Breast Cancer Intrinsic Subtypes: A Critical Conception in Bioinformatics

Heloisa Helena Zaccaron Milioli

B.Sc. in Biological Sciences

M.Sc. in Genetics

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Newcastle

Faculty of Science and Information Technology

School of Environmental and Life Sciences

Callaghan, NSW

Australia

September, 2016

Statement of Originality

The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository[1]*, subject to the provisions of the Copyright Act 1968.*

September, 2016

_____

Heloisa Helena Zaccaron Milioli

_____

Prof. Pablo Moscato

[1] Unless an Embargo has been approved for a determined period

Statement of Authorship

I hereby certify that the work embodied in this thesis contains a published paper/s/scholarly work of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publication/s/scholarly work.

September, 2016

_____

Heloisa Helena Zaccaron Milioli

_____

Prof. Pablo Moscato

Acknowledgements

I would like to express my deep gratitude to Prof Pablo Moscato. I appreciated the guidance and encouragement he has provided throughout my time at Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine (CIBM). I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. I would like to thank my co-supervisors, A/Prof Regina Berretta and Dr Jannette Sakoff, for their professional advices and useful critiques.

Special thanks should be given to Dr Carlos Riveros for his patient assistance and unconditional support. For all the extra hours he spent working with me, the constructive criticism and friendly advice during my PhD. I am sincerely grateful for sharing his truthful and illuminating views on a number of issues related to the breast cancer project. I am also grateful to Dr Renato Vimieiro for the valuable help in data management, extensive collaboration and magic proofreading. In particular, I thank Inna Tishchenko for her precious effort and intelligence in the analysis of the data.

I would like to extend my thanks to all CIBM students and collaborators who contributed with valuable discussions and enthusiastic encouragements: Ademir Cristiano Gabardo, Ahmed Shamsul Arefin, Amer Abu Zaher, Amir Salehipour, Chloe Warren, Claudio Sanhueza, Francia Jimenez, Leila Moslemi Naeni, Luke Mathieson, Mohammad Nazmul Haque, Nasimul Noman, Natalie de Vries, Nisha Puthiyedth, Shannon Fenn. Thanks for sharing the experience, either positive or negative. I also acknowledge the Hunter Medical Research Institute (HMRI) and University of Newcastle (UoN) staffs for sharing an amazing productive environment.

I express my warm thanks to Jennie Thomas for her enthusiastic support of students and researchers through a number of grants and scholarships. Being part of your *family* is a great honour and an enormous pleasure. Thanks for believing in my research and for funding my dreams. I am also grateful to A/Prof David Wild for guiding me throughout my visit to Bloomington, US.

Special thanks to my beloved family for their unconditional support!

Words cannot express how grateful I am to my mother, father, aunt, brother and nephew for all of the sacrifices that you've made on my behalf. Your positive energies have sustained me thus far. I would also like to thank my in-laws for striving towards my goal. Finally, I would like express appreciation to my beloved husband, Jorge André Martins. I would not be here without him, his patience and love.

To the best grandmother,

Helena Mafioleti Zaccaron

(Wherever you are)

Table of Contents

List of Figures

List of Tables

List of Equations

Abbreviations

AACR	Australasian Association of Cancer Registries
ACS	American Cancer Society
AIHW	Australian Institute of Health and Welfare
AJCC	American Joint Committee on Cancer
AR	Androgen receptor
ARI	Adjusted Rand Index
BL1	Basal-like 1
BL2	Basal-like 2
BLBC	Basal-like breast cancer
BLIA	Basal-like immune-activated
BLIS	Basal-like immune-suppressed
ChIP-chip	Chromatin immunoprecipitation on chip
CIBEX	Center for information biology gene expression database
CIBM	Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine
CGH	Comparative genomic hybridization
CNA	Copy number aberration
CNV	Copy number variation
CTD	Comparative Toxicogenomic Database
DamID	DNA adenine methyltransferase identification
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
EGA	European Genome-Phenome Archive
EpCAM	Epithelial cell adhesion molecule
ER	Oestrogen receptor
FGED	Functional Genomics Data Society
FOIPPA	Freedom of Information and Protection of Privacy Act
FS	Feature Selection
GEO	Gene Expression Omnibus
HER2	Human epidermal growth factor receptor 2
HREC	Human Research Ethics Committee

