

Jagiellonian University

Faculty of Physics, Astronomy and Applied Computer Science

Department of Medical Physics

Sonia Wróbel

Student ID: 1065409

**Canonical correlation analysis of multi-omics data obtained from
microvesicles of human skin melanoma cell lines and its possible
application in biomarker discovery**

Thesis submitted for the degree of PhD in biophysics

Supervisors:

Prof. Ewa Łucja Stępień, PhD

Monika Piwowar, PhD with habilitation

Cracow, 2021

Oświadczenie

Ja niżej podpisana Sonia Wróbel (nr indeksu: 1065409) doktorantka Wydziału Fizyki, Astronomii i Informatyki Stosowanej Uniwersytetu Jagiellońskiego oświadczam, że przedłożona przeze mnie rozprawa doktorska pt. „Canonical correlation analysis of multi-omics data obtained from microvesicles of human skin melanoma cell lines and its possible application in biomarker discovery” jest oryginalna i przedstawia wyniki badań wykonanych przeze mnie osobiście, pod kierunkiem prof. dr hab. Ewy Stępień oraz dr hab. Moniki Piwowar. Pracę napisałam samodzielnie.

Oświadczam, że moja rozprawa doktorska została opracowana zgodnie z Ustawą o prawie autorskim i prawach pokrewnych z dnia 4 lutego 1994 r. (Dziennik Ustaw 1994 nr 24 poz. 83 wraz z późniejszymi zmianami).

Jestem świadoma, że niezgodność niniejszego oświadczenia z prawdą ujawniona w dowolnym czasie, niezależnie od skutków prawnych wynikających z ww. ustawy, może spowodować unieważnienie stopnia nabytego na podstawie tej rozprawy.

.....

Kraków, dnia

.....

Podpis autora pracy

Podziękowania

Pragnę podziękować promotorkom mojej pracy doktorskiej: prof. dr hab. Ewie Stępień oraz dr hab. Monice Piwovar, za opiekę naukową, niezliczone godziny konsultacji oraz niewyczerpaną cierpliwość i wyrozumiałość.

Dziękuję prof. dr hab. Małgorzacie Przybyło za umożliwienie wykonania hodowli komórkowych w Instytucie Zoologii i Badań Biomedycznych, Uniwersytetu Jagiellońskiego oraz za wsparcie merytoryczne.

Dziękuję dr Magdalenie Surman, dr Agnieszce Kamińskiej oraz mgr inż. Annie Drożdż za nieocenioną pomoc w realizacji części eksperymentalnej mojej pracy, za życzliwość oraz ciekawe rozmowy o nauce i nie tylko.

Chciałabym również podziękować dr. hab. Jakubowi Mielczarkowi, dr Katarzynie Dziedzic-Kocurek, dr. Piotrowi Warchołowi oraz dr. Marcinowi Krzykawskiemu za stworzenie i prowadzenie akademickiego laboratorium "Garaż Złożoności". To właśnie dzięki Wam i "Garażowi Złożoności" uwierzyłam w siebie, pokochałam naukę i odkryłam czym jest bioinformatyka. Zawsze z dużym wzruszeniem będę wspominać nasze "garażowe" seminaria, warsztaty i konferencje.

Dziękuję Michałowi Kańskiemu za korektę językową oraz bardzo cenne uwagi merytoryczne.

Pracę dedykuję mojej Mamie i mojemu Tacie. Dziękuję Wam za wsparcie w realizacji wszystkich moich celów.

Acknowledgements

This work was supported by:

- Polish Ministry of Science and Higher Education Grant No.: 7150/E-338/M/2018
to Sonia Wróbel
- Grant No.: N18/DBS/000007 to Prof. Małgorzata Przybyło
- National Science Center of Poland Grant OPUS 17 No.: 2019/33/B/NZ3/01004
to Prof. Ewa Stępień
- National Science Center of Poland Grant OPUS 4 No.: 2012/07B/NZ5/02510
to Prof. Ewa Stępień
- National Science Center of Poland Grant OPUS 6 No.: 2013/11/B/NZ4/04315
to Prof. Małgorzata Przybyło
- National Science Center of Poland Grant PRELUDIUM No.: 018/31/N/NZ4/03787
to Dr Magdalena Surman
- National Centre for Research and Development Grant LIDER No. LIDER/9/0031/L-
9/NCBR/2018 to Dr Agnieszka Kamińska

List of Author's full-papers

- Cezary Turek, **Sonia Wróbel** and Monika Piwowar, "OmicsON – Integration of omics data with molecular networks and statistical procedures", *PLOS ONE*, vol. 15, p. e0235398, July 2020. IF= 2.740, MNiSW= 100
- Jan Majta, Krzysztof Odrzywołek, Bożena Milanovic, Vladyslav Hubar, **Sonia Wróbel**, Emilia Strycharz-Angrecka, Szymon Wojciechowski and Kaja Milanowska, "Identification of Differentiating Metabolic Pathways between Infant Gut Microbiome Populations Reveals Depletion of Function-Level Adaptation to Human Milk in the Finnish Population", *mSphere*, vol. 4, Mar. 2019. IF= 3.680, MNiSW= 100
- **Sonia Wróbel**, Małgorzata Przybyło, Ewa Stępień, "The clinical trial landscape for melanoma therapies", *Journal of Clinical Medicine*, vol. 8, p. 368, Mar. 2019. IF= 4.241, MNiSW= 140
- Piotr Ciochoń, Natalia Olszowska, **Sonia Wróbel** and Jacek Kołodziej, "Reconstructions of the sulfur-passivated InSb (100) surface", *Applied Surface Science*, vol. 400, pp. 154-161, Apr. 2017. IF= 6.707, MNiSW= 140
- Jakub Mielczarek, Grzegorz Gazdowicz, Jakub Kramarz, Piotr Łątka, Marcin Krzykowski, Artur Miroszewski, Paweł Pieczarko, Renata Szczelina, Piotr Warchoń and **Sonia Wróbel**, "A Prototype of a 3D Bioprinter", *Solid State Phenomena*, vol. 237, pp. 221-226, Aug. 2015, IF= 0.660

Conference contributions

- Conference on Cybernetic Modeling of Biological Systems (MCSB) 2021, **Sonia Wróbel**, Cezary Turek, Ewa Stępień, Monika Piwowar, "An analysis of molecular information transfer in Melanoma Metastasis Formation by combining Bioinformatics and Systems Biology", Kraków 2021 (**oral presentation**)
- Autumn Workshop Polish Bioinformatics Society (PBIS), **Sonia Wróbel**, Ewa Stępień, Cezary Turek, Monika Piwowar, "Canonical Correlation - based bioinformatic analysis for effective melanoma biomarker discovery ", Warszawa 2020 (**poster presentation**)
- XII Interdyscyplinarna Konferencja Naukowa TYGIEL 2020, Cezary Turek, **Sonia Wróbel**, Monika Piwowar, "Relacja danych typu "omics" z wykorzystaniem autorskiej biblioteki do integracji danych", 2020 (**chapter in monograph** "Wybrane zagadnienia z zakresu onkologii i zdrowia człowieka", pages: 326-343)
- Microbiome *R&D* and Business Collaboration Forum, Jan Majta, Krzysztof Odrzywołek, Szymon Wojciechowski, Vladyslav Hubar, **Sonia Wróbel**, Kaja Milanowska, "Gut feeling: function based microbiome research", San Diego 2018 (**poster presentation**)
- Microbiome Congress, Jan Majta, Krzysztof Odrzywołek, Szymon Wojciechowski, Vladyslav Hubar, **Sonia Wróbel**, Kaja Milanowska, "Gut feeling: function based microbiome research", London 2018 (**poster co-author**)
- Translational Microbiome Conference, Jan Majta, Łukasz Maziarka, Szymon Wojciechowski, **Sonia Wróbel**, Piotr Stępnia, Kaja Milanowska, "Empowering microbiome research with Artificial Intelligence", Boston 2018 (**poster co-author**)

Abstract

This PhD Thesis demonstrates a new canonical correlation-based method that uses multi-omics data to address the challenge of melanoma biomarker discovery. Proposed solution allows to identify circulating biomarkers candidates by detection of highly significant molecular signatures like cancer-associated miRNAs and oncoproteins.

The working hypothesis is that canonical correlation analysis can be successfully used to integrate omics data such as transcriptome and proteome and, as a result, efficiently correlate molecular signatures. The algorithm highlights are biological features which are candidate biomarkers in cutaneous melanoma.

The proposed hypothesis was verified using human melanoma models for primary and metastatic tumors. Omics data from conducted experiments and from open access molecular databases were collected and analyzed.

Canonical correlation analysis has successfully integrated three different types of data: 1) miRNA transcripts and 2) proteomics from melanoma microvesicles with 3) whole cancer cells RNA transcripts. A list of 24 molecular signatures (out of thousands of possible) was obtained and a vast majority of them have a proven role in the cancer process. This result allowed the generation of biologically valid functional and mechanistic hypotheses that could be used by scientists to design experiments necessary to test the clinical potential of the proposed biomarkers.

Presented results confirm the validity of the research hypothesis that canonical correlation analysis can be applied for discovery of extracellular vesicles-derived melanoma biomarkers.

Streszczenie

Niniejsza rozprawa doktorska prezentuje nową metodę opartą o analizę korelacji kanonicznej, która adresuje wyzwania związane z odkrywaniem biomarkerów czerniaka skóry. Zaproponowane rozwiązanie identyfikuje wysoce prawdopodobnych kandydatów na biomarkery, wykrywając molekularne sygnatury takie jak miRNA czy onkoproteiny związane z procesem nowotworzenia.

Hipoteza robocza zakłada, że analiza korelacji kanonicznej może być z powodzeniem użyta do integracji danych omicznych takich jak transkryptom czy proteom i efektywnie korelować sygnatury molekularne. W rezultacie identyfikowane są wysoce istotne cechy biologiczne, które mogą być kandydatami na biomarkery w czerniaku skóry.

Zaproponowana hipoteza była weryfikowana z użyciem modeli linii komórkowych ludzkiego czerniaka skóry o różnym stopniu złośliwości. Wykorzystano dane pochodzące z przeprowadzonych eksperymentów oraz z publicznie dostępnych baz danych molekularnych.

W wyniku przeprowadzonej analizy korelacji kanonicznej z sukcesem zintegrowano trzy różne rodzaje danych: 1) transkryptom miRNA i 2) proteomikę pochodzące z mikropęcherzyków zewnątrzkomórkowych czerniaka oraz 3) transkryptom RNA z całych komórek nowotworowych. Otrzymano listę 24 sygnatur molekularnych (spośród tysięcy możliwych), z których znakomita większość ma udowodnioną rolę w procesie nowotworzenia. Rezultat ten pozwolił na wygenerowanie biologicznie uzasadnionych hipotez funkcjonalnych i mechanistycznych, które mogą posłużyć naukowcom w projektowaniu eksperymentów koniecznych do zbadania potencjału klinicznego proponowanych biomarkerów.

Zaprezentowane efekty potwierdzają hipotezę badawczą, że zastosowana metoda bioinformatyczna oparta o analizę korelacji kanonicznej może być stosowana jako narzędzie identyfikacji nowych biomarkerów w czerniaku skóry.

Glossary

B

bioinformatics - interdisciplinary field concerned with *in silico* collection, storage, integration, and analysis of biological data (usually large and complex datasets like: DNA, RNA, and protein sequences)

biological pathway - a sequence of processes taking place among biomolecules that results in a specific phenotype

biomarker - biological molecule related to the selected phenotype

C

canonical correlation - correlation between variables originated from two sets of multisets of data

CCA - canonical correlation analysis; a multivariate statistical technique used to determine the maximum correlation of variables from two sets or multisets of data

canonical variable - pairs of linear combinations of variables, that have the largest correlation

counts - a number of reads (sequence measurements) of specific feature (DNA, RNA, protein)

E

extracellular vesicle - membranous particle secreted by cell into extracellular environment, containing functional molecular content like RNAs, proteins and lipids

G

gene expression - measurement of RNA levels

genome - collection of organism's complete DNA sequence

genome sequencing - the process of identifying the exact order of base pairs in DNA molecule

genomics - field of data science dedicated to the detection and decoding of functional information from the DNA sequence

GWAS - genome-wide association study; an approach of examining large cohorts (group of patients with a shared phenotype) to identify genetic variants associated with a specific phenotype

I

interactome - collection of (ideally all) protein-protein interactions in cell

M

metabolome - collection of cell's complete small-molecules

microRNA - miRNA; non-coding single-stranded RNA molecule, composed of 21-23 nucleotides, regulating the expression of genes

molecular data - data from biological molecules like DNA, RNA, or proteins

multi-omics - multiple omics datasets collected from the same experimental sample (see omics)

multiset - more than two sets of data

N

NGS - Next Generation Sequencing; massively parallel DNA/RNA sequencing technology

nutrigenomics - a field that concerns understanding the relationship between nutrients and genes

O

omics - a range of biomolecular domains like genomics, transcriptomics, proteomics, or metabolomics. This term is usually used to refer to measurements made on the same sample (patient or cell line)

P

penalty (functions) - class of algorithms for constrained optimization

personalized medicine - promote treatment, prediction and prevention of diseases based on detailed molecular information about a patient

phenotype - a trait or other observable feature of the condition

proteome - collection of a cell proteins

S

sCCA - sparse Canonical Correlation Analysis; penalized version of CCA with a sparsity term that improves variable selection

systems biology - interdisciplinary area of research focused on interactions and interdependencies in complex biological systems

T

totalRNA - complete cell's RNA

transcriptome - collection of cell's complete RNA sequence

translational medicine - area of research that uses, integrates, and implements knowledge, resources and technology from various biomedical areas in order to develop and improve methods of treatment, prevention, and prediction of diseases

tumor microenvironment - TME; environment of cancer cells like blood vessels, extracellular matrix (ECM), biological molecules, or other cell types

Abbreviations and Acronyms

AAbs	autoantibodies
ABCA9	ABC transporter A family member 9
ABCB5	ATP-binding cassette sub-family B member 5
ABCC2	ATP-binding cassette sub-family C member 2
APP	amyloid precursor protein
ARF6	ADP-ribosylation factor 6
BMDC	bone marrow-derived cell
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase
CCA	canonical correlation analysis
CD4	cluster of differentiation 4
CD8	cluster of differentiation 8
CD40	cluster of differentiation 40
CD86	cluster of differentiation 86
CDK4	Cyclin Dependent Kinase 4
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
CisPt	Cisplatin
circRNA	circular RNA
CXCR4	CXC-chemokine receptor 4
dCCA	deep canonical correlation analysis
DNA	deoxyribonucleic acid
ECM	extracellular matrix
EGFR	epidermal growth factor receptor

ERK	extracellular signal-regulated kinase
EVs	extracellular vesicles
EXs	exosomes
FDA	Food and Drug Administration
GAPDH	glyceraldehyde-3-phosphate
GEO	Gene Expression Omnibus
GnRH	gonadotropin-releasing hormone
HSPG	heparan sulfate proteoglycan
ICAM	intracellular adhesion molecule
IGF1R	type 1 insulin-like growth factor receptor
ILVs	intraluminal vesicles
kCCA	kernel canonical correlation analysis
lasso	least absolute shrinkage and selection operator
LBPA	lyso-bis-phosphatidyl acid
LFA1	lymphocyte function-associated antigen 1
lncRNA	long non-coding RNA
MET	MET Proto-Oncogene; hepatocyte growth factor receptor
MHC I	major histocompatibility complex protein I
MHC II	major histocompatibility complex protein II
MIF	migration inhibitory factor
miRNA	microRNA
MVBs	multivesicular bodies
MVEs	multi vesicular endosomes
MVs	microvesicles
mtDNA	mitochondrial DNA
mtRNA	mitochondrial RNA
MTEX	melanoma cell-derived exosomes
NCBI	National Center for Biotechnology Information
NGI60	NCI-60 Human Tumor Cell Lines Screens
NGS	next generation sequencing

NIH	National Institute of Health
pCCA	penalized canonical correlation analysis
PD-L1	programmed death-ligand 1
PECAM1	platelet endothelial cell adhesion molecule
PIK3R1	Phosphoinositide-3-Kinase Regulatory Subunit 1
piRNA	piwi-interacting RNA
PLD	phospholipase D
PMEL	premelanosome protein
PPI	protein-protein interaction
pre-miRNA	precursor mRNA
PrP	prion protein
PTEN	Phosphatase And Tensin Homolog
RDA	redundancy analysis
RNA	ribonucleic acid
ROCK	RHO-associated protein kinase
rRNA	ribosomal RNA
RTK	receptor tyrosine kinase
sCCA	sparse canonical correlation analysis
snRNA	small nuclear RNA
ssDNA	single stranded DNA
TCGA	The Cancer Genome Atlas
TCR	T cell receptor
TDP43	TAR DNA-binding protein 43
TEM	transmission electron microscopy
TME	tumor microenvironment
TPM	transcripts per kilobase million
Treg	regulatory T cell
TRF	transferrin receptor
TSG101	tumor susceptibility gene 101 protein
TSPAN	tetraspanin

TWIST1	twist transcription factor 1
VPS	vacuolar protein sorting-associated protein
vtRNA	valut RNA

Contents

Abstract	vii
Streszczenie	viii
Glossary	ix
Abbreviations and Acronyms	xii
1 Introduction	1
1.1 Thesis objectives	4
THEORETICAL BACKGROUND	7
2 Canonical correlation analysis	7
2.1 Theoretical foundations	7
2.2 CCA extensions	9
2.3 CCA applications to omics studies	9
2.3.1 Cancer biomarkers discovery by CCA	12
2.4 CCA implementations in bioinformatics - state of the art	15
3 Extracellular vesicles	16
3.1 Classification and function of extracellular vesicles	16
3.2 Extracellular vesicles cargo	18
3.3 Extracellular vesicles in cancer	23
3.3.1 EVs in melanoma	24

3.4	EVs as melanoma biomarker	25
MATERIALS AND METHODS		30
4	Sparse multiset CCA computational implementation	30
5	Multi-omics melanoma datasets	32
5.1	Microvesicles miRNA data	33
5.1.1	Cell culture	34
5.1.2	Isolation of microvesicles	34
5.1.3	Isolation of microvesicles miRNA and library preparation	35
5.1.4	NGS sequencing and raw data preprocessing	35
5.2	Microvesicles proteomics data	37
5.3	Cell totalRNA data	37
6	Data integration	38
6.1	Sparse multiset CCA biomarkers discovery pipeline	38
RESULTS		42
7	Datasets quality control and preprocessing	42
7.1	Melanoma cell lines genetic background	42
7.2	miRNA, proteins, and totalRNA datasets preparation	44
8	Sparse multiset CCA optimization	49
9	Sparse multiset CCA results	51
9.1	Microvesicles miRNA biomarkers	53
9.2	Microvesicles proteomics biomarkers	54
9.3	Cell totalRNA biomarkers	55
9.4	Biomarkers supporting biological findings	56

10 Discussion	59
10.1 Combination of centrifugation and hydrostatic filtration dialysis is an efficient protocol for microvesicles miRNAs isolation	60
10.2 Melanoma models contain proto-oncogenes	60
10.3 Sparse multiset CCA - based bioinformatics pipeline is highly effective method for melanoma biomarkers candidates selection	61
10.4 Sparse multiset CCA - derived functional and mechanistic insights into microvesicles role in melanoma metastasis formation	62
10.5 Identified molecular signatures were used to design a panel of candidates for liquid biopsy biomarkers	63
10.6 Possible area for improvements	63
11 Conclusions	65
Bibliography	85
Appendices	87
.1 List of CCA-based bioinformatics tools	87
.2 List of omics databases	88
.3 Melanoma molecular profile	89
.4 Sparse multiset CCA results	91
.5 Optimization of sparse multiset CCA results	93
.6 Melanoma EVs biomarkers research	103

Chapter 1

Introduction

New generation sequencing (NGS) and other novel molecular measurement methods provide a huge number of multidimensional data [1]. They enable to gathering information on millions of molecular variables from biological sample. Simultaneously, large-scale databases such as the National Center for Biotechnology Information (NCBI) Human Genome Project Resources [2,3], 1000 Genomes Browser [4], Universal Protein Resource (UniProt) [5,6] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [7,8] are developing dynamically. Observed innovations give new meaning to the challenges of collecting, storing, managing, and analyzing biological data [9,10]. In literature, this phenomenon can be found under the collective name: multi-omics, where omics is shorthand for the following -omics domains: genomics, transcriptomics, proteomics, metabolomics, lipidomics, epigenomics and microbiomics [11,12].

An optimal systems biology approach needs to be proposed for integrating several sets of different types of biomolecular data. In this contexts, a fundamental matter is understanding the interactions between transcriptomics, proteomics, and other types of omics and non-omics data, generated using different platforms. In such analyzes, a selection of variables is a key to obtain interpretable results, which has been applied and developed in my previous omics study [13]. The systems biology approach allows scientists to understand the role of genetic factors and their impact on such complex diseases as cancer. Canonical correlation analysis (CCA) is one of the most powerful methods for this bioinformatic challenge. Using CCA in biological data integration for cross-platform studies has been reported recently in a number of publications as a very promising methodology [14]. The application of these

novel methods of simultaneous analysis of multiset biomolecular data opens new possibilities for understanding the hitherto unknown mechanisms and interdependencies occurring in the tumor microenvironment (TME) related to intercellular communication.

Effective study of TME and the processes of metastasis formation requires a deep understanding of the key mechanisms involved in transfer of biological information. Cells communicate with each other using not only single molecules, but also by means of structures called extracellular vesicles (EVs) [15, 16]. Cells are capable of secreting EVs during inward budding of multi vesicular endosomes (MVEs) (exosomes: 50-100 nm), outward budding of plasma membrane (microvesicles, MVs or ectosomes: 100-500 nm) as well as a result of other biological processes, e.g. programmed cell death, called apoptosis (apoptotic blebs) [17].

Systematic analyses have proved that EVs molecular cargo consist of selected proteins, lipids, and nucleic acids. It reflects the "status" of the secreting cell which plays a crucial role in a cell-to-cell crosstalk [18]. There is growing evidence supporting this new communication paradigm not only in physiological conditions but also in different pathological states such as tumor formation process [19]. EVs are involved in cells interactions in tumor microenvironment to control processes such as pre-metastatic niche construction [20].

Cancer cells release EVs containing rare RNA species and other nucleic acids fragments with driver mutations [21]. In addition, a higher number of EVs are observed in the blood of cancer patients compared to healthy population [22]. Clearly, this uniquely detected EVs cargo profile has great promise as a new source of biomarkers for liquid biopsy, to offer a high level of sensitivity and specificity of a the non-invasive medical procedure through the collection of body fluid samples such as blood [23]. This is a typical precision medicine approach where the diagnosis or selection of appropriate treatment is identified based on molecular tests of patient's genetic factors [24].

Melanoma is one of the best described cancers in which a significant EVs influence on oncogenesis is observed [25, 26]. Great progress has been made in understanding the biology of melanoma-derived EVs in mouse models but also in humans. Despite many studies, melanoma remains a huge clinical challenge [27]. Based on European Cancer Information System (ECIS), estimated incidence of skin melanoma in European Union 28 countries (EU28) in 2018 is about 120,505 new cases (60,524 in males and 59,981 in females) (Source:

<https://ecis.jrc.ec.europa.eu>, accessed on 2/08/2020). In the same year in Poland, 3,723 people were diagnosed with skin melanoma and 1,728 died, this makes Poland one of the countries with the highest mortality rate from skin melanoma in Europe (the 7th place in 31 countries in 2018), despite the incidence per 100,000 inhabitants is very low (the 4th place in Europe: ASR (age-adjusted rates) = 10.3). The main reason for this is a lack of effective biomarker candidates which are easy to use [28,29]. Currently, dermatoscopy (subjective assessment of moles and other pigmented patches on the skin by a dermatologist using an optical device called a dermatoscope) is the only reliable test to diagnose skin melanoma, while the test that clearly confirms or excludes the presence of cancer cells is still the histopathological examination of the resected tissue, which is invasive procedure and is not suitable for patients screening or stratification [27,30]. Scientists' attention is focused mainly on biomarkers from blood samples, which can be used in routine screening, especially in high-risk patients populations (high number of moles, family history of melanoma, Fitzpatrick skin phototype: ivory white skin with freckles, light blue eyes and platinum or red hair, weakened and suppressed immune system or accompanying diseases for example xeroderma pigmentosum) [27]. In 2018, very promising results were reported in the literature for blood biomarkers panels consisting of 10 autoantibodies (AABs) that can be used to detect early stages of melanoma with sensitivity of 79 % and specificity of 84 % [28].

The above-described factors cause that scientists and clinicians to search for new biomarkers that could improve the diagnosis and treatment of melanoma [28,31,32]. A detailed study of the molecular content of cancer-derived EVs is one of the most promising area for new diagnostic, prognostic, or predictive biomarkers. This increased focus on bioactive molecules from circulating EVs is accompanied with new challenges in development of state-of-the-art bioinformatics methodology for EVs-based biomarker discovery [33].

The presented doctoral thesis takes up the challenge of integrating multi-omics data derived from EVs and whole cancer cells in order to identify important molecular signatures that may be candidates for new liquid biopsy biomarkers in cutaneous melanoma.

1.1 Thesis objectives

The aforementioned unaddressed scientific and clinical challenges were the inspiration for this PhD thesis. The presented thesis is focused on hypothesis that canonical correlation analysis-based bioinformatics algorithm can be an effective method for identification of the extracellular vesicle-based biomarker candidates in human melanoma. The main goal of the thesis is to develop a new bioinformatics pipeline with strong support of CCA component and study its application for melanoma multi-omics dataset.

To achieve this research goal, the following steps were taken:

1. Preparation of proprietary and public datasets obtained from microvesicles and whole cells of human skin melanoma cell lines.
2. Development of a bioinformatics method based on sparse multiset canonical correlation analysis for selecting highly correlated biological features.
3. Application of the sparse multiset canonical correlation analysis method to integrated experimental melanoma data.
4. Data validation using information regarding biological interrelationships, bioactive molecules, and their functions.
5. Interpretation of obtained results and creation of mechanistic and functional hypothesis for biomarker candidates.

The assumptions described above are presented on a graphical abstract (Figure 1.1).

