D. (2001), "Acquisition-time image quality control in digital dermatoscopy of skin lesions", *Computerized Medical Imaging and Graphics*, Vol. 25 No. 6, pp. 495–499.

- Hajian-Tilaki, K. (2013), "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation", *Caspian J Intern Med*, Vol. 4 No. 2, pp. 627–35.
- Hansen, G.L., Sparrow, E.M., Kokate, J.Y., Leland, K.J. and Iaizzo, P.A. (1997), "Wound status evaluation using color image processing", *IEEE Transactions on Medical Imaging*, Vol. 16 No. 1, pp. 78–86.
- Hansen, N.F., Gartner, J.J., Mei, L., Samuels, Y. and Mullikin, J.C. (2013), "Shimmer: detection of genetic alterations in tumors using next-generation sequence data", *Bioinformatics (Oxford, England)*, Vol. 29 No. 12, pp. 1498–1503.
- Hart, I.R., Birch, M. and Marshall, J.F. (1991), "Cell adhesion receptor expression during melanoma progression and metastasis", *Cancer Metastasis Rev*, Vol. 10 No. 2, pp. 115–28.
- Hayward, N.K., Wilmott, J.S., Waddell, N., Johansson, P.A., Field, M.A., Nones, K., Patch, A.-M., et al. (2017), "Whole-genome landscapes of major melanoma subtypes", *Nature*, Vol. 545 No. 7653, pp. 175–180.
- He, K.Y., Ge, D. and He, M.M. (2017), "Big Data Analytics for Genomic Medicine", *International Journal of Molecular Sciences*, Vol. 18 No. 2, available at:<https://doi.org/10.3390/ijms18020412>.
- Helgadottir, H., Rocha Trocoli Drakensjö, I. and Girnita, A. (2018), "Personalized Medicine in Malignant Melanoma: Towards Patient Tailored Treatment", *Frontiers in Oncology*, Vol. 8, p. 202.
- Herbin, M., Bon, F.-X., Venot, A., Jeanlouis, F., Dubertret, M., Dubertret, L. and Strauch, G. (1993), "Assessment of healing kinetics through true color image processing", *IEEE Transactions on Medical Imaging*, Vol. 12 No. 1, pp. 39–43.
- Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R. and Taylor, R. (2005), "Can electronic medical record systems transform health care? Potential health benefits, savings, and costs", *Health Affairs (Project Hope)*, Vol. 24 No. 5, pp. 1103–1117.



- Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., et al. (2014), "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin", *Cell*, Vol. 158 No. 4, pp. 929–944.
- Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., et al. (2012), "A landscape of driver mutations in melanoma", *Cell*, Vol. 150 No. 2, pp. 251–63.
- Hoek, K., Rimm, D.L., Williams, K.R., Zhao, H., Ariyan, S., Lin, A., Kluger, H.M., et al. (2004), "Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas", *Cancer Research*, Vol. 64 No. 15, pp. 5270–5282.
- "Home - SNP - NCBI". (n.d.). , available at: <https://www.ncbi.nlm.nih.gov/snp> (accessed 14 February 2019).
- Hormozdiari, F., Hajirasouliha, I., McPherson, A., Eichler, E.E. and Sahinalp, S.C. (2011), "Simultaneous structural variation discovery among multiple paired-end sequenced genomes", *Genome Research*, Vol. 21 No. 12, pp. 2203–2212.
- Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., et al. (2013), "TERT promoter mutations in familial and sporadic melanoma", *Science (New York, N.Y.)*, Vol. 339 No. 6122, pp. 959–961.
- Hsu, Y.-C., Hsiao, Y.-T., Kao, T.-Y., Chang, J.-G. and Shieh, G.S. (2017), "Detection of Somatic Mutations in Exome Sequencing of Tumor-only Samples", *Scientific Reports*, Vol. 7 No. 1, p. 15959.
- Hua, X., Xu, H., Yang, Y., Zhu, J., Liu, P. and Lu, Y. (2013), "DrGaP: A Powerful Tool for Identifying Driver Genes and Pathways in Cancer Sequencing Studies", *American Journal of Human Genetics*, Vol. 93 No. 3, pp. 439–451.
- Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L. and Garraway, L.A. (2013), "Highly recurrent TERT promoter mutations in human melanoma", *Science (New York, N.Y.)*, Vol. 339 No. 6122, pp. 957–959.

- Huang, Q., Jing, S., Yi, J. and Zhen, W. (2015), *Innovative Testing and Measurement Solutions for Smart Grid*, John Wiley & Sons.
- Hussussian, C.J., Struewing, J.P., Goldstein, A.M., Higgins, P.A., Ally, D.S., Sheahan, M.D., Clark, W.H., et al. (1994), "Germline p16 mutations in familial melanoma", *Nature Genetics*, Vol. 8 No. 1, pp. 15–21.
- Iakovidis, D.K., Pelekis, N., Kotsifakos, E.E., Kopanakis, I., Karanikas, H. and Theodoridis, Y. (2009), "A pattern similarity scheme for medical image retrieval", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 13 No. 4, pp. 442–450.
- "Indelocator". (n.d.). *MuTect2*, available at: <http://archive.broadinstitute.org/cancer/cga/indelocator> (accessed 1 January 2016).
- Jaffe, C.C. (2012), "Imaging and Genomics: Is There a Synergy?", *Radiology*, Vol. 264 No. 2, pp. 329–331.
- Jain, S., Jagtap, V. and Pise, N. (2015), "Computer Aided Melanoma Skin Cancer Detection Using Image Processing", *Procedia Computer Science*, Vol. 48, pp. 735–740.
- Jamshidi, N., Diehn, M., Bredel, M. and Kuo, M.D. (2013), "Illuminating Radiogenomic Characteristics of Glioblastoma Multiforme through Integration of MR Imaging, Messenger RNA Expression, and DNA Copy Number Variation", *Radiology*, Vol. 270 No. 1, pp. 1–2.
- Jensen, E.H., Lewis, J.M., McLoughlin, J.M., Alvarado, M.D., Daud, A., Messina, J., Enkemann, S., et al. (2007), "Down-regulation of pro-apoptotic genes is an early event in the progression of malignant melanoma", *Annals of Surgical Oncology*, Vol. 14 No. 4, pp. 1416–1423.
- Jimenez-Sanchez, G. (2015), "Genomics: the Power and the Promise", *Genome*, Vol. 58 No. 12, pp. vii–x.
- Jogi, A., Vaapil, M., Johansson, M. and Pahlman, S. (2012), "Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors", *Ups J Med Sci*, Vol. 117 No. 2, pp. 217–24.
- Jones, D., Raine, K.M., Davies, H., Tarpey, P.S., Butler, A.P., Teague, J.W., Nik-Zainal, S., et al. (2016), "cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single

- Nucleotide Variants in NGS Data”, *Current Protocols in Bioinformatics*, Vol. 56, pp. 15.10.1-15.10.18.
- Kalatskaya, I., Trinh, Q.M., Spears, M., McPherson, J.D., Bartlett, J.M.S. and Stein, L. (2017), “ISOWN: accurate somatic mutation identification in the absence of normal tissue controls”, *Genome Medicine*, Vol. 9 No. 1, p. 59.
- Kamb, A., Shattuck-Eidens, D., Eeles, R., Liu, Q., Gruis, N.A., Ding, W., Hussey, C., et al. (1994), “Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus”, *Nat Genet*, Vol. 8 No. 1, pp. 23–6.
- Kaminker, J.S., Zhang, Y., Watanabe, C. and Zhang, Z. (2007), “CanPredict: a computational tool for predicting cancer-associated missense mutations”, *Nucleic Acids Research*, Vol. 35 No. Web Server issue, pp. W595-598.
- Kaminker, J.S., Zhang, Y., Waugh, A., Haverty, P.M., Peters, B., Sebisano, D., Stinson, J., et al. (2007), “Distinguishing cancer-associated missense mutations from common polymorphisms”, *Cancer Research*, Vol. 67 No. 2, pp. 465–473.
- Kashani-Sabet, M., Venna, S., Nosrati, M., Rangel, J., Sucker, A., Egberts, F., Baehner, F.L., et al. (2009), “A multimarker prognostic assay for primary cutaneous melanoma”, *Clinical Cancer Research*, Vol. 15 No. 22, pp. 6987–6992.
- Kasmi, R. and Mokrani, K. (2016), “Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule”, *IET Image Processing*, Vol. 10 No. 6, pp. 448–455.
- Kassahn, K.S., Holmes, O., Nones, K., Patch, A.-M., Miller, D.K., Christ, A.N., Harliwong, I., et al. (2013), “Somatic point mutation calling in low cellularity tumors”, *PLoS One*, Vol. 8 No. 11, p. e74380.
- Kato, T., Kawai, K., Egami, Y., Kakehi, Y. and Araki, N. (2014), “Rac1-dependent lamellipodial motility in prostate cancer PC-3 cells revealed by optogenetic control of Rac1 activity”, *PLoS One*, Vol. 9 No. 5, p. e97749.

- Katrib, A., Hsu, W., Bui, A. and Xing, Y. (2016), “‘RADIOTRANSCRIPTOMICS’: A synergy of imaging and transcriptomics in clinical assessment”, *Quantitative Biology (Beijing, China)*, Vol. 4 No. 1, pp. 1–12.
- Kaur, R., Albano, P.P., Cole, J.G., Hagerty, J., LeAnder, R.W., Moss, R.H. and Stoecker, W.V. (2015), “Real-time supervised detection of pink areas in dermoscopic images of melanoma: importance of color shades, texture and location”, *Skin Research and Technology*, Vol. 21 No. 4, pp. 466–473.
- Kavakiotis, I., Xochelli, A., Agathangelidis, A., Tsoumakas, G., Maglaveras, N., Stamatopoulos, K., Hadzidimitriou, A., et al. (2016), “Integrating multiple immunogenetic data sources for feature extraction and mining somatic hypermutation patterns: the case of ‘towards analysis’ in chronic lymphocytic leukaemia”, *BMC Bioinformatics*, Vol. 17 No. 5, p. 173.
- Kim, S., Jeong, K., Bhutani, K., Lee, J., Patel, A., Scott, E., Nam, H., et al. (2013), “Virmid: accurate detection of somatic mutations with sample impurity inference”, *Genome Biology*, Vol. 14 No. 8, p. R90.
- Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., et al. (2012), “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing”, *Genome Res*, Vol. 22 No. 3, pp. 568–76.
- Kockan, C., Hach, F., Sarrafi, I., Bell, R.H., McConeghy, B., Beja, K., Haegert, A., et al. (2017), “SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA”, *Bioinformatics (Oxford, England)*, Vol. 33 No. 1, pp. 26–34.
- Kong, Y., Kumar, S.M. and Xu, X. (2010), “Molecular Pathogenesis of Sporadic Melanoma and Melanoma-Initiating Cells”, *Archives of Pathology & Laboratory Medicine*, Vol. 134 No. 12, pp. 1740–1749.

