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 Repository¹, 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

Acknowledgements V  
Table of Contents IX  
List of Figures XIII  
List of Tables XV  
List of Equations XVII  
Abbreviations XIX  
Achievements XXIII

## CHAPTER 1  

1. **INTRODUCTION AND OVERVIEW**  
1.1 Breast Cancer: an Overview 2  
1.2 Bioinformatics Resources and Tools 4  
1.3 Research Motivation 6  
   1.3.1 Research Questions 7  
1.4 Research Aims and Thesis Structure 7  
1.5 References 11

## CHAPTER 2  

2. **BREAST CANCER: CURRENT STATUS AND PERSPECTIVES**  
2.1 Breast Carcinogenesis 17  
2.2 The Breast Tumour Classification 19  
2.3 Intrinsic Subtypes 24  
   2.3.1 Luminal A and B 25  
   2.3.2 HER2-enriched 26  
   2.3.3 Basal-like 26  
   2.3.4 Normal-like 28  
   2.3.5 Other groups 28  
2.4 Novel Integrative Clusters 29  
2.5 Predicting Molecular Subtypes 30  
2.6 References 32
CHAPTER 5

5. **ITERATIVELY REFINING THE METABRIC SUBTYPE LABELS**

5.1 Introduction
5.2 Methods
  5.2.1 Transcriptomic Data Set
  5.2.2 The Refinement Method
  5.2.3 The CM1 Score
  5.2.4 Statistical Analysis
  5.2.5 Clinical Data and Survival Curves
5.3 Results and Discussion
  5.3.1 Discriminative Probes Used to Assign Intrinsic Subtype Labels in the Refinement Process
  5.3.2 New Subtype Labels Reveal More Reliable Distribution of Clinical Markers and Survival Outcomes
5.4 Conclusion
5.5 References
5.6 Supporting Information

CHAPTER 6

6. **META-FEATURES FOR PREDICTING BREAST CANCER INTRINSIC SUBTYPES**

6.1 Introduction
6.2 Methods
  6.2.1 Ethics Statement and Data Description
  6.2.2 Study Design and Computing Resources
  6.2.3 Statistical Analysis
6.3 Results and Discussion
  6.3.1 Thirteen Meta-features Define Breast Cancer Intrinsic Subtypes
  6.3.2 An Ensemble Learning Approach Validates the Quality of Meta-features for Predicting Subtypes
  6.3.3 Expanding Prediction Models Based on Microarray Data
6.4 References
6.5 Supporting Information

CHAPTER 7
### 7. Basal-like Breast Cancer Subtype

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Introduction</td>
<td>184</td>
</tr>
<tr>
<td>7.2 Methods</td>
<td>186</td>
</tr>
<tr>
<td>7.2.1 Breast Cancer Data Sets</td>
<td>186</td>
</tr>
<tr>
<td>7.2.2 Probe Selection Approach</td>
<td>187</td>
</tr>
<tr>
<td>7.2.3 Clustering Basal-like Breast Cancer Samples</td>
<td>188</td>
</tr>
<tr>
<td>7.2.4 Validation across Data Sets</td>
<td>189</td>
</tr>
<tr>
<td>7.2.5 Network Analysis</td>
<td>189</td>
</tr>
<tr>
<td>7.2.6 MicroRNA Differential Expression</td>
<td>190</td>
</tr>
<tr>
<td>7.2.7 Copy Number Aberration Profiles</td>
<td>190</td>
</tr>
<tr>
<td>7.3 Results</td>
<td>191</td>
</tr>
<tr>
<td>7.3.1 Survival-related Probes Defining Basal-like Subgroups</td>
<td>191</td>
</tr>
<tr>
<td>7.3.2 Basal I and Basal II Validated across Independent Data Sets and Microarray Platforms</td>
<td>200</td>
</tr>
<tr>
<td>7.3.3 Clinical Features and Survival Outcomes Supporting the Basal-like Subgroups</td>
<td>200</td>
</tr>
<tr>
<td>7.3.4 MicroRNAs Differentially Expressed between Basal I and Basal II</td>
<td>203</td>
</tr>
<tr>
<td>7.3.5 Copy Number Aberration Profiles Further Differentiating Basal-like Subgroups</td>
<td>206</td>
</tr>
<tr>
<td>7.4 Discussion</td>
<td>209</td>
</tr>
<tr>
<td>7.4.1 Survival-related Probes Defining the Molecular Signature of Basal-like Breast Cancer Subgroups</td>
<td>209</td>
</tr>
<tr>
<td>7.4.2 MicroRNA Expression Levels Differentiating Basal I from Basal II</td>
<td>210</td>
</tr>
<tr>
<td>7.4.3 Genomic Aberrations Further Characterise Basal II and Basal I Subgroups</td>
<td>212</td>
</tr>
<tr>
<td>7.4.4 Consensus on the Analysis of Basal-like Breast Cancer Subtypes: a Literature Overview</td>
<td>213</td>
</tr>
<tr>
<td>7.5 Conclusion</td>
<td>215</td>
</tr>
<tr>
<td>7.6 References</td>
<td>216</td>
</tr>
<tr>
<td>7.7 Supporting Information</td>
<td>225</td>
</tr>
</tbody>
</table>