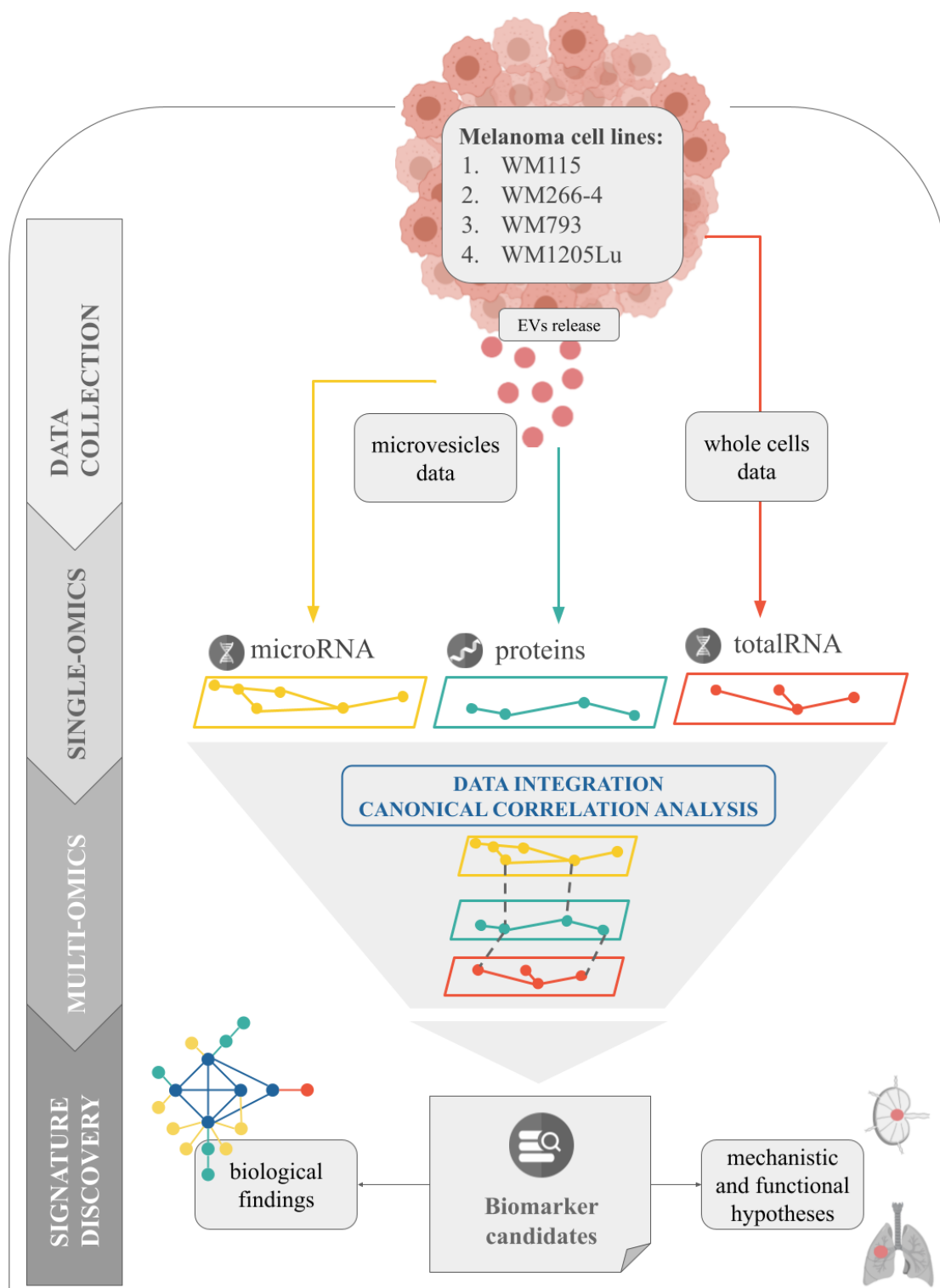


Figure 1.1: Graphical abstract.

THEORETICAL BACKGROUND

Chapter 2

Canonical correlation analysis

The first description of canonical correlation analysis (CCA) was published by Harold Hotelling in 1936 [34]. He introduced CCA as a method to identify relation between two datasets by searching for linear combinations of variables that have the largest correlation. These pairs of linear combinations are called the canonical variables and correlations between them are called canonical correlations. CCA helps identify and quantify the strength of association by reducing a high-dimensional relationship between studied sets of variables into more intuitive and easily interpretable pairs of canonical variables.

With the development of next-generation sequencing, scientists are facing new difficulties related to the large amount of multi-variate omics data (where "omics" phrase is shorthand for: genomics, transcriptomics, epigenomics, proteomics, lipidomics, metabolomics and microbiomics) [35]. Methods based on conventional CCA or its variants and extensions perfectly address these challenges and, as a consequence, work well in molecular biology studies e.g. integration of omics data or network analysis [36–38]. CCA-based computational tools are one of the most promising and powerful approaches that can decode complex relationships between datasets and find relevant biological signatures [39, 40].

2.1 Theoretical foundations

As stated before, the main objective of conventional CCA is to find pairs of weight vectors $\mathbf{a}$ and $\mathbf{b}$ which will yield maximum correlation between canonical variables $\mathbf{V}_{\mathbf{X}} = \mathbf{X}\mathbf{a}$ and

$\mathbf{V}_Y = \mathbf{Y}\mathbf{b}$. $\mathbf{X}$ and $\mathbf{Y}$ are the data matrices containing measurements of analyzed datasets. Columns in these matrices contain independent dimensions, while rows represents observations (measurements performed on the same entities).

If columns of both data matrices have zero mean, formula for calculating CCA weights can be described as:

$$(\mathbf{a}, \mathbf{b}) = \arg \max_{\mathbf{a}, \mathbf{b}} \text{corr}(\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b}) = \arg \max_{\mathbf{a}, \mathbf{b}} \frac{\mathbf{a}^\top \Sigma_{\mathbf{XY}} \mathbf{b}}{\sqrt{\mathbf{a}^\top \Sigma_{\mathbf{XX}} \mathbf{a}} \sqrt{\mathbf{b}^\top \Sigma_{\mathbf{YY}} \mathbf{b}}}, \quad (2.1)$$

where $\Sigma_{\mathbf{XX}} = \mathbf{X}^\top \mathbf{X}$, $\Sigma_{\mathbf{YY}} = \mathbf{Y}^\top \mathbf{Y}$, $\Sigma_{\mathbf{XY}} = \mathbf{X}^\top \mathbf{Y}$ are covariance matrices. Because correlation coefficient is invariant when we multiply $\mathbf{a}$ or $\mathbf{b}$ by a constant value, we need to impose additional constraints on the problem. Mathematically convenient choice for such constraints is:

$$\mathbf{a}^\top \Sigma_{\mathbf{XX}} \mathbf{a} = 1, \mathbf{b}^\top \Sigma_{\mathbf{YY}} \mathbf{b} = 1. \quad (2.2)$$

Maximalization problem from equation 2.1 with constraints (2.2) can be solved using Lagrange multipliers method. In Lagrangian form, this can be stated as:

$$\mathcal{L}(\lambda_1, \lambda_2, \mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \Sigma_{\mathbf{YX}} \mathbf{b} - \frac{\lambda_1}{2} (\mathbf{a}^\top \Sigma_{\mathbf{XX}} \mathbf{a} - 1) - \frac{\lambda_2}{2} (\mathbf{b}^\top \Sigma_{\mathbf{YY}} \mathbf{b} - 1), \quad (2.3)$$

where λ_1 and λ_2 are Lagrange multipliers. The solution is obtained by taking derivatives of $\mathcal{L}$ with respect to $\mathbf{a}$ and $\mathbf{b}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = \Sigma_{\mathbf{XY}} \mathbf{b} - \lambda_1 \Sigma_{\mathbf{XX}} \mathbf{a} = 0, \quad (2.4)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \Sigma_{\mathbf{YX}} \mathbf{a} - \lambda_2 \Sigma_{\mathbf{YY}} \mathbf{b} = 0. \quad (2.5)$$

Subtracting equation 2.4 times $\mathbf{a}^\top$ from equation 2.5 times $\mathbf{b}^\top$ gives:

$$0 = \lambda_1 \mathbf{a}^\top \Sigma_{\mathbf{XX}} \mathbf{a} - \lambda_2 \mathbf{b}^\top \Sigma_{\mathbf{YY}} \mathbf{b}. \quad (2.6)$$

This equation with constraints (2.2) leads to: $\lambda_1 = \lambda_2 = \lambda$. Applying this equality to equations 2.4, 2.5 will lead after rearranging, to final canonical correlation analysis equations:

$$\Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-1} \Sigma_{\mathbf{YX}}^{\top} \mathbf{a} = \lambda^2 \mathbf{a}, \quad (2.7)$$

$$\Sigma_{\mathbf{YY}}^{-1} \Sigma_{\mathbf{XY}}^{\top} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}} \mathbf{b} = \lambda^2 \mathbf{b}. \quad (2.8)$$

The form of these equation is a standard eigenvalue problem ($\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$) for which many solving algorithms exist.

2.2 CCA extensions

An emergent research directions make conventional CCA insufficient to analyze high-dimensional multiset omics data. As a result, subsequent articles proposed variants of the CCA method to meet the challenges of multi-omics data [1]. Some of the most used are:

1. for nonlinear associations between two datasets collected on the same samples: kernel CCA (kCCA) or deep CCA (dCCA) (green color on Figure 2.1 and Figure 2.2),
2. for datasets with number of samples grater than the number of variables: penalized CCA (pCCA) or sparse CCA (sCCA) (blue color on Figure 2.1 and Figure 2.2),
3. for linear associations between more than two datasets collected on the same samples: multiset CCA or sparse multiset CCA (violet color on Figure 2.1 and Figure 2.2).

The technical relationships between CCA versions as well as the decision flowchart in selecting the optimal method for specific data is presented on Figure 2.1 and Figure 2.2.

2.3 CCA applications to omics studies

Next Generation Sequencing (NGS), together with other ultra-high throughput experimental methods provide enormous amounts of multi-dimensional datasets. Term multi-omics or integrative omics refers to usage of multiple biological data like genome, transcriptome, proteome, metabolome etc. data for a complex system biology anaysis [43, 44]. Decoding the interpretable interactions between data generated using different platforms is "the holy

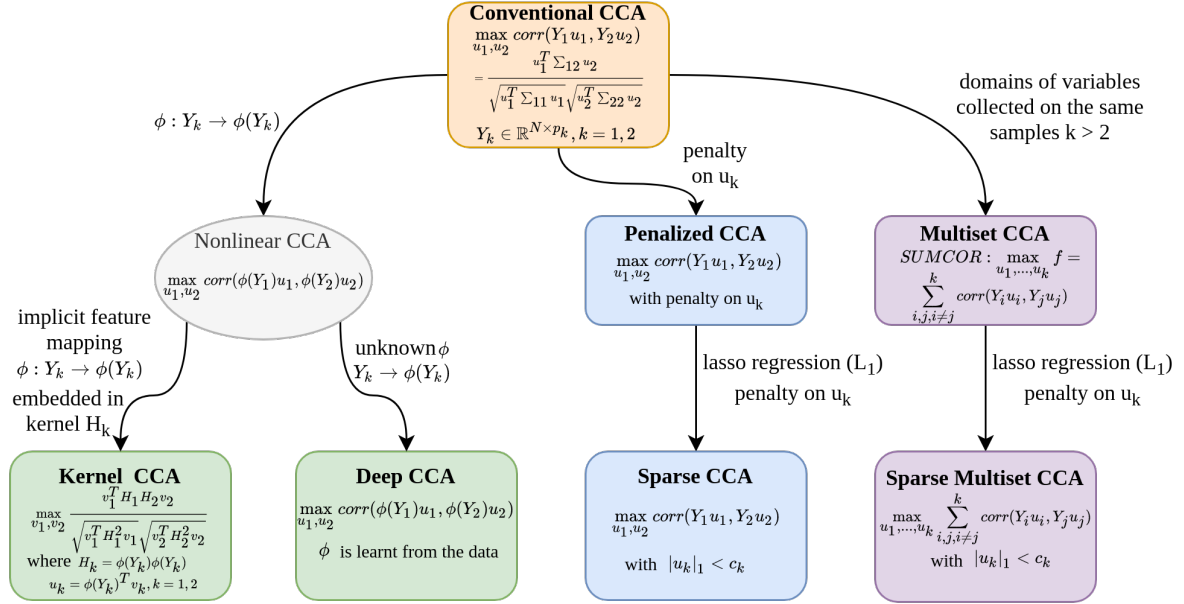


Figure 2.1: Relationship between conventional CCA (orange) and three main CCA variants: nonlinear CCAs (green), penalty CCAs (blue) and multiset CCAs (violet) [41]. To simplify the mathematical notation, Y with subscripts correspond to matrices and u with subscripts correspond to weight vectors. Figure created with diagrams.net.

grail" of bioinformatics [45]. Another challenge is the high-dimensionality of the data with a small sample support. Typical "omics" dataset has a significantly lower number of examined samples (cell lines, patients) than the number of analyzed features (miRNA, RNA, proteins). In such analyses, the integration of multiple omics datasets together and the selection of variables is a key to identify biological interrelationships [46]. Canonical correlation analysis (CCA) is one of the most effective method to address this research challenge.

One of the first studies on CCA as a molecular data exploration tool was the article published in 2004 by Yamanishi et al. They used kernel CCA to predict the protein networks based on multi-omics data and metadata like gene expression, protein-protein interactions (PPIs), phylogenetic profiling, or protein subcellular localization [47]. In 2009 Parkhomenko and colleagues demonstrated new sparse CCA algorithm in study of two datasets simultaneously (SNPs and gene expression) to indicate natural variations in human genome [36]. One year later (2010), Naylor et al. introduced sparse CCA to study regulatory variants in gene expression data of children with asthma by finding functionally relevant SNPs and quanti-

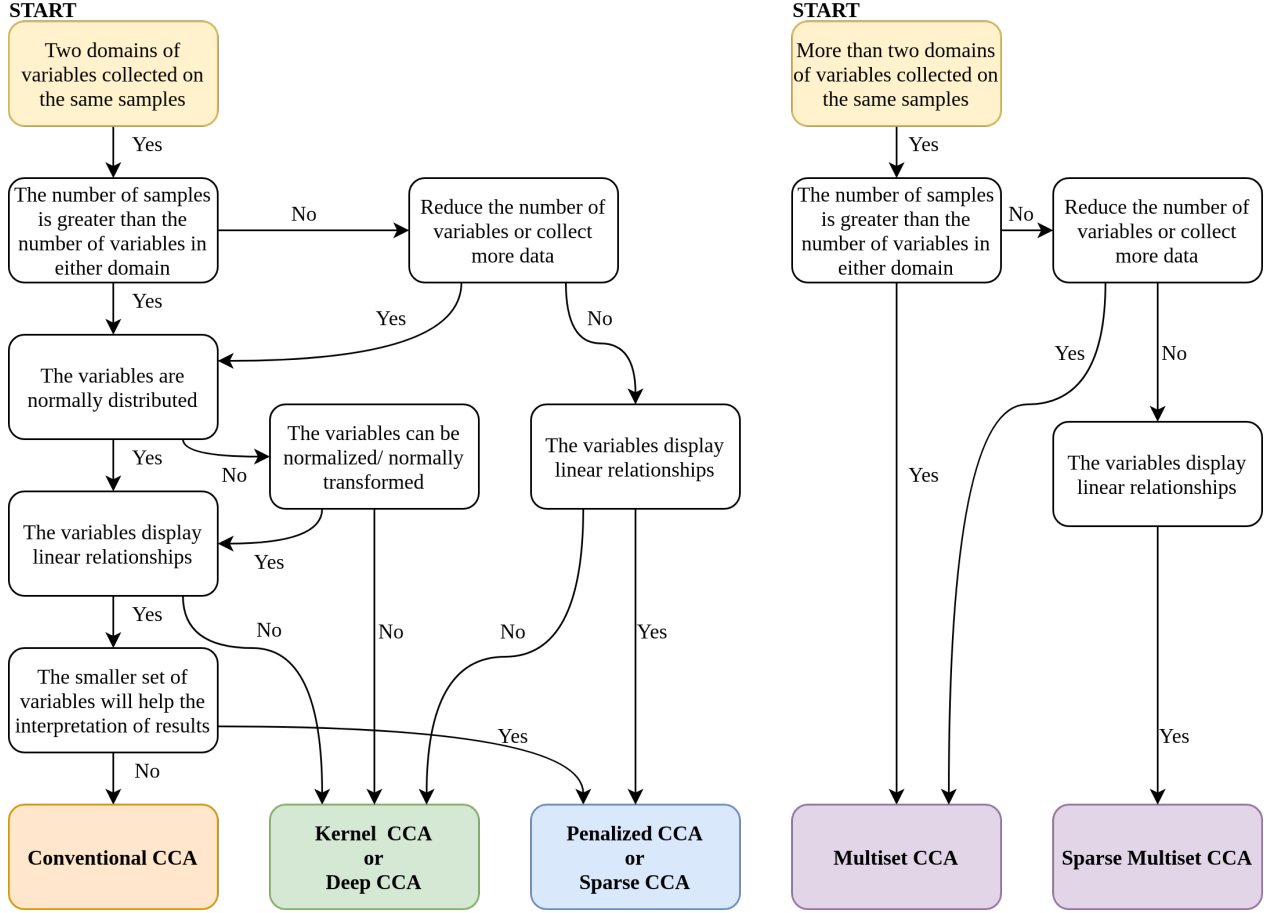


Figure 2.2: The decision flowchart presenting the possible choices when considering implementation of CCA for a omics dataset [42]. Figure created with diagrams.net.

tative trait loci (eQTLs) [38]. Later on, in 2012, Tang and his research team applied sparse CCA to white blood cells genome-wide association studies (GWASs) to conduct gene-based test of association [48]. In 2013, Rousu et al. described sparse CCA biomarker discovery pipeline with asymmetrical penalty for tuberculosis and malaria datasets by use proteomics data and complex clinical phenotypic profiles [39]. Jun et al. showed, in 2018, very promising results with multiset CCA used on multi-block genomic data [49]. Their method identified single-nucleotide polymorphisms (SNPs) related to analysed diseases (diabetes, coronary heart disease or Parkinson disease). In 2020, Revilla and colleagues reported sparse multiset CCA method, applied to host RNA and microbiome DNA data originated from inflammatory bowel disease patients [50].

CCA-based methods have been heavily explored in the field of metabolomics and nutrige-

nomics research. In 2009, González et al. presented regularized CCA (rCCA) to study nutrigenomics (gene expression combined with concentrations of fatty acids published by Martin et al. [51]) and reported new transcription factor pathways that could potentially link fatty acids with gene expression [52]. In 2015, Cao et al. introduce sparse CCA method validated on the same nutrigenomic dataset [40]. In the same year, Piwowar et al. published workflow combining conventional CCA as well as regularized CCA. It outlined, in the Martin’s et al. nutrigenomics data, previously unreported connections important for lipid metabolism [53]. One of the latest applications (2017) on Martin and colleagues experimental data is Yoshida et al. publication presenting two-stage kernel CCA (TSKCCA) [54]. They identified novel important lipid metabolism factors (like specific types of Enoyl-CoA isomerase). Later on, in 2020, Rodosthenous and his research group also evaluated proprietary sparse CCA method on this "golden standard" nutrigenomics panel. The most recent demonstrations (2020), using these dataset is a novel CCA method proposed by the Turek et al. [13].

2.3.1 Cancer biomarkers discovery by CCA

In 2008, Waaijenborg et al. proposed a penalized CCA (pCCA) to study co-regulated and co-expressed genetic networks in glioma patients data from National Center for Biotechnology Information - Gene Expression Omnibus (GEO) database [55]. They used the elastic net-based penalization and demonstrated effective analysis of association between gene expression and DNA-markers (they outlined insulin growth factor receptor type I - IGF1R, as key glioblastoma development gene). The year after this publication (in 2009), Waaijenborg and colleagues updated the sparse CCA method with asymmetric penalty (ridge penalization for first variables set and elastic net penalization for the second variables set) and evaluated that on the same glioblastoma patients data (from REpository for Molecular BRAin Neoplasia DaTa- REMBRANDT) to study completeness of brain cancer-related gene networks [56]. In the third quarter of the 2009, the same research team reported identification of important groups of genes and SNPs in glioma patients by using non-linear sCCA [57].

Year 2009 was full of articles on the application of CCA in the context of omics data in cancer. At the beginning of the year, Witten et al. proposed two novel extensions of sparse CCA in lymphoma data analysis (from GEO database): sparse supervised CCA (to

identify variables not only associated with each other but also with the outcome) and sparse multiple CCA (to study more than two matrices) [37]. In the same month, Lê Cao et al. introduced sparse CCA with elastic-net penalization and co-inertia analysis (CIA) to study relationships between two gene expression matrices based on The NCI-60 Human Tumor Cell Lines Screens (NCI60) [58]. In the same year, Witten’s group reported use of a sparse CCA pipeline to a breast cancer gene expression and DNA copy number dataset (from International Cancer Benchmarking Partnership resources) [59]. They applied sCCA with a least absolute shrinkage and selection operator (lasso) penalty and successfully identified gene regions of gain and loss. Two months after that, González et al. demonstrated regularized CCA for the pharmacogenomic cancer panel (from NCI60) and emerged novel ATP-binding cassette transporter candidate substrates [52].

In 2010 Soneson et al. used conventional CCA and regularized dual CCA to childhood leukemia copy-number variants and genes expression data corresponding to ten subtypes of this disease [60]. Three years after this very promising results (2013), Hong et al. published single and bivariate CCA methods. They validated algorithms on lung squamous cell cancer datasets from The Cancer Genome Atlas (TCGA) and constructed RNA-seq co-expression networks [61]. In the same year, Lin and colleagues presented two new sparse CCAs with elastic net penalties. They analysed SNPs and gene expression of gliomas patients (from the GEO database) and NCI60 [62]. In 2020, Rodosthenous et al. compared CCAs solutions introduced by Parkhomenko et al., Witten et al., its extensions and proposed a new approach to multiple sCCA for analysis of more than two matrices [14]. They evaluated the method by using gene expression, miRNA and methylation TCGA sets for breast, kidney, and lung cancer types.

From the literature review described above, it can be certainly concluded that the variants of sparse canonical correlation analysis are the most frequently implemented methods in cancer research, which is summarized on Figure 2.3.

Few of described CCA studies used skin melanoma data [52, 58, 62]. Lê Cao group reported overexpression of genes associated with tyrosinase, tyrosinase related proteins 1 and 2, dopachrome tautomerase, glycosphingolipid (ganglioside and globosid), melanocyte inducing transcription factor, vimentin, S-100A1, S-100B and melan-A [58]. González and oth-

ers described high expression of ABCC2, ABCA9 and ABCB5 genes based on melanoma cell lines pharmacogenomic study [52]. Lin et al. identified epidermal growth factor receptor (EGFR), transcription factor p53 and gonadotropin-releasing hormone (GnRH) pathways as key in their melanoma data analysis [62]. None of the publications mentioned above analyzed melanoma data using more than two datasets (multiset CCA).

Sparse CCA				Sparse Multiset CCA			
Publication Year	Reference	Cancer type	Database	Publication Year	Reference	Cancer type	Database
2008 2009 2013	[55] [56] [62]	glioma	GEO	2009	[37]	lymphoma	GEO
2009	[57]	glial	REMBRANDT	2020	[14]	breast kidney colon lung	TCGA
2009 2009 2013	[58] [52] [62]	leukemia melanoma ovarian breast prostate lung renal colon CNS	NCI60				
2009	[59]	breast	ICBP				
2013	[61]	lung	TCGA				

Figure 2.3: The most commonly used CCA-based method in cancer research is sparse canonical correlation analysis. The table shows 10 publications that reported sCCA application for different cancer types data, including two articles showing results for three datasets collected on the same samples (sparse multiset canonical correlation analysis). Researchers most often used data resources such as: The Gene Expression Omnibus (GEO), The Cancer Genome Atlas (TCGA) or The NCI-60 Human Tumor Cell Lines Screens (NCI60), and less often from databases dedicated to specific cancer type like: REpository for Molecular BRAin Neoplasia DaTa (REMBRANDT) or International Cancer Benchmarking Partnership (ICBP). Figure created with diagrams.net.

2.4 CCA implementations in bioinformatics - state of the art

In publicly available bioinformatics resources there are about 16 updated and supported CCA-based tools, dedicated to omics data (see Appendix 1). Each of the methods is based on well-described functions, so that the user does not use them as so-called "black boxes". The tools are available on the websites of research groups as well as in repositories popular among the bioinformatics scientists community, such as: Bioconductor (<https://www.bioconductor.org/>) [63,64], The Comprehensive R Archive Network (CRAN, <https://cran.r-project.org/>) [65] or GitHub (<https://github.com/>).

A big problem in developing and validating CCA-based bioinformatics tools is a small amount of relevant publicly available experimental data. Many groups use the Martin et al. nutrigenomic dataset [51], which has become one of the most frequently used data in articles and technical manuals [13, 14, 40, 52–54, 66]. In the case of cancer research, the following databases are in the lead: GEO [37, 55, 56, 62], NCI60 [52, 58, 62, 67] and TCGA [14, 61, 68] (Figure 2.3). There is a great need to create new data collections that could be used for CCA methods development and improvement.

Of the publicly available implementations, only two are dedicated to multiple datasets: 'sCCA' and 'MultiCCA' [14,59] (see Appendix 1). Both solutions are based on the algorithm described by Witten et al. in an article from 2009. In 2017, Csala et al. proposed an implementation of redundancy analysis (RDA) to multiset omics data [69, 70]. The method may become an alternative to multiset CCA, but at the moment it is still insufficiently validated on clinical data to draw appropriate conclusions on its effectiveness (currently validation was performed only on the data for the Marfan syndrome: methylation, gene expression, and cytokine markers).

Chapter 3

Extracellular vesicles

Extracellular vesicles (EVs) are cell-derived particles, involved in physiological and pathological processes, including oncogenesis [20]. Their molecular content, such as: miRNA, mRNA, long non-coding RNA (lncRNA), DNA fragments, proteins, and lipids, allows cells to exchange biological information [15–18]. EV cargo is considered as a key orchestrator in crosstalk between cells in tumor microenvironment (TME) [19,22]. However, detailed description of the EV mode of action in tumor progression and metastasis remains unknown. EVs are secreted by all types of cells and they are present in all body fluids (including blood), furthermore cancer cells are now indicated to produce more EVs than nonmalignant ones [71]. Based on this knowledge, EVs are recognized as promising liquid biopsy-based cancer biomarkers [20,72].

3.1 Classification and function of extracellular vesicles

The term extracellular vesicles is used to refer to all membrane vesicles but they consist a heterogeneous group. Therefore, it remains a big challenge to properly characterize EVs and describe their properties and functions [17]. Current studies have been mostly focused on two major populations of EVs: exosomes (EXs) originated from the endosomal membrane and microvesicles (MVs or ectosomes) formed from the plasma membrane, but EVs may also have different biogenesis (Figure 3.1: a and b) [18]. A cell can secrete cellular particles during different biological processes, such as: apoptosis (apoptotic bodies), migration (migrasomes) by synthesis and releasing of chemotactic molecules or cancer progression (oncosomes) [73,74].

Although the formation of subpopulations of EVs occurs in different manner states of a cell cycle, the same cellular processes and sorting machines are involved in the biogenesis of these entities and the cargo is the first regulatory factor in all EV subtypes formation. The composition of EVs is controlled by a donor cell type and its physiological state. Each class of EVs can be distinguished by their distinct molecular messengers like functional nucleic acids or proteins [18]. According to the literature, the typical steps of the vesicle formation include: cargo clustering, membrane budding, fission process and EV release, which have been described in a number of papers [75].

Exosomes are intraluminal vesicles (ILVs) generated by inward budding of membrane of multivesicular endosomes (MVEs) also called multivesicular bodies (MVBs) and they are secreted into extracellular space (Figure 3.1: a). Their size in diameter is usually between 50 nm and 100 nm. Ectosomes, also called microvesicles, are created by outward budding of the plasma membrane and secretion into extracellular environment (Figure 3.1: b). They are typically larger than exosomes and ranging from 100 nm to 500 nm in diameter. Ectosomes are a more mysterious EV fraction than exosomes. There are relatively small amount of studies describing ectosomes cargo and functions in detail [76–80].

Previous experiments have identified different ways in which the EVs may internalize and affect the function of the recipient cell: (1) direct fusion of EVs with recipient cell plasma membrane [81]; (2) internalization of EVs via macropinocytosis and endocytosis including clathrin-mediated endocytosis, caveolin-dependent endocytosis, and lipid-rafts interaction [18,82,83]; (3) receptor-ligand interactions on the recipient cell surface as observed in phagocytosis or endocytosis [84]. EV functional cargo (including miRNA, mRNA, DNA fragments, cytosolic and nuclear proteins, metabolites, membrane, and extracellular matrix proteins) can induce complex biological response in all these processes [85]. Near and long-distance cell-to-cell communication via EVs has been widely reported in various aspects:

1. **Gene transcription and translation.** | Studies showed that EV mRNAs and miRNAs (called “EV shuttled RNA”) can be transported to distant cells, allowing horizontal genetic transfer. Some of this RNA species are exclusively packed into EVs and stay functional in the new location, resulting in specific protein production (mRNA) or regulation of gene expression in recipient cells (miRNA) e.g. angiogenesis-related genes

[78, 86].