- Kontogianni, G., Papadodima, O., Maglogiannis, I., Frangia-Tsivou, K. and Chatziioannou, A. (2016), "Integrative Bioinformatic Analysis of a Greek Epidemiological Cohort Provides Insight into the Pathogenesis of Primary Cutaneous Melanoma".
- Kontogianni, G., Piroti, G., Maglogiannis, I., Chatziioannou, A. and Papadodima, O. (2018), "Dissecting the Mutational Landscape of Cutaneous Melanoma: An Omic Analysis Based on Patients from Greece", *Cancers*, Vol. 10 No. 4, p. 96.
- Kor, S. and Tiwary, U. (2004), "Feature level fusion of multimodal medical images in lifting wavelet transform domain", *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, Vol. 2, pp. 1479–1482.
- Körner, H., Epanchintsev, A., Berking, C., Schuler-Thurner, B., Speicher, M.R., Menssen, A. and Hermeking, H. (2007), "Digital karyotyping reveals frequent inactivation of the dystrophin/DMD gene in malignant melanoma", *Cell Cycle*, Vol. 6 No. 2, pp. 189–198.
- Korotkov, K. and Garcia, R. (2012), "Computerized analysis of pigmented skin lesions: a review", *Artificial Intelligence in Medicine*, Vol. 56 No. 2, pp. 69–90.
- Koutsandreas, T., Binenbaum, I., Pilalis, E., Valavanis, I., Papadodima, O. and Chatziioannou, A. (2016), "Analyzing and visualizing genomic complexity for the derivation of the emergent molecular networks", *International Journal of Monitoring and Surveillance Technologies Research (IJMSTR)*, Vol. 4 No. 2, pp. 30–49.
- Krauthammer, M., Kong, Y., Ha, B.H., Evans, P., Bacchiocchi, A., McCusker, J.P., Cheng, E., et al. (2012), "Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma", *Nat Genet*, Vol. 44 No. 9, pp. 1006–14.

- Krumm, N., Sudmant, P.H., Ko, A., O’Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., et al. (2012), “Copy number variation detection and genotyping from exome sequence data”, *Genome Research*, Vol. 22 No. 8, pp. 1525–1532.
- Kuhn, M., Weston, S., Culp, M., Coulter, N., code), R.Q. (Author of imported C., code), R.R. (Copyright holder of imported C. and code), R.R.P.L. (Copyright holder of imported C. (2018), *C50: C5.0 Decision Trees and Rule-Based Models*, available at: <https://CRAN.R-project.org/package=C50> (accessed 15 February 2019).
- Kwasnicka, H. and Paradowski, M. (2005), “Melanocytic lesion images segmentation enforcing by spatial relations based declarative knowledge”, presented at the Intelligent Systems Design and Applications, 2005. ISDA’05. Proceedings. 5th International Conference on, IEEE, pp. 286–291.
- Kypreou, K.P., Stefanaki, I., Antonopoulou, K., Karagianni, F., Ntritsos, G., Zaras, A., Nikolaou, V., et al. (2016), “Prediction of Melanoma Risk in a Southern European Population Based on a Weighted Genetic Risk Score”, *J Invest Dermatol*, Vol. 136 No. 3, pp. 690–5.
- Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., et al. (2016), “VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research”, *Nucleic Acids Research*, Vol. 44 No. 11, p. e108.
- Lambin, P., van Stiphout, R.G.P.M., Starmans, M.H.W., Rios-Velazquez, E., Nalbantov, G., Aerts, H.J.W.L., Roelofs, E., et al. (2013), “Predicting outcomes in radiation oncology—multifactorial decision support systems”, *Nature Reviews Clinical Oncology*, Vol. 10 No. 1, pp. 27–40.
- Lanckriet, G.R., De Bie, T., Cristianini, N., Jordan, M.I. and Noble, W.S. (2004), “A statistical framework for genomic data fusion”, *Bioinformatics*, Vol. 20 No. 16, pp. 2626–2635.
- Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., et al. (2012), “SomaticSniper: identification of somatic point mutations in whole genome sequencing data”, *Bioinformatics (Oxford, England)*, Vol. 28 No. 3, pp. 311–317.

- Law, M.H., Bishop, D.T., Lee, J.E., Brossard, M., Martin, N.G., Moses, E.K., Song, F., et al. (2015), "Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma", *Nat Genet*, Vol. 47 No. 9, pp. 987–95.
- Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., et al. (2013), "Mutational heterogeneity in cancer and the search for new cancer-associated genes", *Nature*, Vol. 499 No. 7457, pp. 214–8.
- Lazova, R., Pornputtapong, N., Halaban, R., Bosenberg, M., Bai, Y., Chai, H. and Krauthammer, M. (2017), "Spitz nevi and Spitzoid melanomas: exome sequencing and comparison with conventional melanocytic nevi and melanomas", *Modern Pathology*, Vol. 30 No. 5, pp. 640–649.
- LeDell, E., Petersen, M. and Laan, M. van der. (2014), *CvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals*, available at: <https://CRAN.R-project.org/package=cvAUC> (accessed 15 February 2019).
- Lee, T., Ng, V., Gallagher, R., Coldman, A. and McLean, D. (1997), "DullRazor: a software approach to hair removal from images", *Computers in Biology and Medicine*, Vol. 27 No. 6, pp. 533–543.
- Lee, T.K., Atkins, M.S., King, M.A., Lau, S. and McLean, D.I. (2005), "Counting moles automatically from back images", *IEEE Transactions on Biomedical Engineering*, Vol. 52 No. 11, pp. 1966–1969.
- Lefevre, E., Colot, O., Vannoorenberghe, P. and de Bruçq, D. (2000), "Knowledge modeling methods in the framework of evidence theory: an experimental comparison for melanoma detection", Vol. 4, presented at the Systems, Man, and Cybernetics, 2000 IEEE International Conference on, IEEE, pp. 2806–2811.
- Leiserson, M.D.M., Blokh, D., Sharan, R. and Raphael, B.J. (2013), "Simultaneous identification of multiple driver pathways in cancer", *PLoS Computational Biology*, Vol. 9 No. 5, p. e1003054.

- Li, H. (2011), "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data", *Bioinformatics (Oxford, England)*, Vol. 27 No. 21, pp. 2987–2993.
- Li, H. and Durbin, R. (2010), "Fast and accurate long-read alignment with Burrows-Wheeler transform", *Bioinformatics*, Vol. 26 No. 5, pp. 589–95.
- Li, Y. and Patra, J.C. (2010), "Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network", *Bioinformatics*, Vol. 26 No. 9, pp. 1219–1224.
- Liu, W., Peng, Y. and Tobin, D.J. (2013), "A new 12-gene diagnostic biomarker signature of melanoma revealed by integrated microarray analysis", *PeerJ*, Vol. 1, p. e49.
- Liu, X., Jian, X. and Boerwinkle, E. (2013), "dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations", *Hum Mutat*, Vol. 34 No. 9, pp. E2393-402.
- Liu, X., Wu, C., Li, C. and Boerwinkle, E. (2016), "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs", *Hum Mutat*, Vol. 37 No. 3, pp. 235–41.
- Liu, Y., Loewer, M., Aluru, S. and Schmidt, B. (2016), "SNVsniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations", *BMC Systems Biology*, Vol. 10 Suppl 2, p. 47.
- Lohmann, W. and Paul, E. (1988), "In situ detection of melanomas by fluorescence measurements", *Naturwissenschaften*, Vol. 75 No. 4, pp. 201–202.
- Luce, L.N., Abbate, M., Cotignola, J. and Giliberto, F. (2017), "Non-myogenic tumors display altered expression of dystrophin (DMD) and a high frequency of genetic alterations", *Oncotarget*, Vol. 8 No. 1, p. 145.
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., et al. (2016), "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)", *Nucleic Acids Research*, Vol. 45 No. D1, pp. D896–D901.



- Machesky, L.M. (2008), "Lamellipodia and filopodia in metastasis and invasion", *FEBS Lett*, Vol. 582 No. 14, pp. 2102–11.
- Maclin, P.S., Dempsey, J., Brooks, J. and Rand, J. (1991), "Using neural networks to diagnose cancer", *Journal of Medical Systems*, Vol. 15 No. 1, pp. 11–19.
- Madhunapantula, S.V. and Robertson, G.P. (2009), "The PTEN-AKT3 Signaling Cascade as a Therapeutic Target in Melanoma", *Pigment Cell & Melanoma Research*, Vol. 22 No. 4, pp. 400–419.
- Maglogiannis, I., Caroni, C., Pavlopoulos, S., Karioti, V. and Koutsouris, D. (2001), "Utilizing artificial intelligence for the characterization of dermatological images", presented at the 4th International Conference Neural Networks and Expert Systems in Medicine and Healthcare, Milos Island, Greece, pp. 362–368.
- Maglogiannis, I. and Doukas, C.N. (2009), "Overview of advanced computer vision systems for skin lesions characterization", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 13 No. 5, pp. 721–733.
- Maglogiannis, I. and Kosmopoulos, D.I. (2006), "Computational vision systems for the detection of malignant melanoma", *Oncology Reports*, Vol. 15 No. 4, pp. 1027–1032.
- Maglogiannis, I., Pavlopoulos, S. and Koutsouris, D. (2005), "An integrated computer supported acquisition, handling, and characterization system for pigmented skin lesions in dermatological images", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 9 No. 1, pp. 86–98.
- Maglogiannis, I.G. and Zafiroopoulos, E.P. (2004), "Characterization of digital medical images utilizing support vector machines", *BMC Medical Informatics and Decision Making*, Vol. 4 No. 1, p. 4.
- Majewski, J., Schwartzenuber, J., Lalonde, E., Montpetit, A. and Jabado, N. (2011), "What can exome sequencing do for you?", *J Med Genet*, Vol. 48 No. 9, pp. 580–9.

