UNIVERSITY OF CALIFORNIA SAN DIEGO

Profiling archival samples to examine the molecular underpinnings of early carcinogenesis

A Dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Daniela Nachmanson

Committee in charge:

Professor Olivier Harismendy, Chair
Professor Ludmil Alexandrov, Co-Chair
Professor Hannah Carter
Professor Silvio Gutkind
Professor Melissa Gymrek

2022

The Dissertation of Daniela Nachmanson is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego
2022

DEDICATION

This dissertation is dedicated to my parents, Lev and Elena Nachmanson. A mathematician and a musician who together found a delightful balance between logic and feeling. After migrating twice, they fostered a creative, curious, and loving home for me and my brothers in the U.S. Without them, this work would not exist.

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

Supplemental_Tables

Table S1.1. Description of the specimen used in the study.

Table S1.2. Adapters and primers used in targeted sequencing library preparation methods.

Table S1.3. Summary statistics of the sequencing data.

Table S1.4. Copy number segments and overall burden for test specimen.

Table S1.5. Default variant calling filters for raw ensemble variant calls.

Table S1.6. Benchmarking variant calls from the test specimen. The number of raw ensemble variants, after germline filtering, and the effect of the FFPE filtering are reported.

Table S1.7. Summary of high confidence somatic mutations and probabilistic evaluation of their distribution across regions.

Table S2.1. Subject and sample Information.

Table S2.2a. Sample RNA-seq statistics and PAM50 probabilities.

Table S2.2b. Region RNA-seq statistics and PAM50 probabilities.

Table S2.3. Summary of somatic genetic analysis.

Table S2.4. Chromosome arm-level CNA.

Table S2.5. Mutational signatures identified in DCIS somatic mutations.

Table S2.6. Non-silent mutations in breast cancer driver genes.

Table S2.7a. Cell densities according to lymphocytes type, state, and histological compartment.

Table S2.7b. Median T-cell, B-cell, and T-regs densities in each immune-state and histological compartment.

Table S2.7c. Key summary metrics and resulting immune-state for all DCIS regions.

Table S2.8. Scores of MHC1 immunostaining in DCIS and adjacent normal regions.

Table S2.9. Exome sequencing statistics.

Table S2.10. DCIS and breast cancer gene list.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DCIS	Ductal carcinoma in situ
IBC	Invasive breast carcinoma
FFPE	Formalin-fixed paraffin-embedded
PML	Premalignant lesion
LCM	Laser-capture microdissection
VAF	Variant allele fraction
CNA	Copy number alteration
SNP	Single nucleotide polymorphism
Indel	Insertion or deletion
MAF	Minor allele frequency
PRS	Polygenic risk score

ACKNOWLEDGEMENTS

First and foremost I would like to sincerely thank my mentor and chair, Professor Olivier Harismendy. Olivier was the first professor at UCSD with whom I had an interview, the first rotation lab I joined, and continued to be an integral part of my time here as I joined the lab. Throughout my time in the Oncogenomics lab he ran, he helped me grow to be a stronger scientist by helping me hone creative methodological problem-solving skills, persistence in research problems, but also to respect the data. He encouraged my attendance and participation in dozens of conferences, meetings, journal clubs, poster sessions, and talks, helping me both network, become engaged with the scientific community, and improve my scientific communication skills. He has been an absolutely fantastic mentor and teacher, even throughout a global pandemic when motivation and progress were constantly challenged for everyone, he provided consistency, accountability, and caring check-ins. In general, Olivier has had a strong reputation for being an engaged mentor and advocate among the student body, who always has insightful advice, feedback, or perspectives to offer, almost regardless of the topic. UCSD, the Bioinformatics and Systems Biology Program, and I were very lucky to have him and I am enormously grateful to him for helping me become the scientist I am today.

Next, I would also like to thank my excellent committee members. Ludmil Alexandrov provided consistent support, engaged and critically important feedback on all my research projects, and overall excellent mentorship, his expertise in somatic mutation accumulation and their patterns are hard to parallel and I feel tremendously fortunate to have worked with him during my time at UCSD. I would also like to thank Professor Melissa Gymrek, even though my research had only a sprinkle of population genetics, she provided me with organized and clear feedback, insightful questions, and encouragement which has helped to build my scientific confidence, she continues

to be a scientific role model to me. In the past year, I have fortunately worked more closely with Professor Hannah Carter and am very grateful for her genomics advice and her inspired suggested research directions, she has been remarkably helpful and generous with lab resources for some collaborative projects and I thank her for her advice and support. Last but not least, thank you to Professor Silvio Gutkind, he gave me the opportunity to work on several incredibly interesting projects regarding oral premalignancy and mTOR inhibition in collaboration with his lab, he is a continuous source of research inspiration and motivation and I am very grateful to have had him as both a collaborator and committee member.

I would like to thank all the members of the Oncogenomics lab that I had the pleasure to work with, in particular, Adam Officer, Joseph Steward, Max Xu, and Young Choi. Your scientific (and non-scientific) advice, encouragement, and friendship made my time at UCSD much more enjoyable. In addition, I would like to thank the past and present leaders of the Cancer Genomics Journal club that I believe I participated and presented in approximately 14 consecutive academic quarters. In particular, Professors Olivier Harismendy, Hannah Carter, Jill Mesirov, and, Kathleen Curtius. This journal club helped me to stay up to date with new research and technologies in cancer genomics and also provided an opportunity to get research insights and hear critical feedback on recent research from several leaders in the field.

I would also like to thank several collaborators. I thank the breast group of the Molecular Characterization of Screen Detected Lesions (MCL) consortium, in particular, Gill Hirst and Sandy Borowsky, for samples, research feedback, and opportunities to present my research several times to larger audiences. For some work I did not get the opportunity to feature here, I would also like to thank members of the Gutkind lab, in particular, Mike Allevato, Nadia Arang, and Sam Wu for giving me the opportunity to participate in several exciting cancer-related research projects.

Additionally, to my early mentors in Seattle, at the University of Washington and TwinStrand Biosciences, some of whom I recently got to work with on a carcinogenesis detection project not featured here, in particular Professors Rosana Risques and Scott Kennedy, Jesse Salk, and Clint Valentine, thank you for your long term scientific and mentorship scientific support who early on inspired my interest in early cancer evolution.

Of course, I would like to thank my wonderful family - including my parents, Lev and Elena Nachmanson, my three brothers: Michael, David, and Ben, and sister-in-law Laura. I am forever grateful to my babushka Nina, who insistently supported me through the first 4 years of my Ph.D. via Skype calls from Russia, she passed away last year but her legacy will be continued through her family, and my newborn niece, Nina, named in her honor. I would also be remiss not to mention the fantastic friend support network I have had to help me. Particularly, my childhood friend Carmen Lopez, thank you for the million walks and wine nights. All my UCSD-associated pals who kept me sane and my life fun, especially Holly Sullivan, Meghana Pagadala, Nadia Arang, Lindsey Bergh, Sara Mirza, Ariel Wang, Miranda Diaz, Jeff Granados, and Michael Wiest. To Sebastian Cifuentes, my fantastic partner, who has had amazing patience for my research-related frustration and always helped me celebrate each little win on the way, you made San Diego feel like home. He also taught me that I love dogs, particularly Archie and Dora, who I also have to thank for forcing me outside for walks. Lastly, my undergraduate friends who never failed to support me from afar, Maddi Miller, Lydia Yekalam, Yasameen Fardi, Katie Kooi, Kaila Turner, Molly Deutsch, Hannah Hainline, and Lauren Ratto, love you all so much.

Chapter 1, is a reformatted reprint of the materials in two different works. The first is adapted from as it appears in "Mutational profiling of micro-dissected pre-malignant lesions from archived specimens" in *BMC Medical Genomics*, 2020 by Daniela Nachmanson, Joseph Steward,

Huazhen Yao, Adam Officer, Eliza Jeong, Thomas J. O'Keefe, Farnaz Hasteh, Kristen Jepsen, Gillian L. Hirst, Laura J. Esserman, Alexander D. Borowsky and Olivier Harismendy. The second is adapted from a manuscript in submission titled, "PROCEED - Predicting exome sequencing depth of suboptimal DNA samples", by Daniela Nachmanson, Kristen Jepsen, Huazhen Yao, Ben Nachmanson, and Olivier Harismnedy. The dissertation author was the primary investigator and author of both papers.

Chapter 2, in full, is a reformatted presentation of the material as it appears as "The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ." by Daniela Nachmanson, Adam Officer, Hidetoshi Mori, Jonathan Gordon, Mark F. Evans, Joseph Steward, Huazhen Yao, Thomas O'Keefe, Farnaz Hasteh, Gary S. Stein, Kristen Jepsen, Donald L. Weaver, Gillian L. Hirst, Brian L. Sprague, Laura J. Esserman, Alexander D. Borowsky, Janet L. Stein, Olivier Harismendy. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is a reformatted presentation of the material currently under review titled, "Accurate genome-wide germline profiling from decade-old archival tissue DNA reveals the contribution of common variants to precancer disease outcome" by Daniela Nachmanson, Meghana Pagadala, Joseph Steward, Callie Cheung, Lauryn Keeler Bruce, Nicole Q. Lee, Thomas J. O'Keefe, Grace Y. Lin, Farnaz Hasteh, Gerald P. Morris, Hannah Carter, Olivier Harismendy. The dissertation author was the primary investigator and author of this material.

VITA

2015 University of Washington
Bachelor of Science, Biochemistry

2022 University of California San Diego
Doctor of Philosophy, Bioinformatics and Systems Biology

PUBLICATIONS

Nachmanson, D., Pagadala, M., Steward, J., Chueng, C., Bruce, L. K., Lee, N. Q., O'Keefe, T., Lin, G. Y., Hasteh, F., Morris, G. P., Carter, H., Harismendy, O*. "Accurate genome-wide germline profiling from decade-old archival tissue DNA reveals the contribution of common variants to precancer disease outcome." *In Submission 2022.* ↟

Nachmanson, D., Jepsen, K., Yao, H., Nachmanson, B., Harismendy, O*. "PROCEED - Predicting exome sequencing depth of suboptimal DNA samples." *In Submission 2022.* ↟

Nachmanson, D., Officer, A., Mori, H., Gordon, J., Evans, M. F., Steward, J., Yao, H., O'Keefe, T., Hasteh, F., Stein, G. S., Jepsen, K., Weaver, D. L., Hirst, G. L., Sprague, B. L., Esserman, L. J., Borowsky, A. D., Stein, J. L. & Harismendy, O*. The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. NPJ Breast Cancer 8, 6 (2022). ↟

Gutkind, J. S., Molinolo, A. A., Wu, X., Wang, Z., **Nachmanson, D.**, Harismendy, O., Alexandrov, L. B., Wuertz, B. R., Ondrey, F. G., Laronde, D., Rock, L. D., Rosin, M., Coffey, C., Butler, V. D., Bengtson, L., Hsu, C.-H., Bauman, J. E., Hewitt, S. M., Cohen, E. E., Chow, H.-H. S., Lippman, S. M. & Szabo, E*. Inhibition of mTOR signaling and clinical activity of metformin in oral premalignant lesions. JCI Insight 6, (2021).

Choi, Y. Y., Shin, S.-J., Lee, J. E., Madlensky, L., Lee, S.-T., Park, J. S., Jo, J.-H., Kim, H., **Nachmanson, D.,** Xu, X., Noh, S. H., Cheong, J.-H. & Harismendy, O*. Prevalence of cancer susceptibility variants in patients with multiple Lynch syndrome related cancers. Scientific Reports 11, (2021).

Zhang, Y., Kohrn, B. F., Yang, M., **Nachmanson, D.,** Soong, T. R., Lee, I.-H., Tao, Y., Clevers,H., Swisher, E. M., Brentnall, T. A., Loeb, L. A., Kennedy, S. R., Salk, J. J., Naxerova, K. & Risques, R*. A. PolyG-DS: An ultrasensitive polyguanine tract-profiling method to detect clonal expansions and trace cell lineage. *Proc. Natl. Acad. Sci. U. S. A.* 118, (2021).

Nachmanson, D., Steward, J., Yao, H., Officer, A., Jeong, E., O'Keefe, T. J., Hasteh, F., Jepsen, K., Hirst, G. L., Esserman, L. J., Borowsky, A. D. & Harismendy, O*. Mutational profiling of micro-dissected pre-malignant lesions from archived specimens. BMC Med. Genomics 13, 173 (2020). ↟

Krimmel-Morrison, J. D., Ghezelayagh, T. S., Lian, S., Zhang, Y., Fredrickson, J., **Nachmanson, D.,** Baker, K. T., Radke, M. R., Hun, E., Norquist, B. M., Emond, M. J., Swisher, E. M. & Risques, R*. A. Characterization of TP53 mutations in Pap test DNA of women with and without serous ovarian carcinoma. *Gynecol. Oncol.* 156, 407–414 (2020).

Salk, J. J., Loubet-Senear, K., Maritschnegg, E., Valentine, C. C., Williams, L. N., Higgins, J. E., Horvat, R., Vanderstichele, A., **Nachmanson, D.,** Baker, K. T., Emond, M. J., Loter, E., Tretiakova, M., Soussi, T., Loeb, L. A., Zeillinger, R., Speiser, P. & Risques, R*. A. Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Rep.* 28, 132–144.e3 (2019).

Baker, K. T., **Nachmanson, D.,** Kumar, S., Emond, M. J., Ussakli, C., Brentnall, T. A., Kennedy, S. R. & Risques, R*. A. Mitochondrial DNA Mutations are Associated with Ulcerative Colitis Preneoplasia but Tend to be Negatively Selected in Cancer. *Mol. Cancer Res.* **17,** 488–498 (2019).

Nachmanson, D., Lian, S., Schmidt, E. K., Hipp, M. J., Baker, K. T., Zhang, Y., Tretiakova, M., Loubet-Senear, K., Kohrn, B. F., Salk, J. J., Kennedy, S. R. & Risques, R*. A. Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). *Genome Res.* 28, 1589–1599 (2018).

*Co-corresponding authors.

↟Included in dissertation.

ABSTRACT OF THE DISSERTATION

Profiling archival samples to examine the molecular underpinnings of early carcinogenesis

by

Daniela Nachmanson

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2022

Professor Olivier Harismendy, Chair
Professor Ludmil Alexandrov, Co-Chair

Precancer can represent a benign condition in some and a potent precursor to lethal cancer in others. As cancer screening and early detection of cancer improves, precancer diagnoses are becoming quite common, yet our poor understanding of precancer leaves clinicians unclear on how to treat them. Precancerous lesions are often tiny and archived, formalin-fixed, and paraffin-embedded (FFPE), which degrades and damages the small amount of nucleic acid available. This

makes the genomic and transcriptomic analysis of precancers quite challenging as many of the lesions, especially the smallest, are often incompatible with standard DNA or RNA sequencing assays. My work reports the optimization of both experimental and computational methods to perform genetic profiling of precancer lesions - including acquired (somatic) and inherited (germline) variation - and the application of these methods to improve our understanding of breast precancer, ductal carcinoma in situ (DCIS).

A low-input FFPE DNA optimized exome sequencing workflow was paired with RNA sequencing, histological analysis, and immune microenvironment spatial profiling to generate a multimodal breast precancer atlas across 85 histological regions microdissected from the lesions of 39 patients diagnosed with pure DCIS. This allowed the measurement of relationships between molecular and phenotypic features, phylogenetic analysis, and identification of immune states revealed by stromal and epithelial-specific profiling. Further, a low-coverage whole-genome sequencing with reference-based imputation method, which provided reliable germline genotyping directly from archival FFPE tissues, was leveraged to evaluate the role of germline genetics in long-term precancer progression. I applied this methodology to samples from 36 patients with DCIS and identified that increased established breast polygenic risk scores were predictive of adverse outcomes.

The improved mapping provided by the breast precancer atlas, and suggestive evidence that germline variants contribute to DCIS outcome represent valuable resources in the study of the natural history of breast precancer. In particular, the findings emphasize the importance of including contextual information - spatial, microenvironmental, and inherited genetic factors - to fully characterize lesions. By only using archival tissue, I set the stage for the large retrospective studies that will be needed to guide the clinical management of precancer.

INTRODUCTION

Nearly every cancer type has a premalignant precursor. Precancer is considered an intermediate state between normal and cancer, with increased risk but non-guaranteed progression to cancer. Precancer offers a window of clinical intervention in the earliest moments of tumorigenesis. Historically, research efforts have focused on characterizing and improving the treatment of late-stage cancers This is understandably so, as these patients have an urgent life-threatening need for intervention, as opposed to precancer or early-stage cancers. Recently, more focus has shifted towards cancer prevention, also referred to as chemoprevention. Adoption of HPV vaccination has resulted in an age-dependent 34-97% rate reduction of cervical cancer in young women in the UK, and reduced tobacco smoking has decreased lung cancer by an estimated 32% in the US, both representing cancer prevention successes [1,2]. Detection, monitoring, and treatment of premalignancy represent another forefront of cancer prevention serving as an early indicator for potential cancer formation. However, precancer is poorly characterized at the molecular level, there is no precancer analog to the cancer genome atlas (TCGA) just yet, though characterizations for individual tissue types are beginning to emerge - such as in skin, esophagus, and lung [3–5]. The current lack of molecular, spatial, and microenvironmental characterization and their interplay in the earliest stages of tumorigenesis, hinder the development of prevention strategies, risk stratification models, and novel therapies. Characterization of these precancers can enable proper risk assessment and treatment to make cancer prevention the standard and reduce cancer morbidity.

There are very valid reasons why precancer is so poorly characterized - the samples are typically small and damaged, and progression is rare and slow [6]. Premalignant lesions (PML) in

precancer are small and to ensure the absence of invasive cancer, tissues are formalin-fixed and paraffin-embedded (FFPE) for histologic analysis. As such, fresh or frozen PML are rarely available. Formalin is known to create adducts in the DNA and lead to spurious substitutions. These DNA aberrations contribute to low coverage and make it difficult to distinguish artifacts from true somatic variants [7,8]. Almost every molecular-based assay which relies on nucleic acid readout, whether it be genetic, transcriptional, or epigenetic, struggles with low input amount and FFPE damage, resulting in many assays being suboptimal or simply not possible. The other primary challenge is logistical. While the probability and rate of progressing from precancer to cancer vary from tissue to tissue, oftentimes this transition will not occur and if it does, the process can take decades. The rarity of the outcome and the long-followup time require any study relating precancer molecular features to outcome to choose one of two suboptimal scenarios. Either, conduct a prospective study and recruit very large sample sizes to account for the low expected incidence and wait a decade, or collect old archival samples to generate a retrospective cohort, which will likely lack typical important controls specimen for genetic studies such as germline blood. While the time and sample size required for a prospective study can not be improved, the methodology to overcome the challenges of archival tissue required in a retrospective study can. Encouragingly, recent advances have improved molecular profiling of low-input FFPE tissue [9–11]. To continue advancing molecular profiling from retrospective archival tissue I will improve upon genomic profiling of archival tissue and establish the feasibility of retrospective study in epithelial premalignancy.

In the breast, ductal carcinoma in situ (DCIS) is a non-obligate precursor to deadly invasive breast carcinoma (IBC). DCIS is sensitively detected through screening and imaging and represents nearly a quarter of all breast cancer-related diagnoses [12]. Despite DCIS screening and

diagnostic success, there has been no parallel improvement in breast cancer mortality, indicating suboptimal risk stratification and therapeutic intervention of DCIS [13]. Current factors that impact the risk of breast cancer progression include age, size, grade of the lesion, or hormone receptor status, but they currently do not comprise a reliable prediction model [14,15]. It is increasingly accepted that low-risk DCIS patients are likely undergoing unnecessarily aggressive treatments including surgery, chemotherapy, radiation, and/or adjuvant therapies [16,17]. The high incidence of DCIS combined with the high mortality of IBC underlines the urgent need for improved molecular and histological characterization and subsequent improved therapeutic intervention.

Studies have identified that all transcriptional subtypes and most acquired genetic alterations in IBC are already established in DCIS, including the most frequently mutated driver genes[18–24]. Further, inherited variants that contribute to IBC risk, also contribute to DCIS risk [25]. This suggests that the presence or absence of particular genetic or molecular markers will likely be unable to distinguish the biology of DCIS vs IBC and more integrative measurements may be necessary. DCIS is also remarkably diverse - presenting with varying nuclear grades, sizes (<1cm to >15cm), histological architectures (solid, cribriform, micropapillary), and hormone receptor status (ER, HER2). Even the abundant genetic heterogeneity in IBC has been shown to already exist in DCIS, albeit the majority of these estimates have been performed in DCIS synchronous to invasive cancer which biases towards high-risk lesions [23,26–28]. In addition, DCIS does not grow in isolation. The duct where it develops exists in the context of complex and dynamic tissue and cellular architecture and immune microenvironment, under varying hormonal and metabolic contexts. Studies examining the roles of the basal layer, fibroblasts, adipocytes, stromal cells, and extracellular matrix have identified distinguishing features between DCIS and IBC [29–32]. The immune microenvironment has also been evidenced as having a key role in progression to IBC,

with evidence for high stromal tumor-infiltrating lymphocytes and B lymphocytes, or specific immunological make-up, being associated with a higher risk of progression [24,33–38]. The relationship between genetic and transcriptomic features, histological architecture, and immune microenvironment, and their diversity is poorly understood in DCIS. I will contribute to the mapping of these relationships and the underlying molecular profile of DCIS to help describe these early stages in carcinogenesis and provide an important basis for biomarker development.

In this dissertation, the primary goals are to improve the molecular characterization of precancer and apply the methodologies to better describe pure DCIS. From benchtop to the downstream bioinformatics analysis, I aim to optimize experimental and computational approaches to generate and analyze genetic data - both acquired and inherited (Chapters 1 and 3 respectively) - from precancer lesions. I then aim to leverage these methodologies, complemented with other precancer compatible methods, to profile novel DCIS samples and generate a multi-modal breast precancer atlas (Chapter 2). Lastly, I aim to evaluate the contribution of inherited variants to DCIS outcomes in a case-control study (Chapter 3). Notably, all of the above will be performed using just archival tissues, representing a framework for future retrospective precancer studies which will guide and advance chemo-preventative strategies.

References

1. Falcaro, M., Castañon, A., Ndlela, B., Checchi, M., Soldan, K., Lopez-Bernal, J., Elliss-Brookes, L. & Sasieni, P. The effects of the national HPV vaccination programme in England, UK, on cervical cancer and grade 3 cervical intraepithelial neoplasia incidence: a register-based observational study. *Lancet* **398,** 2084–2092 (2021).

2. Moolgavkar, S. H., Holford, T. R. & Levy, D. T. Impact of reduced tobacco smoking on lung cancer mortality in the United States during 1975–2000. *Journal of the* (2012). at <https://academic.oup.com/jnci/article-abstract/104/7/541/2517461>

3. Shain, A. H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., Dummer, R., North, J., Pincus, L., Ruben, B., Rickaby, W., D'Arrigo, C., Robson, A. & Bastian, B. C. The Genetic Evolution of Melanoma from Precursor Lesions. *N. Engl. J. Med.* **373,** 1926–1936 (2015).

4. Vinayanuwattikun, C., Le Calvez-Kelm, F., Abedi-Ardekani, B., Zaridze, D., Mukeria, A., Voegele, C., Vallée, M., Purnomosari, D., Forey, N., Durand, G., Byrnes, G., Mckay, J., Brennan, P. & Scelo, G. Elucidating Genomic Characteristics of Lung Cancer Progression from In Situ to Invasive Adenocarcinoma. *Sci. Rep.* **6,** 31628 (2016).

5. Ross-Innes, C. S., Becq, J., Warren, A., Cheetham, R. K., Northen, H., O'Donovan, M., Malhotra, S., di Pietro, M., Ivakhno, S., He, M., Weaver, J. M. J., Lynch, A. G., Kingsbury, Z., Ross, M., Humphray, S., Bentley, D. & Fitzgerald, R. C. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat. Genet.* **47,** 1038–1046 (2015).

6. Kelloff, G. J., Crowell, J. A., Steele, V. E., Lubet, R. A., Malone, W. A., Boone, C. W., Kopelovich, L., Hawk, E. T., Lieberman, R., Lawrence, J. A., Ali, I., Viner, J. L. & Sigman, C. C. Progress in cancer chemoprevention: development of diet-derived chemopreventive agents. *J. Nutr.* **130,** 467S–471S (2000).

7. Arreaza, G., Qiu, P., Pang, L., Albright, A., Hong, L. Z., Marton, M. J. & Levitan, D. Pre-Analytical Considerations for Successful Next-Generation Sequencing (NGS): Challenges and Opportunities for Formalin-Fixed and Paraffin-Embedded Tumor Tissue (FFPE) Samples. *Int. J. Mol. Sci.* **17,** (2016).

8. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61,** 64–71 (2015).