2. **Immune response.**| EVs released by immune cells (like B cells and dendritic cells), mesenchymal stem cells or cancer cells, contain immunomodulatory cargo: major histocompatibility complex proteins (MHC I and MHC II); membrane organisers (CD8, CD86, CD40 ligand, and programmed death-ligand 1: PD-L1) or neoantigens. This molecular composition stimulates or suppresses immune response by influencing CD4⁺ and CD8⁺ T cells, natural killer cells, regulatory T cells (Tregs), and macrophages. For example, B lymphocyte exosomes have been found to carry the MHC class I and II proteins on their surface for antigen presentation which results in specific T cells activation [87]. This mechanism is of great importance in antitumor influence of immune cells EVs.
3. **Disease progression.**| EV integrins and tetraspanins can be targeted to docking specific cells (in selected organs). This phenomenon is key in pre-metastatic niche formation [20]. Tumor-derived particles molecular content can affect disease progression also through signaling proteins like mesenchymal-to-epithelial transition (MET) receptor tyrosine kinase (RTK) and macrophage migration inhibitory factor (MIF) or RNA.
4. **Other:**| cell survival and proliferation, metabolism, angiogenesis, wound healing, or host-microbiome interaction, and viral immunity [22].

3.2 Extracellular vesicles cargo

It has been shown that cells are able to regulate EVs cargo generation and selection depending on the physiological or pathological process. This molecular composition can contain proteins (membrane organisers, biogenesis factors, oncoproteins), lipids, nucleic acids including DNA and RNA, MHC class I and II, extracellular matrix adhesion molecules, and many others illustrated in more detail on the Figure 3.2 and Figure 3.3.

The cargo of EVs can contain a unique nucleic acids composition but the mechanism of cargo selection and loading remains unknown and needs investigation. Accumulating evidence indicates that this molecular content is very complex and contains DNA fragments [88] like

mitochondrial DNA (mtDNA), double stranded DNA (dsDNA), single stranded DNA (ssDNA), viral DNA and RNA types [89] like messenger RNA (mRNA), micro RNA (miRNA), precursor miRNA (pre-miRNA), circular RNA (circRNA), mitochondrial RNA (mtRNA), transfer RNA (tRNA or sRNA, for soluble RNA), tRNA-derived small RNA (tsRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), piwi-interacting RNA (piRNA), ribosomal RNA (rRNA), vault RNAs (vtRNAs) or long non-coding RNA (lncRNA). These nucleic acids can significantly affect the biology of a recipient cell e.g. EV miRNAs pair with target mRNA and inhibit gene expression in cells.

The protein composition has a direct impact on the fate and function of extracellular vesicles, enhancing the importance of selective content-sorting mechanisms. Depending on the cell type, EVs will be characterized by a set of proteins unique to a given cells population and physiological or pathological conditions. EVs can deliver proteins like: extracellular, cytoskeletal, heatshock and nuclear proteins, enzymes, RNA binding proteins, apoptotic, and signal-transducers [18].

Based on all these findings, EVs hold a great promise as a source of molecular information for clinical application [90].

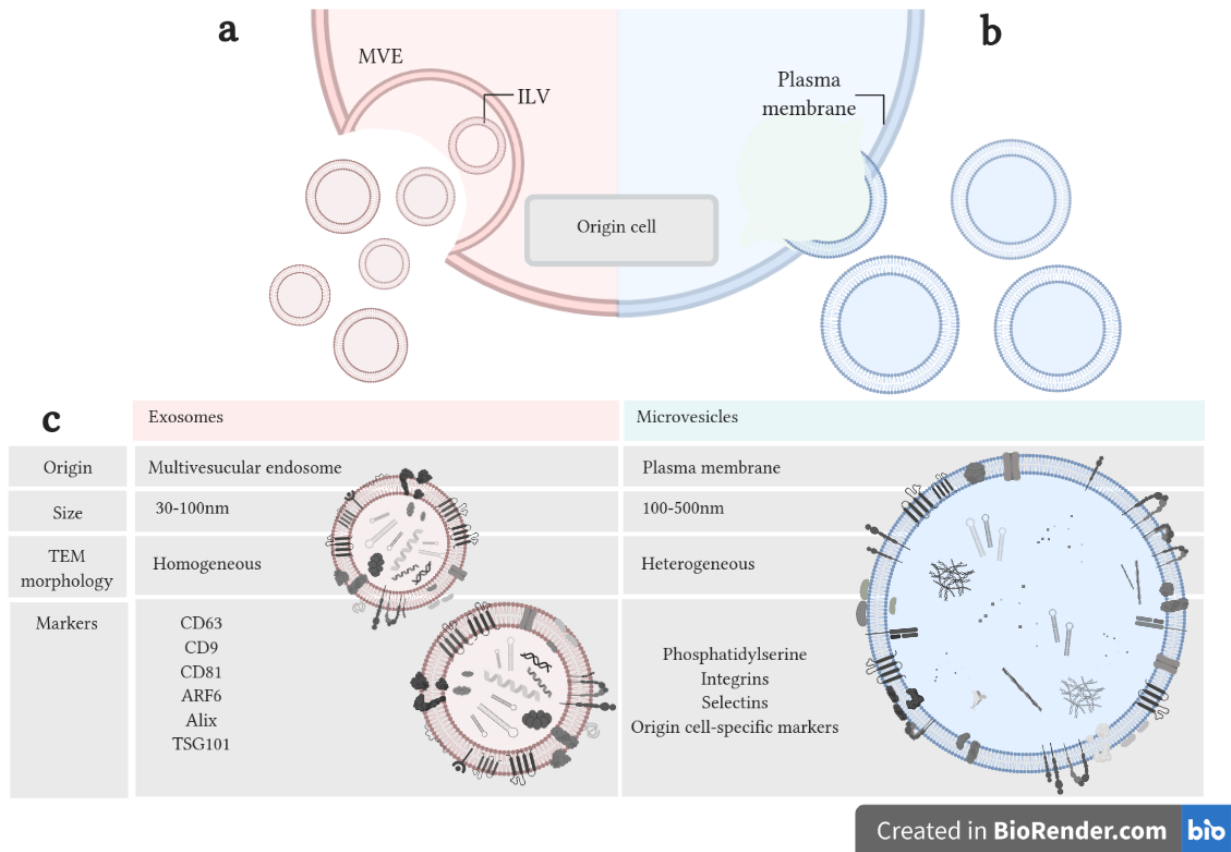


Figure 3.1: | **Extracellular vesicles biogenesis and main features [18]**. Almost all cell types release extracellular vesicles in physiological or pathological conditions. a| Exosomes are formed by inward budding of membrane of the matured multivesicular endosomes (MVEs). Particles formed in this process are called intraluminal vesicles (ILV). ILVs released into the extracellular space are then referred to as exosomes. b| Microvesicles also called ectosomes are created by direct budding of the plasma membrane and they are released into the extracellular space. c| Extracellular vesicles can be divided into two main classes: exosomes and microvesicles. The average size of the EVs is from 30 nm to 500 nm in diameter. EVs have their characteristic transmission electron microscopy (TEM) morphology (homogeneous for exosomes and heterogeneous for microvesicles) and signature set of markers. Figure created with BioRender.com.

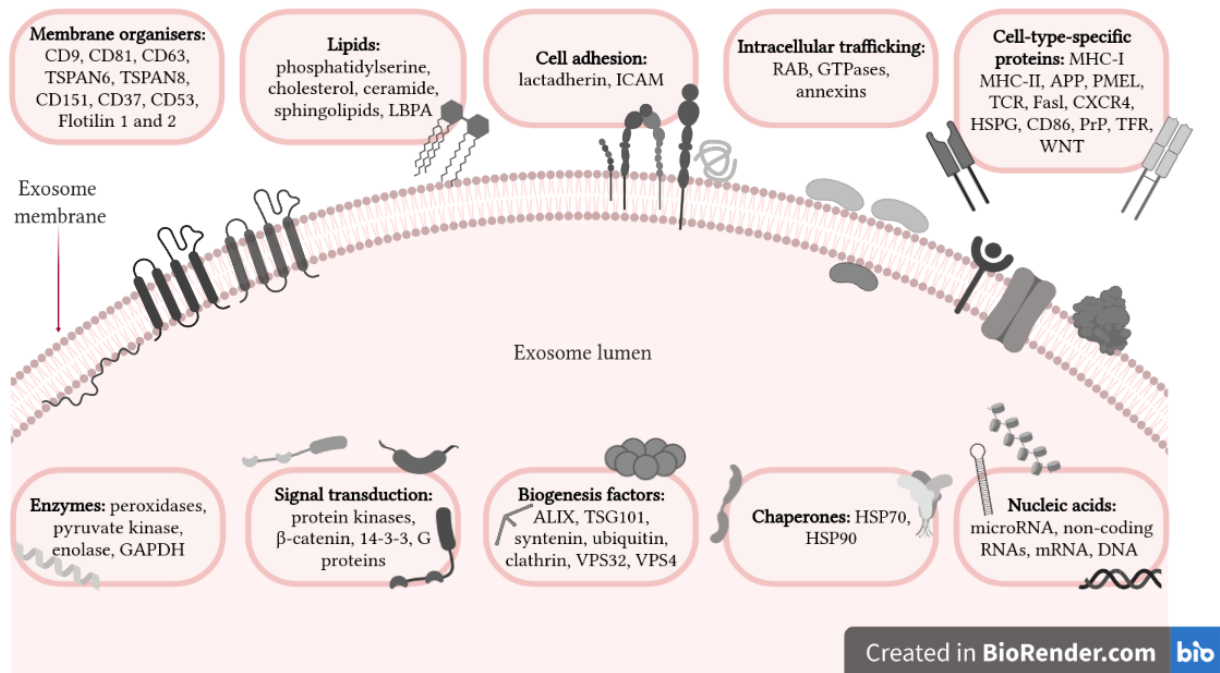


Figure 3.2: | **Exosome molecular content.** Exosomes cargo include for example: membrane organisers (TSPAN, tetraspanin); lipids (LBPA, lyso-bis-phosphatidyl acid); cell adhesion molecules (ICAM, intracellular adhesion molecule); intracellular trafficking proteins; cell-type-specific proteins (MHC, major histocompatibility complex; APP, amyloid precursor protein; PMEL, premelanosome protein; TCR, T cell receptor; CXCR4, CXC-chemokine receptor 4; HSPG, heparan sulfate proteoglycan; PrP, prion protein; TFR, transferrin receptor); enzymes (GAPDH, glyceraldehyde-3-phosphate, dehydrogenase); signal transduction proteins; biogenesis factors (ALIX, ALG-2 interacting protein X; TSG101, tumor susceptibility gene 101 protein; VPS, vacuolar protein sorting-associated protein); chaperones (HSP70-heat shock protein 70 kDa); nucleic acids. Figure created with BioRender.com.

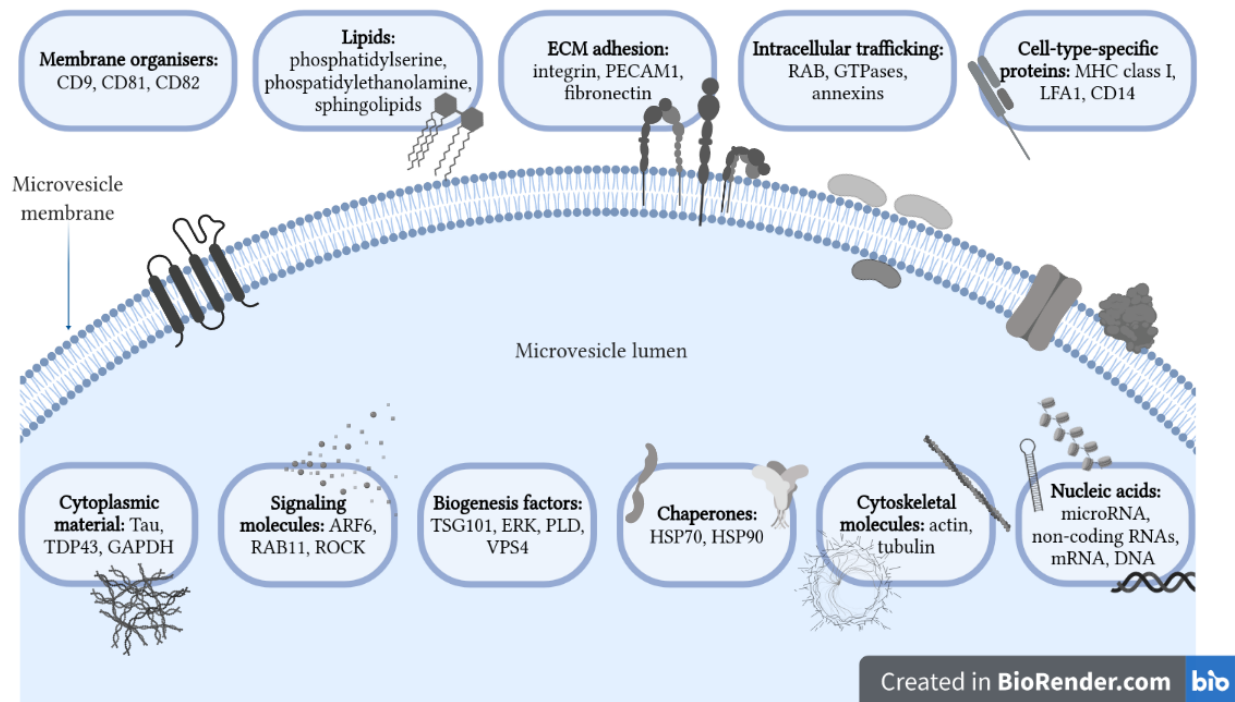


Figure 3.3: | **Ectosomes molecular content.** Ectosomes cargo includes for example: membrane organisers; lipids; extracellular matrix (ECM) adhesion molecules (PECAM1, platelet endothelial cell adhesion molecule); cell-type-specific proteins (MHC, major histocompatibility complex; LFA1, lymphocyte function-associated antigen 1); cytoplasmic material (TDP43, TAR DNA-binding protein 43; GAPDH, glyceraldehyde-3-phosphate); signaling molecules (ARF6, ADP-ribosylation factor 6; ROCK, RHO-associated protein kinase); biogenesis factors (TSG101, tumor susceptibility gene 101 protein; ERK, extracellular signal-regulated kinase, PLD, phospholipase D; VPS, vacuolar protein sorting-associated protein); chaperones (HSP90, heat shock 90 kDa protein); cytoskeletal molecules and nucleic acids. Figure created with BioRender.com.

3.3 Extracellular vesicles in cancer

The processes of carcinogenesis and shaping the TME (known as “hallmarks of cancer”) are inextricably linked with intercellular communication. Extracellular vesicles, as the main regulatory entities transporting tumor-specific functional cargo in cell-to-cell crosstalk, are a key to understanding the contributions behind cancer growth, invasion and metastasis.

There are about 10^{10} EVs per mL in the circulation in cancer patients (twice as many as in a healthy population) [91]. The mechanism of increased vesicle secretion in neoplastic diseases is not fully understood, but scientists are looking for the cause in altered cell physiology. The proposed hypothesis explaining this phenomenon assumes that it is mediated by reduction of pH in TME, which is a factor stimulating the production of EVs [92–94]. Tumor-derived exosomes demonstrate preference for certain organs which are potential pre-metastatic niches [20]. This phenomenon is called organ tropism. Metastasis is predetermined by bioactive components of EVs including oncoproteins, oncopeptides or nucleic acids which promotes inflammation, immune suppression, invasion, cell migration and proliferation, angiogenesis vascular leakiness, lymphogenesis, coagulation, and stromal cell activation [20, 79, 80].

Cancer EVs molecular content is different than in normal cells. Enriched biological components, such as cancer-related proteins or nucleic acids, were detected in the tumor EVs compared to their origin cells. Additionally, preliminary results in human cancer shows many highly enriched miRNAs ($>1\,000$ Transcripts Per Kilobase Million- TPM) in extracellular vesicles (EXs and MVs) and undetectable in donor cancer cells [20]. Data from these and other currently conducted experiments suggest that extracellular vesicles cargo (especially miRNAs and proteins) can be one of the most promising and unique signatures for use in cancer diagnostics. Of particular interest is approach combining information about different types of EVs biomolecules like RNA, microRNA, proteins, or lipids. A multicomponent analysis of EVs and origin/target cells would allow to explore and interpret biological interrelationships in TME or pre-metastatic niche formation.

3.3.1 EVs in melanoma

To date, several mechanisms involved in the biogenesis, release and cell-cell transmission of EVs in melanoma have been described. The trafficking of EVs by melanoma cells is a low-pH dependent process [92]. Lowering of pH in the melanoma TME has been shown to increase the release and uptake of EVs and the enhanced uptake of cargo through EV fusion. The fusion efficiency of melanoma EVs was higher in cells of metastatic origin than in cells derived from primary tumors or control cells [95]. In addition, caveolin-1 (a protein involved in melanoma progression) was found to be highly delivered by EVs released in the acidic TME.

Melanoma EVs show organ tropism to the brain, lungs, liver, or bones. Studies showed that melanoma-derived exosomes in mice during formation of a pre-metastatic niche distribute to the organs that mirror the most common metastatic sites in human melanoma (Figure 3.4) [20].

In mouse melanoma model, based on the B16F10 cell line, exosomes are involved in pre-metastatic niche development in lungs [96]. Tumor exosomes contained MET Proto-Oncogene (type of RTK hepatocyte growth factor receptor) were transferred to bone marrow-derived cells (BMDCs) [96]. EXs permanently "educate" BMDCs via MET receptor. Then, the mobilization of MET+ BMDCs into the circulation was observed. As a result, melanoma-derived exosomes induced ECM remodelling, tissue inflammation (increased levels of proteins S100A8 and S100A9), and vascular leakiness at pre-metastatic sites in lungs. Also in human clinical trials it was observed that patients with stage 3 and 4 melanoma had a significantly increased amount of melanoma cell-derived exosomes (MTEX) containing MET and phosphorylated MET (Tyr1349) compared to the healthy donors (Figure 3.4) [96]. MTEX suppress functions of T cell receptor (TCR) and immune cells like T cells or NK cells in melanoma patients. MTEX are also enriched in immunosuppressive proteins and are currently indicated as the primary mechanism of melanoma-induced immunosuppression. This may be one of the key reasons that some patients do not benefit from cancer immunotherapy [97]. Exosomes also contribute to the resistance of human melanoma to chemotherapy. They have been reported to reduce cytotoxic drug activity (Cisplatin, CisPt) by lowering the tumor pH and by removing the drug from melanoma cells via MTEX [98].

3.4 EVs as melanoma biomarker

In 2016 U.S. Food and Drug Administration (FDA) and the National Institutes of Health (NIH) Biomarker Working Group defined biomarker as “characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions” [99]. They described 7 categories of biomarkers: 1) diagnostics, 2) monitoring, 3) pharmacodynamic/ response, 4) predictive, 5) prognostic, 6) safety and 7) susceptibility/ risk biomarker. According to these regulations EVs can be considered as diagnostic or prognostic biomarkers [22].

EVs have been detected in all body fluids including blood, urine, breast milk, bile, synovial, lacrimal, seminal, ascites, and bronchoalveolar lavage fluids. Human metastatic melanoma cells secrete a higher amount of EXs than primary ones [95]. One of emerging application of this findings is the use EVs as cancer biomarkers. Strategies based on isolation of molecular markers including tumor-derived microvesicles from blood or other biofluid are proposed as powerful, noninvasive diagnostic tool and called ‘liquid biopsy’ (Figure 3.5).

Enriched miRNAs were identified in EVs, including hsa-miR-214-3p, hsa-miR-199a-3p, hsa-miR-155-5p, hsa-miR-142-3p, hsa-miR-150-5p, hsa-miR-223-3p and hsa-miR-486-5p. All this bioactive molecules are associated with melanoma progression [100]. In addition, miR-199a and miR-214 are involved in TWIST1 (twist transcription factor 1) gene regulation. TWIST1 expression is high in melanoma cell lines generated from pre-metastatic tumors [100]. TWIST1 enhances melanoma cells migratory and invasive functions. MTEX with elevated levels of proteins like housekeeping CD63 and melanoma-associated caveolin-1 have also been identified in the blood of melanoma patients [101].

Melanoma EVs present immunoregulatory components like programmed death-ligand 1 (PD-L1) and Fas ligand (FasL). PD-L1 on MTEX contributes to immunosuppression by inhibiting the function of CD8⁺ T cells and facilitating tumor progression [97]. FasL on melanoma-derived MVs contributes to T cells apoptosis [102].

Databases cataloging data and metadata from available scientific studies increasingly help in the exploration of melanoma EVs cargo diversity. The largest databases are: ExoCarta (www.exocarta.org) [103], Vesiclepedia (www.microvesicles.org) [104], EVpedia (www.evpedia.info)

; this database is no longer supported and updated) [105], miRandola (mirandola.iit.cnr.it) [106] and exRNA Atlas (www.exrna-atlas.org/) [107].

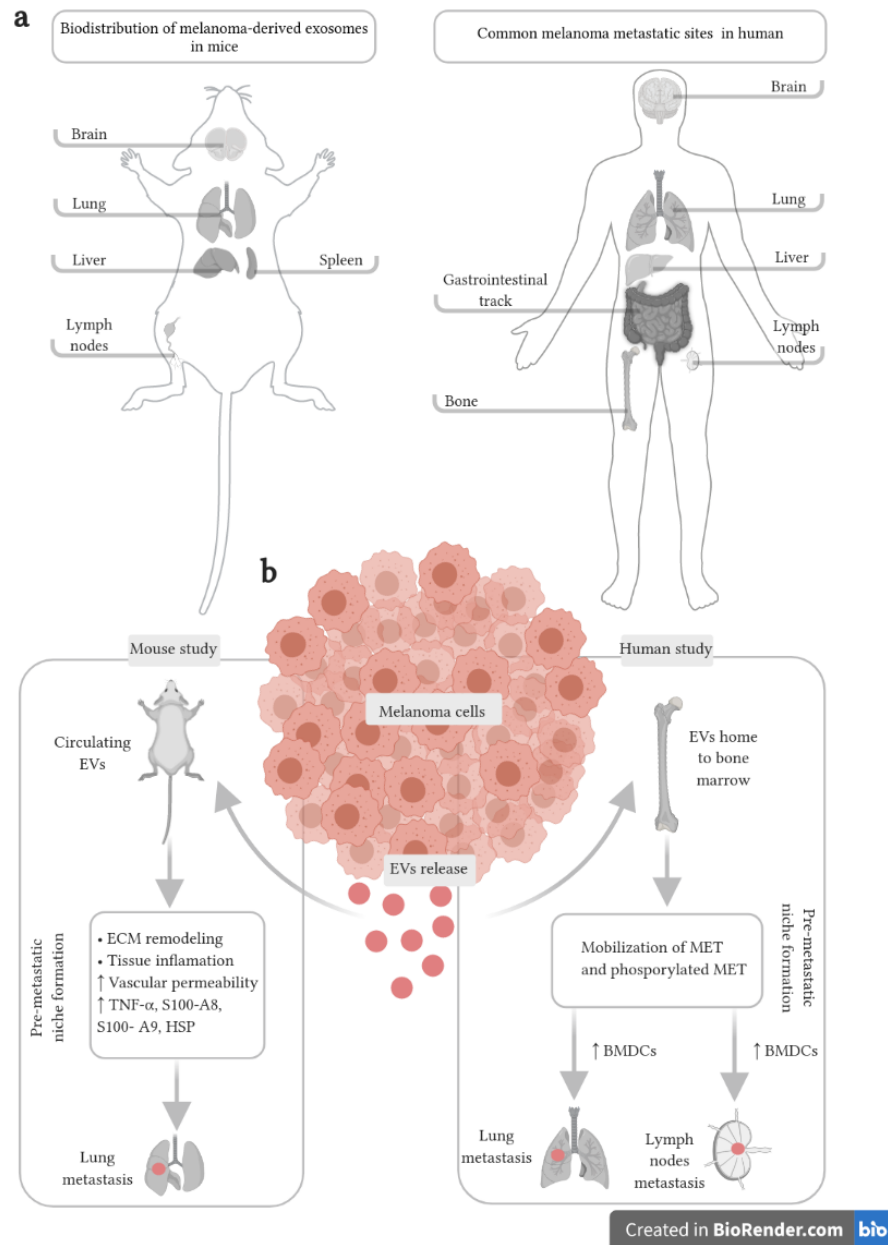


Figure 3.4: | **Key findings supporting a role for extracellular vesicles in melanoma.**

a| Systemic biodistribution of melanoma-derived exosomes in mice during formation of a pre-metastatic niche reflects most common metastatic sites in human melanoma [20]. b|

Mouse study: Circulating melanoma cell-derived exosomes (MTEX) home to lungs and induce extracellular matrix (ECM) remodeling, inflammation (higher expression of inflammatory genes: tumor necrosis factor- TNF, chemoattractants- S100 or heat shock protein- HSP).

Human study: MTEXs home to bone marrow and mobilize MET⁺ bone marrow-derived cells (BMDCs) which home to lungs and lymph nodes. Figure created with BioRender.com.

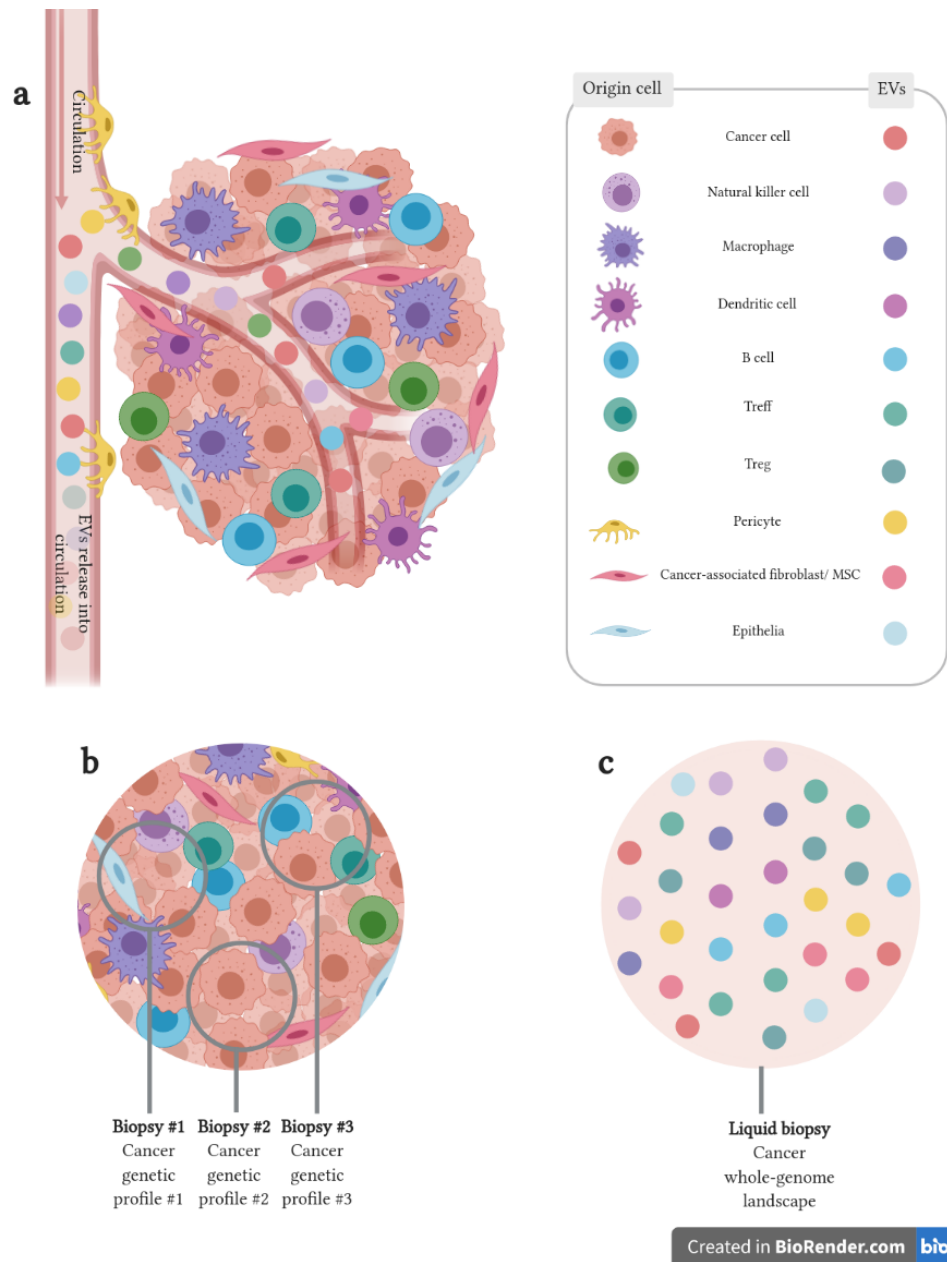


Figure 3.5: | **Extracellular vesicles in cancer.** a| Tumor microenvironment (TME) includes cancer cells with different genetic profiles, normal cells (for example immune cells) and blood vessels. All TME cells secrete extracellular vesicles (EVs) into circulation. b| The diverse environment of a cancer has advantages in terms of performing a complete tumor characterization based on traditional biopsy. The biopsy cannot provide a complete landscape of the cancer-related genetic signatures. c| Liquid biopsy diagnostics based on EVs is emerging as one of the most powerful tools to be used in the clinic, promising a complete analysis of tumor heterogeneity. Figure created with BioRender.com.