- Maldonado, J.L., Fridlyand, J., Patel, H., Jain, A.N., Busam, K., Kageshita, T., Ono, T., et al. (2003), "Determinants of BRAF mutations in primary melanomas", *Journal of the National Cancer Institute*, Vol. 95 No. 24, pp. 1878–1890.
- Malhotra, A., Lindberg, M., Faust, G.G., Leibowitz, M.L., Clark, R.A., Layer, R.M., Quinlan, A.R., et al. (2013), "Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms", *Genome Research*, Vol. 23 No. 5, pp. 762–776.
- Mann, G.J., Pupo, G.M., Campain, A.E., Carter, C.D., Schramm, S.-J., Pianova, S., Gerega, S.K., et al. (2013), "BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma", *Journal of Investigative Dermatology*, Vol. 133 No. 2, pp. 509–517.
- Manousaki, A.G., Manios, A.G., Tsompanaki, E.I., Panayiotides, J.G., Tsiftsis, D.D., Kostaki, A.K. and Tosca, A.D. (2006), "A simple digital image processing system to aid in melanoma diagnosis in an everyday melanocytic skin lesion unit. A preliminary report", *International Journal of Dermatology*, Vol. 45 No. 4, pp. 402–410.
- Martin-Belmonte, F. and Perez-Moreno, M. (2011), "Epithelial cell polarity, stem cells and cancer", *Nat Rev Cancer*, Vol. 12 No. 1, pp. 23–38.
- Maurer, M., Su, T., Saal, L.H., Koujak, S., Hopkins, B.D., Barkley, C.R., Wu, J., et al. (2009), "3-Phosphoinositide-dependent kinase 1 potentiates upstream lesions on the phosphatidylinositol 3-kinase pathway in breast carcinoma", *Cancer Research*, Vol. 69 No. 15, pp. 6299–6306.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., et al. (2010), "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data", *Genome Res*, Vol. 20 No. 9, pp. 1297–303.

- Mehrotra, A., Mehta, G., Aras, S., Trivedi, A. and de la Serna, I.L. (2014), "SWI/SNF chromatin remodeling enzymes in melanocyte differentiation and melanoma", *Critical Reviews in Eukaryotic Gene Expression*, Vol. 24 No. 2, pp. 151–161.
- Melamed, R.D., Aydin, I.T., Rajan, G.S., Phelps, R., Silvers, D.N., Emmett, K.J., Brunner, G., et al. (2017), "Genomic characterization of dysplastic nevi unveils implications for diagnosis of melanoma", *Journal of Investigative Dermatology*, Vol. 137 No. 4, pp. 905–909.
- "Melanoma Research Foundation". (n.d.). *Melanoma Research Foundation*, available at: <https://www.melanoma.org/home-page> (accessed 11 February 2019).
- Menon, R., Deng, M., Boehm, D., Braun, M., Fend, F., Boehm, D., Biskup, S., et al. (2012), "Exome enrichment and SOLiD sequencing of formalin fixed paraffin embedded (FFPE) prostate cancer tissue", *International Journal of Molecular Sciences*, Vol. 13 No. 7, pp. 8933–8942.
- Menzies, S.W., Ingvar, C., Crotty, K.A. and McCarthy, W.H. (1996), "Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features", *Archives of Dermatology*, Vol. 132 No. 10, pp. 1178–1182.
- Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhi, R. and Getz, G. (2011), "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers", *Genome Biology*, Vol. 12 No. 4, p. R41.
- Messadi, M., Bessaid, A. and Taleb-Ahmed, A. (2009), "Extraction of specific parameters for skin tumour classification", *Journal of Medical Engineering & Technology*, Vol. 33 No. 4, pp. 288–295.
- Mete, M., Kockara, S. and Aydin, K. (2011), "Fast density-based lesion detection in dermoscopy images", *Computerized Medical Imaging and Graphics*, Vol. 35 No. 2, pp. 128–136.
- Metsis, V., Huang, H., Andronesi, O.C., Makedon, F. and Tzika, A. (2012), "Heterogeneous data fusion for brain tumor classification", *Oncology Reports*, Vol. 28 No. 4, pp. 1413–1416.

- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., C++-code), C.-C.C. (libsvm and C++-code), C.-C.L. (libsvm. (2019), *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, available at: <https://CRAN.R-project.org/package=e1071> (accessed 15 February 2019).
- Miller, C.A., Settle, S.H., Sulman, E.P., Aldape, K.D. and Milosavljevic, A. (2011), “Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors”, *BMC Medical Genomics*, Vol. 4, p. 34.
- Mirmohammadsadegh, A., Marini, A., Nambiar, S., Hassan, M., Tannapfel, A., Ruzicka, T. and Hengge, U.R. (2006), “Epigenetic silencing of the PTEN gene in melanoma”, *Cancer Research*, Vol. 66 No. 13, pp. 6546–6552.
- Mishra, N.K. and Celebi, M.E. (2016), “An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning”, *ArXiv:1601.07843 [Cs, Stat]*, available at: <http://arxiv.org/abs/1601.07843> (accessed 16 May 2019).
- Mitchell, H.B. (2012), *Data Fusion: Concepts and Ideas*, Springer Science & Business Media.
- Mort, R.L., Jackson, I.J. and Patton, E.E. (2015), “The melanocyte lineage in development and disease”, *Development (Cambridge, England)*, Vol. 142 No. 4, pp. 620–632.
- Motoyama, H., Tanaka, T., Tanaka, M. and Oka, H. (2004), “Feature of malignant melanoma based on color information”, Vol. 1, presented at the SICE 2004 Annual Conference, IEEE, pp. 230–233.
- Moutselos, K., Maglogiannis, I. and Chatziioannou, A. (2011), “GOrevenge: a novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies”, *IEEE Trans Biomed Eng*, Vol. 58 No. 12, pp. 3522–7.
- Moutselos, K., Maglogiannis, I. and Chatziioannou, A. (2014), “Integration of high-volume molecular and imaging data for composite biomarker discovery in the study of melanoma”, *Biomed Res Int*, Vol. 2014, p. 145243.

- Muller, E., Goardon, N., Brault, B., Rousselin, A., Paimparay, G., Legros, A., Fouillet, R., et al. (2016), "OutLyzer: software for extracting low-allele-frequency tumor mutations from sequencing background noise in clinical practice", *Oncotarget*, Vol. 7 No. 48, pp. 79485–79493.
- Nan, H., Kraft, P., Hunter, D.J. and Han, J. (2009), "Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians", *International Journal of Cancer*, Vol. 125 No. 4, pp. 909–917.
- Nasir, M., Attique Khan, M., Sharif, M., Lali, I.U., Saba, T. and Iqbal, T. (2018), "An improved strategy for skin lesion detection and classification using uniform segmentation and feature selection based approach", *Microscopy Research and Technique*, Vol. 81 No. 6, pp. 528–543.
- Ng, P.C. and Henikoff, S. (2001), "Predicting deleterious amino acid substitutions", *Genome Research*, Vol. 11 No. 5, pp. 863–874.
- Nikolaev, S.I., Rimoldi, D., Iseli, C., Valsesia, A., Robyr, D., Gehrig, C., Harshman, K., et al. (2011), "Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma", *Nat Genet*, Vol. 44 No. 2, pp. 133–9.
- Nikolaou, V. and Stratigos, A. (2014), "Emerging trends in the epidemiology of melanoma", *British Journal of Dermatology*, Vol. 170 No. 1, pp. 11–19.
- Ogorzałek, M., Nowak, L., Surowka, G. and Alekseenko, A. (2011), "Melanoma in the clinic—diagnosis, management and complications of malignancy", *Modern Techniques for Computer-Aided Melanoma Diagnosis*.
- Organization, W.H. and Safety, I.P. on C. (2001), *Biomarkers in Risk Assessment : Validity and Validation*, Geneva : World Health Organization, available at:  
<https://apps.who.int/iris/handle/10665/42363> (accessed 23 February 2019).
- Ortega-Molina, A. and Serrano, M. (2013), "PTEN in cancer, metabolism, and aging", *Trends in Endocrinology and Metabolism: TEM*, Vol. 24 No. 4, pp. 184–189.

- Otsu, N. (1975), "A threshold selection method from gray-level histograms", *Automatica*, Vol. 11 No. 285–296, pp. 23–27.
- Padua, R.A., Barrass, N. and Currie, G.A. (1984), "A novel transforming gene in a human malignant melanoma cell line", *Nature*, Vol. 311 No. 5987, pp. 671–673.
- Palmieri, G., Colombino, M., Casula, M., Manca, A., Mandalà, M., Cossu, A. and Italian Melanoma Intergroup (IMI). (2018), "Molecular Pathways in Melanomagenesis: What We Learned from Next-Generation Sequencing Approaches", *Current Oncology Reports*, Vol. 20 No. 11, p. 86.
- Parmar, C., Grossmann, P., Bussink, J., Lambin, P. and Aerts, H.J.W.L. (2015), "Machine Learning methods for Quantitative Radiomic Biomarkers", *Scientific Reports*, Vol. 5, p. 13087.
- Pasquale, E.B. (2010), "Eph receptors and ephrins in cancer: bidirectional signalling and beyond", *Nat Rev Cancer*, Vol. 10 No. 3, pp. 165–80.
- Patwardhan, S.V., Dai, S. and Dhawan, A.P. (2005), "Multi-spectral image analysis and classification of melanoma using fuzzy membership based partitions", *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, Vol. 29 No. 4, pp. 287–296.
- Pehamberger, H., Steiner, A. and Wolff, K. (1987), "In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions", *Journal of the American Academy of Dermatology*, Vol. 17 No. 4, pp. 571–583.
- "Picard Tools - By Broad Institute". (n.d.). , available at: <http://broadinstitute.github.io/picard/> (accessed 12 February 2019).
- Pikor, L.A., Enfield, K.S., Cameron, H. and Lam, W.L. (2011), "DNA extraction from paraffin embedded material for genetic and epigenetic analyses", *JoVE (Journal of Visualized Experiments)*, No. 49, pp. e2763–e2763.