HTC	High content screening
HTS	High-throughput screening
ICGC	International Cancer Genomics Consortium
IDC	Invasive ductal carcinoma
IHC	Immunohistochemical
IHGSC	International Human Genome Sequencing Consortium
ILC	Invasive lobular carcinoma
IM	Immunomodulatory
JS	Jensen Shannon
Ki-67	Antigen identified by monoclonal antibody Ki-67
***k*NN**	*k* nearest neighbours
LAR	Luminal androgen receptor
lincRNA	long intergenic non-coding RNA
MA	Memetic algorithm
MCC	Matthews' Correlation Coefficient
MDL	Minimum Description Length Principle
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MIAME	Minimum Information About a Microarray Experiment
microRNA	miRNA
MGED	Microarray Gene Expression Data Society
MS	Menopausal status
MST	Minimum Spanning Tree
NCBI	National Center for Biotechnology Information
NOS	Not otherwise specified
NPI	Nottingham prognostic score
NSC	Nearest Shrunken Centroids
NST	No special type
ORF	Open reading frame
PIPA	Personal Information Protection Act
PIPEDA	Personal Information Protection and Electronic Documents Act
PR	Progesterone receptor
PRC	Priority Research Centres
RHD	Research Higher Degree
RNA	Ribonucleic acid
ROCK	Research Online Cancer Knowledgebase
RT-PCR	Reverse-transcriptase Polymerase chain reaction

SAM	Sentrix® Array Matrix
SCM	Subtype Classification Model
SNP	Single nucleotide polymorphism
SSP	Single Sample Predictor
TCGA	The Cancer Genome Atlas
TEND	Trends in the Exploration of Novel Drug targets
TNBC	Triple-negative breast cancer
TNM	Tumour size, nodes, metastasis
TTD	Therapeutic Target Database
UCSC	University of California Santa Cruz
WEKA	Waikato Environment for Knowledge Analysis

Achievements

During my PhD, I applied for grants; submitted manuscripts for publication; and attended workshops, conferences and seminars. The relevant achievements are listed as follows:

Grants Awarded

- Hunter Medical Research Institute, 2014.

JENNIE THOMAS MEDICAL RESEARCH TRAVEL GRANT (AUD $10,000)

- Hunter Cancer Research Alliance, 2015.

HCRA TRAVEL GRANT (AUD $1,000)

- Hunter Cancer Research Alliance, 2016.

HCRA PhD Research Award 2016 (AUD $5,000).

- EMBL Australia PhD, 2016.

Travel Grant to attend the 18th EMBL PhD Symposium (AUD $3,000).

- XII ELAG Course Fellowship (USD $700)

Instituto Genética Para Todos – Brazil (unable to attend)

Papers Published in Journals

MILIOLI, H.H.; VIMIEIRO, R.; RIVEROS, C.; TISHCHENKO, I.; BERRETTA, R.; MOSCATO, P. (2015) The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the original PAM50 labels in the METABRIC data set. *PLoS One*; 10(7): 0129711. doi: 10.1371/journal.pone.0129711

MILIOLI, H.H. (2015). The IMPAKT of breast cancer research: fundamental science and clinical medicine. *Future Science OA*; (0). doi: 10.4155/fso.15.69

MILIOLI, H.H.; VIMIEIRO, R.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. (2016) Iteratively refining breast cancer intrinsic subtypes in the METABRIC dataset *BioData Mining*; 9:2. doi: 10.1186/s13040-015-0078-9

TISHCHENKO, I.; **MILIOLI, H.H.**; RIVEROS, C.; MOSCATO, P. (2016) Extensive Transcriptomic and Genomic Analysis Provides New Insights about Luminal Breast Cancers. *PLoS One; 11*(6): e0158259. doi: 10.1371/journal.pone.0158259

MILIOLI, H.H. Life as an early career researcher: interview with Heloisa Helena Milioli**.** *Future Science OA*; 1(4) (2016). doi: 10.4155/fsoa-2016-0033.

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC Med Genomics*; 10(1):19 (2017). doi: 10.1186/s12920-017-0250-9.

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; BERRETTA, R.; MOSCATO, P. Meta-features modelling gene expression imbalances: an innovative strategy for breast cancer subtype prediction. Manuscript in preparation to be submitted for publication at Genomics, Proteomics and Bioinformatics (GPB).