MATERIALS AND METHODS

Chapter 4

Sparse multiset CCA computational implementation

Based on the literature study described in the Theoretical Background Chapter 2, I chose sparse CCA as the optimal and best-validated method for biomarkers research in cancer (Figure 2.3). Additionally, as a result of the sCCA analysis, I obtain canonical variables only for the most important observables, the rest of the results are given as zero. This allows to eliminate data "noise" and focus attention only on the most significant outputs of the performed procedure. That's the reason why correlation variables of the sCCA are easier to interpret compared to other CCA extensions. It has been also shown in a comparative (benchmark) study [108]. This is important in research on biological systems, characterized by high complexity, which makes interpretation very challenging.

In my bioinformatic pipeline, I used sparse multiset CCA implementation called 'sCCA' (see Appendix .1) [59]. 'sCCA' is built on the basis of the algorithm proposed by Witten et al. in 2009 [37]. The bioinformatics implementation of this tool was made in the R programming language [65].

A detailed description of the derivation of the algorithm can be found in the original publication [37]. Sparse CCA is defined as:

$$\begin{aligned}
(\mathbf{a}, \mathbf{b}) &= \arg \min_{\mathbf{a}, \mathbf{b}} -\text{corr}(\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b}) + \tau_a \|\mathbf{a}\|_1 + \tau_b \|\mathbf{b}\|_1 \\
&\text{subject to } \|a\|_2 \leq 1; \|b\|_2 \leq 1,
\end{aligned} \tag{4.1}$$

where τ_a and τ_b are penalty constants, the values that control the sparsity level of obtained canonical vectors. Symbol $\|\cdot\|_p$ denotes p-norm which is defined as $\|v\|_p = (\sum_i |x_i|^p)^{1/p}$. $\|v\|_2 = \sqrt{\sum_i v_i^2}$ is standard Euclidean norm and $\|v\|_1 = \sum_i |v_i|$ is called lasso penalty. It is the application of lasso penalty that leads to sparsity of the obtained solution.

Unlike the standard CCA, the presented equation doesn't have closed form solution. However, this equation is bi-convex with respect to weight vectors. For such a case, the solution can be obtained by an iterative algorithm. At each step we can fix values of one of the weight vector, which make equation 4.1 strictly convex and solvable for the second weight vector.

When we want to get sparse CCA for more than 2 datasets, the equation 4.1 is defined as:

$$(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M) = \arg \min_{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M} - \sum_{i < j} \text{corr}(\mathbf{X}_i \mathbf{w}_i, \mathbf{X}_j \mathbf{w}_j) + \sum_m^M \tau_i \|\mathbf{w}_m\|_1, \tag{4.2}$$

where $\mathbf{w}_i$, $\mathbf{X}_i$, τ_i are weight vector, data matrix and penalty are constant for the i-th dataset, respectively. This equation is multi-convex, so it can be solved in similar fashion as 4.1 with the difference that at each iteration we fix all weight vector with exception of one.

Prepared data matrices should have samples on rows and the biological features on columns. 'sCCA' solution requires three input matrices representing different sets of features for the same set of melanoma cell lines (Figure 6.1: a), so datasets had the same number of rows, but different numbers of columns. In presented analysis I used:

1. The first input data: microvesicle miRNA transcripts,
2. The second input data: microvesicle proteins,
3. The third input data: cell totalRNA transcripts.

Chapter 5

Multi-omics melanoma datasets

The experiments and analyses were performed on cutaneous melanoma. Melanoma is a model tumor in research on cancer immunology, in particular in the context of disease progression, which was the key for my research hypothesis [109]. Melanoma is also a cancer type that has not been previously described in the context of sparse multiset CCA (Figure 2.3). I used four melanoma cell lines: WM115, WM266-4, WM793 and WM1205Lu. WM115 and WM266-4 were established from primary and metastatic lesions from the same patient, representing primary vertical growth phase and lymph node metastasis vertical growth phase, respectively [110]. WM793 was obtained from primary lesion from a patient, representing primary vertical growth phase and the corresponding metastatic vertical growth phase variant of this cell line was isolated from nude mice lung metastases and was designated as WM1205Lu [111]. Additional information on the used cell line models is gathered in the Table 5.1.

The cBio Cancer Genomics Portal data resources for cutaneous melanoma was used to analyze the genetic background of the cell lines in relation to the melanoma patients population (see Appendix .3 and Table 5.1) [112, 113].

In order to obtain the broadest possible biological context of the studied phenomenon, various types of data were used to generate multi-omics melanoma dataset. I integrated proprietary microvesicle miRNA transcriptomics with publicly available microvesicle proteomics and cell totalRNA transcriptomics collected for the same melanoma cell lines. The use of standardized cell lines allowed for the consistency and comparability of the genetic profiles.

Table 5.1: Melanoma cell lines characteristics: WM115, WM266-4, WM793 and WM1205Lu.

Cell Line Information				
Name	WM115	WM266-4	WM793	WM1205Lu
Type	primary	metastatic	primary	metastatic
Originated from	skin	lymph node	skin	lung
Patient Clinical Information				
Cell line sex	Female		Male	- (mouse model)
Age of sampling	55		37	- (mouse model)
Important Mutated Genes				
cBioPortal	BRAF		BRAF	
Cell Model Passports	CDKN2A	BRAF	CDK4	BRAF
	PTEN	CDKN2A	PTEN	CDK4
	PIK3R1			

5.1 Microvesicles miRNA data

Laboratory work necessary to generate the microvesicle miRNA data was carried out in the Extracellular Vesicles Group in Department of Medical Physics, Marian Smoluchowski Institute of Physics at Jagiellonian University and in cooperation with the Department of Glycoconjugate Biochemistry, Institute of Zoology and Biomedical Research at Jagiellonian University. Next-generation sequencing was carried out by Intelliseq Ltd. based in Cracow.

In this thesis, a new method of miRNA isolation from microvesicles, which uses centrifugation and Hydrostatic Filtration Dialysis (HFD), has been proposed. This method is based on protocols and experiments developed for the isolation of EVs totalRNA and proteins in the group of Professor Stępień and Professor Przybyło [77, 80, 114, 115].

5.1.1 Cell culture

Cell lines originate from the European Searchable Tumour Cell Line and Data Bank (ESTDAB)-A Collection of Immunologically Characterised Melanoma Cell Lines and Databank (Tübingen, Germany). Cell banks of WM115, WM266-4, WM793, WM1205Lu are divided and stored in multiple, validated, liquid nitrogen dewars. Cells were carefully retrieved from bank and transferred to cell culture medium composed of: RPMI 1640 GlutaMAX-I medium, with 10% foetal bovine serum (FBS), streptomycin (100 g/mL) and penicillin (100 unit/mL). The mixture was centrifuged in a Micro 22R, Hettich centrifuge for 5 min at 4 °C at 140 g. In this way, dimethylsulfoxide (DMSO), which protects cells during banking, was disposed of. The obtained melanoma cell pellets were suspended in culture medium and transferred to 75 cm² cell culture flasks with vented filter cap with a capacity of approx. 20 mL each. Cells were cultured under sterile conditions using a laminar chamber (BV-100, Telstar Life Science Solutions). Cells were grown in monolayers in a culture incubator (490-ICE, Lab-Line, Instruments Inc.) under stable conditions: 5% CO₂ atmosphere at 37 °C and passaged after reaching approximately 80-90% confluence.

5.1.2 Isolation of microvesicles

The culture medium collected from 40 culture flasks for each cell line was used for the isolation of miRNA. In order to isolate the microvesicles, a series of centrifugations was performed. First centrifugation at a value of 2.000 g for 10 min at room temperature (RT) to remove cells, cell fragments, organelles and apoptotic bodies. The resulting supernatant was centrifugated for second time at 4.500 g for 20 min at RT to clear residual precipitate. After completion, the supernatant was collected for further procedures.

In order to concentrate and purify the material, the Hydrostatic Filtration Dialysis (HFD) method was used, which is an alternative EVs isolation method [114,115]. It is based on filtration using a cellulose dialysis membrane with a large pore diameter with molecular weight cut-off (MWCO) 1,000 kDa (catalog number 131486, Spectrum Laboratories, California). Filtration is assisted by negative pressure which speeds up filtration process and minimizes changes in the sample over time. After the volume was reduced, the sample is washed with

deionized water and again reduced to a volume of few milliliters. Ultimately, about 3 ml of concentrated and purified EVs fraction are obtained from a sample with a volume of about 800 ml.

In the last step, the concentrate was centrifuged at 18 000 g for 1 hour at 4 °C (Optima MAX-XP, Beckman Coulter Life Sciences). Centrifugation separated microvesicles and exosomes. Both fractions were frozen at -80 °C. Only microvesicles were used for further research because their molecular cargo and function in cancer microenvironment is not yet well understood. The scheme of microvesicles isolation protocol is presented on Figure 5.1.

5.1.3 Isolation of microvesicles miRNA and library preparation

MiRNA was isolated from microvesicles according to the mirVanaTM miRNA Isolation Kit (ThermoFisher Scientific USA) instructions. MiRNA quality and quantity was assessed using the Nanodrop and Agilent Bioanalyzer. Samples were deep-frozen in liquid nitrogen.

5.1.4 NGS sequencing and raw data preprocessing

Samples were transported in a container intended for biological material, with care to maintain a constant temperature. Secured material was transferred to the external company Intelliseq Ltd., that performed the sequencing service.

Next generation sequencing of miRNA was performed by using Illumina HiSeqTM Sequencing System. Illumina products were the first NGS-based solutions in the world to be approved by the Food and Drug Administration (FDA) and validated for meeting stringent performance criteria in molecular diagnostics [116].

Quality control (QC) checking on raw sequence data was performed using FastQC software (version 0.11.08). Based on QC report low quality sequences and adapters (AGATCGGAA-GAGCACACGTCT) were filtered by TrimGalore (version 0.6.3). Mapping of clean reads to the Homo sapiens genome hg38 (version 97) were performed using Burrows-Wheeler Aligner-BWA (version 0.7.17). Quantification were carried out using HTSeq-count (version 0.11.2) with "-nonunique all" parameter and GFF file from miRBase database.

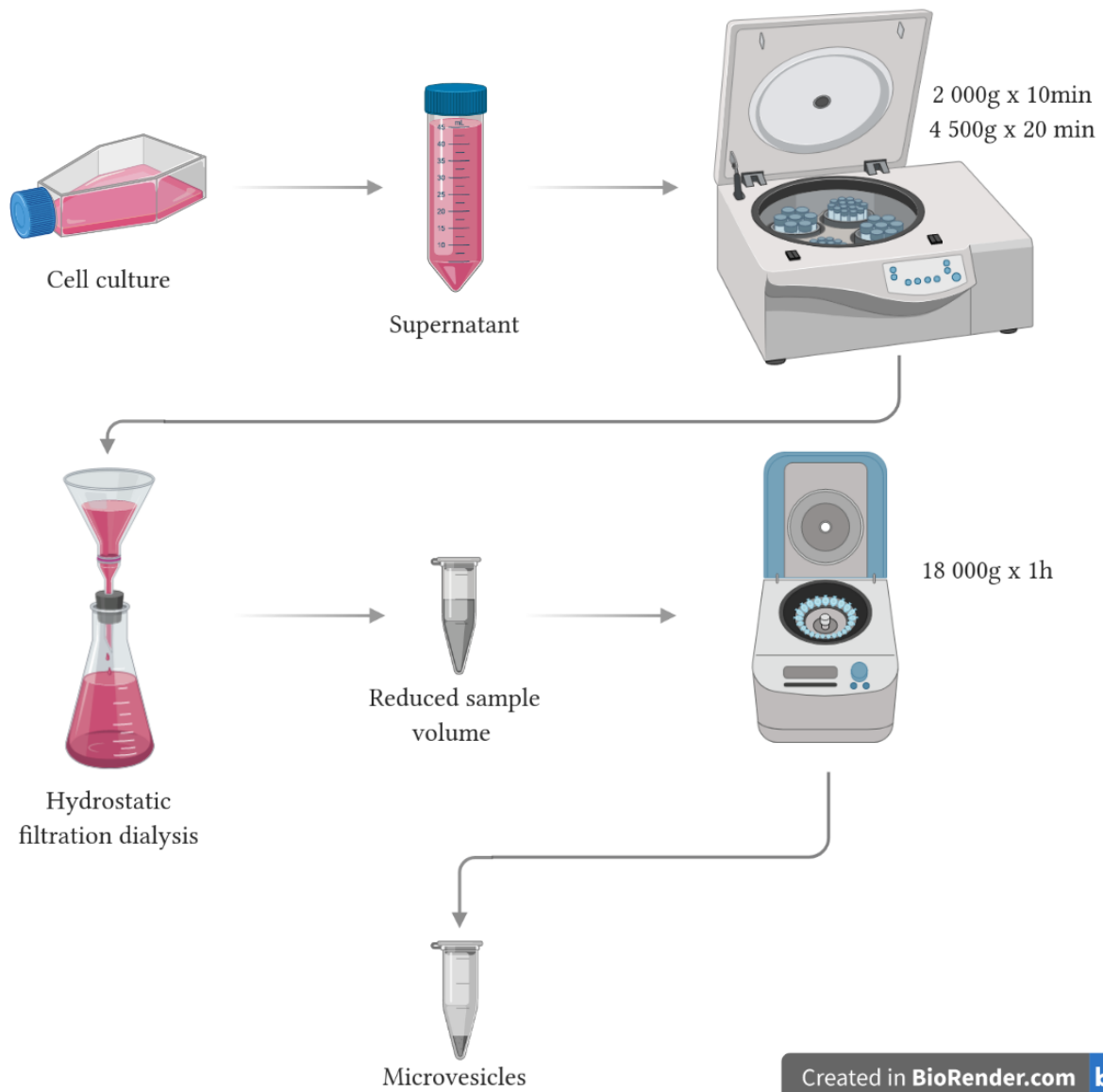


Figure 5.1: | Scheme of microvesicles isolation protocol. Figure created with BioRender.com.

MiRNA datasets are available in the Jagiellonian University Repository under the link: <https://ruj.uj.edu.pl/xmlui/handle/item/260896> [117].

5.2 Microvesicles proteomics data

The Mass Spectrometry proteomic data used in the analysis originated from the open access ProteomeXchange database and are available under accession number: PXD017366 (link: <http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PX017366>) and were generated in the collaboration of research teams of Professor Stępień and Professor Przybyło [80]. All laboratory methods, materials and protocols are consistent with those used to generate the miRNA data. The exact characterization of the data as well as the laboratory details are presented in the original publication by Surman et al. from 2020 [80].

5.3 Cell totalRNA data

The only data type that characterizes whole-cell molecular information is totalRNA transcriptome. The data was obtained from open access resources: for WM115 (link <https://cellmodelpassports.sanger.ac.uk/passports/SIDM01000>), WM266-4 (link <https://cellmodelpassports.sanger.ac.uk/passports/SIDM00979>) and WM793 (link <https://cellmodelpassports.sanger.ac.uk/passports/SIDM00973>) from Cell Model Passport and for WM1205Lu from Gene Expression Omnibus under accession number: GSM5012180 (link <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5012180>).

Chapter 6

Data integration

All datasets were downloaded from the databases in text format (Figure 6.1). Three-stage filtering was used for the miRNA and totalRNA data. In the first step, transcripts that were not expressed in more than three samples were filtered out. In second step, logarithm of fold change ($\log_2FC$) was calculated. The results that didn't meet the criteria $|\log_2FC| > 2$ were filtered out. The last step was to filter the results below the count value of 20. The cut-off point used was derived from the data distribution.

Prepared miRNA, proteomics and totalRNA data was saved to files in comma-separated values (CSV) text format, creating multi-omics input for the biomarker discovery bioinformatics pipeline.

6.1 Sparse multiset CCA biomarkers discovery pipeline

Based on the prepared sparse multiset CCA implementation (Chapter 4) and integrated multi-omics data (Chapter 5), the melanoma biomarker discovery pipeline was constructed. An overview of the method is presented in the Figure 6.1. The novelties of this procedure are:

1. use of three melanoma data matrices instead of two,
2. intergration of data not only from whole cells but also from microvesicles,
3. harnessing sparse multiset CCA results and evidences from EVs and cancer dedicated

Table 6.1: List of input datasets used in sparse multiset CCA: miRNA, proteins and total-RNA.

	Cell line	Database	Year	Reference
miRNA	WM115	Jagiellonian University Repository	2021	[117]
	WM244-6			
	WM793			
	WM1205Lu			
proteins	WM115	ProteomeXchange	2020	[80]
	WM244-6			
	WM793			
	WM1205Lu			
totalRNA	WM115	Cell Model Passport	2018	[118]
	WM244-6			
	WM793			
	WM1205Lu	GEO	2021	[119]

databases to generate biomarkers candidates list with functional and mechanistic hypotheses.

The result of the method was a list of the most significant biomarker candidates along with biological insights that allow the generation of functional and mechanistic hypotheses (Figure 6.1 c). Significant pathways were detected using interaction databases and bioinformatics tools (like STRING database and TarBase: mirPath). The evidence supporting the findings was obtained from databases dedicated to EVs (like ExoCarta, exoRBase or exRNA Atlas) and neoplasms (like cBioPortal, DepMap or TCGA). The full list of databases is included in the Appendix Table 2. The most important biological processes were visualized with BioRender.com and Cytoscape applications: EnrichmentMap and AutoAnnotate.

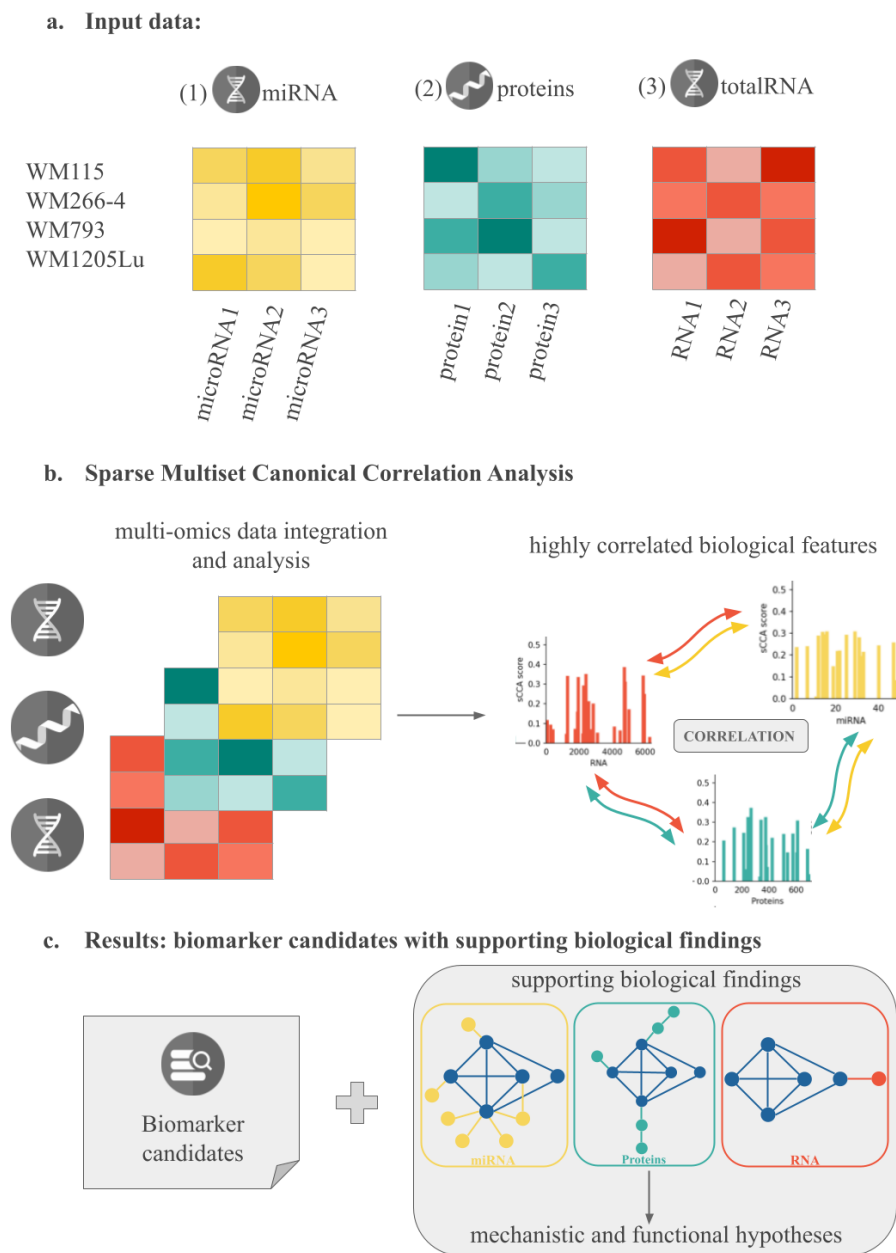


Figure 6.1: | **Sparse multiset canonical correlation analysis- based biomarker discovery workflow.** |a. Sparse multiset CCA requires three input datasets: (1) microvesicle miRNA transcriptomics, (2) microvesicle proteomics, (3) cell totalRNA transcriptomics. |b. Results of sparse multiset CCA are visualized on a histogram plot, where x axis corresponds to miRNA, proteins and totalRNA; y axis represents normalized canonical weights form 0 to 1. |c. Sparse multiset CCA identifies biomarkers candidates. To validate results, workflow is supported by enriched pathway analysis and evidences for each biomarker candidate. Based on that biological findings, mechanistic and functional hypotheses are constructed. Figure created with BioRender.com.

RESULTS

Chapter 7

Datasets quality control and preprocessing

An efficient isolation generated a pure fraction of short RNA fragments. The concentration measured using Nanodrop and Agilent Bioanalyzer of microvesicles samples (on average 35 $ng/\mu l$ per sample) were assessed as sufficient for sequencing. Next-generation sequencing was successful and resulted in obtaining proprietary microvesicles miRNA data for the studied melanoma cell lines. Combined with microvesicles proteomic data and cell total transcriptome from databases, a unique multi-omic set was obtained that was used in a sparse multiset CCA analysis. This is the first such dataset for melanoma with molecular information from whole cells and EVs for three different biological omics signatures (based on the literature analysis summarized in the Figure 2.3).

7.1 Melanoma cell lines genetic background

In parallel to the sparse multiset CCA pipeline, an analysis of the genetic background of the used melanoma cell lines was performed. The results allowed for identification of key signatures involved in oncogenesis in WM115, WM266-4, WM793 and WM1205Lu. This is particularly important for the generated functional and mechanistic hypotheses, because general molecular profile is the key to interpret the sparse multiset CCA outputs. This approach can decode complex biological insights and lay the groundwork for further validation

of biomarker candidates in a wet laboratory.

Based on analysis described in Appendices .3, the most important oncogenic mutations in melanoma cell lines WM115, WM266-4, WM793 and WM1205Lu include (Tables 7.1 a and b):

- **BRAF, missence mutation** form V600D in WM115 and WM266-4 plus form V600E in WM115, WM266-4, WM793 and WM1205Lu,
- **CDKN2A, deep deletion** in WM115 and WM266-4,
- **PTEN, deep deletion** in WM266-4 and PTEN, nonsense in WM793 form W274 ,
- **CDK4, missence mutation** in WM793 and WM1205Lu form K22Q,
- **PIK3R1, nonsense mutation** in WM115 form R642 .

For WM115 primary melanoma cell line and WM266-4 metastatic melanoma cell line (Table 7.1 a), the most important mutations and at the same time one of the most common in the patients population (approximately 50 %) are the BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase) missence mutations [113]. Identified BRAF gene perturbations are characterized by a amino acid change from valine (V) to glutamic acid (E) or aspartic acid (D) at codon 600. For WM115 only, PTEN (Phosphatase and Tensin Homolog) deep deletion and PI3KR1 (Phosphoinositide-3-Kinase Regulatory Subunit 1) nonsense form R642* have been detected. In this case, tumor development was very much dependent on these genetic factors, as it has been proven that BRAF V600E mutation in conjunction with the loss of the inhibitor of PI3K-activation PTEN promotes the formation of melanoma [120]. Additionally, the CDKN2A (Cyclin Dependent Kinase Inhibitor 2A) deep deletion is associated with the invasiveness of melanoma tumors and the tendency to distant metastasis (e.g. to lymph nodes or brain) [121].

Oncogenic mutations: BRAF (form V600E) and CDK4 (Cyclin Dependent Kinase 4 form K22Q), were found in WM793 primary melanoma cell line and WM1205Lu metastatic melanoma cell line (Table 7.1 b). In addition, a nonsense mutation in PTEN form W274* has been identified as important in WM793 tumor progression processes.

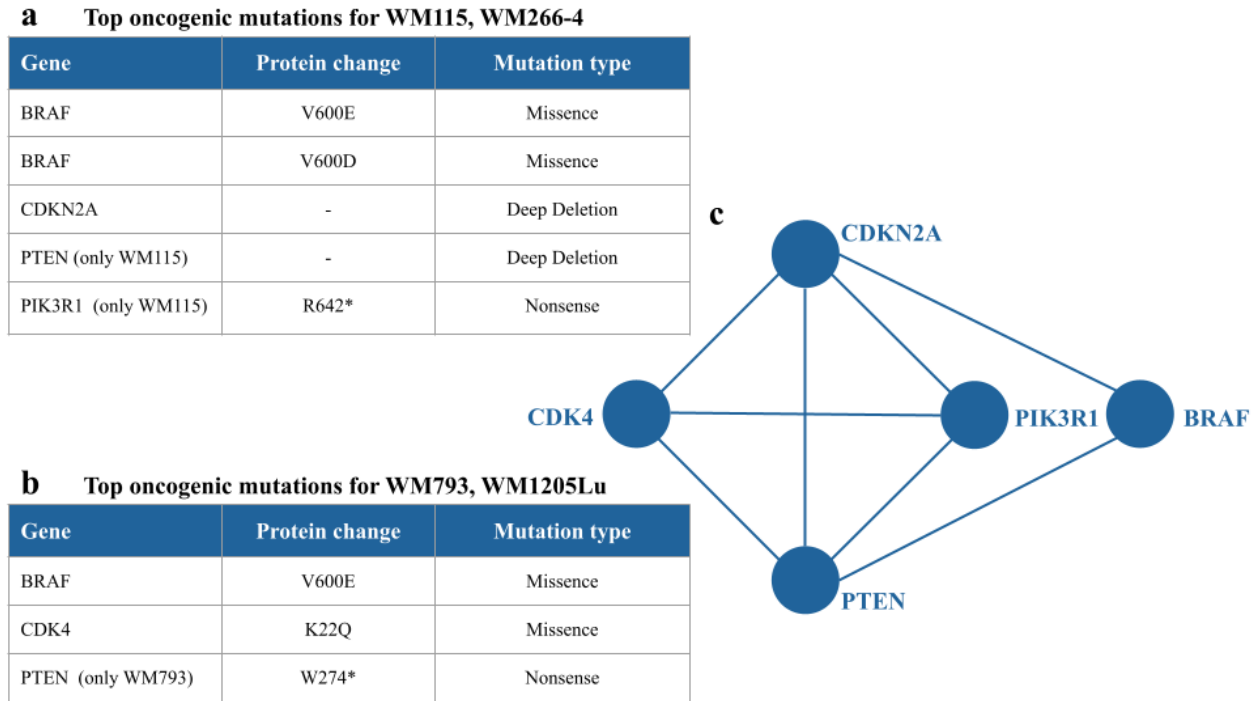


Figure 7.1: | **Top oncogenic mutations** for a| WM115, WM266-4 and b| WM793, WM1205Lu cell lines. c| Interactome of top oncogenic genes were visualized based on the STRING database analysis. Nodes corresponds to proetin coding genes and lines between them represent protein-protein interactions. Genes are connected into one network visualizing a broader biological context.