- Pilalis, E.D. and Chatziioannou, A.A. (2013), "Prioritized functional analysis of biological experiments using resampling and noise control methodologies", presented at the Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, IEEE, pp. 1–3.
- Piskol, R., Ramaswami, G. and Li, J.B. (2013), "Reliable identification of genomic variants from RNA-seq data", *The American Journal of Human Genetics*, Vol. 93 No. 4, pp. 641–651.
- Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., et al. (2010), "A comprehensive catalogue of somatic mutations from a human cancer genome", *Nature*, Vol. 463 No. 7278, pp. 191–196.
- Prelich, G. (2012), "Gene Overexpression: Uses, Mechanisms, and Interpretation", *Genetics*, Vol. 190 No. 3, pp. 841–854.
- Przybyła-Kasperek, M., Nowak-Brzezińska, A. and Simiński, R. (2017), "Decision Fusion Methods in a Dispersed Decision System - A Comparison on Medical Data", in Nguyen, N.T., Papadopoulos, G.A., Jędrzejowicz, P., Trawiński, B. and Vossen, G. (Eds.), *Computational Collective Intelligence*, Springer International Publishing, pp. 139–149.
- Radenbaugh, A.J., Ma, S., Ewing, A., Stuart, J.M., Collisson, E.A., Zhu, J. and Haussler, D. (2014), "RADIA: RNA and DNA integrated analysis for somatic mutation detection", *PLoS One*, Vol. 9 No. 11, p. e111516.
- Ragnarsson-Olding, B., Platz, A., Olding, L. and Ringborg, U. (2004), "p53 protein expression and TP53 mutations in malignant melanomas of sun-sheltered mucosal membranes versus chronically sun-exposed skin", *Melanoma Research*, Vol. 14 No. 5, pp. 395–401.
- Rahman, M.M. and Bhattacharya, P. (2010), "An integrated and interactive decision support system for automated melanoma recognition of dermoscopic images", *Computerized Medical Imaging and Graphics*, Vol. 34 No. 6, pp. 479–486.

- Ramos, A.H., Lichtenstein, L., Gupta, M., Lawrence, M.S., Pugh, T.J., Saksena, G., Meyerson, M., et al. (2015), "Oncotator: cancer variant annotation tool", *Hum Mutat*, Vol. 36 No. 4, pp. E2423-9.
- Raphael, B.J., Dobson, J.R., Oesper, L. and Vandin, F. (2014), "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine", *Genome Medicine*, Vol. 6 No. 1, p. 5.
- Raskin, L., Fullen, D.R., Giordano, T.J., Thomas, D.G., Frohm, M.L., Cha, K.B., Ahn, J., et al. (2013), "Transcriptome profiling identifies HMGA2 as a biomarker of melanoma progression and prognosis", *J Invest Dermatol*, Vol. 133 No. 11, pp. 2585–92.
- Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. and Korbel, J.O. (2012), "DELLY: structural variant discovery by integrated paired-end and split-read analysis", *Bioinformatics*, Vol. 28 No. 18, pp. i333–i339.
- Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S.K., Tourna, A., Yakovleva, A., et al. (2019), "The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens", *Genome Biology*, Vol. 20 No. 1, p. 1.
- Reva, B., Antipin, Y. and Sander, C. (2007), "Determinants of protein function revealed by combinatorial entropy optimization", *Genome Biology*, Vol. 8 No. 11, p. R232.
- Reva, B., Antipin, Y. and Sander, C. (2011), "Predicting the functional impact of protein mutations: application to cancer genomics", *Nucleic Acids Research*, p. gkr407.
- Ridley, A.J. (2004), "Rho proteins and cancer", *Breast Cancer Res Treat*, Vol. 84 No. 1, pp. 13–9.
- Riker, A.I., Enkemann, S.A., Fodstad, O., Liu, S., Ren, S., Morris, C., Xi, Y., et al. (2008), "The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis", *BMC Medical Genomics*, Vol. 1 No. 1, p. 13.



- Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., WGS500 Consortium, Wilkie, A.O.M., et al. (2014), "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications", *Nature Genetics*, Vol. 46 No. 8, pp. 912–918.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015), "limma powers differential expression analyses for RNA-sequencing and microarray studies", *Nucleic Acids Research*, Vol. 43 No. 7, p. e47.
- Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G. and Zentgraf, M. (2018), *KlaR: Classification and Visualization*, available at: <https://CRAN.R-project.org/package=klaR> (accessed 15 February 2019).
- Romero, R., Espinoza, J., Gotsch, F., Kusanovic, J.P., Friel, L.A., Erez, O., Mazaki-Tovi, S., et al. (2006), "The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome", *BJOG: An International Journal of Obstetrics and Gynaecology*, Vol. 113 Suppl 3, pp. 118–135.
- Rose, A.E., Poliseno, L., Wang, J., Clark, M., Pearlman, A., Wang, G., Medicherla, R., et al. (2011), "Integrative genomics identifies molecular alterations that challenge the linear model of melanoma progression", *Cancer Research*, Vol. 71 No. 7, pp. 2561–2571.
- Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., Bashashati, A., et al. (2012), "JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data", *Bioinformatics (Oxford, England)*, Vol. 28 No. 7, pp. 907–913.
- Rothberg, B.E.G., Bracken, M.B. and Rimm, D.L. (2009), "Tissue biomarkers for prognosis in cutaneous melanoma: a systematic review and meta-analysis", *Journal of the National Cancer Institute*.
- Rubegni, P., Cevenini, G., Burrioni, M., Perotti, R., Dell'Eva, G., Sbano, P., Miracco, C., et al. (2002), "Automated diagnosis of pigmented skin lesions", *International Journal of Cancer*, Vol. 101 No. 6, pp. 576–580.

- Ruela, M., Barata, C., Marques, J.S. and Rozeira, J. (2017), "A system for the detection of melanomas in dermoscopy images using shape and symmetry features", *Computer Methods in Biomechanics and Biomedical Engineering-Imaging and Visualization*, Vol. 5 No. 2, pp. 127–137.
- Ryslik, G.A., Cheng, Y., Cheung, K.-H., Modis, Y. and Zhao, H. (2013), "Utilizing protein structure to identify non-random somatic mutations", *BMC Bioinformatics*, Vol. 14, p. 190.
- Sadeghi, M., Lee, T.K., McLean, D., Lui, H. and Atkins, M.S. (2013), "Detection and Analysis of Irregular Streaks in Dermoscopic Images of Skin Lesions", *IEEE Transactions on Medical Imaging*, Vol. 32 No. 5, pp. 849–861.
- Sadeghi, M., Razmara, M., Lee, T.K. and Atkins, M.S. (2011), "A novel method for detection of pigment network in dermoscopic images using graphs", *Computerized Medical Imaging and Graphics*, Vol. 35 No. 2, pp. 137–143.
- Saladi, S.V., Keenen, B., Marathe, H.G., Qi, H., Chin, K.V. and de la Serna, I.L. (2010), "Modulation of extracellular matrix/adhesion molecule expression by BRG1 is associated with increased melanoma invasiveness", *Mol Cancer*, Vol. 9, p. 280.
- Sanchez-Garcia, F., Akavia, U.D., Mozes, E. and Pe'er, D. (2010), "JISTIC: Identification of Significant Targets in Cancer", *BMC Bioinformatics*, Vol. 11, p. 189.
- Sander, E.E. and Collard, J.G. (1999), "Rho-like GTPases: their role in epithelial cell-cell adhesion and invasion", *Eur J Cancer*, Vol. 35 No. 14, pp. 1905–11.
- Sanders, J., Goldstein, B., Leotta, D. and Richards, K. (1999), "Image processing techniques for quantitative analysis of skin structures", *Computer Methods and Programs in Biomedicine*, Vol. 59 No. 3, pp. 167–180.
- Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J. and Cheetham, R.K. (2012), "Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs", *Bioinformatics (Oxford, England)*, Vol. 28 No. 14, pp. 1811–1817.

- Scatolini, M., Grand, M.M., Grosso, E., Venesio, T., Pisacane, A., Balsamo, A., Sirovich, R., et al. (2010), "Altered molecular pathways in melanocytic lesions", *International Journal of Cancer*, Vol. 126 No. 8, pp. 1869–1881.
- Schaefer, G., Krawczyk, B., Celebi, M.E. and Iyatomi, H. (2014), "An ensemble classification approach for melanoma diagnosis", *Memetic Computing*, Vol. 6 No. 4, pp. 233–240.
- Schrama, D., Scherer, D., Schneider, M., Zapatka, M., Bröcker, E.-B., Schadendorf, D., Ugurel, S., et al. (2011), "ERCC5 p. Asp1104His and ERCC2 p. Lys751Gln polymorphisms are independent prognostic factors for the clinical course of melanoma", *Journal of Investigative Dermatology*, Vol. 131 No. 6, pp. 1280–1290.
- Schramm, S.-J. and Mann, G.J. (2011), "Melanoma prognosis: a REMARK-based systematic review and bioinformatic analysis of immunohistochemical and gene microarray studies", *Molecular Cancer Therapeutics*, Vol. 10 No. 8, pp. 1520–1528.
- Schwarz, J.M., Cooper, D.N., Schuelke, M. and Seelow, D. (2014), "MutationTaster2: mutation prediction for the deep-sequencing age", *Nature Methods*, Vol. 11 No. 4, pp. 361–362.
- "Scitable | Learn Science at Nature". (n.d.). , available at: <https://www.nature.com/scitable> (accessed 27 February 2019).
- Scortegagna, M., Ruller, C., Feng, Y., Lazova, R., Kluger, H., Li, J.-L., De, S.K., et al. (2014), "Genetic inactivation or pharmacological inhibition of Pdk1 delays development and inhibits metastasis of BrafV600E:: Pten<sup>-/-</sup>-melanoma", *Oncogene*, Vol. 33 No. 34, pp. 4330–4339.
- Sengupta, S., Gulukota, K., Zhu, Y., Ober, C., Naughton, K., Wentworth-Sheilds, W. and Ji, Y. (2016), "Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples", *Nucleic Acids Research*, Vol. 44 No. 3, p. e25.