9. Foley, J. W., Zhu, C., Jolivet, P., Zhu, S. X., Lu, P., Meaney, M. J. & West, R. B. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome Res.* **29,** 1816–1825 (2019).

10. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H. & Campbell, P. J. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).

11. Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R. J., Abascal, F., Hall, M. W. J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., Fitzgerald, R. C., Handford, P. A., Campbell, P. J., Saeb-Parsy, K. & Jones, P. H. Somatic mutant clones colonize the human esophagus with age. *Science* **362,** 911–917 (2018).

12. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70,** 7–30 (2020).

13. Bleyer, A. & Welch, H. G. Effect of three decades of screening mammography on breast-cancer incidence. *N. Engl. J. Med.* **367,** 1998–2005 (2012).

14. Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F. & Antoniou, A. C. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21,** 1708–1718 (2019).

15. Silverstein, M. J. The University of Southern California/Van Nuys prognostic index for ductal carcinoma in situ of the breast. *Am. J. Surg.* **186,** 337–343 (2003).

16. Esserman, L. J., Thompson, I. M., Jr & Reid, B. Overdiagnosis and overtreatment in cancer: an opportunity for improvement. *JAMA* **310,** 797–798 (2013).

17. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* **380,** 1778–1786 (2012).

18. Pang, J.-M. B., Savas, P., Fellowes, A. P., Mir Arnau, G., Kader, T., Vedururu, R., Hewitt, C., Takano, E. A., Byrne, D. J., Choong, D. Y., Millar, E. K., Lee, C. S., O'Toole, S. A., Lakhani, S. R., Cummings, M. C., Mann, G. B., Campbell, I. G., Dobrovic, A., Loi, S., Gorringe, K. L. & Fox, S. B. Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. *Mod. Pathol.* **30,** 952–963 (2017).

19. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J. & Forbes, S. A. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47,** D941–D947 (2019).

20. Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M. & Bernard, P. S. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

21. Lin, C.-Y., Vennam, S., Purington, N., Lin, E., Varma, S., Han, S., Desa, M., Seto, T., Wang, N. J., Stehr, H., Troxell, M. L., Kurian, A. W. & West, R. B. Genomic landscape of ductal carcinoma in situ and association with progression. *Breast Cancer Res. Treat.* **178,** 307–316 (2019).

22. Nagasawa, S., Kuze, Y., Maeda, I., Kojima, Y., Motoyoshi, A., Onishi, T., Iwatani, T., Yokoe, T., Koike, J., Chosokabe, M., Kubota, M., Seino, H., Suzuki, A., Seki, M., Tsuchihara, K., Inoue, E., Tsugawa, K., Ohta, T. & Suzuki, Y. Genomic profiling reveals heterogeneous populations of ductal carcinoma in situ of the breast. *Commun Biol* **4,** 438 (2021).

23. Pareja, F., Brown, D. N., Lee, J. Y., Da Cruz Paula, A., Selenica, P., Bi, R., Geyer, F. C., Gazzo, A., da Silva, E. M., Vahdatinia, M., Stylianou, A. A., Ferrando, L., Wen, H. Y., Hicks, J. B., Weigelt, B. & Reis-Filho, J. S. Whole-Exome Sequencing Analysis of the Progression from Non-Low-Grade Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. *Clin. Cancer Res.* **26,** 3682–3693 (2020).

24. Abba, M. C., Gong, T., Lu, Y., Lee, J., Zhong, Y., Lacunza, E., Butti, M., Takata, Y., Gaddis, S., Shen, J., Estecio, M. R., Sahin, A. A. & Aldaz, C. M. A Molecular Portrait of High-Grade Ductal Carcinoma In Situ. *Cancer Res.* **75,** 3980–3990 (2015).

25. Petridis, C., Brook, M. N., Shah, V., Kohut, K., Gorman, P., Caneppele, M., Levi, D., Papouli, E., Orr, N., Cox, A., Cross, S. S., Dos-Santos-Silva, I., Peto, J., Swerdlow, A., Schoemaker, M. J., Bolla, M. K., Wang, Q., Dennis, J., Michailidou, K., Benitez, J., González-Neira, A., Tessier, D. C., Vincent, D., Li, J., Figueroa, J., Kristensen, V., Borresen-Dale, A.-L., Soucy, P., Simard, J., Milne, R. L., Giles, G. G., Margolin, S., Lindblom, A., Brüning, T., Brauch, H., Southey, M. C., Hopper, J. L., Dörk, T., Bogdanova, N. V., Kabisch, M., Hamann, U., Schmutzler, R. K., Meindl, A., Brenner, H., Arndt, V., Winqvist, R., Pylkäs, K., Fasching, P. A., Beckmann, M. W., Lubinski, J., Jakubowska, A., Mulligan, A. M., Andrulis, I. L., Tollenaar, R. A. E. M., Devilee, P., Le Marchand, L., Haiman, C. A., Mannermaa, A., Kosma, V.-M., Radice, P., Peterlongo, P., Marme, F., Burwinkel, B., van Deurzen, C. H. M., Hollestelle, A., Miller, N., Kerin, M. J., Lambrechts, D., Floris, G., Wesseling, J., Flyger, H., Bojesen, S. E., Yao, S., Ambrosone, C. B., Chenevix-Trench, G., Truong, T., Guénel, P., Rudolph, A., Chang-Claude, J., Nevanlinna, H., Blomqvist, C., Czene, K., Brand, J. S., Olson, J. E., Couch, F. J., Dunning, A. M., Hall, P., Easton, D. F., Pharoah, P. D. P., Pinder, S. E., Schmidt, M. K., Tomlinson, I., Roylance, R., García-Closas, M. & Sawyer, E. J. Genetic predisposition to ductal carcinoma in situ of the breast. *Breast Cancer Res.* **18,** 22 (2016).

26. Foschini, M. P., Morandi, L., Leonardi, E., Flamminio, F., Ishikawa, Y., Masetti, R. & Eusebi, V. Genetic clonal mapping of in situ and invasive ductal carcinoma indicates the field cancerization phenomenon in the breast. *Hum. Pathol.* **44,** 1310–1319 (2013).

27. Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M. E. & Navin, N. E. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172,** 205–217.e12 (2018).

28. Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lønning, P. E., Wedge, D. C. & Campbell, P. J. Subclonal

diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21,** 751–759 (2015).

29. Allen, M. D., Marshall, J. F. & Jones, J. L. αvβ6 Expression in myoepithelial cells: a novel marker for predicting DCIS progression with therapeutic potential. *Cancer Res.* **74,** 5942–5947 (2014).

30. Delort, L., Cholet, J., Decombat, C., Vermerie, M., Dumontet, C., Castelli, F. A., Fenaille, F., Auxenfans, C., Rossary, A. & Caldefie-Chezet, F. The Adipose Microenvironment Dysregulates the Mammary Myoepithelial Cells and Could Participate to the Progression of Breast Cancer. *Front Cell Dev Biol* **8,** 571948 (2020).

31. Allinen, M., Beroukhim, R., Cai, L., Brennan, C., Lahti-Domenici, J., Huang, H., Porter, D., Hu, M., Chin, L., Richardson, A., Schnitt, S., Sellers, W. R. & Polyak, K. Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* **6,** 17–32 (2004).

32. Hu, M., Yao, J., Carroll, D. K., Weremowicz, S., Chen, H., Carrasco, D., Richardson, A., Violette, S., Nikolskaya, T., Nikolsky, Y., Bauerlein, E. L., Hahn, W. C., Gelman, R. S., Allred, C., Bissell, M. J., Schnitt, S. & Polyak, K. Regulation of in situ to invasive breast carcinoma transition. *Cancer Cell* **13,** 394–406 (2008).

33. Gerdes, M. J., Gökmen-Polar, Y., Sui, Y., Pang, A. S., LaPlante, N., Harris, A. L., Tan, P.-H., Ginty, F. & Badve, S. S. Single-cell heterogeneity in ductal carcinoma in situ of breast. *Mod. Pathol.* **31,** 406–417 (2018).

34. Pruneri, G., Lazzeroni, M., Bagnardi, V., Tiburzio, G. B., Rotmensz, N., DeCensi, A., Guerrieri-Gonzaga, A., Vingiani, A., Curigliano, G., Zurrida, S., Bassi, F., Salgado, R., Van den Eynden, G., Loi, S., Denkert, C., Bonanni, B. & Viale, G. The prevalence and clinical relevance of tumor-infiltrating lymphocytes (TILs) in ductal carcinoma in situ of the breast. *Ann. Oncol.* **28,** 321–328 (2017).

35. Campbell, M. J., Baehner, F., O'Meara, T., Ojukwu, E., Han, B., Mukhtar, R., Tandon, V., Endicott, M., Zhu, Z., Wong, J., Krings, G., Au, A., Gray, J. W. & Esserman, L. Characterizing the immune microenvironment in high-risk ductal carcinoma in situ of the breast. *Breast Cancer Res. Treat.* **161,** 17–28 (2017).

36. Trinh, A., Gil Del Alcazar, C. R., Shukla, S. A., Chin, K., Chang, Y. H., Thibault, G., Eng, J., Jovanović, B., Aldaz, C. M., Park, S. Y., Jeong, J., Wu, C., Gray, J. & Polyak, K. Genomic Alterations during the In Situ to Invasive Ductal Breast Carcinoma Transition Shaped by the Immune System. *Mol. Cancer Res.* **19,** 623–635 (2021).

37. Lesurf, R., Aure, M. R., Mørk, H. H., Vitelli, V., Oslo Breast Cancer Research Consortium (OSBREAC), Lundgren, S., Børresen-Dale, A.-L., Kristensen, V., Wärnberg, F., Hallett, M. & Sørlie, T. Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Rep.* **16,** 1166–1179 (2016).

38. Gil Del Alcazar, C. R., Huh, S. J., Ekram, M. B., Trinh, A., Liu, L. L., Beca, F., Zi, X., Kwak, M., Bergholtz, H., Su, Y., Ding, L., Russnes, H. G., Richardson, A. L., Babski, K.,

Min Hui Kim, E., McDonnell, C. H., 3rd, Wagner, J., Rowberry, R., Freeman, G. J., Dillon, D., Sorlie, T., Coussens, L. M., Garber, J. E., Fan, R., Bobolis, K., Allred, D. C., Jeong, J., Park, S. Y., Michor, F. & Polyak, K. Immune Escape in Breast Cancer During In Situ to Invasive Carcinoma Transition. *Cancer Discov.* **7,** 1098–1115 (2017).

CHAPTER 1: Mutational profiling of micro-dissected pre-malignant lesions from archived specimens

1.1 Abstract

Systematic cancer screening has led to the increased detection of precancer. The absence of reliable prognostic markers has led mostly to overtreatment resulting in potentially unnecessary stress, or potentially insufficient treatment and avoidable progression. Importantly, most mutational profiling studies have relied on precancer synchronous to invasive cancer, or performed in patients without outcome information, hence limiting their utility for biomarker discovery. The limitations in comprehensive mutational profiling of precancer are in large part due to the significant technical and methodological challenges: most premalignant lesions (PML) are small, fixed in formalin and paraffin-embedded (FFPE), and lack matching normal DNA. Using test DNA from a highly degraded FFPE specimen, multiple targeted sequencing approaches were evaluated, varying DNA input amount (3-200 ng), library preparation strategy (BE: Blunt-End, SS: Single-Strand, AT: A-Tailing), and target size (whole exome vs cancer gene panel). Variants in high-input DNA from FFPE and mirrored frozen specimens were used for precancer-specific variant calling training and testing, respectively. The resulting approach was applied to profile and compare multiple regions micro-dissected (mean area 5 mm^2) from 3 breast ductal carcinoma in situ (DCIS). Using low-input FFPE DNA, BE and SS libraries resulted in 4.9 and 3.7 increase over AT libraries in the fraction of whole-exome covered at 20x (BE:87%, SS:63%, AT:17%). Compared to high-confidence somatic mutations from frozen specimens, precancer-specific variant filtering increased recall (BE:85%, SS:80%, AT:75%) and precision (BE:93%, SS:91%, AT:84%) to levels expected from sampling variation. Copy number alterations were consistent across all tested approaches and only impacted by the design of the capture probe-set. Applied to DNA extracted

from 9 micro-dissected regions (8 DCIS, 1 normal epithelium), the approach achieved comparable performance, illustrated the data adequacy to identify candidate driver events (*GATA3* mutations, *ERBB2* or *FGFR1* gains, *TP53* loss) and measure intra-lesion genetic heterogeneity. Lastly, we developed and trained PROjeCt ExomE Depth (PROCEED), an exome sequencing depth prediction model trained on pre-sequencing metrics from 166 library preparation condition, shared via a web-application to assist experimental users in quality assessment prior to sequencing. Alternate experimental and analytical strategies increased the accuracy of DNA sequencing from archived micro-dissected PML regions, supporting the deeper molecular characterization of early cancer lesions and achieving a critical milestone in the development of biology-informed prognostic markers and precision chemo-prevention strategies.

1.2 Introduction

For some cancer types, the wide-spread adoption of cancer screening has increased the detection of precancer [1]. Today, breast ductal carcinoma in situ (DCIS) comprises nearly ~25% of all breast cancer diagnoses in the United States [2]. In the case of DCIS, disease-specific guidelines recommend surgical excision and radiation, and endocrine risk reducing therapy. While treatment prevents rate of second events, it has not translated into increase survival rates, which are very high for most patients with DCIS, suggesting overtreatment of precancer and highlighting a critical need to improve risk models and identify prognostic markers [1,3,4]. Current precancer risk models rarely account for molecular biomarkers such as mutations or copy number alterations, which are seldom profiled. This is in large part due to technical challenges in profiling premalignant lesions (PMLs), specimens from PML biopsies are typically formalin-fixed and paraffin-embedded (FFPE) in their entirety, to verify the absence of any invasive component. As a consequence, no fresh or frozen material is available for research. Moreover, many PMLs observed in absence of invasive lesions are very small or have low overall cellularity, sometimes less than a millimeter in diameter or containing fewer than 1000 cells. Hence, while FFPE specimens have successfully been used in high-throughput sequencing, the challenges posed by excessive formalin-induced deoxyribonucleic acid (DNA) damage – detailed below - are typically overcome by an increase in DNA input quantity [5–9], a solution not available for archival PML profiling. Thus, small FFPE PML specimens pose significant challenges in the generation of high throughput sequencing libraries and preclude the investigation of genetic biomarkers. To overcome these limitations, previous studies have been performed in fresh PML from areas adjacent to invasive disease instead of on pure PML, ignoring the vast majority of PML that are less likely to progress [10–15]. Profiling

of pure precancer in the absence of invasive disease is required to avoid such biases and thus necessitates methodology to work with archival FFPE specimens.

One of the main challenges of library preparation from damaged, low input DNA samples is to preserve the library complexity: the faithful and unbiased representation of all fragments in the starting DNA sample. Indeed the multiple steps of the library preparation, including the repair of the input DNA, the ligation of adapter, target enrichment and the multiple rounds of purification and PCR amplification can all act as bottlenecks, and introduce strong skews that will reduce the library complexity and eventually impact precision and recall of variant calling. Moreover, formalin is known to create adducts in the DNA and lead to spurious substitutions, which can be difficult to distinguish from true somatic variants, especially at low allelic fractions [16,17]. Finally, the most insightful prognostic biomarker studies of precancer progression require long follow up to rely on actual outcome (recurrence, second events, survival) rather than proxy risk markers (grade, subtype, histological markers). As a consequence, most studies are retrospective and rely on old archived material without matching germline DNA sample, rendering the identification of high confidence somatic mutations more difficult. Hence, both technical and experimental challenges are hampering progress in precancer mutational profiling.

Here we present the development of DNA library preparation and variant calling strategies specifically optimized for low abundance, damaged DNA, commonly extracted from PMLs. Using highly damaged DNA, we compared the effect of the input amount, the size of the captured genomic region, and library preparation strategy on the quality of coverage depth and breadth. We determined that library preparation using blunt-end (BE) adapter ligation strategy maximizes the library complexity down to 3 nanograms (ng) of input DNA and is compatible with whole exome capture. Further, we generated a web-tool PROjeCt ExomE Depth (PROCEED), to help predict

the coverage of exome sequencing libraries soley using pre-sequencing metrics. Using a set of DNA variants called from a frozen mirrored tissue specimen, we optimized the variant calling strategy to maximize its accuracy. We further demonstrated its validity on 10 DNA samples extracted from laser capture micro-dissected regions of PML or adjacent normal from DCIS of 3 independent patient FFPE specimen. We illustrate the utility of the approach to identify somatic mutations in candidate genes and characterize precancer clonal heterogeneity within a specimen.

1.3 Results

Evaluation of targeted sequencing approaches for low input FFPE DNA.

Regions of PML on a histological section can contain as few as 500 cells, corresponding to 3.3 ng of haploid DNA. The expected reduced DNA extraction yield can be mitigated by combining regions matched across sequential sections. Hence, optimizing targeted sequencing down to 3 ng of input DNA is a reasonable objective to identify high confidence mutations in PML. To develop the methodology, a test DNA sample was extracted from a 4 year-old FFPE HER2 positive breast invasive carcinoma which showed significant fragmentation (DNA integrity number of 2.4), likely representative of DNA extracted from old archived specimens. High-throughput sequencing libraries were prepared using traditional A-tailing adapter ligation protocol optimized for low-input, damaged DNA (referred to as AT method—Figure 1.1a) with a decreasing amount of input DNA from 200 down to 10 ng. After capture of the whole exome by hybrid selection, the DNA libraries were amplified and sequenced. The performance was evaluated in comparison to whole exome data generated from 200 ng of DNA extracted from a mirrored frozen tissue specimen [18]. Libraries generated from 200 or 50 ng FFPE DNA achieved reasonable coverage, with nearly all targeted bases covered at least 20-fold (referred to as Cov20). In contrast, the sequencing libraries from 10 ng FFPE DNA lacked complexity (79% read PCR duplicates), resulting in 17% Cov20 after elimination of the duplicate reads (Figure 1.1b, c). Such poor performance at 10 ng, precluded us from further decreasing the DNA input and suggested that perhaps a smaller capture panel, restricted to cancer genes (710 kb total size) would elicit the goal of 3 ng input. Unfortunately, 3 ng AT libraries sequencing with a cancer panel had a high percentage of duplicate reads (83%) and 1.6% Cov20 (Figure 1.1d, e). Hence, the consistently poor performance of both low input strategies (3 ng and 10 ng) irrespective of the capture panel

size suggests that the bottleneck reducing the library complexity originates upstream of the targeted capture, which led us to examine the initial ligation of the sequencing adapters.

Library preparation methods which utilize alternate ligation techniques have previously been shown to increase the complexity of the library notably for the analysis of cell free DNA [19,20], ancient DNA [21–23] and other applications [24,25]. We therefore evaluated blunt-end ligation (BE) and single strand ligation (SS) library preparation strategies to increase the number of input DNA fragments incorporated in the library and enhance the resulting complexity and usable sequence coverage (Figure 1.1a, Figure S1). The cancer panel capture of BE libraries prepared from 3 ng of test DNA showed a reduced percentage of duplicate reads compared to AT libraries with 10 ng input (73% vs. 83% respectively Figure 1.1e), leading to a dramatic increase in Cov20 (99.7% vs. 2%) and bringing it to levels comparable to high input (50 ng) AT libraries (Figure 1.1d), albeit with higher duplicate rates. In turn, the SS library offered a lesser, but measurable, improvement over the AT library for low input (Figure 1.1d, e). The superiority of the BE and SS strategies were further confirmed using whole exome capture with BE and SS libraries resulting in 84% and 63% Cov20 at low input (3 ng), respectively, higher than 17% observed for low input (10 ng) AT libraries (Figure 1.1b). Compared to the cancer panel capture, the improvements of these alternative ligation strategies on whole exome sequencing were milder but remained remarkable and, in the case of BE strategy, likely to support more sensitive mutational profiling of archived PMLs.

Predicting exome coverage depth.

In order to help identify which FFPE DNA libraries would generate sufficient coverage for downstream analyses, we constructed an interpretive model for predicting mean exome coverage from FFPE samples such as PML using multiple linear regression trained on 166 FFPE DNA

exome sequencing libraries (Table 1.1). Libraries were derived from FFPE blocks stored for a median of 4 years with a mean DNA yield of 17 ng (range 2-309) and were subsequently prepared with several different library preparation kits representative of various exome library preparation techniques (AT, BE, and others, see Methods). The mean exome coverage was predicted as a function of the number of cycles and total nanograms (ng) of DNA produced in the pre-hybridization PCR, the total sequencing reads, and the type of sequencing kit used (Figure 1.3a). Overall the mean exome coverage depth D for a single FFPE DNA library was modeled as:

$$D = \beta_0 + \beta_1 PCR\ cycles + \beta_2 PCR\ yield + \beta_3 Reads + \beta_4 Kit$$

The parameters of the model were solved by ordinary least squares. Overall, we found the above model to be predictive of mean coverage, with an adjusted R^2 value of 0.85 (Table 1). While all variables were found to be significant ($p<0.001$), we quantified the contribution of each variable to the predictive performance by removing that variable and evaluating the partial model (Figure 1.3b). Overall, models excluding PCR yield or excluding PCR cycles, had the largest decrease in accuracy, R^2=0.61 and 0.43 respectively, with the removal of both resulting in R^2=0.16, suggesting these had the greatest contribution in the prediction. In order to evaluate the overall goodness of the fit we ran a 5-fold cross-validation over a hundred iterations, and estimated the average root mean squared error (RMSE) of the predicted depth to be 13.7 (Figure 1.3c). Despite the extreme variability of FFPE derived DNA quantity and quality, as well as many other factors that contribute to sequencing performance excluded from this model, we find reasonable coverage predictions from very limited measurements that do not require measurement of original DNA input or sample sequencing. PROCEED thus provides a valuable quality control tool for experimental scientists in order to prevent sequencing samples that will produce insufficient coverage for downstream analysis such as mutation and copy number assessment.

Somatic mutation profiling from low input FFPE DNA.

At a minimum, both SNVs and indels are necessary to evaluate the mutational landscape of PML. Unfiltered SNVs identified in FFPE DNA showed both a high overall abundance (422,322) of low variant allelic fraction substitutions (VAF < 5%) and bias of C to T substitutions (53%), which is expected from the cytosine deamination resulting from formalin fixation [16,17]. In contrast, low VAF variants from frozen DNA were much lower in abundance (175,364) but contained a high-prevalence (52%) of C to A substitutions consistent with previously reported 8-oxoguanine damage observed in frozen samples (Figure S2) [26]. We hypothesized that we could discriminate against artifactual FFPE variants in the test specimens using stringent filtering criteria including high strand bias, low allelic fraction and poor concordance between multiple variant callers in order to call accurate somatic variants in FFPE preserved PML (Methods).

First, we established a set of benchmarking somatic mutations from the test DNA extracted from a mirroring frozen tissue specimen. Its whole exome sequence resulted in 247 SNVs and 10 indels that were used to measure performance of the variant calling from the FFPE libraries generated above. Prior to filtering, the analysis of variants from low input AT libraries resulted in an average of 7,475 false positive somatic mutations. In contrast, the analysis of variants from BE and SS libraries resulted in an average of 1,967 and 3,137 false positive mutations, respectively (Figure 1.2a). We developed additional filtering criteria, trained on variants from the high-input (200 ng) FFPE library and used variants called from a publicly available panel of normal samples to remove additional artifacts. This approach considerably reduced the fraction of false positives (Figure 1.2b), increasing precision from less than 20% to 84%, 93% and 91% for the AT, BE and SS low input libraries, respectively (Figure 1.2c, d). The variant recall increased from 75% in the AT low input library to 85% and 80% for the BE and SS low-input libraries, respectively,

consistent with differences observed in Cov20 (Table S1.6). Importantly, these values were similar to the theoretical maximum values obtained from down-sampling the frozen sample itself to an equivalent number of reads (90% precision, 88% recall), indicating that the differences mainly come from sampling rather than from technical artifacts. The improvement in accuracy was similar for the small number of indels (Table S1.6). These data suggest that FFPE specific filtering of ensemble variant calls paired with BE library preparation enables accurate clonal somatic SNV and indel variant calling of whole exome sequencing data from 3 ng FFPE test DNA.

In contrast to SNVs and indels, Copy Number Alterations (CNA) can be accurately identified using lower coverage depth whole genome sequencing data, though has been more challenging in targeted sequencing [27,28]. We evaluated the ability of all input quantity and library preparation methods to reliably identify CNA (Methods). The genome-wide CNA burden obtained from low input SS and BE library was consistent with the one of high-input (200 ng) AT libraries (8.1%, 7% and 8.4% respectively—Figure 1.2e and Table S1.4), and consistent with the results of the cancer panel sequencing and the lower CNA burden observed in the 10 and 50 ng AT libraries are still within the expected confidence interval (Figure 1.2f, Figure S3). The resulting level of copy number gains and losses estimated for each chromosome arms and more focal areas were highly reproducible between all tested library preparation strategies (Figure S4, S5). Additionally, the copy number status of known cancer genes was also consistent between exome and cancer panel, including the expected *ERBB2* amplification which was correctly determined in all cases. In a few instances, the denser tiling of the exome probes helped identify a copy number breakpoint missed in the cancer panel, resulting in discordant copy number estimate for a few genes (Figure S6). This suggests that the input amount and quality of the sample have little impact on the

accuracy of the copy number profiling, albeit this observation was limited to a specimen with few CNAs

Mutational profiling of breast PML.