The interactions between proteins coded by selected top oncogenic genes were also investigated based on the protein-protein interaction (PPI) database STRING [122]. As a result, a network of connections was obtained, which is visualized in the Figure 7.1 c. From the interactome, it can be deduced that investigated genes are involved in common biological pathways. Most interactions were detected for PTEN and CDKN2A nodes.

7.2 miRNA, proteins, and totalRNA datasets preparation

In the first step, the miRNA data obtained after sequencing and totalRNA data downloaded from database were visualized without filtering at Figure 7.2 a and Figure 7.4 a. From the

graphs, it can be concluded that most of the counts are close to zero or very low values. After applying the filtering described in 'Materials and Methods' Chapter 6, the undesirable low counts were cut off, resulting in the distribution shown in the Figure 7.2 b and Figure 7.4 b. After this operation, box and violin plots are shifted towards more normal distributions which is desirable and guarantees increased data homogeneity. This is necessary to perform an effective CCA analysis which identifies "patterns" in the data and it is important to clear them of any "noise" or falsely amplified signals that may be misinterpreted by the algorithm. Proteins data had very good quality, therefore the filtering procedure was skipped (Figure 7.3).

The effect of the applied filtering is also presented in the form of a histogram, for easy and intuitive representation for the cut-off of low count values (Figure 7.5). In Figure 7.5 a and c it is noticeable how the distribution changes from a single dominant bar for values close to zero and a flattened tail, to a more homogeneous distribution.

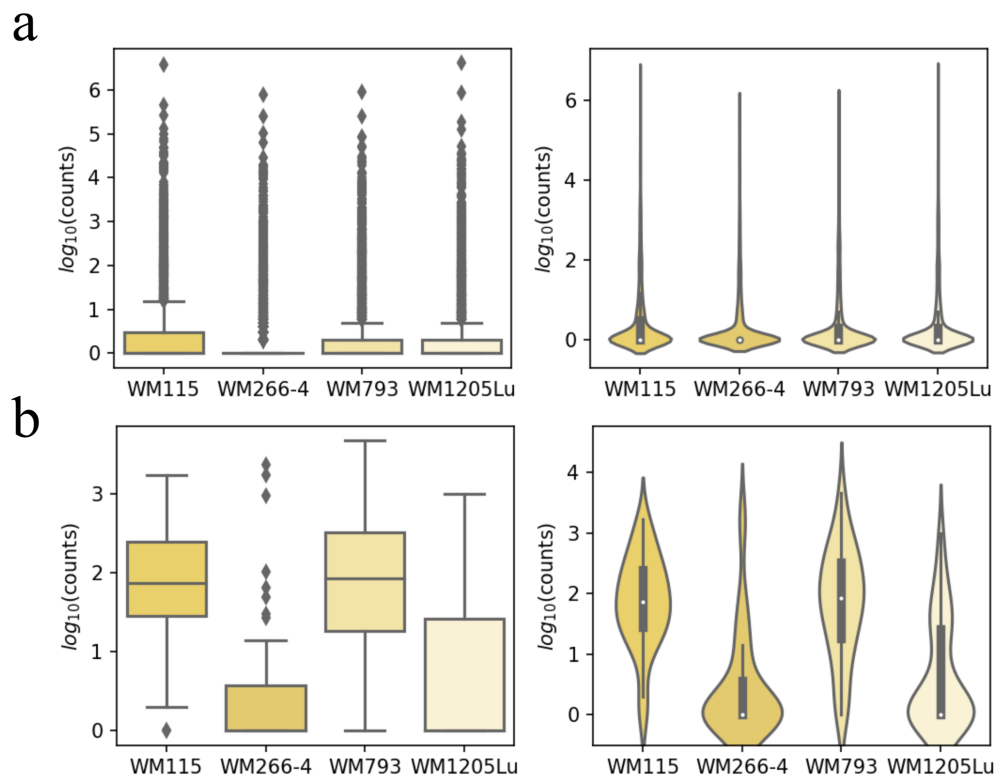


Figure 7.2: Box plot and violin plot for miRNA data a|(left column) before and b| (right column) after filtering. Vertical line inside box plot and white dot inside violin plot indicate the median values. Plot created with proprietary Python script using Mathplotlib version 3.4.2 and Seaborn version 0.11.1 libraries.

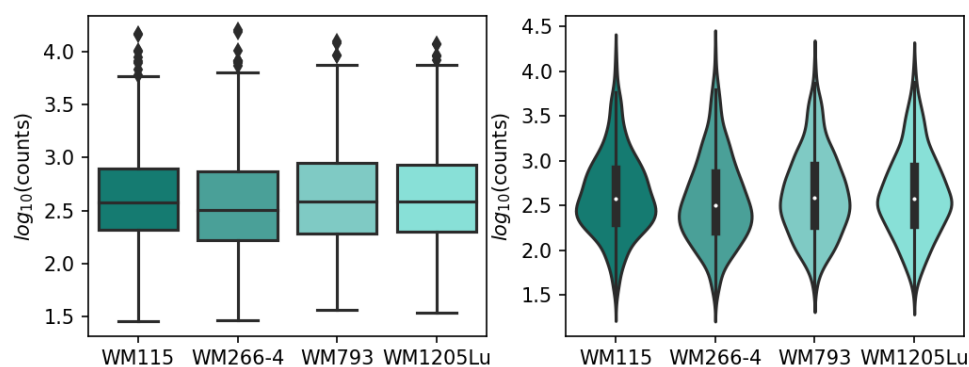


Figure 7.3: Box plot and violin plot for proteins raw data. Vertical line inside box plot and white dot inside violin plot indicate the median values. Plot created with proprietary Python script using Mathplotlib version 3.4.2 and Seaborn version 0.11.1 libraries.

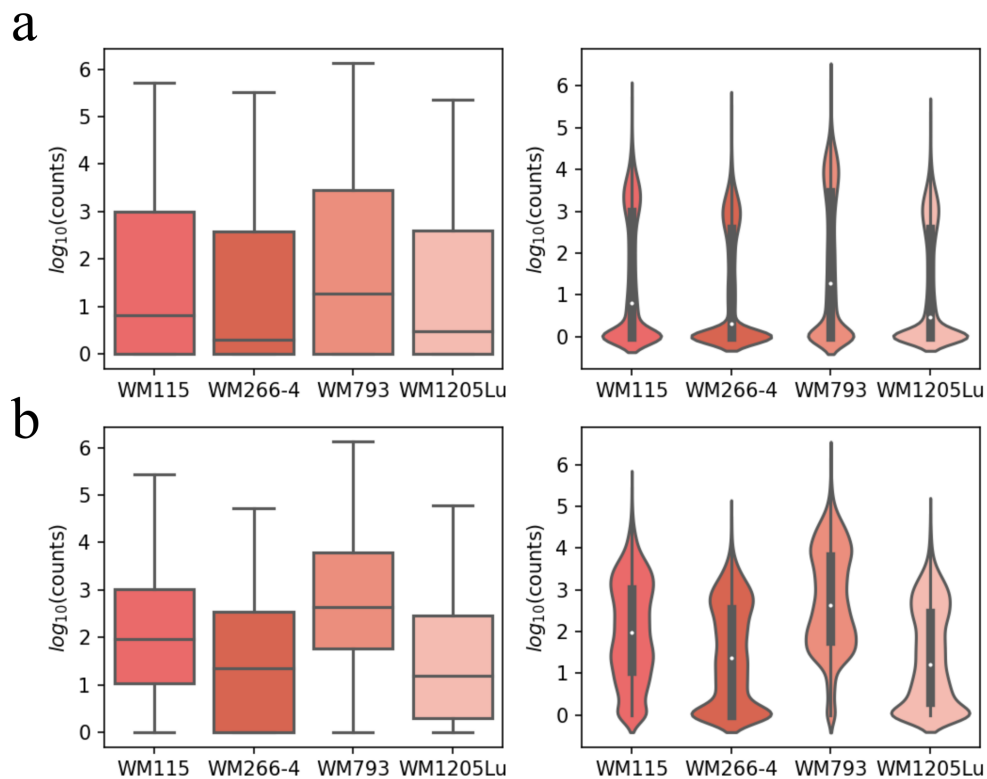


Figure 7.4: Box plot and violin plot for totalRNA data a|(left column) before and b| (right column) after filtering. Vertical line inside box plot and white dot inside violin plot indicate the median values. Plot created with proprietary Python script using Mathplotlib version 3.4.2 and Seaborn version 0.11.1 libraries.

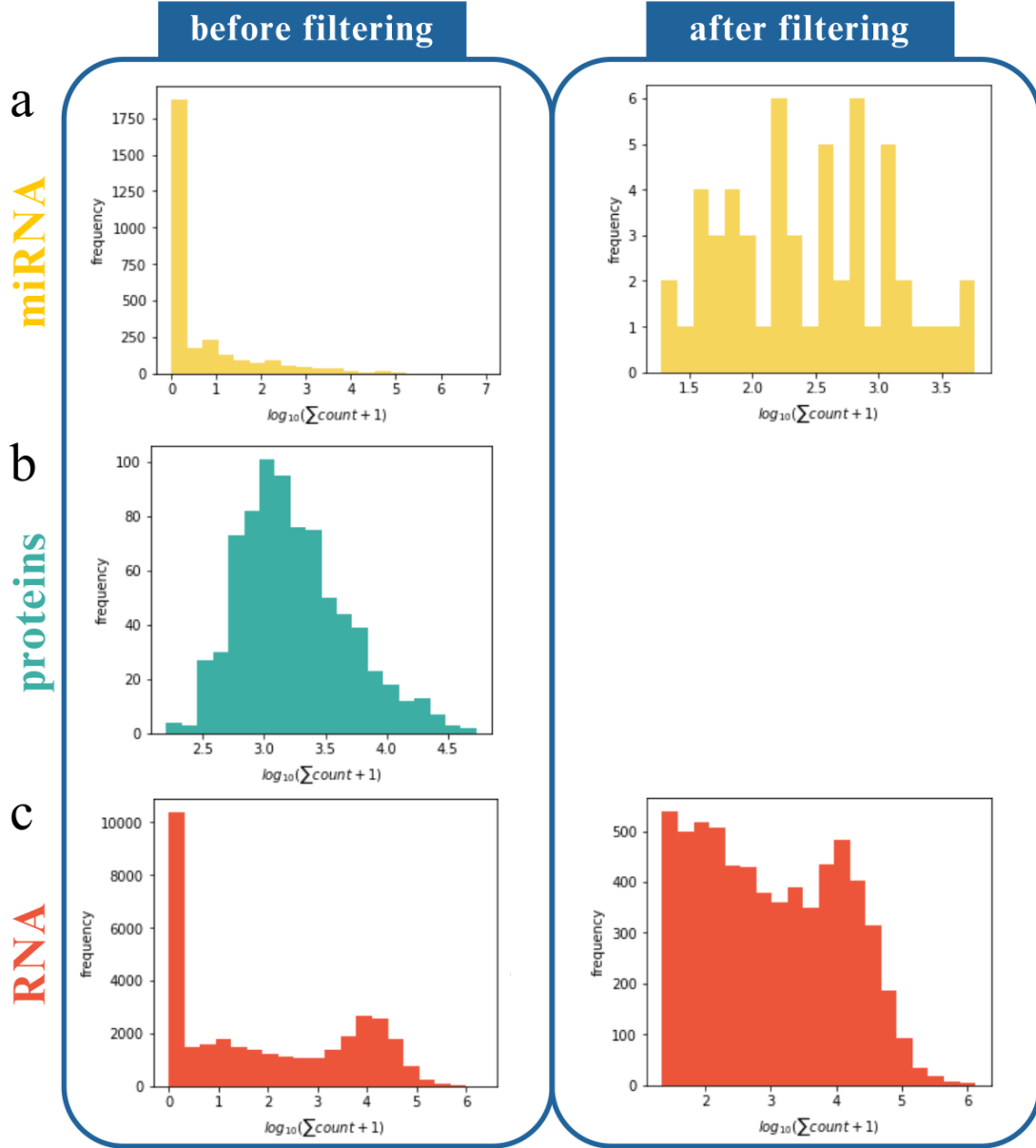


Figure 7.5: Histogram for logarithmic number of counts (sequence measurements) of a| miRNAs b| proteins and c| totalRNAs before (left plot) and after (right plot) filtering. For proteins there was no need to perform filtering due to the high quality of the data. Three-stage filtering was used, details of which are described in the Chapter 6 "Data integration". Plots created with proprietary Python script using Mathplotlib version 3.4.2 and Seaborn version 0.11.1 libraries.

Chapter 8

Sparse multiset CCA optimization

The penalty used in sparse multiset CCA is a regression analysis method to improve the accuracy of prediction and interpretation of the resulting biomarkers. In order to obtain properly selected biological signatures, it was necessary to identify the optimal penalty for presented CCA statistical model.

A number of tests were carried out. The best results were obtained for penalties below 10. The figure 8.1 shows the process of selecting an appropriate penalty value. A penalty of 7 was used on the upper panel, 4 on the medium and 2.5 on the lower panel on Figure 8.1. The charts evolve and the restrictive signature selection results in only the most significant biological features. A penalty of 2.5 was chosen as the a optimal for this analysis. After applying such an optimized sparse multiset CCA algorithm, 8 miRNAs, 7 proteins and 9 RNAs were obtained as biomarkers candidates. For the optimization steps where the lasso penalty was equal to 7 and 4, the results are presented in the tables in the Appendix .5.

As the value of the penalty is lowered, the canonical weights (CCA scores) increase their values, which effectively results in a stronger signal for the signatures sought. Only the most promising biological features receive a CCA score, the rest receive a score of zero. Combined with choosing the right penalty, the result is free from "distractions", which is clearly visible on the bottom panel on Figure 8.1. The prepared "noise-free" results are easy to biological interpret.

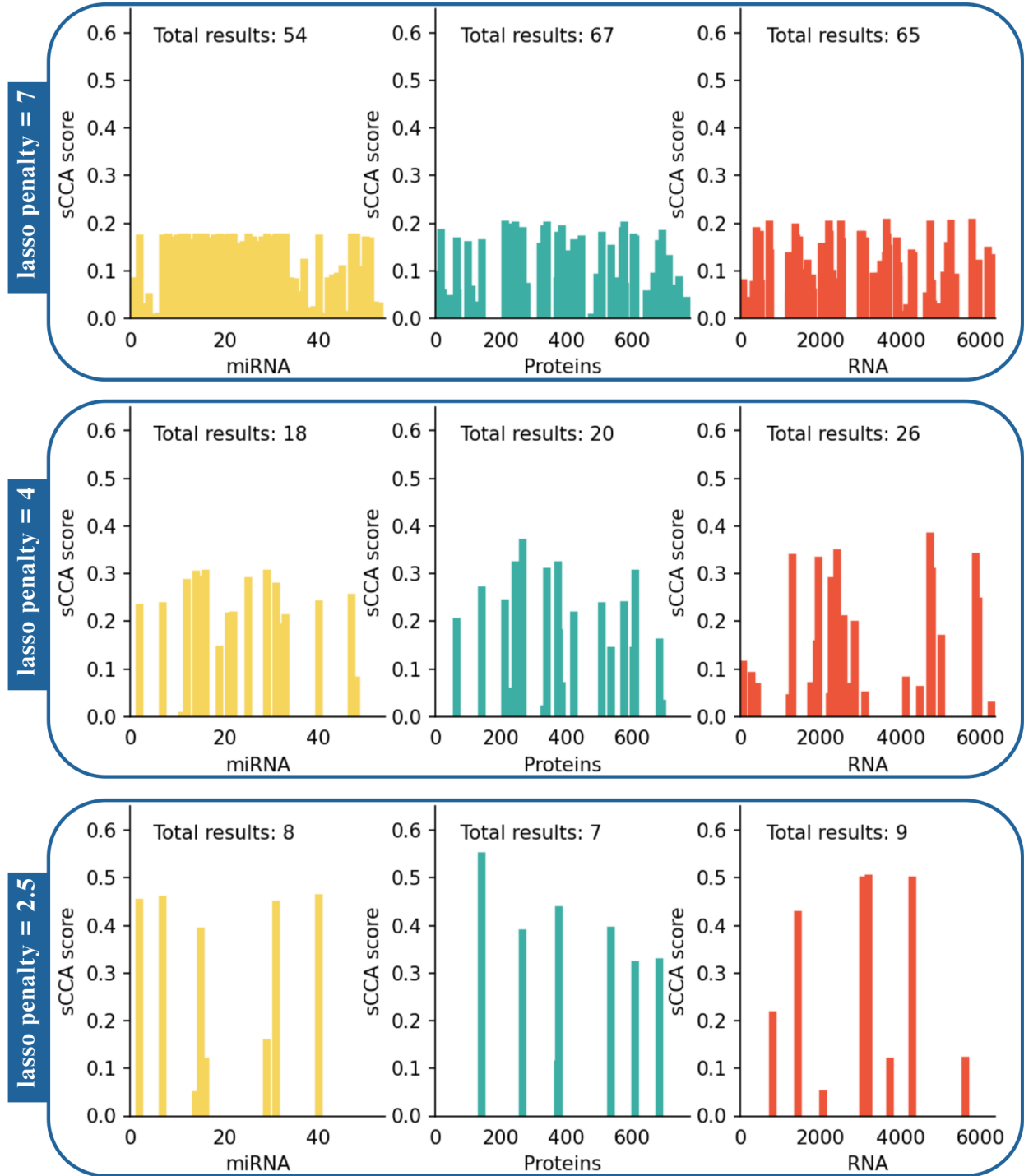


Figure 8.1: | **Sparse multiset CCA lasso penalty optimization.** X axes on presented plots correspond to number of miRNA, proteins and totalRNA and y axes to canonical weights called "sCCA scores". Plots created with proprietary Python script using Mathplotlib version 3.4.2 and Seaborn version 0.11.1 libraries.

Chapter 9

Sparse multiset CCA results

Sparse multiset CCA pipeline identified 24 biomarker candidates including 8 miRNAs, 7 proteins and 9 RNAs. A list of final result signatures is provided in Appendix .4. Results were listed in order of decreasing canonical weights (sCCA score). The higher the canonical weight, the higher the correlation the biomarker achieved (it is more significant in the context of the sparse multiset CCA). However, presented final molecular signatures are the result of the optimization of the sparse multiset CCA method, which assumed very restrictive rules (described in Section 8) in order to obtain a narrow group of biological features with the greatest impact. Therefore, it can be safely assumed that for the selected biomarkers, each one has a very high potential according to the applied bioinformatics procedure.

Among the results, top 13 gained canonical weights around 0.4, including:

- 5 miRNAs: hsa-miR-371a-5p (0.45751), hsa-miR-372-3p (0.45339), hsa-miR-224-5p (0.44775), hsa-miR-517c-3p (0.44331), hsa-miR-519b-3p (0.38675)
- 4 proteins: EFTUD2 (0.54569), ADRM1 (0.43178), APEX1 (0.39002), HNRNPD (0.38322)
- 4 RNAs: MIR181B1 (0.49818), MAP2K4P1 (0.49432), RAB-EGLN2 (0.494332), EFEMP1 (0.42259).

A detailed description of each category of biomarkers with biological evidences is provided in the sections below: miRNA in Section 9.1, proteins in Section 9.2, RNA in Section 9.3 and biological findings in Section 9.4.

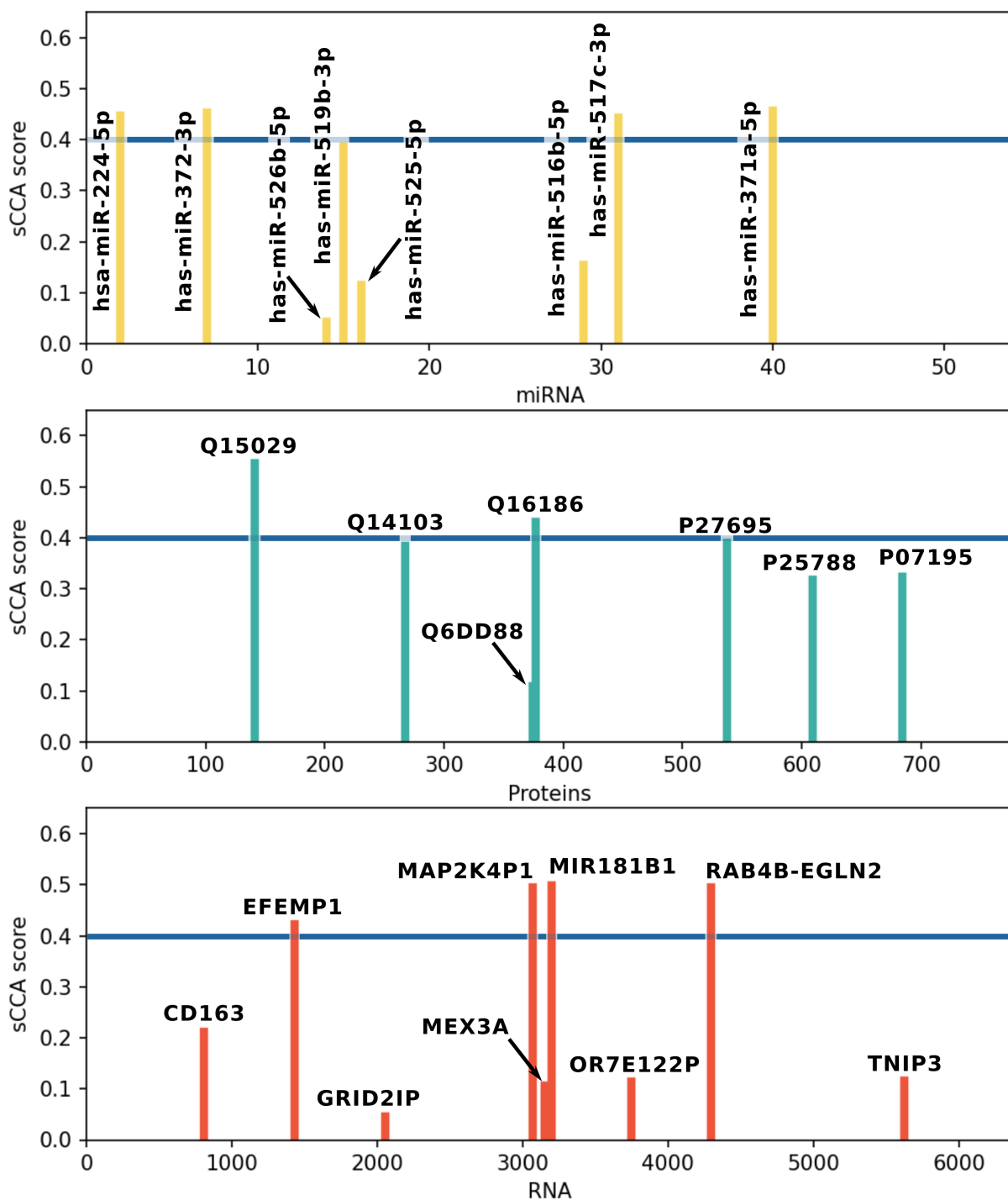


Figure 9.1: | Sparse multiset CCA results for melanoma biomarkers candidates. X axes correspond to miRNA, proteins, and totalRNA and y axes to canonical weights called "sCCA scores". The blue line marks 0.4 score for the best biomarkers. Detailed biomarkers annotations are described in Tables 9.1, 9.2, 9.3, and in Appendices Table 9.

9.1 Microvesicles miRNA biomarkers

Sparse multiset CCA pipeline identified 8 miRNA biomarker candidates presented in Table 9.1. Using TarBase: mirPath, the most important pathways influenced by the selected miRNA were investigated. The following melanoma related pathways have been identified for miRNAs in KEGG database [7]:

- hsa-miR-224-5p: melanoma pathway (KEGG id: hsa05218),
- hsa-miR-372-3p, hsa-miR-224-5p, hsa-miR-519b-3p and hsa-miR-516b-5p: microRNAs in cancer (KEGG id: hsa05206) or pathways in cancer (KEGG id: map05200),
- hsa-miR-317a-5p, hsa-miR-525-5p: TNF signaling pathway (KEGG id: map04668),
- hsa-miR-224-5p: p53 signaling pathway (KEGG id: map04115),
- hsa-miR-224-5p, hsa-miR-517c-3p and hsa-miR-519b-3p: glioma (KEGG id: map05214),
- hsa-miR-372-3p: MAPK signaling pathway (KEGG id: map04010).

The next step was to check whether the miRNA candidates for biomarkers were previously reported in the EVs molecular cargo. Based on Vesiclepedia, hsa-miR-317a-5p, hsa-miR-224-5p and hsa-miR-516b-5p were identified in exosomes [104].

Table 9.1: List of significant miRNAs identified by sparse multiset CCA.

miRBase accession	Mature sequence	Score
MIMAT0004687	hsa-miR-371a-5p	0.45751
MIMAT0000724	hsa-miR-372-3p	0.45339
MIMAT0000281	hsa-miR-224-5p	0.44775
MIMAT0002866	hsa-miR-517c-3p	0.44331
MIMAT0002837	hsa-miR-519b-3p	0.38675
MIMAT0002859	hsa-miR-516b-5p	0.15347
MIMAT0002838	hsa-miR-525-5p	0.11448
MIMAT0002835	hsa-miR-526b-5p	0.04333

9.2 Microvesicles proteomics biomarkers

Proposed sparse multiset CCA algorithm detected 7 proteins as promising biomarkers (Table 9.2). UniProt database reported APEX1, LDHB and ATL3 as melanoma related oncoproteins [6]. Additionally, PSMA3 was described in literature as significant protein in melanoma [123]. HNRNPD and PSMA3 were reported as important in oncogenesis in general [124–126].

According to Vesiclepedia database six of protein biomarker candidates (EFTUD2, ADRM1, APEX1, HNRNPD, LDHB and ATL3) are present in melanoma exosomes [104]. All of them are present in various types of cancers exosomes (for example in glioblastoma and different brain tumors).

Due to the large amount of evidence supporting the important role of selected proteins in tumorigenesis and the fact that the functional effects of proteins in TME are the most accessible and intuitive for biological interpretation, they are very promising biomarkers for further laboratory verification. It would be highly recommended to start a wet-laboratory verification of functional and mechanistic hypotheses from this group.

Table 9.2: List of significant proteins identified by sparse multiset CCA.

UniProt accession	Gene name	Score
Q15029	EFTUD2	0.54569
Q16186	ADRM1	0.43178
P27695	APEX1	0.39002
Q14103	HNRNPD	0.38322
P07195	LDHB	0.32278
P25788	PSMA3	0.31736
Q6DD88	ATL3	0.10916

9.3 Cell totalRNA biomarkers

Presented bioinformatics sparse multiset CCA pipeline selected 9 RNA biomarkers candidates (Table 9.3).

In this data category, the algorithm assigned the highest canonical weight (0.49818) to MIR181B1. MIR181B1 belongs to miR-181 family and is affiliated with KEGGs "microRNAs in cancer" pathways (KEGG identifier: hsa05206). MIR181B1 with MIR181A1 have already been described in the literature as an essential regulatory genes of PTEN/PI3K signaling pathway [127, 128].

EFEMP1, CD163, and TNIP3 are of vital in the development of cutaneous melanoma [129–131]. MEX3A is very important in other types of cancer development and progression, including glioblastoma [132, 133].