- Sengüven, B., Baris, E., Oygur, T. and Berktaş, M. (2014), "Comparison of methods for the extraction of DNA from formalin-fixed, paraffin-embedded archival tissues", *International Journal of Medical Sciences*, Vol. 11 No. 5, p. 494.
- Serrano, C. and Acha, B. (2009), "Pattern analysis of dermoscopic images based on Markov random fields", *Pattern Recognition*, Vol. 42 No. 6, pp. 1052–1057.
- Shain, A.H. and Bastian, B.C. (2016), "From melanocytes to melanomas", *Nature Reviews. Cancer*, Vol. 16 No. 6, pp. 345–358.
- Shain, A.H., Garrido, M., Botton, T., Talevich, E., Yeh, I., Sanborn, J.Z., Chung, J., et al. (2015), "Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway", *Nat Genet*, Vol. 47 No. 10, pp. 1194–9.
- Shain, A.H., Joseph, N.M., Yu, R., Benhamida, J., Liu, S., Prow, T., Ruben, B., et al. (2018), "Genomic and Transcriptomic Analysis Reveals Incremental Disruption of Key Signaling Pathways during Melanoma Evolution", *Cancer Cell*, Vol. 34 No. 1, pp. 45-55.e4.
- Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., et al. (2013), "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models", *Human Mutation*, Vol. 34 No. 1, pp. 57–65.
- Shiraishi, Y., Sato, Y., Chiba, K., Okuno, Y., Nagata, Y., Yoshida, K., Shiba, N., et al. (2013), "An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data", *Nucleic Acids Research*, Vol. 41 No. 7, p. e89.
- Simanshu, D.K., Nissley, D.V. and McCormick, F. (2017), "RAS Proteins and Their Regulators in Human Disease", *Cell*, Vol. 170 No. 1, pp. 17–33.
- Simes, R.J. (1985), "Treatment selection for cancer patients: Application of statistical decision theory to the treatment of advanced ovarian cancer", *Journal of Chronic Diseases*, Vol. 38 No. 2, pp. 171–186.

- Sims, D., Sudbery, I., Illott, N.E., Heger, A. and Ponting, C.P. (2014), "Sequencing depth and coverage: key considerations in genomic analyses", *Nature Reviews Genetics*, Vol. 15 No. 2, pp. 121–132.
- Sindi, S., Helman, E., Bashir, A. and Raphael, B.J. (2009), "A geometric approach for classification and comparison of structural variants", *Bioinformatics*, Vol. 25 No. 12, pp. i222–i230.
- Sindi, S.S., Önal, S., Peng, L.C., Wu, H.-T. and Raphael, B.J. (2012), "An integrative probabilistic model for identification of structural variation in sequencing data", *Genome Biology*, Vol. 13 No. 3, p. R22.
- Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2015), *ROCR: Visualizing the Performance of Scoring Classifiers*, available at: <https://CRAN.R-project.org/package=ROCR> (accessed 15 February 2019).
- Smith, A.P., Hoek, K. and Becker, D. (2005), "Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas", *Cancer Biology & Therapy*, Vol. 4 No. 9, pp. 1018–1029.
- Smith, K.S., Yadav, V.K., Pei, S., Pollyea, D.A., Jordan, C.T. and De, S. (2016), "SomVarIUS: somatic variant identification from unpaired tissue samples", *Bioinformatics (Oxford, England)*, Vol. 32 No. 6, pp. 808–813.
- de Snoo, F.A. and Hayward, N.K. (2005), "Cutaneous melanoma susceptibility and progression genes", *Cancer Letters*, Vol. 230 No. 2, pp. 153–186.
- "SOAP :: Short Oligonucleotide Analysis Package". (n.d.). , available at: <http://soap.genomics.org.cn/SOAPsnv.html> (accessed 20 September 2018).
- Spencer, D.H., Sehn, J.K., Abel, H.J., Watson, M.A., Pfeifer, J.D. and Duncavage, E.J. (2013), "Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens", *J Mol Diagn*, Vol. 15 No. 5, pp. 623–33.

- Spinella, J.-F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., Ouimet, M., et al. (2016), "SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing", *BMC Genomics*, Vol. 17 No. 1, p. 912.
- Stanley, R.J., Moss, R.H., Van Stoecker, W. and Aggarwal, C. (2003), "A fuzzy-based histogram analysis technique for skin lesion discrimination in dermatology clinical images", *Computerized Medical Imaging and Graphics*, Vol. 27 No. 5, pp. 387–396.
- Stark, M.S., Klein, K., Weide, B., Haydu, L.E., Pflugfelder, A., Tang, Y.H., Palmer, J.M., et al. (2015), "The prognostic and predictive value of melanoma-related microRNAs using tissue and serum: a microRNA expression analysis", *EBioMedicine*, Vol. 2 No. 7, pp. 671–680.
- Stark, M.S., Woods, S.L., Gartside, M.G., Bonazzi, V.F., Dutton-Regester, K., Aoude, L.G., Chow, D., et al. (2012), "Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing", *Nature Genetics*, Vol. 44 No. 2, pp. 165–169.
- Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., et al. (2015), "Big Data: Astronomical or Genomical?", *PLoS Biology*, Vol. 13 No. 7, available at:<https://doi.org/10.1371/journal.pbio.1002195>.
- Stolz, W. and Landthaler, M. (1994), "Classification, diagnosis and differential diagnosis of malignant melanoma", *Der Chirurg; Zeitschrift Fur Alle Gebiete Der Operativen Medizin*, Vol. 65 No. 3, pp. 145–152.
- Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009), "The cancer genome", *Nature*, Vol. 458 No. 7239, pp. 719–724.
- Strimbu, K. and Tavel, J.A. (2010), "What are Biomarkers?", *Current Opinion in HIV and AIDS*, Vol. 5 No. 6, pp. 463–466.
- Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P., Stark, M.S., Hayward, N.K., Martin, N.G., et al. (2008), "A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines

- human blue-brown eye color”, *The American Journal of Human Genetics*, Vol. 82 No. 2, pp. 424–431.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., et al. (2005), “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102 No. 43, pp. 15545–15550.
- Talantov, D., Mazumder, A., Yu, J.X., Briggs, T., Jiang, Y., Backus, J., Atkins, D., et al. (2005), “Novel genes associated with malignant melanoma but not benign melanocytic lesions”, *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, Vol. 11 No. 20, pp. 7234–7242.
- Talevich, E., Shain, A.H., Botton, T. and Bastian, B.C. (2016), “CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing”, *PLoS Comput Biol*, Vol. 12 No. 4, p. e1004873.
- Tan, H., Bao, J. and Zhou, X. (2012), “A novel missense-mutation-related feature extraction scheme for ‘driver’ mutation identification”, *Bioinformatics (Oxford, England)*, Vol. 28 No. 22, pp. 2948–2955.
- Tanaka, T., Yamada, R., Tanaka, M., Shimizu, K. and Oka, H. (2004), “A study on the image diagnosis of melanoma”, Vol. 1, presented at the Engineering in Medicine and Biology Society, 2004. IEMBS’04. 26th Annual International Conference of the IEEE, IEEE, pp. 1597–1600.
- Tang, X., Baheti, S., Shameer, K., Thompson, K.J., Wills, Q., Niu, N., Holcomb, I.N., et al. (2014), “The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data”, *Nucleic Acids Research*, Vol. 42 No. 22, p. e172.

Taouil, K., Romdhane, N.B. and Bouhlef, M.S. (2006), "A new automatic approach for edge detection of skin lesion images", Vol. 1, presented at the Information and Communication Technologies, 2006. ICTTA'06. 2nd, IEEE, pp. 212–220.

"The Cancer Genome Atlas". (n.d.) , available at: <https://cancergenome.nih.gov/>.

"The Skin Cancer Foundation". (n.d.) , available at: <http://www.skincancer.org>.

The Gene Ontology Consortium. (2019), "The Gene Ontology Resource: 20 years and still GOing strong", *Nucleic Acids Research*, Vol. 47 No. D1, pp. D330–D338.

Tomatis, S., Bono, A., Bartoli, C., Tragni, G., Farina, B. and Marchesini, R. (1998), "Image analysis in the RGB and HS colour planes for a computer-assisted diagnosis of cutaneous pigmented lesions", *Tumori*, Vol. 84 No. 1, pp. 29–32.

Umbaugh, S.E., Moss, R.H. and Stoecker, W.V. (1991), "Applying artificial intelligence to the identification of variegated coloring in skin tumors", *IEEE Engineering in Medicine and Biology Magazine*, Vol. 10 No. 4, pp. 57–62.

Umbaugh, S.E., Wei, Y.-S. and Zuke, M. (1997), "Feature extraction in image analysis. A program for facilitating data reduction in medical image classification", *IEEE Engineering in Medicine and Biology Magazine*, Vol. 16 No. 4, pp. 62–73.

de Unamuno Bustos, B., Murria Estal, R., Pérez Simó, G., de Juan Jimenez, I., Escutia Muñoz, B., Rodríguez Serna, M., Alegre de Miquel, V., et al. (2017), "Towards Personalized Medicine in Melanoma: Implementation of a Clinical Next-Generation Sequencing Panel", *Scientific Reports*, Vol. 7 No. 1, p. 495.

Uong, A. and Zon, L.I. (2010), "Melanocytes in development and cancer", *Journal of Cellular Physiology*, Vol. 222 No. 1, pp. 38–41.



- Usuyama, N., Shiraishi, Y., Sato, Y., Kume, H., Homma, Y., Ogawa, S., Miyano, S., et al. (2014), “HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations”, *Bioinformatics (Oxford, England)*, Vol. 30 No. 23, pp. 3302–3309.
- Valavanis, I., Maglogiannis, I. and Chatziioannou, A. (2015), “Exploring robust diagnostic signatures for cutaneous melanoma utilizing genetic and imaging data”, *IEEE Journal of Biomedical and Health Informatics*, pp. 190–198.
- Vallone, M.G., Tell-Marti, G., Potrony, M., Rebollo-Morell, A., Badenas, C., Puig-Butille, J.A., Gimenez-Xavier, P., et al. (2018), “Melanocortin 1 receptor (MC1R) polymorphisms’ influence on size and dermoscopic features of nevi”, *Pigment Cell & Melanoma Research*, Vol. 31 No. 1, pp. 39–50.
- Van Allen, E.M., Wagle, N., Stojanov, P., Perrin, D.L., Cibulskis, K., Marlow, S., Jane-Valbuena, J., et al. (2014), “Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine”, *Nat Med*, Vol. 20 No. 6, pp. 682–8.
- Vandin, F., Upfal, E. and Raphael, B.J. (2011), “Algorithms for detecting significantly mutated pathways in cancer”, *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, Vol. 18 No. 3, pp. 507–522.
- Venot, A., Devaux, J., Herbin, M., Lebruchec, J., Dubertret, L., Raulo, Y. and Roucayrol, J. (1988), “An automated system for the registration and comparison of photographic images in medicine”, *IEEE Transactions on Medical Imaging*, Vol. 7 No. 4, pp. 298–303.
- Vigil, D., Cherfils, J., Rossman, K.L. and Der, C.J. (2010), “Ras superfamily GEFs and GAPs: validated and tractable targets for cancer therapy?”, *Nature Reviews. Cancer*, Vol. 10 No. 12, pp. 842–857.
- Wachsman, W., Morhenn, V., Palmer, T., Walls, L., Hata, T., Zalla, J., Scheinberg, R., et al. (2011), “Noninvasive genomic detection of melanoma”, *British Journal of Dermatology*, Vol. 164 No. 4, pp. 797–806.

