Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

8-19-2022 8:00 AM

The molecular landscape of early-stage breast cancer with lymph node metastasis

Farhad Ghasemi, *The University of Western Ontario*

Supervisor: Brackstone, M, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Surgery

© Farhad Ghasemi 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

 Part of the Genetic Processes Commons, Medical Genetics Commons, Oncology Commons, and the Surgery Commons

Recommended Citation

Ghasemi, Farhad, "The molecular landscape of early-stage breast cancer with lymph node metastasis" (2022). *Electronic Thesis and Dissertation Repository*. 8729.
https://ir.lib.uwo.ca/etd/8729

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

ABSTRACT

Axillary lymph nodes (ALNs) are the primary site of metastasis in breast cancer, and their involvement has implications in disease staging, prognostication, and treatment decisions. A non-invasive modality of assessing the risk of ALN metastasis can improve care in patients with early-stage breast cancer by omitting the morbidity and costs associated with axillary surgery.

This thesis explores the molecular landscape of early-stage breast cancers with ALN metastasis and shows the potential of tumour molecular signatures in predicting ALN involvement. After a systematic review of the literature, we use data from The Cancer Genome Atlas (TCGA) to develop molecular signatures associated with ALN metastasis. We then use machine-learning to develop predictive models. We show that the predictive performance of models may be improved by accounting for the intrinsic molecular subtype of breast cancer. If validated externally, these models have the potential to reduce the rates of axillary surgery in patients with early-stage breast cancer.

KEYWORDS

Breast cancer, lymph node, metastasis, predictive model, nomogram, early stage, molecular signature, differentially expressed genes, messenger RNA, micro-RNA, bioinformatics

SUMMARY FOR LAY AUDIENCE

The lymph nodes underneath the armpit are the most common site of spread in breast cancer. In each patient, it is important to determine if these lymph nodes contain cancer as this information helps clinicians assign a stage to the cancer and suggest appropriate treatments. Clinical examination is not enough to rule out the presence of cancer in these lymph nodes. Most patients require surgery to remove several representative lymph nodes from the armpit area, so that these lymph nodes can be examined by a pathologist underneath a microscope for the presence of breast cancer. There is an opportunity to improve care, as surgery has risks for patients and costs for healthcare system. A solution to this problem could be a computer-generated predictive model that uses the genetic information from the cancer biopsy sample and provides an estimation for risk of cancer spread to lymph nodes for each patient.

We first searched the literature for available evidence on the topic of lymph node spread prediction in early-stage breast cancers. We included 59 articles and discussed the various patient and tumour factors studied in connection to the lymph node spread of breast cancer. We then used the publicly available genetic databases from The Cancer Genome Atlas (TCGA) collaborative to find the differences in the genetic information of early-stage breast tumours with lymph node spread, compared to those without. Our study also highlights that the genetic differences seen in cancers with lymph node spread are not consistent between the four previously established subgroups of breast cancer, known as the “intrinsic molecular subtypes”, and emphasizes the heterogeneity in the genetic information of breast cancers.

Based on the discovered molecular differences, we use computer-generated predictive models of lymph node spread in early-stage breast cancer. We show that the accuracy of these

predictive models can be improved using a new approach that takes into account the intrinsic molecular subtype of the cancer. If validated in other populations, these models can be useful in reducing the rates of lymph node surgery and improve care in early-stage breast cancer.

THE CO-AUTHORSHIP

The co-authors listed below made notable contributions to this thesis.

Muriel Brackstone, MD, MSc, PhD, FRCSC as my supervisor provided me with direction and guidance on study inception, design, data analysis and manuscript writing.

Nadeesha Samarasinghe, MD assisted me with the systematic review portion of the manuscript and was involved with data extraction and interpretation for this segment.

Joseph S Mymryk, Ph.D. has provided critical revisions and edits on the writing and provided guidance with data interpretation.

ACKNOWLEDGEMENTS

I am grateful for my supervisor's support, Dr. Muriel Brackstone, who encouraged me to pursue the Master of Science in Surgery. I am a more competent researcher because of her mentorship.

I am thankful to the Division of General Surgery and in particular my program directors, Dr. Michael Ott and Julie Ann Van Koughnett, for their continuous support of academia, and allowing me the time to develop my research skills in conjunction with my clinical training.

Lastly, I have the utmost appreciation of Prof. Joseph Mymryk and Dr. Anthony Nichols who have both showed continuous support and encouragement through several stages of my academic journey so far.

DEDICATION

My family, Taylor and Otis, provide my life with balance and joy. I am indebted for their support and patience through the process of writing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY TABLES

Supplementary Table 3.9. Top 50 differentially expressed genes between node-positive and node-negative in Luminal B subtype of early-stage breast cancers.

Supplementary Table 3.10. Top 50 differentially expressed genes between node-positive and node-negative in Basal subtype of early-stage breast cancers.

Supplementary Table 3.11. Top 50 differentially expressed genes between node-positive and node-negative in Her2 subtype of early-stage breast cancers.

Supplementary Table 3.12. Gene Ontology pathway enrichment analysis using the DEGs identified comparing node-positive vs. negative samples in all early-stage breast cancer molecular subtypes.

Supplementary Table 3.13. Reactome pathway enrichment analysis using the DEGs identified comparing node-positive vs. negative samples in all early-stage breast cancer molecular subtypes.

Supplementary Table 3.14: Gene Ontology pathway enrichment analysis using the DEGs identified comparing node-positive vs. negative samples in early-stage breast cancer Luminal A and Luminal B molecular subtypes separately. No pathways were identified in basal and Her2 subtypes.

Supplementary Table 3.15: Reactome enrichment analysis using the DEGs identified comparing node-positive vs. negative samples in early-stage breast cancer Luminal A and Luminal B molecular subtypes separately. No pathways were identified in basal and Her2 subtypes.

Supplementary Table 3.16. Top 10 differentially expressed proteins between node-positive and negative in early-stage breast cancer.

Supplementary Table 3.17. Top 10 differentially expressed proteins between node-positive and negative early-stage breast cancers, in each molecular subtype.

Supplementary Table 4.1. Clinical characteristics of training and validation cohorts

Supplementary Table 4.2: Model performance measures with the training cohort with molecular signature inputs varying in length including 50, 75 and 100 genes.

Supplementary Table 4.3. Input variables, training cohort size and final variables included after feature selection in predictive models of axillary lymph node status in early-stage breast cancers.

Supplementary Table 4.4 – Performance of predictive models in the validation cohort with subtypes combined and analyzed individually.

Supplementary Table 4.5 – Variable coefficients of predictive models.

Supplementary Table 4.6. Performance of predictive model 4 (based on uniform molecular signature and clinical characteristics) in the basal subtype of validation dataset (n=46).

LIST OF ABBREVIATIONS

ALN, Axillary Lymph Node

ALND, Axillary Lymph Node Dissection

AUC, Area under the curve

BMI, Body Mass Index

CI, Confidence Interval

CNA, Copy number alteration

DCE, Dynamic contrast-enhanced

DFS, Disease-Free Survival

DWI, Diffusion-weighted images

ER, Estrogen Receptor

FDG-PET, Fluorodeoxyglucose-positron emission tomography

FDR, False discovery rate

FNR, False negative rate

Her2, Human epidermal growth factor receptor 2

HR, Hormone Receptor

IHC, Immunohistochemistry

miRNA, microRNA

MRI, Magnetic resonance imaging

mRNA, messenger RNA

MSKCC, Memorial Sloan Kettering Cancer Center

NSABP, National Surgical Adjuvant Breast and Bowel Project

OR, Odds ratio

OS, Overall Survival

PR, Progesterone Receptor

RPPA, Reverse Phase Protein Array

ROC, Receiver-operator characteristics

SLN, Sentinel Lymph Node

SLNB, Sentinel Lymph Node Biopsy

SNV, Single nucleotide variation

TCGA, The Cancer Genome Atlas

US, Ultrasound

CHAPTER 1

INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1 BREAST CANCER: EPIDEMIOLOGY AND PROGNOSIS

Breast cancer has become the most commonly diagnosed cancer worldwide with over 2.2 million new cases per year(1). Globally, higher incidence rates are seen in high-income regions such as North America, Northern/Western Europe and Australia/New Zealand compared to Asia and sub-Saharan Africa(2). This trend is attributed to risk factors associated with urbanization and economic development including obesity, higher fat intake and physical inactivity(3,4).

Breast cancer remains as the leading cause of cancer death in women worldwide(5). While advancement in screening and systemic therapies have improved survival from breast cancer in developed countries, the rates of mortality has been increasing along with the incidence of the disease in developing countries(6). Mortality and recurrence risk depends on disease stage(7). The 8th-edition of the American Joint Committee on Cancer's Staging System for breast cancer includes two staging systems(8). Firstly, the clinical stage is determined based on the pre-operative tumour size, nodal status, and presence of distant metastasis. Secondly, the pathological stage includes the results of post-operative pathology findings including the derived tumour size, nodal involvement, tumour grade, hormone and oncogene expression profiles and the results of multi-gene panel testing, and this stage is more accurate in predicting individualized outcomes(9).

1.2 AXILLARY LYMPH NODES: SIGNIFICANCE AND ROLE OF SURGERY

Axillary surgery in breast cancer has been evolving for centuries(10). In the 19th century, a German pathologist, Rudolf Virchow, noted the presence of ipsilateral axillary lymph node (ALN) involvement in the autopsy of women who died of metastatic breast cancer(11). Virchow

suspected that ALNs were a nidus for distant metastatic disease. In line with Virchow's hypothesis, William Halsted, an American Surgeon in the 19th century, advocated for complete removal of the ALNs to improve outcomes in all breast cancer patients(12). The radical mastectomy, which included the removal of breast, pectoralis muscle and ipsilateral lymph nodes became the standard of care for decades to come.

The concept that ALN metastasis was simply an indicator of tumour chronology was questioned by several observations(13). First was the emergence of distant metastasis in patients without axillary involvement after radical mastectomy, which contradicted the idea that the ALNs served as a nidus for all distant metastases. Second were several studies including the National Surgical Adjuvant Breast and Bowel Project (NSABP)-04 which systematically showed no overall survival advantage to early ALN clearance in patients without clinically palpable disease(14,15). This suggested that the presence of ALN metastasis was perhaps a marker of aggressive tumour biology and not just chronology(16). In line with this philosophy, was the finding that even in patients with breast cancer recurrence, lymph node involvement in the primary cancer presentation predicted unfavourable outcomes(17). Although the mechanisms behind lymph node and distant metastasis continue to be studied, ALN involvement remains as one of the most important prognostic factors in breast cancer(7,18).

Clinical examination is inadequate in determining ALN involvement(19). In the NSABP B-04 trial, up to 40% of patients with clinically negative axilla who were randomized to receive an axillary lymph node dissection (ALND) had evidence of lymph node metastasis on final pathology(14). ALND however is accompanied by a high risk of post-operative morbidity for the

patients. Studies have shown that over 70% of patients may experience symptoms such as arm and shoulder pain/weakness, numbness, tingling and lymphedema (20,21).

The need for a less morbid modality of axillary staging led to the introduction of sentinel lymph node biopsy (SLNB) as an alternative to ALND. The sentinel lymph nodes (SLN) are the first nodes in the chain that drain lymph from an organ. SLNB was introduced in breast cancer following reassuring results in parotid, melanoma, and penile cancers(22). Guiliano *et al.* published on a series of SLNBs in 1994 and showed SLNB to accurately predict axillary nodal status in breast cancer(23).

The NSABP B-32 trial compared overall survival (OS) and disease-free survival (DFS) between 5,611 clinically node-negative breast cancer patients receiving SLNB+ALND or SLNB alone with ALND only if sentinel lymph nodes were positive(24). After 10-years, no differences in OS or DFS were reported between the two groups (25). This suggested that SLNB was a viable alternative to staging the axilla in patients with clinically negative axilla.

ALN involvement can be as infrequent as 20% in early-stage breast cancer patients without palpable lymphadenopathy(26). Yet, due to the essential role of an accurate nodal stage in clinical decision-making and prognosis, major guidelines continue to recommend SLNB in this patient group(27,28). The establishment of SLNB as standard of care reduced post-operative morbidity and improved quality of life metrics for patients compared to ALND(21,29). Still, significant rates of residual morbidity were reported in up to 1 in 6 patients after SLNB in the NSABP B-32 trial(30). These included residual shoulder abduction deficit in 13.2% of patients at 6 months, arm volume differences in 16.7%, arm numbness in 7.5% and tingling in 6.7% of patients at 36 months follow up.

Identification of patients at low risk of axillary involvement therefore presents the potential for omitting axillary surgery in this group and sparing these patients from the associated risks. SLNB omission has public-health benefits arising from reduced operative time(31). This highlights a clinical gap in our ability to stratify early-stage breast cancer patients based on their risk for ALN involvement.

Accurate recognition of patients at high-risk of axillary metastasis would also contribute to decisions regarding pre-operative systemic therapy. Neoadjuvant chemotherapy has been shown to be effective at downstaging the axilla in patients with biopsy-proven axillary metastases(32). In a retrospective study of 630 biospy-proven node-positive breast cancer patients from Memorial Sloan Kettering who received neoadjuvant chemotherapy, 91% converted to clinically negative axilla and 46% achieved complete pathologic response(33). SLNB can appropriately stage the axilla after neoadjuvant chemotherapy but cannot be done twice (before and after neoadjuvant chemotherapy) due the low detection rate (60.8%, 95% CI 55.6-65.9) and high false-negative rate (51.6%, 95% CI 38.7-64.2) as shown in the SENTINA trial(34). Hence, non-invasive classification of early-stage breast cancer patients with high risk of axillary metastases, but no palpable lymphadenopathy, can enable clinicians to identify candidates for neoadjuvant chemotherapy administration, potentially sparing them of further ALND or irradiation. Gene expression-based assays such as the Oncotype DX™(35) have been established within care pathways in informing treatment decisions for adjuvant chemotherapy(36). Similar molecular-based risk assessment tools may help select optimal patients for systemic treatment in the pre-operative setting.

1.3 CLINICAL AND MOLECULAR SUBTYPES OF BREAST CANCER

The discovery of the importance of hormone receptors (HRs) in breast cancer and the development of antibodies against estrogen receptor (ER) allowed immune histochemistry (IHC)-based classification of breast cancer based on HRs(37–39). Identification of the human epidermal growth factor receptor 2 (Her2) gene activation in the 1980s and its significance in predicting poor prognosis in affected patients led to further categorization of breast cancer based on this oncogene(40). The more aggressive Her2-positive breast cancers became a target for the therapeutic anti-Her2 molecule monoclonal antibody, trastuzumab, in 1998(40). These discoveries led to the establishment of clinical subtypes based the expression levels of hormone receptors such as ER, progesterone receptor (PR) and Her2 on immune histochemistry (IHC). These classifications allowed for subtype-based approaches to treatment decisions such as hormone therapy for ER/PR-positive patients and Her2-directed treatments in those with Her2-positive disease (36).

Since then, molecular techniques such as RNA-seq and microarray gene expression analyses have advanced our knowledge of heterogeneity within breast cancer. In 2000, Perou and Sorlie classified breast cancer into 4 distinct molecular subtypes including luminal A, luminal B, basal-like and Her2-enriched subtypes (**Table 1.1**)(41). This classification was based on a 50-gene expression signature (known as the PAM50 signature). A “normal-like” subtype was also proposed but the presence of this subtype is questioned from attribution of results to artifact from normal breast tissue(42). The 4 intrinsic subtypes have shown value in disease prognostication(43). Luminal tumours are associated with IHC-based HR-positive tumours, and the Ki-67 level on IHC was used to distinguish between Luminal A and Luminal B tumours on pathology(37,44). HER2-enriched subtype mapped to HER2-positive, ER/PR-negative disease

and Basal-like to triple-negative cancers. Notably, the clinical IHC-based subtypes and intrinsic molecular subtypes do not completely overlap (**Table 1.2**)(45), and several projects have attempted to increase concordance between these two sets of classifications(46,47). Due to the better accuracy and reproducibility of gene expression-based subtypes, the St. Gallen international expert consensus panel have advocated for the use these molecular subtypes in developing therapy concepts for early-stage breast cancer(48). The intrinsic molecular subtypes have utility beyond the pathology-based classifications and have shown value in improving predictions regarding response to neoadjuvant systemic therapy and prognosis(45).

Growing evidence suggests that the different molecular subtypes of breast cancer have distinct metastatic behaviour as well(49). While luminal subtypes of breast cancer metastasize to ALNs more frequently, systemic spread have been associated with the HER2 and basal subtypes(50). As such, it is imperative to explore the performance of any predictive model in breast cancer in the different disease subtypes. Understanding the molecular differences that are seen in the presence or absence of nodal metastasis in each subtype independently can potentially reduce the confounding effects of inter-subtype heterogeneity which may ultimately improve model performance.

1.4 THE APPLICATION OF MACHINE LEARNING IN PREDICTION OF CANCER PROGNOSIS

Advancements in our understanding of tumour biology has fueled a push towards cancer care that is tailored to each individual patient(51). Improved technologies in genomics, transcriptomics, proteomics and epigenomics can generate complex sets of data on individual tumours, but achieving clinical utility with this data requires advanced statistical techniques that can facilitate the interpretation process. Machine learning algorithms which adopt a

myriad of statistical and optimization techniques have been utilized to interpret the growing body of cancer data to generate clinically relevant information such as risk of cancer recurrence and survival(52). Reviews of available machine learning models in cancer suggest a 15-20% improvement in the ability of these algorithms to predict cancer susceptibility, recurrence and mortality compared to traditional statistical methods such as logistic regression analyses(53).

1.5 OBJECTIVES AND OVERVIEW OF CHAPTERS

The objective of this thesis is to develop a predictive model of axillary metastasis based on molecular data that is specific to patients with early-stage breast cancers (size ≤5 cm and no clinical lymphadenopathy). My approach will be to explore the available evidence, analyze the molecular signatures associated with nodal metastasis, and finally, to develop predictive algorithms.

To date, several models have been developed aimed at predicting the presence of nodal metastasis without invasive axillary procedures. In the following chapter of this thesis, I will discuss the available evidence on this topic. My literature search will explore various clinical, pathological, radiological, and molecular factors associated with axillary status, and the performance of the developed predictive models.

The third chapter will focus on a molecular analysis of differences contributing to axillary involvement in early-stage breast cancer using genomic data. As tumour size is the primary factor associated with lymph node involvement, we will focus our analysis on early-stage tumours (patients with size ≤ 5cm) to reduce the impact of large tumour size as a confounder. Our analysis will account for the previously established intrinsic molecular subtypes.

In the fourth chapter we will utilize the identified molecular differences based on nodal involvement to construct predictive models of axillary metastasis in early-stage breast cancer. Considering the various molecular subtypes as separate diseases, I will develop predictive models for each subtype independently to assess if this approach improves performance. In the final chapter I will summarize my findings, integrate the individual conclusions, and discuss their impact in the context of the current literature. I will also describe potential ways that my work could be extended in the future.

1.5 TABLES

Table 1.1 The four intrinsic molecular subtypes of breast cancer, and surrogate clinical and pathological markers. Table from Szymiczek *et al.* 2021(37).

Intrinsic subtype	**Surrogate IHC definition**				**Prevalence [%]**	**Prognosis**
	ER	PR	HER2	Ki-67 Level		
Luminal A	+	+	−	Low (<14%)	30–70	Good
Luminal B	+	+ or −	+ or −	High (≥14%)	10–20	Intermediate
HER2-enriched	−	−	+	Any	5–15	Poor but improved with anti-HER2 treatment
Basal-like	−	−	−	Any	15–20	Poor

Table 1.2 The consensus between PAM50 intrinsic molecular subtypes and IHC-based clinical subtypes. Table from Prat *et al.* 2015(45).

IHC-based group	**N**	**PAM50 intrinsic subtype distribution**			
		Luminal A	Luminal B	HER2-enriched	Basal-like
HR+/HER2−	4295	60.3%	31.9%	6.6%	1.2%
Luminal A	637	62.2%	27.0%	10.2%	0.6%
Luminal B	317	34.1%	51.1%	11.0%	3.8%
HER2+	831	17.6%	26.8%	44.6%	11.0%
HER2+/HR+	182	33.0%	46.2%	18.7%	2.2%
HER2+/HR−	168	19.0%	4.2%	66.1%	10.7%
TNBC	868	1.6%	3.2%	9.1%	86.1%

1.6 BIBLIOGRAPHY

1. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. Int J Cancer. 2021;149(4):778–89.

2. DeSantis CE, Bray F, Ferlay J, Lortet-Tieulent J, Anderson BO, Jemal A. International Variation in Female Breast Cancer Incidence and Mortality Rates. Cancer Epidemiology Prev Biomarkers. 2015;24(10):1495–506.

3. Linos E, Spanos D, Rosner BA, Linos K, Hesketh T, Qu JD, et al. Effects of Reproductive and Demographic Changes on Breast Cancer Incidence in China: A Modeling Analysis. Jnci J National Cancer Inst. 2008;100(19):1352–60.

4. Boyle P, Howell A. The globalisation of breast cancer. Breast Cancer Res Bcr. 2010;12(Suppl 4):S7–S7.

5. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. Ca Cancer J Clin. 2021;71(3):209–49.

6. Youlden DR, Cramb SM, Dunn NAM, Muller JM, Pyke CM, Baade PD. The descriptive epidemiology of female breast cancer: An international comparison of screening, incidence, survival and mortality. Cancer Epidemiol. 2012;36(3):237–48.

7. Abdel-Rahman O. Validation of the 8th AJCC prognostic staging system for breast cancer in a population-based setting. Breast Cancer Res Tr. 2018;168(1):269–75.

8. N H Gabriel, JL C, CJ D, SB E, EA M, HS R, et al. AJCC cancer staging manual. 2017:589–636.

9. Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast Cancer—Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. Ca Cancer J Clin. 2017;67(4):290–303.

10. Ekmektzoglou KA, Xanthos T, German V, Zografos GC. Breast cancer: From the earliest times through to the end of the 20th century. Eur J Obstet Gyn R B. 2009;145(1):3–8.

11. Jatoi I. The natural history of breast cancer. Surg Clin N Am. 1999;79(5):949–60.

12. Halsted WS. The results of operations for the cure of cancer of the breast performed at the Johns Hopkins Hospital from June, 1889, to January, 1894. Ann Surg. 1894;20(NA;):497–555.

13. Fisher B. The surgical dilemma in the primary therapy of invasive breast cancer: A critical appraisal. Curr Prob Surg. 1970;7(10):3–53.

14. Fisher B, Jeong JH, Anderson S, Bryant J, Fisher ER, Wolmark N. Twenty-Five-Year Follow-up of a Randomized Trial Comparing Radical Mastectomy, Total Mastectomy, and Total Mastectomy Followed by Irradiation. New Engl J Medicine. 2002;347(8):567–75.

15. Cancer Research Campaign Working Party. Cancer research campaign (King's/Cambridge) trial for early breast cancer. A detailed update at the tenth year. Lancet. 1980;316(8185):55–60.

16. Hellman S. Karnofsky Memorial Lecture. Natural history of small breast cancers. J Clin Oncol. 1994;12(10):2229–34.

17. Rack B, Janni W, Gerber B, Strobl B, Schindlbeck C, Klanner E, et al. Patients with Recurrent Breast Cancer: Does the Primary Axillary Lymph node Status Predict more Aggressive Tumor Progression? Breast Cancer Res Tr. 2003;82(2):83–92.

18. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. Ca Cancer J Clin. 2017;67(2):93–9.

19. Weiss A, Chavez-MacGregor M, Lichtensztajn DY, Yi M, Tadros A, Hortobagyi GN, et al. Validation Study of the American Joint Committee on Cancer Eighth Edition Prognostic Stage Compared With the Anatomic Stage in Breast Cancer. Jama Oncol. 2017;4(2):203.

20. Hack TF, Cohen L, Katz J, Robson LS, Goss P. Physical and Psychological Morbidity After Axillary Lymph Node Dissection for Breast Cancer. J Clin Oncol. 1999;17(1):143–143.

21. Peintinger F, Reitsamer R, Stranzl H, Ralph G. Comparison of quality of life and arm complaints after axillary lymph node dissection vs sentinel lymph node biopsy in breast cancer patients. Brit J Cancer. 2003;89(4):648–52.

22. D'Angelo-Donovan DD, Dickson-Witmer D, Petrelli NJ. Sentinel lymph node biopsy in breast cancer: A history and current clinical recommendations. Surg Oncol. 2012;21(3):196–200.

23. Giuliano AE, Kirgan DM, Guenther JM, Morton DL. Lymphatic Mapping and Sentinel Lymphadenectomy for Breast Cancer. Ann Surg. 1994;220(3):391–401.

24. Krag DN, Julian TB, Harlow SP, Weaver DL, Ashikaga T, Bryant J, et al. NSABP-32: Phase III, randomized trial comparing axillary resection with sentinal lymph node dissection: A description of the trial. Ann Surg Oncol. 2004;11(Suppl 3):208S-210S.

25. Julian TB, Anderson SJ, Krag DN, Harlow SP, Costantino JP, Ashikaga T, et al. 10-yr follow-up results of NSABP B-32, a randomized phase III clinical trial to compare sentinel node resection (SNR) to conventional axillary dissection (AD) in clinically node-negative breast cancer patients. J Clin Oncol. 2013;31(15_suppl):1000–1000.

26. Cabanes PA, Salmon RJ, Vilcoq JR, Durand JC, Fourquet A, Gautier C, et al. Value of axillary dissection in addition to lumpectomy and radiotherapy in early breast cancer. Lancet. 1992;339(8804):1245–8.