To validate the optimized targeted sequencing and somatic variant calling on PMLs, we collected archived tissue specimens from 3 patients diagnosed with breast high grade ductal carcinoma in situ (DCIS—1 low grade, 2 high grade) without evidence of invasive disease. A total of 8 PML regions and one normal breast epithelium region (patient 3, region 3 N) were isolated by laser capture microdissection (LCM—Figure 1.4a). The dissected regions had a mean area of 5 mm2 and, combined over three adjacent sections, contained an average of 80,000 cells (Figure 1.4b). For one large DCIS region, adjacent sections (patient 2, regions 2A1 and 2A2) were processed independently for replication. For each region, between 1.4 and 21 ng of DNA were extracted and used to prepare exome libraries using the BE method. The rate of duplicate reads was between 32 and 82%, and, as expected, inversely correlated with the amount of input DNA (Figure S7). The resulting mean coverage depth was between ~ 2 and 45 fold, which is sufficient for accurate detection of CNA, but likely limiting the sensitivity to identify mutations, particularly in patient 2.

Between 7 and 43% of the regions' genomes were involved in CNAs, predominantly through copy number losses (Figure 1.5a). With the exception of region 2B, the CNA burden was consistent between regions of the same patient and minimal in normal breast epithelium (< 5%). A total of 18 chromosome arms were affected by copy number changes in at least one DCIS region. None of these were altered in the normal breast epithelium (Figure 1.5b). Some of the chromosome arm losses identified, such as 6q, 8p, 16q, 17p or 22q are frequent in DCIS [29]. Within patient 1 and 3, all regions have consistent CNA suggesting a common clonal ancestor. In contrast, regions 2A

and 2B have a different number of arms altered (5 vs. 13, respectively) with only 3 in common. Both regions featured high nuclear grade, but only 2B showed comedo-necrosis, a marker of more advanced and worse prognosis DCIS. Region 2B was the only region affected by chromosome arm gains at 8q, 20q and 21q. The absence of these gains in 2A as well as its additional losses of 4p and 9p, suggest the independent clonal evolution of 2A and 2B. Finally, all PML regions from patient 2 and 3 displayed a loss of 17p, generally associated with *TP53* loss of heterozygosity (LOH) frequently observed in high-grade specimens. At a higher resolution, out of 98 cancer genes evaluated, 38 had a CNA in at least one region (Figure 1.5c). The most notable ones were the amplification of *ERBB2* in all regions of patient 2, amplification of *FGFR1* in all regions of patient 3 and loss of *TP53* in patient 2 and 3. This latter observation was consistent with 17p LOH and confirmed by change in B-allele frequency at heterozygous SNPs (Figure S8). The relative gene expression levels measured for 3 genes in 4 matching regions were consistent with their copy number in 10/12 cases, with possible discrepancies due to spatial variation in histology or transcriptional regulation (Figure S9) [30]. Similar to chromosome arms, none of the genes were altered in the normal epithelium and most had consistent copy numbers between regions of the same patient, with the exception of 12 genes distinguishing regions 2A and 2B and further supporting separate clonal evolution.

We next identified somatic mutations in each region of patient 1 and 3's specimens. After additional quality filtering using cross-patient information, we identified between 18 and 154 somatic mutations per PML region, of which 14 to 108 were non-silent (Table S1.7). The resulting mutational burden (0.46–3.97 mutations/Mb) is a range similar to what has been observed in invasive breast cancer [31]. The somatic mutations were then used to characterize the clonal relationships between regions. To account for uneven sequencing coverage between regions and

the possibility of random allelic dropout, we used a maximum likelihood approach, comparing allelic fraction and total read depth of any mutated position in all regions of patients 1 and 3 (Figure 1.5d, e) [32]. Mutations in patient 1 were evaluated at 155 mutated positions, of which 141 were confidently identified in all 3 regions, and 14 were either missing or absent from one or more regions (Table S1.8). While a portion of these may be shared mutations may be germline variants, we identified a *GATA3* splice-site deletion in regions 1A and 1B previously observed in DCIS studies, disrupting a canonical splice site and for which the resulting transcript has been shown to lead to an abnormally high number of neoantigens [33,34]. Similarly, mutations in patient 3 were evaluated at 183 positions, 160 of which were shared by all 4 regions, including normal, and likely residual germline variants. The results were used to build a phylogenic tree illustrating the clonal relationship between regions (Figure 1.5f). Interestingly, region 3 N is mutated at 4 positions not found in the PML regions. These could represent mutations acquired in aging normal tissue or residual germline variants lost in the PMLs [35–37]. The 3 PML regions gained an additional 8 mutations before diverging including an *ERBB3* Ile763Leu substitution was exclusively observed in all 3 PML regions of patient 3. This mutation is predicted to be deleterious, possibly activating this uncommon driver of breast cancer [38,39] and may have contributed, in concert with *FGFR1* gain, to the clonal expansion observed in this patient. Region 3A gained an additional 2 mutations, while 3C and 3B shared an additional 4 and each gained between 1 and 3 mutations that were not shared. Overall, our analysis suggests that even in absence of normal DNA, and akin to experiments in large tumors, variants from multi-region sequencing can be used to trace evolutionary relationships between areas of pre-malignancy [32,40].

1.4 Discussion

The results presented here demonstrate our ability to perform comprehensive mutational profiling from some of the most challenging clinical specimens with DNA in limited quantity—down to 3 ng—and of poor quality—highly degraded and chemically altered. In particular, this demonstration relied on a thorough benchmarking study using DNA from mirrored matching frozen versus FFPE specimens, which provided a real-world experimental framework to guide the process development.

We demonstrated that by utilizing a sequencing library preparation that uses a non-standard adapter ligation, we can drastically improve sequencing performance from these challenging specimens. A-tailing, together with transposon-mediated construction, is one of the most popular methods to prepare high-throughput sequencing libraries. While the latter necessitates longer DNA fragments and is not-suitable for highly degraded DNA, A-tailing has been broadly used in library preparation and is compatible with highly degraded specimens so long as DNA input is increased. We illustrated this limitation, analyzing targeted sequencing from limited dilutions and saw a drop in coverage and variant calling accuracy below 50 ng DNA input. Target coverage cannot be rescued by sequencing of a smaller panel, since the bottleneck resulting in lack of library complexity occurs prior to capture. For clinical reasons, to allow the thorough inspection of all tissue parts to formally exclude invasive disease, all pure precancer are fixed in formalin and archived. Furthermore, areas of PML are typically small, limiting the quantity of material available for analysis. For the largest lesions, a labor-intensive dissection and pooling can increase the amount of DNA extracted but would preclude the study of their genetic heterogeneity. Similar to the benefit demonstrated on ancient DNA [21,23], we showed that an alternative library preparation using either single-strand and even more so for blunt-end ligation strategies considerably improved

both coverage and consequently variant calling performances. Despite the recognized high efficiency of sticky-end ligation [41], end-repair and addition of an overhanging adenosine is likely a rate-limiting step on highly damaged DNA.

In addition to improving experimental preparation of targeted sequencing libraries, we also wanted a way to predict the coverage of a given sample prior to sequencing. As insufficient coverage for CNA or mutational analysis results in a high sample failure rate with wasted time and resources. As such we utilized limited measurements after the first PCR in the library preparation, including PCR cycles and yield, to provide a projection of mean exome coverage, which can aid in deciding which samples to sequence. This model is made available with a web application, PROCEED. Some important limitations of PROCEED exist, the training data does not represent all possible exome library preparation methods, as there are countless. However, we included a diverse array of samples prepared with three different library preparation kits, and expect the exome data is somewhat representative of commonly used protocols. Overall, PROCEED was designed to provide easy usage for users without programming backgrounds, and believe that it represents a useful tool in the quality prediction of exome data from FFPE samples such as precancer, saving users time and resources in their FFPE exome library preparations.

Another challenge we faced and addressed was calling mutations in absence of a matching normal DNA in both the test and PML DNA, likely leading to some ambiguous mutation classification [42]. The most useful precancer specimens for biomarker studies are the ones with long follow-up, and aside from logistical challenges of collecting matching blood or saliva as part of routine clinical workflow, these traditional sources of normal DNA are generally not available for archival precancer. The dissection of adjacent normal tissue, performed for one specimen in this study, can be sometimes used. While sufficient material can be found at the margin of the surgical

specimen, the cellularity of the normal tissue will vary greatly between organs and histological context leading to even smaller quantities of DNA. In the breast, the normal ductal tree is poorly cellular in comparison to dysplastic and in situ proliferative lesions. In some instances, an area of high lymphocytic infiltration, for example at the location of a previous biopsy, can be used. Such histologically normal tissue are also formalin fixed and their mutational profiling presents similar, if not more, challenges than for precancer. Furthermore, and as demonstrated elsewhere, some histologically normal specimens will contain a few somatic mutations at low-allelic fraction, resulting from early clonal selection, such as the few private mutations identified in sample 3N [35–37]. Nevertheless, the inclusion of such a matched normal DNA in the analysis and interpretation would greatly aid in the removal of residual germline DNA and additional sequencing artifacts. In absence of matching normal DNA, the parallel sequencing of a panel of unmatched normal DNA, from the same ethnic background and processed using the exact same protocol and analysis is recommended, especially in a clinical setting [42]. This was not available for this benchmarking study. Instead, we combined the use of a publicly obtained pool of normal with filtering for common variants using public databases following recommendation provided elsewhere [43,44]. This approach will miss rare germline variants, especially present in rare, or under-studied ethnicities. Previous studies have observed that tumor-only exomes may lead to ~300 residual germline variants after careful filtering [42,45]. Importantly, coding and deleterious germline variants could be a source of false positives. In our study, multi-region sampling provided additional information to help us classify mutations, and we determined that 155–160 mutations were shared and likely represented residual germline variants. This approach would however not remove the ambiguity for shared mutations and, in the event this mutation is an oncogenic driver common to all regions, additional validation steps may be required, including sequencing from a more distant region

located on a separate tissue block, or comparison to its allelic fraction and copy number status in the bulk DNA.

Independent of the possible residual germline variants, we took specific steps to benchmark variant calling in highly degraded FFPE DNA, comparing the results to high quality variants called from the DNA of an adjacent frozen specimen. After carefully down-sampling the data to obtain the same number of raw sequencing reads—including in silico "replicates" from the deeper reference (frozen tissue) dataset itself—we identified two main mode of errors. First, a large number of false negative variants, leading to lower recall, were directly associated to the uneven and lower coverage depth, particularly in the low input FFPE AT library. The high and more even coverage observed in the BE libraries, for the same number of raw reads, remediated this issue, resulting in recall similar to the one expected from sampling bias observed in other studies [46]. Second, the false positives were mostly due to C to T substitutions as a consequence of formalin fixation, as previously observed. Such substitutions however remained at low allelic fraction and displayed strong strand bias, which could be remediated using stringent heuristic filtering. Alternate solutions have been described elsewhere to remove same or similar sequencing artifacts using machine learning [47] or relying on the precise substitution signature of the artifacts [48,49]. These would likely yield similar or superior results. DNA damage can also be repaired prior to library preparation altogether using cocktails of DNA repair enzymes, such as UDG and Fpg [50,51]. This strategy would decrease false positive mutations [52] but typically require more than 5 ng FFPE DNA and the extra enzymatic step would likely add bottleneck and decrease library complexity.

Of the 3 patients studied, all displayed chromosomal copy number changes previously observed in DCIS, leading to losses of known tumor suppressors (*TP53*) or gains of known oncogenes (*ERBB2* or *FGFR1*). In 2/3 patients, we observed a comparable number of somatic

mutations to pure DCIS previously studies [53,54]. The identification of only two known or likely breast cancer driver mutations (*GATA3* and *ERBB3*) in two of the specimen was not surprising, as these specimens represent pure PML lacking any invasive component, and thus should bear less genomic resemblance to breast cancer [53]. Interestingly *GATA3* has been reported as mutated at a higher frequency in pure DCIS than in invasive cancer, suggesting a negative selection during transition to invasive cancer [34]. This particular splice-site mutation produces an alternative transcript resulting in a high numbers of neoantigens, perhaps subjecting mutated lesions to a more effective immune-surveillance [33]. The relative contribution of gene copy number alterations versus somatic mutations to cellular proliferation and clonal selection in normal and pre-malignant tissue is an active field of study [35–37] and progress in this field will require multi-modal molecular profiling approaches compatible with small amounts of archived tissue, such as the one described here.

Beyond the identification of drivers of growth and proliferation, the proposed approach can help measure the genetic heterogeneity in PML lesions. In breast DCIS, phenotypic heterogeneity associated with subtype and grade can co-exist within a specimen [55]. Additionally, immunostaining of key markers has revealed spatial heterogeneity within a duct and between ducts of a patient [56]. But the mutational landscape underlying such heterogeneity has not been thoroughly studied in pure DCIS at a genome-wide level for the technical reasons mentioned herein. Multi-region assessment of karyotypes, select gene copy number [57] and mitochondrial mutations [58] suggested significant heterogeneity in DCIS but its clinical significance and association with other progression risk factors has not been assessed. Copy-number heterogeneity has also been observed via single-cell sequencing from frozen DCIS patient specimens, albeit in the presence of an invasive lesion [15]. A similar approach has been developed in archived tissue specimens, but none

have been used in large studies or in pure DCIS [59]. While single-cell sequencing may be able to scale up—both number of cells and number of samples, it cannot yet identify point mutations with high sensitivity. Hence, in the context of a clinical study, multi-region sequencing enabled by LCM may be preferable as it would increase the accuracy of the clonal evolution and enable the identification of driver mutations and mutational signatures [60]. Moreover, the results of multi-region genomic profiling would enable us to place somatic mutations and copy number alterations in the context of the surrounding extra-cellular matrix or stromal composition obtained via imaging and morphological studies, providing a granular view of the premalignant landscape as aspired by the pre-cancer atlas [61].

1.5 Materials and Methods

Samples.

Test specimen.

Mirrored frozen-FFPE tissue specimen from a HER2 positive invasive breast cancer was obtained from Asterand Biosciences (Detroit, MI). The length distribution of the DNA fragments was measured by capillary electrophoresis (Agilent BioAnalyzer) and used to calculate the DNA integrity number (DIN) between 1 (very degraded) to 10 (intact genomic DNA)

PML specimen.

FFPE blocks were obtained from UCSD Health Anatomic Pathology after surgical biopsy, excision or mastectomy. The UCSD institutional review board approved the retrospective study and waived the requirement for consent. Consecutive sections of the blocks were used for Hematoxylin–Eosin staining (N = 1; 4 μM glass slide) then for Laser Capture Microdissection (LCM; N = 3; 7 μM glass slide coated with polyethylene naphthalate—ThermoFisher #LCM0522). The slides were stored at − 20 °C in an airtight container with desiccant until ready for dissection (1 day to 3 months). The LCM sections were thawed and stained with eosin, sections were kept in xylene and dissected within 2 h of staining. Laser Capture Microdissection was performed using the Arcturus Laser Capture Microdissection System. Matching regions from 3 adjacent sections were collected on one Capsure Macro Cap (Thermofisher), region size permitting. Post-dissection, all caps were covered and stored at − 20 °C.

DNA extraction and QC.

The DNA was extracted from FFPE tissue using the QIAamp DNA FFPE tissue kit and QIAamp DNA Micro Kit (Qiagen) for the test specimen or LCM specimen, respectively. For the

LCM sample, the membrane and adhering tissue were peeled off the caps using a razor blade and the peeled membrane was incubated in proteinase K digestion reaction overnight for 16 h at 56 °C to maximize DNA yield after cell lysis and the elution was done in 20 µL. The extracted DNA was quantified by fluorometry (HS dsDNA kit Qbit—Thermofisher). All samples used in the study are described in Table S1.1.

Targeted sequencing.

DNA fragmentation.

DNA was sheared down to 200 base pairs (bp) using Adaptive Focused Acoustics on the Covaris E220 (Covaris Inc) following manufacturer recommendations with the following modifications: 50 µL of Low TE buffer in microTUBE-130 tubes (AT libraries) or 10 µL Low EDTA TE buffer supplemented with 5 µL of truSHEAR buffer using a microTUBE-15 (SS and BE libraries).

Library preparation.

AT libraries were prepared with the SureSelect XT HS protocol (Agilent Technologies) extending the adapter ligation time to 45 min (min). After ligation, excess adapters were removed using a 0.8 × SPRI bead clean up with Agencourt AMPure XP beads (Beckman Coulter), then eluted into 21 µL of nuclease-free water. SS libraries were prepared using the Accel-NGS 1S Plus DNA Library Kit (Swift Biosciences). Prior to the single-strand ligation protocol, 15 µL of fragmented DNA was denatured at 95 °C for 2 min, then set on ice. The adaptase and extension steps were performed by kit specifications followed by a purification step using 1.2 × AMPure XP, eluted into 20 µL of nuclease free water. The subsequent ligation step incorporates SWIFT-1S P5 and SWIFT-1S P7 adapters, followed by a 1 × AMPure XP bead clean-up and elution into 20 µL of nuclease free water. BE libraries were prepared using the Accel-NGS 2S PCR-Free DNA

Library Kit (Swift Biosciences). Repair I was followed by a 1 × AMPure bead cleanup, Repair II was followed by a 1 × PEG NaCl cleanup, and Ligation I (P7 index adapter) and Ligation II (P5 UMI adapter) were followed by a 0.85 × PEG NaCl cleanups. Only Ligation II cleanup was eluted into 20 µL Low EDTA TE, the other cleanups proceeded directly into the next reaction. Adapters used in the study are summarized in Table S1.2.

Pre-capture PCR amplification.

Ligated and purified libraries were amplified using KAPA HiFi HotStart Real-time PCR 2X Master Mix (KAPA Biosystems). AT libraries were amplified with 2 µL of KAPA P5 primer and 2 µL of SureSelect P7 Index primer. SS libraries were amplified with 5 µL of SWIFT-1S P5 Index and P7 Index primers. BE samples were amplified with 5 µL of KAPA P5 and KAPA P7 primers. The reactions were denatured for 45 s (s) at 98 °C and amplified 13–15 cycles for 15 s at 98 °C, for 30 s at 65 °C, and for 30 s at 72 °C, followed by final extension for 1 min at 72 °C. Samples were amplified until they reached Fluorescent Standard 3, cycles being dependent on input DNA quantity and quality. PCR reactions were then purified using 1 × AMPure XP bead clean-up and eluted into 20 µL of nuclease-free water. The resulting libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qubit dsDNA HS assay). Primers used in the study are summarized in Table S1.2.

Targeted capture hybridization and post-capture PCR.

Samples were paired and combined (12 µL total) to yield a capture "pond" of at least 350 ng, and supplemented with 5 µL of SureSelect XTHS and XT Low Input Blocker Mix. The baits for target enrichment consisted of either Agilent SureSelect Clinical Research Exome panel (S06588914), Human All Exon V7 panel (S31285117) or Cancer All-In-One Solid Tumor (A3131601). The hybridization and capture was performed using Agilent SureSelect XT HS

Target Enrichment Kit following manufacturer's recommendations. Post-capture amplification was performed on the beads in a 25 μL reaction: 12.5 μL of nuclease-free water, 10 μL 5 × Herculase II Reaction Buffer, 1 μL Herculase II Fusion DNA Polymerase, 0.5 μL 100 mM (mM) dNTP Mix and 1 μL SureSelect Post-Capture Primer Mix. The reaction was denatured for 30 s at 98 °C, then amplified for 12 cycles of 98 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min, followed by an extension at 72 °C for 5 min and a final hold at 4 °C. Libraries were purified with a 1 × AMPure XP bead clean up and eluted into 20 μL nuclease free water in preparation for sequencing. The resulting libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qbit—ThermoFisher).

RNA-sequencing.

Expression profiling was performed on select dissected PML regions: 1A, 2A2, 2B and 3C. RNA library preparation was performed with SMART-3Seq, a 3′ tagging strategy specifically designed for degraded RNA directly from FFPE LCM specimen [30]. Read count data was obtained using a dedicated analysis workflow (https://github.com/danielanach/SMART-3SEQ-smk). Count data was then normalized for read depth and scaled by a million to give transcripts per million (TPM) counts.

Sequencing.

All libraries were sequenced using the HiSeq 4000 sequencer (Illumina) for 100 cycles in Paired-End mode. Libraries with distinct indexes were pooled in equimolar amounts. The sequencing and capture pools were later deconvoluted using program bcl2fastq [62].

Sequencing reads processing and coverage quality control.

Sequencing data was analyzed using bcbio-nextgen (v1.1.6) as a workflow manager [43]. Samples prepared with identical targeted panels were down-sampled to have equal number of reads

using seqtk sample (v1.3) [63]. Adapter sequences were trimmed using Atropos (v1.1.22), the trimmed reads were subsequently aligned with bwa-mem (v0.7.17) to reference genome hg19, then PCR duplicates were removed using biobambam2 (v2.0.87) [64–66]. Additional BAM file manipulation and collection of QC metrics was performed with picard (v2.20.4) and samtools (v1.9). The summary statistics of the sequencing and coverage results are presented in Table S1.3.

Copy number analysis.

Copy number alterations (CNAs) were called using CNVkit (v0.9.6) [67] using equal sized bins of ~ 250 bp. Any bins with log2 copy ratio lower than − 15, were considered artifacts and removed. Breakpoints between copy number segment were determined using the circular binary segmentation algorithm ($p < 10-4$) [68]. Low quality segments were removed from downstream analysis (less than 10 probes, biweight midvariance more than 2 or log2 copy ratio confidence interval contains 0). Copy number genomic burden was computed as the sum of sizes of segments in a gain (log2(ratio) > 0.3) or loss (log2(ratio) < − 0.3) over the sum of the sizes of all segments. The summary statistics of CNA calling on the test specimen are reported in Table S1.4. Chromosomal arm gains and losses were called when more than half of their total length was involved in a gained and lost segment, respectively. Gene copy number estimates were assigned based on the segment that covered the gene. For the test specimen, if more than one segment covered a gene then the higher confidence segment was used. For the DCIS specimen, copy number alterations were determined for all autosomal genes containing at least 3 bins whose segments were covered by at least 110 probes (N = 17,750 total). Additionally, due to the imprecision of segmentation breakpoints, any genes with a breakpoint identified in one region of a patient, were removed from the comparison to other regions. Loss of heterozygosity (LOH) was called for segments with B-allele frequencies lower than 0.3 or greater than 0.7.

Variant calling and initial filtering.

Single nucleotide variants (SNVs) and short insertions and deletions (indels) were called with VarDictJava (v1.6.0), and Mutect2 (v2.2) [69,70]. Variants were required to fall within a 10 bp boundary of targeted regions that overlapped with RefSeq genes (v 109.20190905). A publicly available list of variants observed in a pool of normal DNA exome sequencing was obtained from the GATK resource (https://console.cloud.google.com/storage/browser/details/gatk-best-practices/somatic-b37/Mutect2-exome-panel.vcf) and used to eliminate artifacts and common germline variants. Only variants called by both algorithms were considered (ensemble calling). These ensemble variants were then subjected to an initial filtering step with default bcbio-nextgen tumor-only variant calling filters listed in Table S1.5. Functional effects were predicted using SnpEff (v4.3.1) [71]. The resulting variants are referred to as raw ensemble variants.

Germline variant filtering.

In absence of a matched normal control for both test (frozen and FFPE) and DCIS specimens, somatic mutations were prioritized computationally using the approach from the bcbio-nextgen tumor-only configuration then additionally subjected to more stringent filtering [43]. Briefly, common variants (MAF > 10−3) present in population databases—1000 genomes (v2.8), ExAC (v0.3), or gnomAD exome (v2.1)—were removed unless in a tier 1 gene from the cancer gene consensus and present in either COSMIC (v68) or clinvar (20190513) [43,72–76]. Additionally, variants were removed as likely germline if found at a variant allelic fraction (VAF) greater or equal to 0.9 in non-LOH genomic segments—as determined by CNA analysis (above). The remaining variants are referred to as candidate somatic mutations.

Analysis specific filtering of candidate somatic mutations.