- Wachsman, W., Zapala, M., Udall, D., Paik, A., Hata, T., Walls, L., Wong, R., et al. (2007), "Differentiation of melanoma from dysplastic nevi in suspicious pigmented skin lesions by non-invasive tape stripping", *Training*, Vol. 100 No. 100, p. 100.
- Wajapeyee, N., Serra, R.W., Zhu, X., Mahalingam, M. and Green, M.R. (2008), "Oncogenic BRAF Induces Senescence and Apoptosis through Pathways Mediated by the Secreted Protein IGFBP7", *Cell*, Vol. 132 No. 3, pp. 363–374.
- Walia, V., Mu, E.W., Lin, J.C. and Samuels, Y. (2012), "Delving into somatic variation in sporadic melanoma", *Pigment Cell & Melanoma Research*, Vol. 25 No. 2, pp. 155–170.
- Wang, W., Wang, P., Xu, F., Luo, R., Wong, M.P., Lam, T.-W. and Wang, J. (2014), "FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data", *Bioinformatics (Oxford, England)*, Vol. 30 No. 17, pp. 2498–2500.
- Wang, Y., Marino-Enriquez, A., Bennett, R.R., Zhu, M., Shen, Y., Eilers, G., Lee, J.-C., et al. (2014), "Dystrophin is a tumor suppressor in human cancers with myogenic programs", *Nature Genetics*, Vol. 46 No. 6, pp. 601–606.
- Ward, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, Vol. 58 No. 301, p. 236.
- Weaver, O. and Leung, J.W.T. (2017), "Biomarkers and Imaging of Breast Cancer", *American Journal of Roentgenology*, Vol. 210 No. 2, pp. 271–278.
- Wei, X., Walia, V., Lin, J.C., Teer, J.K., Prickett, T.D., Gartner, J., Davis, S., et al. (2011), "Exome sequencing identifies GRIN2A as frequently mutated in melanoma", *Nature Genetics*, Vol. 43 No. 5, pp. 442–446.
- Wendl, M.C., Wallis, J.W., Lin, L., Kandoth, C., Mardis, E.R., Wilson, R.K. and Ding, L. (2011), "PathScan: a tool for discerning mutational significance in groups of putative cancer genes", *Bioinformatics*, Vol. 27 No. 12, pp. 1595–1602.

“WHO | WHO definitions of genetics and genomics”. (n.d.). WHO, available at:

<http://www.who.int/genomics/geneticsVSgenomics/en/> (accessed 10 April 2019).

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., et al. (2012), “LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets”, *Nucleic Acids Research*, Vol. 40 No. 22, pp. 11189–11201.

Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S.R., Avril, M.F., et al. (2006), “Gene expression profiling of primary cutaneous melanoma and clinical outcome”, *J Natl Cancer Inst*, Vol. 98 No. 7, pp. 472–82.

Wong, R., Tran, V., Talwalker, S. and Benson, N.R. (2006), “Analysis of RNA recovery and gene expression in the epidermis using non-invasive tape stripping”, *Journal of Dermatological Science*, Vol. 44 No. 2, pp. 81–92.

Wong, W.C., Kim, D., Carter, H., Diekhans, M., Ryan, M.C. and Karchin, R. (2011), “CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer”, *Bioinformatics (Oxford, England)*, Vol. 27 No. 15, pp. 2147–2148.

Wu, P.-Y., Cheng, C.-W., Kaddi, C.D., Venugopalan, J., Hoffman, R. and Wang, M.D. (2017), “-Omic and Electronic Health Records Big Data Analytics for Precision Medicine”, *IEEE Transactions on Bio-Medical Engineering*, Vol. 64 No. 2, p. 263.

Xi, R., Luquette, J., Hadjipanayis, A., Kim, T.-M. and Park, P.J. (2010), “BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data”, *Genome Biology*, Vol. 11 No. Suppl 1, p. O10.

Xu, C. (2018), “A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data”, *Computational and Structural Biotechnology Journal*, Vol. 16, pp. 15–24.

- Xu, H., DiCarlo, J., Satya, R.V., Peng, Q. and Wang, Y. (2014), "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data", *BMC Genomics*, Vol. 15, p. 244.
- Xu, Z., Zhou, Y., Cao, Y., Dinh, T.L., Wan, J. and Zhao, M. (2016), "Identification of candidate biomarkers and analysis of prognostic values in ovarian cancer by integrated bioinformatics analysis", *Med Oncol*, Vol. 33 No. 11, p. 130.
- Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.-H., Zhang, C., et al. (2013), "Diverse mechanisms of somatic structural variations in human cancer genomes", *Cell*, Vol. 153 No. 4, pp. 919–929.
- Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., et al. (2008), "Heterogeneous data fusion for alzheimer's disease study", presented at the Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1025–1033.
- Ye, J., Pavlicek, A., Lunney, E.A., Rejto, P.A. and Teng, C.-H. (2010), "Statistical method on nonrandom clustering with application to somatic mutations in cancer", *BMC Bioinformatics*, Vol. 11, p. 11.
- Youn, A. and Simon, R. (2011), "Identifying cancer driver genes in tumor genome sequencing studies", *Bioinformatics*, Vol. 27 No. 2, pp. 175–181.
- Yuan, X., Yang, Z., Zouridakis, G. and Mullani, N. (2006), "SVM-based texture classification and application to early melanoma detection", presented at the Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, IEEE, pp. 4775–4778.
- Yue, P., Melamud, E. and Moulton, J. (2006), "SNPs3D: candidate gene and SNP selection for association studies", *BMC Bioinformatics*, Vol. 7, p. 166.
- Zalaudek, I., Argenziano, G., Mordente, I., Moscarella, E., Corona, R., Sera, F., Blum, A., et al. (2007), "Nevus type in dermoscopy is related to skin type in white persons", *Archives of Dermatology*, Vol. 143 No. 3, pp. 351–356.

- Zhang, J., Liu, J., Sun, J., Chen, C., Foltz, G. and Lin, B. (2014), "Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing", *Briefings in Bioinformatics*, Vol. 15 No. 2, pp. 244–255.
- Zhang, P., Lehmann, B.D., Shyr, Y. and Guo, Y. (2017), "The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies", *International Journal of Genomics*, Vol. 2017.
- Zhang, Q., Ding, L., Larson, D.E., Koboldt, D.C., McLellan, M.D., Chen, K., Shi, X., et al. (2010), "CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data", *Bioinformatics*, Vol. 26 No. 4, pp. 464–9.
- Zhang, T., Dutton-Regester, K., Brown, K.M. and Hayward, N.K. (2016), "The genomic landscape of cutaneous melanoma", *Pigment Cell & Melanoma Research*.
- Zhang, Z., Moss, R.H. and Stoecker, W.V. (2003), "Neural networks skin tumor diagnostic system", Vol. 1, presented at the Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on, IEEE, pp. 191–192.
- Zhao, R., Choi, B.Y., Lee, M.-H., Bode, A.M. and Dong, Z. (2016), "Implications of Genetic and Epigenetic Alterations of CDKN2A (p16(INK4a)) in Cancer", *EBioMedicine*, Vol. 8, pp. 30–39.
- Zhavoronkov, A. (2018), "Artificial Intelligence for Drug Discovery, Biomarker Development, and Generation of Novel Chemistry", *Molecular Pharmaceutics*, Vol. 15 No. 10, pp. 4311–4313.
- Zhou, H., Li, X., Schaefer, G., Celebi, M.E. and Miller, P. (2013), "Mean shift based gradient vector flow for image segmentation", *Computer Vision and Image Understanding*, Vol. 117 No. 9, pp. 1004–1016.
- Zhou, X.-P., Gimm, O., Hampel, H., Niemann, T., Walker, M.J. and Eng, C. (2000), "Epigenetic PTEN Silencing in Malignant Melanomas without PTEN Mutation", *The American Journal of Pathology*, Vol. 157 No. 4, pp. 1123–1128.

Zingg, D., Arenas-Ramirez, N., Sahin, D., Rosalia, R.A., Antunes, A.T., Haeusel, J., Sommer, L., et al.

(2017), “The Histone Methyltransferase Ezh2 Controls Mechanisms of Adaptive Resistance to Tumor Immunotherapy”, *Cell Reports*, Vol. 20 No. 4, pp. 854–867.

Zinn, P.O., Majadan, B., Sathyan, P., Singh, S.K., Majumder, S., Jolesz, F.A. and Colen, R.R. (2011),

“Radiogenomic Mapping of Edema/Cellular Invasion MRI-Phenotypes in Glioblastoma Multiforme”, *PLOS ONE*, Vol. 6 No. 10, p. e25451.

Zouridakis, G., Doshi, M. and Mullani, N. (2004), “Early diagnosis of skin cancer based on segmentation

and measurement of vascularization and pigmentation in nevoscope images”, Vol. 1, presented

at the Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International

Conference of the IEEE, IEEE, pp. 1593–1596.