Abstracts Published

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; SAKOFF, J.; BERRETTA, R.; MOSCATO, P. Consensus on breast cancer cell lines classification for an effective and efficient clinical decision-making. *IMPAKT 2015 Breast Cancer Conference. Annals of Oncology* 26 (suppl 3):iii32-iii33 (2015). doi: 10.1093/annonc/mdv121.08

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Molecular classification of basal-like breast cancer subtypes based on predictive survival markers. *IMPAKT 2015 Breast Cancer Conference. Annals of Oncology.* 26 (suppl 3):iii17-iii18 (2015). doi: 10.1093/annonc/mdv117.11

MILIOLI, H.H., TISHCHENKO, I., RIVEROS, C., BERRETTA, R. & MOSCATO, P. (2015) Basal-like breast cancer subgroups uncovered by genomic and transcriptomic profiles and overall survival outcomes. *Hunter Cancer Research Alliance Annual Symposium*. *Asia-Pacific Journal of Clinical Oncology* 11(Suppl. 5):6-19 (2015). doi: 10.1111/ajco.12444

TISHCHENKO, I., **MILIOLI, H.H.**, RIVEROS, C. & MOSCATO, P. How intrinsic are luminal breast cancer subtypes? *Hunter Cancer Research Alliance Annual Symposium*. *Asia-Pacific Journal of Clinical Oncology* 11(Suppl. 5):6-19 (2015). doi: 10.1111/ajco.12444

MILIOLI, H.H., SANHUEZA, C., BERRETTA, R. & MOSCATO, P. (2015) ABSTRACT P40 Breast Cancer Molecular Portraits of Intrinsic Subtypes and Integrative Clusters in the METABRIC Data Set. *Hunter Cancer Research Alliance Annual Symposium*. *Asia-Pacific Journal of Clinical Oncology* 12(Suppl. 6):13-34 (2016). doi: 10.1111/ajco.12618

Oral Presentations

MILIOLI, H.H.; VIMIEIRO, R.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Refining the breast cancer molecular subtypes in the METABRIC data set. *World Congress on Controversies in Breast Cancer (CoBRA), 2015. Melbourne, AU.*

MILIOLI, H.H.; SANHUEZA, C.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Breast cancer molecular portraits of intrinsic subtypes and integrative clusters in the METABRIC data set. **Young Scientist Award** *2nd World Congress on Controversies in Breast Cancer (CoBrCa) 2016. Barcelona, Spain.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancers uncovered by genomic and transcriptomic profiles and patients' overall survival. *Sydney Cancer Conference (SCC) 2016. Sydney, AU.*

Poster Sessions

MILIOLI, H.H.; VIMIEIRO, R.; RIVEROS, C.; SAKOFF, J.; BERRETTA, R.; MOSCATO, P. Breast Cancer Subtypes Individuation Driving Novel Drug Targets for Tailored Therapies. *Translational Cancer Research Conference, 2013. Newcastle, AU.*

MILIOLI, H.H.; VIMIEIRO, R.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Identification of novel biomarkers for predicting breast cancer intrinsic subtypes. *ASMR Satellite Scientific Meeting, 2014. Newcastle, AU.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; MOSCATO, P. Meta-features as predictors of breast cancer intrinsic subtype in the METABRIC gene expression data set. **Best Poster Award (Bronze Medal)** *International Conference on Bioinformatics, 2014. Sydney, AU.*

RIVEROS, C.; **MILIOLI, H.H.**; VIMIEIRO, R.; BERRETTA, R.; MOSCATO, P. Discovery of gene interactions by GPU-enabled computation of pairwise expression level metafeatures. *International Conference on Bioinformatics, 2014. Sydney, AU.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; TISHCHENKO, I.; BERRETTA, R.; MOSCATO, P. Using an iterative approach to reclassify sample subtypes in the METABRIC breast cancer data set. **Best Poster Award (Third place)** *BioInfoSummer, 2014. Melbourne, AU.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancer subsets revealed by survival predictor genes. *ASMR Satellite Scientific Meeting, 2015. Newcastle, AU.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Molecular classification of basal-like breast cancer subtypes based on predictive survival markers. *IMPAKT 2015 Breast Cancer Conference. Brussels, BE.*