Additional filtering was implemented in a context specific to the analysis presented. (1) Calling "gold standard" mutations from the frozen test specimen: the candidate somatic mutations in the DNA of the frozen test specimens were filtered for high-quality variants: ensemble quality score greater than 175, average number of read mismatches less than 2.5, position covered by at least 25 reads, mean position in read greater than 20, microsatellite length less than 5, and VAF more than 0.14. This resulted in 247 SNVs and 10 indels. (2) Benchmarking mutations in FFPE test specimens: mutations in the DNA of the FFPE test specimens required specific filtering due to the abundant low-frequency damage as well as lower coverage depth. The following parameters were used: position covered by at least 5 reads, mapping quality more than 45, mean position in read greater than 15, number of average read mis-matches less than 2.5, microsatellite length less than 5, tumor log odds threshold more than 10, Fisher strand bias Phred-scaled probability less than 10 and VAF more than 0.14. The accuracy of resulting DNA variants from the test FFPE specimen was measured against the set of "gold standard" variants from the mirrored frozen specimen using vcfeval by RTG-tools (v3.10.1), using variant ensemble quality as the score [77]. The results of the benchmarking analysis are reported in Table S1.6. (3) Profiling dissected regions from the DCIS specimen: in addition to the filtering of FFPE candidate somatic mutations presented above, the following steps were implemented. Any variants found at high VAF (> 0.9) in non-LOH segments in one region were also excluded from the variants from all other regions of same patient. Candidate somatic mutations with ensemble quality score lower than 115 were excluded, corresponding to the optimal F-score obtained for low-input BE libraries in the benchmarking analysis. To the exception of few well-described hotspot mutations in breast cancer (*PIK3CA, TP53, GATA3*), somatic mutations identified in more than one patient were removed

Clonality analysis.

To allow the analysis of clonal relationships between regions of the same patient, the coverage depth of each allele at any remaining mutated position in any region was extracted using Mutect2 joint variant caller on the sets of aligned reads from each region. In order to call a mutation either absent or present in a region, we used a Bayesian inference model specifically designed for multi-region variant calling [32]. Treeomics (v1.7.10) was run with the default parameters except for e = 0.02.

PROCEED model and web-tool.

Library preparation of training samples.

DNA used for model training was all derived from oral or breast premalignant lesions which were formalin-fixed and paraffin-embedded (FFPE) and stored for a median of 6 years, then sectioned and laser-capture microdissected [78,79]. DNA libraries were prepared with various library preparation sequencing kits: AT: Agilent SureSelect XT HS (N=22), BE: Accel-NGS 2S PCR-Free DNA Library Kit (N=90), Other: Ultra II NEB FS (N=54). All samples underwent pre-hybridization PCR amplification with KAPA HiFi HotStart Real-time PCR 2X Master Mix (KAPA Biosystems) and the PCR was terminated just prior to the plateau. The total amount of DNA produced in the pre-hybridization was measured via fluorometry (Qbit - Thermofisher) after a post-PCR cleanup to remove adapters, and any leftover impurities. Exome hybridization was performed with the Agilent SureSelect XT HS Target Enrichment Kit and XT Low Input Blocker Mix paired with the Human All Exon V7 panel (S31285117). Samples were all sequenced on the Illumina NovaSeq or HiSeq4000 with 100bp paired-end reads.

Model training.

A multiple linear regression model was solved using ordinary least squares (OLS) with Statsmodels (v0.12.1) implemented in python (v3.7). The PCR cycles, the PCR yield and the number of sequencing reads were all Box-Cox normalized, and the average depth was square-root transformed prior to parameter estimation. Code for the model training and testing can be found here: https://github.com/danielanach/PROCEED/blob/89d6c182760df30e44a26c7628bb0f21bee5149f/model_training/build_PROCEED_model.ipynb

Web application.

PROCEED was implemented with Streamlit (v0.8.2) which was both used to build and host the web application. PROCEED can be accessed here: https://share.streamlit.io/danielanach/proceed/main/PROCEED.py

1.6 Figures

a

	AT		SS	BE
Specimen	**Frozen**	**FFPE**	**FFPE**	**FFPE**
Library	Standard A-tailing	Standard A-tailing	Single-Strand	Blunt-End
Targeted panel	Exome	Exome / Cancer panel	Exome / Cancer panel	Exome / Cancer panel
Input (ng)	200	200*, 50, 10*, 3	50† , 3	50† , 3

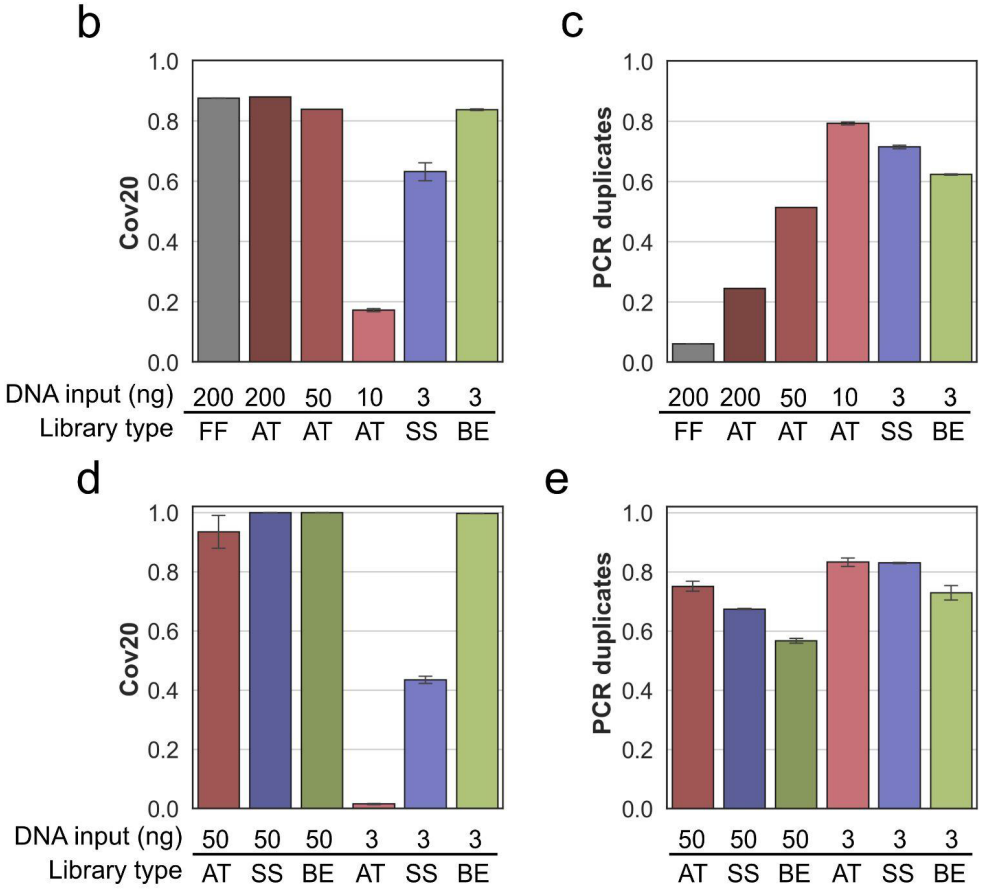


Figure 1.1. Benchmarking results for sequencing performance.
(a) Experimental design for performance evaluation using a test DNA specimen (* Exome and † Cancer panel). **(b, c)** Fraction of targeted bases covered by a minimum of 20 reads (b) and fraction of PCR duplicates (c) observed in whole exome sequencing. **(d, e)** Fraction of targeted bases covered by a minimum of 20 reads (d) and fraction of PCR duplicates (e) observed in cancer panel sequencing. All error bars represent standard deviation.

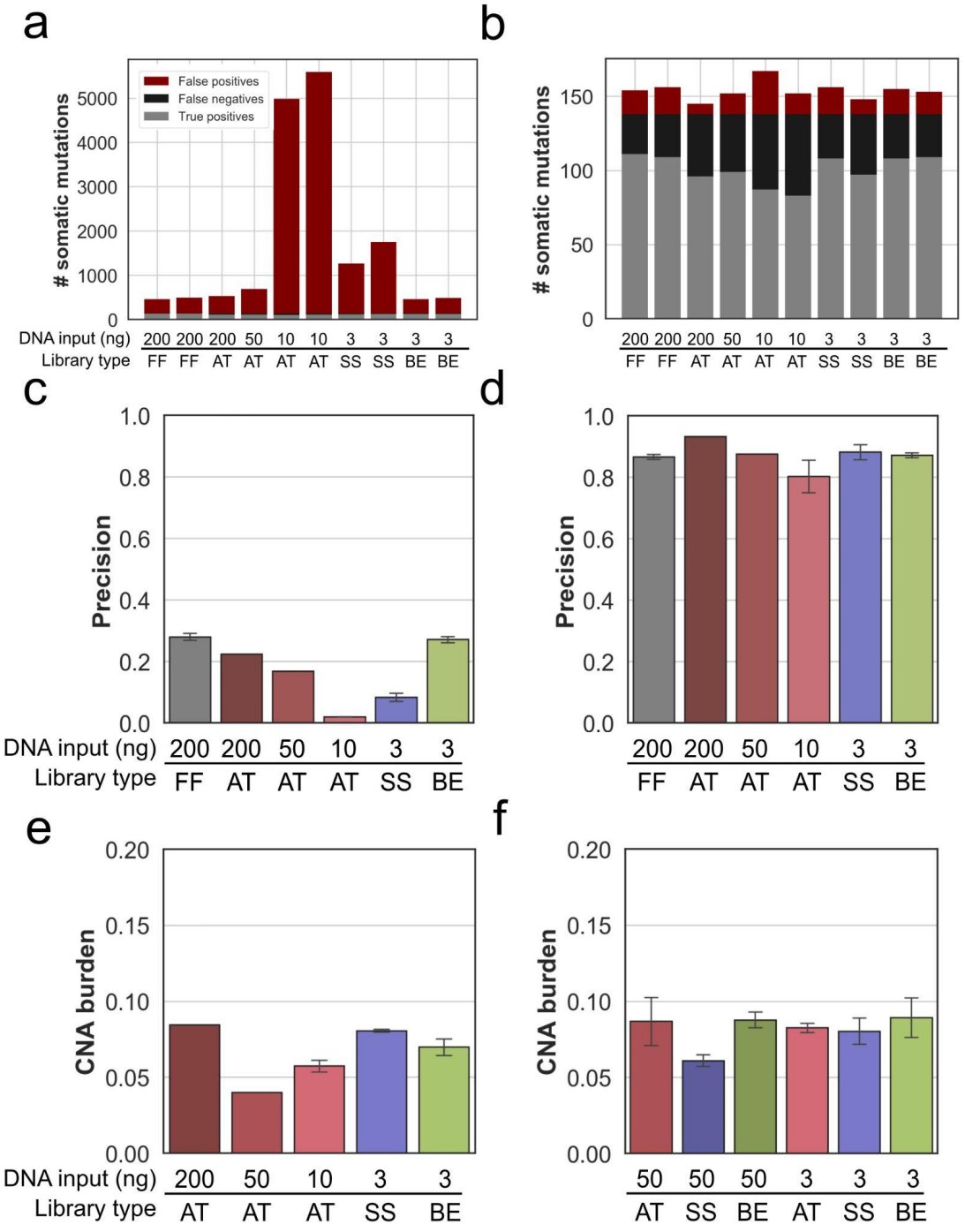


Figure 1.2. Benchmarking results for variant calling.
(a, b) Count of total variants from whole exome sequencing, separated by false positives (red), false negatives (black) and true positives (grey), before (a) and after (b) PML specific filtering. **(c, d)** Exome variant calling precision for various library preparation strategies and amount of input DNA (x-axis) before (c) and after (d) PML specific filtering. **(e, f)** Fraction of the genome involved in a copy number alteration (CNA burden—y axis) for all exome (e) and cancer panel (f) library preparation strategies and DNA input amounts. All error bars represent standard deviation.

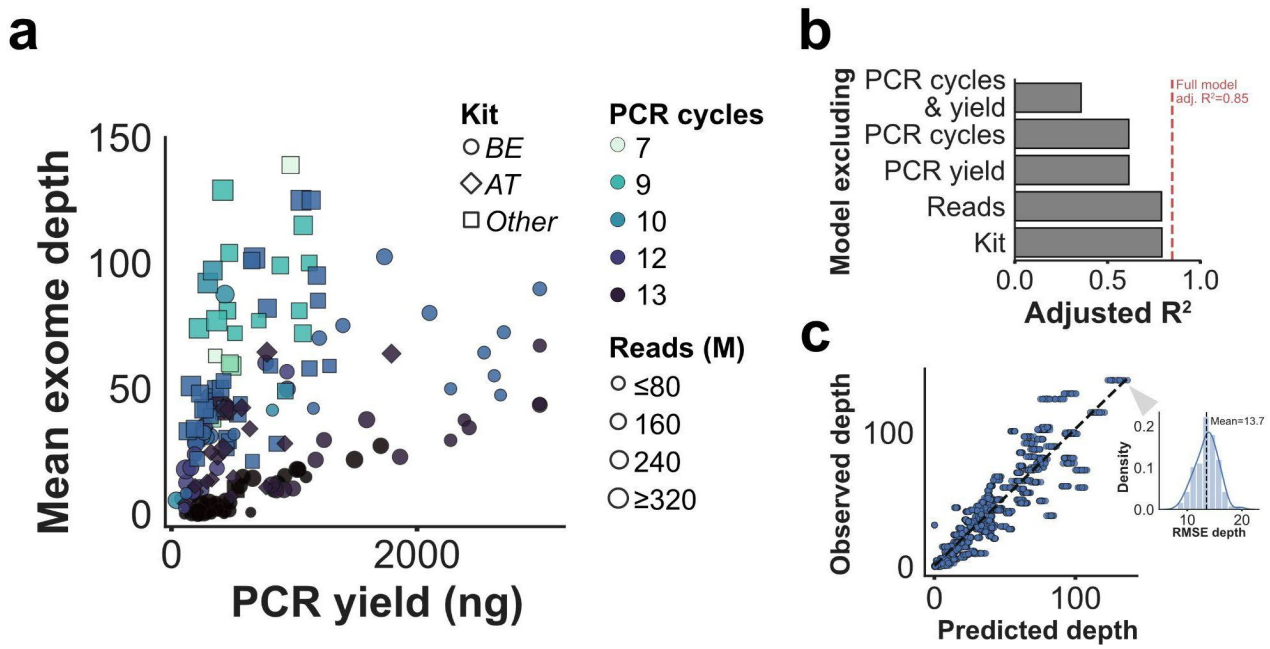


Figure 1.3. Exome sequencing depth predictive model description and evaluation.
(a) Scatterplot of all variables included in the model to predict the mean exome depth (y-axis), including PCR yield (x-axis), PCR cycles (color), sequencing reads in millions (M) (size), and library preparation kit used (shape). **(b)** Result of ordinary least squares model fit as measured by adjusted R^2 for models excluding each or a combination of the features. Vertical red dashed line indicates the fit of the model including all features. **(c)** Result of the 5-fold cross validation on the 166 FFPE samples over a hundred iterations, with a scatterplot of all predicted and observed values on the left, and density plot of the root mean squared error (RMSE) between predictions and observations on the right.

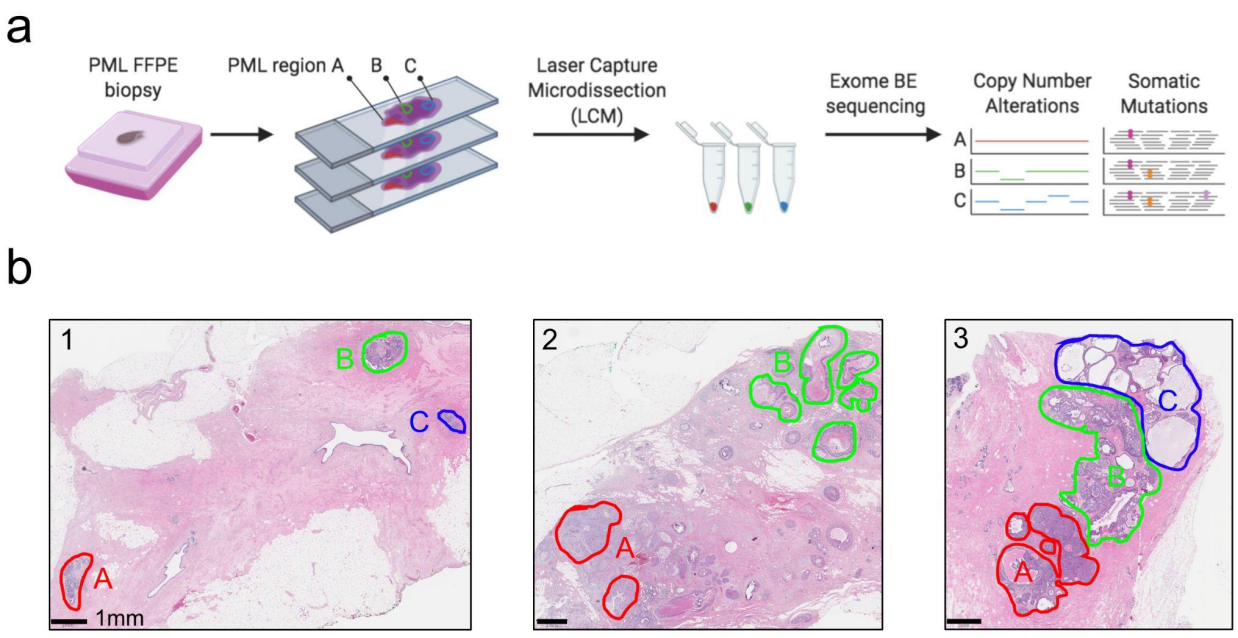


Figure 1.4. Overview of the PML regional sequencing strategy.
(a) Overall experimental and analytical workflow of the validation study. **(b)** Images showing the Hematoxylin and Eosin stained sections from the three DCIS patientspatient studied. Dissected PML regions are highlighted in color to the exception of region 3N consisting of multiple areas of normal epithelium outside the selected field of view.

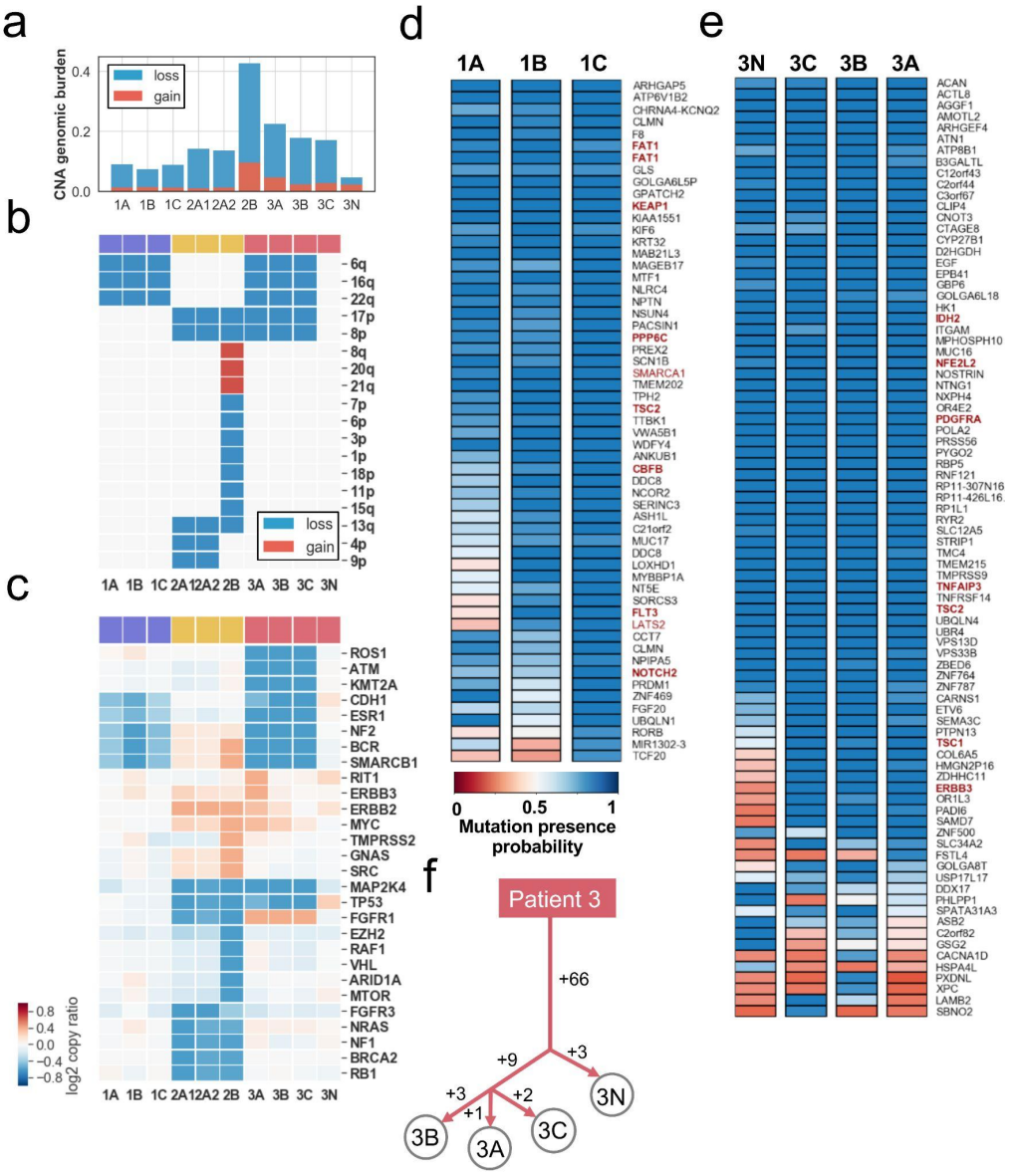


Figure 1.5. Mutational profile and clonal analysis from multi-region DNA sequencing of DCIS patients.

(a) Fraction of genome involved in copy number losses (blue) or gains (red) for each sequenced region. **(b)** Chromosome arm copy number status in each sequenced region: lost (blue) or gained (red). **(c)** Cancer gene copy number (log2 ratio—blue red gradient). Genes from the cancer panel with copy number gain (log2(ratio) > 0.4) or loss (log2(ratio) < -0.6) in at least one region, indicated with an asterisk, are displayed. **(d, e)** Bayesian probabilistic variant classification of selected high confidence somatic variants (represented by their cognate gene—rows) across all dissected regions of same patient (columns). Variants are shown for patient 1 (d) and patient 3 (e). The color gradient indicates the posterior probability of mutation presence in each region. Genes from the Cancer Gene Census are indicated in red font. **(f)** Maximum likelihood tree generated using somatic mutations identified in patient 3's regions.

1.7 Tables

Table 1.1. Result of multiple linear regression fit with ordinary least squares.

	Beta coefficient	Standard error	p>|t|
Intercept	-4.9	1.056	<0.001
Reads[1]	0.05	0.006	<0.001
PCR yield[1]	2.7	0.001	<0.001
PCR cycles[1]	-0.02	0.178	<0.001
Kit - AT[2]	2.5	0.299	<0.001

Adj. $R^2 = 0.85$
$p = 1.69 \times 10^{-52}$

1. Data are normalized thus take caution in interpreting the beta coefficient values, see methods.
2. Though the coefficient is positive for libraries prepared with AT, this is not at odds with the result showing that AT libraries produce decreased coverage from the same amount of DNA as opposed to BE libraries, as input DNA is not a parameter in this model.

1.8 Supplemental Data, Tables and Figures

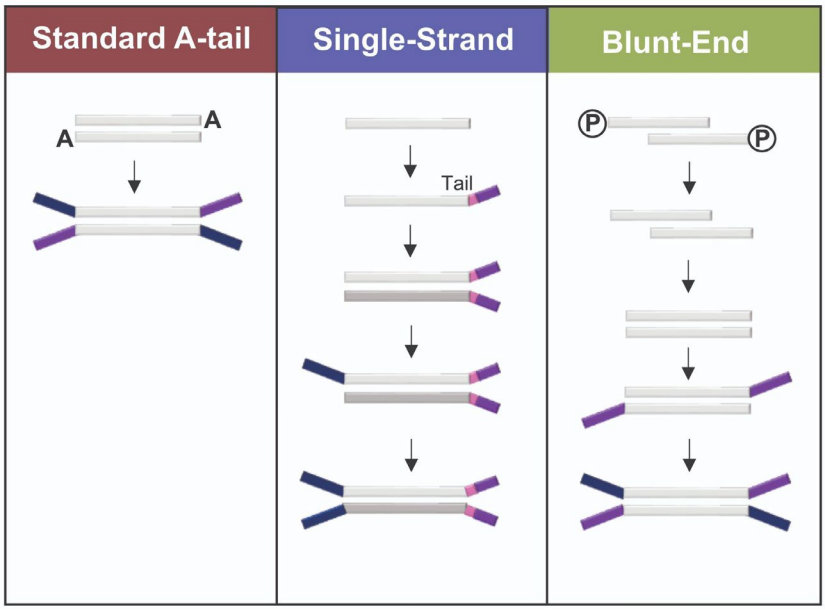


Figure S1.1. Schematic overview of adapter ligation across library preparation strategies. Purple and blue colors correspond to sequences containing Illumina P5 and P7 adapter sequences respectively.