### 8. Concluding Remarks

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1 Final Statements</td>
<td>242</td>
</tr>
<tr>
<td>8.2 Future Directions</td>
<td>246</td>
</tr>
<tr>
<td>8.3 Closing Note</td>
<td>248</td>
</tr>
</tbody>
</table>
List of Figures

Figure 3.1 Conceptual view of a cRNA microarray processing. ........................................ 42
Figure 4.1 The step-by-step process ............................................................................ 59
Figure 4.2 The gene expression profile of the balanced top ten probes selected for each of the five breast cancer intrinsic subtypes across 997 samples from the discovery set .................. 69
Figure 4.3 Gene expression patterns of the 42 probes selected using the CM1 score ........ 70
Figure 4.4 The mRNA log2 normalised expression values of 7 novel highly discriminative biomarkers across the five intrinsic subtypes .......................................................... 71
Figure 4.5 Class distribution in the METABRIC discovery and validation, and ROCK set .... 77
Figure 4.6 Similarity between subtypes distribution in the METABRIC discovery and validation sets, and in the ROCK set .................................................................................. 79
Figure 4.7 ER marker distribution across subtypes in the METABRIC data sets ............. 81
Figure 4.8 PR marker distribution across subtypes in the METABRIC data sets ............. 82
Figure 4.9 HER2 distribution across subtypes in the METABRIC data sets ..................... 83
Figure 4.10 The survival curves for METABRIC discovery and validation sets .......... 84
Figure 4.11 The mRNA log2 normalised expression values of 42 probes (A and B) in the CM1 list across the five intrinsic subtypes in the METABRIC discovery and validation, and ROCK 97
Figure 5.1 Refinement Method ..................................................................................... 128
Figure 5.2 The heat map of refined intrinsic features selected using CM1 score ............ 131
Figure 5.3 The survival curves for original and refined labels in the METABRIC discovery and validation sets ............................................................................................................. 133
Figure 5.4 Mean Final Classifier Performance, as measured by Fleiss' κ against the final ensemble learning labels of all samples, across the 10 different refinement runs .............. 141
Figure 5.5 Evolution of performance of classifiers along iterations in a typical refinement run. The κ values are measured against final ensemble learning labels ............................................................. 142
Figure 5.6 MST- kNN clustering, coloured according to the original METABRIC labels defined by the PAM50 method ................................................................................................. 145
Figure 5.7 MST- kNN clustering, coloured according to the refined labels using an iterative process ................................................................................................................................. 146
Figure 5.8 MST- kNN clustering, coloured according to the IntClust classification proposed by Curtis et al. (2012) .............................................................................................................. 147
Figure 6.1 Summary systematic approach ..................................................................... 159
Figure 6.2 Meta-features selected with the CM1 score in the METABRIC discovery set .... 164
Figure 6.3 Gene expression patterns of the 13 meta-features selected using the CM1 score and $(\alpha,\beta)$-k-Feature set

Figure 6.4 Pairwise expression patterns across intrinsic subtypes in the METABRIC discovery and validation sets

Figure 6.5 Individual expression patterns across intrinsic subtypes in the METABRIC discovery and validation sets

Figure 6.6 Graph representing an instance of the $(\alpha,\beta)$-k-Feature Set; as per the data defined in Table 6.5.