27. W. C Robert, Craig A D, O. A Benjamin, J. B Harold, Bradford C W, B. E Stephen, et al. Invasive Breast Cancer. J Natl Compr Canc Ne [Internet]. 2011;9(2):136–222. Available from: https://jnccn.org/view/journals/jnccn/9/2/article-p136.xml

28. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30(8):1194–220.

29. Mansel RE, Fallowfield L, Kissin M, Goyal A, Newcombe RG, Dixon JM, et al. Randomized Multicenter Trial of Sentinel Node Biopsy Versus Standard Axillary Treatment in Operable Breast Cancer: The ALMANAC Trial. Jnci J National Cancer Inst. 2006;98(9):599–609.

30. Ashikaga T, Krag DN, Land SR, Julian TB, Anderson SJ, Brown AM, et al. Morbidity results from the NSABP B-32 trial comparing sentinel lymph node dissection versus axillary dissection. J Surg Oncol. 2010;102(2):111–8.

31. Perrier L, Nessah K, Morelle M, Mignotte H, Carrère MO, Brémond A. Cost comparison of two surgical strategies in the treatment of breast cancer: Sentinel lymph node biopsy versus axillary lymph node dissection. Int J Technol Assess. 2004;20(4):449–54.

32. Bear HD, Anderson S, Brown A, Smith R, Mamounas EP, Fisher B, et al. The Effect on Tumor Response of Adding Sequential Preoperative Docetaxel to Preoperative Doxorubicin and Cyclophosphamide: Preliminary Results From National Surgical Adjuvant Breast and Bowel Project Protocol B-27. J Clin Oncol. 2003;21(22):4165–74.

33. Montagna G, Mamtani A, Knezevic A, Brogi E, Barrio AV, Morrow M. Selecting Node-Positive Patients for Axillary Downstaging with Neoadjuvant Chemotherapy. Ann Surg Oncol. 2020;27(11):4515–22.

34. Kuehn T, Bauerfeind I, Fehm T, Fleige B, Hausschild M, Helms G, et al. Sentinel-lymph-node biopsy in patients with breast cancer before and after neoadjuvant chemotherapy (SENTINA): a prospective, multicentre cohort study. Lancet Oncol. 2013;14(7):609–18.

35. Syed YY. Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer. Mol Diagn Ther. 2020;24(5):621–32.

36. National Comprehensive Cancer Network. Breast Cancer (version 2.2022) [Internet]. [cited 2022 Feb 19]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

37. Szymiczek A, Lone A, Akbari MR. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. Clin Genet. 2021;99(5):613–37.

38. Osborne CK, Yochmowitz MG, Knight WA, McGuire WL. The value of estrogen and progesterone receptors in the treatment of breast cancer. Cancer. 1980;46(S12):2884–8.

39. Greene GL. In Memoriam: Elwood Jensen (1920–2012). Endocr Rev. 2013;34(6):761–3.

40. Ross JS, Fletcher JA, Linette GP, Stec J, Clark E, Ayers M, et al. The HER-2/neu Gene and Protein in Breast Cancer 2003: Biomarker and Target of Therapy. Oncol. 2003;8(4):307–25.

41. Perou CM, Sørlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52.

42. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014;5(3):412.

43. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009;27(8):1160–7.

44. Soliman NA, Yussif SM. Ki-67 as a prognostic marker according to breast cancer molecular subtype. Cancer Biology Medicine. 2016;13(4):496–504.

45. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast. 2015;24:S26–35.

46. Raj-Kumar PK, Liu J, Hooke JA, Kovatich AJ, Kvecher L, Shriver CD, et al. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. Sci Rep-uk. 2019;9(1):7956.

47. Bastien RR, Rodríguez-Lescure Á, Ebbert MT, Prat A, Munárriz B, Rowe L, et al. PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. Bmc Med Genomics. 2012;5(1):44–44.

48. Harbeck N, Thomssen C, Gnant M. St. Gallen 2013: Brief Preliminary Summary of the Consensus Discussion. Breast Care. 2013;8(2):102–9.

49. Kennecke H, Yerushalmi R, Woods R, Cheang MCU, Voduc D, Speers CH, et al. Metastatic Behavior of Breast Cancer Subtypes. J Clin Oncol. 2010;28(20):3271–7.

50. Buonomo OC, Caredda E, Portarena I, Vanni G, Orlandi A, Bagni C, et al. New insights into the metastatic behavior after breast cancer surgery, according to well-established clinicopathological variables and molecular subtypes. Plos One. 2017;12(9):e0184680.

51. Verma M. Personalized Medicine and Cancer. J Personalized Medicine. 2012;2(1):1–14.

52. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnology J. 2015;13:8–17.

53. Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics. 2006;2:117693510600200030.

CHAPTER 2

PREDICTORS OF AXILLARY INVOLVEMENT IN EARLY-STAGE BREAST CANCER: SYSTEMATIC REVIEW

CHAPTER 2: PREDICTORS OF AXILLARY INVOLVEMENT IN EARLY-STAGE BREAST CANCER: SYSTEMATIC REVIEW

2.1 INTRODUCTION

The treatment of breast cancer and its lymph node basin in the axilla has evolved significantly over the past decades(1). In parallel to a shift towards breast conserving surgery, studies supporting de-intensification of surgery in the axilla have moved practice away from axillary lymph node dissections (ALNDs) in early-stage breast cancers. The National Surgical Adjuvant Breast and Bowel Project (NASBP) B-32 trial revealed that sampling sentinel lymph nodes (SLNs) can provide reliable staging of the axilla in patients without clinically evident lymphadenopathy(2), avoiding the significant morbidity associated with ALNDs(3,4).

Axillary lymph node involvement can be as infrequent as 20% in early-stage breast cancer patients without palpable lymphadenopathy(5). Despite this, available major guidelines recommend SLN biopsy (SLNB) for staging of the axilla in this patient group(6,7). Although less morbid than an ALND, SLNB still exposes patients to risks of complication, including wound infections, seroma formations, hematoma, nerve injury and lymphedema(8,9). In addition, SLNB is a resource intensive operation often requiring pre-operative localization of the lymph nodes by radiology and extended time in the operating room. This identifies a clinical need for predictive models that would distinguish patients with low risk of axillary involvement in whom invasive axillary staging can be omitted.

Recognition of patients at high risk of axillary involvement also has important value in pre-operative treatment decisions. Neoadjuvant chemotherapy has been shown to be effective at downstaging the axilla in patients with biopsy-proven axillary metastases(10). SLNB can

appropriately stage the axilla after neoadjuvant chemotherapy but is inaccurate if done twice due to the high false negative rates found in the SENTINA trial(11). Hence, non-invasive classification of early-stage breast cancer patients with high risk of axillary metastases, but no palpable lymphadenopathy, can enable clinicians to identify candidates for neoadjuvant chemotherapy administration, potentially sparing them of further ALND or irradiation.

In this review we explored the literature to identify the various clinical, histopathological, radiological, and molecular factors that have been associated with axillary lymph node (ALN) involvement in patients with early-stage breast cancer and examine the models developed for predicting cancer metastasis in the ALNs .

2.2 MATERIALS AND METHODS

A comprehensive search of the literature was completed through PubMed and Web of Science to identify publications examining axillary metastases in early-stage breast cancers (search terms provided in **Table 2.1**). Inclusion criteria was set for primary research articles in English, with full text available, published before July 2021, patient population comprised of women with invasive breast carcinoma, no clinically palpable lymphadenopathy (or multivariate analysis conducted with clinical nodal stage), tumour size of ≤ 5cm (or multivariate analysis conducted with tumour size or T stage). Exclusion criteria included studies investigating patients receiving neoadjuvant chemotherapy, patient cohort with only micro-invasive disease or uncommon breast cancer subtypes (such as metaplastic carcinoma), or those employing invasive axillary sampling (such as fine needle aspiration). A manual search of the bibliographies of the selected articles was also conducted. Abstract and full-text review were completed by two authors independently. The Prediction model study Risk of Bias Assessment Tool (PROBAST) was used to assess for risk of bias and explore study quality(12).

2.3 RESULTS

Of the 1809 studies that resulted from the literature search, 59 met the selection criteria (**Figure 2.1**). **Figure 2.2** outlines the various clinical, histopathological, molecular and radiological variables considered in the included articles. **Supplementary Table 2.1** provides a comprehensive description of cohort characteristics and the models derived in each of the studies. Below, we highlight some of these variables and models.

2.3.1 Clinical patient and tumour factors

Age

Younger age has been associated with a statistically significant increased risk of ALN involvement in multiple studies (**Figure 2.2**). Age as a variable was treated differently in various cohorts, either as a continuous variable or as a categorical variable with inconsistent age groupings. Reyal *et al.* showed in their cohort of 1543 patients that age of diagnosis > 60 had an odds ratio (OR) of 0.56 (95% CI [0.4-0.7]) for axillary metastasis in the training model, and similar results were shown with two validation datasets(13). Ding *et al.* found that age < 40 was associated with a hazard ratio (HR) of 2.188 (95% CI [1.198–4.001])(14). Other studies such as Choi *et al.* and Dihge *et al*. treated age as a continuous variable and showed OR of 0.96 (95% CI [0.92–0.99]) and 0.98 (95% CI [0.96-1.00]) per year respectively(15,16). A bimodal effect, with a return of increasing risk of lymph node involvement in the older patients has also been suggested in other studies not specific to early-stage breast cancers(17). More aggressive tumour biology in younger patients, and a potential lack of appropriate immunologic response

in the elderly have been proposed as potential explanations for a biphasic association with lymph node involvement.

Menopause Status

Menopausal status can relate to both age and hormone exposure. Post-menopausal women were found to have a reduced risk of sentinel lymph node metastasis in several studies (in Chen *et al.* OR=0.78, 95% CI [0.66-0.93])(18,19). In 2 other cohorts where age did not correlate with sentinel lymph node involvement, neither did menopausal status(20,21).

Race

The rate of ALN metastasis was found to be highest in black women under 50 years of age in a multivariate model based on age, tumour size and race(22), underlining the potential importance of racial differences in disease presentation.

Diabetes and Obesity

Diabetes may be associated with an elevated risk of breast cancer(23). Minami *et al.* investigated the correlation between impaired glucose tolerance (as defined by hemoglobin A_{1c} >6.0%), and found it to be an independent predictor of axillary metastasis in multivariate model (OR 2.560, 95% CI[1.11–5.88])(20). Increased Body Mass Index (BMI) has not been shown to be a predictor of axillary involvement in several models of axillary involvement (studies outlined in **Figure 2.2**).

Tumour Size

Tumour size is the most widely utilized factor in predictive models. Like age, size of the tumour has been treated either as a continuous variable or a categorical variable (based on T staging). Furthermore, the size can be derived either from the pre-operative clinical information

or from the post-operative pathology specimen. Regardless of the methodology, most studies found an association between larger tumours and increased risk of ALN involvement. Martin *et al.* investigated the significance of clinically examined tumour size in 795 women with tumours ≤ 4cm and clinically negative axilla, and found the clinical T stage to be a statistically significant factor for axillary metastasis in univariate and multivariate models (p=0.0003 and 0.0007 respectively)(24). In a Chinese cohort of 1000 patients undergoing SLNB, tumour size on pathology specimen was treated as a continuous variable, and yielded an OR of 1.409 (per cm, 95% CI [1.203-1.651], p<0.001) in their multivariate model (25). Other studies incorporating the tumour size in their models are outlined in **Figure 2.2**. Smaller breast size(24) and larger tumour-to-breast volume ratio(26) have also been proposed as predictive factors, although these have not been routinely utilized in most models.

Tumour Location

While the laterality of the tumour (left or right breast) has no correlation with axillary status(14,20,27), tumour location within the breast can change the likelihood of axillary metastasis. Zhang *et al.* showed central tumours to be associated with axillary involvement, noting the abundance of lymphatics in this area as a possible explanation(28). In another multivariate model, upper-inner quadrant tumour location was an independent predictor of lower risk of SLN metastasis compared to upper-outer quadrant tumours (OR 0.563, 95%CI [0.397-0.895], p=0.002)(25). In the same study, no significance was noted between central and upper-outer quadrant tumours (p=0.377)(25). Martin *et al.*'s cohort of patients had axillary involvement in 30.2% of outer tumours, 21.3% of central tumours and 19.2% of inner tumours (chi-squared p=0.0041)(24). Minami *et al.* utilized nipple-to-tumour distance measurements to

quantify the tumour location, and found that increased nipple-to-tumour distance to be independently associated with a decreased risk of SLN metastasis in their multivariate model (OR=0.773 per 1 cm increase in distance, 95% CI [0.638-0.937], p=0.009)(20). Other studies did not find tumour location to be an independent variable in their multivariate models(29,30).

Other clinical factors

Notable factors in patient history such as gravidity and family history have not been found to be associated with sentinel lymph node status(20). Studies utilizing more infrequent patient and tumour characteristics, such as palpability of primary tumour, bilateral cancer, mode of cancer detection and mode of SLN detection are illustrated in **Figure 2.2**. Three groups in our included studies did not exclude patients with clinically positive axilla, and instead, integrated clinical nodal stage in their multivariate model. In 2 of these models, clinical nodal stage was independently associated with pathologically positive axilla (31,32). Yu *et al.* incorporated clinical nodal stage with other clinical variables and radiological factors to create their pre-operative predictive model of axillary involvement(33).

2.3.2 Histopathological factors

Lymphovascular Invasion

The presence of lymphovascular invasion (LVI) suggests tumour access to pathways for metastasis, and as such, has been a strong predictor of axillary involvement in many included models (**Figure 2**). In Chen *et al.*'s training cohort of 1000 patients undergoing SLNB, the presence of LVI had an OR of 8.856 (95% CI [6.112-12.833]) for SLN metastasis(25). A multivariate model based on a Canadian cohort of 405 SLN biopsies showed LVI to be an

independent predictor of SLN involvement (OR 10.736, 95% CI [6.065-19.004])(34). In a model based on patients with T1c tumours from Croatia, LVI showed an OR of 3.681 (95% CI [1.393-9.724])(26). In another study utilizing ultrasound findings with clinicopathological variables, LVI had an OR of 6.755 (95% CI [4.248–10.741]) in their final model(35). Vascular invasion specifically reduced the likelihood of node negative disease in other models(36–38). On the other hand, in the model developed in Fujii *et al.*, only lymphatic invasion was an independent variable (HR 8.381; 95%CI 4.023-17.436, $p<0.001$) and the presence of vascular invasion was not statistically significant(39).

Histologic Type

The histologic type of the tumour was investigated as a variable within several studies (**Figure 2.2**), with some showing association with axillary status. Viale *et al.* noted that favourable histology (medullary, cribriform, tubular, mucinous tumours) had lower odds of axillary involvement compared to ductal carcinoma (OR=0.55 95%CI[0.39–0.78], p=0.007) in their multivariate logistic regression model(40). Special tumour types (defined as colloid, medullary or tubular) were independently associated with negative axilla in the Memorial Sloan Kettering Cancer Center (MSKCC) cohort(41). Similarly, tubular carcinomas had less axillary involvement in a multivariate model constructed from the Korean Breast Cancer Registry(42). Infiltrating lobular carcinoma was an independent predictor of lower rates of SLN involvement in a multivariate model of 1506 patients undergoing SLNB in Belgium (OR 0.49, p=0.003)(43). In contrast, histologic type of tumour as categorized into “invasive ductal carcinoma” and “others”, was not associated with SLN status in univariate or multivariate analyses in another study (20).

Tumour Grade

Nuclear grade was not associated with axillary involvement in most included studies (**Figure 2.2**). Histologic grade on the other hand was built into multiple multivariate models and showed statistical significance. In a cohort of 324 patients undergoing SLNB, advanced histologic grade was an independent predictor of axillary status in the final model with an OR of 1.415 (95% CI [1.004-1.996], p=0.048)(44). The model from the cohort in Qiu *et al.* incorporated histologic grade with an OR of 1.696 (1.316-2.186)(30).

Multifocality

Tumour multifocality may signify higher tumour burden than suggested by the disease T stage, which only considers the diameter of the largest invasive focus. In an Italian study with 4351 breast cancer patients undergoing SLNB, the presence of multifocality had an OR of 1.78 (95% CI [1.41–2.24]) in the multivariate model of SLN involvement (40). Similarly, Qiu *et al.* showed multifocality to be an independent predictor of SLN positivity in their cohort of 1227 patients with an OR of 6.578 (95% CI [1.787-24.219])(30). In a Swedish multivariate model based on 692 patients, unifocal disease was a favourable independent predictor for node negative disease (OR=1.72, 95% CI [1.11-2.65])(16). Multifocality is also one of the independent variables within the Bevilacqua *et al.'s* MSKCC nomogram(41). Other studies utilizing tumour multifocality in their analyses are outlined in **Figure 2.2**.

Other Histopathological Factors

Across the various studies, many other histopathological variables were inconsistently explored, including tumour margin characteristics, characteristics/extent of ductal carcinoma in-situ (DCIS) or lobular carcinoma in-situ (LCIS) within the tumour specimen, neuroinvasion,

microvascular density (MVD), lymphovascular density (LVD), host immune reaction/tumour infiltrating lymphocytes (TILs), presence of extensive intraductal carcinoma (EIC), breast glandular content percentage and calcifications (**Figure 2.2**).

2.3.3 Molecular factors

Hormone Receptors (HRs)

The predictive value of HR positivity for lymph node involvement is controversial (**Figure 2.2**). In the MSKCC cohort, estrogen receptor (ER) and progesterone receptor (PR) positive status were both included in the final nomogram as independent predictors of SLN metastasis(41). The multivariate model based on the Korean Breast Cancer Registry also showed positive ER and PR status to be statistically notable variables with OR of 1.37 (95% CI [1.24-1.50]) and 1.16 (95% CI [1.06-1.26]) for axillary involvement, respectively(42). Qiu *et al.* also found ER and PR to be independent predictors in their model (ER OR 1.698, 95% CI [1.22-2.335] and PR OR 1.517 95% CI [1.110-2.074])(30). Viale *et al.* found only PR status to be statistically significant in their multivariate model (PR negative had OR=0.73 95% CI [0.59-0.90] for SLN involvement)(40). In another cohort, the combined hormone receptor status (ER+ or PR+) was only significant in univariate analysis(45). In comparison, several other studies did not find ER or PR status to be associated with axillary involvement in their analyses(14,15,19,20,39,45–49).

Her2 Receptor Status

Most included studies did not find the tumour Her2/neu status to be independently significant as a variable in their statistical models of SLN metastasis (**Figure 2.2**). The subtype of

the tumour, as determined by the status of ER, PR and Her2, has been utilized in several nomograms. Triple-negative tumour type (defined as negative for ER, PR and Her2) was a favourable factor in predicting node negative disease in the Dihge *et al.* cohort (OR 5.06, 95% CI [1.89-13.50])(16). Similarly, Mao *et al.*'s predictive model of SLN metastasis included triple-negative subtype status as an associated variable with axillary metastasis, showing an OR of 0.506 (95% CI [0.307–0.835])(44). Marrazzo *et al.* also found triple-negative status to be an independent negative predictor of SLN metastasis(38). Zhang *et al.* found tumors with the luminal subtype to exhibit an increased odds of axillary involvement in their multivariate model compared to triple-negative tumours (OR 1.380, 95% CI [1.059-1.799]), but Her2-enriched vs. triple-negative subtype was not statistically significant (OR 1.152, 95% CI [0.764-1.737])(28).

Markers of Cell Proliferation

Several studies have investigated variables linked with tumour cell proliferation, such as S-phase fraction, Ki-67 index and mitotic index (**Figure 2.2**). The analysis of tumour cells by flow cytometry by *Ahlgren* et al. showed that samples with an S-phase fraction ≥ 10% was associated with increased axillary involvement in tumours < 5cm and clinically negative axilla (OR 1.68 [95% CI 1.15-2.42], p=0.0073)(50). Similarly, Ki-67 expression was an independent predictor of axillary involvement in multivariate models when treated as a continuous variable (per 1% increase, OR=1.02, 95% CI [1.00-1.04])(51) or as a categorical variable (≥10% expression, OR=1.35, 95% CI [1.08-1.24])(42). In contrast, other studies incorporating Ki-67 expression did not find this variable to be an independent predictor within their models(20,40,48,52).

Other Molecular Factors

Britto *et al.* evaluated D2-40 and VEGF-A expression with immunohistochemistry (IHC) and did not find them to improve SLN status prediction in early breast cancer(19). Yoo *et al.* evaluated D2-40 and CD34 by IHC, and similarly did not find either variable to be independent predictors of axillary involvement in their model(48). P53 expression was evaluated with IHC in 2 studies, and its expression level was not associated with axillary involvement(15,21).

IHC with CD31 antibodies can be used as markers of neovascularization, and previous work has shown its expression to correlate with tumour cell spreading within breast ductal systems(53). Choi *et al.* found that CD31 staining was an independent predictor of ALN involvement (OR 2.90, 95%CI [1.04-8.92]) within their multivariate model, which included MRI features(15). Kiss-1, nm-23 and Cath-D expressions were evaluated by IHC, and only Kiss-1 expression was found to have value in predicting lowered risk of ALN metastasis in the multivariate model (OR 0.114, 95%CI [0.019-0.693])(21).

Okuno *et al.* investigated microRNA (miRNA) expression levels in cT1-3N0 ER+, Her2- tumours. Through microarray, they identified that miRNA-98, 22 and 223 were differentially expressed between SLN+ and SLN- patients(54). A multivariate model was constructed showing that miRNA-98 expression level was an independent predictor of SLN status (p=0.001)(54). In a separate study, the hypermethylation status of the *RAR-b2* gene was an independent predictor in the multivariate model, with a greater risk for a macro-SLN metastasis compared with micro-SLN or no SLN involvement (OR=1.595, 95%CI [1.16-1.93], and OR=3.86, 95% CI [1.65-9.00] respectively)(55).

Predictive markers have also been sought in peripheral blood samples. In a multivariate logistic regression model that included tumour size, Ki67 index and molecular subtypes,

increased blood levels of CA153 (OR 1.165, 95% CI [1.061-1.279]), CEA (OR 3.440, 95% CI [1.859-6.366]) and white blood cells (OR 1.475, 95% CI [1.077-2.022]) were independent predictors of axillary involvement(52). Takada *et al.* found that a high peripheral blood platelet to lymphocyte ratio to be predictive of SLN metastasis in univariate and multivariate analyses(45).

2.3.4 Radiological factors

Imaging modalities such as mammography, lymphoscintigraphy, ultrasound (US) of the tumour and axilla, MRI of the tumour and axilla and fluorodeoxyglucose-positron emission tomography (FDG-PET) scans have been utilized to extract variables of value in predicting the status of the axilla (**Figure 2.2**).

Mammography

On mammography, radiating spiculations was not an independent predictor of axillary involvement in a multivariate logistic regression model based on a Japanese cohort(31).

US

US characteristics derived from both tumour and ALNs can be important variables to consider. Jiang *et al*. looked at 130 early-stage breast cancer patients and found that on univariate analysis with the training group, tumour circularity, internal microcalcification and US-reported axillary status differed between patients with and without ALN involvement(56). The constructed multivariate model was only based on tumour circularity and US-ALN status (both factors independently significant, $p<0.001$). Hu *et al.* found indistinct tumour margins, calcifications, and tumour aspect ratio of ≥ 1 as seen on US to be independent variables in their

model(51). Yu *et al.* used radiomics to extract features from US images of the tumour and combined it with US-reported axillary status and clinicopathological variables to construct their model of axillary involvement(49). Other studies utilizing US images in their analyses are included in **Figure 2.2**.

MRI

An increasing number of studies have been incorporating MRI findings in their models of axillary involvement. Choi *et al.* combined findings from dynamic contrast-enhanced (DCE) and diffusion-weighted images (DWI) from MRI with clinicopathological variables in their study(15). Three other authors incorporated DWI images in their analyses(33,37,57). Others utilized MRI sequences include T1 images with contrast(33), fat-suppressed T2 images(58) and T2-weighted images(33,37,57). Characteristics extracted from MRI images of the axilla were also utilized (**Figure 2**). Irregularity of lymph node margins and lymph node asymmetry on MRI both reduced the odds of lymph node negative disease in 397 early-stage breast cancers, with OR of 0.17 (95% CI [0.047-0.609]) and 0.258 (95% CI [0.114-0.585]) in their multivariate model(37).

Lymphoscintigraphy and FDG-PET

In the pre-operative lymphoscintigraphy, the presence of abnormal lymphatic pathways was associated with an increase in SLN metastasis(59). The authors discuss metastasis-related obstruction of lymphatic pathways and subsequent development of new lymphatic pathways as the result of increased hydro-pressure as a possible explanation. Noguchi *et al.* found sentinel lymph node count measures after lymphoscintigraphy in T1N0M0 breast cancers to be lower in patients with SLN metastasis(60). FDG PET/CT scans also provided radiologic variables included

in predictive models, namely the total lesion glycosis, metabolic tumour volume, and SUVmax measurements (**Figure 2**).

2.3.5 Predictive models of axillary metastasis

The most frequently validated model was the MSKCC nomogram that developed based on a 3786-patient modeling cohort(41). This model, which relies on tumour size, histologic type, location, age, multifocality, nuclear grade and ER/PR receptors, achieved an area under the curve (AUC) for Receiver-Operator Characteristics (ROC) curve of 0.754 in the validation cohort (n=1545). The MSKCC nomogram has been tested in independent Australian(61), Canadian(34), Chinese(25,30,62), German(63), Japanese(54) and Dutch (64) cohorts.