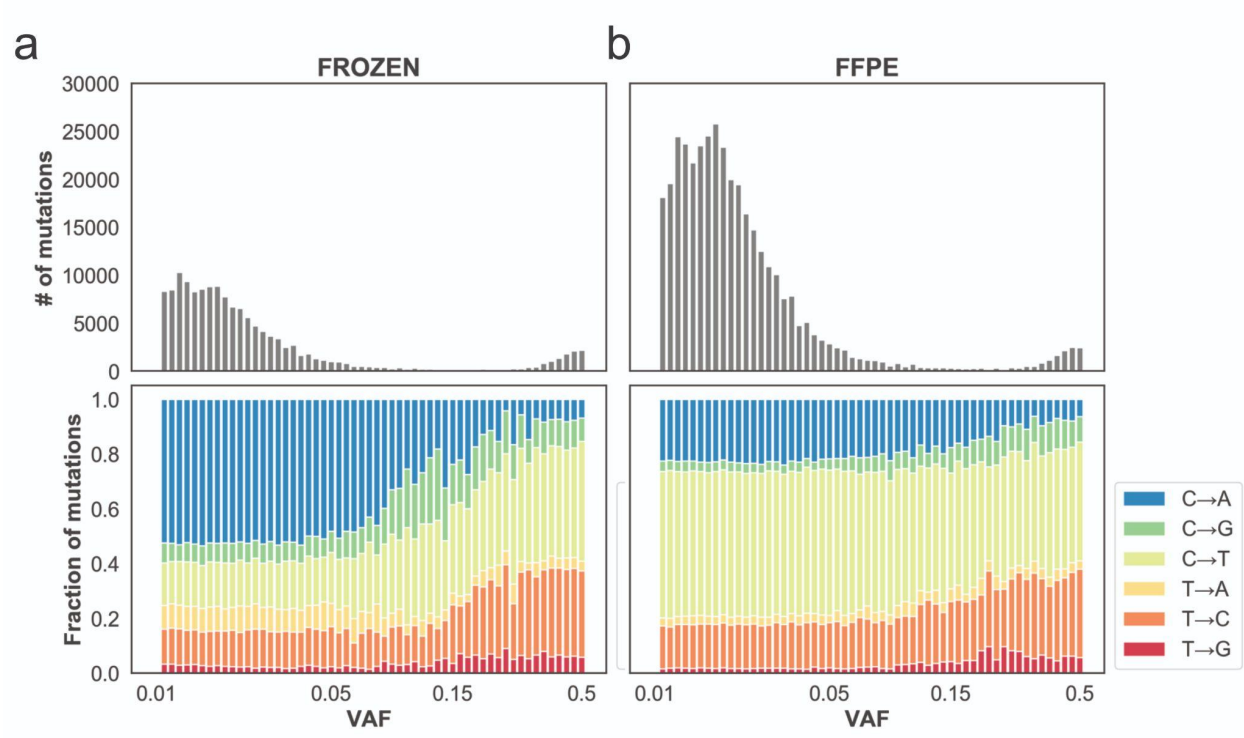


Figure S1.2. Raw nucleotide substitution evaluation.
(a, b) The number (top) and substitution pattern (bottom) at variable variant allelic fractions (VAF - x-axis) observed in test specimen (200 ng DNA input) sequenced with AT library strategy across whole exome for frozen sample (a) and mirrored FFPE (b). Raw SNV substitutions were identified from VarDict output in absence of any filters. The proportion of each substitution is shown for each VAF bin in the bottom panels.

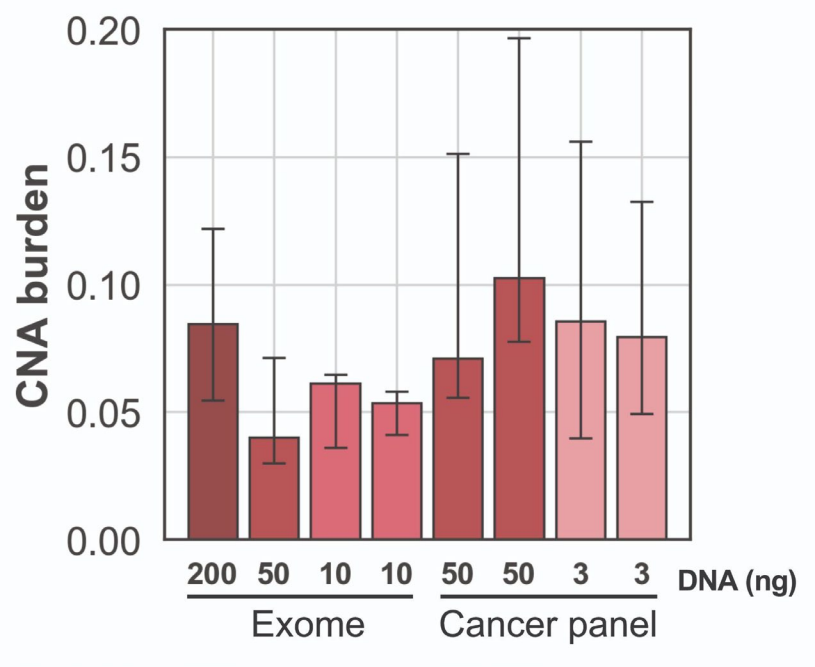


Figure S1.3. Copy number burden estimates for AT libraries with varying DNA input. Copy number alteration (CNA) burden was approximated as the fraction of genome in a CNA (-$\log_2$ copy<-0.3 or >0.3). Error bars represent CNA burden computed with either the upper or lower bounds of the 95% confidence interval around the $\log_2$ copy ratio of each segment.

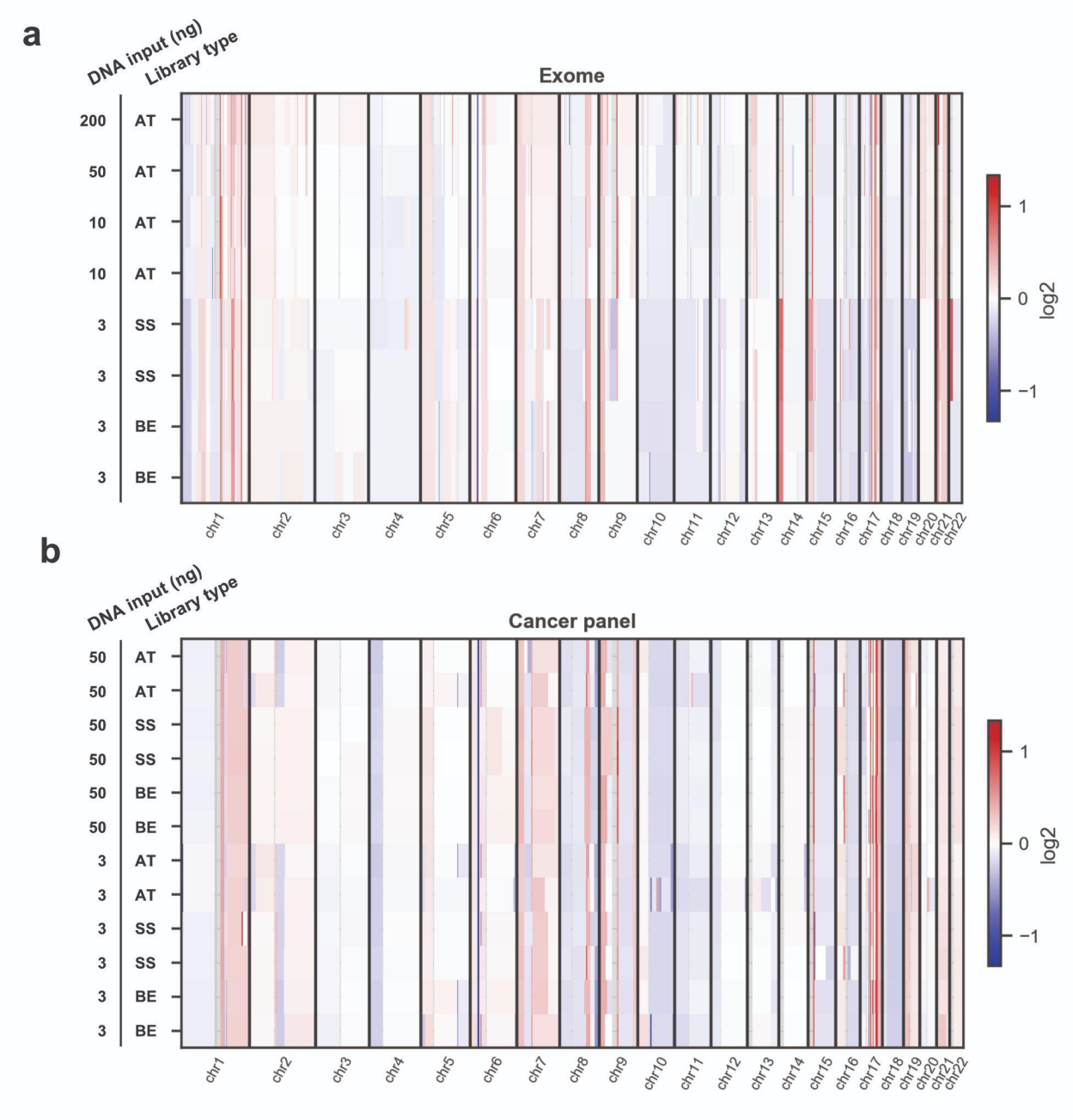


Figure S1.4. Genome wide copy number profile across library preparation strategy and DNA input amount.
Log_2 copy ratio across entire genome for samples prepared with exome **(a)** and cancer panel **(b)**.

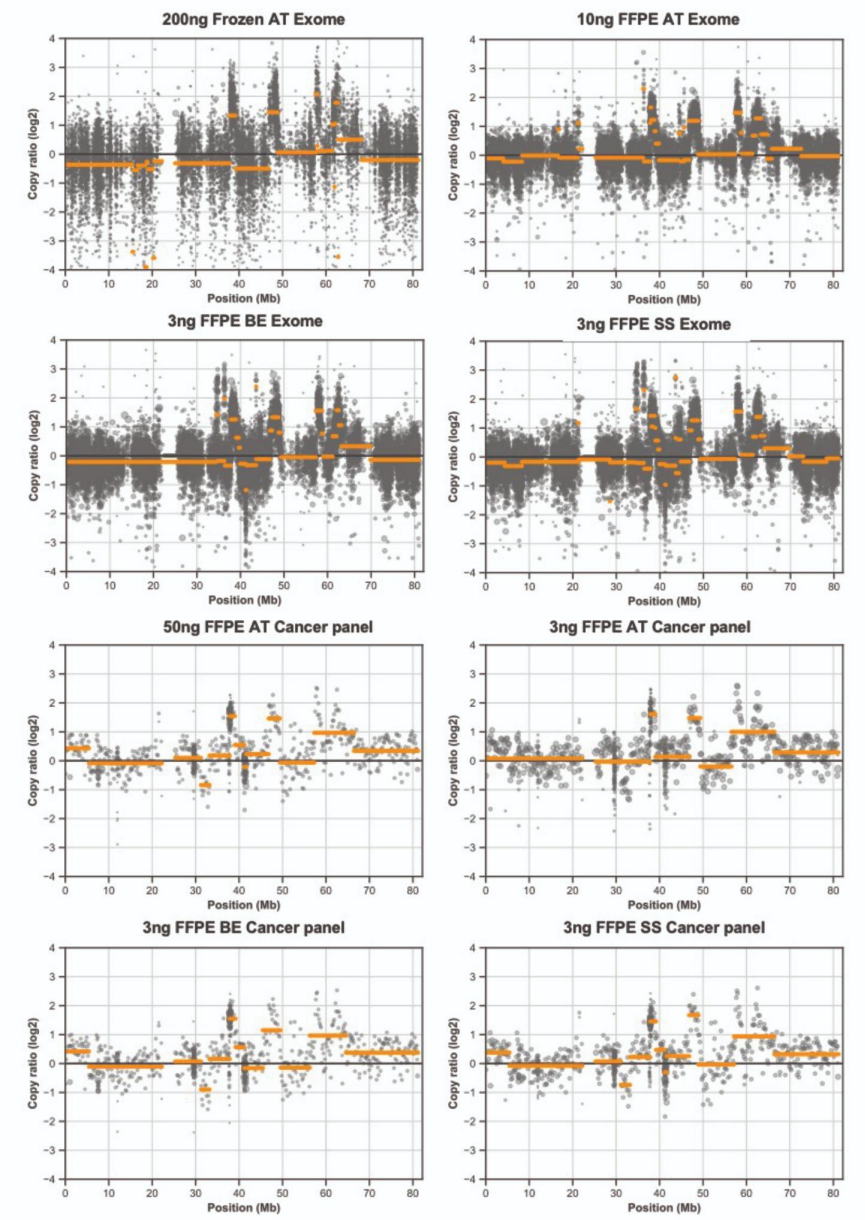


Figure S1.5. Copy number profile of chromosome 17 across library preparation strategies and DNA input amount in FFPE test specimen.
Scatter plots show log2 copy ratios for bins (grey) and segments (orange). Library preparation strategy and DNA input amount are indicated above each panel.

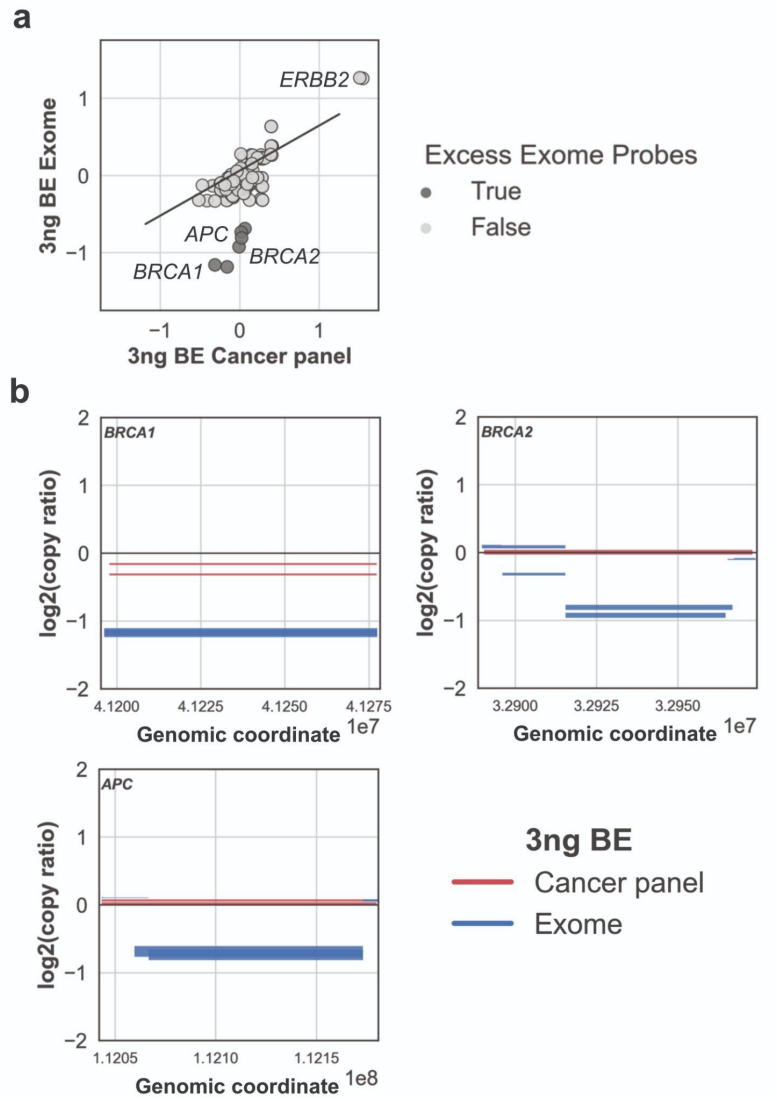


Figure S1.6. Discordant copy number ratio between exome and cancer panel in low input BE library preparation strategy.
(a) Comparison of log2 copy number ratio observed in cancer panel (x-axis) and exome (y-axis) for 98 cancer genes. Genes covered by excess number of probes (>6x) in exome as compared to the cancer panel are colored in dark gray. **(b)** Copy number levels (log2 ratio - y axis) of genomic segments overlapping genes with discordant copy numbers (BRCA1, BRCA2, APC) between exome (blue) and cancer panel (red). All experimental replicates are displayed. Line thickness indicates the confidence level (thick=high) of the segment called.

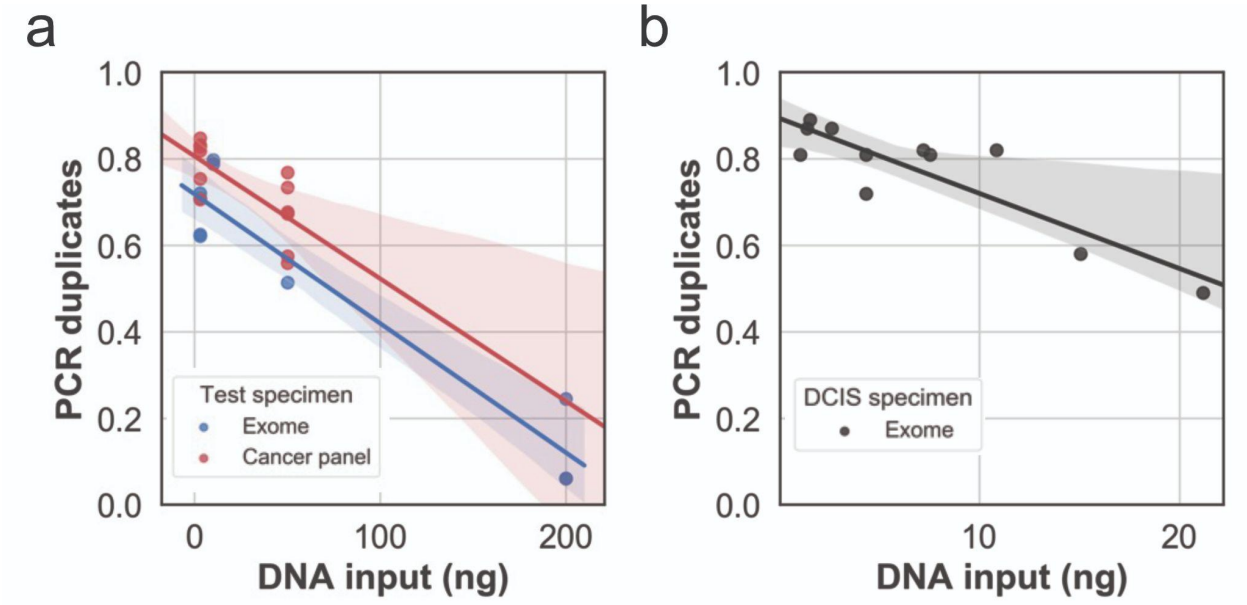


Figure S1.7. PCR duplicate rate as a function of DNA input amount.
Fraction of PCR duplicate reads on y-axis as a function of DNA input on x-axis (p = 1.4e-05) in test FFPE specimen **(a)** and for DCIS specimen (p = 3e-03) **(b)**.

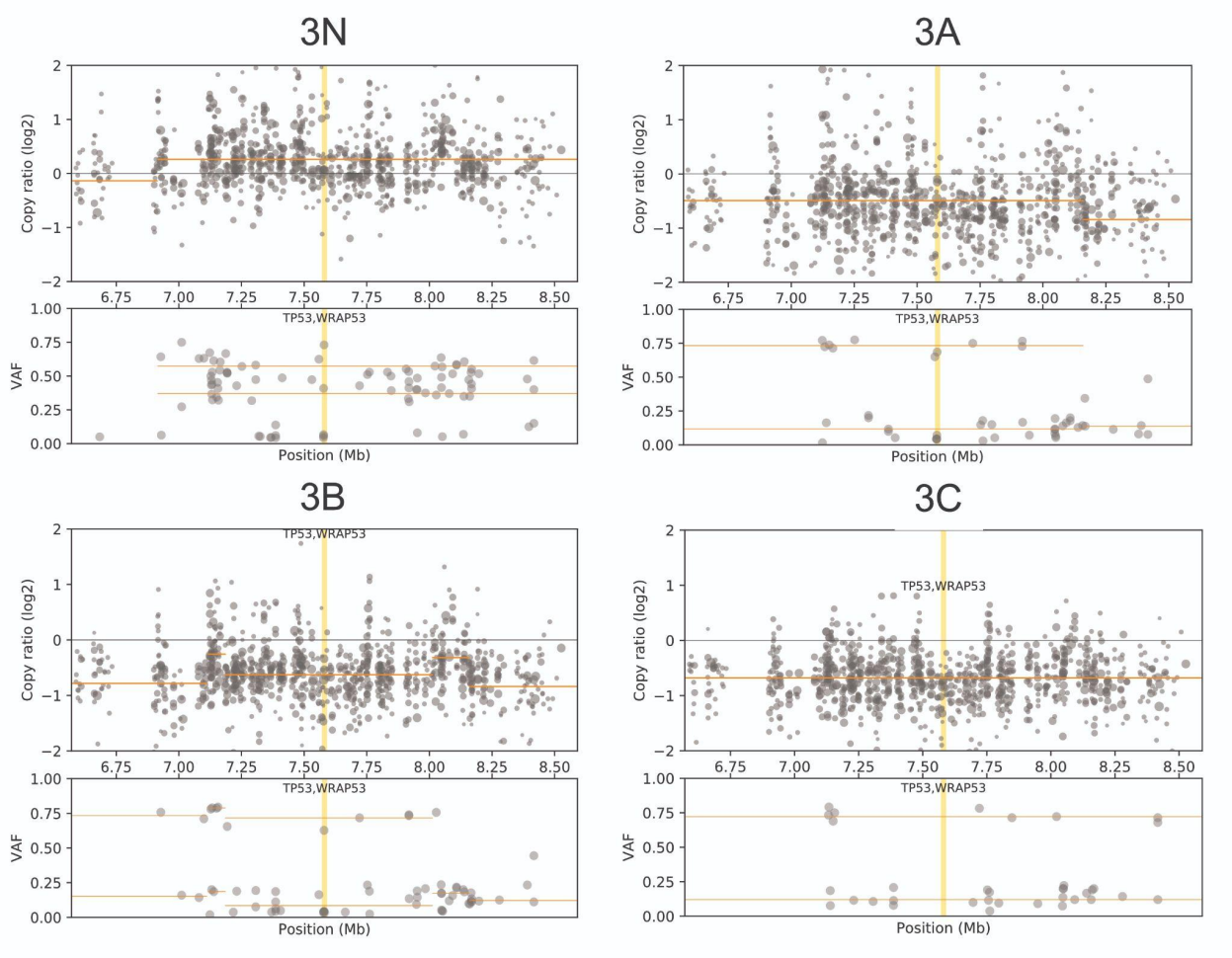


Figure S1.8. *TP53* LOH in regions of patient 3.
Scatter plot of log2 copy number ratio (upper panels) and B-allele frequency (lower panels) of coverage bins (grey dots) and resulting genomic segments (orange lines) in a window of 2 Mb around *TP53* (yellow stripe).

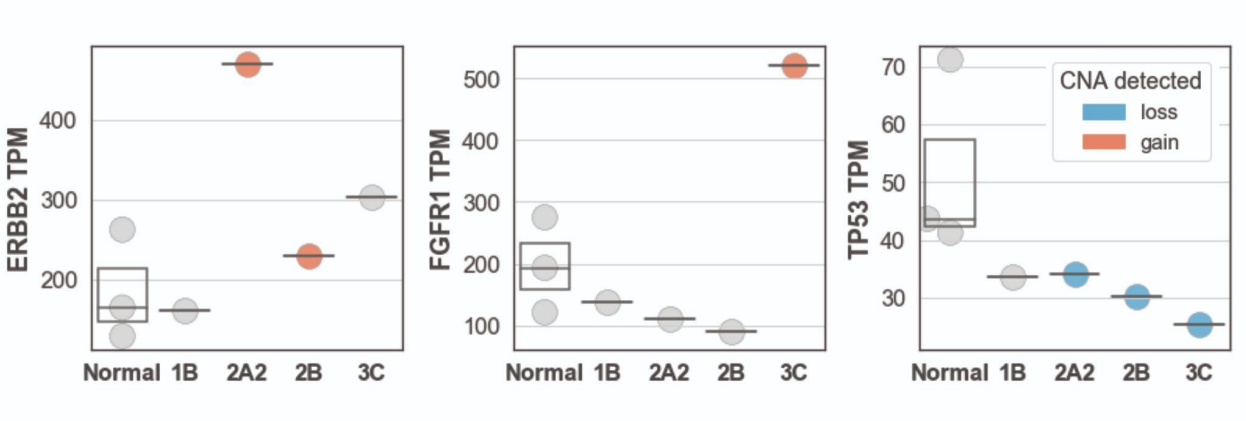


Figure S1.9. Expression level of selected genes affected with CNA.
Normalized read counts, transcripts per million (TPM), from SMART-3Seq expression profiling of select DCIS regions for *ERBB2* (left), *FGFR1* (middle) and *TP53* (right), and as compared to unrelated normal dissected breast epithelium. Specimen in which a gene with a copy number gain was detected are shown in red, copy number loss in blue and, copy neutral in gray.