MILIOLI, H.H.; TISHCHENKO, I.; RIVEROS, C.; SAKOFF, J.; BERRETTA, R.; MOSCATO, P. Consensus on breast cancer cell lines classification for an effective and efficient clinical decision-making. *IMPAKT 2015 Breast Cancer Conference. Brussels, BE.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; MOSCATO, P. Meta-features predicting gene expression imbalances across breast cancer intrinsic subtypes. *EMBL Australia PhD Symposium, 2015. Melbourne, AU.*

TISHCHENKO, I., **MILIOLI, H.H.**, RIVEROS, C. & MOSCATO, P. How intrinsic are luminal breast cancer subtypes? *Hunter Cancer Research Alliance Annual Symposium, 2015. Newcastle, AU.*

MILIOLI, H.H., TISHCHENKO, I., RIVEROS, C., BERRETTA, R. & MOSCATO, P. Basal-like breast cancer subgroups uncovered by genomic and transcriptomic profiles and overall survival outcomes. *Hunter Cancer Research Alliance Annual Symposium, 2015. Newcastle, AU.*

NAENI, L., **MILIOLI, H.H.**, TISHCHENKO, BERRETTA, R. & MOSCATO, P. (2015) A New Clustering Approach Identifies Candidate Biomarkers for Breast Cancer Subtyping. *BioInfoSummer, 2015*. *Sydney, AU.*

MILIOLI, H.H.; RIVEROS, C.; VIMIEIRO, R.; MOSCATO, P. Meta-features predicting gene expression imbalances across breast cancer intrinsic subtypes. **Best Poster Presentation** *BioInfoSummer, 2015. Sydney, AU.*

Other Presentations

Confirmation Year Presentation

Faculty of Science and IT. The University of Newcastle, 2013.

RHD candidates are required to submit the 'Confirmation Year Report' and present the research overview. In August 2013, I presented the preliminary results in the Faculty of Science and IT as an open seminar.

HCRA, ECR and PhD Student (HEAPS) Seminar Series

Hunter Medical Research Institute. The University of Newcastle, 2014 and 2015.

The HEAPS seminar series are organised by the Hunter Cancer Research Alliance (HCRA) for RHD students and supervisors. It is an opportunity for researchers to practice presenting (and critiquing) work in a local and highly supportive environment. In 2014 and 2015, I presented and discussed the results of my research as well as supported other researchers' work.

HUBS3302 Bioinformatics Mini-Conference

Faculty of Health and Medicine. The University of Newcastle, 2014 and 2015.

The purpose of this event is to inspire students in the field and, specially, in their final project for the discipline. In the 2014 and 2015 Bioinformatics Mini-Conference, organised by Belinda Goldie, I presented my research on breast cancer.

Science and Engineering Challenge

Faculty of Engineering and Built Environment. The University of Newcastle, 2014, 2015 and 2016.

The 'Science and Engineering Challenge' organise a number of events aimed at challenging students of all different ages in Science and Engineering. As part of the team, I coordinated activities in Tamworth (2014), Muswellbrook (2014), Dubbo (2015), Newcastle (2015), Central Coast (2016) and Narrabri (2016), and presented my research to the Rotary International (Australian Rotary Districts) in Tamworth and Dubbo.

Faculty Progress Seminar

Faculty of Science and IT. The University of Newcastle, 2015.

Students in the Faculty of Science and IT are required to present a Progress Seminar after completing 2 to 3 years of a PhD. In June 2015, I discussed the overall aims and results of my research and outlined my thesis to fellow RHD candidates and academics in the school.

Google Computer Science for High Schools

Faculty of Engineering and Built Environment. The University of Newcastle, 2015 and 2016.

The University of Newcastle's Computer Science 4 High Schools (CS4HS) is an introductory workshop for in-service and pre-service teachers (both at primary and secondary level), and career advisors focused on developing competencies included in the recently approved Digital Technologies curriculum and is accredited by BOSTES. In three events, I had the opportunity to explain the relevance of computer science to analyse biological/medical data.

Relevant Activities

Course: **Winter School in Mathematical and Computational Biology**

University of Queensland (UQ), Brisbane, 2013.