Figure 6.7 Graph containing a feasible solution for the $(\alpha,\beta)$-k-Feature Set problem; as per the data defined in Table 6.5.

Figure 7.1 Heat map of the 80-genes signature in METABRIC training set

Figure 7.2 Minimum Spanning Tree of the 80-probe signature

Figure 7.3 Survival curves in the METABRIC and ROCK data sets

Figure 7.4 The box plot of miRNAs differentiating Basal I and Basal II subgroups

Figure 7.5 Copy number aberration of basal subgroups in METABRIC data set

Figure 7.6 The heat map of 400 probes in METABRIC training set

Figure 7.7 Network analysis of multiple drug targets for breast cancer therapy

Figure 8.1 t-SNE graph of METABRIC samples coloured according to PAM50

Figure 8.2 t-SNE graph of METABRIC samples coloured using the refined labels
List of Tables

Table 2.1 Primary Tumour (T) ................................................................. 21
Table 2.2 Regional Lymph Nodes (N) ...................................................... 22
Table 2.3 Distant Metastasis (M) ............................................................ 22
Table 2.4 Anatomic stage/prognostic groups .......................................... 23
Table 3.1 METABRIC microarray data description ............................... 47
Table 3.2 Data accession – gene expression and genotyping information ......................................................... 48
Table 3.3 Data accession – microRNA expression information ................ 49
Table 3.4 Overview of the ten data sets in the ROCK online portal ......... 51
Table 4.1 CM1 List ................................................................................. 66
Table 4.2 Scores and ranks for the CM1 list ........................................... 67
Table 4.3 The ensemble learning overall performance on assigning labels to samples in the METABRIC discovery and validation sets, and ROCK test set ......................................................... 73
Table 4.4 Contingency tables for predicted labels using classifiers trained with the CM1 list .. 73
Table 4.5 Contingency tables for predicted labels using classifiers trained with the PAM50 list .............................................................. 73
Table 4.6 Contingency tables for predicted labels using classifiers trained with CM1 and PAM50 lists ................................................................................. 74
Table 4.7 Agreement of the 24 classifiers on assigning labels using Fleiss' kappa statistic .......... 75
Table 4.8 Agreement measured by the Adjusted Rand Index between different labelling .... 76
Table 4.9 The CM1 score calculated for each breast cancer subtype .......... 91
Table 4.10 Summary performance of the classifiers using the CM1 list .... 92
Table 4.11 Summary performance of the classifiers using the PAM50 list .. 94
Table 4.12 The agreement between sample labelling with Fleiss’ Kappa measure and the Jensen-Shannon divergence of two probability distributions ......................................................... 95
Table 4.13 The Jensen-Shannon divergence of two probability distributions ................ 96
Table 5.1 Contingency table for predicted labels vs. initial subtypes (rows and columns, respectively) ............................................................................. 130
Table 5.2 Number of samples for each clinical marker in the METABRIC data set according to the PAM50 method and refinement process ......................................................... 132
Table 5.3 Refined subtype labels in the METABRIC data set .................. 137
Table 5.4 List of the 24 classifiers used in the ensemble learning .......... 137
Table 5.5 Average agreement of classifiers per subtype ......................... 138
Table 5.6 Probe appearance after ten iterative processes and the respective annotation based on Dunning et al. (2010) and Illumina array data................................................................. 139
Table 5.7 The percentage of PAM50 labels matching integrative clusters (IntClust 1-10) in the METABRIC study.................................................................................................................. 148
Table 5.8 The percentage of Refined labels matching integrative clusters (IntClust 1-10) in the METABRIC study .................................................................................................................. 149
Table 6.1 List of meta-features selected with CM1 score and $\alpha,\beta$-k Feature set ............... 163
Table 6.2 Contingency tables for predicted labels using ensemble learning trained with 13 meta-features Discovery set Validation set ........................................................................................ 168
Table 6.3 Performance of 22 Weka classifiers on predicting labels in the METABRIC discovery and validation sets .................................................................................................................. 169
Table 6.4 Fleiss’ kappa values and Adjusted Rand Index for the discovery and validation sets170
Table 6.5 An example of numerical matrix with five features and six samples belonging to class $F$ or $G.$ .......................................................................................................................... 177
Table 7.1 The 80-genes signature related to survival................................................................. 198
Table 7.2 Clinical information of patients and tumour samples in the METABRIC data set ... 202
Table 7.3 MicroRNAs differentiating basal-like breast cancer subgroups............................... 203
Table 7.4 MicroRNAs and corresponding target genes............................................................. 204
Table 7.5 Cytobands associated with significant CNA acquisitions ........................................ 208
Table 7.6 Basal-like samples classification for the validation set ............................................. 225
Table 7.7 Basal-like samples classification for the validation set ............................................. 225
Table 7.8 The centroids computed for differentiating Basal I and Basal II............................... 225
Table 7.9 The functional annotation of G1 probes according to DAVID ................................. 225
Table 7.10 The functional annotation of G2 probes according to DAVID ................................. 225
Table 7.11 The functional annotation of G3 probes according to DAVID ................................. 225
Table 7.12 MicroRNAs differentiating Basal I and Basal II...................................................... 226
Table 7.13 MicroRNAs and gene targets in Basal I................................................................. 227
Table 7.14 MicroRNAs and gene targets in Basal II................................................................. 230
Table 7.15 Summary gene targets and corresponding drugs................................................. 237
List of Equations