Numerous other studies have utilized a combination of clinical and histopathological variables to construct their predictive models. In a Chinese study, multivariate logistic regression analysis with only 4 variables (age, tumour size, tumour location and LVI) from the modelling cohort (n=1000) led to a final model with an AUC of 0.7649 in the validation cohort (n=545)(25). By comparison, the MSKCC nomogram yielded an inferior AUC of 0.7105 in this modeling cohort. Elmadahm *et al*. validated this model in their cohort of 982 patients with an AUC of 0.71 (95% CI of 0.67-0.75) for the ROC curve(61).

Datasets from the Korean Breast Cancer Registry yielded a clinicopathological model with an AUC of 0.750 in both training (n=29326) and validation (n=12569) datasets(42). In another study, Houvenaeghel *et al*. constructed multivariate logistic models of SLN status in 12572 patients with small (≤ 30mm) invasive breast cancer and clinically negative axilla using clinicopathological variables(65). The resulting model had and AUC of 0.798 (95% CI 0.78-0.815)

for the ROC curve in the validation dataset. Given that the size of tumour on pathology and LVI are unavailable on biopsy specimens, they also constructed a “pre-operative” model substituting clinical T stage with size and removing LVI. This “pre-operative” model had a lower AUC of 0.727 (95% CI of 0.707-0.746).

Another clinicopathological model developed from a large cohort of breast cancer patients from 7 centres across China led to a final nomogram based on age, clinical T stage, tumour location, local invasion, histologic and molecular subtypes (28). This model achieved an AUC of 0.7157 in the training (n=1869) and 0.7007 in the validation cohorts (n=642)(28).

Chen *et al.* utilized clinicopathological variables to develop a predictive model of axillary involvement in patients with no ALN involvement based on clinical and US examinations(18). Despite negative US results, 25.6% of patients had SLN metastasis, and their predictive model (based on tumour size, menstrual status, histologic grade, and ER status) had poor predictive value with an AUC of 0.658 for the ROC curve.

The addition of molecular markers can improve the performance of predictive models. Okuno *et al.* investigated miRNA expression levels in clinically T1-3N0 ER+/Her2- tumours, and after stepwise analysis with multivariate logistic regression, the resulting model was based on tumour size, LVI and miRNA-98 expression. This model showed an AUC of 0.883 (0.807-0.958) in the validation cohort, and performed better than a model based on tumour size and LVI alone, and better than the MSKCC models in both training and validation cohorts. Xie *et al.* incorporated immunostaining for nm-23 and Kiss-1 proteins along with clinicopathological variables to achieve an AUC of 0.849 in the training (n=50) and 0.702 in the validation (n=20) cohorts of patients with pathological T1-2 disease and clinically negative ALNs(21).

Although most of the included studies used multivariate logistic regression to arrive at their predictive models, others utilized more complex machine learning techniques. Dihge *et al.* developed an artificial neural network model of axillary status based on clinicopathological factors in 800 breast cancer patients with clinically negative axilla(36). Their model had a better AUC (0.727, 95% CI [0.708-0.746]) in internal validation tests compared to linear multivariate logistic models, although the superiority of this model was not statistically significant (p=0.09). Liu *et al.* used Bagged-trees machine learning algorithms to create a model based on 12 clinicopathological variables on early-stage breast cancer patients with an AUC of 0.801(27) for the ROC curve. This approach for predictive model development was better than the traditional logistic regression model, which showed an AUC of 0.660 and 0.580 for training and validation cohorts respectively.

The inclusion of radiological variables can also improve the performance of predictive models. In a multivariate model constructed with clinicopathological and US findings in breast cancer patients, an AUC of 0.92 and 0.82 for ROC curves was achieved in internal and external validations cohorts(51). Another model based on tumour circularity and US-ALN status achieved an AUC of 0.89 (95% CI [0.84–0.94]) and 0.90 (95% CI [0.80–0.99]) on training and validation cohorts(56). Zong *et al.* developed a nomogram for predicting ALN involvement based on US features of tumour and axilla with an AUC of 0.873 (95% CI, 0.836-0.910) in the development (n=847), and 0.802 (95% CI, 0.740-0.865) in the validation cohort (n=481) (35). Interestingly, the inclusion of clinical factors did not improve the performance of their models.

Recent studies have been utilizing machine-learning algorithms with radiomics feature extraction to derive predictive models of axillary involvement. Pre-operative US-based

radiomics analysis was shown to improve the performance of the MSKCC nomogram in cohort of 452 patients with breast cancer undergoing SLNB(62). Zhou *et al*. trained various convolutional neural network algorithms to use US images in cT1-2N0 patients (n=680), and the best model achieved an AUC of 0.89 (95% CI, 0.83-0.95) in the independent validation cohort (n=78)(66). Notably, this model outperformed clinical interpretation by trained radiologists in predicting ALN involvement based on primary tumour's US features. Yu *et al.* extracted a radiomics score from US images of the primary breast tumour in clinically T1-2N0 patients using the LASSO algorithm, and in combination with age, US reported tumour size, US reported ALN status, developed a model for ALN prediction with AUC of 0.84 (95% CI 0.71-0.82) in the primary (n=300) and 0.81 (95% CI 0.74-0.88) in the validation (n=126) cohorts(49).

A clinicoradiomics model of ALN metastasis was developed using a combination of clinical (cT stage, cN stage, histologic grade, age and Her2 status) and radiomic (T1+C, T2WI, and DWI-ADC sequences from MRI images of breast tumour and axilla) signatures and yielded an AUC for ROC curve of 0.92 in the development (n=849) and 0.90 in the validation (n=365) cohorts(33). Intravoxel incoherent motion MRI is a DWI technique, which in addition to T2 weighted-imaging features led to a multivariate model with AUC of 0.785 for prediction of ALN involvement a cohort of patients with T1-2 disease and clinically negative axilla(57). In a separate study incorporating clinicopathological, MRI and US imaging factors in 397 early-stage breast cancer patients with clinically negative ALNs, Li *et al.* developed a model with an AUC of 0.809 (95% CI of 0.756-0.863). The addition of pathological variables including vascular invasion and Her2 status did not improve the model. Ding *et al.* noted that optimization of peritumoral feature inclusion could improve model performance from an AUC of 0.704 to 0.796(67).

2.3.6 Quality assessment

PROBAST was used to assess the risk of bias, with details provided in **Supplementary Table 2.1**. Most studies (33 of 59, 55.9%) only used one patient cohort, and did not validate their findings. 23 of 59 studies (39.0%) included both development and validation cohorts, and the remaining 3 studies (5.1%) validated previous models in independent cohorts. The lack of validation datasets was the main expressed concern regarding the risk of bias in the included papers. From the applicability perspective, LVI was frequently included in predictive models despite known low pre-operative accuracy when assessed on core biopsy specimens.

2.4 DISCUSSION

Non-invasive determination of ALN metastasis in early-stage breast cancers can have an impact on clinical decision making. Patients with clinically occult axillary involvement can benefit from pre-operative chemotherapy, while still retaining the possibility of a SLN biopsy after their treatment. On the other hand, patients with a low risk of axillary involvement may be spared an invasive sampling of their ALNs, and thus, be spared from the morbidity and operative resources required for these procedures.

Despite the development of a multitude of nomograms and predictive models, SLN biopsy remains the standard of care for staging of the ALNs in most early-stage breast cancer patients. Recent guidelines support a discussion with older patients (70 years or older) and HR-positive tumours regarding omission of axillary surgery, citing no difference in terms of survival outcomes for these patients. The ongoing SOUND trial (Sentinel Node vs Observation After Axillary Ultra-souND, https://clinicaltrials.gov/ct2/show/NCT02167490) is investigating if SLNB can be safely avoided in patients with clinically T1N0 breast cancers and negative axilla on pre-operative imaging. In another ongoing trial, the INSEMA trial (Comparison of Axillary Sentinel Lymph Node Biopsy Versus no Axillary Surgery, https://clinicaltrials.gov/ct2/show/NCT02466737), early-stage breast cancer patients with clinically negative axilla are randomized to axillary sampling or no axillary surgery, and the primary outcome of invasive disease-free survival is being investigated between these groups. The results of these ongoing trials may lead to a reduction in the rates of ALN surgery performed in this cohort of patients.

A limitation of this study is the notable number of papers that were excluded from this review due to failure to either restrict the population to tumours ≤ 5cm and clinically negative axilla, or to conduct multivariate analyses with these variables. Although we understand that there can be subjective bias in the examination of the axilla in breast cancer patients, inclusion of patients with large tumours or high metastatic burden in their axilla can introduce variability in the modelling populations, and hence, such models may not be readily applicable to the patient population under investigation in the current study.

This review highlighted the various patient factors and the numerous resulting models that have been developed to predict axillary involvement in patients with early-stage breast cancer and clinically negative axilla. We expect that further development of these predictive models will have a consequential impact on the decision to undergo axillary surgery and the timing of these procedures for patients. Future prospective randomized-controlled trials are needed to confirm the clinical utility of pre-operative predictive models of axillary involvement.

2.5 FIGURES

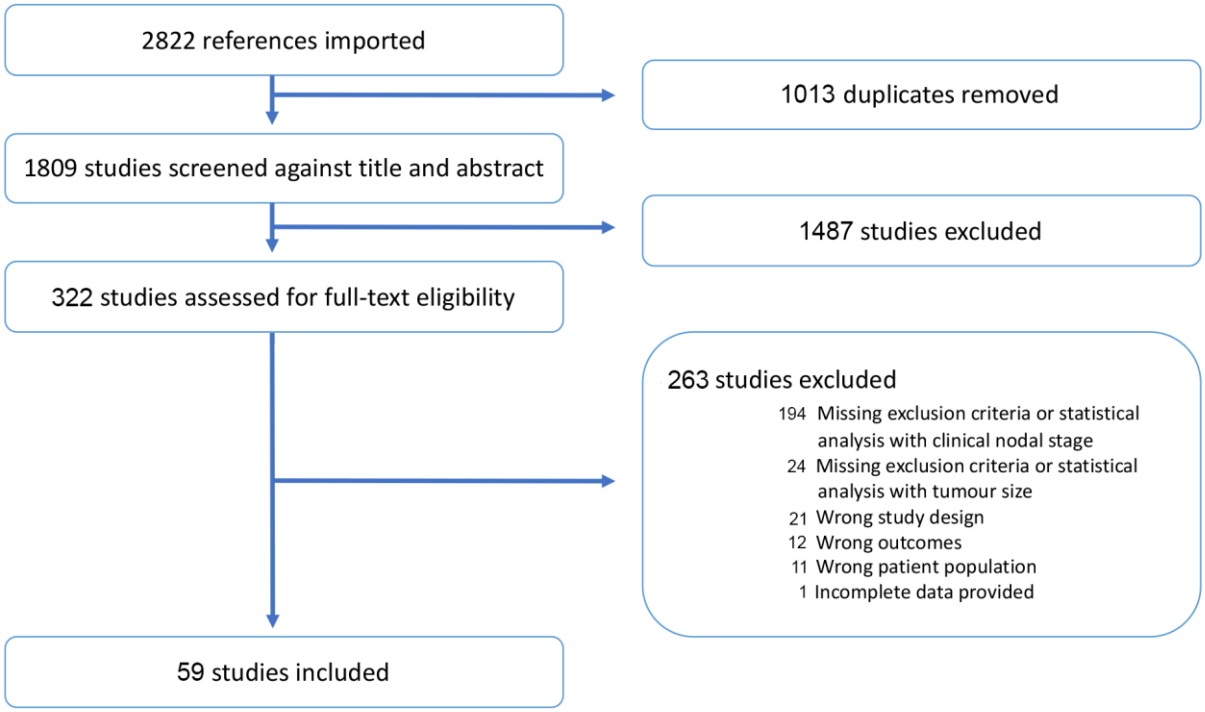


Figure 2.1 – PRISMA chart outlining the number of studies included and excluded in each step of the systematic review.

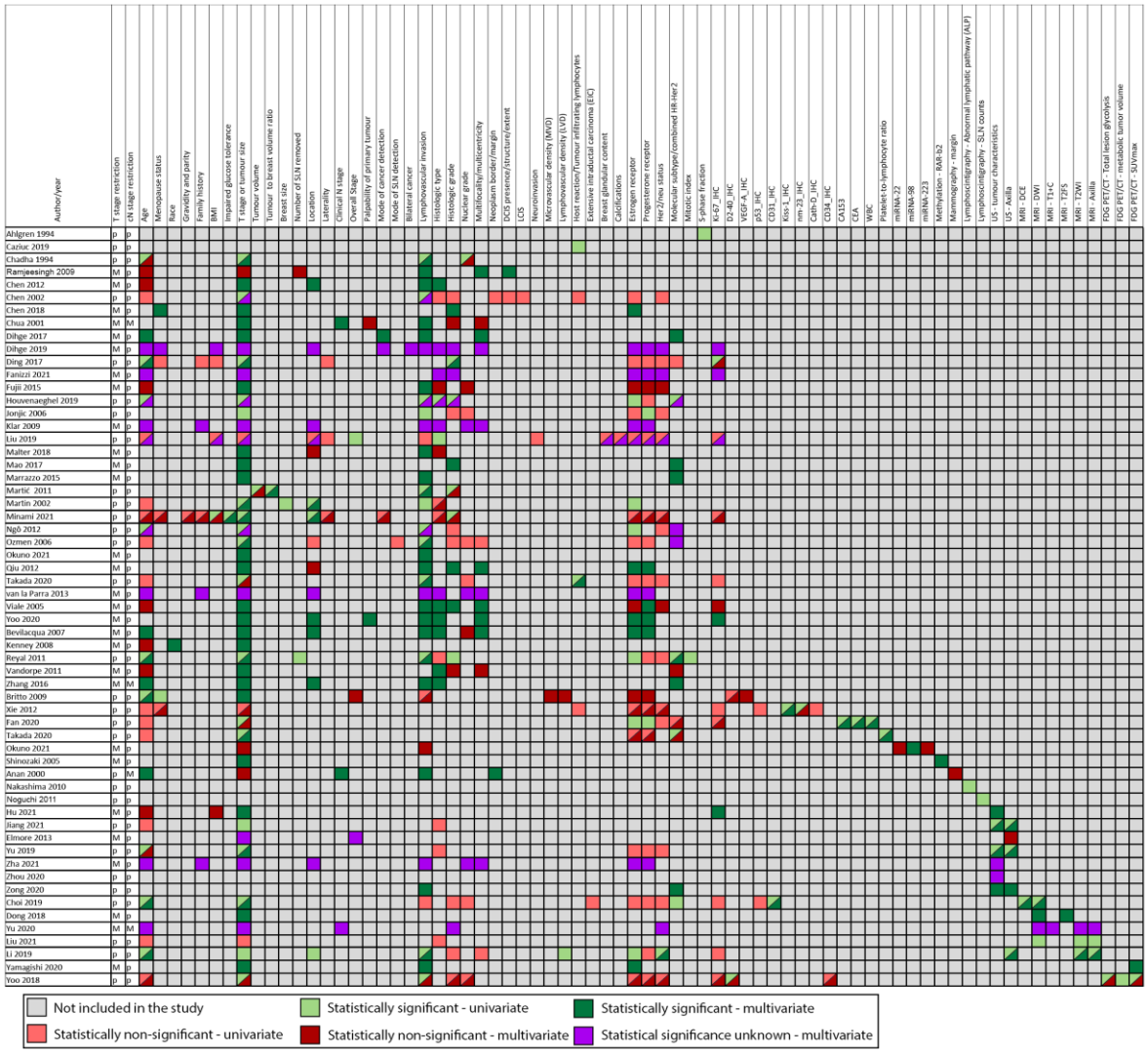


Figure 2.2 – Compilation of factors studied in relation to axillary involvement in early-stage breast cancers.

2.6 TABLES

Table 2.1 – Search terms used to find articles in the databases for the systematic review.

Database	Search terms
PubMed	(predict*[Title] OR model*[Title] OR tool*[Title] OR nomogram*[Title] OR decide*[Title] OR decision*[Title] OR signature*[Title] OR profile*[Title] OR marker*[Title] OR biomarker*[Title] OR "risk factor"*[Title]) AND (axilla*[Title] OR lymph*[Title] OR node*[Title]) AND (breast[Title] OR Ductal[Title] OR Lobular[Title]) AND ("0001/01/01"[PDat] : "2021/07/01"[PDat]) AND English[lang]
Web of Science	(TI=(predict* OR model* OR tool* OR nomogram* OR decide* OR decision* OR signature* OR profile* OR marker* OR biomarker*) AND TI=(axilla* OR lymph OR lymphatic* OR node*) AND TI=(breast OR Ductal OR Lobular) NOT TI=pancrea*) AND LANGUAGE: (English) Refined by: DOCUMENT TYPES: (ARTICLE) Timespan: 1900-2021. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

2.7 BIBLIOGRAPHY

1. Dixon JM, Cartlidge CWJ. Twenty-five years of change in the management of the axilla in breast cancer. Breast J [Internet]. 2020;26:22–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31854498

2. Krag DN, Anderson SJ, Julian TB, Brown AM, Harlow SP, Costantino JP, et al. Sentinel-lymph-node resection compared with conventional axillary-lymph-node dissection in clinically node-negative patients with breast cancer: overall survival findings from the NSABP B-32 randomised phase 3 trial. Lancet Oncol [Internet]. 2010;11:927–33. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20863759

3. Schijven MP, Vingerhoets AJJM, Rutten HJT, Nieuwenhuijzen GAP, Roumen RMH, Bussel ME van, et al. Comparison of morbidity between axillary lymph node dissection and sentinel node biopsy. European J Surg Oncol Ejso. 2003;29(4):341–50.

4. Schrenk P, Rieger R, Shamiyeh A, Wayand W. Morbidity following sentinel lymph node biopsy versus axillary lymph node dissection for patients with breast carcinoma. Cancer. 2000;88(3):608–14.

5. Cabanes PA, Salmon RJ, Vilcoq JR, Durand JC, Fourquet A, Gautier C, et al. Value of axillary dissection in addition to lumpectomy and radiotherapy in early breast cancer. Lancet. 1992;339(8804):1245–8.

6. W. C Robert, Craig A D, O. A Benjamin, J. B Harold, Bradford C W, B. E Stephen, et al. Invasive Breast Cancer. J Natl Compr Canc Ne [Internet]. 2011;9(2):136–222. Available from: https://jnccn.org/view/journals/jnccn/9/2/article-p136.xml

7. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30(8):1194–220.

8. Bianco PD, Zavagno G, Burelli P, Scalco G, Barutta L, Carraro P, et al. Morbidity comparison of sentinel lymph node biopsy versus conventional axillary lymph node dissection for breast cancer patients: Results of the sentinella–GIVOM Italian randomised clinical trial. European J Surg Oncol Ejso. 2008;34(5):508–13.

9. Baron RH, Fey JV, Borgen PI, Stempel MM, Hardick KR, Zee KJV. Eighteen Sensations After Breast Cancer Surgery: A 5-Year Comparison of Sentinel Lymph Node Biopsy and Axillary Lymph Node Dissection. Ann Surg Oncol. 2007;14(5):1653.

10. Bear HD, Anderson S, Brown A, Smith R, Mamounas EP, Fisher B, et al. The Effect on Tumor Response of Adding Sequential Preoperative Docetaxel to Preoperative Doxorubicin and Cyclophosphamide: Preliminary Results From National Surgical Adjuvant Breast and Bowel Project Protocol B-27. J Clin Oncol. 2003;21(22):4165–74.

11. Kuehn T, Bauerfeind I, Fehm T, Fleige B, Hausschild M, Helms G, et al. Sentinel-lymph-node biopsy in patients with breast cancer before and after neoadjuvant chemotherapy (SENTINA): a prospective, multicentre cohort study. Lancet Oncol. 2013;14(7):609–18.

12. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51.

13. Reyal F, Rouzier R, Depont-Hazelzet B, Bollet MA, Pierga JY, Alran S, et al. The Molecular Subtype Classification Is a Determinant of Sentinel Node Positivity in Early Breast Carcinoma. Plos One. 2011;6(5):e20297.

14. Ding J, Jiang L, Wu W. Predictive Value of Clinicopathological Characteristics for Sentinel Lymph Node Metastasis in Early Breast Cancer. Med Sci Monit [Internet]. 2017;23:4102–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28839123

15. Choi EJ, Youk JH, Choi H, Song JS. Dynamic contrast-enhanced and diffusion-weighted MRI of invasive breast cancer for the prediction of sentinel lymph node status. J Magn Reson Imaging. 2020;51(2):615–26.

16. Dihge L, Bendahl P -O., Rydén L. Nomograms for preoperative prediction of axillary nodal status in breast cancer. Brit J Surg. 2017;104(11):1494–505.

17. Wildiers H, Calster BV, Poll-Franse LV van de, Hendrickx W, Røislien J, Smeets A, et al. Relationship Between Age and Axillary Lymph Node Involvement in Women With Breast Cancer. J Clin Oncol. 2009;27(18):2931–7.

18. Chen X, He Y, Wang J, Huo L, Fan Z, Li J, et al. Feasibility of using negative ultrasonography results of axillary lymph nodes to predict sentinel lymph node metastasis in breast cancer patients. Cancer Med-us. 2018;7(7):3066–72.

19. Britto AV, Schenka AA, Moraes-Schenka NG, Alvarenga M, Shinzato JY, Vassallo J, et al. Immunostaining with D2–40 improves evaluation of lymphovascular invasion, but may not predict sentinel lymph node status in early breast cancer. Bmc Cancer. 2009;9(1):109.

20. Minami S, Sakimura C, Irie J, Tokai Y, Okubo H, Ohno T. Predictive Factors Among Clinicopathological Characteristics for Sentinel Lymph Node Metastasis in T1-T2 Breast Cancer. Cancer Management Res. 2021;Volume 13:215–23.

21. Xie F, Yang H, Wang S, Zhou B, Tong F, Yang D, et al. A Logistic Regression Model for Predicting Axillary Lymph Node Metastases in Early Breast Carcinoma Patients. Sensors Basel Switz. 2012;12(7):9936–50.

22. Kenney RJ, Marszalek JM, McNally ME, Nelson BV, Herati AS, Talboy GE. The effects of race and age on axillary lymph node involvement in breast cancer patients at a Midwestern safety-net hospital. Am J Surg. 2008;196(1):64–9.

23. Larsson SC, Mantzoros CS, Wolk A. Diabetes mellitus and risk of breast cancer: A meta-analysis. Int J Cancer. 2007;121(4):856–62.

24. Martin C, Cutuli B, Velten M. Predictive model of axillary lymph node involvement in women with small invasive breast carcinoma. Cancer. 2002;94(2):314–22.

25. Chen J ying, Chen J jian, Yang B long, Liu Z bin, Huang X yan, Liu G yu, et al. Predicting sentinel lymph node metastasis in a Chinese breast cancer population: assessment of an existing nomogram and a new predictive nomogram. Breast Cancer Res Tr. 2012;135(3):839–48.

26. Martić K, Vlajčić Z, Rudman F, Lambaša S, Tomasović-Lončarić Č, Stanec Z. Tumor and Breast Volume Ratio as a Predictive Factor for Axillary Lymph Node Metastases in T1c Ductal Invasive Breast Cancer: Prospective Observational Clinico-pathological Study. Jpn J Clin Oncol. 2011;41(12):1322–6.

27. Liu C, Zhao Z, Gu X, Sun L, Chen G, Zhang H, et al. Establishment and Verification of a Bagged-Trees-Based Model for Prediction of Sentinel Lymph Node Metastasis for Early Breast Cancer Patients. Frontiers Oncol. 2019;9:282.

28. Zhang J, Li X, Huang R, Feng WL, Kong YN, Xu F, et al. A nomogram to predict the probability of axillary lymph node metastasis in female patients with breast cancer in China: A nationwide, multicenter, 10-year epidemiological study. Oncotarget. 2015;5(0):35311–25.

29. Malter W, Hellmich M, Badian M, Kirn V, Mallmann P, Krämer S. Factors Predictive of Sentinel Lymph Node Involvement in Primary Breast Cancer. Anticancer Res. 2018;38(6):3657–62.

30. Qiu P fei, Liu J juan, Wang Y sheng, Yang G ren, Liu Y bing, Sun X, et al. Risk Factors for Sentinel Lymph Node Metastasis and Validation Study of the MSKCC Nomogram in Breast Cancer Patients. Jpn J Clin Oncol. 2012;42(11):1002–7.

31. Anan K, Mitsuyama S, Tamae K, Nishihara K, Iwashita T, Abe Y, et al. Axillary lymph node metastases in patients with small carcinomas of the breast: is accurate prediction possible? Eur J Surg. 2000;166(8):610–5.

32. Chua B, Ung O, Taylor R, Boyages J. Frequency and predictors of axillary lymph node metastases in invasive breast cancer. Anz J Surg. 2001;71(12):723–8.

33. Yu Y, Tan Y, Xie C, Hu Q, Ouyang J, Chen Y, et al. Development and Validation of a Preoperative Magnetic Resonance Imaging Radiomics–Based Signature to Predict Axillary Lymph Node Metastasis and Disease-Free Survival in Patients With Early-Stage Breast Cancer. Jama Netw Open. 2020;3(12):e2028086.

34. Ramjeesingh R, Quan ML, Gardner S, Holloway CMB. Prediction of involvement of sentinel and nonsentinel lymph nodes in a Canadian population with breast cancer. Can J Surg J Can De Chir. 2009;52(1):23–30.

35. Zong Q, Deng J, Ge W, Chen J, Xu D. Establishment of Simple Nomograms for Predicting Axillary Lymph Node Involvement in Early Breast Cancer. Cancer Management Res. 2020;12:2025–35.