Based on the Vesiclepedia, 2 RNAs were identified in melanoma EVs: RAB4B and EFEMP1. In addition, 4 of them were detected in other types of cancer (RAB4B in brain cancer, EFEMP1 in glioblastoma, brain and lung cancers, CD163 in colorectal cancer and MEX3A in glioblastoma and colorectal cancer) and 2 in the exosomes of a healthy population (CD163 and MEX3A) [104].

Table 9.3: List of significant RNAs identified by spare multiset CCA.

Ensembl accession	Gene name	Score
ENSG00000207975	MIR181B1	0.49818
ENSG00000269904	MAP2K4P1	0.49432
ENSG00000171570	RAB4B-EGLN2	0.49432
ENSG00000115380	EFEMP1	0.42259
ENSG00000177575	CD163	0.21231
ENSG00000050730	TNIP3	0.11515
ENSG00000215160	OR7E122P	0.11325
ENSG00000254726	MEX3A	0.10581
ENSG00000215045	GRID2IP	0.04656

9.4 Biomarkers supporting biological findings

The process of generating functional and mechanistic hypotheses began with the analysis of individual biological relationships of selected biomarkers in context of melanoma cell lines genetic background. The following databases and bioinformatics resources were used to examine the interactions: STRING [122], TarBase: mirPath and g:Profiler [134]. The interactions were visualized for each group of biomarkers separately on Figure 9.2. In the next step, all interactions were mapped to one shared network (Figure 9.3).

Particularly enriched sCCA-detected nodes are PTEN and its adjacent PIK3R1. Strong influence of miRNA biomarkers (hsa-miR-525-5p, hsa-miR-317a-5p, hsa-miR-224-5p and hsa-miR-519b-3p) in this network area has been observed. Based on the findings described in the Sections 7.1, 9.1, 9.2 and 9.3 and networks analysis (Figure 7.1 c, 9.2 and 9.3), the most significant impact on development and progression of the cutaneous melanoma in WM115, WM266-4, WM793 and WM1205Lu was recorded in PTEN related PI3K/AKT/mTOR and tyrosine kinase pathways.

Based on these results and the conducted literature research, the hypothesis was defined that microvesicles behave like exosomes in the process of melanoma progression. Many parallels can be found to the work describing the exosomal transfer of functional miRNA and proteins in melanoma, resulting in pre-metastatic niche formation for example:

- exosomes prepare lymph nodes for melanoma macrometastases [135],
- exosomes increase metastatic invasion of primary melanoma and home to lungs [96].

Based on these analogies, detailed functional and mechanistic hypotheses for microvesicles were created, which can be used by scientists to design wet laboratory experiments validating the clinical potential of selected CCA biomarkers. These findings are illustrated on the Figure 9.3 a and b.

Important observation is that among the signatures identified in the sparse multiset CCA analysis, molecules vital in brain tumors and glioblastomas development and progression appear very often. Therefore this hypothesis is considered as impactful on the Figure 9.3 c.

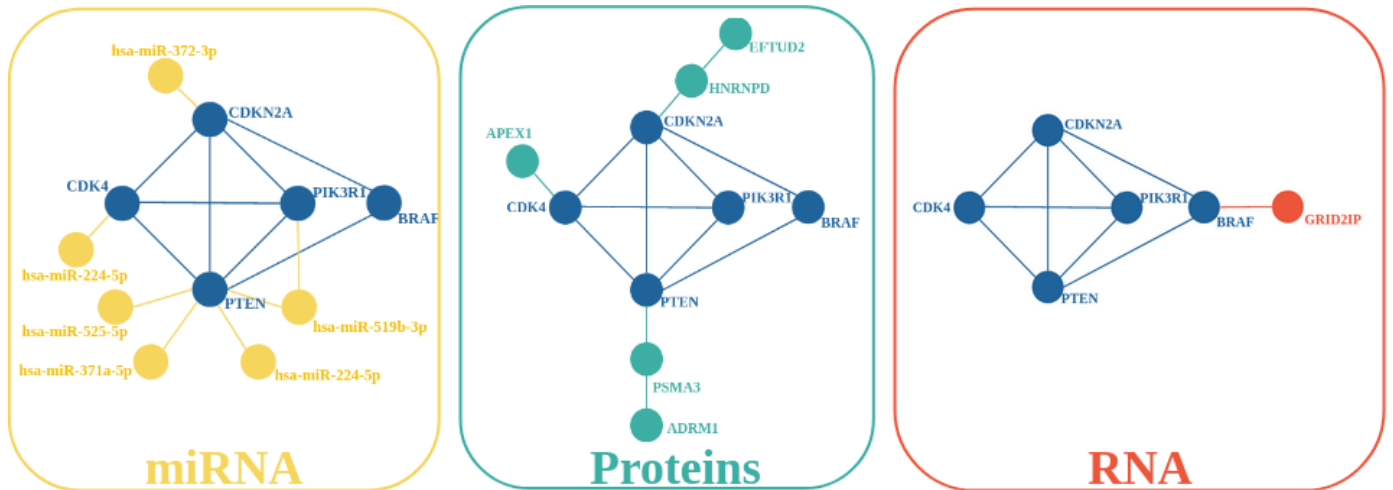


Figure 9.2: | **Interactome panels for sparse multiset CCA results- derived molecular signatures: miRNAs, proteins and RNAs.**

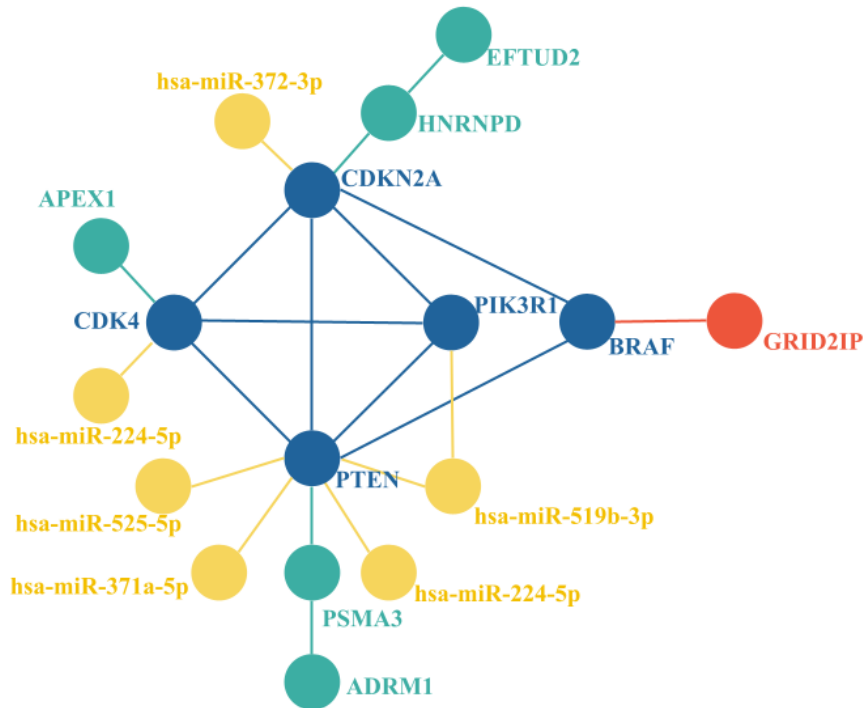


Figure 9.3: | **Interactome for sparse multiset CCA results- derived molecular signatures.** Color blue represents gene-gene interactions network with key oncogenes in analyzed cell lines; yellow represents miRNA nodes and connections, green represents proteins nodes and connections and red represents RNAs nodes and connections.

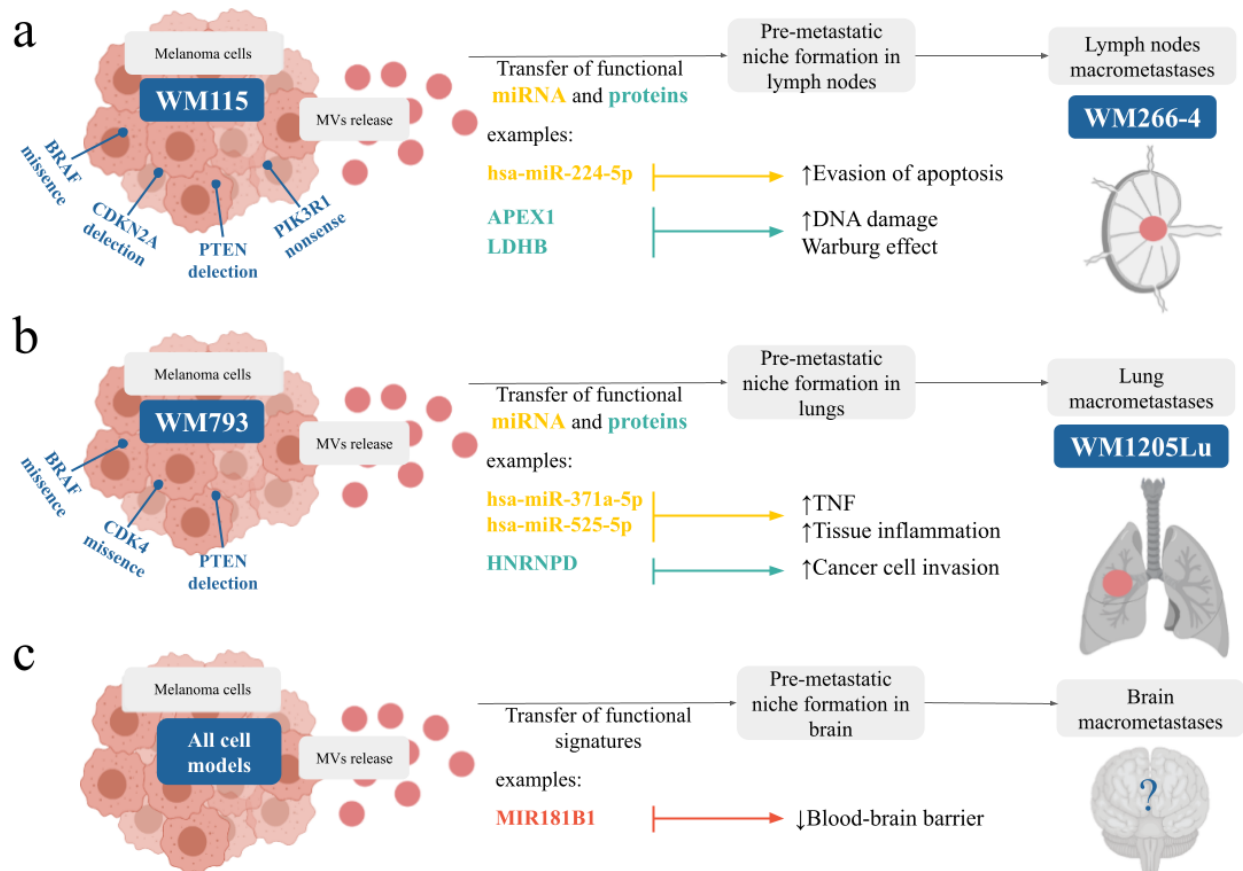


Figure 9.4: | **Functional and mechanistic hypotheses.** Color blue represents key oncogenes in analyzed cell lines; yellow represents miRNA signatures; green represents protein signatures and red represents RNAs signatures. a| Primary WM115 cell line possible transformation process to metastatic WM266-4. b| Primary WM793 cell line possible transformation process to metastatic WM1205Lu. c| The process of metastasis to the brain with the participation of microvesicles functional signatures.

Chapter 10

Discussion

The goal of new CCA-based bioinformatics pipeline development and its validation on real experimental data for melanoma cell lines has been achieved. It has been proven that this approach can be successfully applied to discover new melanoma biomarkers candidates. Of the thousands of possible biological signatures, sparse multiset CCA selected those with strongest correlation to cutaneous melanoma. Majority of them have proven importance in the development and progression of melanoma or other types of cancer. Additionally, sCCA detected new molecules, highly correlated with melanoma which were not previously reported. Therefore, the proposed pipeline has the potential to reveal completely new biomarkers.

The novelty presented in this thesis is the performance of sparse multiset canonical correlation analysis on three matrices with the use of data from extracellular vesicles. There are very few articles describing sparse multiset CCA, and no one has yet used similar combination of data types: information from EVs and whole cells. This synthesis of data allowed to exploration across different biologies which wouldn't be possible using only one biological location. It is an approach promoted by systems biology that is not limited to just one type of information.

The greatest advantage of proposed workflow is the ability to integrate multiple different omics datasets rather than analyzing them separately, which allows to study biological interdependencies. Because of that, the new method reflects the real complexity of tumor microenvironment. Moreover, from the obtained results it can be concluded that sparse multiset CCA is able to maintain the biological context. In the case of the presented analysis,

it was intercellular communication and processes related to carcinogenesis and metastasis formation.

10.1 Combination of centrifugation and hydrostatic filtration dialysis is an efficient protocol for microvesicles miRNAs isolation

The proposed new method of isolation of miRNA from micvesicles, based on hydrostatic filtration dialysis and certifugation protocol, is effective and resulted in obtaining a pure fraction of short RNA fragments that were subjected to successive next generation sequencing. The use of hydrostatic filtration dialysis significantly accelerated the process of miRNA isolation by reducing the amount of liquid material that could be centrifuged more efficiently in small volumes in the next steps of isolation. The miRNA data obtained in this procedure was made available in the open access Jagillonian University Repository [117].

10.2 Melanoma models contain proto-oncogenes

The following mutated genes have been identified that are of key importance in the studied cell lines that affect neoplastic processes: BRAF, CDKN2A, PTEN, CDK4 and PIK3R1.

Missence mutation in BRAF was identified as one of top genetic factors in all tested cells [120, 136]. Cyclin-dependent kinase (CDK) inhibitors (CDKN2A in WM115/WM266-4 and CDK4 in WM793/WM1205Lu) also play an important role in all studied melanoma models [121, 137, 138].

Oncogenic mutation in the PTEN is present in both primary lines studied (deep deletion in WM115 and nonsense mutation in WM793). PTEN is a tumor suppressor and its loss and perturbations are very often observed in neoplasms [120, 139]. PTEN antagonizes the PI3K/AKT/mTOR pathway [140]. Additionally, in the WM115 primary line there is a mutation in the PIK3R1 gene, which also belongs to the phosphoinositide 3-kinases (PI3Ks) family.

Recognition of the genetic background of the cell lines used allowed to study the map of interactions that are key to oncogenesis. Starting the analysis from this step provided the right direction for further functional and mechanistic considerations in biological space.

10.3 Sparse multiset CCA - based bioinformatics pipeline is highly effective method for melanoma biomarkers candidates selection

Novel CCA-based bioinformatics method for melanoma biomarker discovery was proposed. Among the existing CCA extensions, sparse CCA variant was selected, due to its proven effectiveness in the analysis of cancer data and multiple validation in biomarker discovery (Figure 2.3). Sparse multiset CCA method has also been demonstrated to successfully integrate more than two tumor datasets [14,37]. The presented method works perfectly in the case of data with a very large number of variables and a small sample support [141].

The solution presented in the dissertation requires 3 matrices of input data. The operation of the tool was demonstrated on human melanoma: (1) microvesicle miRNA transcriptomics, (2) microvesicle proteomics and (3) whole cell totalRNA transcriptomics. The skin melanoma models used were characterized by various degrees of malignancy.

Sparse multiset CCA pipeline was optimized by examining different lasso penalty values (Figure 8.1). A lasso penalty equal to 2.5 was chosen as the most optimal in conducted analysis. After using this value, 24 biological signatures were obtained as a result: 8 miRNAs, 7 proteins and 9 RNAs (Appendix .4). On these results, a functional analysis was carried out in order to identify biological features.

10.4 Sparse multiset CCA - derived functional and mechanistic insights into microvesicles role in melanoma metastasis formation

The CCA- based bioinformatic pipeline described in this dissertation allowed for the discovery of important functional and mechanistic biological hypotheses. This addresses the challenge faced by the experimenters who, from the enormous space of available options, must choose the most promising for analysis in a wet laboratory.

It also avoids biases resulting from the availability of information or resources, because sparse multiset CCA searches the entire space of available data and examines relationships between datasets, which is very challenging (or even impossible in case of omics data) to do without the use of bioinformatics. As a result, the obtained molecular signatures are not only extremely accurately selected from among all available possibilities, but also the very process of biomarkers selection is relatively fast and less resource-consuming compared to testing each option separately in a laboratory.

PTEN related PI3K/AKT/mTOR and tyrosine kinase pathways were identified as the most important molecular interactions in analysis [120,139,142]. Additionally considering supporting evidence for detected biomarkers candidates, it can be concluded that microvesicles behave very much like exosomes in melanoma pre-metastatic niche formation and progression.

Significant number of molecular signatures related to brain cancers (mostly glioblastoma) was identified. This provides the basis for linking these observations to melanoma strong organo tropism to the brain. Melanoma metastases most often of all cancer types home to the brain. This phenomenon is observed in 40-50 % of patients with stage IV melanoma. EVs are capable of crossing the blood-brain barrier. Moreover, the miRNA miR-181 family (in my analysis MIR181B1 RNA gene was indicated as the most significant in totalRNA panel) has the ability to destroy the blood-brain barrier [143,144]. Signal pathways identified as important in sparse multiset CCA: PTEN pathway and PI3K/AKT/mTOR pathway, are also reported to be crucial in the metastasis of melanoma to the brain.

10.5 Identified molecular signatures were used to design a panel of candidates for liquid biopsy biomarkers

The main finding of the presented sparse multiset canonical correlation analysis is a list of 24 biomarkers candidates. The complete list of selected molecular signatures is presented in Appendices 9. Based on network analysis and verification of supporting biological evidences, this list may be further narrowed down to only those components that have a proven role in skin melanoma or other cancers. This biological findings have determined the following biomarker panel:

1. miRNAs: hsa-miR-224-5p, hsa-miR-525-5p, hsa-miR-371a-5p, hsa-miR-224-5p, hsa-miR-519b-3p, hsa-miR-372-3p
2. proteins: EFTUD2, ADRM1, APEX1, HNRNPD, LDHB, PSMA3.

TotalRNA was omitted in the panel, firstly because of the small contribution of this component in the interactome analysis, and secondly because by designing the panel only with miRNA and proteins, I proposed a signature panel with potential for liquid biopsy. It is a great facilitation for the possible use of the signatures for clinical applications. However, it should be emphasized that the RNA signatures have strong supporting evidence in the processes involved in brain metastasis.

10.6 Possible area for improvements

The analysis presented in the doctoral dissertation was a demonstration of the method's possibilities. However, validations in cohorts (minimum 50-100 patients) are required to generate solid clinical hypotheses. Generating such a large amount of data of three types, i.e. miRNA, proteins and totalRNA, is associated with enormous costs, but amount of open source data that can be used is increasing. The solution may be to integrate data for more melanoma cell lines and generate only the missing ones, which will significantly reduce the cost of experimental work.

As described in the Theoretical Background, redundancy analysis (RDA) is a very promising method that is comparable to CCA. The next step of presented research may be a comparative analysis with the use of both methods: sparse multiset CCA and multiset RDA.

Chapter 11

Conclusions

The obtained results validated positively the hypothesis that that canonical correlation analysis can be applied in extracellular vesicle-based melanoma biomarker discovery and can detect important signatures in multi-omics datasets. The most important conclusions of the presented doctoral dissertation are:

1. Experimental and public sources of proteomic and transcriptomic datasets from microvesicles and whole cells can be successfully used for canonical correlation analysis.
2. Proposed method identified biomarkers candidates like circulating cancer-associated miRNAs and oncoproteins. The presented analytic method allowed to examine the differences between primary and malignant melanoma cells by selecting molecular signatures that have supporting evidence to effect pre-metastatic niche development.
3. Certifugation combine with hydrostatic filtration dialysis method of small RNA isolation and sequencing protocol allowed to obtain good quality miRNA data.
4. The proposed method is extensible and can be used in research on other types of cancer.
5. The used methodology allowed to generate functional and mechanical hypotheses that can be a starting point to research in wet laboratory and to validate the proposed candidates for biomarkers.

Bibliography

- [1] A. Csala and A. H. Zwinderman, “Multivariate statistical methods for high-dimensional multiset omics data analysis,” in *Computational Biology*, pp. 71–83, Codon Publications, Nov. 2019.
- [2] F. S. Collins, “The human genome project: Lessons from large-scale biology,” *Science*, vol. 300, pp. 286–290, Apr. 2003.
- [3] E. D. Green and M. S. Guyer, “Charting a course for genomic medicine from base pairs to bedside,” *Nature*, vol. 470, pp. 204–213, Feb. 2011.
- [4] The 1000 Genomes Project Consortium, “A global reference for human genetic variation,” *Nature*, vol. 526, pp. 68–74, Sept. 2015.
- [5] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, *et al.*, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Research*, vol. 32, pp. 115D–119, Jan. 2004.
- [6] The UniProt Consortium, A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, R. Agivetova, *et al.*, “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Research*, vol. 49, pp. D480–D489, Nov. 2020.
- [7] M. Kanehisa, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 28, pp. 27–30, Jan. 2000.
- [8] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, “KEGG: integrating viruses and cellular organisms,” *Nucleic Acids Research*, vol. 49, pp. D545–D551, Oct. 2020.

- [9] B. Berger, J. Peng, and M. Singh, “Computational solutions for omics data,” *Nature Reviews Genetics*, vol. 14, pp. 333–346, Apr. 2013.
- [10] B. Schmidt and A. Hildebrandt, “Next-generation sequencing: big data meets high performance computing,” *Drug Discovery Today*, vol. 22, pp. 712–717, Apr. 2017.
- [11] C. Manzonì, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, *et al.*, “Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences,” *Briefings in Bioinformatics*, vol. 19, pp. 286–302, Nov. 2016.
- [12] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease,” *Genome Biology*, vol. 18, May 2017.
- [13] C. Turek, S. Wróbel, and M. Piwowar, “OmicsON – integration of omics data with molecular networks and statistical procedures,” *PLOS ONE*, vol. 15, p. e0235398, July 2020.
- [14] T. Rodosthenous, V. Shahrezaei, and M. Evangelou, “Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study,” *Bioinformatics*, vol. 36, pp. 4616–4625, May 2020.
- [15] M. Mathieu, L. Martin-Jaular, G. Lavieu, and C. Théry, “Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication,” *Nature Cell Biology*, vol. 21, pp. 9–17, Jan. 2019.
- [16] G. Raposo and P. D. Stahl, “Extracellular vesicles: a new communication paradigm?,” *Nature Reviews Molecular Cell Biology*, vol. 20, pp. 509–510, July 2019.
- [17] E. Cocucci and J. Meldolesi, “Ectosomes and exosomes: shedding the confusion between extracellular vesicles,” *Trends in Cell Biology*, vol. 25, pp. 364–372, June 2015.
- [18] G. van Niel, G. D’Angelo, and G. Raposo, “Shedding light on the cell biology of extracellular vesicles,” *Nature Reviews Molecular Cell Biology*, vol. 19, pp. 213–228, Jan. 2018.

- [19] M. P. Bebelman, M. J. Smit, D. M. Pegtel, and S. R. Baglio, “Biogenesis and function of extracellular vesicles in cancer,” *Pharmacology & Therapeutics*, vol. 188, pp. 1–11, Aug. 2018.
- [20] R. Xu, A. Rai, M. Chen, W. Suwakulsiri, D. W. Greening, and R. J. Simpson, “Extracellular vesicles in cancer — implications for future improvements in cancer care,” *Nature Reviews Clinical Oncology*, vol. 15, pp. 617–638, May 2018.
- [21] M. J. Shurtleff, M. M. Temoche-Diaz, and R. Schekman, “Extracellular vesicles and cancer: Caveat lector,” *Annual Review of Cancer Biology*, vol. 2, pp. 395–411, Mar. 2018.
- [22] R. Kalluri, “The biology and function of exosomes in cancer,” *Journal of Clinical Investigation*, vol. 126, pp. 1208–1215, Apr. 2016.
- [23] R. Rahbarghazi, N. Jabbari, N. A. Sani, R. Asghari, L. Salimi, S. A. Kalashani, *et al.*, “Tumor-derived extracellular vesicles: reliable tools for cancer diagnosis and clinical applications,” *Cell Communication and Signaling*, vol. 17, July 2019.
- [24] M. Shang, J. S. Ji, C. Song, B. J. Gao, J. G. Jin, W. P. Kuo, *et al.*, “Extracellular vesicles: A brief overview and its role in precision medicine,” in *Methods in Molecular Biology*, pp. 1–14, Springer New York, 2017.
- [25] L. N. de Sousa Andrade, A. H. Otake, S. G. B. Cardim, F. I. da Silva, M. M. I. Sakamoto, T. K. Furuya, *et al.*, “Extracellular vesicles shedding promotes melanoma growth in response to chemotherapy,” *Scientific Reports*, vol. 9, Oct. 2019.
- [26] P. Sharma, B. Diergaarde, S. Ferrone, J. M. Kirkwood, and T. L. Whiteside, “Melanoma cell-derived exosomes in plasma of melanoma patients suppress functions of immune effector cells,” *Scientific Reports*, vol. 10, Jan. 2020.
- [27] S. Wróbel, M. Przybyło, and E. Stepień, “The clinical trial landscape for melanoma therapies,” *Journal of Clinical Medicine*, vol. 8, p. 368, Mar. 2019.

- [28] P. Zaenker, J. Lo, R. Pearce, P. Cantwell, L. Cowell, M. Lee, *et al.*, “A diagnostic autoantibody signature for primary cutaneous melanoma,” *Oncotarget*, vol. 9, pp. 30539–30551, July 2018.
- [29] J. J. Luke and P. A. Ascierto, “Biology confirmed but biomarkers elusive in melanoma immunotherapy,” *Nature Reviews Clinical Oncology*, vol. 17, pp. 198–199, Jan. 2020.
- [30] T. Okabe, T. Fujimura, J. Okajima, Y. Kambayashi, S. Aiba, and S. Maruyama, “First-in-human clinical study of novel technique to diagnose malignant melanoma via thermal conductivity measurements,” *Scientific Reports*, vol. 9, Mar. 2019.
- [31] R. V. Laar, M. Lincoln, and B. V. Laar, “Development and validation of a plasma-based melanoma biomarker suitable for clinical use,” *British Journal of Cancer*, vol. 118, pp. 857–866, Jan. 2018.
- [32] S. Y. Lim, J. H. Lee, R. J. Diefenbach, R. F. Kefford, and H. Rizos, “Liquid biomarkers in melanoma: detection and discovery,” *Molecular Cancer*, vol. 17, Jan. 2018.
- [33] A. Gézsi, Á. Kovács, T. Visnovitz, and E. I. Buzás, “Systems biology approaches to investigating the roles of extracellular vesicles in human diseases,” *Experimental & Molecular Medicine*, vol. 51, pp. 1–11, Mar. 2019.
- [34] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, p. 321, Dec. 1936.
- [35] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, “The next-generation sequencing revolution and its impact on genomics,” *Cell*, vol. 155, pp. 27–38, Sept. 2013.
- [36] E. Parkhomenko, D. Tritchler, and J. Beyene, “Sparse canonical correlation analysis with application to genomic data integration,” *Statistical Applications in Genetics and Molecular Biology*, vol. 8, pp. 1–34, Jan. 2009.
- [37] D. M. Witten and R. J. Tibshirani, “Extensions of sparse canonical correlation analysis with applications to genomic data,” *Statistical Applications in Genetics and Molecular Biology*, vol. 8, pp. 1–27, Jan. 2009.