# Conferences & Publications

## Presentations & Posters on International Conferences

- 45th Annual Hellenic Medical Conference, Athens, 15-18/05/2019, Presentation: Karanikas, H., Maglogiannis, I., Papadodima, O., **Kontogianni, G.**, Piroti, G., Billiris, A., Zaras, A., Kypreou, K., Stratigos, A., Chatziioannou, A., 'Translating the diagnostic complexity of melanoma into rational therapeutic stratification - TRANSITION'
- 17th European Conference on Computational Biology (ECCB 2018), Athens, 8-12/09/2018, Poster: **Kontogianni, G.**, Papadodima, O., Liampa, I., de Lastic, H.X., Maglogiannis, I., Chatziioannou, A., 'Development of a classification system for the translational analysis of cancer genomic and imaging data for melanoma prognosis'
- 68th Congress of the Hellenic Society of Biochemistry and Molecular Biology (HSBMB 2017), Athens, 10-12/11/2017, Presentation: **Kontogianni, G.**, Piroti, G., Maglogiannis, I., Chatziioannou, A. & Papadodima, O., 'Characterisation of somatic mutations in Greek patients with cutaneous melanoma'
- 12th Conference of the Hellenic Society for Computational Biology and Bioinformatics (HSCBB17), Athens, 11-13/10/2017, Poster: **Kontogianni, G.**, Piroti, G., Maglogiannis, I., Papadodima, O. & Chatziioannou, A., 'Comprehensive analysis of somatic mutations in Greek patients with cutaneous melanoma using multiple bioinformatics tools'
- 10th Swedish-Hellenic Life Science Research Conference, Athens, 9-10 October 2017, Presentation: **Kontogianni, G.**, Pyroti, G., Maglogiannis, I., Chatziioannou, A. & Papadodima, O., 'Characterisation of somatic mutations in Greek patients with cutaneous melanoma'
- Workshop organised in the frame of the cooperation of the German Cancer Research Centre (DKFZ) with the Athens Cancer Comprehensive Centre (ACCC) at NHRF, June 2017, Athens, Greece, Poster: **Kontogianni, G.**, Pyroti, G., Maglogiannis, I.,

Chatziioannou, A. & Papadodima, O., 'Characterisation of somatic mutations in Greek patients with cutaneous melanoma'

- 12th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI 2016), Thessaloniki, Greece, September 16th-18th 2016, Presentation: **Kontogianni, G.**, Papadodima, O., Maglogiannis, I., Frangia-Tsivou, K., & Chatziioannou, A., 'Integrative bioinformatic analysis of a Greek epidemiological cohort provides insight into the pathogenesis of primary cutaneous melanoma'

- XIV Mediterranean Conference on Medical and Biological Engineering and Computing, Paphos, Cyprus, March 31st- April 2nd 2016, Presentation: **Kontogianni, G.**, Papadodima, O., Mitrakas, A., Maglogiannis, I., Koukourakis, M. I., Giatromanolaki, A., & Chatziioannou, A., 'Exploring the Molecular Determinants of Tumor-Stroma Interaction in Non-small Cell Lung Cancer Through the Utilization of RNA-seq Data from Lung Biopsies'

- 16th International Conference on Engineering Applications of Neural Networks (EANN 2015), Rhodes, 25-28/09/2015, Presentation: Maglogiannis, I., Goudas, T., Billiris, A., Karanikas, H., Valavanis, I., Papadodima, O., **Kontogianni, G.**, Chatziioannou, A., 'Redesigning EHRs and Clinical Decision Support Systems for the Precision Medicine Era'

- 65th Congress of The Hellenic Society of Biochemistry and Molecular Biology, Thessaloniki, Greece, November 2014, Poster: **Kontogianni, G.**, Papadodima, O., Maglogiannis, I. & Chatziioannou, A., 'Development of a pipeline for the translational analysis of melanoma genomic variation'

- 5th Melanoma Congress, Nafplio, Greece, June 2014, Poster: Papadodima, O., **Kontogianni, G.**, Maglogiannis, I. & Chatziioannou, A., 'Development of a pipeline for the translational analysis of melanoma genomic variation'



## Publications on Conference Papers

**Kontogianni, G.**, Papadodima, O., Maglogiannis, I., Frangia-Tsivou, K., & Chatziioannou, A. (2016, September). Integrative Bioinformatic Analysis of a Greek Epidemiological Cohort Provides Insight into the Pathogenesis of Primary Cutaneous Melanoma. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 39-52). Springer International Publishing. [https://doi.org/10.1007/978-3-319-44944-9\\_4](https://doi.org/10.1007/978-3-319-44944-9_4)

Maglogiannis, I., Goudas, T., Billiris, A., Karanikas, H., Valavanis, I., Papadodima, O., **Kontogianni, G.**, Chatziioannou, A. (2015, September). Redesigning EHRs and Clinical Decision Support Systems for the Precision Medicine Era. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) (p. 14). ACM. <https://doi.org/10.1145/2797143.2797158>

## Publications on International Journals

Maglogiannis, I., **Kontogianni, G.**, Karanikas, H., Billiris, A., Papadodima, O. & Chatziioannou, A., An integrated platform for Melanoma related healthcare data Management (submitted in Journal of Medical Systems, IF: 2.098, on April 2019)

Papadodima, O., **Kontogianni, G.**, Piroti, G., Maglogiannis, I. & Chatziioannou, A., Genomics of Cutaneous Melanoma: Focus on Next-Generation Sequencing Approaches and Bioinformatics. J Transl Genet Genom 2019;3:7. [10.20517/jtgg.2018.33](https://doi.org/10.20517/jtgg.2018.33)

**Kontogianni, G.**, Piroti, G., Maglogiannis, I., Chatziioannou, A., & Papadodima, O. (2018). Dissecting the Mutational Landscape of Cutaneous Melanoma: An Omic Analysis Based on Patients from Greece. Cancers, 10(4), 96. IF: 6.162 [10.3390/cancers10040096](https://doi.org/10.3390/cancers10040096)



# Appendix

Tables A1: Patient and melanoma lesion characteristics (extended)

Patient	REGISTRY CODE AND DATE	AGE	SEX	ETHNIC ORIGIN	PRE-X NEVUS	ASYMMETRY	BORDER	SIZE INCREASED	DIAMETER
3	8521/15	52	♀	Caucasian	yes	yes	Irregular	yes	>5mm
5	4549/14	82	♂	Caucasian	yes	no	Regular	yes	>5mm
8	5078/14	80	♂	Caucasian	no	yes	Irregular	yes	>5mm
10	8152/14	77	♀	Caucasian	no	yes	Irregular	yes	>5mm
11	12116/14	72	♂	Caucasian	yes	yes	Irregular	yes	>5mm
12	10719/14	56	♀	Caucasian	no	yes	Irregular	yes	>5mm
13	13762/15	69	♀	Caucasian	no	yes	Regular	yes	>5mm
14	4764/11	73	♀	Caucasian	yes	yes	Irregular	yes	>5mm
15	2395/12	38	♂	Caucasian	no	no	Regular	yes	>5mm

Patient	COLOR CHANGE	SITE	HISTOLOGY	CLARK	ULCERATION	GROWTH PHASE VERTICAL	GROWTH PHASE RADIAL	HISTO-GENESIS
3	yes	waste	melanoma & dysplastic nevus	IV	no	yes	yes	superficial spreading melanoma
5	yes	back	melanoma & nevus	IV	yes	yes	no	nodular melanoma
8	yes	head	melanoma	III	no	no	invasive	lentigo maligna melanoma
10	yes	subungual (foot)	melanoma	NA	no	no	invasive	acral melanoma
11	yes	back	melanoma & dysplastic nevus	I	no	no	in situ	superficial spreading melanoma
12	yes	head	melanoma	I	no	no	in situ	lentigo maligna melanoma
13	no	back	melanoma	V	no	yes	no	other

14	NA	tibia	melanoma & congenital nevus	IV	>3mm deep	yes	yes	superficial spreading melanoma
15	no	abdomen	melanoma metastatic	NA	no	yes	NA	metastatic melanoma

Patient	BRESLOW	MITOSES	LYMPH REACTION	CELL TYPE	NEURO-TROPISM	MICRO-SATELLITE	VESSEL INVASION	SATELLITES	UV EXPOSURE
3	1-2	>6	non brisk	epithelioid	no	no	no	no	intermittent
5	>4	>6	non brisk	epithelioid	no	no	yes	no	NA
8	<1	absent	non brisk	epithelioid	no	no	no	no	intermittent
10	<1	absent	non brisk	epithelioid	no	no	no	no	NA
11	NA	NA	non brisk	mixed	NA	NA	NA	NA	intermittent
12	NA	NA	non brisk	mixed	NA	NA	NA	NA	intermittent
13	>4	<6	non brisk	spindle	yes	no	no	no	intermittent
14	2-4	>6	non brisk	epithelioid	no	yes	yes	yes	NA
15	NA	>6	brisk	mixed	no	NA	yes	NA	intermittent

## Tables A2: Quality control for whole-exome sequencing of 18 samples

Samples	Raw reads	Raw bases (Mb)	Clean reads	Clean bases (Mb)	Clean data rate (%)	Clean read1 Q20 (%)	Clean read2 Q20 (%)	Clean read1 Q30 (%)	Clean read2 Q30 (%)	GC content (%)
10N	81,908,911	12284.46	77,892,314	10721.15	87.27	98.2	95.67	94.88	89.52	47.37
10T	87,850,010	13177.5	82,700,876	11410.9	86.59	97.58	94.8	93.59	87.85	47.45
11N	77,075,954	11560.04	73,717,912	10377.04	89.77	98.24	95.47	94.96	89.11	46.49
11T	81,648,488	12245.6	78,585,616	11010.32	89.91	98.28	95.62	95.01	89.36	47.67
12N	95,155,256	14271.78	89,782,142	12649.44	88.63	98.22	95.33	94.88	88.82	47.88
12T	76,428,719	11462.92	73,704,040	10375.49	90.51	98.27	95.5	94.95	89.07	48.26
13N	89,284,608	13391.6	85,843,526	12326.39	92.05	98.28	94.9	95.02	87.99	47.98
13T	80,205,589	12029.76	77,464,946	11150.37	92.69	98.28	94.67	94.99	87.47	48.75
14N	98,722,715	14808.4	92,258,398	12592.06	85.03	97.56	92.91	93.72	84.97	48.15
14T	85,691,122	12853.64	81,740,696	11192.18	87.07	97.72	93.78	94.03	86.25	47.34
15N	95,142,746	14271.41	91,161,502	12745.86	89.31	98.22	95.2	94.77	88.55	48.42
15T	94,552,990	14182.95	90,329,490	12509.01	88.2	98.12	95	94.55	88.12	51.22
3N	106,854,782	16028.22	99,723,498	14174.59	88.44	97.63	93.85	93.74	85.96	46.59
3T	92,450,892	13867.63	85,702,424	12163.33	87.71	97.66	93.83	93.73	85.9	47.92
5N	113,855,594	17078.27	104,767,924	14210.26	83.21	96.71	92.25	92.24	83.94	46.84
5T	87,318,334	13097.73	82,535,914	10854.94	82.88	97.53	94.52	93.67	87.53	49.16
8N	265,848,869	39876.8	240,421,486	32397.09	81.24	96.09	90.47	91.27	81.17	46.9
8T	79,430,894	11914.63	74,667,412	10308.63	86.52	97.63	94.77	93.69	87.76	47.39
Average	99,412,581	14911.3	93,500,006	12953.84	87.61	97.79	94.36	94.09	87.19	47.88