36. Dihge L, Ohlsson M, Edén P, Bendahl PO, Rydén L. Artificial neural network models to predict nodal status in clinically node-negative breast cancer. Bmc Cancer. 2019;19(1):610.

37. Li J, Ma W, Jiang X, Cui C, Wang H, Chen J, et al. Development and Validation of Nomograms Predictive of Axillary Nodal Status to Guide Surgical Decision-Making in Early-Stage Breast Cancer. J Cancer. 2019;10(5):1263–74.

38. Marrazzo A, Boscaino G, Marrazzo E, Taormina P, Toesca A. Breast cancer subtypes can be determinant in the decision making process to avoid surgical axillary staging: A retrospective cohort study. Int J Surg. 2015;21:156–61.

39. Fujii T, Yajima R, Tatsuki H, Suto T, Morita H, Tsutsumi S, et al. Significance of lymphatic invasion combined with size of primary tumor for predicting sentinel lymph node metastasis in patients with breast cancer. Anticancer Res. 2015;35(6):3581–4.

40. Viale G, Zurrida S, Maiorano E, Mazzarol G, Pruneri G, Paganelli G, et al. Predicting the status of axillary sentinel lymph nodes in 4351 patients with invasive breast carcinoma treated in a single institution. Cancer. 2005;103(3):492–500.

41. Bevilacqua JLB, Kattan MW, Fey JV, III HSC, Borgen PI, Zee KJV. Doctor, What Are My Chances of Having a Positive Sentinel Node? A Validated Nomogram for Risk Estimation. J Clin Oncol. 2007;25(24):3670–9.

42. Yoo TK, Kim SJ, Lee J, Lee SB, Lee SJ, Park HY, et al. A N0 Predicting Model for Sentinel Lymph Node Biopsy Omission in Early Breast Cancer Upstaged From Ductal Carcinoma in Situ. Clin Breast Cancer. 2020;20(3):e281–9.

43. Vandorpe T, Smeets A, Calster BV, Hoorde KV, Leunen K, Amant F, et al. Lobular and non-lobular breast cancers differ regarding axillary lymph node metastasis: a cross-sectional study on 4,292 consecutive patients. Breast Cancer Res Tr. 2011;128(2):429–35.

44. Mao F, Yao R, Peng L, Zhao JL, Liang ZY, Sun Q. Predictive clinicopathological characteristics affecting sentinel lymph node metastasis in early breast cancer patients. Transl Cancer Res. 2017;6(5):968–75.

45. TAKADA K, KASHIWAGI S, ASANO Y, GOTO W, KOUHASHI R, YABUMOTO A, et al. Prediction of Sentinel Lymph Node Metastasis Using the Platelet-to-lymphocyte Ratio in T1 Breast Cancer. Anticancer Res. 2020;40(4):2343–9.

46. Chen M, Palleschi S, Khoynezhad A, Gecelter G, Marini CP, Simms HH. Role of Primary Breast Cancer Characteristics in Predicting Positive Sentinel Lymph Node Biopsy Results: A Multivariate Analysis. Arch Surg-chicago. 2002;137(5):606–10.

47. Ozmen V, Karanlik H, Cabioglu N, Igci A, Kecer M, Asoglu O, et al. Factors predicting the sentinel and non-sentinel lymph node metastases in breast cancer. Breast Cancer Res Tr. 2006;95(1):1–6.

48. Yoo J, Kim BS, Yoon HJ. Predictive value of primary tumor parameters using 18F-FDG PET/CT for occult lymph node metastasis in breast cancer with clinically negative axillary lymph node. Ann Nucl Med. 2018;32(9):642–8.

49. Yu FH, Wang JX, Ye XH, Deng J, Hang J, Yang B. Ultrasound-based radiomics nomogram: A potential biomarker to predict axillary lymph node metastasis in early-stage invasive breast cancer. Eur J Radiol. 2019;119:108658.

50. Ahlgren J, Westman G, Stål O, Arnesson LG. Prediction of Axillary Lymph Node Metastases in a Screened Breast Cancer Population. Acta Oncol. 2009;33(6):603–8.

51. Hu X, Xue J, Peng S, Yang P, Yang Z, Yang L, et al. Preoperative Nomogram for Predicting Sentinel Lymph Node Metastasis Risk in Breast Cancer: A Potential Application on Omitting Sentinel Lymph Node Biopsy. Frontiers Oncol. 2021;11:665240.

52. Fan Y, Chen X, Li H. Clinical value of serum biomarkers CA153, CEA, and white blood cells in predicting sentinel lymph node metastasis of breast cancer. Int J Clin Exp Patho. 2020;13(11):2889–94.

53. Sapino A, Righi L, Cassoni P, Bongiovanni M, Deaglio S, Malavasi F, et al. CD31 expression by cells of extensive ductal in situ and invasive carcinomas of the breast. Breast Cancer Res. 2001;3(Suppl 1):A57.

54. Okuno J, Miyake T, Sota Y, Tanei T, Kagara N, Naoi Y, et al. Development of Prediction Model Including MicroRNA Expression for Sentinel Lymph Node Metastasis in ER-Positive and HER2-Negative Breast Cancer. Ann Surg Oncol. 2021;28(1):310–9.

55. Shinozaki M, Hoon DSB, Giuliano AE, Hansen NM, Wang HJ, Turner R, et al. Distinct Hypermethylation Profile of Primary Breast Cancer Is Associated with Sentinel Lymph Node Metastasis. Clin Cancer Res. 2005;11(6):2156–62.

56. Jiang T, Su W, Zhao Y, Li Q, Huang P. Non-invasive prediction of lymph node status for patients with early-stage invasive breast cancer based on a morphological feature from ultrasound images. Quantitative Imaging Medicine Surg. 2021;11(8):3399407–3393407.

57. Liu Y, Luo H, Wang C, Chen X, Wang M, Zhou P, et al. Diagnostic performance of T2-weighted imaging and intravoxel incoherent motion diffusion-weighted MRI for predicting metastatic axillary lymph nodes in T1 and T2 stage breast cancer. Acta Radiol. 2021;028418512110028.

58. Dong Y, Feng Q, Yang W, Lu Z, Deng C, Zhang L, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. Eur Radiol. 2018;28(2):582–91.

59. Nakashima K, Kurebayashi J, Sonoo H, Tanaka K, Ikeda M, Shiiki S, et al. Preoperative dynamic lymphoscintigraphy predicts sentinel lymph node metastasis in patients with early breast cancer. Breast Cancer-tokyo. 2009;17(1):17.

60. Noguchi A, Onoguchi M, Ohnishi T, Hashizume T, Kajita A, Funauchi M, et al. Predicting sentinel lymph node metastasis in breast cancer with lymphoscintigraphy. Ann Nucl Med. 2011;25(3):221–6.

61. Elmadahm A, Lord SJ, Hudson HM, Lee CK, Buizen L, Farshid G, et al. Performance of four published risk models to predict sentinel lymph-node involvement in Australian women with early breast cancer. Breast. 2018;41:82–8.

62. Zha H ling, Zong M, Liu X pei, Pan J zhen, Wang H, Gong H yan, et al. Preoperative ultrasound-based radiomics score can improve the accuracy of the Memorial Sloan Kettering Cancer Center nomogram for predicting sentinel lymph node metastasis in breast cancer. Eur J Radiol. 2021;135:109512.

63. Klar M, Foeldi M, Markert S, Gitsch G, Stickeler E, Watermann D. Good Prediction of the Likelihood for Sentinel Lymph Node Metastasis by Using the MSKCC Nomogram in a German Breast Cancer Population. Ann Surg Oncol. 2009;16(5):1136–42.

64. Parra RFD van la, Francissen CMTP, Peer PGM, Ernst MF, Roos WK de, Zee KJV, et al. Assessment of the Memorial Sloan-Kettering Cancer Center nomogram to predict sentinel lymph node metastases in a Dutch breast cancer population. Eur J Cancer. 2013;49(3):564–71.

65. Houvenaeghel G, Lambaudie E, Classe JM, Mazouni C, Giard S, Cohen M, et al. Lymph node positivity in different early breast carcinoma phenotypes: a predictive model. BMC Cancer [Internet]. 2019;19:45. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30630443

66. Zhou LQ, Wu XL, Huang SY, Wu GG, Ye HR, Wei Q, et al. Lymph Node Metastasis Prediction from Primary Breast Cancer US Images Using Deep Learning. Radiology [Internet]. 2020;294:19–28. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31746687

67. Ding J, Chen S, Serrano Sosa M, Cattell R, Lei L, Sun J, et al. Optimizing the Peritumoral Region Size in Radiomics Analysis for Sentinel Lymph Node Status Prediction in Breast Cancer. Acad Radiol. 2020;

CHAPTER 3

THE MOLECULAR LANDSCAPE OF EARLY-STAGE BREAST CANCER WITH AXILLARY METASTASIS

CHAPTER 3: THE MOLECULAR LANDSCAPE OF EARLY-STAGE BREAST CANCER WITH AXILLARY METASTASIS

3.1 INTRODUCTION

Breast cancer has become the most diagnosed cancer world-wide(1). Axillary lymph nodes (ALNs) are the primary site of metastasis for breast cancer, and the involvement of ALNs at the time of diagnosis relays necessary information about disease stage and prognosis(2). For that reason, current guidelines rely on the status of ALNs to recommend important clinical decisions regarding cancer treatment, including neoadjuvant and adjuvant chemoradiation and the extent of axillary surgery(3,4).

Lack of palpable lymphadenopathy on clinical examination is an inaccurate predictor of axillary status(5,6). As such, invasive surgery to sample lymph-nodes in the form of sentinel lymph-node biopsy (SLNB) remains the standard of care for patients with early-stage breast cancer and clinically negative axilla(4). The consequences are far-reaching, including morbidity for patients(7) and expenditure of the limited operative time existing within the healthcare system.

Characterization of differences in the genetic landscape of the early-stage breast cancers with nodal involvement is clinically important, as it may aid with risk stratification of tumours based on the pre-operative tissue biopsy. Expression-based nomograms, such as Onco*type* DX™ (a commercial test developed by Genomic Health, Redwood City, CA, USA), have already proven useful in selecting patients that would benefit most from adjuvant systemic therapy, but none have been established to inform neoadjuvant therapy decisions(8). Additionally, an accurate

pre-operative model may spare low-risk patients from an invasive operation to stage their disease.

Several molecular subtypes with divergent biology have been established in breast cancer, and growing evidence supports a subtype-specific approach to their diagnosis and treatment. As such, it would be important for an investigation of genetic differences that contribute to ALN metastasis to account for the heterogeneity that is inherently related to each established molecular subtype. A previously published exploration for a molecular signature of nodal metastasis in breast cancer was not successful, despite accounting for the molecular subtypes, with the authors concluding that factors outside of the primary tumour such as alterations in ALN microenvironment can lead to the establishment of metastatic deposits(9). In contrast, another study employing machine learning algorithms to RNA sequencing data led to predictive models of lymph-node involvement with improved performance compared to using clinical variables alone(10). It is notable however, that these models were not developed or validated specifically for the patient population that would most benefit from risk stratification, those with early-stage disease and no clinically evident lymphadenopathy. These analyses were also limited to data from RNA expression and did not account for potential differences in other molecular components, such as DNA mutations or microRNA (miRNA) profiles.

To better understand the molecular changes associated with lymph node metastasis, the molecular profile of early-stage tumours for patients undergoing SLNB was analyzed in this study. To maximize chances of finding molecular signature associated with ALN-positive disease, our comparison accounted for the intrinsic subtype of cancer, and included multiple

platforms, including single-nucleotide variation (SNV), copy number alterations (CNA) in DNA, as well as messenger RNA (mRNA), miRNA, and protein quantification.

3.2 MATERIALS AND METHODS

Clinical data analysis

Survival data was downloaded from Liu *et al.* supplementary tables (11). PAM50 molecular subtypes were downloaded from Berger *et al.* (12), and samples with subtype "normal" (normal breast-like) were excluded. Clinical datasets were downloaded from Broad Institute Firehose databases (Version 2016_01_28, firebrowse.org, n=1097) [13]. All tables were merged, removing samples not included in all tables. Patients who received neoadjuvant chemotherapy were removed.

Tumour size was restricted to T-stages of T1 and T2, and samples missing this information were excluded. TCGA data does not provide information regarding clinical stage of the axilla at the time of presentation. To select for patients without clinically palpable ALNs, we investigated the axillary staging method for the remaining patients and selected for patients who received SLNB +/- ALND (n=431). Patients with metaplastic carcinoma or unassigned histologic diagnosis were removed (n=414). To clarify the Her2 receptor status, assignment was based on IHC method, but if IHC results were equivocal or not available, then fluorescent *in-situ* hybridization results were included. For the menopause state, *pre* was defined as <6 months since LMP without prior bilateral ovariectomy and not on estrogen replacement, *peri* as 6-12 months since last menstrual period, and *post* as prior bilateral ovariectomy or >12 months since last menstrual period (LMP), with no prior hysterectomy.

Clinical characteristics were compared between node positive and negative patients. Mann-Whitney U test was used to compare age. Fisher's exact test was used to compare histology, T stage, M stage, ER status, PR status and Her2 status. Pearson's chi-squared test was

used to compare menopause status and molecular subtypes. The α threshold of significance was set at 0.05.

DNA data analysis

Level 3 oncotated SNV data and copy number alteration data were downloaded from Broad Institute firebrowse website(13). Samples not included in the previously created clinical datasets of early-stage patients were excluded. This left 220 breast cancer patients with negative and 143 patients with positive ALNs. *Maftools* package (version 2.6.05) was used in R Statistical Environment for DNA analysis(14). Duplicated and silent SNVs were removed. The prevalence of mutation variants was compared by the status of ALNs using Mann–Whitney U tests, with a p value of 0.05 set as threshold of significance. Box plots were generated illustrating median, first and third quartiles, and minimum and maximum values within 1.5 times the interquartile range below or above the first and third quartiles respectively. SNVs were compared between node-positive and negative early-stage tumours using Fisher's test on 2x2 contingency tables. A minimum mutation count was set at 5% of the smallest comparison group size to reduce noise from rarely mutated genes. Derived p-values were corrected for multiple testing using the Benjamini-Hochberg method and a false discovery rate (FDR) of 0.05 was set as the threshold.

Copy number alterations (CNA) were downloaded from Broad Institute's GISTIC 2.0 results(13). Deep amplifications and deletions were used. A total of 158 node-positive and 245 node-negative early-stage samples met the clinical criteria. Fisher's exact test were used to compare amplification counts per gene in the node-positive vs. node-negative groups. Genes

with total amplifications of less than 5% of the smallest cohort were removed. P values were then corrected for multiple-testing using Benjamini-Hochberg method. Similar process was repeated for deletions.

miRNA data analysis

Non-normalized miRNA count data was downloaded from the Broad Institute Firebrowse platform. A total of 407 patients (node status was negative in 248 and positive in 159) and 503 miRNAs were analyzed. Differentially expressed miRNAs were derived using DESeq2 package (version 1.30)(15). miRNAs with sum of reads of less than 10 across the samples in the comparison were excluded. Log_2 fold changes were reported comparing node positive vs. negative patients. Independent filtering was applied through the DESeq2 *results* function. Adjustment for multiple-testing was done through Benjamini-Hochberg method. FDR threshold of 0.05 was set. Analysis was repeated similar for each molecular subtype separately. Bar plots were generated using *ggplot2* package (version 3.3.5)(16) in R Statistical Environment.

Messenger RNA (mRNA) data analysis

RNAseq counts were downloaded from the University of California Santa Cruz Xena platform(17). The available transcripts were filtered for protein-coding genes using the *biomaRT* package(18) in R ("hsapiens_gene_ensembl" dataset utilized). This reduced the total number of transcripts from 60483 to 19556 protein-coding mRNAs. Raw count reads were converted from the log_2(counts+1) scale to integer. Clinical exclusion criteria were applied as

described above to yield a total of 250 node-negative and 162 node-positive early-stage samples.

Differentially expressed mRNAs were analyzed using DESeq2 package (version 1.30)(15). Log_2 fold changes were reported comparing node positive vs. negative patients. FDR threshold of 0.05 was set. Transcripts with sum of reads of less than 10 across the samples included in the comparison were excluded. Analysis was repeated for each molecular subtype separately.

Pathway analysis was completed using the *clusterProfiler* package (version 3.18.1)(19), based on the Gene Ontology Biological Process(20) and Reactome(21) databases. Statistically significant differentially expressed genes were assessed for enrichment within each pathway. p values were adjusted for multiple-testing using the Benjamini-Hochberg method and a threshold of 0.05 was set for the FDR.

Protein data analysis

Reverse Phase Protein Array (RPPA) data was downloaded from Xena platform(17). Retrieved data was Z-score normalized. Samples were filtered for primary tumours (removal of normal and metastatic tissue samples). After refining the dataset based on previously described clinical criteria, 128 node-positive and 201 node-negative early-stage samples remained. RPPA scores were compared using t-test with Welch's correction. "NA" values were omitted in the calculations. False-discovery rates were calculated using Benjamini-Hochberg method, with FDR threshold value set at 0.05.

3.3 RESULTS

3.3.1 Clinical characteristics

The clinical characteristics of the patient cohort is detailed in **Table 3.1**. After inclusion of only patients with T1 and T2 tumours undergoing SLNB, 251 did not have any nodal metastasis, while 163 patients did. Patients with axillary involvement were significantly younger (p=0.002), had larger tumours (p=0.004) and were ER-receptor positive (p=0.044) compared to those without. A significantly higher proportion of women with axillary involvement were pre-menopausal (p=0.026). There were no statistically significant differences between histology, molecular subtype, PR status or HER2 status (p>0.05).

3.3.2 Single Nucleotide Variations (SNVs) and copy-number alterations (CNAs)

After exclusion of silent and duplicated variants, SNV data was available for 363 early-stage tumours (n=143 and 220 for node-positive and negative patients respectively). **Figure 3.1** outlines the various variant classifications by nodal involvement. Total mutation load was higher in node-negative tumours compared to node-positive samples (median of 32 vs. 24 SNVs per sample respectively, p=0.0063). Frame-shift deletions and missense mutations were significantly more frequent in node-negative disease (p=0.00085 and 0.0072, **Supplementary Table 3.1**).

In the subtype specific analysis, lower frame-shift deletion and missense mutation variant counts in node-positive disease was seen in Basal and Her2-enriched tumours (p=0.00063 and 0.0496 in Basal; p=0.0301 and <0.0001 in Her2-enriched respectively, **Figure 3.2**). Luminal A and B cohorts on the other hand did not show significant differences in the frequency of different

classes of SNVs ($p>0.05$). Splice-site and nonsense mutation variants were more frequent in node-negative Her2-enriched tumours ($p=0.0086$ and <0.0001 respectively).

A total of 58 genes met the minimum mutation count threshold and were included in the SNV comparison between patients with and without nodal involvement. After adjustment for the FDR, no genes showed SNV frequencies that had a statistically significant difference (Top 10 differentially mutated genes are provided in **Supplementary Table 3.2**). Within subtypes as well, there were no statistically significant SNV differences between node-positive and negative patients (FDR>0.05).

There were no differences in the deep amplification or deletions between node positive and negative samples that met the threshold of significance (**Supplementary Tables 3.3 and 3.4**). Similarly, no differences were identified comparing CNAs in subtypes individually (FDR>0.05).

3.3.3 miRNA expression

A total of 503 separate miRNA expression levels were compared between 159 node-positive early-stage breast cancer samples and 248 node-negative samples. After correction for multiple comparisons, 40 miRNAs were differentially expressed (**Figure 3.3, Supplementary Table 3.5**). Of these 40 miRNAs, 10 were overexpressed in node-positive patients compared to node-negative patients, and 30 were under-expressed.

Examining these 40 miRNAs in each molecular subtype individually, the differential expression pattern within subtypes with the highest prevalence in our dataset most closely resemble the combined analysis. The expression of 95% (38/40) of the identified miRNAs in

Luminal A, the most prevalent subtype, mirrored the results in combined analysis. In contrast, 22.5% (9 of 40) miRNAs showed the opposite expression trend in the least common subtypes, Luminal B and Her2-enriched (ie. miRNAs were overexpressed in node-positive disease in the subtype analysis, but under-expressed in combined analysis, or *vice versa*). Basal samples which present the second most available subtype, shared 82.5% (33/40) of the expression trend as the overall analysis provided.

The comparison of miRNA expression between node-positive and node-negative tumours was repeated in each molecular subtype individually (**Figure 3.4, Supplementary Table 3.6**). miRNA 517a, 517b, 206 and 105-2 were significantly over-expressed in Luminal A subtype with node-involvement, while miRNA 221 was significantly under-expressed. Of note, while statistically significant under-expression of miRNA 221 in node-positive patients was also found on the analysis of the entire cohort, all molecular subtypes other than Luminal A showed the opposite pattern of expression.

In the Luminal B comparison, miRNA 184 and 224 were overexpressed and 30a was under-expressed in node-positive samples. None of these miRNAs met the criteria for statistical significance in the overall comparison or within other subtypes. In the Basal subtype, miRNA 3150b and 3065 were both over-expressed in node-positive samples. miRNA 3150b met the threshold of significance in the overall analysis. No miRNAs were differentially expressed at a statistically significant rate in the Her2 subtype comparison (FDR>0.05).

3.3.4 mRNA expression

In the overall analysis with subtypes combined, transcription levels were compared between 250 node-negative and 162 node-positive samples for 19180 mRNAs. 766 mRNAs were differentially expressed by ALN status (**Figure 3.5, Supplementary Table 3.7**). Of these, 249 were over-expressed and 517 were under-expressed in node-positive disease. The comparison was repeated in subtypes individually. Of the 766 identified differentially expressed genes (DEGs) with all subtypes combined, only 33.2% (254 transcripts) were either over-expressed or under-expressed across all subtypes with nodal involvement. Expression trends in node-positive *vs.* node-negative in the combined subtype comparison was similar to the trend seen in 86.9% of transcripts in Luminal A (665 of 765 transcripts), 68.6% in Luminal B (524 of 764 transcripts), 80.9 % in Basal (619 of 765 transcripts), and 69.1% in Her2 samples (526 of 761 transcripts). Its notable that some transcripts were excluded from the comparison in each subtype due to low overall counts in those samples, as discussed in the methods.

In Luminal A (18983 transcripts compared in node-negative n=143, node-positive n=99), 185 statistically significant DEGs were identified (**Figure 3.6**, **Supplementary Table 3.8**). In Luminal B (18549 transcripts compared in node-negative n=29, node-positive n=27), 272 statistically significant DEGs were seen (**Figure 3.7**, **Supplementary Table 3.9**). In Basal (18803 transcripts compared in node-negative n=60, node-positive n=27), 96 statistically significant DEGs were noted (**Figure 3.8**, **Supplementary Table 3.10**). The comparison in Her2-enriched (18231 transcripts compared in node-negative n=18, node-positive n=9) yielded 126 statistically significant DEGs (**Figure 3.9**, **Supplementary Table 3.11**).

Utilizing the identified DEGs, pathway enrichment analysis was completed in combined patients and individual subtypes (**Supplementary Tables 3.12-3.15**). Several immune response related pathways were highlighted in the overall analysis (**Supplementary Table 3.12**). In Luminal A subtype, Nucleosome and chromatin-related pathways were enriched in the DEGs (**Supplementary Table 3.14**).

3.3.5 Protein expression

Expression level of 281 proteins were compared between 128 node-positive and 201 node-negative early-stage samples (**Supplementary Table 3.16**). After correction for multiple comparisons, no significant differences were found between node-positive and negative samples (FDR>0.05). Comparison was repeated in subtypes separately, not revealing any statistically significant differences based on nodal-metastasis in any of the 4 molecular subtypes. (**Supplementary Table 3.17**).

3.4 DISCUSSION

Our study found significant differences in clinical characteristics, RNA expression and miRNA expression associated with lymph-node involvement in early-stage breast cancer. Increased tumour size, younger age and ER-positive status were all associated with ALN metastasis in this cohort of patients with early-stage breast cancers, compatible with findings of previous studies(22,23). Notably, lymphovascular invasion has been reported as a strong predictor of ALN metastasis(22,24), but was not included in the clinical datasets available for this analysis. That said, lymphovascular invasion assessment on pre-operative biopsy alone is not accurate(25), and hence the applicability in the pre-operative setting is questionable.

Of the 766 identified DEGs, only 33.2% were consistently over or under-expressed in node-positive tumours across all molecular subtypes. This heterogeneity was seen also in the miRNA analysis. The lack of consistency between the combined-subtype analysis and each individual subtype was higher in the less common subtypes (Luminal B and Her2-enriched samples) as expected. This highlights the baseline molecular heterogeneity that exists in breast cancer and supports a subtype-specific approach to molecular comparisons in early-stage breast cancers to reduce noise from the analysis.