1.9 Author Contributions

ADB and FH selected specimen and performed pathology annotation, EJ and TOK identified breast cancer patients and reviewed associated clinical data. JS, KJ and HY performed the targeted sequencing. KJ, DN, OH designed the test study, OH, GH, LE, ADB designed the validation study. AO performed the RNA-seq analysis. DN and OH performed the analysis and wrote the manuscript. All authors read and approved the final manuscript.

1.10 Acknowledgements

We express our gratitude to members of the Moores Cancer Center Biotechnology and Tissue Technology Shared Resource: Drs Molinolo, Kaushal, Estrada and Kimberly McIntyre for their technical assistance. All sequencing was conducted at the IGM Genomics Center, University of California, San Diego, La Jolla, CA. Figure 3a was created with BioRender.com and printed with permission.

Chapter 1, in full, is a reformatted reprint of the material as it appears as "Mutational profiling of micro-dissected pre-malignant lesions from archived specimens" in *BMC Medical Genomics*, 2020 Chapter 1, is a reformatted reprint of the material as it appears as "Mutational profiling of micro-dissected pre-malignant lesions from archived specimens" in *BMC Medical Genomics*, 2020 by Daniela Nachmanson, Joseph Steward, Huazhen Yao, Adam Officer, Eliza Jeong, Thomas J. O'Keefe, Farnaz Hasteh, Kristen Jepsen, Gillian L. Hirst, Laura J. Esserman, Alexander D. Borowsky and Olivier Harismendy. The dissertation author was the primary investigator and author of this paper.

1.11 References

1. Esserman, L. J., Thompson, I. M., Jr & Reid, B. Overdiagnosis and overtreatment in cancer: an opportunity for improvement. *JAMA* **310,** 797–798 (2013).

2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70,** 7–30 (2020).

3. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* **380,** 1778–1786 (2012).

4. Bleyer, A. & Welch, H. G. Effect of three decades of screening mammography on breast-cancer incidence. *N. Engl. J. Med.* **367,** 1998–2005 (2012).

5. Menon, R., Deng, M., Boehm, D., Braun, M., Fend, F., Boehm, D., Biskup, S. & Perner, S. Exome enrichment and SOLiD sequencing of formalin fixed paraffin embedded (FFPE) prostate cancer tissue. *Int. J. Mol. Sci.* **13,** 8933–8942 (2012).

6. Hedegaard, J., Thorsen, K., Lund, M. K., Hein, A.-M. K., Hamilton-Dutoit, S. J., Vang, S., Nordentoft, I., Birkenkamp-Demtröder, K., Kruhøffer, M., Hager, H., Knudsen, B., Andersen, C. L., Sørensen, K. D., Pedersen, J. S., Ørntoft, T. F. & Dyrskjøt, L. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS One* **9,** e98187 (2014).

7. Van Allen, E. M., Wagle, N., Stojanov, P., Perrin, D. L., Cibulskis, K., Marlow, S., Jane-Valbuena, J., Friedrich, D. C., Kryukov, G., Carter, S. L., McKenna, A., Sivachenko, A., Rosenberg, M., Kiezun, A., Voet, D., Lawrence, M., Lichtenstein, L. T., Gentry, J. G., Huang, F. W., Fostel, J., Farlow, D., Barbie, D., Gandhi, L., Lander, E. S., Gray, S. W., Joffe, S., Janne, P., Garber, J., MacConaill, L., Lindeman, N., Rollins, B., Kantoff, P., Fisher, S. A., Gabriel, S., Getz, G. & Garraway, L. A. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **20,** 682–688 (2014).

8. Munchel, S., Hoang, Y., Zhao, Y., Cottrell, J., Klotzle, B., Godwin, A. K., Koestler, D., Beyerlein, P., Fan, J.-B., Bibikova, M. & Chien, J. Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. *Oncotarget* **6,** 25943–25961 (2015).

9. Astolfi, A., Urbini, M., Indio, V., Nannini, M., Genovese, C. G., Santini, D., Saponara, M., Mandrioli, A., Ercolani, G., Brandi, G., Biasco, G. & Pantaleo, M. A. Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST). *BMC Genomics* **16,** 892 (2015).

10. Miron, A., Varadi, M., Carrasco, D., Li, H., Luongo, L., Kim, H. J., Park, S. Y., Cho, E. Y., Lewis, G., Kehoe, S., Iglehart, J. D., Dillon, D., Allred, D. C., Macconaill, L., Gelman, R. & Polyak, K. PIK3CA mutations in in situ and invasive breast carcinomas. *Cancer Res.* **70,** 5674–5678 (2010).

11. Sontag, L. & Axelrod, D. E. Evaluation of pathways for progression of heterogeneous breast tumors. *J. Theor. Biol.* **232,** 179–189 (2005).

12. Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lønning, P. E., Wedge, D. C. & Campbell, P. J. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21,** 751–759 (2015).

13. Newburger, D. E., Kashef-Haghighi, D., Weng, Z., Salari, R., Sweeney, R. T., Brunner, A. L., Zhu, S. X., Guo, X., Varma, S., Troxell, M. L., West, R. B., Batzoglou, S. & Sidow, A. Genome evolution during progression to breast cancer. *Genome Res.* **23,** 1097–1108 (2013).

14. Oikawa, M., Yano, H., Matsumoto, M., Otsubo, R., Shibata, K., Hayashi, T., Abe, K., Kinoshita, N., Yoshiura, K.-I. & Nagayasu, T. A novel diagnostic method targeting genomic instability in intracystic tumors of the breast. *Breast Cancer* **22,** 529–535 (2015).

15. Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M. E. & Navin, N. E. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172,** 205–217.e12 (2018).

16. Arreaza, G., Qiu, P., Pang, L., Albright, A., Hong, L. Z., Marton, M. J. & Levitan, D. Pre-Analytical Considerations for Successful Next-Generation Sequencing (NGS): Challenges and Opportunities for Formalin-Fixed and Paraffin-Embedded Tumor Tissue (FFPE) Samples. *Int. J. Mol. Sci.* **17,** (2016).

17. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61,** 64–71 (2015).

18. Levy, E., Marty, R., Gárate Calderón, V., Woo, B., Dow, M., Armisen, R., Carter, H. & Harismendy, O. Immune DNA signature of T-cell infiltration in breast tumor exomes. *Sci. Rep.* **6,** 30064 (2016).

19. Troll, C. J., Kapp, J., Rao, V., Harkins, K. M., Cole, C., Naughton, C., Morgan, J. M., Shapiro, B. & Green, R. E. A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos. *BMC Genomics* **20,** 1023 (2019).

20. Bennett, C. W., Berchem, G., Kim, Y. J. & El-Khoury, V. Cell-free DNA and next-generation sequencing in the service of personalized medicine for lung cancer. *Oncotarget* **7,** 71013–71035 (2016).

21. Bennett, E. A., Massilani, D., Lizzo, G., Daligault, J., Geigl, E.-M. & Grange, T. Library construction for ancient genomics: single strand or double strand? *Biotechniques* **56,** 289–90, 292–6, 298, passim (2014).

22. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L. & Meyer, M. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15758–15763 (2013).

23. Gansauge, M.-T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **8,** 737–748 (2013).

24. Sundaram, A. Y. M., Hughes, T., Biondi, S., Bolduc, N., Bowman, S. K., Camilli, A., Chew, Y. C., Couture, C., Farmer, A., Jerome, J. P., Lazinski, D. W., McUsic, A., Peng, X., Shazand, K., Xu, F., Lyle, R. & Gilfillan, G. D. A comparative study of ChIP-seq sequencing library preparation methods. *BMC Genomics* **17,** 816 (2016).

25. Ma, S., Hsieh, Y.-P., Ma, J. & Lu, C. Low-input and multiplexed microfluidic assay reveals epigenomic variation across cerebellum and prefrontal cortex. *Sci Adv* **4,** eaar8187 (2018).

26. Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., Fostel, J. L., Friedrich, D. C., Perrin, D., Dionne, D., Kim, S., Gabriel, S. B., Lander, E. S., Fisher, S. & Getz, G. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).

27. Kader, T., Goode, D. L., Wong, S. Q., Connaughton, J., Rowley, S. M., Devereux, L., Byrne, D., Fox, S. B., Mir Arnau, G., Tothill, R. W., Campbell, I. G. & Gorringe, K. L. Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue. *Genome Med.* **8,** 121 (2016).

28. Rieber, N., Bohnert, R., Ziehm, U. & Jansen, G. Reliability of algorithmic somatic copy number alteration detection from targeted capture data. *Bioinformatics* **33,** 2791–2798 (2017).

29. Reis-Filho, J. S. & Lakhani, S. R. The diagnosis and management of pre-invasive breast disease: genetic alterations in pre-invasive lesions. *Breast Cancer Res.* **5,** 313–319 (2003).

30. Foley, J. W., Zhu, C., Jolivet, P., Zhu, S. X., Lu, P., Meaney, M. J. & West, R. B. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome Res.* **29,** 1816–1825 (2019).

31. Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).

32. Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C. A., Vogelstein, B. & Nowak, M. A. Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8,** 14114 (2017).

33. Jayasinghe, R. G., Cao, S., Gao, Q., Wendl, M. C., Vo, N. S., Reynolds, S. M., Zhao, Y., Climente-González, H., Chai, S., Wang, F., Varghese, R., Huang, M., Liang, W.-W.,

Wyczalkowski, M. A., Sengupta, S., Li, Z., Payne, S. H., Fenyö, D., Miner, J. H., Walter, M. J., Cancer Genome Atlas Research Network, Vincent, B., Eyras, E., Chen, K., Shmulevich, I., Chen, F. & Ding, L. Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep.* **23,** 270–281.e3 (2018).

34. Pang, J.-M. B., Savas, P., Fellowes, A. P., Mir Arnau, G., Kader, T., Vedururu, R., Hewitt, C., Takano, E. A., Byrne, D. J., Choong, D. Y., Millar, E. K., Lee, C. S., O'Toole, S. A., Lakhani, S. R., Cummings, M. C., Mann, G. B., Campbell, I. G., Dobrovic, A., Loi, S., Gorringe, K. L. & Fox, S. B. Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. *Mod. Pathol.* **30,** 952–963 (2017).

35. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H. & Campbell, P. J. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).

36. Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R. J., Abascal, F., Hall, M. W. J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., Fitzgerald, R. C., Handford, P. A., Campbell, P. J., Saeb-Parsy, K. & Jones, P. H. Somatic mutant clones colonize the human esophagus with age. *Science* **362,** 911–917 (2018).

37. Salk, J. J., Loubet-Senear, K., Maritschnegg, E., Valentine, C. C., Williams, L. N., Higgins, J. E., Horvat, R., Vanderstichele, A., Nachmanson, D., Baker, K. T., Emond, M. J., Loter, E., Tretiakova, M., Soussi, T., Loeb, L. A., Zeillinger, R., Speiser, P. & Risques, R. A. Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Rep.* **28,** 132–144.e3 (2019).

38. Bose, R., Kavuri, S. M., Searleman, A. C., Shen, W., Shen, D., Koboldt, D. C., Monsey, J., Goel, N., Aronson, A. B., Li, S., Ma, C. X., Ding, L., Mardis, E. R. & Ellis, M. J. Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discov.* **3,** 224–237 (2013).

39. Mujoo, K., Choi, B.-K., Huang, Z., Zhang, N. & An, Z. Regulation of ERBB3/HER3 signaling in cancer. *Oncotarget* **5,** 10222–10236 (2014).

40. Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. A., Stamp, G., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Futreal, P. A., Stewart, A. & Swanton, C. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46,** 225–233 (2014).

41. Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H. & Turner, D. J. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5,** 1005–1010 (2008).

42. Jones, S., Anagnostou, V., Lytle, K., Parpart-Li, S., Nesselbush, M., Riley, D. R., Shukla, M., Chesnick, B., Kadan, M., Papp, E., Galens, K. G., Murphy, D., Zhang, T., Kann, L., Sausen, M., Angiuoli, S. V., Diaz, L. A., Jr & Velculescu, V. E. Personalized genomic

analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7,** 283ra53 (2015).

43. Guimera, R. V. bcbio-nextgen: Automated, distributed next-gen sequencing pipeline. *EMBnet.journal* **17,** 30 (2011).

44. Kalatskaya, I., Trinh, Q. M., Spears, M., McPherson, J. D., Bartlett, J. M. S. & Stein, L. ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. *Genome Med.* **9,** 59 (2017).

45. Garofalo, A., Sholl, L., Reardon, B., Taylor-Weiner, A., Amin-Mansour, A., Miao, D., Liu, D., Oliver, N., MacConaill, L., Ducar, M., Rojas-Rudilla, V., Giannakis, M., Ghazani, A., Gray, S., Janne, P., Garber, J., Joffe, S., Lindeman, N., Wagle, N., Garraway, L. A. & Van Allen, E. M. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med.* **8,** 79 (2016).

46. Chen, Z., Yuan, Y., Chen, X., Chen, J., Lin, S., Li, X. & Du, H. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci. Rep.* **10,** 3501 (2020).

47. Anzar, I., Sverchkova, A., Stratford, R. & Clancy, T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics* **12,** 63 (2019).

48. Yost, S. E., Smith, E. N., Schwab, R. B., Bao, L., Jung, H., Wang, X., Voest, E., Pierce, J. P., Messer, K., Parker, B. A., Harismendy, O. & Frazer, K. A. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40,** e107 (2012).

49. Kim, H., Lee, A. J., Lee, J., Chun, H., Ju, Y. S. & Hong, D. FIREVAT: finding reliable variants without artifacts in human cancer samples using etiologically relevant mutational signatures. *Genome Med.* **11,** 81 (2019).

50. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Pääbo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29,** 4793–4799 (2001).

51. Tchou, J. & Grollman, A. P. Repair of DNA containing the oxidatively-damaged base, 8-oxoguanine. *Mutat. Res.* **299,** 277–287 (1993).

52. Chen, L., Liu, P., Evans, T. C., Jr & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355,** 752–756 (2017).

53. Kim, S. Y., Jung, S.-H., Kim, M. S., Baek, I.-P., Lee, S. H., Kim, T.-M., Chung, Y.-J. & Lee, S. H. Genomic differences between pure ductal carcinoma in situ and synchronous ductal carcinoma in situ with invasive breast cancer. *Oncotarget* **6,** 7597–7607 (2015).

54. Pareja, F., Brown, D. N., Lee, J. Y., Da Cruz Paula, A., Selenica, P., Bi, R., Geyer, F. C., Gazzo, A., da Silva, E. M., Vahdatinia, M., Stylianou, A. A., Ferrando, L., Wen, H. Y., Hicks,

J. B., Weigelt, B. & Reis-Filho, J. S. Whole-Exome Sequencing Analysis of the Progression from Non-Low-Grade Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. *Clin. Cancer Res.* **26,** 3682–3693 (2020).

55. Sinha, V. C. & Piwnica-Worms, H. Intratumoral Heterogeneity in Ductal Carcinoma In Situ: Chaos and Consequence. *J. Mammary Gland Biol. Neoplasia* **23,** 191–205 (2018).

56. Gerdes, M. J., Gökmen-Polar, Y., Sui, Y., Pang, A. S., LaPlante, N., Harris, A. L., Tan, P.-H., Ginty, F. & Badve, S. S. Single-cell heterogeneity in ductal carcinoma in situ of breast. *Mod. Pathol.* **31,** 406–417 (2018).

57. Chin, K., de Solorzano, C. O., Knowles, D., Jones, A., Chou, W., Rodriguez, E. G., Kuo, W.-L., Ljung, B.-M., Chew, K., Myambo, K., Miranda, M., Krig, S., Garbe, J., Stampfer, M., Yaswen, P., Gray, J. W. & Lockett, S. J. In situ analyses of genome instability in breast cancer. *Nat. Genet.* **36,** 984–988 (2004).

58. Foschini, M. P., Morandi, L., Leonardi, E., Flamminio, F., Ishikawa, Y., Masetti, R. & Eusebi, V. Genetic clonal mapping of in situ and invasive ductal carcinoma indicates the field cancerization phenomenon in the breast. *Hum. Pathol.* **44,** 1310–1319 (2013).

59. Martelotto, L. G., Baslan, T., Kendall, J., Geyer, F. C., Burke, K. A., Spraggon, L., Piscuoglio, S., Chadalavada, K., Nanjangud, G., Ng, C. K. Y., Moody, P., D'Italia, S., Rodgers, L., Cox, H., da Cruz Paula, A., Stepansky, A., Schizas, M., Wen, H. Y., King, T. A., Norton, L., Weigelt, B., Hicks, J. B. & Reis-Filho, J. S. Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat. Med.* **23,** 376–385 (2017).

60. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24,** 52–60 (2014).

61. Srivastava, S., Ghosh, S., Kagan, J., Mazurchuk, R. & National Cancer Institute's HTAN Implementation. The Making of a PreCancer Atlas: Promises, Challenges, and Opportunities. *Trends Cancer Res.* **4,** 523–536 (2018).

62. bcl2fastq. at <https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html>

63. Li, H. seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub* **767,** 69 (2012).

64. Didion, J. P., Martin, M. & Collins, F. S. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5,** e3720 (2017).

65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

66. Institute, B. Picard tools. (2016).

67. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.*

12, e1004873 (2016).

68. Olshen, A. B., Bengtsson, H., Neuvial, P., Spellman, P. T., Olshen, R. A. & Seshan, V. E. Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* **27,** 2038–2046 (2011).

69. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. C. & Dry, J. R. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44,** e108 (2016).

70. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C. & Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054 (2019). doi:10.1101/861054

71. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6: 80--92. (2012).

72. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. & Abecasis, G. R. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

73. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G. & Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

74. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., The Genome Aggregation Database Consortium, Neale, B. M., Daly, M. J. & MacArthur, D. G. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human

protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210

75. Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T. & Campbell, P. J. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45,** D777–D783 (2017).

76. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L. & Maglott, D. R. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46,** D1062–D1067 (2018).

77. Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Nohzadeh-Malakshah, S., Rathod, M., Ware, D., Trigg, L. & De La Vega, F. M. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol.* **21,** 405–419 (2014).

78. Gutkind, J. S., Molinolo, A. A., Wu, X., Wang, Z., Nachmanson, D., Harismendy, O., Alexandrov, L. B., Wuertz, B. R., Ondrey, F. G., Laronde, D., Rock, L. D., Rosin, M., Coffey, C., Butler, V. D., Bengtson, L., Hsu, C.-H., Bauman, J. E., Hewitt, S. M., Cohen, E. E., Chow, H.-H. S., Lippman, S. M. & Szabo, E. Inhibition of mTOR signaling and clinical activity of metformin in oral premalignant lesions. *JCI Insight* **6,** (2021).

79. Nachmanson, D., Officer, A., Mori, H., Gordon, J., Evans, M. F., Steward, J., Yao, H., O'Keefe, T., Hasteh, F., Stein, G. S., Jepsen, K., Weaver, D. L., Hirst, G. L., Sprague, B. L., Esserman, L. J., Borowsky, A. D., Stein, J. L. & Harismendy, O. The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. *NPJ Breast Cancer* **8,** 6 (2022).

CHAPTER 2: The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ

2.1 Abstract

Micro-environmental and molecular factors mediating the progression of Breast Ductal Carcinoma In Situ (DCIS) are not well understood, impeding the development of prevention strategies and the safe testing of treatment de-escalation. We addressed methodological barriers and characterized the mutational, transcriptional, histological and microenvironmental landscape across 85 multiple micro-dissected regions from 39 cases. Most somatic alterations, including whole genome duplications, were clonal, but genetic divergence increased with physical distance. Phenotypic and subtype heterogeneity was frequently associated with underlying genetic heterogeneity and regions with low-risk features preceded those with high-risk features according to the inferred phylogeny. B- and T-lymphocytes spatial analysis identified 3 immune states, including an epithelial excluded state located preferentially at DCIS regions, and characterized by histological and molecular features of immune escape, independently from molecular subtypes. Such breast pre-cancer atlas with uniquely integrated observations will help scope future expansion studies and build finer models of outcomes and progression risk.

2.2 Introduction

Increasing adoption of breast cancer screening and advances in imaging capabilities have improved our ability to identify breast ductal carcinoma in situ (DCIS). Rarely diagnosed 40 years ago, DCIS now comprises nearly 20% of all breast cancer-related diagnoses [1,2]. Unfortunately, this progress has not resulted in decreased breast cancer mortality. Standard treatment, involving surgical excision often complemented with radiation therapy (in the setting of breast conserving surgery) and endocrine recurrence risk reduction (particularly with ER+ DCIS), therefore constitutes overtreatment, and not without treatment-related consequences for many [2,3]. DCIS progression is particularly difficult to study longitudinally due to the current standard of surgical excision of the lesion and the infrequent progression and/or occurrence of new primary lesions over a long timespan (5-10% after 10 years) [4]. Clinicopathological risk factors such as large size, dense breast, younger age, high pathological grade, presence of comedo-necrosis or Her2 positivity have been associated with increased risk of recurrence, but the resulting predictive models, or those relying on gene expression signatures, are currently insufficient to safely distinguish patients to watch from patients to treat [5].

Contrary to models of progression in other tissue types, there is little evidence for the sequential accumulation of somatic alterations during progression from in situ to invasive breast cancer (IBC), but rather all IBC intrinsic subtypes and known driver mutations have been identified in DCIS, albeit at variable prevalence [6–12]. Moreover, both single-cell and bulk studies have shown similar clonal make-up of synchronous invasive and in situ lesions, convoluting the idea that clonal selection drives invasion [11,13]. The role of the immune environment has also been investigated, highlighting the higher lymphocyte infiltration in Her2+ or Triple Negative DCIS, or specific immunological make-up of samples at higher risk of progression [12,14–19]. Similarly, the role of the

basal layer, fibroblasts, adipocytes, other stromal cells or overall extracellular matrix has identified features that are different between DCIS and IBC, likely mediated by chemokine signaling and can be associated with known progression risk factors [20–23]. Their active participation in the malignant transformation of the breast epithelium remains to be established as similar mechanisms are typically involved in normal development, activity and aging of the mammary gland [12,24].

Progress in our understanding of the processes mediating DCIS onset and progression has been considerably hindered by technical and logistical limitations. Indeed, pure DCIS lesions are commonly small in size, formalin and paraffin embedded (which damages nucleic acids) and can display significant histological heterogeneity [25]. As a consequence, comprehensive molecular and cellular assays and their integrated analysis have seldom been performed in pure DCIS cohorts. Capturing evidence of phenotypic, genetic and cellular heterogeneity, and how they relate to each other is necessary to develop a better spatial, temporal and functional understanding of the mechanisms at play. Recent advances in genome-wide assays, becoming compatible with ever more challenging samples [26–29], have improved our ability to connect histological and molecular observations and enabled such application even to individual microbiopsies from a histological slide of pure DCIS.

Here we describe the combined, parallel histological, molecular and immunological profiling of pre-malignant lesions from 39 patients diagnosed with DCIS, including multiple epithelial micro-biopsies within a subset of samples. The dissection of specific epithelial lesions provided a detailed assessment of the association of their histological architecture with intrinsic subtypes, mutational landscape, driver mutations and immunological states. Multi-region profiling resulted in the inference of clonal relationships, illustrating how genotypes related to phenotypes within a specimen. We therefore report a multi-modal and sub-histological profiling of a cohort of

pure DCIS, illustrating spatial heterogeneity and placing diverse states of immune-activity observed in their specific molecular and histological context.

2.3 Results

Histological and molecular characterization.