The table includes the numbers for raw reads/ bases, clean reads/ bases, the fraction of clean reads in raw reads, the fraction of sequencing bases with quality score  $\geq 20$  and 30 (Q20/Q30) in clean read1 and 2, and GC content

ID	Initial bases on target	Total effective reads	Total effective bases (Mb)	Effective sequences on target (Mb)	Capture specificity (%)	Mapping rate (%)	Duplicate rate (%)	Mismatch rate in target region (%)	target >= 10x (%)	target >= 20x (%)
10N	50,390,601	67,869,977	8,947.49	5,237.52	58.54	99.98	14.73	0.38	98.78	96.44
10T	50,390,601	74,315,959	9,723.28	5,620.69	57.81	99.79	12.31	0.43	98.84	96.75
11N	50,390,601	65,733,769	9,091.66	5,113.14	56.24	99.97	11.5	0.39	98.24	95.03
11T	50,390,601	70,620,657	9,709.05	5,613.45	57.82	99.98	10.91	0.38	98.89	96.74
12N	50,390,601	77,499,635	10,713.00	6,214.33	58.01	99.97	14.37	0.4	99.07	97.53
12T	50,390,601	66,311,938	9,123.91	5,280.29	57.87	99.99	11.03	0.39	98.87	96.62
13N	50,390,601	76,783,384	10,881.64	5,917.48	54.38	99.71	10.76	0.41	98.94	97
13T	50,390,601	70,205,023	9,911.95	5,637.03	56.87	99.99	10.17	0.42	98.8	96.16
14N	50,390,601	76,729,921	9,481.72	5,164.47	54.47	99.69	20.54	0.5	99.12	97.38
14T	50,390,601	72,748,635	9,141.41	5,152.08	56.36	99.89	14.91	0.45	98.86	96.3
15N	50,390,601	80,768,568	10,976.32	6,468.64	58.93	99.94	12.58	0.4	99.36	97.96
15T	50,390,601	80,533,837	10,521.19	6,084.10	57.83	99.97	13.57	0.42	99.14	96.66
3N	50,390,601	87,124,794	12,052.24	6,541.83	54.28	99.97	13.71	0.46	98.95	97.37
3T	50,390,601	76,161,797	10,498.24	5,993.05	57.09	99.98	12.35	0.47	99.11	97.4
5N	50,390,601	82,886,127	9,743.16	4,733.55	48.58	99.54	25.38	0.54	98.9	96.37
5T	50,390,601	71,306,354	8,755.80	5,079.65	58.01	99.94	16.44	0.44	98.89	95.86
8N	50,390,601	162,496,497	14,782.48	3,637.26	24.61	99.13	40.89	0.87	98.78	94.41
8T	50,390,601	66,920,301	8,843.59	5,151.36	58.25	99.99	12.35	0.44	98.93	96.48
Average	50,390,601	79,278,731	10,161.01	5,480.00	54.78	99.86	15.47	0.46	98.92	96.58

The table includes the length of target regions, the number of effective reads (mapped, nonduplicate reads), the number of bases in total effective reads, the number of effective bases located on target regions, the fraction of effective bases on target regions (capture specificity), the mapping rate, the duplicate rate, the percentage of mismatch bases in effective bases on targets, and the percentage of targeted bases that were covered by at least ten/twenty reads

ID	read1 Q20 (%)	read2 Q20 (%)	read1 Q30 (%)	read2 Q30 (%)	GC (%)	Mapping rate (%)	Duplicate rate (%)	Capture specificity (%)	Mismatch rate in target region (%)	Average sequencing depth on target	Fraction of target covered $\geq 1x$ (%)	Fraction of target covered $\geq 4x$ (%)
10N	98.2	95.67	94.88	89.52	47.37	99.98	14.73	58.54	0.38	103.94	99.79	99.56
10T	97.58	94.8	93.59	87.85	47.45	99.79	12.31	57.81	0.43	111.54	99.81	99.58
11N	98.24	95.47	94.96	89.11	46.49	99.97	11.5	56.24	0.39	101.47	99.81	99.44
11T	98.28	95.62	95.01	89.36	47.67	99.98	10.91	57.82	0.38	111.4	99.89	99.65
12N	98.22	95.33	94.88	88.82	47.88	99.97	14.37	58.01	0.4	123.32	99.88	99.66
12T	98.27	95.5	94.95	89.07	48.26	99.99	11.03	57.87	0.39	104.79	99.8	99.6
13N	98.28	94.9	95.02	87.99	47.98	99.71	10.76	54.38	0.41	117.43	99.81	99.6
13T	98.28	94.67	94.99	87.47	48.75	99.99	10.17	56.87	0.42	111.87	99.8	99.58
14N	97.56	92.91	93.72	84.97	48.15	99.69	20.54	54.47	0.5	102.49	99.83	99.67
14T	97.72	93.78	94.03	86.25	47.34	99.89	14.91	56.36	0.45	102.24	99.82	99.63
15N	98.22	95.2	94.77	88.55	48.42	99.94	12.58	58.93	0.4	128.37	99.93	99.81
15T	98.12	95	94.55	88.12	51.22	99.97	13.57	57.83	0.42	120.74	99.92	99.78
3N	97.63	93.85	93.74	85.96	46.59	99.97	13.71	54.28	0.46	129.82	99.82	99.59
3T	97.66	93.83	93.73	85.9	47.92	99.98	12.35	57.09	0.47	118.93	99.83	99.65
5N	96.71	92.25	92.24	83.94	46.84	99.54	25.38	48.58	0.54	93.94	99.9	99.68
5T	97.53	94.52	93.67	87.53	49.16	99.94	16.44	58.01	0.44	100.81	99.89	99.71
8N	96.09	90.47	91.27	81.17	46.9	99.13	40.89	24.61	0.87	72.18	99.94	99.76
8T	97.63	94.77	93.69	87.76	47.39	99.99	12.35	58.25	0.44	102.23	99.91	99.71

The table includes the percentage of sequencing bases with quality score  $\geq 20$  and  $30$  (Q20/Q30) in clean read1 and 2, GC content, the percentage of mapped reads in total clean reads, the percentage of duplicate reads, the percentage of mismatch bases in effective bases on targets, the average sequencing coverage on target regions (calculated as Effective bases on targets divided by Initial bases on targets), and the percentage of targeted bases that were covered by at least one/four reads

Table A3: Binary tables of patients carrying mutations for MAPK pathway and cell cycle

MAPK pathway									
	patient3	patient5	patient8	patient10	patient11	patient12	patient13	patient14	patient15
ANGPT1	0	0	1	0	0	0	0	0	0
BRAF	1	1	0	0	0	0	0	1	0
CUL1	1	0	0	0	0	0	0	0	0
CUL3	0	0	0	0	1	1	0	0	0
GPS1	0	1	0	0	0	0	0	0	0
GRIN2A	0	1	0	0	0	1	0	1	0
GRIN2B	0	0	1	0	1	1	0	0	0
ITGA2B	0	0	0	0	1	1	0	0	0
JAK1	0	1	0	0	0	0	0	0	0
KALRN	0	0	0	0	0	0	1	0	1
KL	1	0	0	0	1	0	0	0	0
KSR1	1	0	0	0	0	0	0	0	0
KSR2	0	0	1	0	0	1	0	0	0
LRRK2	0	0	1	0	0	0	0	0	0
MAP2K5	0	0	0	0	0	0	0	0	1
MAP3K1	0	0	1	0	0	0	0	0	0
MAP3K5	1	0	0	0	0	0	0	0	0
MAPK8IP3	0	0	0	0	0	0	0	1	0
PTK2B	0	0	0	0	0	0	0	1	0
RAG1	0	1	0	0	0	1	0	0	0
RASA3	0	0	0	0	0	0	0	0	1
SPTB	0	0	0	0	0	1	0	0	1
SPTBN1	0	0	1	0	0	0	0	0	0
TNRC6C	0	0	0	0	0	1	0	0	1



Cell Cycle									
	patient3	patient5	patient8	patient10	patient11	patient12	patient13	patient14	patient15
ANAPC5	1	0	0	0	0	0	0	0	0
ARF6	0	0	1	0	0	0	0	0	0
ATM	0	0	0	0	1	0	0	0	0
CDC45	0	0	0	0	1	0	0	0	0
CDKN1B	0	0	0	0	0	0	1	0	0
CUL1	1	0	0	0	0	0	0	0	0
CUL3	0	0	0	0	1	1	0	0	0
DNA2	0	0	0	0	1	0	0	0	0
DYNC1H1	0	0	1	0	0	1	0	0	0
FBXO5	1	0	0	0	0	0	0	0	0
GPS1	0	1	0	0	0	0	0	0	0
HIST1H2BN	0	0	0	0	0	0	1	0	0
NASP	0	0	1	0	0	0	0	0	0
NBN	0	0	0	0	0	0	0	1	0
NCAPG2	0	0	0	0	1	1	0	0	0
NEK6	0	0	0	0	0	0	1	0	0
NEK9	0	0	0	0	1	0	0	0	0
NUP153	0	1	0	0	0	0	0	0	0
NUP155	0	1	0	0	0	0	0	0	0
PARD3	0	0	1	0	0	0	0	0	0
POLD3	0	0	0	0	1	0	0	0	0
RB1CC1	0	0	0	0	0	0	0	1	1
RFC3	0	0	0	0	0	0	0	0	1
SKA1	0	0	0	0	0	0	1	0	0
SPECC1L	0	0	1	0	0	0	0	0	0
STAG2	0	0	0	0	1	0	0	0	0
TDRD1	1	0	0	0	0	1	0	0	0
TERT	0	1	0	0	0	0	0	0	0
TET2	1	0	0	0	0	0	0	0	0
TOPBP1	0	0	1	0	0	1	0	0	0
TP53	0	1	0	0	0	0	1	0	0
TPR	0	1	0	0	0	0	0	0	0
TPX2	0	0	0	0	0	0	0	1	0
TXLNG	0	0	0	0	0	0	0	1	0
UTP14C	1	0	0	0	0	1	0	0	0