- [38] M. G. Naylor, X. Lin, S. T. Weiss, B. A. Raby, and C. Lange, “Using canonical correlation analysis to discover genetic regulatory variants,” *PLoS ONE*, vol. 5, p. e10395, May 2010.
- [39] J. Rousu, D. D. Agranoff, O. Sodeinde, J. Shawe-Taylor, and D. Fernandez-Reyes, “Biomarker discovery by sparse canonical correlation analysis of complex clinical phenotypes of tuberculosis and malaria,” *PLoS Computational Biology*, vol. 9, p. e1003018, Apr. 2013.
- [40] D.-S. Cao, S. Liu, W.-B. Zeng, and Y.-Z. Liang, “Sparse canonical correlation analysis applied to -omics studies for integrative analysis and biomarker discovery,” *Journal of Chemometrics*, vol. 29, pp. 371–378, Apr. 2015.
- [41] X. Zhuang, Z. Yang, and D. Cordes, “A technical review of canonical correlation analysis for neuroscience applications,” *Human Brain Mapping*, vol. 41, pp. 3807–3833, June 2020.
- [42] H.-T. Wang, J. Smallwood, J. Mourao-Miranda, C. H. Xia, T. D. Satterthwaite, D. S. Bassett, *et al.*, “Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists,” *NeuroImage*, vol. 216, p. 116745, Aug. 2020.
- [43] A. Ebrahim, E. Brunk, J. Tan, E. J. O’Brien, D. Kim, R. Szubin, *et al.*, “Multi-omic data integration enables discovery of hidden biological regularities,” *Nature Communications*, vol. 7, Oct. 2016.
- [44] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, “Multi-omics data integration, interpretation, and its application,” *Bioinformatics and Biology Insights*, vol. 14, p. 117793221989905, Jan. 2020.
- [45] C. Zang, T. Wang, K. Deng, B. Li, S. Hu, Q. Qin, *et al.*, “High-dimensional genomic data bias correction and data integration using MANCIE,” *Nature Communications*, vol. 7, Apr. 2016.
- [46] K. Tadist, S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi, “Feature selection methods and genomic big data: a systematic review,” *Journal of Big Data*, vol. 6, Aug. 2019.

- [47] Y. Yamanishi, J.-P. Vert, and M. Kanehisa, “Protein network inference from multiple genomic data: a supervised approach,” *Bioinformatics*, vol. 20, pp. i363–i370, July 2004.
- [48] C. S. Tang and M. A. R. Ferreira, “A gene-based test of association using canonical correlation analysis,” *Bioinformatics*, vol. 28, pp. 845–850, Jan. 2012.
- [49] I. Jun, W. Choi, and M. Park, “Multi-block analysis of genomic data using generalized canonical correlation analysis,” *Genomics & Informatics*, vol. 16, p. e33, Dec. 2018.
- [50] L. Revilla, A. Mayorgas, A. M. Corraliza, M. C. Masamunt, A. Metwaly, D. Haller, *et al.*, “Multi-omic modelling of inflammatory bowel disease with regularized canonical correlation analysis,” *PLOS ONE*, vol. 16, p. e0246367, Feb. 2021.
- [51] P. G. P. Martin, H. Guillou, F. Lasserre, S. Déjean, A. Lan, J.-M. Pascussi, *et al.*, “Novel aspects of PPAR α -mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study,” *Hepatology*, vol. 45, pp. 767–777, Mar. 2007.
- [52] I. González, S. Déjean, P. G. P. Martin, O. Gonçalves, P. Besse, and A. Baccini, “Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis,” *Journal of Biological Systems*, vol. 17, pp. 173–199, June 2009.
- [53] M. Piwowar and W. Jurkowski, “ONION: Functional approach for integration of lipidomics and transcriptomics data,” *PLOS ONE*, vol. 10, p. e0128854, June 2015.
- [54] K. Yoshida, J. Yoshimoto, and K. Doya, “Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data,” *BMC Bioinformatics*, vol. 18, Feb. 2017.
- [55] S. Waaijenborg, P. C. V. de Witt Hamer, and A. H. Zwinderman, “Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis,” *Statistical Applications in Genetics and Molecular Biology*, vol. 7, Jan. 2008.

- [56] S. Waaijenborg and A. H. Zwinderman, “Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks,” *BMC Bioinformatics*, vol. 10, Sept. 2009.
- [57] S. Waaijenborg and A. H. Zwinderman, “Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis,” *Bioinformatics*, vol. 25, pp. 2764–2771, Aug. 2009.
- [58] K.-A. L. Cao, P. G. Martin, C. Robert-Granié, and P. Besse, “Sparse canonical methods for biological data integration: application to a cross-platform study,” *BMC Bioinformatics*, vol. 10, Jan. 2009.
- [59] D. M. Witten, R. J. Tibshirani, and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, vol. 10, pp. 515–534, Apr. 2009.
- [60] C. Soneson, H. Lilljebjörn, T. Fioretos, and M. Fontes, “Integrative analysis of gene expression and copy number alterations using canonical correlation analysis,” *BMC Bioinformatics*, vol. 11, Apr. 2010.
- [61] S. Hong, X. Chen, L. Jin, and M. Xiong, “Canonical correlation analysis for RNA-seq co-expression networks,” *Nucleic Acids Research*, vol. 41, pp. e95–e95, Mar. 2013.
- [62] D. Lin, J. Zhang, J. Li, V. D. Calhoun, H.-W. Deng, and Y.-P. Wang, “Group sparse canonical correlation analysis for genomic data integration,” *BMC Bioinformatics*, vol. 14, no. 1, p. 245, 2013.
- [63] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, *et al.*, “Bioconductor: open software development for computational biology and bioinformatics,” *Genome Biology*, vol. 5, no. 10, p. R80, 2004.
- [64] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, *et al.*, “Orchestrating high-throughput genomic analysis with bioconductor,” *Nature Methods*, vol. 12, pp. 115–121, Jan. 2015.

- [65] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.
- [66] Y. Im, H. Gang, and J. K. Yoo, “High-throughput data dimension reduction via seeded canonical correlation analysis,” *Journal of Chemometrics*, vol. 29, pp. 193–199, Nov. 2014.
- [67] A. Tenenhaus and M. Tenenhaus, “Regularized generalized canonical correlation analysis,” *Psychometrika*, vol. 76, pp. 257–284, Mar. 2011.
- [68] T. Jendoubi and K. Strimmer, “A whitening approach to probabilistic canonical correlation analysis for omics data integration,” *BMC Bioinformatics*, vol. 20, Jan. 2019.
- [69] A. Csala, F. P. J. M. Voorbraak, A. H. Zwinderman, and M. H. Hof, “Sparse redundancy analysis of high-dimensional genetic and genomic data,” *Bioinformatics*, vol. 33, pp. 3228–3234, June 2017.
- [70] A. Csala, M. H. Hof, and A. H. Zwinderman, “Multiset sparse redundancy analysis for high-dimensional omics data,” *Biometrical Journal*, vol. 61, pp. 406–423, Dec. 2018.
- [71] T. L. Whiteside, “Tumor-derived exosomes and their role in cancer progression,” in *Advances in Clinical Chemistry*, pp. 103–141, Elsevier, 2016.
- [72] A. Möller and R. J. Lobb, “The evolving translational potential of small extracellular vesicles in cancer,” *Nature Reviews Cancer*, vol. 20, pp. 697–709, Sept. 2020.
- [73] L. Ma, Y. Li, J. Peng, D. Wu, X. Zhao, Y. Cui, *et al.*, “Discovery of the migrasome, an organelle mediating release of cytoplasmic contents during cell migration,” *Cell Research*, vol. 25, pp. 24–38, Oct. 2014.
- [74] P. W. Kriebel, R. Majumdar, L. M. Jenkins, H. Senoo, W. Wang, S. Ammu, *et al.*, “Extracellular vesicles direct migration by synthesizing and releasing chemotactic signals,” *Journal of Cell Biology*, vol. 217, pp. 2891–2910, June 2018.
- [75] F. Teng and M. Fussenegger, “Shedding light on extracellular vesicle biogenesis and bioengineering,” *Advanced Science*, vol. 8, p. 2003505, Nov. 2020.

- [76] M. Surman, E. Stępień, D. Hoja-Łukowicz, and M. Przybyło, “Deciphering the role of ectosomes in cancer development and progression: focus on the proteome,” *Clinical & Experimental Metastasis*, vol. 34, pp. 273–289, Mar. 2017.
- [77] M. Surman, D. Hoja-Łukowicz, S. Szwed, A. Drożdż, E. Stępień, and M. Przybyło, “Human melanoma-derived ectosomes are enriched with specific glycan epitopes,” *Life Sciences*, vol. 207, pp. 395–411, Aug. 2018.
- [78] E. Ł. Stępień, M. Durak-Kozica, A. Kamińska, M. Targosz-Korecka, M. Libera, G. Tylko, *et al.*, “Circulating ectosomes: Determination of angiogenic microRNAs in type 2 diabetes,” *Theranostics*, vol. 8, no. 14, pp. 3874–3890, 2018.
- [79] M. Surman, D. Hoja-Łukowicz, S. Szwed, S. Kędracka-Krok, U. Jankowska, M. Kurtyka, *et al.*, “An insight into the proteome of uveal melanoma-derived ectosomes reveals the presence of potentially useful biomarkers,” *International Journal of Molecular Sciences*, vol. 20, p. 3789, Aug. 2019.
- [80] M. Surman, S. Kędracka-Krok, D. Hoja-Łukowicz, U. Jankowska, A. Drożdż, E. Ł. Stępień, *et al.*, “Mass spectrometry-based proteomic characterization of cutaneous melanoma ectosomes reveals the presence of cancer-related molecules,” *International Journal of Molecular Sciences*, vol. 21, p. 2934, Apr. 2020.
- [81] I. Prada and J. Meldolesi, “Binding and fusion of extracellular vesicles to the plasma membrane of their cell targets,” *International Journal of Molecular Sciences*, vol. 17, p. 1296, Aug. 2016.
- [82] T. Tian, Y.-L. Zhu, Y.-Y. Zhou, G.-F. Liang, Y.-Y. Wang, F.-H. Hu, *et al.*, “Exosome uptake through clathrin-mediated endocytosis and macropinocytosis and mediating miR-21 delivery,” *Journal of Biological Chemistry*, vol. 289, pp. 22258–22267, Aug. 2014.
- [83] T. Janas, M. M. Janas, K. Sapoń, and T. Janas, “Mechanisms of RNA loading into exosomes,” *FEBS Letters*, vol. 589, pp. 1391–1398, Apr. 2015.

- [84] A. Gonda, J. Kabagwira, G. N. Senthil, and N. R. Wall, “Internalization of exosomes through receptor-mediated endocytosis,” *Molecular Cancer Research*, vol. 17, pp. 337–347, Nov. 2018.
- [85] R. Kalluri and V. S. LeBleu, “The biology, function, and biomedical applications of exosomes,” *Science*, vol. 367, p. eaau6977, Feb. 2020.
- [86] H. Valadi, K. Ekström, A. Bossios, M. Sjöstrand, J. J. Lee, and J. O. Lötval, “Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells,” *Nature Cell Biology*, vol. 9, pp. 654–659, May 2007.
- [87] P. Kurywachak, J. Tavormina, and R. Kalluri, “The emerging roles of exosomes in the modulation of immune responses in cancer,” *Genome Medicine*, vol. 10, Mar. 2018.
- [88] E. Z. Malkin and S. V. Bratman, “Bioactive DNA from extracellular vesicles and particles,” *Cell Death & Disease*, vol. 11, July 2020.
- [89] A. Turchinovich, O. Drapkina, and A. Tonevitsky, “Transcriptome of extracellular vesicles: State-of-the-art,” *Frontiers in Immunology*, vol. 10, Feb. 2019.
- [90] T. L. Whiteside, “Validation of plasma-derived small extracellular vesicles as cancer biomarkers,” *Nature Reviews Clinical Oncology*, vol. 17, pp. 719–720, Sept. 2020.
- [91] K. B. Johnsen, J. M. Gudbergsson, T. L. Andresen, and J. B. Simonsen, “What is the blood concentration of extracellular vesicles? implications for the use of extracellular vesicles as blood-borne biomarkers of cancer,” *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1871, pp. 109–116, Jan. 2019.
- [92] I. Parolini, C. Federici, C. Raggi, L. Lugini, S. Palleschi, A. D. Milito, *et al.*, “Microenvironmental pH is a key factor for exosome traffic in tumor cells,” *Journal of Biological Chemistry*, vol. 284, pp. 34211–34222, Sept. 2009.
- [93] S. Fais, G. Venturi, and B. Gatenby, “Microenvironmental acidosis in carcinogenesis and metastases: new strategies in prevention and therapy,” *Cancer and Metastasis Reviews*, vol. 33, pp. 1095–1108, Nov. 2014.

- [94] M. Logozzi, D. Mizzoni, D. Angelini, R. D. Raimo, M. Falchi, L. Battistini, *et al.*, “Microenvironmental pH and exosome levels interplay in human cancer cell lines of different histotypes,” *Cancers*, vol. 10, p. 370, Oct. 2018.
- [95] Z. Boussadia, J. Lamberti, F. Mattei, E. Pizzi, R. Puglisi, C. Zanetti, *et al.*, “Acidic microenvironment plays a key role in human melanoma progression through a sustained exosome mediated transfer of clinically relevant metastatic molecules,” *Journal of Experimental & Clinical Cancer Research*, vol. 37, Oct. 2018.
- [96] H. Peinado, M. Alečković, S. Lavotshkin, I. Matei, B. Costa-Silva, G. Moreno-Bueno, *et al.*, “Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET,” *Nature Medicine*, vol. 18, pp. 883–891, May 2012.
- [97] G. Chen, A. C. Huang, W. Zhang, G. Zhang, M. Wu, W. Xu, *et al.*, “Exosomal PD-11 contributes to immunosuppression and is associated with anti-PD-1 response,” *Nature*, vol. 560, pp. 382–386, Aug. 2018.
- [98] C. Federici, F. Petrucci, S. Caimi, A. Cesolini, M. Logozzi, M. Borghi, *et al.*, “Exosome release and low pH belong to a framework of resistance of human melanoma cells to cisplatin,” *PLoS ONE*, vol. 9, p. e88193, Feb. 2014.
- [99] F.-N. B. W. Group, “Best (biomarkers, endpoints, and other tools),” *Resource [Internet: <https://www.ncbi.nlm.nih.gov/books/NBK326791/>]*, 2016.
- [100] T. R. Lunavat, L. Cheng, D.-K. Kim, J. Bhadury, S. C. Jang, C. Lässer, *et al.*, “Small RNA deep sequencing discriminates subsets of extracellular vesicles released by melanoma cells – evidence of unique microRNA cargos,” *RNA Biology*, vol. 12, pp. 810–823, July 2015.
- [101] M. Logozzi, A. D. Milito, L. Lugini, M. Borghi, L. Calabrò, M. Spada, *et al.*, “High levels of exosomes expressing CD63 and caveolin-1 in plasma of melanoma patients,” *PLoS ONE*, vol. 4, p. e5219, Apr. 2009.

- [102] G. Andreola, L. Rivoltini, C. Castelli, V. Huber, P. Perego, P. Deho, *et al.*, “Induction of lymphocyte apoptosis by tumor cell secretion of FasL-bearing microvesicles,” *Journal of Experimental Medicine*, vol. 195, pp. 1303–1316, May 2002.
- [103] S. Keerthikumar, D. Chisanga, D. Ariyaratne, H. A. Saffar, S. Anand, K. Zhao, *et al.*, “ExoCarta: A web-based compendium of exosomal cargo,” *Journal of Molecular Biology*, vol. 428, pp. 688–692, Feb. 2016.
- [104] M. Pathan, P. Fonseka, S. V. Chitti, T. Kang, R. Sanwlani, J. V. Deun, *et al.*, “Vesiclepedia 2019: a compendium of RNA, proteins, lipids and metabolites in extracellular vesicles,” *Nucleic Acids Research*, vol. 47, pp. D516–D519, Nov. 2018.
- [105] D.-K. Kim, J. Lee, S. R. Kim, D.-S. Choi, Y. J. Yoon, J. H. Kim, *et al.*, “EVpedia: a community web portal for extracellular vesicles research,” *Bioinformatics*, vol. 31, pp. 933–939, Nov. 2014.
- [106] F. Russo, S. D. Bella, F. Vannini, G. Berti, F. Scoyni, H. V. Cook, *et al.*, “miRandola 2017: a curated knowledge base of non-invasive biomarkers,” *Nucleic Acids Research*, vol. 46, pp. D354–D359, Sept. 2017.
- [107] O. D. Murillo, W. Thistlethwaite, J. Rozowsky, S. L. Subramanian, R. Lucero, N. Shah, *et al.*, “exRNA atlas analysis reveals distinct extracellular RNA cargo types and their carriers present across human biofluids,” *Cell*, vol. 177, pp. 463–477.e15, Apr. 2019.
- [108] R. Shikder, P. Irani, and P. Hu, “Genome-wide canonical correlation analysis-based computational methods for mining information from microbiome and gene expression data,” in *Advances in Artificial Intelligence*, pp. 511–517, Springer International Publishing, 2019.
- [109] M. Maio, “Melanoma as a model tumour for immuno-oncology,” *Annals of Oncology*, vol. 23, pp. viii10–viii14, Sept. 2012.
- [110] M. Herlyn, G. Balaban, J. Bannicelli, D. G. IV, R. Halaban, D. Herlyn, *et al.*, “Primary melanoma cells of the vertical growth phase: Similarities to metastatic cells,” *JNCI: Journal of the National Cancer Institute*, Feb. 1985.

- [111] I. Cornil, D. Theodorescu, S. Man, M. Herlyn, J. Jambrosic, and R. S. Kerbel, “Fibroblast cell interactions with human melanoma cells affect tumor cell growth as a function of tumor progression,” *Proceedings of the National Academy of Sciences*, vol. 88, pp. 6028–6032, July 1991.
- [112] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, *et al.*, “The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data: Figure 1,” *Cancer Discovery*, vol. 2, pp. 401–404, May 2012.
- [113] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, *et al.*, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,” *Science Signaling*, vol. 6, pp. pl1–pl1, Mar. 2013.
- [114] M. Roman, A. Kamińska, A. Drożdż, M. Platt, M. Kuźniewski, M. T. Małecki, *et al.*, “Raman spectral signatures of urinary extracellular vesicles from diabetic patients and hyperglycemic endothelial cells as potential biomarkers in diabetes,” *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 17, pp. 137–149, Apr. 2019.
- [115] A. Drożdż, A. Kamińska, M. Surman, A. Gonet-Surówka, R. Jach, H. Huras, *et al.*, “Low-vacuum filtration as an alternative extracellular vesicle concentration method: A comparison with ultracentrifugation and differential centrifugation,” *Pharmaceutics*, vol. 12, p. 872, Sept. 2020.
- [116] A. R. Sepulveda, S. R. Hamilton, C. J. Allegra, W. Grody, A. M. Cushman-Vokoun, W. K. Funkhouser, *et al.*, “Molecular biomarkers for the evaluation of colorectal cancer,” *The Journal of Molecular Diagnostics*, vol. 19, pp. 187–225, Mar. 2017.
- [117] S. Wróbel, E. Stepień, C. Turek, M. Piwowar, M. Przybyło, and M. Surman, “Microvesicles microrna transcriptomes dataset for human melanoma cell cultures.” Jagiellonian University Repository, 2021.
- [118] D. van der Meer, S. Barthorpe, W. Yang, H. Lightfoot, C. Hall, J. Gilbert, *et al.*, “Cell model passports—a hub for clinical, genetic and functional datasets of preclinical cancer models,” *Nucleic Acids Research*, vol. 47, pp. D923–D929, Sept. 2018.

- [119] “Bioproject:prjna690945; geo: Gse164495; transcriptome analysis of melanoma cells stably expressing bmp6 (human).” NCBI Resources, 2021.
- [120] D. Dankort, D. P. Curley, R. A. Cartlidge, B. Nelson, A. N. Karnezis, W. E. D. Jr, *et al.*, “BrafV600e cooperates with pten loss to induce metastatic melanoma,” *Nature Genetics*, vol. 41, pp. 544–552, Mar. 2009.
- [121] H. Zeng, A. Jorapur, A. H. Shain, U. E. Lang, R. Torres, Y. Zhang, *et al.*, “Bi-allelic loss of CDKN2a initiates melanoma invasion via BRN2 activation,” *Cancer Cell*, vol. 34, pp. 56–68.e9, July 2018.
- [122] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, *et al.*, “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” *Nucleic Acids Research*, vol. 47, pp. D607–D613, Nov. 2018.
- [123] Y. Li, J. Huang, J. Sun, S. Xiang, D. Yang, X. Ying, *et al.*, “The transcription levels and prognostic values of seven proteasome alpha subunits in human cancers,” *Oncotarget*, vol. 8, pp. 4501–4519, Dec. 2016.
- [124] M. Kumar, A. Matta, O. Masui, G. Srivastava, J. Kaur, A. Thakar, *et al.*, “Nuclear heterogeneous nuclear ribonucleoprotein d is associated with poor prognosis and interactome analysis reveals its novel binding partners in oral cancer,” *Journal of Translational Medicine*, vol. 13, Aug. 2015.
- [125] L. Alfano, A. Caporaso, A. Altieri, M. Dell’Aquila, C. Landi, L. Bini, *et al.*, “Depletion of the RNA binding protein HNRNPD impairs homologous recombination by inhibiting DNA-end resection and inducing r-loop accumulation,” *Nucleic Acids Research*, vol. 47, pp. 4068–4085, Feb. 2019.
- [126] H. Xu, H. Han, S. Song, N. Yi, C. Qian, Y. Qiu, *et al.*, “Exosome-transmitted PSMA3 and PSMA3-AS1 promote proteasome inhibitor resistance in multiple myeloma,” *Clinical Cancer Research*, vol. 25, pp. 1923–1935, Jan. 2019.

- [127] J. Liu, Y. Xing, and L. Rong, “miR-181 regulates cisplatin-resistant non-small cell lung cancer via downregulation of autophagy through the PTEN/PI3k/AKT pathway,” *Oncology Reports*, Feb. 2018.
- [128] J. Henao-Mejia, A. Williams, L. A. Goff, M. Staron, P. Licona-Limón, S. M. Kaeck, *et al.*, “The MicroRNA miR-181 is a critical cellular metabolic rheostat essential for NKT cell ontogenesis and lymphocyte development and homeostasis,” *Immunity*, vol. 38, pp. 984–997, May 2013.
- [129] H.-W. Tseng, S.-C. Li, and K.-W. Tsai, “Metformin treatment suppresses melanoma cell growth and motility through modulation of microRNA expression,” *Cancers*, vol. 11, p. 209, Feb. 2019.
- [130] W. J. Lee, M. H. Lee, H. T. Kim, C. H. Won, M. W. Lee, J. H. Choi, *et al.*, “Prognostic significance of CD163 expression and its correlation with cyclooxygenase-2 and vascular endothelial growth factor expression in cutaneous melanoma,” *Melanoma Research*, vol. 29, pp. 501–509, Oct. 2019.
- [131] C. Roufas, I. Georgakopoulos-Soares, and A. Zaravinos, “Distinct genomic features across cytolytic subgroups in skin melanoma,” *Cancer Immunology, Immunotherapy*, Mar. 2021.
- [132] C. Yang, H. Zhan, Y. Zhao, Y. Wu, L. Li, and H. Wang, “MEX3a contributes to development and progression of glioma through regulating cell proliferation and cell migration and targeting CCL2,” *Cell Death & Disease*, vol. 12, Jan. 2021.
- [133] F. Bufalieri, M. Caimano, L. L. Severini, I. Basili, F. Paglia, L. Sampirisi, *et al.*, “The RNA-binding ubiquitin ligase MEX3a affects glioblastoma tumorigenesis by inducing ubiquitylation and degradation of RIG-i,” *Cancers*, vol. 12, p. 321, Jan. 2020.
- [134] H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, *et al.*, “miR-TarBase 2020: updates to the experimentally validated microRNA–target interaction database,” *Nucleic Acids Research*, Oct. 2019.

- [135] J. L. Hood, R. S. San, and S. A. Wickline, “Exosomes released by melanoma cells prepare sentinel lymph nodes for tumor metastasis,” *Cancer Research*, vol. 71, pp. 3792–3801, Apr. 2011.
- [136] B. Y. Kong, M. S. Carlino, and A. M. Menzies, “Biology and treatment of BRAF mutant metastatic melanoma,” *Melanoma Management*, vol. 3, pp. 33–45, Mar. 2016.
- [137] B. Lee and G. A. McArthur, “CDK4 inhibitors an emerging strategy for the treatment of melanoma,” *Melanoma Management*, vol. 2, pp. 255–266, Aug. 2015.
- [138] M. Julve, J. J. Clark, and M. P. Lythgoe, “Advances in cyclin-dependent kinase inhibitors for the treatment of melanoma,” *Expert Opinion on Pharmacotherapy*, vol. 22, pp. 351–361, Oct. 2020.
- [139] H. Wu, V. Goel, and F. G. Haluska, “PTEN signaling pathways in melanoma,” *Oncogene*, vol. 22, pp. 3113–3122, May 2003.
- [140] N. Chalhoub and S. J. Baker, “PTEN and the PI3-kinase pathway in cancer,” *Annual Review of Pathology: Mechanisms of Disease*, vol. 4, pp. 127–150, Feb. 2009.
- [141] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasiija, “Canonical correlation analysis of high-dimensional data with very small sample support,” *Signal Processing*, vol. 128, pp. 449–458, Nov. 2016.
- [142] Y. Dong, J.-A. Richards, R. Gupta, P. P. Aung, A. Emley, Y. Kluger, *et al.*, “PTEN functions as a melanoma tumor suppressor by promoting host immune response,” *Oncogene*, vol. 33, pp. 4632–4642, Oct. 2013.
- [143] J. Ma, Y. Yao, P. Wang, Y. Liu, L. Zhao, Z. Li, *et al.*, “MiR-181a regulates blood-tumor barrier permeability by targeting krüppel-like factor 6,” *Journal of Cerebral Blood Flow & Metabolism*, vol. 34, pp. 1826–1836, Sept. 2014.
- [144] N. Tominaga, N. Kosaka, M. Ono, T. Katsuda, Y. Yoshioka, K. Tamura, *et al.*, “Brain metastatic cancer cells release microRNA-181c-containing extracellular vesicles capable of destructing blood-brain barrier,” *Nature Communications*, vol. 6, Apr. 2015.

- [145] A. Cichonska, J. Rousu, P. Marttinen, A. J. Kangas, P. Soinen, T. Lehtimäki, *et al.*, “metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis,” *Bioinformatics*, vol. 32, pp. 1981–1989, Feb. 2016.
- [146] G. Abraham and M. Inouye, “FlashPCA: fast sparse canonical correlation analysis of genomic data,” Apr. 2016.
- [147] X. Cao, K. R. Crews, J. Downing, J. Lamba, and S. B. Pounds, “CC-PROMISE effectively integrates two forms of molecular data with multiple biologically related endpoints,” *BMC Bioinformatics*, vol. 17, Oct. 2016.
- [148] P. Madrigal, “fCCAC: functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets,” *Bioinformatics*, p. btw724, Dec. 2016.
- [149] N. Lin, Y. Zhu, R. Fan, and M. Xiong, “A quadratically regularized functional canonical correlation analysis for identifying the global structure of pleiotropy with NGS data,” *PLOS Computational Biology*, vol. 13, p. e1005788, Oct. 2017.
- [150] A. Mandal and P. Maji, “FaRoC: Fast and robust supervised canonical correlation analysis for multimodal omics data,” *IEEE Transactions on Cybernetics*, vol. 48, pp. 1229–1241, Apr. 2018.
- [151] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, *et al.*, “The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity,” *Nature*, vol. 483, pp. 603–607, Mar. 2012.
- [152] C. F. Malone, N. V. Dharia, G. Kugener, A. B. Forman, M. V. Rothberg, M. Abdusamad, *et al.*, “Selective modulation of a pan-essential protein as a therapeutic strategy in cancer,” *Cancer Discovery*, p. candisc.1213.2020, Apr. 2021.
- [153] S. Li, Y. Li, B. Chen, J. Zhao, S. Yu, Y. Tang, *et al.*, “exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes,” *Nucleic Acids Research*, vol. 46, pp. D106–D112, Oct. 2017.