Equation 4.1 CM1 score ........................................................................................................ 60
Equation 4.2 Cramer's V .................................................................................................... 62
Equation 4.3 Average sensitivity ....................................................................................... 62
Equation 4.4 Fleiss' kappa ................................................................................................. 63
Equation 4.5 Adjusted Rand Index ..................................................................................... 63
Equation 7.1 Normalisation ............................................................................................... 189
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACR</td>
<td>Australasian Association of Cancer Registries</td>
</tr>
<tr>
<td>ACS</td>
<td>American Cancer Society</td>
</tr>
<tr>
<td>AIHW</td>
<td>Australian Institute of Health and Welfare</td>
</tr>
<tr>
<td>AJCC</td>
<td>American Joint Committee on Cancer</td>
</tr>
<tr>
<td>AR</td>
<td>Androgen receptor</td>
</tr>
<tr>
<td>ARI</td>
<td>Adjusted Rand Index</td>
</tr>
<tr>
<td>BL1</td>
<td>Basal-like 1</td>
</tr>
<tr>
<td>BL2</td>
<td>Basal-like 2</td>
</tr>
<tr>
<td>BLBC</td>
<td>Basal-like breast cancer</td>
</tr>
<tr>
<td>BLIA</td>
<td>Basal-like immune-activated</td>
</tr>
<tr>
<td>BLIS</td>
<td>Basal-like immune-suppressed</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>Chromatin immunoprecipitation on chip</td>
</tr>
<tr>
<td>CIBEX</td>
<td>Center for information biology gene expression database</td>
</tr>
<tr>
<td>CIBM</td>
<td>Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine</td>
</tr>
<tr>
<td>CGH</td>
<td>Comparative genomic hybridization</td>
</tr>
<tr>
<td>CNA</td>
<td>Copy number aberration</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy number variation</td>
</tr>
<tr>
<td>CTD</td>
<td>Comparative Toxicogenomic Database</td>
</tr>
<tr>
<td>DamID</td>
<td>DNA adenine methyltransferase identification</td>
</tr>
<tr>
<td>DAVID</td>
<td>Database for Annotation, Visualization and Integrated Discovery</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>EBI</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>EGA</td>
<td>European Genome-Phenome Archive</td>
</tr>
<tr>
<td>EpCAM</td>
<td>Epithelial cell adhesion molecule</td>
</tr>
<tr>
<td>ER</td>
<td>Oestrogen receptor</td>
</tr>
<tr>
<td>FGED</td>
<td>Functional Genomics Data Society</td>
</tr>
<tr>
<td>FOIPPA</td>
<td>Freedom of Information and Protection of Privacy Act</td>
</tr>
<tr>
<td>FS</td>
<td>Feature Selection</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>HER2</td>
<td>Human epidermal growth factor receptor 2</td>
</tr>
<tr>
<td>HREC</td>
<td>Human Research Ethics Committee</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>HTC</td>
<td>High content screening</td>
</tr>
<tr>
<td>HTS</td>
<td>High-throughput screening</td>
</tr>
<tr>
<td>ICGC</td>
<td>International Cancer Genomics Consortium</td>