Pathway enrichment analysis in Luminal A subtypes highlighted the differential expression of several chromatin-related genes in tumours with ALN metastasis (**Supplementary Table 3.14**). Amongst these were *H1.1, H1.4* and H1.5 mRNAs that all encode for subtypes of linker histone H1. Linker Histone H1 is involved in higher-order formation of chromatin(26), and in humans, histone H1 has 11 subtypes(27). *In-vitro* and animal studies have shown that while the knockout of one of the H1 subtype genes results in compensation in the overall level of

expressed histone H1 by the upregulation of other subtypes, depleting single H1 variants can cause alterations in chromatin structure and the expression level of other gene subsets(28,29). Human cancer cohorts have shown correlations between expression of H1 subtypes and tumour aggressiveness, although this correlation can be positive or negative depending on the variant(30). The expression of H1.1-1.5 subtypes are tightly associated with S-phase and DNA-replication, while other H1 subtypes such as H1.0 are expressed throughout the cell cycle(31). In prostate cancer, increased tumour grade has been associated with increased H1.5 expression on immunohistochemistry(32). In our dataset however, Luminal A early-stage breast cancer patients with nodal involvement had a decreased levels of *H1.5* transcript (Log_2Fold change of -1.51, **Supplementary Table 3.8**), although the levels of H1.5 protein were not available for analysis in the TCGA dataset. The connection between linker histone H1 variant expression and metastasis in breast cancer may be a productive avenue for future study.

miRNAs are small single stranded RNAs that can inhibit protein expression by binding to target mRNAs(33). In our analysis, miRNA 577 was under-expressed in node-positive early-stage breast cancers (**Figure 3.3**). Lower miRNA 577 levels were similarly reported to be associated with lymph-node metastasis and larger tumour size in a separate cohort of 120 breast tumours(34). *In-vitro* experiments have suggestion a role for miRNA 577 in inhibiting epithelial-mesenchymal transition (EMT), and down-regulating Rab25 protein levels(34). In non-small cell lung cancer, decreased miRNA 577 has been associated with increased cell proliferation and invasion(35). Notably however the expression pattern of miRNA 577 in node-positive disease compared to node-negative disease varied between the four molecular subtypes in the TCGA dataset (**Figure 3.3**). Further highlighting the differences between subtypes is the dramatic

differences in mean count level of miRNA 577, which was only 3.81, 3.72 and 10.07 in Luminal A, Luminal B and Her2 subtypes, compared to 223.50 in the Basal subtype.

Another example of heterogeneity within miRNA expression patterns was seen with miRNA 206. The under-expression of miRNA 206 has been previously associated with advanced clinical stage in breast cancer(36). More specifically, this was shown in the Basal subtype of breast cancer and its mechanism of action was attributed to the miRNA's role in inhibition of *TM5SF1* expression, an oncogene involved in cell migration(37). In our dataset, miRNA 206 levels were higher in node-positive Basal early-stage breast cancers, although this was not statistically significant (**Figure 3.4**). However, in subtypes other than Basal, the levels of miRNA 206 appear to be higher in node-positive disease, possibly suggesting a different role for miRNA 206 in the non-basal subtypes.

After correction for multiple comparisons, no statistically significant protein level differences were found between node-positive and negative disease, however strong trends were seen. p38MAPK is a mitogen-activated protein (MAP) kinase involved in many processes including inflammation, cell growth, differentiation, and death(38). p38MAPK had a trend towards increased levels in node-positive tumours. Previous studies support this and have shown p38MAPK signalling to be associated with invasive and metastatic behaviour in breast carcinoma (39–41). Additionally, the NF2 tumour suppressor (also known as Merlin or schwannomin) showed a trend towards down-regulation in node-positive early-stage breast tumours (**Supplementary Table 3.16**). NF2 regulates contact-dependent inhibition of proliferation(42). Chromosomal alterations inactivating NF2 have been associated with increased metastatic potential in prostate cancer cell lines(43). NF2 down-regulation has also

been seen in advanced disease, and in tumour tissue compared to adjacent normal breast tissue(44). There may be value in investigating NF2 protein levels as a clinical marker of metastatic potential in breast cancer.

The value of the discovered molecular differences as predictive signatures of nodal metastasis will need to be validated in independent cohorts. Several factors served as limitations to our analyses. Clinical information such as the pre-operative nodal stage was inferred based on the operation each patient received (namely a SLNB) as this information was not directly available. Other tumour characteristics such as lymphovascular invasion or Ki67 positivity that have been previously associated with lymph-node metastasis were also not available for this cohort(22,45).

The molecular data derived from the TCGA is bulk tumour analysis, include heterogenous tissue which includes cells other the breast carcinoma(46). As such, the expression patterns of extracellular matrix cells or immune cells were included in the analyzed tissue. Single-cell data can further delineate the origin of the cells contributing to the molecular differences seen in node-positive disease. A molecular signature was extracted using single-cell RNAseq data in breast carcinoma and performed well when validated in the TCGA dataset with accuracy of 91%(47). The performance of these models in early-stage breast cancers in each molecular subtype remains to be assessed.

This comparison between molecular alterations in early-stage breast cancer patients with nodal-involvement revealed distinct heterogeneity between the established molecular subtypes, and identified numerous molecular differences associated with nodal metastasis. The

potential molecular signatures identified in this study need to be further validated in independent datasets and may prove valuable in the development of predictive models.

3.5 BIBLIOGRAPHY

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. Ca Cancer J Clin. 2021;71(3):209–49.

2. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. Ca Cancer J Clin. 2017;67(2):93–9.

3. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30(8):1194–220.

4. National Comprehensive Cancer Network. Breast Cancer (version 2.2022) [Internet]. [cited 2022 Feb 19]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

5. Weiss A, Chavez-MacGregor M, Lichtensztajn DY, Yi M, Tadros A, Hortobagyi GN, et al. Validation Study of the American Joint Committee on Cancer Eighth Edition Prognostic Stage Compared With the Anatomic Stage in Breast Cancer. Jama Oncol. 2017;4(2):203.

6. Fisher B, Jeong JH, Anderson S, Bryant J, Fisher ER, Wolmark N. Twenty-Five-Year Follow-up of a Randomized Trial Comparing Radical Mastectomy, Total Mastectomy, and Total Mastectomy Followed by Irradiation. New Engl J Medicine. 2002;347(8):567–75.

7. Ashikaga T, Krag DN, Land SR, Julian TB, Anderson SJ, Brown AM, et al. Morbidity results from the NSABP B-32 trial comparing sentinel lymph node dissection versus axillary dissection. J Surg Oncol. 2010;102(2):111–8.

8. Syed YY. Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer. Mol Diagn Ther. 2020;24(5):621–32.

9. Shriver CD, Hueman MT, Ellsworth RE. Molecular signatures of lymph node status by intrinsic subtype: gene expression analysis of primary breast tumors from patients with and without metastatic lymph nodes. J Exp Clin Cancer Res Cr. 2014;33(1):116.

10. Dihge L, Vallon-Christersson J, Hegardt C, Saal LH, Hakkinen J, Larsson C, et al. Prediction of Lymph Node Metastasis in Breast Cancer by Gene Expression and Clinicopathological Models: Development and Validation within a Population-Based Cohort. Clin Cancer Res [Internet]. 2019;25:6368–81. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31340938

11. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173(2):400-416.e11.

12. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018;33(4):690-705.e9.

13. Harvard BI of M and. TCGA Genome Data Analysis Center (2016): SNP6 Copy number analysis (GISTIC2). Broad Institute of MIT and Harvard. 2016;

14. Mayakonda A, Koeffler HP. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. Biorxiv. 2016;052662.

15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

16. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

17. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8.

18. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–91.

19. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omics J Integr Biology. 2012;16(5):284–7.

20. Consortium TGO, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2020;49(D1):D325–34.

21. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):D498–503.

22. Bevilacqua JLB, Kattan MW, Fey JV, III HSC, Borgen PI, Zee KJV. Doctor, What Are My Chances of Having a Positive Sentinel Node? A Validated Nomogram for Risk Estimation. J Clin Oncol. 2007;25(24):3670–9.

23. Dihge L, Bendahl P -O., Rydén L. Nomograms for preoperative prediction of axillary nodal status in breast cancer. Brit J Surg. 2017;104(11):1494–505.

24. Viale G, Zurrida S, Maiorano E, Mazzarol G, Pruneri G, Paganelli G, et al. Predicting the status of axillary sentinel lymph nodes in 4351 patients with invasive breast carcinoma treated in a single institution. Cancer. 2005;103(3):492–500.

25. Harris GC, Denley HE, Pinder SE, Lee AHS, Ellis IO, Elston CW, et al. Correlation of Histologic Prognostic Factors in Core Biopsies and Therapeutic Excisions of Invasive Breast Carcinoma. Am J Surg Pathology. 2003;27(1):11–5.

26. Robinson PJ, Rhodes D. Structure of the '30nm' chromatin fibre: A key role for the linker histone. Curr Opin Struc Biol. 2006;16(3):336–43.

27. Izzo A, Kamieniarz K, Schneider R. The histone H1 family: specific members, specific functions? Biol Chem. 2008;389(4):333–43.

28. Fan Y, Nikitina T, Zhao J, Fleury TJ, Bhattacharyya R, Bouhassira EE, et al. Histone H1 Depletion in Mammals Alters Global Chromatin Structure but Causes Specific Changes in Gene Regulation. Cell. 2005;123(7):1199–212.

29. Sancho M, Diani E, Beato M, Jordan A. Depletion of Human Histone H1 Variants Uncovers Specific Roles in Gene Expression and Cell Growth. Plos Genet. 2008;4(10):e1000227.

30. Scaffidi P. Histone H1 alterations in cancer. Biochimica Et Biophysica Acta Bba - Gene Regul Mech. 2016;1859(3):533–9.

31. Biterge B, Schneider R. Histone variants: key players of chromatin. Cell Tissue Res. 2014;356(3):457–66.

32. Khachaturov V, Xiao GQ, Kinoshita Y, Unger PD, Burstein DE. Histone H1.5, a novel prostatic cancer marker: an immunohistochemical study. Hum Pathol. 2014;45(10):2115–9.

33. Bartel DP. MicroRNAs Genomics, Biogenesis, Mechanism, and Function. Cell. 2004;116(2):281–97.

34. Yin C, Mou Q, Pan X, Zhang G, Li H, Sun Y. MiR-577 suppresses epithelial-mesenchymal transition and metastasis of breast cancer by targeting Rab25. Thorac Cancer. 2018;9(4):472–9.

35. Men L, Nie D, Nie H. microRNA-577 inhibits cell proliferation and invasion in non-small cell lung cancer by directly targeting homeobox A1. Mol Med Rep. 2019;19(3):1875–82.

36. Amir S, Simion C, Umeh-Garcia M, Krig S, Moss T, Carraway KL, et al. Regulation of the T-box transcription factor Tbx3 by the tumour suppressor microRNA-206 in breast cancer. Brit J Cancer. 2016;114(10):1125–34.

37. Fan C, Liu N, Zheng D, Du J, Wang K. MicroRNA-206 inhibits metastasis of triple-negative breast cancer by targeting transmembrane 4 L6 family member 1. Cancer Management Res. 2019;11:6755–64.

38. Ono K, Han J. The p38 signal transduction pathway Activation and function. Cell Signal. 2000;12(1):1–13.

39. Limoge M, Safina A, Truskinovsky AM, Aljahdali I, Zonneville J, Gruevski A, et al. Tumor p38MAPK signaling enhances breast carcinoma vascularization and growth by promoting expression and deposition of pro-tumorigenic factors. Oncotarget. 2017;8(37):61969–81.

40. Bakin AV, Rinehart C, Tomlinson AK, Arteaga CL. p38 mitogen-activated protein kinase is required for TGFβ-mediated fibroblastic transdifferentiation and cell migration. J Cell Sci. 2002;115(15):3193–206.

41. Wu X, Zhang W, Font-Burgada J, Palmer T, Hamil AS, Biswas SK, et al. Ubiquitin-conjugating enzyme Ubc13 controls breast cancer metastasis through a TAK1-p38 MAP kinase cascade. Proceedings of the National Academy of Sciences [Internet]. 2014;111(38):13870–5. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1414358111

42. Curto M, McClatchey AI. Nf2/Merlin: a coordinator of receptor signalling and intercellular contact. Brit J Cancer. 2008;98(2):256–62.

43. Malhotra A, Shibata Y, Hall IM, Dutta A. Chromosomal structural variations during progression of a prostate epithelial cell line to a malignant metastatic state inactivate the NF2, NIPSNAP1, UGT2B17, and LPIN2 genes. Cancer Biol Ther. 2013;14(9):840–52.

44. Wang Z, Zhou Z, Wang Z, Cui Y. NF2 inhibits proliferation and cancer stemness in breast cancer. Open Med-warsaw. 2020;15(1):302–8.

45. Yoo TK, Kim SJ, Lee J, Lee SB, Lee SJ, Park HY, et al. A N0 Predicting Model for Sentinel Lymph Node Biopsy Omission in Early Breast Cancer Upstaged From Ductal Carcinoma in Situ. Clin Breast Cancer. 2020;20(3):e281–9.

46. Cui W, Xue H, Wei L, Jin J, Tian X, Wang Q. High heterogeneity undermines generalization of differential expression results in RNA-Seq analysis. Hum Genomics. 2021;15(1):7.

47. Kim BC, Kim J, Lim I, Kim DH, Lim SM, Woo SK. Machine Learning Model for Lymph Node Metastasis Prediction in Breast Cancer Using Random Forest Algorithm and Mitochondrial Metabolism Hub Genes. Appl Sci. 2021;11(7):2897.

3.6 FIGURES

Figure 3.1. Frequency of various classes of single nucleotide variation compared by nodal status.

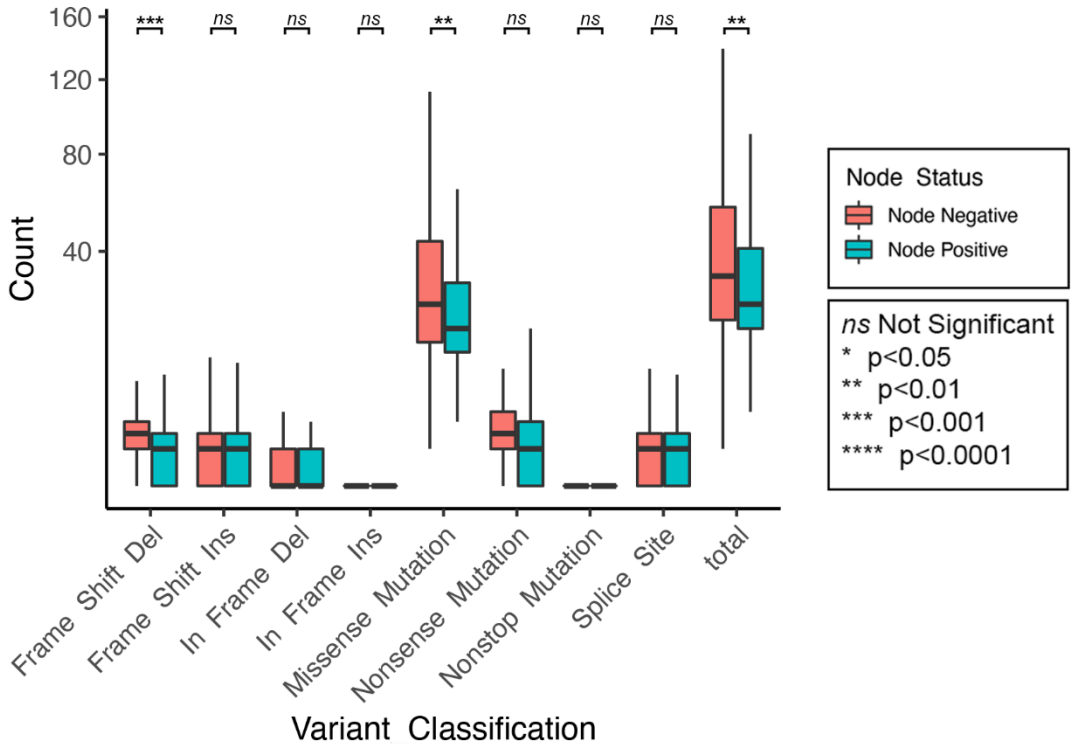


Figure 3.2. SNV classification compared based on ALN status in each subtype

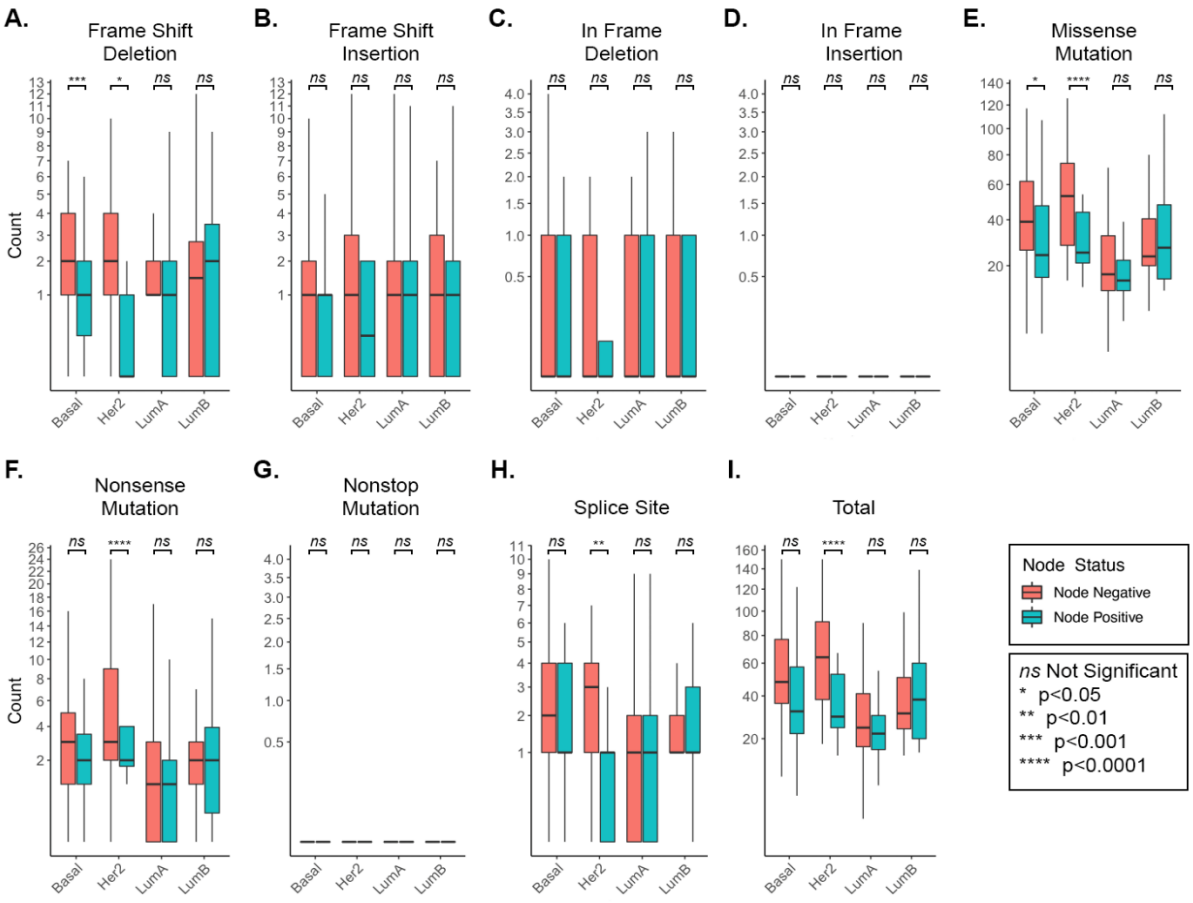


Figure 3.3. Differentially expressed miRNAs between node-positive and node-negative groups in early-stage samples. Forty miRNAs met criteria for statistical significance (FDR<0.05). Fold change and level of significance of each miRNA based on nodal status is also illustrated within each molecular subtype for comparison.

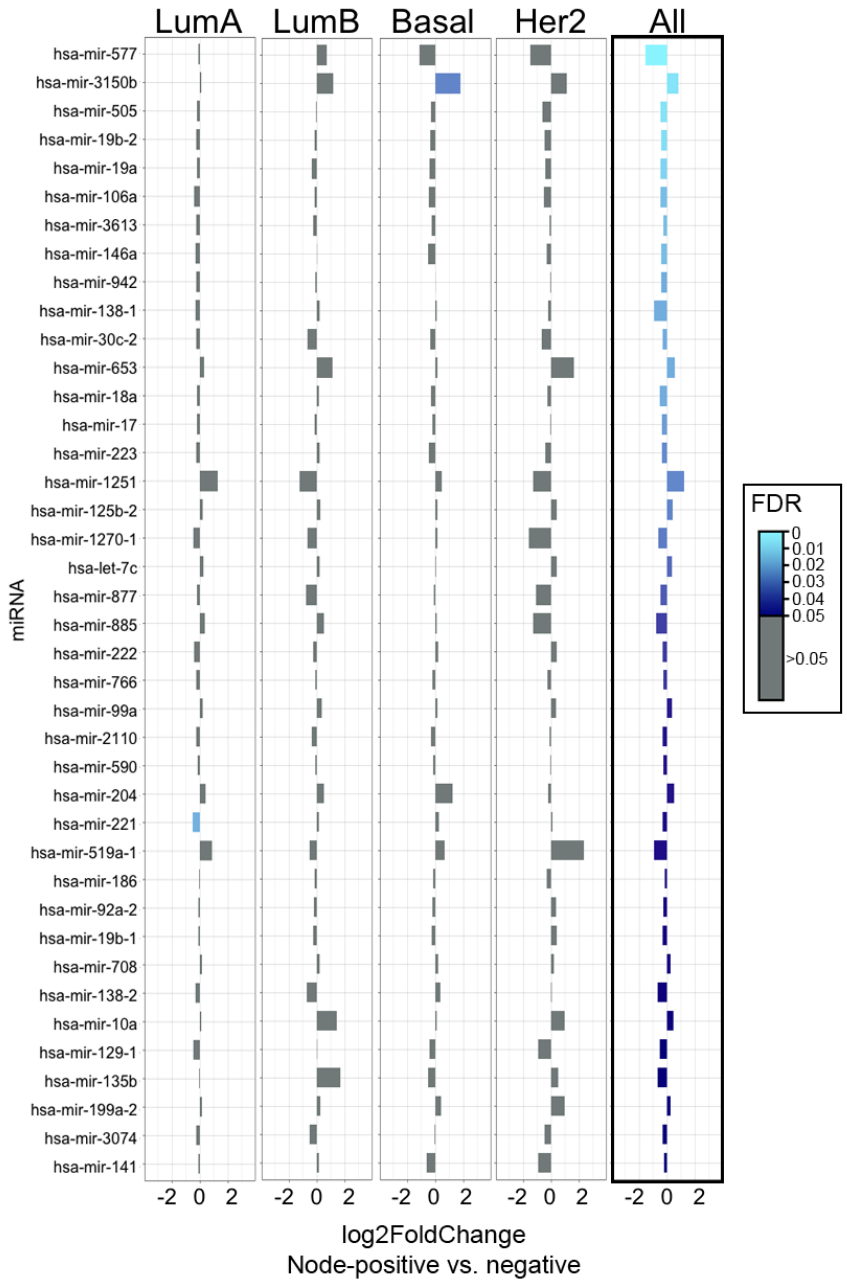


Figure 3.4. The statistically significant differentially expressed miRNAs between node-positive and node-negative early-stage breast cancers analyzed in Luminal A (**A**), Luminal B (**B**) and Basal (**C**) molecular subtypes separately. Differences in these miRNA levels between node-positive and node-negative tumours in other subtypes and all subtypes combined are provided for comparison. No statistically significant differentially expressed miRNAs were found in Her2 subtype.