The winter school introduced mathematical and computational biology and bioinformatics to advanced undergraduate and postgraduate students, postdoctoral researchers and others working in the field. Important topics, such as mathematics, statistics, computer science, information technology, biology, chemistry and medical sciences and engineering, were selected for each day. Lectures and interactive discussions were ministered by national and international authorities.

Course: **European Molecular Biology Laboratory (EMBL) Australia PhD Course**

Australian National University (ANU), Canberra, 2014.

EMBL Australia offered to sixty students a unique introduction to research with the annual EMBL Australia PhD Course. The two-week program shows students how their research fits into the bigger picture of science, and introduces a range of fields including: bioinformatics, developmental biology, genomics, systems biology and regenerative medicine.

Course: **European Molecular Biology Laboratory (EMBL) Australia PhD Course**

Welcome Genome Campus, Hinxton, UK, 2016.

This course introduced a wide range of post-genome techniques including practical experience in performing (1) high-throughput RNAi screening, (2) microarray gene expression analysis and interpretation, using a range of commercial and academic software tools, (3) next-generation sequencing and alignment; (4) protein-protein interaction networks and integration with other data sources, and (5) pathway analysis. Laboratory work was based on the training of state-of-the-art methods and complementary approaches to address biological and medical questions.

Training: **Collaborative Research Training in Human Genetics and Bioinformatics**

Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine (CIBM). The University of Newcastle, 2014.

The CIBM established a research-training program in 2014 that contributed to improve the capacity of young investigators to conduct human genetics and bioinformatics research. The training promoted scientific collaborations between the University of Newcastle and international (undergraduate) students. The proposed program provided opportunities to generate expertise that could contribute to the

long-term goal of harnessing genetic knowledge and bioinformatics skills to diagnose, prevent, or treat diseases. Training activities were coordinated, facilitated and monitored by Prof. Pablo Moscato, A/Prof Regina Berretta and PhD student Heloisa Helena Milioli.

Short-term Exchange Program: **Cheminformatics and Chemogenomics Research Group (CCRG)**

Indiana University (IU), Bloomington USA, 2015.

Further investigation on cheminformatics and toxicogenomics has been developed in collaboration with A/Prof. David J. Wild (May/June 2015), at the School of Informatics and Computing in Bloomington (USA). These approaches were used to delineate drug-targets for basal-like breast cancer, one of the most aggressive subtypes with limited therapy response. Further research, however, is required to design and perform in vitro tests.

Organising Committee: **Australian Society for Medical Research (ASMR) Satellite Scientific Meeting**

Hunter Medical Research Institute (HMRI), Newcastle, 2015.

This event showcases the recent research achievements of Hunter scientists, encourages postgraduate and student interactions and fosters collaboration between researchers within the Faculty of Health and Medicine, HMRI and the international community. In the 2015 edition, I was member of the committee.

Abstract

Breast cancers have been uncovered by high-throughput technologies that allow the investigation at the genomic, transcriptomic and proteomic levels. In the early 2000s, the gene expression profiling has led to the classification of five intrinsic subtypes: luminal A, luminal B, HER2-enriched, normal-like and basal-like. A decade later, the spectrum of copy number aberrations has further expanded the heterogeneous architecture of this disease with the identification of 10 integrative clusters (IntClusts). The referred classifications aim at explaining the diverse phenotypes and independent outcomes that impact clinical decision-making. However, intrinsic subtypes and IntClusts show limited overlap. In this context, novel methodologies in bioinformatics to analyse large-scale microarray data will contribute to further understanding the molecular subtypes. In this study, we focus on developing new approaches to cover multi-perspective, highly dimensional, and highly complex data analysis in breast cancer. Our goal is to review and reconcile the disease classification, underlying the differences across clinicopathological features and survival outcomes. For this purpose, we have explored the information processed by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC); one of the largest of its type and depth, with over 2000 samples. A series of distinct approaches combining computer science, statistics, mathematics, and engineering have been applied in order to bring new insights to cancer biology. The translational strategy will facilitate a more efficient and effective incorporation of bioinformatics research into laboratory assays. Further applications of this knowledge are, therefore, critical in order to support novel implementations in the clinical setting; paving the way for future progress in medicine.

Keywords

Breast cancer, Intrinsic subtypes, Integrative clusters, IntClusts, Microarray, Gene expression, Copy number aberration, MicroRNA, METABRIC, Feature selection, Data mining, Ensemble learning, Prediction models, Classification