We collected a total of 43 specimens (referred to as samples) from 39 patients diagnosed with pure DCIS, including three samples from subsequent DCIS diagnosed between 14 and 70 months after the index DCIS (Figure 2.1a-b, Table 2.1, Supplementary Table 2.1). Sixty-nine percent (29/42) of the samples were positive for estrogen receptor (ER) expression and 40% (16/40) had ERBB2 gene overexpression or amplification (Supplementary Figure 2.1a). Each sample was further annotated for grade and histological architecture and the annotations were used to identify regions of interest, guide the micro-biopsies of the epithelial areas and the immuno-histological analysis. On the basis of their studied regions, the cohort consisted of 32 high or intermediate grade DCIS (HG-DCIS), 9 low-grade DCIS (LG-DCIS) and 2 low-grade atypical ductal hyperplasia (ADH). The DCIS regions could be further annotated according to their dominant histological architecture (17 cribriform, 19 solid, 3 mixed, 2 micropapillary) and the presence of necrosis (10 comedo-necrosis, 17 other). LG-DCIS were more frequently of cribriform architecture (8/9), while HG-DCIS were frequently necrotic (25/32). The relative area of adipose tissue in each sample varied between 4 and 91 percent as estimated by segmental classification of the whole slide digital image (Figure 2.1c, Methods). The lower adipose fraction was associated with higher mammographic breast density ($p=0.0067$) suggesting the sample histology was representative of the whole breast texture. Interestingly, solid DCIS were associated with a higher adipose fraction (median 69% vs 40%, $p=0.008$), suggesting a contribution of the breast microenvironment to the growth architecture. Overall, the cohort represents a diverse set of pure in situ lesions identified in absence of any detectable invasive component. The studied samples are enriched for DCIS lesions and specifically annotated for their histological architecture. Each

sample was profiled using multiple assays, performed on sequential histological sections (4-7 μm) used for whole transcriptome, whole-exome and spatial immune profiling. Whenever possible, the investigated regions were matched across assays to preserve the spatial information in the analysis and limit the variation due to spatial heterogeneity. Spatial heterogeneity was further addressed in 21 samples for which multiple sub-regions were profiled independently.

The expression of genes was measured using high-throughput sequencing of RNA-seq libraries directly prepared from the micro-biopsied regions 27 (Supplementary Table 2.2). The samples were classified according to the PAM50 intrinsic subtypes used for invasive breast cancer (IBC), which identified Basal (N=5), Luminal A (N=10), Luminal B (N=6), Her2-like (N=7) and Normal-like (N=10) samples (Figure 2.1d). Consistent with IBC classification, Luminal A and B were enriched for samples from ER+ cases, while Her2-like were enriched for Her2+ cases. Similarly, Luminal B and Her2-like were enriched in HG-DCIS, while Luminal A was almost exclusively composed of cribriform LG-DCIS. Luminal A and Normal-like represented closely related classes and together comprised the majority of the samples (20/38), which is not unexpected given the higher fraction of low-grade and pure in-situ lesions in the cohort, in contrast with IBC and previous DCIS expression profiling studies [18,30]. The PAM50 subtype of two independent sub-regions with matching histology and grade was determined in 10 samples, and observed to be discordant in 5 samples (Supplementary Table 2.2B), which was associated with larger distances between the regions (Mann-Whitney, p=0.005, Supplementary Figure 2.1b). Interestingly, matched index and recurrent samples from two patients had at least one region with concordant subtype. Across all samples the distribution of probabilities for each PAM50 subtype likely captures such heterogeneity. Normal-like were truly a mix of Normal and Luminal A, while Her2-like tended to have two main subsets: Her2/Basal and Her2/Luminal B. This suggests that

subtypes inferred from bulk analysis, even after epithelial micro-dissection, are frequently the result of a variable mixture of pure subtypes.

Subtype differences in the mutational landscape.

To determine whether any of the histological or molecular subtypes described above were associated with specific genetic alterations, we characterized their mutational landscape. Whole exome sequencing was carried out on micro-biopsies from 30 samples using a procedure specifically optimized for low amount of damaged DNA [26]. Mutations and copy number alterations (CNA) were identified in 27 and 30 samples, respectively (Supplementary Table 2.3). The median copy number burden - or fraction of the genome involved in CNA – was 0.14 and was 2.5 fold higher in HG-DCIS (Mann-Whitney, p=0.017, Figure 2.2a-b). Whole genome doubling (WGD) events were detected in 3/8 eligible samples, all of which were low or intermediate grade cribriform DCIS consistent with its early timing in breast carcinogenesis [31] (Supplementary Table 2.3). Consistent with previous studies, loss of 16q (13/30) and 17p (12/30) or gain of 1q (12/30) were among the most frequent chromosomal alterations and, while many events were more frequent in HG-DCIS (Figure 2.2c, Supplementary Table 2.4), these hallmarks were also observed in low-grade or benign lesions, including ADH: (1q gain: 1/9, 16q loss: 7/9, 17p loss: 3/9).

We identified between 74 and 207 coding mutations per sample. The mutational burden was higher in HG-DCIS (Mann-Whitney, p=0.003) and Her2-like subtypes (Mann-Whitney, p=0.025), recognizing that these categories are overlapping. The HG-DCIS burden (4.4 mut/Mb) was higher than previous reports, possibly due to residual germline variants in our study [11,13]. We identified aging-associated mutational signatures (SBS1 and SBS5) in all samples eligible for analysis (N=13), APOBEC signature (SBS2 and SBS13) on 1 intermediate grade solid DCIS and mismatch repair signature (SBS15 or SBS21) in 3 DCIS of variable grade and architecture

(Supplementary Table 2.5). The APOBEC signature is therefore more rare in DCIS than IBC (~8% vs >75%), but can be present in premalignant lesions. Interestingly, this sample also displayed clustered mutations (N=3 within 1,416 bp) in chromosome 17q (Supplementary Figure 2.2), an APOBEC-driven kataegis site frequently seen in IBC [32]. The most recurrently mutated genes were *PIK3CA* (44%), *TP53* (31%), and *GATA3* (20%), and were all affected by known somatic mutations in breast cancer at similar rates to previous studies of pure DCIS [6,7,9–11] (Figure 2.2d, Table & Supplementary Table S2.6). *TP53* mutations were only found in HG-DCIS and associated with high CNA burden (Mann-Whitney, p=0.018), while *GATA3* mutations were only found in cribriform or ADH histologies and associated with LG-DCIS (Fisher Exact, p=0.005). Interestingly, *GATA3* mutations were identified in larger lesions (Mann-Whitney, p=0.038), consistent with a similar observation in invasive cancer and the larger tumor size of *GATA3* mutated xenograft models [33,34]. Another 9 selected genes known to be mutated in IBC were recurrently affected by 18 mutations predicted to be deleterious, 4 of which are known somatic mutations [7]. The result suggests that oncogenic driver mutations are already present at the premalignant stage, including in LG-DCIS (e.g. *SF3B1* c.2098A>G) or ADH (*GATA3* c.925-3_925-2del). This is consistent with previous reports and reports of field effect mutations in normal ducts or benign lesions [35,36], though the contribution of these mutations to the lesion progression remains to be determined.

Genetic heterogeneity and clonal diversity.

The histologic assessment and expression profiling have revealed variable levels of phenotypic heterogeneity across the samples. In order to determine whether such heterogeneity is present at the genetic level, we measured genetic heterogeneity in two distinct ways: 1) divergence, which measures the genetic distance between regions of a sample and, 2) clonal relationships,

which uses phylogenetic tree construction to establish evolutionary order to genetic alterations (Supplementary Table 2.3). We measured divergence by computing a CNA-based score on 19 pairs of histologically matching regions in 11 samples (Supplementary Table 2.3, Methods). With no pairs completely independent, the spatial distance separating dissected DCIS regions was correlated with the extent of their genetic divergence (R2=0.65, p=0.00017, Supplementary Figure 2.3), while this could simply be a result of local proliferation, it could also be a consequence of selective pressures of the micro-environment, migratory capacity or genomic instability of particular clones. Interestingly, 1 ADH had the lowest divergence despite a large distance, suggesting either a different pattern in ADH or a distance threshold for the extent of the correlation. Divergence was not associated with grade, Her2/ER status, or adipose fraction suggesting that local genetic heterogeneity is not associated with progression risk factors.

More precise clonal relationships between regions were evaluated using phylogenetic analysis in 12 samples, comparing CNA, and mutations when available (Figure 2.3, Supplementary Figure 2.4, Methods). While the majority (88.4%) of CNA were shared across all regions of a sample, 11.6% were private to some regions, as observed in 7/12 samples. Multiple samples (3/12) contained mutations in putative cancer driver genes that were private to one region only. These included known and likely pathogenic mutations in *ATR, PIK3CA, MET, KDM5C*, suggesting that not all driver mutations are acquired early. Interestingly, the three samples with the most private CNA displayed discordant histological architecture or discordant PAM50 subtypes between regions, suggesting that within a sample, genetic and phenotypic differences are linked. Furthermore, in 4/5 samples containing regions with discordant histology and 3/4 with discordant PAM50 subtypes, features historically associated with low-risk of progression (benign histology, Normal or LumA subtype), appeared earlier than regions with high-risk features (Her2

or Basal subtype, presence of necrosis). Overall, these results illustrate that in these samples, regions evolved to acquire distinct histological and molecular features, and in particular, regions with low-risk features can precede regions with high-risk features.

Substantial heterogeneity and evolutionary patterns are evident in samples like MCL76_061_16200 (Figure 2.3a-c), where a region of benign columnar alterations preceded two cribriform regions. While all regions shared a WGD event as well as several arm-level CNA and pathogenic mutations in *GATA3* and *SF3B1*, the cribriform region A acquired private 5q and 8q gains and necrotic features. While this example shows tandem genetic and histological changes as seen across the cohort, it also illustrates that despite occurring earlier, the benign region shares many "driver-like" alterations with both cribriform regions. Furthermore, in another example, despite homogeneous cribriform histologies in regions of MCL76_077_15300 (Figure 2.3d-f) only one cribriform region lost a copy number of chromosome 8, and presented with Her2 PAM50 subtype as opposed to its Luminal A predecessor. Notably bulk studies have shown chromosome 8 loss to be more frequent in Her2 vs Luminal A breast cancers [37]. Taken all together we illustrate abundant genetic heterogeneity in pure DCIS of all histologies and grades that parallels the levels of phenotypic heterogeneity and often accompanies it, even in regions that are millimeters apart.

Regional differences in the immune micro-environment.

To measure the diversity of the immune-landscape and to investigate its potential association with molecular or histological features, we used multiplex immuno-histochemistry (mIHC) to measure the number and density of four cell types - T-cells (CD3+), B-cells (CD20+), T-regs (CD3+/FOXP3+) and epithelial cells (PanCK+) - according to their proliferative status (Ki67+). Both epithelial (PanCK+) and adjacent stromal (PanCK- proximal to epithelium) areas from pre-malignant (N=36 regions across 32 samples) or normal (N=21 across 21 samples)

histologies were evaluated. Among pre-malignant regions, the high-grade epithelial areas had lower cell density due to larger cell sizes and frequent central necrosis (median 3.8 vs 6.4 103 cells/mm2 $p<0.03$ – Mann-Whitney). Solid lesions had the highest fraction of proliferating epithelial cells (median 11.5% vs 2.8% $p<0.02$ - Mann-Whitney, Supplementary Figure 2.5a), and interestingly 3/10 HG-DCIS cribriform lesions (2 Her2-like, 1 Luminal B) had markedly higher proliferation. Consistent with previous findings, we observed higher lymphocyte infiltration in ER- and Her2+ samples compared to ER+ ones (Supplementary Figure 2.5b, Mann-Whitney, $p<0.001$). We next classified all regions using non-negative matrix factorization of the stromal and epithelial cell densities, resulting in 3 immune-states characterized by their dominant meta-markers (MM; Figure 2.4a-b Supplementary Table 2.7a-c): 1) "Active" - ubiquitous high T-cells (high MM2), including a subset with elevated T-cell proliferation (high MM1), 2) "Suppressed" - ubiquitous low T-cells (low MM1 and MM2), high B-cells and T-regs (high MM3), and 3) "Excluded" - high stromal, low epithelial densities (high MM4). To further confirm differences between immune-states, we compared the total T-cell, B-cell and T-regs densities in epithelial and stromal compartments. While overall lymphocyte densities were much higher in stroma than in epithelium across all examined regions (median ratio 9.8, Supplementary Table 2.7a), the skew was a distinguishing feature in regions in Excluded state for all three cell types (Figure 2.4c, Supplementary Figure 2.6 and Supplementary Table 2.7b). Furthermore, regions in Active states had the highest epithelial T-cell density (120 cells/mm^2) while regions in Suppressed state had the highest T-regs and B-cells epithelial densities (10.4 and 8.1 cells/mm^2 respectively). A larger fraction of the normal regions were found in Active (7/21) or Suppressed state (9/21) rather than in Excluded state (4/21) and pre-malignant regions in Excluded state were more likely to be high grade (7/15 vs 2/17 $p=0.049$). Interestingly, the immune states of normal and pre-malignant

regions were concordant in 12/19 matched cases and discordant in 7 whose lesions were specifically in the Excluded state (Figure 2.4d). This suggests that the Excluded state may be acquired in response to pre-malignant growth, while other states may be intrinsic to various breast micro-environments. Furthermore, pre-malignant regions in Suppressed state were more likely identified in cases younger than 55 (5/8 vs 4/24 OR=7.6 p=0.02), consistent with the younger age of DCIS patients with infiltrating PD-L1+ lymphocytes [38]. We did not observe any associations between immune states and intrinsic subtype, ER or Her2 status, tumor size, breast density, adipose fraction or DCIS architecture suggesting that they may be independent from traditional histopathological progression risk factors.

In order to identify functional differences between immune-states, we evaluated the differential activity of Hallmark and Reactome processes among the 29 DCIS regions with available gene expression information (Supplementary Figure 2.7). Compared to Active and Suppressed states, the Excluded state was associated with upregulation of Type 1 and 2 Interferon response, PD1 signaling and proliferation-related processes as well as the repression of Calreticulin-Calnexin cycle (Supplementary Figure 2.7). Noting that the epithelium of DCIS in Excluded state were not completely depleted of infiltrating lymphocytes, the upregulated processes were consistent with the higher expression of *PCDC1* or *CTLA4* genes in DCIS in Excluded state (Figure 2.4e), albeit not significant, and suggesting a likely continuum of increasing immuno-suppression from Suppressed to Excluded states. More interestingly, the repression of the Calreticulin-Calnexin cycle was confirmed via single-sample enrichment analysis and showed a progressive repression from Active, to Suppressed, to Excluded states (p=0.022, ANOVA, Figure 2.4f). This suggests that the export of glyco-proteins - including components of MHC1 complex - via the endoplasmic reticulum, impacts immune-surveillance. To verify this hypothesis, we

measured the in situ expression of MHC1 complex in 15 samples (Supplementary Table 2.8 and Supplementary Figure 2.8) and compared its levels in adjacent normal and DCIS in each immune state. While the level of MHC1 expression in DCIS region were not significantly different between Excluded and non-Excluded samples, the change between normal and DCIS was different, with the non-Excluded samples displaying increased expression between normal and DCIS, while the Excluded samples remained constant ($p=0.0009$, Mann-Whitney, $N=15$, Figure 2.4g). This therefore suggests that the Excluded immune state may be mediated by both intrinsic expression level of MHC1 and ability to increase it in DCIS.

2.4 Discussion

There is a compelling requirement for a DCIS atlas that delivers a relatively unbiased, multi-modal perspective of pre-invasive breast cancer. Here, we report the multi-modal profiling of a diverse set of pure DCIS. This comprehensive atlas both confirms previous molecular findings and provides a higher resolution histological and spatial context to interpret them. However, with only 3 known recurrences, the significance of our observations for progression prognosis could not be formally established. Our findings provide a landscape of representative pure DCIS identified in absence of invasive lesions. While some lesions were small, others were quite extended (N=14 larger than 4 cm), which should capture factors that may be associated with robust containment. The cohort therefore spans a variety of clinical, histological, phenotypic and genotypic features. Such variety and contrast are critical to ensure this atlas' utility in designing larger studies, or perhaps providing more cautionary interpretation of observations from cohorts enriched for specific risk factors.

At the heart of our study's innovation was the ability to generate molecular profiles from limited amounts of dissected archival tissue specimen. Similar approaches are used to study clonal expansion in normal tissues [28,29], but generally not performed in parallel for RNA and DNA. Importantly some limitations remain and not all assays were successful. The large variability in success rate was not easy to predict. Likely the age of the specimen, its size, fixation conditions and storage conditions all contribute to success variability which cannot be controlled in a retrospective investigation. Additional limitations are analytical, such as the absence of a matched source of normal DNA from every sample which can result in residual germline variants, perhaps inflating the overall mutation rate observed. The use of adjacent normal tissue can also be problematic and there is ample evidence that they also accumulate somatic mutations [39]. In our

study, we clearly identified known breast cancer driver mutations in samples from ADH or other benign alterations. Overall, while some samples are unlikely to ever contain sufficient material for profiling or dissection of adjacent normal, as methodologies evolve and advance, the success rate and data quality will improve to make molecular premalignant profiling more accessible and as routine as is the case in invasive cancer.

Our report contributes to two major advances for understanding pre-malignant lesions. First, we characterized most samples across four important modalities all within a maximum of 50 μm sequentially sectioned tissue. Such advances were enabled by pre-analytical improvements allowing us to reduce the tissue requirement, to include small lesions, and to precisely match regions of interest across each modality: histology, epithelial gene expression, DNA mutations, and immune landscape. As a result we could isolate regions with different histological features that may coexist within a specimen and more confidently establish their association with expression subtypes, clonal heterogeneity or immune state. For example, the integration of histology and expression subtypes showed clear correlation between cribriform architecture and Luminal A subtype. By integrating histology, expression subtype and immune state we showed that some immune-states are found in normal areas and that there is no clear association between immune state and expression subtype. Hence, the depth and interpretability of the analysis is considerably increased by integrating all modalities at the regional level. This has been clearly the case in large cancer studies such as the TCGA, or, more recently through the integrated analysis of histological and somatic features in normal, aging tissues [28,29,40]. While most studies do not typically include immuno-histochemical or other multiplexed spatial analysis, other important advancements in this field in the past year include spatial proteomics used to evaluate the structure of the myoepithelium in DCIS, and spatial transcriptomics used to identify the transcriptional

effect of driver mutations in DCIS, representing the emerging frontier of premalignant tissue characterization [10,41]. It is therefore likely that additional spatial profiling compatible with FFPE specimens will bring additional prognostic and mechanistic insights in future DCIS studies.

The other important contribution of our study is the sub-histological analysis to compare regions of interest from the same sample and infer phylogenetic relationships between them. While we determined that the majority of the DCIS samples were classified as Normal-like and Luminal A subtype, typically considered less-aggressive subtypes in breast cancer and reflective of the known precursor stage that DCIS represents, we showed evidence for intrinsic heterogeneity in the PAM50 probabilities, either from the distribution of probabilities within a region or from physically separated regions. This is not entirely surprising as bulk expression subtypes are the result of averaging heterogeneity, similar to glioblastoma subtypes [42] or IBC subtypes [43] from single-cell analysis. Such heterogeneity, especially in DCIS, had been proposed before on the basis of marker staining [44] and our results confirm that it may be rather common. Similarly to the frequency of heterogeneity between region subtypes, we identified evidence of genetic heterogeneity in 7/12 cases, including the presence of private putative driver mutations. This fraction may be an underestimate given the close proximity of many selected pairs. However, the majority of putative genetic drivers, copy number hallmarks and even WGD were clonal, shared by all regions investigated, including a few normal regions. This observation supports evolutionary models derived from invasive cancer, including multi-sample studies, that suggest that most driver mutations occur early followed by a phase of clonal expansion. Similar observations were also made in early multi-regional studies in DCIS [44–46] and studies comparing synchronous DCIS-IBC cases using single-cell sequencing [13], providing further evidence that breast cancer genetic evolution starts in the pre-invasive stage and possibly in normal regions. It is likely that driver

alterations may even be present in adjacent histologically normal tissue as observed in field effects studies in normal ducts [39,47]. Such effects support an important contribution of host factors to the initial genetic injury. Hence, unlike previous attempts which were focused on histopathological features, including grade, surgical margins [48,49], future DCIS prognostic models will likely need to be derived from lifetime cancer risk models like GAIL [50] or BOADICEA [51] and incorporate host specific factors, such as polygenic risk scores and reproductive factors, that likely contribute to the DCIS initiation and trajectory.

The immune micro-environment of DCIS has been previously investigated, using both quantification of tumor infiltrating lymphocytes (TIL) and more specific immuno-histochemical approaches and revealed clear quantitative and qualitative variation in lymphocyte infiltration, including higher TIL number and more immunosuppressive features in high risk lesions [19]. Importantly, previous studies in pure DCIS did not quantify stromal and epithelial TILs separately [12,19]. This distinction may be hard to make in IBC, where both compartments interact at the invasive front and pathologist subjectivity can have a major impact [52,53]. However, this separation can be more clearly established in the analysis of DCIS and was critical in the identification of the Excluded immune state in our atlas. While the Active and Suppressed states have been observed before and could readily be identified in our data, the identification of the Excluded state required the use of an analytical method (NMF) to account for the strong correlations that can exist between TILs type and compartments. The inclusion of adjacent normal areas was also important to interpret the significance of the immune-states, as the Excluded state appeared more likely in reaction to the DCIS growth and increased grade. The Excluded state exhibited features of immune evasion and could represent a more advanced level of immuno-suppression than the Suppressed state, with the consequence of a topological exclusion from the duct. The downregulation of

components of the Calreticulin-Calnexin cycle in the epithelium in Excluded state could impact MHC-I export or maturation, as suggested by the lack of MHC-I expression induction in DCIS of the Excluded state, hence providing an evasion mechanism, and contrasting with evasion mediated by MHC-I genetic loss observed in IBC [54,55]. It would be interesting to determine whether the immune states identified can explain the variability of response to local injection of anti-PD1 antibody in DCIS patients, and whether any of the states would elicit, or prevent, the desired ductal infiltration by T-cells [56].

As illustrated by our study and recent advances in the profiling of normal tissues [28,29], histopathology and molecular pathology are becoming more integrated fields, generating deeper and broader datasets at increased cellular and spatial resolution, from the most challenging human samples. Future studies of early transformation and pre-cancer biology such as the one presented here will likely benefit the most from such approaches which capture heterogeneity at scale and can help reconcile analog (optical) and digital (genomics and multiplex) observations. As a result, such multi-dimensional integration may help identify common factors mediating epithelial transformation and progression across multiple glands and organs.

2.5 Materials and Methods

Sample collection and preparation.

FFPE blocks were obtained from UCSD or UVM Pathology Departments after surgical biopsy, excision or mastectomy. The study was reviewed and approved by each institutional review board and they granted a waiver of consent. Eligibility criteria were: 1) adult female, 2) pure DCIS diagnosis (without evidence of invasive disease), 3) with available pathology blocks. Few cases also had bilateral disease or were matching index and recurrent lesion (ipsilateral or contralateral - Table 2.1 and Supplementary Table 2.1). Importantly there was no attempt to enrich for high-risk cases or investigate specifically the role of certain candidate risk factors. Factors such as age, grade, race, ER or Her2 status were not part of the selection criteria and the cohort was designed to reflect patients seen in a regular DCIS clinic. All specimen blocks were de-identified and sectioned sequentially for the following purpose: Hematoxylin-Eosin (H&E) staining (N=1; 4 µM glass slide), Laser Capture Microdissection (LCM; N=3; 7 µM glass slide coated with polyethylene naphthalate – ThermoFisher #LCM0522), multiplex or regular immunohistochemistry (N≥3 4 µM glass slide) and a final H&E staining (N=1; 4 µM glass slide). The H&E slides were scanned at high resolution and reviewed and annotated by the study pathologist. The LCM slides were stored at -20ºC in an airtight container with desiccant until ready for dissection (1 day to 3 months). H&E sections were diagnosed according to standard of care criteria (AJCC TNM 8th ed. / CAP Breast DCIS Reporting Protocol v4.3). DCIS features recorded included lesion grade: Grade I (low), Grade II (intermediate) or Grade III (high), and associated histology: e.g., papillary, cribriform, solid, comedo necrosis. DCIS lesion, normal glands (and in some cases hyperplasia) were delineated on H&E images to assist LCM. DCIS laterality and size, patient age and menopausal status, and lesion mode of detection were obtained from the original

pathology reports or from the Vermont Breast Cancer Surveillance System (UVM specimen) or local cancer registry and chart review (UCSD specimen). Hormone receptor and Her2 statuses (where available) were gathered from the patient reports and/or by de novo IHC staining. The LCM sections were thawed and stained with eosin, sections were kept in xylene and dissected within 2 hours of staining. LCM was performed using the ArcturusXM Laser Capture Microdissection System (ThermoFisher). Matching regions from 6 adjacent sections were collected on CapSure Macro Cap (for DNA, N=3 slides) or HS caps (for RNA, N=3 slides), region size, and unambiguous match permitting. Post-dissection, all caps were covered and stored at -20ºC with desiccant. DNA extraction and QC: The membrane and adhering tissue were peeled off the caps using a razor blade and the peeled membrane was incubated in proteinase K digestion reaction overnight for 16 h at 56°C to maximize DNA yield after cell lysis The DNA was extracted using the QIAamp DNA Micro Kit (Qiagen) and the elution was done in 20 µL. The extracted DNA was quantified by fluorometry (HS dsDNA kit Qbit – Thermofisher).