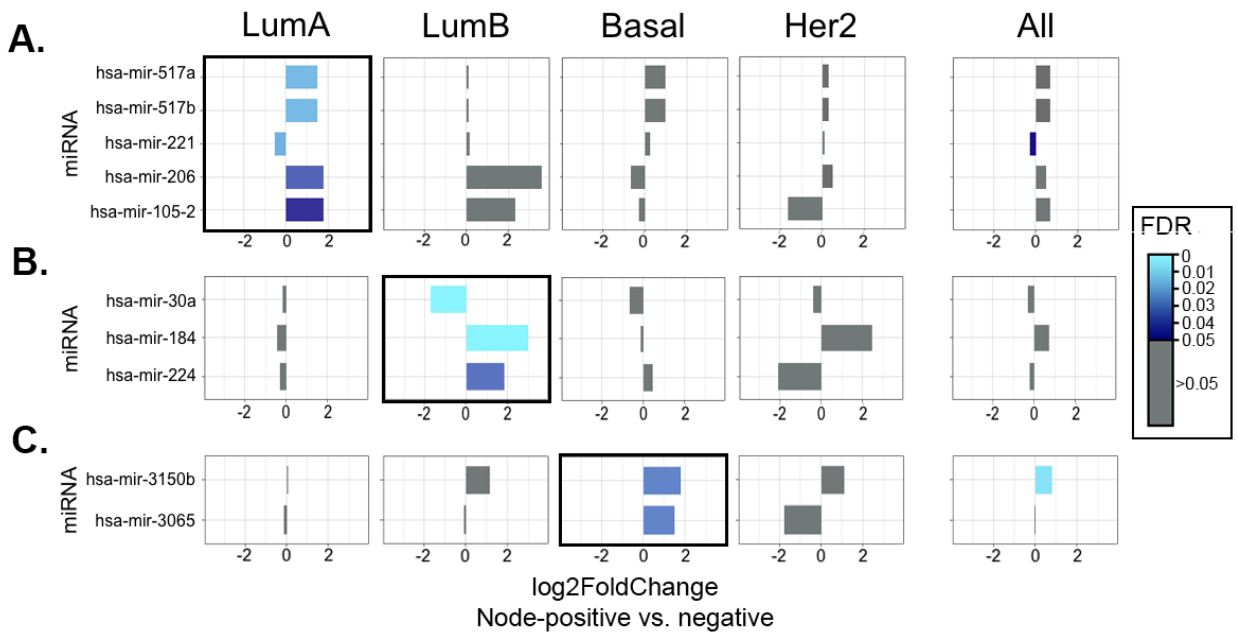
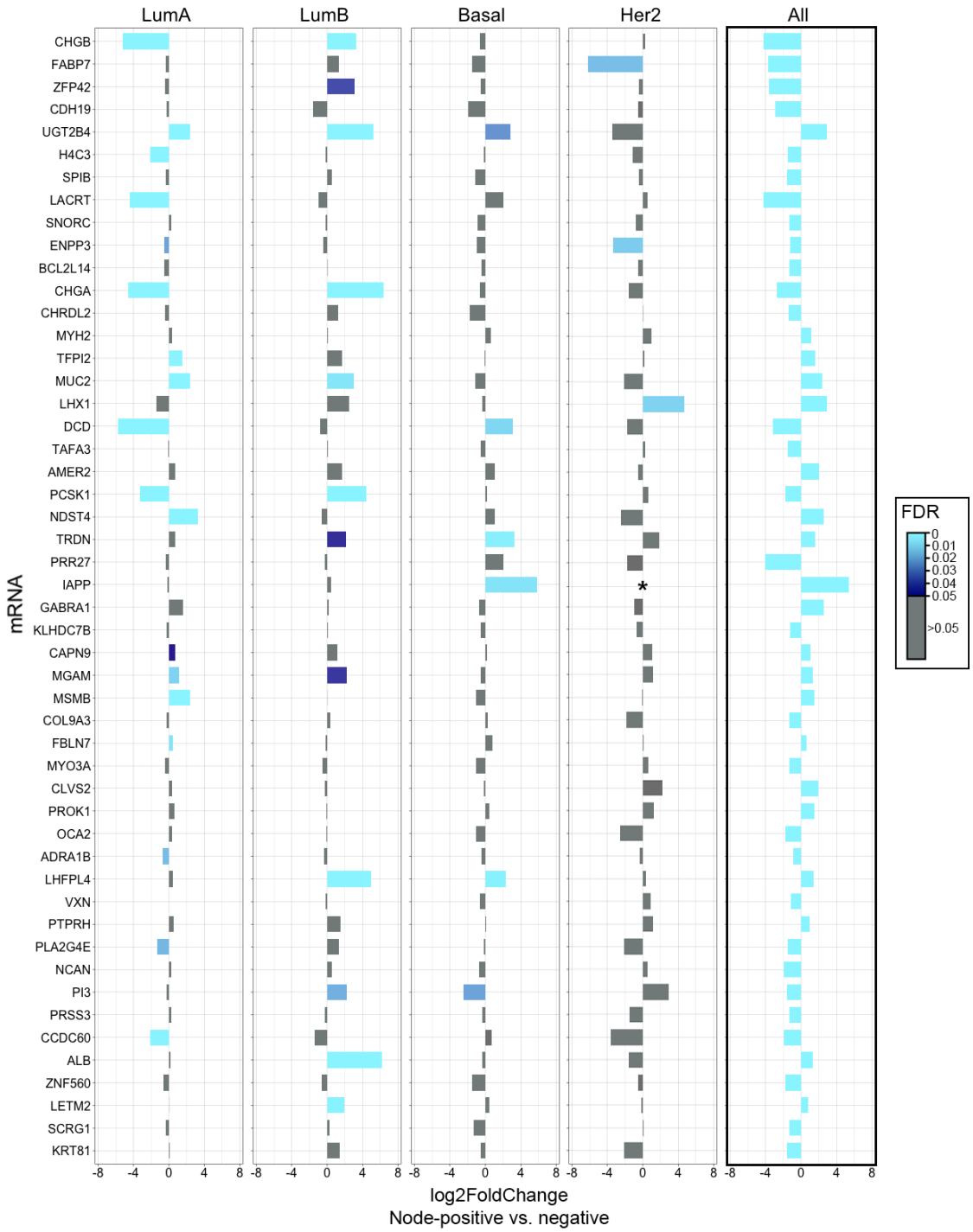


Figure 3.5. Top 50 differentially expressed genes in node-positive compared to node-negative early-stage breast cancers with all subtypes combined. In total, 755 differentially expressed genes were statistically significant in the combined analysis (FDR <0.05).



* excluded from comparison due to low counts.

Figure 3.6. Top 50 of the 185 statistically significant differentially expressed genes in node-positive compared to node-negative Luminal A early-stage breast cancers (FDR <0.05).

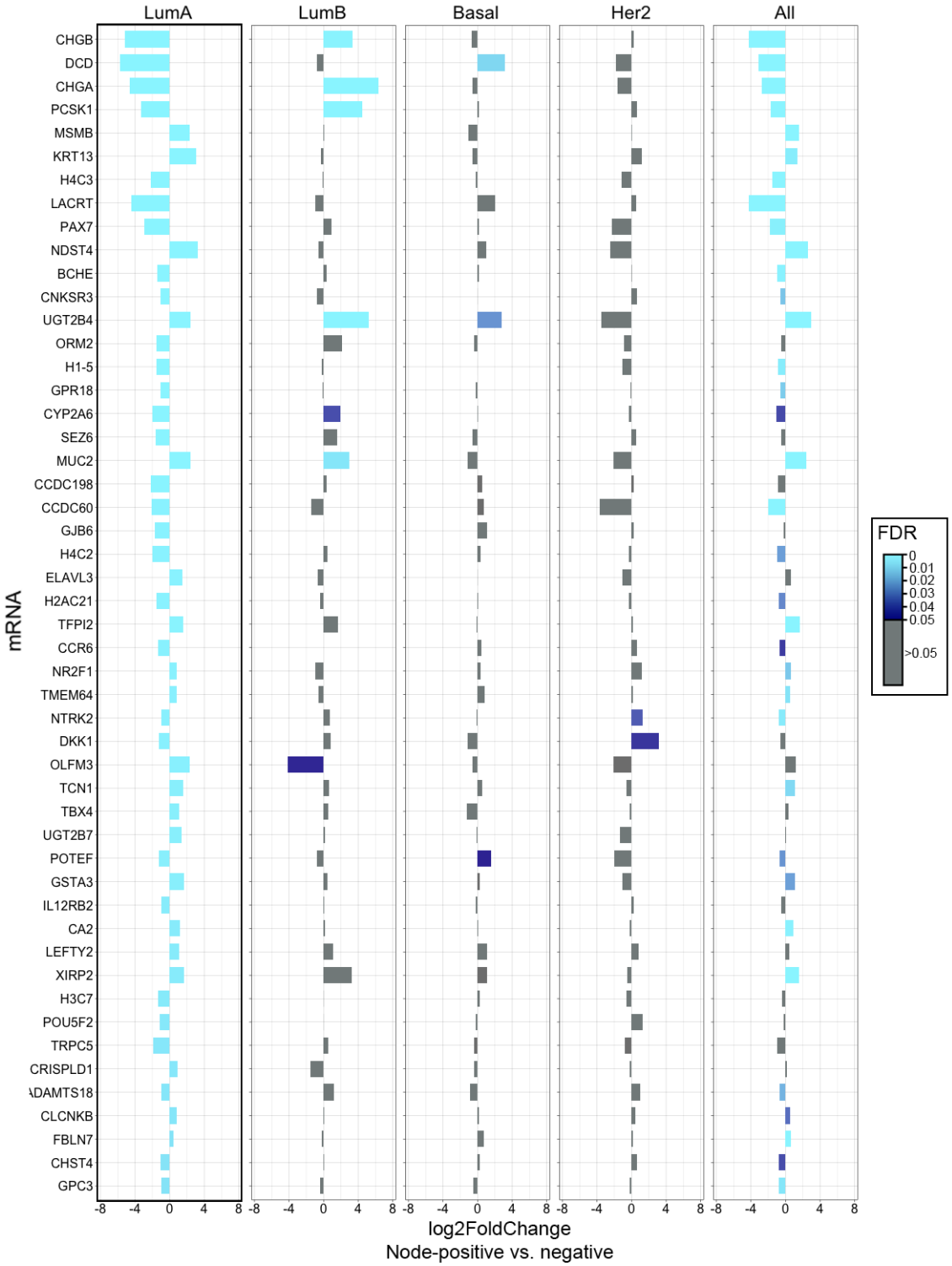


Figure 3.7. Top 50 of 272 statistically significant differentially expressed genes (FDR <0.05) in node-positive compared to node-negative Luminal B early-stage breast cancers.

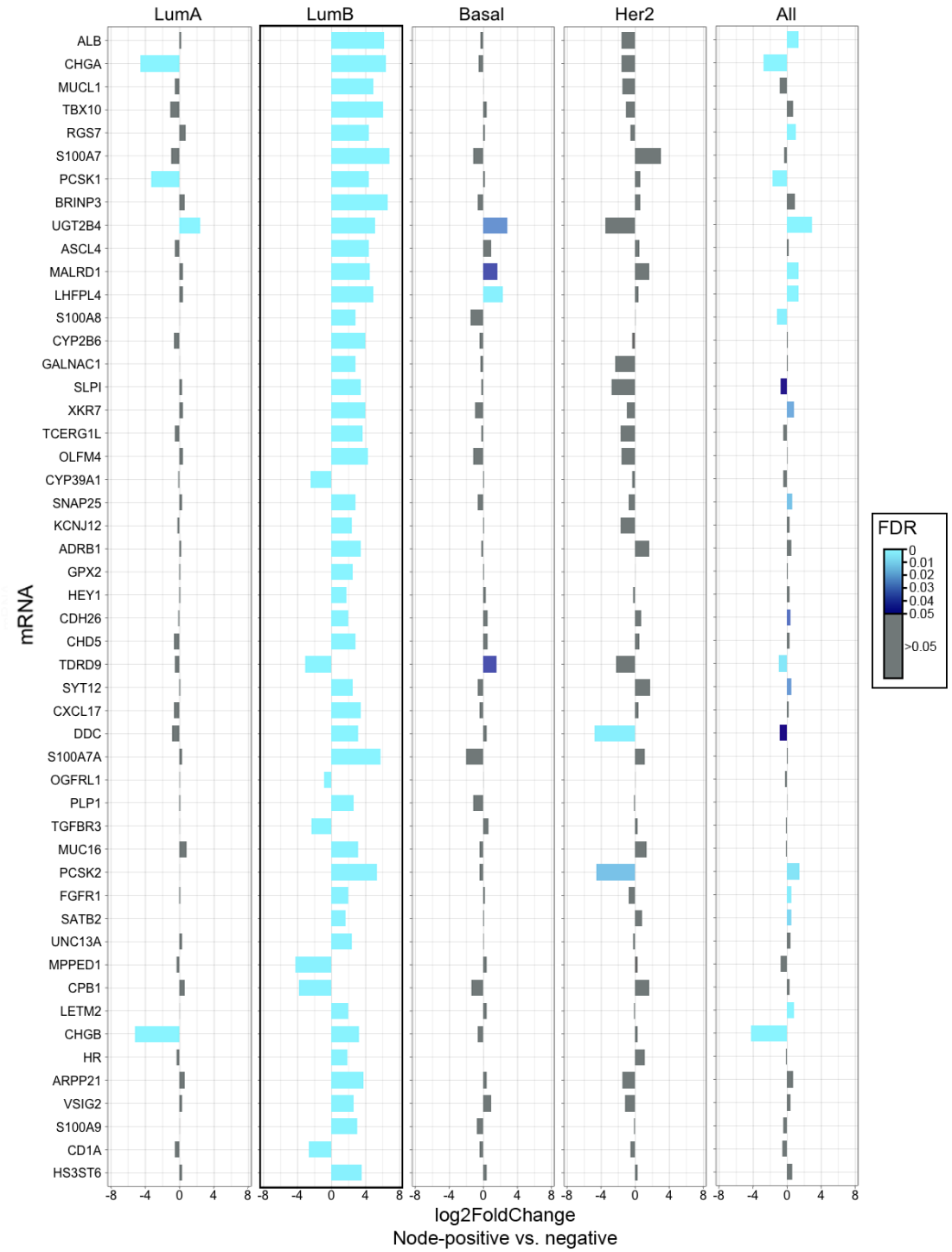


Figure 3.8. Top 50 of 96 statistically significant differentially expressed genes (FDR <0.05) in node-positive compared to node-negative Basal early-stage breast cancers.

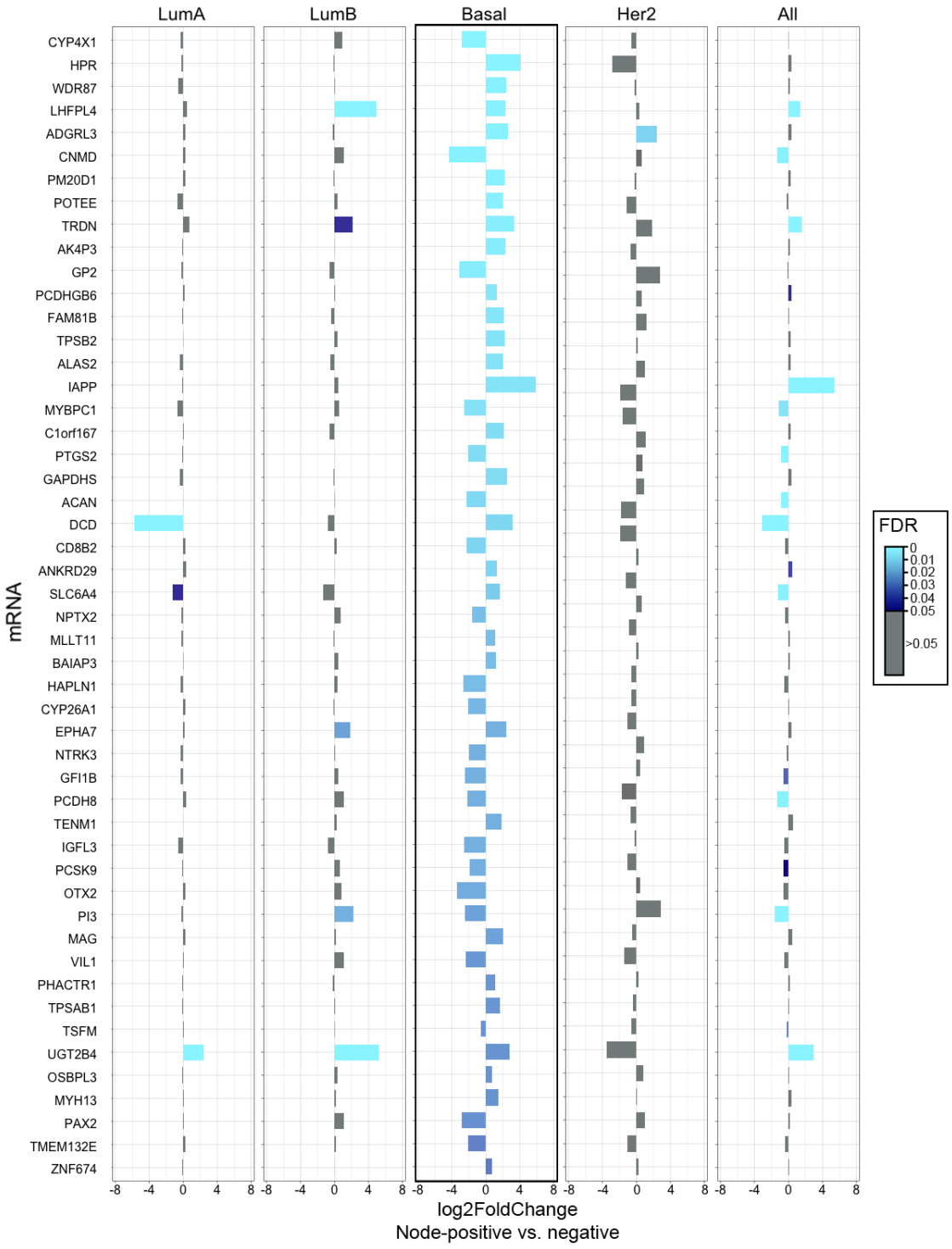
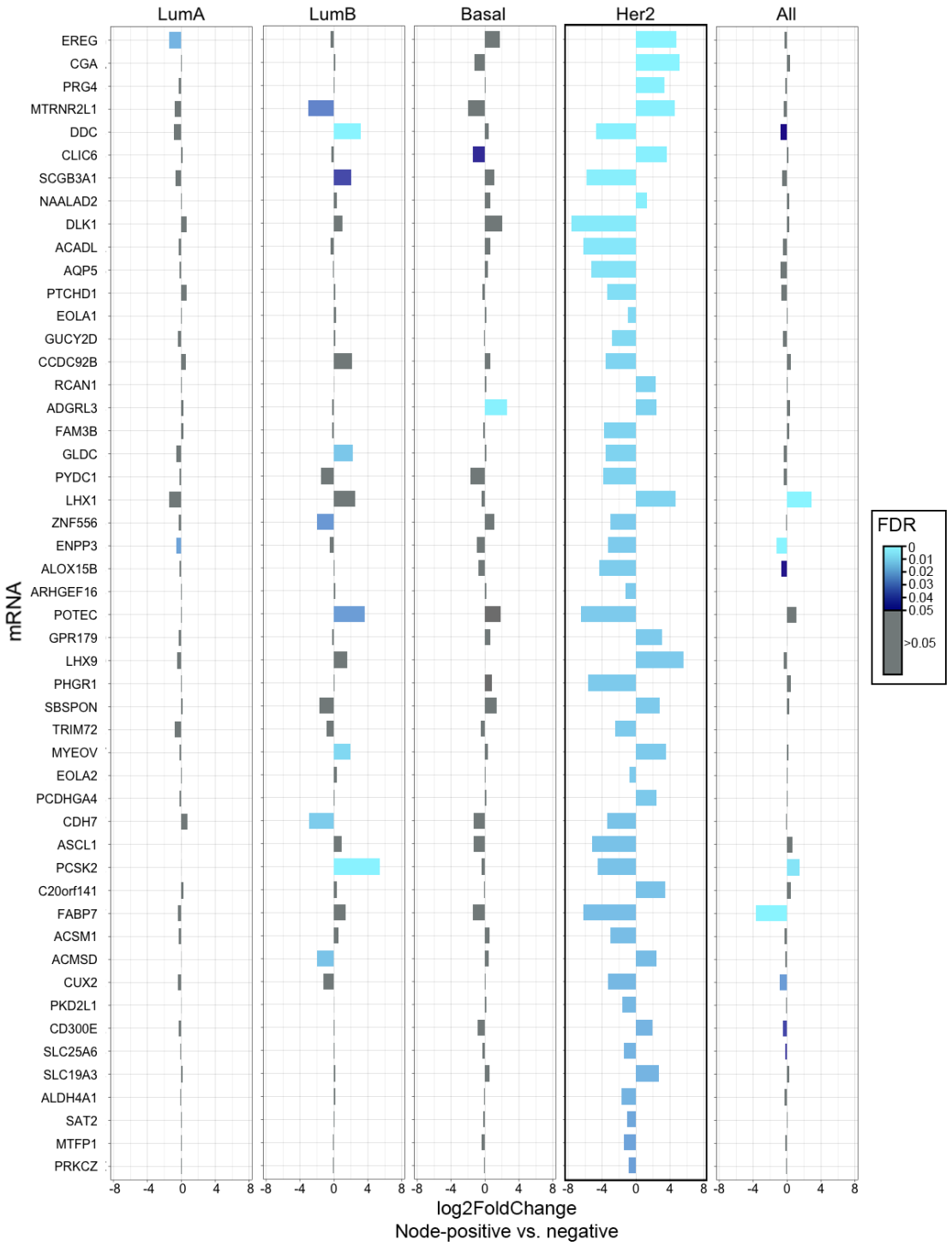


Figure 3.9. Top 50 of the 126 statistically significant differentially expressed genes (FDR<0.05) in node-positive compared to node-negative Her2 early-stage breast cancers.



3.7 TABLES

Table 3.1: Clinical characteristics of early-stage breast cancers divided based on axillary lymph node involvement.

		Nodal status		Comparison
		Negative (n=251)	**Positive (n=163)**	**p Value**
Age	Median (Min-Max)	60 (29-88)	53 (26-87)	**0.002011**[a]
	Unknown	0	0	
Menopause status	Pre	52	51	**0.02597**[b]
	Peri	11	3	
	Post	171	98	
	Indeterminate	17	11	
Histology	Infiltrative Ductal carcinoma	192	127	0.1039[c]
	Infiltrative Lobular carcinoma	44	34	
	Other/Unknown	15	2	
T stage	T1	110	48	**0.0037**[c]
	T2	141	115	
N stage	N0	251	0	
	N1mi	0	25	
	N1	0	106	
	N2	0	22	
	N3	0	10	
M stage	M0	216	133	0.3829[c]
	M1	0	1	
	MX	35	29	
Molecular Subtype	Basal	60	28	0.2325[b]
	Her2	18	9	
	LumA	143	99	
	LumB	30	27	
ER status	Positive	181	132	**0.0444**[c]
	Negative	69	30	
	Unknown	1	1	
PR status	Positive	164	114	0.3339[c]
	Negative	86	48	
	Unknown	1	1	
Her2 status	Positive	29	23	0.4462[c]
	Negative	215	132	
	Unknown	7	8	

[a] Mann-Whitney U test. [b] Pearson's chi-squared test. [c] Fisher's exact test

CHAPTER 4

A new approach to molecular signature processing to improve the predictive performance of axillary lymph node involvement in early-stage breast cancer

CHAPTER 4: A NEW APPROACH TO MOLECULAR SIGNATURE PROCESSING IMPROVES PERFORMANCE IN PREDICTIVE MODELS OF AXILLARY LYMPH-NODE METASTASIS IN EARLY-STAGE BREAST CANCER

4.1 INTRODUCTION

The axillary lymph node (ALN) basin is the most common site of metastasis in breast cancer, and the determination of nodal metastasis is integral to disease staging and prognostication(1,2). Clinical lymphadenopathy is absent in up to 40% of patients with ALN metastasis on pathologic examination(3). As a result, sentinel lymph-node biopsy (SLNB) remains as the standard of care for the staging of axilla in patients presenting with early-stage breast cancer (≤5cm) and absent palpable lymphadenopathy(4,5).

An accurate pre-operative determination of the ALN status that does not entail additional axillary surgery provides several clinical opportunities. Although less morbid than an axillary lymph node dissection (ALND), SLNB still carries risk for patients(6) and requires healthcare resources such as preoperative nuclear medicine radiotracer injection and time in the operating room. Ongoing trials are investigating the value of omitting SLNB in patients with early-stage disease and negative axilla on preoperative assessment(7), and clinical recommendations for SLNB omissions exists in a very select group of older patients with hormone-receptor positive disease(8). Identification of patients at high-risk of axillary involvement also provides an opportunity for downstaging with neoadjuvant systemic therapy(9,10). This can ultimately spare high-risk patients from needing an ALND, as SLNB has a high false negative rate (FNR) when repeated(11).

Molecular predictive models such as Onco*type* DX™ have shown value in the post-operative setting by predicting tumors at risk of recurrence(12), but they have not been well established in the pre-operative stage of care. Previous molecular models of nodal metastasis in breast cancer have also not been specific to the early-stage tumours without clinical lymphadenopathy, which comprises the population that would most benefit from this predictive model(13). Additionally, inherent heterogeneity within various molecular subtypes of breast cancers (14–17) has been noted to be an obstacle in development of molecular signatures of nodal metastasis(18). Accounting for these molecular subtypes may improve predictive model performance.

To address the clinical gap in the non-invasive pre-operative determination of the ALN status in early-stage breast cancer patients without palpable lymphadenopathy, we utilized the datasets from The Cancer Genome Analysis (TCGA) project(17) to develop RNA-based predictive models of axillary lymphadenopathy. We employed several strategies including subtype-specific analyses and additional molecular signature processing to account for inter-subtype molecular heterogeneity with the goal of improving predictive model performance.

4.2 MATERIALS AND METHODS

4.2.1 Data retrieval and processing

TCGA clinical datasets were downloaded from Broad Institute Firehose databases (Version 2016_01_28, firebrowse.org). All tables were merged, removing samples not included in all tables. Stage T1 and T2 samples were selected. Patients who received neoadjuvant chemotherapy were removed. PAM50 molecular subtypes were downloaded from Berger *et al.* (19), and samples with subtype "normal" (normal breast-like) were removed from the dataset.

TCGA data does not provide information regarding clinical stage of the axilla at the time of presentation. To select for patients without clinically palpable axillary lymph nodes, we investigated the axillary staging method for the remaining patients and selected for patients who received sentinel lymph node biopsy +/- axillary dissection. Patients with metaplastic carcinoma or unassigned histologic diagnosis were removed.

RNAseq counts were downloaded from the University of California Santa Cruz Xena platform(20). The available transcripts were filtered for protein-coding genes using the *biomaRT* package(21) in R ("hsapiens_gene_ensembl" dataset utilized). This reduced the total number of transcripts from 60483 to 19556 protein-coding mRNAs. Raw count reads were converted from the $\log_2$(counts+1) scale to integer. Clinical exclusion criteria were applied as described above.

4.2.2 Model development

Training and validation cohorts were created randomly by splitting the TCGA dataset in half in R statistical environment. This resulted in a training cohort of 198 samples, and a

validation cohort of 199 samples. The clinical characteristics of these cohorts are provided in **Supplementary Table 4.1**.

Molecular signatures with the most highly differentially-expressed genes (DEGs) in nodal metastasis were developed using the training cohort and the DESeq2 package (version 1.30)(22) in R statistical environment. RNAseq expression data was then pre-processed for predictive model development through normalization of the integer raw counts with the *variance stabilizing transformation* (vst) algorithm through the DESeq2 package. Normalized counts were then converted to Z-scores.