</tr>
<tr>
<td>IDC</td>
<td>Invasive ductal carcinoma</td>
</tr>
<tr>
<td>IHC</td>
<td>Immunohistochemical</td>
</tr>
<tr>
<td>IHGSC</td>
<td>International Human Genome Sequencing Consortium</td>
</tr>
<tr>
<td>ILC</td>
<td>Invasive lobular carcinoma</td>
</tr>
<tr>
<td>IM</td>
<td>Immunomodulatory</td>
</tr>
<tr>
<td>JS</td>
<td>Jensen Shannon</td>
</tr>
<tr>
<td>Ki-67</td>
<td>Antigen identified by monoclonal antibody Ki-67</td>
</tr>
<tr>
<td>kNN</td>
<td>$k$ nearest neighbours</td>
</tr>
<tr>
<td>LAR</td>
<td>Luminal androgen receptor</td>
</tr>
<tr>
<td>lincRNA</td>
<td>long intergenic non-coding RNA</td>
</tr>
<tr>
<td>MA</td>
<td>Memetic algorithm</td>
</tr>
<tr>
<td>MCC</td>
<td>Matthews’ Correlation Coefficient</td>
</tr>
<tr>
<td>MDL</td>
<td>Minimum Description Length Principle</td>
</tr>
<tr>
<td>METABRIC</td>
<td>Molecular Taxonomy of Breast Cancer International Consortium</td>
</tr>
<tr>
<td>MIAME</td>
<td>Minimum Information About a Microarray Experiment</td>
</tr>
<tr>
<td>microRNA</td>
<td>miRNA</td>
</tr>
<tr>
<td>MGED</td>
<td>Microarray Gene Expression Data Society</td>
</tr>
<tr>
<td>MS</td>
<td>Menopausal status</td>
</tr>
<tr>
<td>MST</td>
<td>Minimum Spanning Tree</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NOS</td>
<td>Not otherwise specified</td>
</tr>
<tr>
<td>NPI</td>
<td>Nottingham prognostic score</td>
</tr>
<tr>
<td>NSC</td>
<td>Nearest Shrunken Centroids</td>
</tr>
<tr>
<td>NST</td>
<td>No special type</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PIPA</td>
<td>Personal Information Protection Act</td>
</tr>
<tr>
<td>PIPEDA</td>
<td>Personal Information Protection and Electronic Documents Act</td>
</tr>
<tr>
<td>PR</td>
<td>Progesterone receptor</td>
</tr>
<tr>
<td>PRC</td>
<td>Priority Research Centres</td>
</tr>
<tr>
<td>RHD</td>
<td>Research Higher Degree</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>ROCK</td>
<td>Research Online Cancer Knowledgebase</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse-transcriptase Polymerase chain reaction</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>SAM</td>
<td>Sentrix® Array Matrix</td>
</tr>
<tr>
<td>SCM</td>
<td>Subtype Classification Model</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SSP</td>
<td>Single Sample Predictor</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TEND</td>
<td>Trends in the Exploration of Novel Drug targets</td>
</tr>
<tr>
<td>TNBC</td>
<td>Triple-negative breast cancer</td>
</tr>
<tr>
<td>TNM</td>
<td>Tumour size, nodes, metastasis</td>
</tr>
<tr>
<td>TTD</td>
<td>Therapeutic Target Database</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California Santa Cruz</td>
</tr>
<tr>
<td>WEKA</td>
<td>Waikato Environment for Knowledge Analysis</td>
</tr>
</tbody>
</table>
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)**


**HCRA TRAVEL GRANT (AUD $1,000)**


**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


**Abstracts Published**


Oral Presentations


Poster Sessions


**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