- [154] T. Liu, Q. Zhang, J. Zhang, C. Li, Y.-R. Miao, Q. Lei, *et al.*, “EVmiRNA: a database of miRNA profiling in extracellular vesicles,” *Nucleic Acids Research*, vol. 47, pp. D89–D93, Oct. 2018.
- [155] R. Edgar, M. Domrachev, and A. E. Lash, “Gene expression omnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Research*, vol. 30, pp. 207–210, Jan. 2002.
- [156] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, *et al.*, “NCBI GEO: archive for functional genomics data sets—update,” *Nucleic Acids Research*, vol. 41, pp. D991–D995, Nov. 2012.
- [157] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, “Gene ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, pp. 25–29, May 2000.
- [158] S. Carbon, E. Douglass, B. M. Good, D. R. Unni, N. L. Harris, C. J. Mungall, *et al.*, “The gene ontology resource: enriching a GOld mine,” *Nucleic Acids Research*, vol. 49, pp. D325–D334, Dec. 2020.
- [159] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, *et al.*, “Toward a shared vision for cancer genomic data,” *New England Journal of Medicine*, vol. 375, pp. 1109–1112, Sept. 2016.
- [160] D. Ochoa, A. Hercules, M. Carmona, D. Suveges, A. Gonzalez-Uriarte, C. Malangone, *et al.*, “Open targets platform: supporting systematic drug–target identification and prioritisation,” *Nucleic Acids Research*, vol. 49, pp. D1302–D1310, Nov. 2020.
- [161] H. Mi, D. Ebert, A. Muruganujan, C. Mills, L.-P. Albou, T. Mushayamaha, *et al.*, “PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API,” *Nucleic Acids Research*, vol. 49, pp. D394–D403, Dec. 2020.

- [162] J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, *et al.*, “ProteomeXchange provides globally coordinated proteomics data submission and dissemination,” *Nature Biotechnology*, vol. 32, pp. 223–226, Mar. 2014.
- [163] E. W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, *et al.*, “The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition,” *Nucleic Acids Research*, vol. 45, pp. D1100–D1106, Oct. 2016.
- [164] A. Fabregat, K. Sidiropoulos, G. Viteri, O. Forner, P. Marin-Garcia, V. Arnau, *et al.*, “Reactome pathway analysis: a high-performance in-memory approach,” *BMC Bioinformatics*, vol. 18, Mar. 2017.
- [165] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, *et al.* *Nucleic Acids Research*, Nov. 2019.
- [166] Y. Lin, T. Liu, T. Cui, Z. Wang, Y. Zhang, P. Tan, *et al.*, “RNAInter in 2020: RNA interactome repository with increased coverage and annotation,” *Nucleic Acids Research*, vol. 48, pp. D189–D197, Sept. 2019.
- [167] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen, “STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene,” *Nucleic Acids Research*, vol. 28, pp. 3442–3444, Sept. 2000.
- [168] D. Karagkouni, M. D. Paraskevopoulou, S. Chatzopoulos, I. S. Vlachos, S. Tastsoglou, I. Kanellos, *et al.*, “DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions,” *Nucleic Acids Research*, vol. 46, pp. D239–D245, Nov. 2017.
- [169] J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, *et al.*, “An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics,” *Cell*, vol. 173, pp. 400–416.e11, Apr. 2018.

- [170] N. Jain, K. F. Mittendorf, M. Holt, M. Lenoue-Newton, I. Maurer, C. Miller, *et al.*, “The my cancer genome clinical trial data model and trial curation workflow,” *Journal of the American Medical Informatics Association*, vol. 27, pp. 1057–1066, June 2020.
- [171] M. F. Berger, E. Hodis, T. P. Heffernan, Y. L. Deribe, M. S. Lawrence, A. Protopopov, *et al.*, “Melanoma genome sequencing reveals frequent PREX2 mutations,” *Nature*, vol. 485, pp. 502–506, May 2012.
- [172] A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, *et al.*, “Genetic basis for clinical response to CTLA-4 blockade in melanoma,” *New England Journal of Medicine*, vol. 371, pp. 2189–2199, Dec. 2014.
- [173] R. Akbani, K. C. Akdemir, B. A. Aksoy, M. Albert, A. Ally, S. B. Amin, *et al.*, “Genomic classification of cutaneous melanoma,” *Cell*, vol. 161, pp. 1681–1696, June 2015.
- [174] E. M. V. Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, *et al.*, “Genomic correlates of response to CTLA-4 blockade in metastatic melanoma,” *Science*, vol. 350, pp. 207–211, Sept. 2015.
- [175] E. Hodis, I. R. Watson, G. V. Kryukov, S. T. Arold, M. Imielinski, J.-P. Theurillat, *et al.*, “A landscape of driver mutations in melanoma,” *Cell*, vol. 150, pp. 251–263, July 2012.
- [176] M. Krauthammer, Y. Kong, B. H. Ha, P. Evans, A. Bacchiocchi, J. P. McCusker, *et al.*, “Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma,” *Nature Genetics*, vol. 44, pp. 1006–1014, July 2012.
- [177] E. M. V. Allen, N. Wagle, A. Sucker, D. J. Treacy, C. M. Johannessen, E. M. Goetz, *et al.*, “The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma,” *Cancer Discovery*, vol. 4, pp. 94–109, Nov. 2013.
- [178] J. V. Deun, P. Mestdagh, P. Agostinis, Özden Akay, S. Anand, J. Anckaert, *et al.*, “EV-TRACK: transparent reporting and centralizing knowledge in extracellular vesicle research,” *Nature Methods*, vol. 14, pp. 228–232, Feb. 2017.

Appendices

.1 List of CCA-based bioinformatics tools

Table 1: List of analyzed bioinformatics and publicly available CCA tools. All tabulated implementations were validated on omics data.

Tool name	Multiset	Year	Reference	Link
MultiCCA	yes	2009	[59]	https://github.com/bnaras/PMA
CCA	no	2009	[52]	https://cran.r-project.org/web/packages/CCA
FRCC	no	2012	[67]	https://cran.r-project.org/web/packages/FRCC
seedCCA	no	2014	[66]	https://cran.r-project.org/web/packages/seedCCA
ONION	no	2015	[53]	https://github.com/wjurkowski/ONION
metaCCA	no	2016	[145]	https://bioconductor.org/packages/metaCCA
FlashPCA	no	2016	[146]	https://github.com/gabraham/flashpca
CCPROMISE	no	2016	[147]	https://bioconductor.org/packages/CCPROMISE
fCCAC	no	2016	[148]	https://bioconductor.org/packages/fCCAC
RGCCA	no	2017	-	https://cran.r-project.org/web/packages/RGCCA
QRFCCA	no	2017	[149]	https://CRAN.R-project.org/packages/QRFCCA
FaRoC	no	2018	[150]	-
whitening	no	2019	[68]	https://cran.r-project.org/web/packages/whitening
SmCCNet	no	2019	-	https://cran.r-project.org/web/packages/SmCCNet
sCCA	yes	2020	[14]	https://github.com/theorod93/sCCA
OmicsON	no	2020	[13]	https://github.com/cmujzbit/OmicsON

.2 List of omics databases

Table 2: List of omics databases used or tested in bioinformatics analysis.

Database		Reference	Link
cBioPortal	-	[113]	https://www.cbioportal.org/
Cell Model Passports	-	[118]	https://cellmodelpassports.sanger.ac.uk/
Cancer Dependency Map	DepMap	[151, 152]	https://depmap.org/
ExoCarta	-	[103]	https://www.exocarta.org/
exoRBase	-	[153]	http://www.exorbase.org/
exRNA Atlas	-	[107]	https://exrna-atlas.org/
Extracellular Vesicles miRNA Database	EVmiRNA	[154]	http://bioinfo.life.hust.edu.cn/EVmiRNA/
Gene Expression Omnibus	GEO	[155, 156]	https://www.ncbi.nlm.nih.gov/geo/
Gene Ontology Resource	GO	[157, 158]	http://geneontology.org/
Genomics Data Commons Portal	GDC	[159]	https://portal.gdc.cancer.gov/
Kyoto Encyclopedia of Genes and Genomes	KEGG	[7, 8]	https://www.genome.jp/kegg/
miRandola	-	[106]	http://mirandola.iit.cnr.it/
miRTarBase	-	[134]	http://mirtarbase.cuhk.edu.cn/
Open Targets Platform	-	[160]	https://www.opentargets.org/
Protein ANalysis THrough Evolutionary Relationships	PANTHER	[161]	http://pantherdb.org/
ProteomeXchange	-	[162, 163]	http://www.proteomexchange.org/
Reactome	-	[164, 165]	https://reactome.org/
RNA Interactome Database	RNAInter	[166]	http://www.rna-society.org/rnainter/
Search Tool for Recurring Instances of Neighbouring Gene	STRING	[122, 167]	https://string-db.org/
TarBase	-	[168]	http://carolina.imis.athena-innovation.gr/
The Cancer Genome Atlas	TCGA	[169]	https://www.cancer.gov/tcga
Vesiclepedia	-	[104]	http://www.microvesicles.org/

.3 Melanoma molecular profile

Many molecular factors are described as important in melanoma oncogenesis such as changes in the mitogen-activated protein kinase (MAPK) pathway or phosphoinositol-3-kinase (PI3K) pathway and the presence of driver mutations in genes: BRAF (37-50 % of melanomas), NRAS (13-25 % of melanomas), NF1 (11.9 % of melanomas), MEK1 (6-7 % of melanomas), KIT (2-8 % of melanomas), CTNNB1 (2-4 % of melanomas), GNAQ (1.3 % of melanomas) or GNA11 (1.2 % of melanomas) [169,170].

The cell lines I used in this dissertation: WM115, WM266-4, WM793 and WM1205Lu are of the cutaneous melanoma type. In order to understand the genetic profile of the cell lines used in the context of the melanoma patients population, I analyzed the expression profiles for cutaneous melanoma clinical data from the following sources:

1. Broad Institute/Dana Farber Cancer Institute, Nature 2012 [171]
2. Memorial Sloan Kettering Cancer Center, 2018
3. Memorial Sloan Kettering Cancer Center, New England Journal of Medicine 2014 [172]
4. TCGA, Cell 2015 [173]
5. Dana-Farber Cancer Institute, Science 2015 [174]
6. Memorial Sloan Kettering Cancer Center, JCO Precision Oncology 2017
7. Broad Institute, Cell 2012 [175]
8. TCGA, Firehose Legacy
9. TCGA, PanCancer Atlas [169]
10. Yale, Nature Genetics 2012 [176]
11. Broad Institute, Cancer Discovery 2014 [177]

All the data I used has been added to the "virtual study" created for this analysis and can be downloaded from the cBioPortal website by using unique link: <https://www.cbioportal.org/study?id=60915411e4b0f0abac8ae8e1>.

The analysis of the broad genetic context allowed for the precise identification of essential molecular signatures in the studied cell lines, which are important in the process of oncogenesis. The findings were supported by literature studies to understand the biological background of the processes. This is crucial for the interpretation and generation of functional and mechanical hypotheses.

.4 Sparse multiset CCA results

Table 3: List of all significant molecular signatures: miRNAs, proteins, and totalRNA identified by sparse multiset CCA for lasso penalty equal to 2.5.

Type	Accession	Name	Score
protein	Q15029	EFTUD2	0.54569
RNA	ENSG00000207975	MIR181B1	0.49818
RNA	ENSG00000269904	MAP2K4P1	0.49432
RNA	ENSG00000171570	RAB4B-EGLN2	0.49432
miRNA	MIMAT0004687	hsa-miR-371a-5p	0.45751
miRNA	MIMAT0000724	hsa-miR-372-3p	0.45339
miRNA	MIMAT0000281	hsa-miR-224-5p	0.44775
miRNA	MIMAT0002866	hsa-miR-517c-3p	0.44331
protein	Q16186	ADRM1	0.43178
RNA	ENSG00000115380	EFEMP1	0.42259
protein	P27695	APEX1	0.39002
miRNA	MIMAT0002837	hsa-miR-519b-3p	0.38675
protein	Q14103	HNRNPD	0.38322
protein	P07195	LDHB	0.32278
protein	P25788	PSMA3	0.31736
RNA	ENSG00000177575	CD163	0.21231
miRNA	MIMAT0002859	hsa-miR-516b-5p	0.15347
RNA	ENSG00000050730	TNIP3	0.11515
miRNA	MIMAT0002838	hsa-miR-525-5p	0.11448
RNA	ENSG00000215160	OR7E122P	0.11325
protein	Q6DD88	ATL3	0.10916
RNA	ENSG00000254726	MEX3A	0.10581
RNA	ENSG00000215045	GRID2IP	0.04656

miRNA	MIMAT0002835	hsa-miR-526b-5p	0.04333
-------	--------------	-----------------	---------

.5 Optimization of sparse multiset CCA results

Table 4: List of all significant molecular miRNA signatures for lasso penalty equal to 4, identified in the optimization process in sparse multiset CCA.

miRBase accession	Name	Score
MIMAT0002838	hsa-miR-525-5p	0.29929
MIMAT0002859 _ 1	hsa-miR-516b-5p	0.29923
MIMAT0002835	hsa-miR-526b-5p	0.29864
MIMAT0002837	hsa-miR-519b-3p	0.28621
MIMAT0002855	hsa-miR-520d-5p	0.28467
MIMAT0002833	hsa-miR-520a-5p	0.28015
MIMAT0002866	hsa-miR-517c-3p	0.27209
MIMAT0005452	hsa-miR-519a*	0.24956
MIMAT0004687	hsa-miR-371a-5p	0.23610
MIMAT0000724	hsa-miR-372-3p	0.23167
MIMAT0000281	hsa-miR-224-5p	0.22885
MIMAT0002849	hsa-miR-524-5p	0.21293
MIMAT0002848	hsa-miR-518c-3p	0.21012
MIMAT0002868	hsa-miR-522-3p	0.20635
MIMAT0002867	hsa-miR-520h	0.18718
MIMAT0002844	hsa-miR-518b	0.13924
MIMAT0005795	hsa-miR-1323	0.07530
MIMAT0002832	hsa-miR-519c-3p	0.00243

Table 5: List of all significant molecular protein signatures for lasso penalty equal to 4, identified in the optimization process in sparse multiset CCA.

UniProt accession	Name	Score
Q14103	HNRNPD	0.36465
Q6DD88	ATL3	0.31732
Q8WUM4	PDCD6IP	0.31697
P42356	PI4KA	0.30403
P25788	PSMA3	0.30022
Q15029	EFTUD2	0.26541
P11717	IGF2R	0.24830
O00232	PSMD12	0.23779
O15498	YKT6	0.23344
P49411	TUFM	0.23242
O95232	LUC7L3	0.21319
Q9Y6E0	STK24	0.19782
Q16186	ADRM1	0.17537
P07195	LDHB	0.15621
P63151	PPP2R2A	0.13903
P27695	APEX1	0.13848
Q03135	CAV1	0.06421
P62158	CALM3	0.05312
P60174	TPI1	0.02721
Q7L576	CYFIP1	0.01481

Table 6: List of all significant molecular RNA signatures for lasso penalty equal to 4, identified in the optimization process in sparse multiset CCA.

Ensembl accession	Name	Score
ENSG00000197956	S100A6	0.37890
ENSG00000185133	INPP5J	0.34362
ENSG00000150630	VEGFC	0.33445
ENSG00000155011	DKK2	0.33241
ENSG00000109738	GLRB	0.32673
ENSG00000252481	SCARNA13	0.30376
ENSG00000134489	HRH4	0.28534
ENSG00000136918	WDR38	0.24202
ENSG00000165115	KIF27	0.20406
ENSG00000229563	LINC01204	0.19210
ENSG00000168003	SLC3A2	0.16311
ENSG00000266010	GATA6-AS1	0.15169
ENSG00000215388	ACTG1P3	0.10816
ENSG00000138356	AOX1	0.08518
ENSG00000239911	PRKAG2-AS1	0.07523
ENSG00000261308	FIGNL2	0.06394
ENSG00000116039	ATP6V1B1	0.06232
ENSG00000197753	LHFPL5	0.06130
ENSG00000213453	FTH1P3	0.05698
ENSG00000264384	RN7SL431P	0.05629
ENSG00000214875	MED28P1	0.04412
ENSG00000226977	HMGN1P24	0.04029
ENSG00000166341	DCHS1	0.03895
ENSG00000242779	ZNF702P	0.02240
ENSG00000233237	LINC00472	0.02133

ENSG00000206538	VGLL3	0.00552
-----------------	-------	---------

Table 7: List of all significant molecular miRNA signatures for lasso penalty equal to 7, identified in the optimization process in sparse multiset CCA.

miRBase accession	Name	Score
MIMAT0002833	hsa-miR-520a-5p	0.17066
MIMAT0005452	hsa-miR-519a*	0.17065
MIMAT0002855	hsa-miR-520d-5p	0.17064
MIMAT0002849	hsa-miR-524-5p	0.17058
MIMAT0002848	hsa-miR-518c-3p	0.17057
MIMAT0002868	hsa-miR-522-3p	0.17056
MIMAT0002867	hsa-miR-520h	0.17047
MIMAT0002835	hsa-miR-526b-5p	0.17039
MIMAT0002838	hsa-miR-525-5p	0.17031
MIMAT0002859 _ 1	hsa-miR-516b-5p	0.17025
MIMAT0002844	hsa-miR-518b	0.17022
MIMAT0005795	hsa-miR-1323	0.16984
MIMAT0002837	hsa-miR-519b-3p	0.16967
MIMAT0002832	hsa-miR-519c-3p	0.16934
MIMAT0002866	hsa-miR-517c-3p	0.16931
MIMAT0000726	hsa-miR-373-3p	0.16924
MIMAT0004687	hsa-miR-371a-5p	0.16856
MIMAT0002842	hsa-miR-518f-3p	0.16853
MIMAT0000281	hsa-miR-224-5p	0.16850
MIMAT0000724	hsa-miR-372-3p	0.16848
MIMAT0002830	hsa-miR-520f-3p	0.16794
MIMAT0002847	hsa-miR-518c-5p	0.16723
MIMAT0002861	hsa-miR-518e-3p	0.16640

MIMAT0005924	hsa-miR-1270	0.16463
MIMAT0002824	hsa-miR-498-5p	0.16404
MIMAT0002856	hsa-miR-520d-3p	0.16360
MIMAT0002859	hsa-miR-516b-5p	0.16316
MIMAT0002840	hsa-miR-523-3p	0.16258
MIMAT0022281 _ 1	hsa-miR-5583-5p	0.16182
MIMAT0002834	hsa-miR-520a-3p	0.15843
MIMAT0002853	hsa-miR-519d-3p	0.15523
MIMAT0002858	hsa-miR-520g-3p	0.15456
MIMAT0002850	hsa-miR-524-3p	0.14950
MIMAT0004549	hsa-miR-148a-5p	0.11714
MIMAT0004951	hsa-miR-887-3p	0.10334
MIMAT0005796	hsa-miR-1271-5p	0.10201
MIMAT0004927	hsa-miR-708-3p	0.08799
MIMAT0004926	hsa-miR-708-5p	0.08502
MIMAT0000268	hsa-miR-211-5p	0.07882
MIMAT0004761	hsa-miR-483-5p	0.07818
MIMAT0002880	hsa-miR-508-3p	0.07809
MIMAT0003250	hsa-miR-585-3p	0.07557
MIMAT0000103	hsa-miR-106a-5p	0.05487
MIMAT0003283	hsa-miR-615-3p	0.04849
MIMAT0000459	hsa-miR-193a-3p	0.04426
MIMAT0025851	hsa-miR-6720-3p	0.02744
MIMAT0031074	hsa-miR-450a-2-3p	0.02567
MIMAT0000417	hsa-miR-15b-5p	0.02287
MIMAT0004614	hsa-miR-193a-5p	0.01718
MIMAT0004569	hsa-miR-222-5p	0.01484
MIMAT0004983	hsa-miR-940	0.00724
MIMAT0004694	hsa-miR-342-5p	0.00405
MIMAT0000686	hsa-miR-34c-5p	0.00358

MIMAT0000683	hsa-miR-302a-5p	0.00274
--------------	-----------------	---------

Table 8: List of molecular protein signatures for lasso penalty equal to 7, identified in the optimization process in sparse multiset CCA.

UniProt accession	Name	Score
O00232	PSMD12	0.19681
Q8WUM4	PDCD6IP	0.19560
P42356	PI4KA	0.19507
O15498	YKT6	0.19448
P62158	CALM3	0.19163
Q03135	CAV1	0.18745
Q7L576	CYFIP1	0.18670
Q14103	HNRNPD	0.18406
P17980	PSMC3	0.18323
Q9Y5X2	SNX8	0.18062
P60174	TPI1	0.17691
P49411	TUFM	0.17444
Q6DD88	ATL3	0.17330
P11717	IGF2R	0.17021
P63151	PPP2R2A	0.17010
P23634	ATP2B4	0.16648
P25788	PSMA3	0.16556
Q9Y6E0	STK24	0.16220
Q15029	EFTUD2	0.15827
O95232	LUC7L3	0.15728
P53985	SLC16A1	0.15709
P52788	SMS	0.15452
Q16186	ADRM1	0.15121

P35606	COPB2	0.15094
P27695	APEX1	0.14619
P07195	LDHB	0.14193
P04632	CAPNS1	0.13993
Q14697	GANAB	0.13576
Q9BSJ8	ESYT1	0.12507
Q9NSD9	FARSB	0.12280
P20073	ANXA7	0.12135
Q8NBS9	TXNDC5	0.11441
O75695	RP2	0.11364
Q05209	PTPN12	0.10324
P19022	CDH2	0.09823
P00387	CYB5R3	0.09322
Q92499	DDX1	0.09282
Q92841	DDX17	0.08759
O95373	IPO7	0.08576
P08134	RHOC	0.08014
P62701	RPS4X	0.07807
P30048	PRDX3	0.07755
P61224	RAP1B	0.07701
Q13045	FLII	0.06718
Q08211	DHX9	0.06690
Q9UHY7	ENOPH1	0.06172
O43795	MYO1B	0.06032
P21266	GSTM3	0.05777
P00568	AK1	0.05216
P63241	EIF5A	0.05198
A0FGR8	ESYT2	0.05011
O14818	PSMA7	0.04172
Q9Y3I0	RTCB	0.04026

Q12905	ILF2	0.03640
P62333	PSMC6	0.03256
P39656	DDOST	0.03079
P06703	S100A6	0.03024
P78371	CCT2	0.02668
Q07021	C1QBP	0.02351
P13473	LAMP2	0.01901
Q13813	SPTAN1	0.00942
Q9Y230	RUVBL2	0.00938
P40926	MDH2	0.00676
P27708	CAD	0.00242
O95747	OXSRI	0.00231
Q01082	SPTBN1	0.00127
P15121	AKR1B1	0.00027

Table 9: List of all significant molecular RNA signatures for lasso penalty equal to 7, identified in the optimization process in sparse multiset CCA.

Ensembl accession	Name	Score
ENSG00000198431	TXNRD1	0.20199
ENSG00000123358	NR4A1	0.20161
ENSG00000137877	SPTBN5	0.19946
ENSG00000204622	HLA-J	0.19717
ENSG00000249395	CASC9	0.19665
ENSG00000162989	KCNJ3	0.19665
ENSG00000171643	S100Z	0.19665
ENSG00000171451	DSEL	0.19090
ENSG00000145246	ATP10D	0.18376
ENSG00000133169	BEX1	0.17644

ENSG00000237646	HLCS-IT1	0.17489
ENSG000001703827	LRRN2	0.17489
ENSG00000261390	MAFTRR	0.17489
ENSG00000162631	NTNG1	0.16981
ENSG00000104047	DTWD1	0.16736
ENSG00000229944	EIF4EP2	0.16444
ENSG00000197461	PDGFA	0.16196
ENSG00000140563	MCTP2	0.16183
ENSG00000145936	KCNMB1	0.15984
ENSG00000246985	SOCS2-AS1	0.15149
ENSG00000166073	GPR176	0.15093
ENSG00000124789	NUP153	0.14598
ENSG00000164185	ZNF474	0.14248
ENSG00000143545	RAB13	0.13610
ENSG00000169193	CCDC126	0.13605
ENSG00000259378	DCAF13P3	0.13029
ENSG00000233232	NPIP7	0.13029
ENSG00000232487	RASA3-IT1	0.13029
ENSG00000235944	ZNF815P	0.12740
ENSG00000157796	WDR19	0.11496
ENSG00000171847	FAM90A1	0.11487
ENSG00000189043	NDUFA4	0.11399
ENSG00000124788	ATXN1	0.11128
ENSG00000174307	PHLDA3	0.10973
ENSG00000219626	FAM228B	0.09536
ENSG00000126550	HTN1	0.09304
ENSG00000224531	SMIM13	0.09014
ENSG00000261971	MMP25-AS1	0.08757
ENSG00000128641	MYO1B	0.08733
ENSG00000173227	SYT12	0.08555

ENSG00000204779	FOXD4L5	0.08389
ENSG00000153823	PID1	0.08106
ENSG00000224975	INE1	0.07684
ENSG00000241547	ACTG1P20	0.07447
ENSG00000188428	BLOC1S5	0.07271
ENSG00000166257	SCN3B	0.07159
ENSG00000275832	ARHGAP23	0.07082
ENSG00000104728	ARHGEF10	0.07036
ENSG00000174353	STAG3L3	0.06874
ENSG00000141738	GRB7	0.05659
ENSG00000169933	FRMPD4	0.05504
ENSG00000124787	RPP40	0.04788
ENSG00000228507	DAP3P2	0.04629
ENSG00000144908	ALDH1L1	0.03779
ENSG00000075213	SEMA3A	0.02884
ENSG00000183032	SLC25A21	0.02286
ENSG00000230461	PROX1-AS1	0.02209
ENSG00000180777	ANKRD30B	0.01857
ENSG00000006715	VPS41	0.01512
ENSG00000198482	ZNF808	0.01427
ENSG00000145390	USP53	0.01055
ENSG00000105223	PLD3	0.00741
ENSG00000182308	DCAF4L1	0.00517
ENSG00000162738	VANGL2	0.00246
ENSG00000145216	FIP1L1	0.00228

.6 Melanoma EVs biomarkers research

To complete the literature review I used EV-TRACK: transparent reporting and centralizing knowledge in extracellular vesicle research [178]. I generated a report that was used to verify theoretical research in the field of biomarkers and omics research for EVs in melanoma. A total of 24 studies were found. All the research on melanoma biomarker or omics data found by EV Track was on exosomes.

First Author	Year	PMID	Study Aim	EV-TRACK ID	EV type
Valenti R.	2006	16982774	Function	EV200173	microvesicles
Temme S.	2010	19949097	Function	EV100029	exosomes
Zhang J.	2010	20845808	Function	EV100096	exosomes
van den Boorn J.G.	2011	21326294	Other	EV110118	exosomes
Ren W.N.	2011	21574092	Function	EV110110	exosomes
Peinado H.	2012	22635005	Function	EV120019	exosomes
Xiao D.	2012	23056502	Omics	EV120097	exosomes
Lee J.H.	2013	23317503	Other	EV130135	EVs
Rappa G.	2013	23767874	Function	EV130053	exosomes
Park S.	2014	25484139	Function	EV140038	exosomes
Eldh M.	2014	25510783	Biomarker	EV140153	exosomes
Guan M.	2014	25480418	Biomarker	EV140162	exosomes
Aspe J.R.	2014	24624263	Function	EV140299	exosomes
Ekström E.J.	2014	24766647	Function	EV140328	exosomes
Cossetti C.	2014	24992257	Function	EV140319	exosomes
Hatanaka M.	2014	25503136	Function	EV140340	exosomes
Federici C.	2014	24516610	Function	EV140330	exosomes
Alegre E.	2014	24878024	Biomarker	EV140294	exosomes
Lazar I.	2015	25950383	Biomarker	EV150011	exosomes
Takeda Y.S.	2015	26248331	Function	EV150056	exosomes
Ragusa M.	2015	25951497	Biomarker	EV150062	exosomes
Klump J.	2018	29288729	Biomarker	EV170037	EVs
Crescitelli R.	2019	32128073	Omics	EV190108	EVs
García-Silva S.	2019	30975894	Biomarker	EV190038	exosome

Figure 1: EV Track report.