Cross-validated Elastic NET predictive models of nodal metastasis were then developed in R using the *glmnet* package (version 4.1-3)(23). The α value for the elastic NET model was set at 0.5. Ten-fold cross-validation was used in model development, and a λ value yielding the minimum mean cross-validated error was selected.

Elastic NET predictive models provide the inherent advantage of variable selection. The optimal number of genes to utilize as input to the models to best fit the training data was first tested using molecular signatures with 50, 75 and 100 genes, and performance values were reported (**Supplementary Table 4.2**). The variables selected in each model, along with the λ values and model performance in training cohorts are provided in **Supplementary Table 4.3**.

The performance of the model was then evaluated in the validation cohort using the *assess.glmnet* function in the *glmnet* package, and performance measures including the area under the curve (AUC) of the receiver-operator characteristic (ROC) curve were reported. ROC curves were generated using the *ROCR* package in R (version 1.0-11)(24). The performance was also separately evaluated in each molecular subtype within the validation cohort.

4.2.3 Development of the uniform molecular signature

A “uniform” molecular signature was generated, which included the top 100 differentially-expressed genes that were over- or under-expressed consistently across all 4 molecular subtypes of node-positive breast cancer. To do so, the previously generated differentially expressed genes in the training cohort were filtered for consistent over- or under-expression in node-positive samples across the entire development cohort, and all molecular subtypes groups within, based on the expression Z-score values. 10-fold cross-validated elastic NET predictive models were again generated using this uniform molecular signature, and the performance of the model was evaluated in the validation cohort.

4.2.4 Development of subtype specific molecular signatures

Subtype-specific molecular signatures were developed by applying the DESeq algorithm within each of the molecular subtypes in the training cohort. Elastic-net predictive models were generated using 3-fold cross-validation given the smaller sample size in the subtype-specific training cohort. Performance of these models was evaluated as previously described.

4.2.5 Addition of clinical factors to predictive models

Predictive models were generated as before, with the additional input of age and tumour T-stage from the clinical datasets to the models in development. The combination of clinical factors was tested with the overall signatures of nodal metastasis, the uniform signature, and subtype-specific signatures of nodal metastasis. Performance in the validation cohort was then assessed as before.

4.3 RESULTS

4.3.1 Training and validation cohorts

The TCGA early-stage breast cancer cohort was randomly divided in half, yielding a total of 198 early-stage samples for training, and 199 samples for validating the training models. The training cohort included 88 node-positive and 110 node-negative samples. The validation cohort included 66 node-positive and 133 node-negative tumours. Detailed information regarding the clinical characteristics of the cohorts are provided in **Supplementary Table 4.1**.

4.3.2 Predictive model based on molecular signature of nodal metastasis

Elastic NET regression models of nodal metastasis were generated in the training cohort. The model fit to the training cohort was assessed with an input of 50, 75, and 100 differentially-expressed genes as the molecular signature, and with 10-fold cross-validation. An input of molecular signature with 100 genes was chosen as it provided the least mean squared error in the training cohort (**Supplementary Table 2**).

After feature selection, the resultant predictive model from the subtype-combined molecular signature was based on 39 genes, and achieved an AUC for the ROC curve of 0.909 in the training cohort, and 0.620 in the validation cohort (**Figure 4.1, Supplementary Tables 4.3-4.5**). Except for the Her2 subtype, the model performed poorly in individual molecular subtypes within the validation cohort (**Figure 2**).

4.3.3 Predictive performance of subtype-specific molecular signatures

There is significant molecular heterogeneity between the four molecular subtypes of breast cancer(16,25), which may introduce noise within the training cohort and diminish the

performance of predictive models. In an attempt to improve model performance by separating this inherent complexity from the molecular signature analysis, models were again developed with subtype-specific signatures and 3-fold cross validation in each individual subtype within the training cohort. It is notable that because of subgrouping, these models were trained based on much smaller training sizes compared to the previous model with all subtypes combined, providing less molecular information for the development of both molecular signatures and predictive models (**Supplementary Table 3**). The subtype specific models achieved an AUC of ROC curve of 0.500 in luminal A, 0.579 in luminal B, 0.493 in basal and 0.695 in Her2 subtypes (**Figure 2**). These are not consistently better compared to the previous combined-subtype predictive model.

4.3.4 Predictive performance of a customized molecular signature of nodal metastasis with uniform expression pattern across molecular subtypes

To both utilize the higher power provided by the large sample size of the entire development cohort compared with each individual molecular subtype group, and to reduce the heterogeneity inherent within the molecular subtypes, an extra step of processing was added to the molecular signature. DEGs in node-positive disease were again identified utilizing the entire subtype-combined training cohort, but only genes that were consistently over- or under-expressed in node-positive disease across the combined cohort and each individual subtype were selected to yield a “uniform” molecular signature. Predictive models were then developed with this unform signature (**Supplementary Table 4.3**). These models slightly outperformed the previous combined-subtype signature in the validation cohort (AUC of 0.664

vs 0.660). However, in the two most common subtypes, luminal A and basal, this uniform molecular signature outperformed both the combined signature and subtype-specific signatures (**Figure 4.3**). Most notably, the predictive performance in the basal subtype of early-stage breast cancer exhibits an AUC of 0.870. The performance numbers remain substantially weaker in subtypes with lower sample count, including luminal B and Her2 (**Supplementary Table 4.4**).

4.3.5 Addition of clinical variables to predictive models

We next assessed the value of adding clinical variables of age and T-stage (shown to be significant in previous studies) to predictive models of nodal metastasis in addition to the previously described combined, uniform and subtype-specific molecular signatures (**Supplementary Table 4.4**). The luminal B, Basal and Her2 signatures did not include these clinical variables after the elastic NET algorithm completed variable selection, while the combined-molecular subtype model included only age. The addition of clinical variables slightly increased the predictive performance of the model (AUC of uniform model improved from 0.664 to 0.669 with the inclusion of age and T-stage, **Figure 4.4**). Of note, at a FNR of less than 10%, the performance of the predictive model of nodal metastasis with the uniform molecular signature and clinical characteristics in the basal subgroup showed an accuracy of 71.74%, with a sensitivity of 91.7% and specificity of 64.7% (**Supplementary table 4.6**).

4.4 DISCUSSION

The involvement of ALN nodes is an important prognostic marker in breast cancer, and is the basis of tumour staging(2). Many important clinical decisions such as recommendation for neoadjuvant chemotherapy and adjuvant chemoradiation also rely on the ALN status of the patient(4,5). Non-invasive prediction of ALN metastasis has the potential of offering several advantages in early-stage breast cancer, including the possibility of sparing low-risk patients from staging axillary surgery, and finding high-risk candidates for neoadjuvant systemic therapy. In this study, we utilized the molecular differences in RNA sequencing data from the TCGA datasets to create predictive models of axillary involvement in early-stage breast cancer patients receiving SLNB (**Figure 4.2**). This predictive model of axillary nodal status achieved an AUC for ROC curve of 0.620 in the validation subset of the TCGA dataset.

Although previously reported molecular models did not specifically investigate patients with early-stage breast cancers with no clinical lymphadenopathy, they showed similar performance values in predicting axillary lymph-node status in breast cancer while encompassing a less-selective group of disease stages. Smeets *et al.* developed a 241 gene signature and used cross-validated weighted Least-Squares Support Vector Machines to train the model on 96 breast cancer patients and validate it in an external dataset with an AUC of 0.651(26). In a separate report using data from the Sweden Cancerome Analysis Network Breast initiative, molecular-based gradient boosted machine models showed an AUC of 0.67 in the validation set(13).

The performance range of the aforementioned models is similar to our combined-subtype model, and is comparable to the performance of the currently well-known clinical nomogram of

nodal metastasis which was developed by Memorial Sloan Kattering in 2007(27). This model has been validated in several external early-stage breast cancer cohorts, showing predictive values ranging from an AUC of 0.67-0.73(28–31). Dihge *et al*. compared the performance of molecular-based predictive models of nodal metastasis and clinical-based models (including variables such as tumour size, lymphovascular invasion, age and multifocality) in their cohort(13). Their predictor assessments showed an AUC of 0.71, 0.67 and 0.72 for clinical, molecular, and clinical/molecular combined models in their local validation cohort. Although this suggests a negligible improvement in predictive performance with the addition of molecular variables, the most frequently utilized clinical variable of lymphovascular invasion can only be accurately determined in post-surgical pathology specimens, due to its multifocal nature(32). This limits the utility of lymphovascular invasion as a variable in the pre-operative setting. Due to frequently missing clinical variables in the TCGA dataset, we only assessed the value of well-established clinical variables of age(33–35) and T-stage(30,36) in advancing the predictive model performance. The inclusion of these clinical variables only had a modest impact on our predictive models (**Figure 4.2**). There is emerging support for the application of machine learning to tumour imaging modalities such as magnetic resonant imaging (MRI) to predict axillary lymph-node stage(37–39). A combined clinical, molecular, and radiologic profiling of the tumour may be the next step in improving predictive performance.

As breast cancer encompasses a heterogenous group of disease with previously identified molecular subtypes showing distinct clinical and biologic differences(14–17), this inherent molecular heterogeneity could be exploited to improve predictive model performance. We utilized two strategies to apply molecular subtype information to develop better models. The

first strategy was to create subtype-specific models by training the machine learning algorithms within each subtype of the training dataset individually. These models did not yield any improved performance in the TCGA datasets (**Figure 4.3**). This likely reflects the limitations inherent in smaller sample sizes created by the division of the training cohort for both the development of the molecular signatures and for training the machine learning models. Nakauchi *et al*. specifically targeted the luminal A subset of breast cancer, and trained a model with 388 patients, and achieved superior performance values with an AUC of 0.717 and 0.749 in validation datasets(40). In contrast, the luminal A subset in our data was comprised of only 114 patients. Repeating our analyses in larger cohorts may provide better predictive performances.

The second strategy was to use an additional processing step, which was applied to the DEGs identified in the node-positive disease across all subtypes combined. With the aim of benefiting from both the additional value of a larger training cohort, and accounting for inter-subtype variations, the DEGs were assessed for conformity in over- or under- expression across all molecular subtypes. This "uniform" signature yielded the best performance of our models in the combined-subtype validation cohort with an AUC of 0.664. This novel uniform signature also showed improvements in the two most common molecular subtypes, Luminal A and Basal (**Figure 2**). This predictive model based on the uniform molecular signature and clinical variables showed notably good performance in basal subtype within the validation dataset with an AUC of 0.870. At a FNR of <10%, this model can yield a SLNB reduction rate of 50% based on the formula suggested by Dihge *et al*(41). The basal molecular subtype of breast cancer typically aligns with triple-negative disease, as identified with standard clinical biomarkers with tumours that lack expression of estrogen, progesterone or Her2 receptors on histopathology(42). There

are existing recommendations by the Surgical Society of Oncology in 2016 for omitting routine sentinel lymph-node biopsy in hormone-receptor positive breast cancer patients of age > 70 years(8). Further external validation of the predictive models proposed in this study can aid in the development of similar strategies in basal/triple-negative breast cancer patients by differentiating tumours with a low risk of nodal metastasis.

Our molecular analysis was subject to the technological limitations of bulk tumour transcriptomics, as the significant intratumorally heterogeneity within breast cancer can add noise to molecular signature identification(43). In addition, factors outside of the primary tumour cells, including the peri-tumoral and lymph node microenvironment may play an important role in progression of metastatic breast disease(44,45). The complex biological processes contributing to nodal metastasis have been noted as a major obstacle to the development of molecular signatures that can stratify patients by nodal involvement(18). Technologies such as single-cell analysis and spatial transcriptomics can further provide us with information about tissue structure and intercellular interactions(46,47), and may be the breakthrough necessary to improve the predictive models.

In conclusion, we utilized RNA sequencing molecular information to create predictive models of nodal metastasis specific to early-stage breast cancer patients undergoing sentinel lymph-node biopsy. Our models showed improved predictive performance by accounting for the molecular heterogeneity between the previously established molecular subtypes of breast cancer, with notably enhanced performance within the Basal-subtype of the disease.

4.5 FIGURES

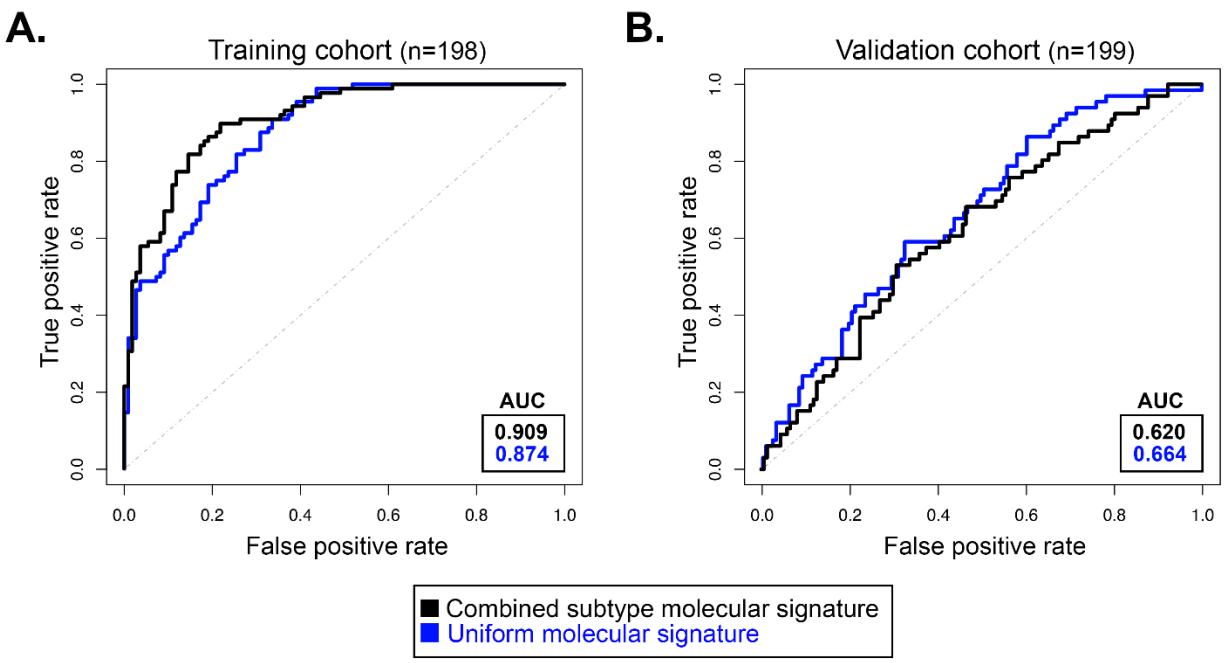


Figure 4.1. Receiver operator characteristic curves showing predictive model performance in A. training cohort and B. validation cohort with molecular signatures generated from all molecular subtypes combined (black), and the "uniform" molecular signature select for consistent over- or under-expression of genes across all node-positive tumour subtypes (blue).

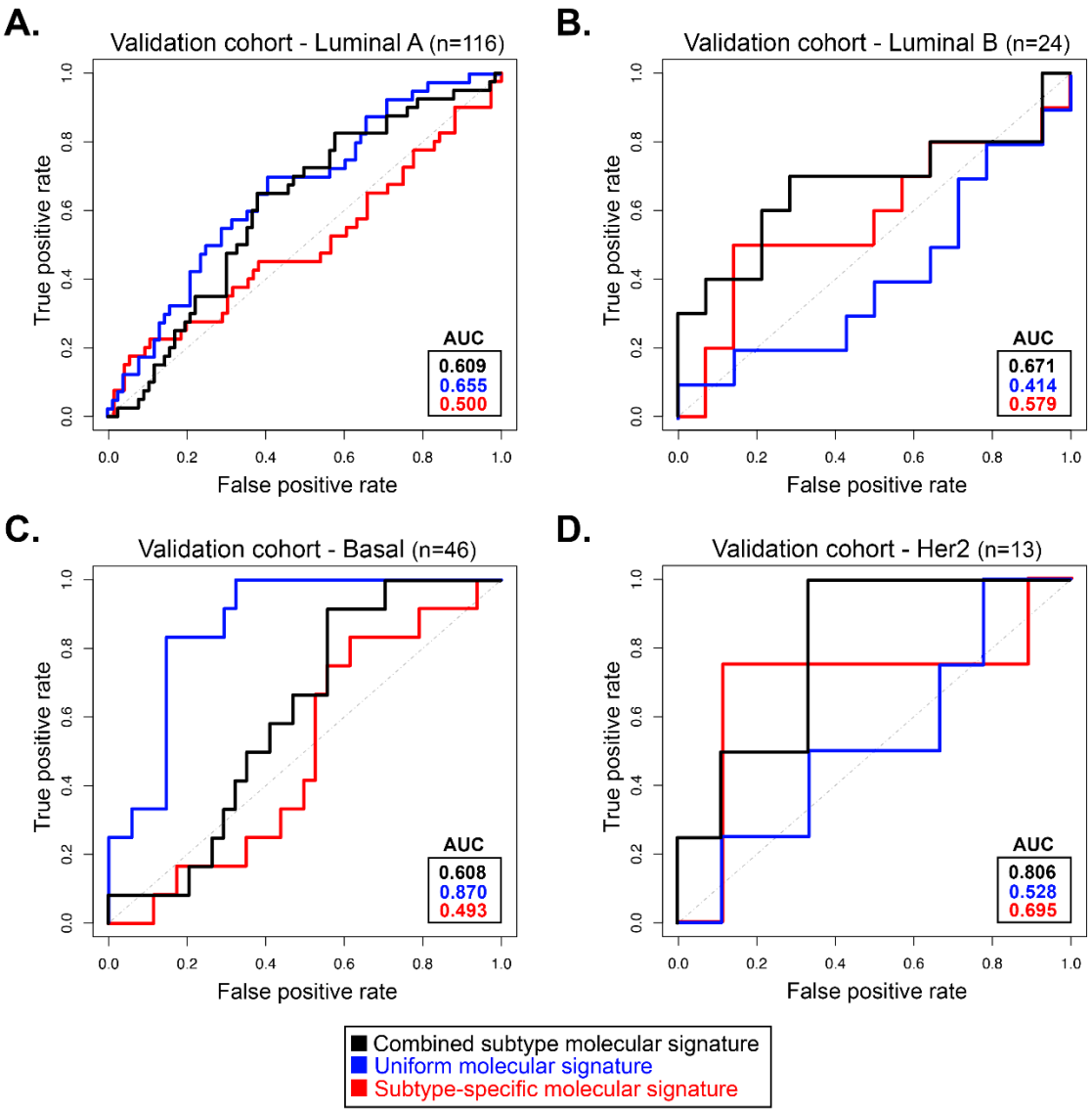


Figure 4.2. Receiver operator characteristic curves showing predictive model performance in A. Luminal A, B. Luminal B, C. Basal, and D. Her2 subtypes within the validation cohort. Models included those generated with the molecular signatures generated from combined-subtypes analysis (black), and the uniform molecular signature (blue), and subtype-specific molecular signatures developed from subtype-specific training cohorts (red). The uniform signature

notably showed improved performance in the two most common molecular subtypes, Luminal A and Basal.

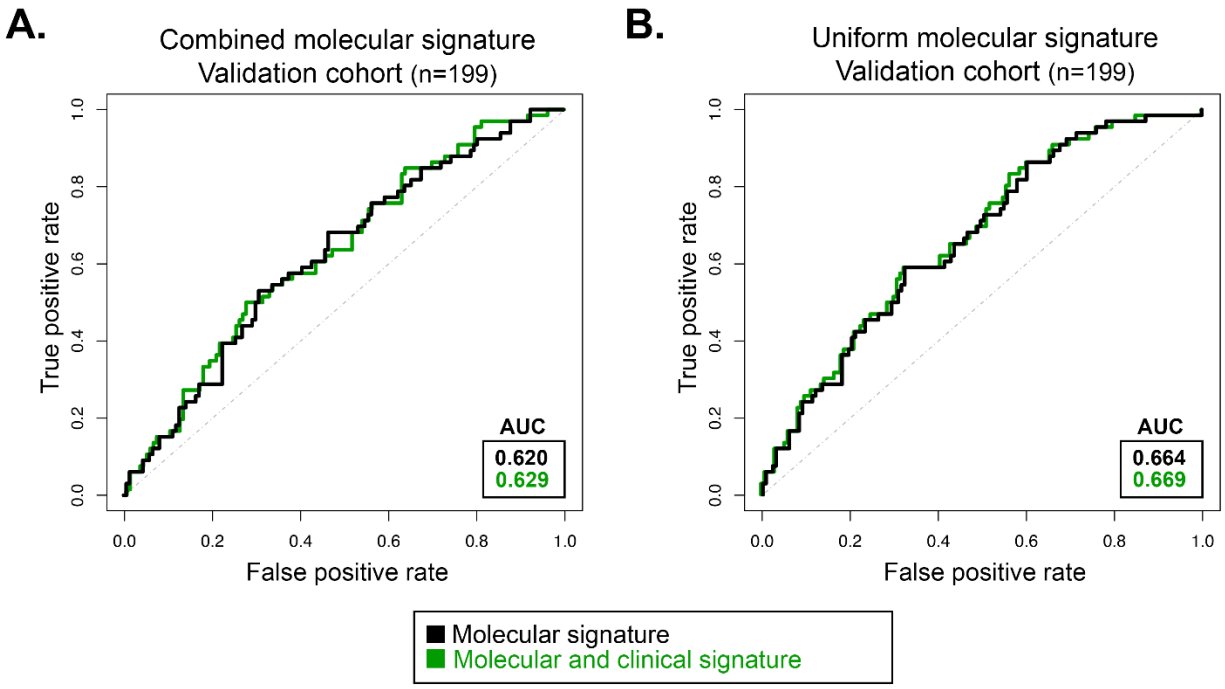


Figure 4.3. Receiver operator characteristic curves showing the performance change with the inclusion of clinical variables of age and T-stage in the training model along with the **A.** combined-subtype molecular signature, and **B.** the uniform molecular signature.

4.6 BIBLIOGRAPHY

1. Abdel-Rahman O. Validation of the 8th AJCC prognostic staging system for breast cancer in a population-based setting. Breast Cancer Res Tr. 2018;168(1):269–75.

2. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. Ca Cancer J Clin. 2017;67(2):93–9.

3. Fisher B, Jeong JH, Anderson S, Bryant J, Fisher ER, Wolmark N. Twenty-Five-Year Follow-up of a Randomized Trial Comparing Radical Mastectomy, Total Mastectomy, and Total Mastectomy Followed by Irradiation. New Engl J Medicine. 2002;347(8):567–75.

4. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30(8):1194–220.

5. National Comprehensive Cancer Network. Breast Cancer (version 2.2022) [Internet]. [cited 2022 Feb 19]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

6. Ashikaga T, Krag DN, Land SR, Julian TB, Anderson SJ, Brown AM, et al. Morbidity results from the NSABP B-32 trial comparing sentinel lymph node dissection versus axillary dissection. J Surg Oncol. 2010;102(2):111–8.

7. Gentilini O, Veronesi U. Abandoning sentinel lymph node biopsy in early breast cancer? A new trial in progress at the European Institute of Oncology of Milan (SOUND: Sentinel node vs Observation after axillary UltraSouND). Breast. 2012;21(5):678–81.

8. ABIM-Foundation. Five Things Physicians and Patients Should Question [Internet]. 2016 [cited 2022 Apr 28]. Available from: https://www.choosingwisely.org/societies/society-of-surgical-oncology/

9. Bear HD, Anderson S, Brown A, Smith R, Mamounas EP, Fisher B, et al. The Effect on Tumor Response of Adding Sequential Preoperative Docetaxel to Preoperative Doxorubicin and Cyclophosphamide: Preliminary Results From National Surgical Adjuvant Breast and Bowel Project Protocol B-27. J Clin Oncol. 2003;21(22):4165–74.

10. Montagna G, Mamtani A, Knezevic A, Brogi E, Barrio AV, Morrow M. Selecting Node-Positive Patients for Axillary Downstaging with Neoadjuvant Chemotherapy. Ann Surg Oncol. 2020;27(11):4515–22.

11. Kuehn T, Bauerfeind I, Fehm T, Fleige B, Hausschild M, Helms G, et al. Sentinel-lymph-node biopsy in patients with breast cancer before and after neoadjuvant chemotherapy (SENTINA): a prospective, multicentre cohort study. Lancet Oncol. 2013;14(7):609–18.

12. Syed YY. Oncotype DX Breast Recurrence Score®: A Review of its Use in Early-Stage Breast Cancer. Mol Diagn Ther. 2020;24(5):621–32.

13. Dihge L, Vallon-Christersson J, Hegardt C, Saal LH, Hakkinen J, Larsson C, et al. Prediction of Lymph Node Metastasis in Breast Cancer by Gene Expression and Clinicopathological Models: Development and Validation within a Population-Based Cohort. Clin Cancer Res [Internet]. 2019;25:6368–81. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31340938

14. Ochoa S, Anda-Jáuregui G de, Hernández-Lemus E. Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes. Frontiers Oncol. 2020;10:845.

15. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast. 2015;24:S26–35.

16. Perou CM, Sørlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52.

17. Atlas NCG. Comprehensive molecular portraits of human breast tumours. Nature [Internet]. 2012;490:61–70. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23000897

18. Shriver CD, Hueman MT, Ellsworth RE. Molecular signatures of lymph node status by intrinsic subtype: gene expression analysis of primary breast tumors from patients with and without metastatic lymph nodes. J Exp Clin Cancer Res Cr. 2014;33(1):116.

19. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018;33(4):690-705.e9.

20. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8.

21. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–91.

22. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

23. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.

24. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940–1.

25. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc National Acad Sci. 2001;98(19):10869–74.

26. Smeets A, Daemen A, Bempt IV, Gevaert O, Claes B, Wildiers H, et al. Prediction of lymph node involvement in breast cancer from primary tumor tissue using gene expression profiling and miRNAs. Breast Cancer Res Tr. 2011;129(3):767–76.

27. Bevilacqua JLB, Kattan MW, Fey JV, III HSC, Borgen PI, Zee KJV. Doctor, What Are My Chances of Having a Positive Sentinel Node? A Validated Nomogram for Risk Estimation. J Clin Oncol. 2007;25(24):3670–9.

28. Qiu P fei, Liu J juan, Wang Y sheng, Yang G ren, Liu Y bing, Sun X, et al. Risk Factors for Sentinel Lymph Node Metastasis and Validation Study of the MSKCC Nomogram in Breast Cancer Patients. Jpn J Clin Oncol. 2012;42(11):1002–7.

29. Parra RFD van la, Francissen CMTP, Peer PGM, Ernst MF, Roos WK de, Zee KJV, et al. Assessment of the Memorial Sloan-Kettering Cancer Center nomogram to predict sentinel lymph node metastases in a Dutch breast cancer population. Eur J Cancer. 2013;49(3):564–71.

30. Chen J ying, Chen J jian, Yang B long, Liu Z bin, Huang X yan, Liu G yu, et al. Predicting sentinel lymph node metastasis in a Chinese breast cancer population: assessment of an existing nomogram and a new predictive nomogram. Breast Cancer Res Tr. 2012;135(3):839–48.

31. Zha H ling, Zong M, Liu X pei, Pan J zhen, Wang H, Gong H yan, et al. Preoperative ultrasound-based radiomics score can improve the accuracy of the Memorial Sloan Kettering Cancer Center nomogram for predicting sentinel lymph node metastasis in breast cancer. Eur J Radiol. 2021;135:109512.

32. Rakha EA, Ellis IO. An overview of assessment of prognostic and predictive factors in breast cancer needle core biopsy specimens. J Clin Pathol. 2007;60(12):1300.

33. Reyal F, Rouzier R, Depont-Hazelzet B, Bollet MA, Pierga JY, Alran S, et al. The Molecular Subtype Classification Is a Determinant of Sentinel Node Positivity in Early Breast Carcinoma. Plos One. 2011;6(5):e20297.

34. Ding J, Jiang L, Wu W. Predictive Value of Clinicopathological Characteristics for Sentinel Lymph Node Metastasis in Early Breast Cancer. Med Sci Monit [Internet]. 2017;23:4102–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28839123

35. Choi EJ, Youk JH, Choi H, Song JS. Dynamic contrast-enhanced and diffusion-weighted MRI of invasive breast cancer for the prediction of sentinel lymph node status. J Magn Reson Imaging. 2020;51(2):615–26.

36. Martin C, Cutuli B, Velten M. Predictive model of axillary lymph node involvement in women with small invasive breast carcinoma. Cancer. 2002;94(2):314–22.

37. Yu Y, Tan Y, Xie C, Hu Q, Ouyang J, Chen Y, et al. Development and Validation of a Preoperative Magnetic Resonance Imaging Radiomics–Based Signature to Predict Axillary Lymph Node Metastasis and Disease-Free Survival in Patients With Early-Stage Breast Cancer. Jama Netw Open. 2020;3(12):e2028086.

38. Liu Y, Luo H, Wang C, Chen X, Wang M, Zhou P, et al. Diagnostic performance of T2-weighted imaging and intravoxel incoherent motion diffusion-weighted MRI for predicting metastatic axillary lymph nodes in T1 and T2 stage breast cancer. Acta Radiol. 2021;028418512110028.

39. Ding J, Chen S, Serrano Sosa M, Cattell R, Lei L, Sun J, et al. Optimizing the Peritumoral Region Size in Radiomics Analysis for Sentinel Lymph Node Status Prediction in Breast Cancer. Acad Radiol. 2020;

40. Nakauchi C, Naoi Y, Shimazu K, Tsunashima R, Nishio M, Maruyama N, et al. Development of a prediction model for lymph node metastasis in luminal A subtype breast cancer: The possibility to omit sentinel lymph node biopsy. Cancer Lett. 2014;353(1):52–8.

41. Dihge L, Ohlsson M, Edén P, Bendahl PO, Rydén L. Artificial neural network models to predict nodal status in clinically node-negative breast cancer. Bmc Cancer. 2019;19(1):610.

42. Bastien RR, Rodríguez-Lescure Á, Ebbert MT, Prat A, Munárriz B, Rowe L, et al. PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. Bmc Med Genomics. 2012;5(1):44–44.

43. Turashvili G, Brogi E. Tumor Heterogeneity in Breast Cancer. Frontiers Medicine. 2017;4:227.

44. Pereira ER, Jones D, Jung K, Padera TP. The lymph node microenvironment and its role in the progression of metastatic cancer. Semin Cell Dev Biol. 2015;38:98–105.

45. Bidard FC, Pierga JY, Vincent-Salomon A, Poupon MF. A “class action” against the microenvironment: do cancer cells cooperate in metastasis? Cancer Metast Rev. 2008;27(1):5–10.

46. Burgess DJ. Spatial transcriptomics coming of age. Nat Rev Genet. 2019;20(6):317–317.

47. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22(10):627–44.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Opportunity to improve care in early-stage breast cancer

Axillary lymph-node (ALN) metastasis in breast cancer is a marker of more advanced disease, and its presence is used to stage the cancer and recommend treatments in current guidelines (1,2). Since clinical examination of lymphadenopathy lacks sensitivity (3), invasive sampling of lymph-nodes remains as the standard of care for most patients with early-stage disease and no clinical lymphadenopathy. A non-invasive mode of predicting the risk of ALN metastasis provides an opportunity to improve care in patients with early-stage breast cancers (tumour size ≤ 5cm, and no palpable lymphadenopathy). Firstly, those patients at low-risk of ALN metastasis may be spared from the operative time and morbidity associated with lymph-node sampling (4). Secondly, neoadjuvant systemic therapy may be considered in patients at high-risk of axillary metastasis to down-stage their axilla (5), leaving them with the option to receive a SLNB to stage their axilla after chemotherapy(6).

5.2 Inconsistencies in literature

Our systematic review revealed a plethora of clinical, pathological, molecular, and radiological variables associated with lymph-node metastasis in the literature for breast cancer patients with early-stage disease. Of note was the inconsistency in the reported significance and strength of association between these variables and ALN involvement (**Figure 2.1**). Although variables such as age, tumour size, tumour location, and hormone-receptor status were frequently included in studies, the lack of uniformity in variable definition and patient selection between these studies was a barrier to conducting a meta-analysis. Advances in meta-analytics through methods such as meta-regressions may enable combining these

heterogeneous datasets, and clarify the relationship between these variables and lymph-node involvement (7).

5.3 Next steps in clarifying the molecular landscape of early-stage breast cancers with nodal metastasis

Noting an absence of predictive molecular models specific to early-stage breast cancer patients, we utilized the TCGA database to identify differences in the molecular landscape of node-positive tumours that can be utilized as a predictive signature. Tumours with axillary metastasis had a lower mutation burden in the TCGA cohort (Figure 3.1). Wang et al. hypothesized a lack of immunogenicity because of reduced neoantigens in tumours with low-mutation burden as a potential explanation for this finding(8). Tumour infiltrating lymphocytes responding to tumour antigens have been proposed as biomarkers for lymph node metastasis in patients, and may be a productive avenue for future studies (9).

Several statistically significant differences in mRNA and miRNA levels were identified between node-positive and node-negative cancers. Importantly, once the analyses were repeated in each of the established intrinsic molecular subtypes of breast cancer (namely Luminal A, Luminal B, Her2 and Basal), only 33% of the identified mRNA expression differences were consistent across all subtypes. This highlighted the intrinsic heterogeneity within breast cancer, supporting existing international recommendations for a subtype-specific approach to early-stage breast cancer(10). The correlation between differences in mRNA expression, miRNA expression and protein quantification in tumours with nodal involvement was beyond the scope of this study, but an important topic to pursue in further studies.

Intertumoral cellular heterogeneity serves as a major limitation to our analyses. TCGA data is comprised of analyses with bulk tumour samples. These surgical specimens include cells of various origins and functions such as immune cells and connective tissue in addition to breast epithelium. Advanced molecular techniques such as single-cell sequencing show promise as the next step in clarifying mechanisms of tumour metastasis despite intertumoral heterogeneity (11). Single-cell sequencing in five breast cancer patients with paired primary tumours and axillary nodes suggested a role for NECTIN2-TIGIT-mediated interaction between cancer cells and tumour microenvironment cells in promoting lymph-node metastasis(12). The lymph-node microenvironment appears to play an important role in the process of metastasis (13). In a breast cancer mouse model, single-cell sequencing of lymph-nodes suggested alterations in the immune and metabolic modulation within the nodes as instrumental to the process of tumour metastasis(14). Exciting developments combining spatial transcriptomics with single-cell analyses show further promise in advancing our knowledge of intercellular interactions within heterogenous tissue (15).

5.4 Predictive models of nodal-metastasis and future directions

After developing algorithms to find molecular differences between node-positive and node-negative early-stage breast cancers, the next step was to utilize these molecular signatures in predictive models of nodal metastasis. To reduce statistical bias, we split the TCGA database in half with computer randomization, into training and validation datasets. Molecular signatures of nodal metastasis were regenerated only based on the training half of the datasets, and machine learning predictive models were developed. Following our previous findings of molecular heterogeneity across the 4 intrinsic molecular subtypes, we trained

models in individual subtypes as well as all subtypes combined, and compared the performance of these models. As a result of sample size reductions, the subtype-specific models did not perform better than the combined-model (**Figure 4.2**). Promising results were seen however with our "uniform" molecular signature models that utilized a new approach to accounting for inter-subtype heterogeneity (**Figures 4.1 and 4.2**).

The uniform signature showed the best performance in the Basal subtype of tumours within the validation cohort. The uniform model which relied on a 31 gene signature, tumour T-stage and age, achieved an AUC of 0.875 in the ROC curve. At a commonly accepted false-negative rate threshold of 10%(16), this model would achieve a 50% SLNB reduction rate (model performance measures provided in **Supplementary Table 4.6**).

The predictive models and molecular signatures identified in our studies will need to be validated in external datasets before clinical application, but if replicated, these predictive models show promise in reducing SLNB procedures. This can have a major impact in reducing associated patient morbidity and healthcare resource use. Finally, to highlight an advantage of machine learning algorithms, the models can be retrained and applied to additional cohorts(17–20). This can possibly reduce model over-fitting, and improve predictive performance with time.

5.5 Bibliography

1. National Comprehensive Cancer Network. Breast Cancer (version 2.2022) [Internet]. [cited 2022 Feb 19]. Available from: https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

2. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30(8):1194–220.

3. Fisher B, Jeong JH, Anderson S, Bryant J, Fisher ER, Wolmark N. Twenty-Five-Year Follow-up of a Randomized Trial Comparing Radical Mastectomy, Total Mastectomy, and Total Mastectomy Followed by Irradiation. New Engl J Medicine. 2002;347(8):567–75.

4. Langer I, Guller U, Berclaz G, Koechli OR, Schaer G, Fehr MK, et al. Morbidity of Sentinel Lymph Node Biopsy (SLN) Alone Versus SLN and Completion Axillary Lymph Node Dissection After Breast Cancer Surgery. Ann Surg. 2007;245(3):452–61.

5. Stafford A, Williams A, Edmiston K, Cocilovo C, Cohen R, Bruce S, et al. Axillary Response in Patients Undergoing Neoadjuvant Endocrine Treatment for Node-Positive Breast Cancer: Systematic Literature Review and NCDB Analysis. Ann Surg Oncol. 2020;27(12):4669–77.

6. Kuehn T, Bauerfeind I, Fehm T, Fleige B, Hausschild M, Helms G, et al. Sentinel-lymph-node biopsy in patients with breast cancer before and after neoadjuvant chemotherapy (SENTINA): a prospective, multicentre cohort study. Lancet Oncol. 2013;14(7):609–18.

7. Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI. Understanding heterogeneity in meta-analysis: the role of meta-regression. Int J Clin Pract. 2009;63(10):1426–34.

8. Wang Z, Liu W, Chen C, Yang X, Luo Y, Zhang B. Low mutation and neoantigen burden and fewer effector tumor infiltrating lymphocytes correlate with breast cancer metastasization to lymph nodes. Sci Rep-uk. 2019;9(1):253.

9. Caziuc A, Schlanger D, Amarinei G, Dindelegan GC. Can Tumor-Infiltrating Lymphocytes (TILs) Be a Predictive Factor for Lymph Nodes Status in Both Early Stage and Locally Advanced Breast Cancer? J Clin Medicine. 2019;8(4):545.

10. Gnant M, Harbeck N, Thomssen C. St. Gallen 2011: Summary of the Consensus Discussion. Breast Care. 2011;6(2):136–41.

11. Han Y, Wang D, Peng L, Huang T, He X, Wang J, et al. Single-cell sequencing: a promising approach for uncovering the mechanisms of tumor metastasis. J Hematol Oncol. 2022;15(1):59.

12. Xu K, Wang R, Xie H, Hu L, Wang C, Xu J, et al. Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. Oncogenesis. 2021;10(10):66.

13. Pereira ER, Jones D, Jung K, Padera TP. The lymph node microenvironment and its role in the progression of metastatic cancer. Semin Cell Dev Biol. 2015;38:98–105.

14. Li YL, Chen CH, Chen JY, Lai YS, Wang SC, Jiang SS, et al. Single-cell analysis reveals immune modulation and metabolic switch in tumor-draining lymph nodes. Oncoimmunology. 2020;9(1):1830513.

15. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22(10):627–44.
16. Boughey JC, Suman VJ, Mittendorf EA, Ahrendt GM, Wilke LG, Taback B, et al. Sentinel Lymph Node Surgery After Neoadjuvant Chemotherapy in Patients With Node-Positive Breast Cancer: The ACOSOG Z1071 (Alliance) Clinical Trial. Jama. 2013;310(14):1455–61.

17. Treboux J, Ingold R, Genoud D. Towards Retraining of Machine Learning Algorithms: An Efficiency Analysis Applied to Smart Agriculture. 2020 Global Internet Things Summit Giots. 2020;00:1–6.

18. Barque M, Martin S, Vianin JEN, Genoud D, Wannier D. Improving wind power prediction with retraining machine learning algorithms. 2018 Int Work Big Data Information Secur Iwbis. 2018;00:43–8.

19. Kitamura G, Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. Clin Imag. 2020;61:15–9.

20. Park JG, Jun HB, Heo TY. Retraining prior state performances of anaerobic digestion improves prediction accuracy of methane yield in various machine learning models. Appl Energ. 2021;298:117250.

Curriculum Vitae

Education

General Surgery 2019-present

Schulich School of Medicine and Dentistry, Western University – London, ON

Master of Science in Surgery (MSc)

2021-present

Schulich School of Medicine and Dentistry, Western University – London, ON

Supervisor: Dr. Muriel Brackstone

Doctor of Medicine (MD) 2015-2019

Schulich School of Medicine and Dentistry, Western University – London, ON

Honours Bachelor of Science in Biology (BSc. H.) 2011-2015

York University – Toronto, ON

Publications

Ghasemi F, Tessier TM, Gameiro SF, Maciver AH, Cecchini MJ, Mymryk J.S. 2020. High MHC-II expression in Epstein–Barr virus-associated gastric cancers suggests that tumor cells serve an important role in antigen presentation. *Scientific reports*. 10(1):1-6.

Ghasemi, F., Gameiro, S.F., Tessier, T.M., Maciver, A.H. and Mymryk, J.S. 2020. High Levels of Class I Major Histocompatibility Complex mRNA Are Present in Epstein–Barr Virus-Associated Gastric Adenocarcinomas. *Cells*, 9(2), 499-503.

Ghasemi, F., Prokpec, S. D., MacNeil, D., Howlett, C., Stecho, W., Platinga, P., …, Boutros, P.C., Nichols, A.C. 2019. Mutational analysis of Head and Neck Squamous Cell Carcinoma stratified by smoking status identifies NSD1 as a biomarker of improved survival. *JCI Insight*, 10;4(1).

Ghasemi, F., Black, M., Sun, R., Vizeacoumar, F., Pinto, N., Ruicci, K., Yoo, J., ..., Boutros, P., Nichols, A.C. 2018. High-throughput testing in head and neck squamous cell carcinoma identifies agents with preferential activity in human papillomavirus-positive or negative cell lines. *Oncotarget*. 9(40), 26064-26071.

Ghasemi, F., Black, M., Vizeacoumar, F., Pinto, N., Ruicci, K. M., Son, H. L. C., MacNeil, D., Lowerison, M.R.., ..., Barrett, J.W., & Nichols, A.C. (2017). Repurposing Albendazole: new potential as a chemotherapeutic agent with preferential activity against HPV-negative head and neck squamous cell cancer. *Oncotarget*. 8(42): 71512–71519.

Ghasemi, F., Anooshirvani, N., Sibbald, R. G., & Alavi, A. (2016). The Point Prevalence of Malignancy in a Wound Clinic. *The International Journal of Lower Extremity Wounds*, 15(1), 58-62.

Ghasemi, F., Wegman, D. W., Kanoatov, M., Yang, B. B., Liu, S. K., Yousef, G. M., & Krylov, S. N. (2013). Improvements to direct quantitative analysis of multiple microRNAs facilitating faster analysis. *Analytical chemistry*, 85(21), 10062-10066.

Gameiro, S. F., **Ghasemi, F.**, Zeng, P. Y., Mundi, N., Howlett, C. J., Plantinga, P., ... & Mymryk, J. S. (2021). Low expression of NSD1, NSD2, and NSD3 define a subset of human papillomavirus-positive oral squamous carcinomas with unfavorable prognosis. *Infectious Agents and Cancer*, *16*(1), 1-10.

Kim, H. A. J., Zeng, P. Y., Shaikh, M. H., Mundi, N., **Ghasemi, F.**, Di Gravio, E., ... & Nichols, A. C. (2021). All HPV-negative head and neck cancers are not the same: Analysis of the TCGA dataset reveals that anatomical sites have distinct mutation, transcriptome, hypoxia, and tumor microenvironment profiles. *Oral Oncology*, *116*, 105260.

Sorgini, A., Kim, H. A. J., Zeng, P. Y., Shaikh, M. H., Mundi, N., **Ghasemi, F.**, ... & Nichols, A. C. (2021). Analysis of the TCGA Dataset Reveals that Subsites of Laryngeal Squamous Cell Carcinoma Are Molecularly Distinct. *Cancers*, *13*(1), 105.

Pinto, N., Prokopec, S.D., **Ghasemi, F**., Meens, J., Ruicci, K.M., Khan, I.M., Mundi, N., Patel, K., Han, M.W., Yoo, J. and Fung, K., 2020. Flavopiridol causes cell cycle inhibition and demonstrates anti-cancer activity in anaplastic thyroid cancer models. *PloS one*, 15(9), p.e0239315.

Black, M., **Ghasemi, F**., Sun, R.X., Stecho, W., Datti, A., Meens, J., Pinto, N., Ruicci, K.M., Khan, M.I., Han, M.W. and Shaikh, M., 2020. Spleen tyrosine kinase expression is correlated with human papillomavirus in head and neck cancer. Oral Oncology, 101, p.104529.

Mundi, N., **Ghasemi, F.**, Zeng, P.Y., Prokopec, S.D., Patel, K., Kim, H.A.J., Di Gravio, E., MacNeil, D., Khan, M.I., Han, M.W. and Shaikh, M., 2020. Sex disparities in head & neck cancer driver genes: An analysis of the TCGA dataset. Oral Oncology, 104, p.104614.

Gameiro, S., **Ghasemi, F.**, Barrett, J., Koropatnick , J., Nichols, A.C., Mymryk, J., Maleki Vareki, S. (2018). Treatment-naïve HPV+ head and neck cancers display a T-cell-inflamed phenotype distinct from their HPV- counterparts that has implications for immunotherapy. *OncoImmunology*. e1498439, 1-14.

Gameiro, S.F., Zhang, A., **Ghasemi, F.**, Barrett, J.W., Nichols, A.C. and Mymryk, J.S., 2017. Analysis of Class I Major Histocompatibility Complex Gene Transcription in Human Tumors Caused by Human Papillomavirus Infection. *Viruses*, *9*(9), p.252.

Gameiro, S.F., **Ghasemi, F.**, Barrett, J.W., Nichols, A.C. and Mymryk, J.S., 2019. High Level Expression of MHC-II in HPV+ Head and Neck Cancers Suggests that Tumor Epithelial Cells Serve an Important Role as Accessory Antigen Presenting Cells. Cancers, 11(8), p.1129.

Mundi, N., Prokopec, S.D., **Ghasemi, F**., Warner, A., Patel, K., MacNeil, D., Howlett, C., Stecho, W., Plantinga, P., Pinto, N. and Ruicci, K.M., 2019. Genomic and human

papillomavirus profiling of an oral cancer cohort identifies TP53 as a predictor of overall survival. Cancers of the head & neck, 4(1), pp.1-8.

Prusinkiewicz, M.A., Gameiro, S.F., **Ghasemi, F.**, Dodge, M.J., Zeng, P.Y., Maekebay, H., Barrett, J.W., Nichols, A.C. and Mymryk, J.S., 2020. Survival-Associated Metabolic Genes in Human Papillomavirus-Positive Head and Neck Cancers. Cancers, 12(1), p.253.

Wegman, D. W., **Ghasemi, F.,** Khorshidi, A., Yang, B. B., Liu, S. K., Yousef, G. M., & Krylov, S. N. (2014). Highly-sensitive amplification-free analysis of multiple miRNAs by capillary electrophoresis. *Analytical chemistry*, 87(2), 1404-1410.

Wegman, D. W., **Ghasemi, F.**, Stasheuski, A. S., Khorshidi, A., Yang, B. B., Liu, S. K., ... & Krylov, S. N. (2016). Achieving single-nucleotide specificity in direct quantitative analysis of multiple MicroRNAs (DQAMmiR). *Analytical chemistry*, 88(4), 2472-2477.

Book Chapters

Gameiro, S.F., **Ghasemi, F**., Barrett, J.W., Koropatnick, J., Nichols, A.C., Mymryk, J.S. and Vareki, S.M., 2020. DIY: Visualizing the immune landscape of tumors using transcriptome and methylome data. In Methods in Enzymology (Vol. 636, pp. 49-76). Academic Press.

Awards

MSc in Surgery Colloquium Award 2022

Department of General Surgery – Senior resident research award 2022

Canadian Graduate Scholarship – Masters (CGS-M) Program 2021

Canadian Institutes of Health Research (CIHR)

Ontario Graduate Scholarship – Offer declined due to conflict with CGS-M 2021

Walker Award – Department of Surgery – London, ON 2020

Given for the best research presentation at the annual research day

Certificate of Merit from Department of Surgery– London, ON 2018
Given for highest achievement during the surgery clerkship block.

Dr. L. DeWitt Wilcox Award in Research - London, ON 2017
Given to a student who has demonstrated initiative, drive, perseverance and awareness of research.

First Place Poster Award, Canadian Society of Otolaryngology 2017
Given annually to the top-ranked poster presentation at the conference.

Summer Research Training Program (SRTP) Award – London, ON 2016-2017
Provided funding for selected medical students to conduct full-time research over two summers.

Undergraduate Research Awards from the Natural Sciences and Engineering Research Council of Canada 2013-2014
Given based on academic achievement to provide funding for full-time research at York University.

Leadership

Student Representative, CBME Committee 2018
Represented the student body to recommend innovations needed to implement Competency-Based Medical Education in the undergraduate medicine curriculum.

Executive, Surgically Oriented Anatomy Prosectors Club 2016
Coordinated hands-on sessions with faculty from various surgical departments to demonstrate surgical procedures for medical students in the anatomy lab.

Executive, Oncology Interest Group 2016
Arranged sessions with staff and residents in Medical Oncology and Radiation Oncology to promote discussion about careers in oncology.

Student Representative, Summer Research Training Program (SRTP) 2016
Represented students on a committee of faculty and clinician-scientists to optimize student learning and promote achievement in research.

Presentations

Virtual presentations

Golden Scalpel presentation, C-CASE 2021
One resident per university nominated by department of surgery to present.

Research Day, Department of Surgery 2020, 2022

Research Day, Department of General Surgery 2020, 2021, 2022

Podium Presentations

Student Research Rounds, Department of Otolaryngology – London, ON 2018

Oncology Grand Rounds Young Cancer Researcher Showcase – London, ON 2017

Head and Neck Cancer Disease Site Retreat – London ON 2017

Poster Presentations

Canadian Society of Otolaryngology – Quebec City, QC 2018

Canadian Society of Otolaryngology – Saskatoon, SK 2017

London Health Research Day – London, ON 2017

Community Involvements

Fundraiser at Princess Margaret's Ride to Conquer Cancer event – Toronto, ON 2013
Recognizing the importance of funding in cancer research, I started and led a team that raised over $6000 for this event. This fundraiser included a 200km bike ride from Toronto to Niagara Falls.

Self-initiated food drive for the Daily Bread Food Bank - Toronto, ON 2014
Collected over 600 lbs of food by going door-to-door with the idea that many are able to donate but cannot make the trip to the food banks.

Volunteer at the GI Clinic, Princess Margaret Hospital –Toronto, ON 2012-2015
Participated in a variety of roles ranging from connecting patients with different support services to creating electronic resources for use by the GI clinic.