RNA-Sequencing and analysis.

Library Preparation.

RNA sequencing was performed using SMART-3Seq, a 3' tagging strategy specifically designed for degraded RNA directly from FFPE LCM specimen 27. LCM dissected SMART-3Seq libraries were prepared using the standard protocol for FFPE tissue on Arcturus HS LCM Cap and the individual library SPRI purification option. All FFPE LCM dissected libraries were amplified using 19 PCR cycles during indexing to minimize over-amplification of high abundance mRNAs in each library. Libraries were individually analyzed for size distribution on an Agilent 2200 TapeStation with High Sensitivity D1000 reagent kits to verify average library size of 190 bp and stored at -20 C until sequencing. When all libraries were ready for sequencing, 1 µL of each library

was then used to create two library pools used for sequencing and quantified by Qubit 2.0 Fluorometer HS DNA assay. Library pools were sequenced with a 1% PhiX spike-in control library and sequenced on an Illumina HiSeq4000, a run type of single read 75 (SR75) and dual index sequencing.

Transcriptome analysis: Read count data was obtained using a dedicated analysis workflow https://github.com/danielanach/SMART-3SEQ-smk. Briefly, sequencing reads were trimmed using cutadapt 1.18, UMIs were processed using the umi_homopolymer.py script in the SMART-3SEQ tools (https://github.com/jwfoley/3SEQtools), aligned using STAR 2.6.1a, deduplicated using the dedup.py script from https://github.com/jwfoley/umi-dedup and read counts were calculated using featureCounts 1.6.3 [57,58]. Count data was then merged and filtered to remove samples with fewer than 55,000 counts and genes with fewer than 10 read counts across all samples. Filtered count data was then loaded into Seurat version 3.2.3 and processed using the SCTransform() function version 0.3.2 to regress out the high mitochondrial content variability across the samples [59]. Batch correction was then performed using ComBat to remove variation attributable to the sequencing center (UCSD vs UVM) [60]. PAM50 subtype probabilities were calculated from the SCTransform and batch normalized data using the genefu package [61]. Gene set enrichment analysis (GSEA) was performed as in [62] and single-sample GSEA as in [63]. Gene sets from the REACTOME and Hallmark collections in MSigDB were used to compare the excluded to the non-excluded groups, a permutation test was performed to assess the significance of the GSEA results [64,65]. ANOVA was used to compare the ssGSEA results between the three mIHC groups. FDR of less than 0.1 and p-values of less than 0.05 were considered significant.

Whole exome sequencing and primary analysis.

Library preparation: DNA was sheared down to 200 base pairs (bp) using Adaptive Focused Acoustics on the Covaris E220 (Covaris Inc) following manufacturer recommendations with 10 µL Low EDTA TE buffer supplemented with 5 µL of truSHEAR buffer using a microTUBE-15. Libraries were prepared using the Accel-NGS 2S PCR-Free DNA Library Kit (Swift Biosciences). Ligated and purified libraries were amplified using KAPA HiFi HotStart Real-time PCR 2X Master Mix (KAPA Biosystems). Samples were amplified with 5 µL of KAPA P5 and KAPA P7 primers. The reactions were denatured for 45 seconds (sec) at 98°C and amplified 13-15 cycles for 15 sec at 98°C, for 30 sec at 65°C, and for 30 sec at 72°C, followed by final extension for 1 min at 72°C. Samples were amplified until they reached Fluorescent Standard 3, cycles being dependent on input DNA quantity and quality. PCR reactions were then purified using 1x AMPure XP bead clean-up and eluted into 20 µL of nuclease-free water. The resulting libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qubit dsDNA HS assay).

Capture and Sequencing: Samples were paired and combined (12 µL total) to yield a capture "pond" of at least 350 ng, and supplemented with 5 µL of SureSelect XTHS and XT Low Input Blocker Mix. The hybridization and capture was performed using the Human All Exon V7 panel (S31285117) paired with the Agilent SureSelect XT HS Target Enrichment Kit following manufacturer's recommendations. Post-capture amplification was performed on the beads in a 25 µL reaction: 12.5 µL of nuclease-free water, 10 µL 5x Herculase II Reaction Buffer, 1 µL Herculase II Fusion DNA Polymerase, 0.5 µL 100 millimolar (mM) dNTP Mix and 1 µL SureSelect Post-Capture Primer Mix. The reaction was denatured for 30 sec at 98°C, then amplified for 12 cycles of 98°C for 30 sec, 60°C for 30 sec and 72°C for 1 min, followed by an

extension at 72°C for 5 minutes and a final hold at 4°C. Libraries were purified with a 1x AMPure XP bead clean up and eluted into 20 μL nuclease free water in preparation for sequencing. The resulting libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qbit – ThermoFisher). All libraries were sequenced using the HiSeq 4000 sequencer (Illumina) for 100 cycles in Paired-End mode. Libraries with distinct indexes were pooled in equimolar amounts. The sequencing and capture pools were later deconvoluted using program bcl2fastq.

Sequencing reads processing and coverage quality control: Sequencing data was analyzed using bcbio-nextgen (v1.1.6) as a workflow manager. Adapter sequences were trimmed using Atropos (v1.1.22), the trimmed reads were subsequently aligned with bwa-mem (v0.7.17) to reference genome hg19, then PCR duplicates were removed using biobambam2 (v2.0.87) [66–68]. Additional BAM file manipulation and collection of QC metrics was performed with picard (v2.20.4) and samtools (v1.9) [69]. The summary statistics of the sequencing and coverage results are presented in Supplementary Table 2.9.

Identification of somatic mutation and copy number alterations.

Variant calling,

Single nucleotide variants (SNVs) and short insertions and deletions (indels) were called with VarDictJava (v1.6.0), and Mutect2 (v2.2) [70,71]. Variants were required to fall within a 10 bp boundary of targeted regions that overlapped with RefSeq genes (v 109.20190905). A pool of normal DNA was created using whole exome sequencing data of blood of 18 unrelated individuals and was used to eliminate artifacts and common germline variants. Only variants called by both algorithms were considered. These variants were then subjected to an initial filtering step with default bcbio-nextgen tumor-only variant calling filters and the following parameters were used:

position covered by at least 5 reads, mapping quality more than 45, mean position in read greater than 15, number of average read mis-matches less than 2.5, microsatellite length less than 5, tumor log odds threshold more than 10, Fisher strand bias Phred-scaled probability less than 10 and VAF more than 0.1 [72]. Functional effects were predicted using SnpEff (v4.3.1) [73]. All samples were re-evaluated for the presence of COSMIC (v91) database mutations which have been previously observed in at least 15 patients and fall within 137 known breast cancer driver genes (Supplementary Table 2.10) [7].

Germline variant filtering.

In absence of matched normal tissue for DCIS samples, somatic mutations were prioritized computationally using the approach from the bcbio-nextgen tumor-only configuration then additionally subjected to more stringent filtering [72]. Briefly, common variants (MAF>10-3 or more than 9 individuals) present in population databases - 1000 genomes (v2.8), ExAC (v0.3), or gnomAD exome (v2.1) - were removed unless in a tier 1 gene from the cancer gene consensus and present in either COSMIC (v91) or clinvar (20190513) [7,74–77]. Variants were removed as likely germline if found at a variant allelic fraction (VAF) greater or equal to 0.9 in non-LOH genomic segments – as determined by CNA analysis (below). Lastly, variants were also removed as potential germline (or artifact) if found in more than two patients in the pool of normal (described above).

Single-sample CNA calling.

CNVkit [78] was used for calling somatic copy number alterations (CNA) to measure both overall CNA burden, arm and gene level CNA and identify LOH as previously described in [26]. Allele-specific copy number calling algorithm, ASCAT, was used on a select number of samples for which there was sufficient coverage and the algorithm converged on a solution, in order to

identify whole-genome doubling events as well as confirm CNA identified by CNVkit [79]. Default parameters were used with ASCAT with the exception of a segmentation penalty of 100 and a gamma of 1.

Multi-region CNA segmentation:

To generate harmonized segmentation breakpoints between regions belonging to the same sample, multi-region segmentation was performed with the R CopyNumber (v1.26.0) package [80]. Outliers in CNVkit bin-level log2 copy ratios were detected and modified using Median Absolute Deviation Winsorization with the winsorize() function, segments were then called using the multipcf() function with a gamma of 40.

Mutational signatures.

Mutational signatures were called on merged region samples using a single-sample variation of SigProfiler with default parameters to decompose into known single-base substitutions (SBS) reported in COSMIC [81,82].

Analysis of the clonal evolution and genetic heterogeneity.

Measurement of genetic divergence.

Divergence was measured on each pair of related regions, a and b, using equation (1):

$$Divergence_{a,b} = \sum_{k=0}^{n} |copy\ ratio\ a_k - copy\ ratio\ b_k| * \left(\frac{bins_k}{total\ bins}\right) \quad (1)$$

Where k is the copy number segment, n is the total segments, and $bins$is the number of bins covered by a segment from the CNVkit input file. The $\frac{bins_k}{total\ bins}$ term was used as a weighted correction factor for the number of bins contributing to a segment. For samples with more than 2 regions, the maximum divergence between any two regions was used to represent the sample.

CNA-based phylogenetic reconstruction.

Construction of phylogenetic trees was performed similarly to the methodology outlined in [83]. Briefly, for each sample the log2 copy ratios from multi-sample copy number segments with at least 12 probes (see above), were translated into a matrix containing -1 for loss (log2 copy ratio<-0.6), 0 for neutral (-0.4<= log2 copy ratio <=0.3) and undetermined for anything else. This matrix was then used to generate Maximum Parsimony trees using phangorn using default parameters [84].

Mutation-based phylogenetic reconstruction.

To allow the analysis of clonal relationships between regions of the same sample, the coverage depth of each allele at any remaining mutated position in any region was extracted using Mutect2 joint variant caller on the sets of aligned reads from each region. In order to call a mutation either absent or present in a region, we used a Bayesian inference model specifically designed for multi-region variant calling [85]. Treeomics (v1.7.10) was run with the default parameters except for e=0.02. The tree solution which matched the CNA-based reconstruction was then integrated into a single tree for Figures 3 and S4.

Multiplex Immunohistochemistry.

Staining.

Tissue sections were prepared from formalin-fixed paraffin embedded tissue blocks and cut to 4 micrometers serial sections and mounted on Superfrost Plus (VWR). The procedure for multiplex immunohistochemistry (mIHC) was followed by a manufacturer's protocol for Opal7-color automation IHC kit (Akoya Bioscience), and the staining was performed with Autostainer DISCOVERY ULTRA (Ventana). Antibodies used in mIHC are anti-CD3 (clone 2GV6, Ventana),

anti-CD20 (clone L26, Ventana), anti-Ki67 (clone 30-9, Ventana), anti-FOXP3 (clone SP97, Spring), anti-pan cytokeratin (CK; clone AE1/AE3, DAKO), anti-CD117 (clone c-kit, DAKO). The molecular markers of immune panel (CD3, CD20, Ki67, CKs, FOXP3 and CD117) were visualized with Opal520, Opal540, Opal570, Opal620, Opal650 and Opal690, respectively. DAPI counterstaining was performed with Discovery QD DAPI (Roche). ProLong Diamond Antifade Mounting (ThermoScientific) was used for mounting the coverslip. Detailed staining conditions and autostainer's protocols are reported in our recent report [86].

Visualization and analysis.

Tissue samples stained with mIHC were scanned with multispectral imaging microscopy (Vectra 3, Akoya Bioscience). Scanned multispectral images were unmixed on inForm software (ver.2.4.0, Akoya Bioscience) to acquire the fluorescence signal from each marker [86]. Imaging analysis was performed on inForm software by identifying tumor (CK+ area) and stroma (CK- area proximal to the epithelium), each nucleated cell and its cell type. Alternatively, QuPath software 2.3.1 [87] was also used to perform similar imaging analysis on unmixed images converted to multi-layered TIFF format by inForm software 86. The images of the regions of the same type (DCIS or normal) from the same case, were typically stitched together and stored and shared as one single larger multilayered TIFF image (data availability below). Scanned image areas were aggregated into up to three histological regions per sample: main pre-invasive lesion, alternate pre-invasive lesion, normal epithelium. In each region, the stromal and epithelial densities of each cell type and state was calculated, including when cells were not present (density=0). Regularized marker densities into distribution deciles were then used to classify samples using non-negative matrix factorization (Supplementary Table 2.7c). The immune-states were assigned and named after the hierarchical clustering of the H matrix (meta-marker values).

MHC1 immunostaining and analysis.

Four micron sections were baked at 60 degrees for 1 hour, followed by deparaffinization through three successive changes of xylene. Tissue was then rehydrated in decreasing grades of alcohol, with two changes of 100%, 95%, and then 70% EtOH, followed by diH20. Antigen retrieval was performed using Antigen Unmasking Solution Citrate Based pH6, H-3300 (Vector) at 95°C for 30 min. Staining was performed using the intelliPATH Automated IHC stainer (Biocare). Endogenous peroxidase was blocked using Bloxall blocking solution, SP-6000 (Vector) for 10 min, followed by 2 washes in TBST. Afterwards, tissue was blocked with a 3% Donkey Serum for 10 min, followed by blocking with Anti-HLA Class I ABC Primary Mouse Antibody, ab70328m (Abcam) at 1:1000 for 1 hour and subsequently washed twice with TBST. Tissues were then blocked with Anti-Mouse HRP UltraPolymer IgG, 2MH-100 (Cell IDx) for 30 min, and washed twice with TBST. The reaction was then developed with 3,3'-Diaminobenzidine Chromogen, 95041-478 (VWR) for 5 min, and then stopped with two washes in diH20. Counterstaining was performed with Mayer's Hematoxylin Solution, 51275 (Sigma) for 5 min. Lastly tissues were washed twice in TBST, and once in diH20, dehydrated in increasing grades of EtOH, then cleared and mounted with xylene based mountant. MHC1 expression was scored from 0 to 3 separately for DCIS and normal epithelium throughout the entire section, away from possible biopsied areas. The scores were established as follows: 0: no staining or weak staining in <50% of cells; 1: weak staining in >50% of cells; 2: intermediate staining in >50% of cells;. 3: strong staining in >50% of cells.

Whole slide image digital analysis.

High resolution whole slide images of H&E stains were loaded into a QuPath (v2.3) project [87]. One analysis area was defined for each specimen, avoiding location of biopsies as well as dust

or marked areas. The analysis areas were segmented into superpixels (sigma=5 μm, spacing=50 μm, maxIterations=10, regularization=0.25) and each superpixel was annotated with both Hematoxylin and Eosin Intensity features (size=2 μm, tile size=25 μm). The mean, median, min, max and standard deviation values were then smoothed (Haralick distance=1, Haralick bins=32). Multiple training areas were annotated from each of the following classes: adipose, stroma, inflammation, epithelium (normal and atypical), void, necrosis, blood vessels. Multiple areas across 2 to 4 samples were used to train a Random Tree classifier. The classifier was then applied to all superpixels included in the analysis area. The accuracy of the classifier was assessed both visually and with multiple test areas for each class. Superpixels of the same class were merged into single annotations and the resulting areas recorded. Separate classifiers were used for images from different institutions, to mitigate possible variation staining, scanning or image format. The fraction of adipose area was compared to breast density using Mann-Whitney test comparing dense & heterogeneously dense breast to other lower densities, or comparing solid DCIS to non-solid DCIS lesions.

Data Availability.

The raw RNA and DNA sequencing data has been deposited in dbGAP phs002225. High resolution whole slide images of the H&E stains and corresponding annotations can be viewed on the JPL LabCAS portal (digital object identifiers included in the Supplementary Table 2.1). Images corresponding to the stitched field of views of the region of interest in the multiplex immunohistochemistry are made available as multilayered tiff files on the JPL LabCAS portal https://doi.org/10.48577/rrry-pj94 (UVM) and https://doi.org/10.48577/3gns-rn74 (UCSD).

2.6 Figures

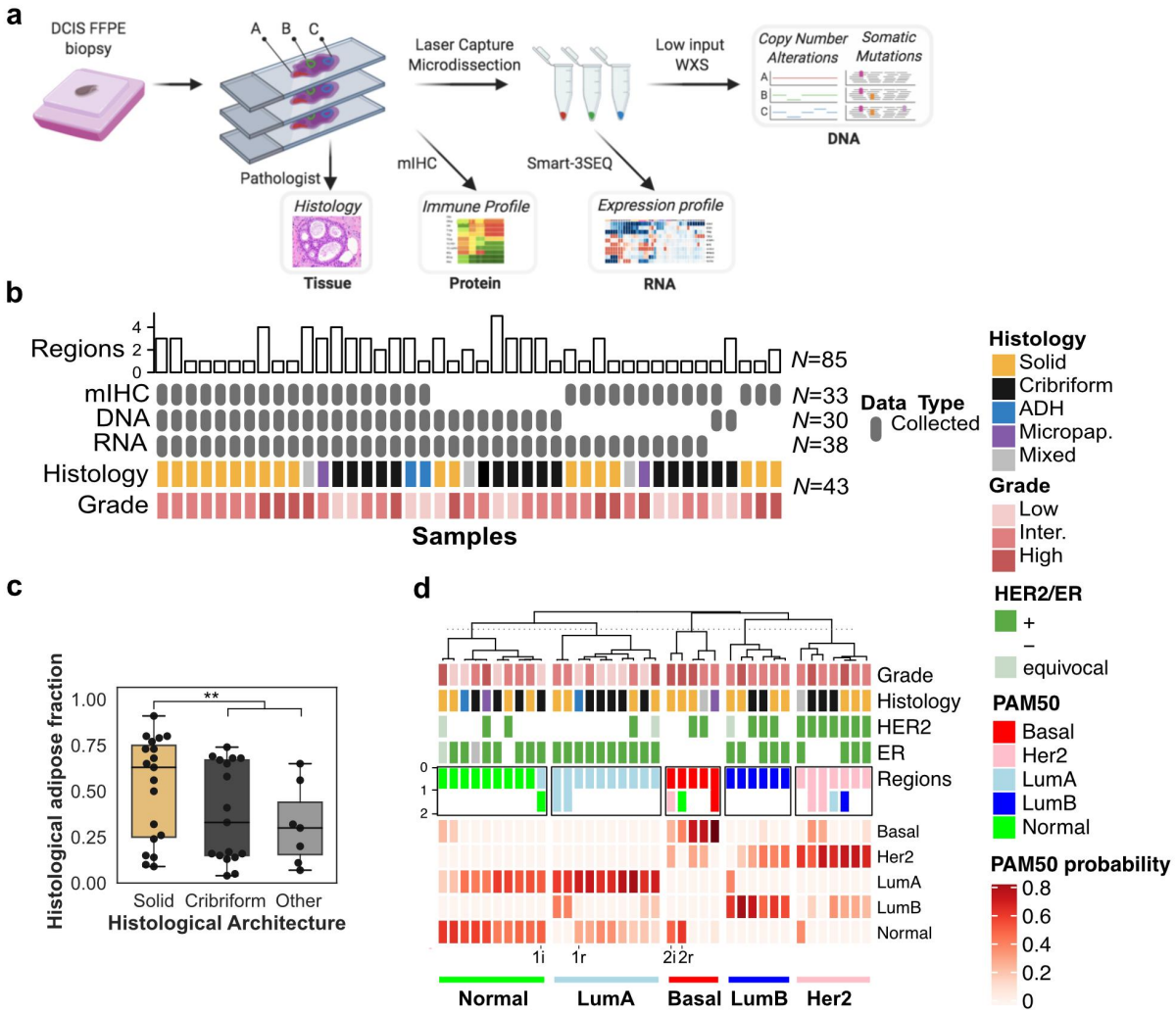


Figure 2.1. Study design and cohort overview.
(a) Archival sample processing and analysis workflow including histology (H&E and mIHC) and microbiopsy-derived whole exome (WXS) and whole transcriptome (Smart-3SEQ) profiling. **(b)** Study cohort overview including histological characteristics (colored rows), data type (grey rows) and number of histological regions (bar chart) investigated. **(c)** Estimate of the fraction of adipose area in H&E images in epithelium of the mIHC images according to each histological architecture, **p<0.01, Mann-Whitney U test. **(d)** Sample classification according to the probabilities of each PAM50 expression subtype. For 10 eligible samples, the intrinsic subtype of a spatially distinct region is indicated. Two patients with recurrence (r) and index (i) samples are indicated at the bottom.

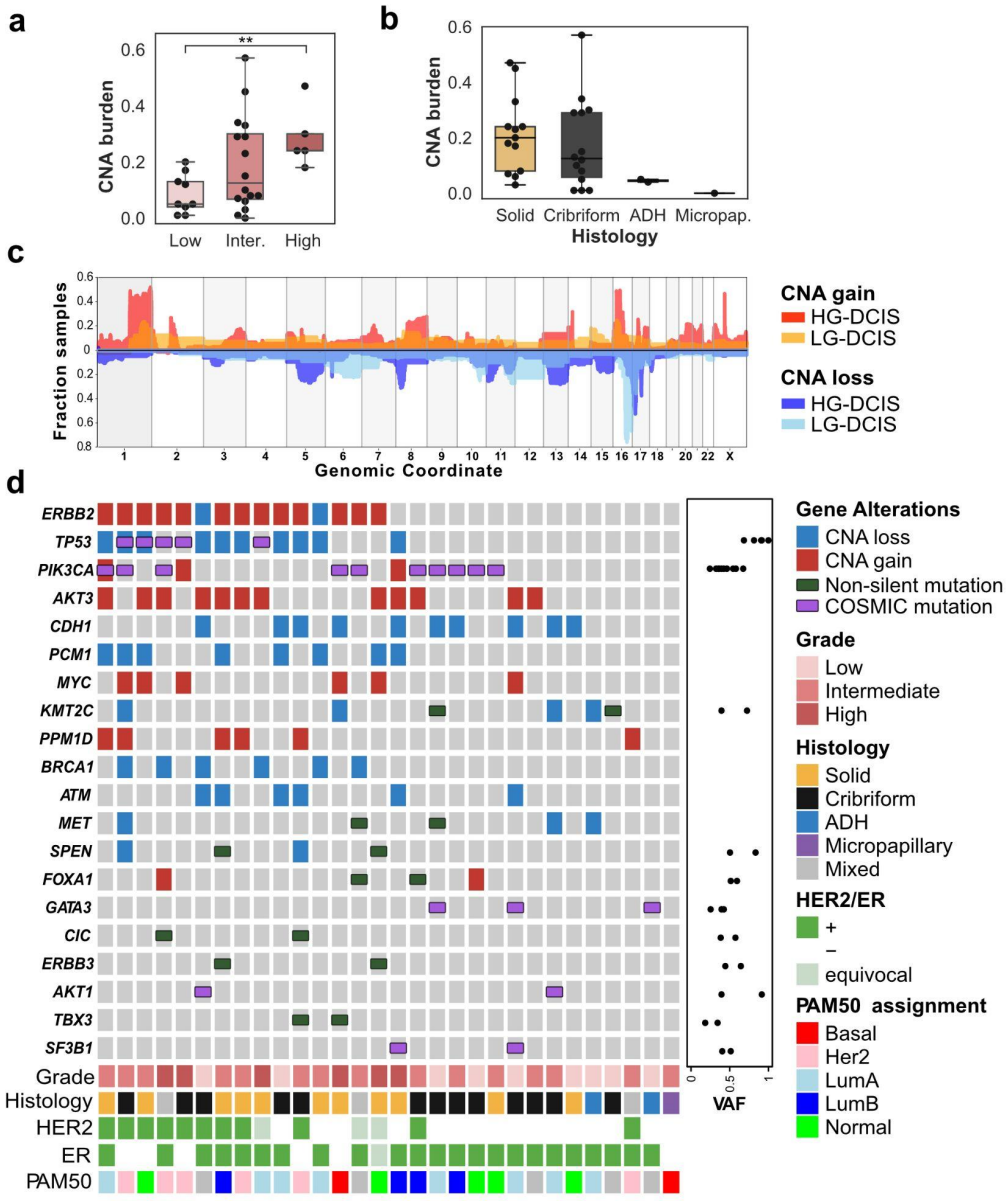


Figure 2.2. Pure DCIS genomic landscape.
(a-b) CNA burden (fraction of base pairs involved in copy number gain or loss) as a function of grade (a) and (b) histological architecture, **p<0.05, ANOVA. **(c)** Smoothed frequency (y-axis) of CNA gains (top) and losses (bottom) smoothed along the genome (x-axis) for HG-DCIS (N=22 - dark colors) and LG-DCIS (N=5 light colors). **(d)** Oncoprint diagram displaying the mutational status of driver genes commonly altered in breast cancer. Genes were included if they were mutated in at least 2 patients or located in a CNA segment present in at least 6 patients, and ordered by frequency of alteration. The variant allele fraction (VAF) of mutations (right panel) and histological characteristics (bottom panel) are indicated.

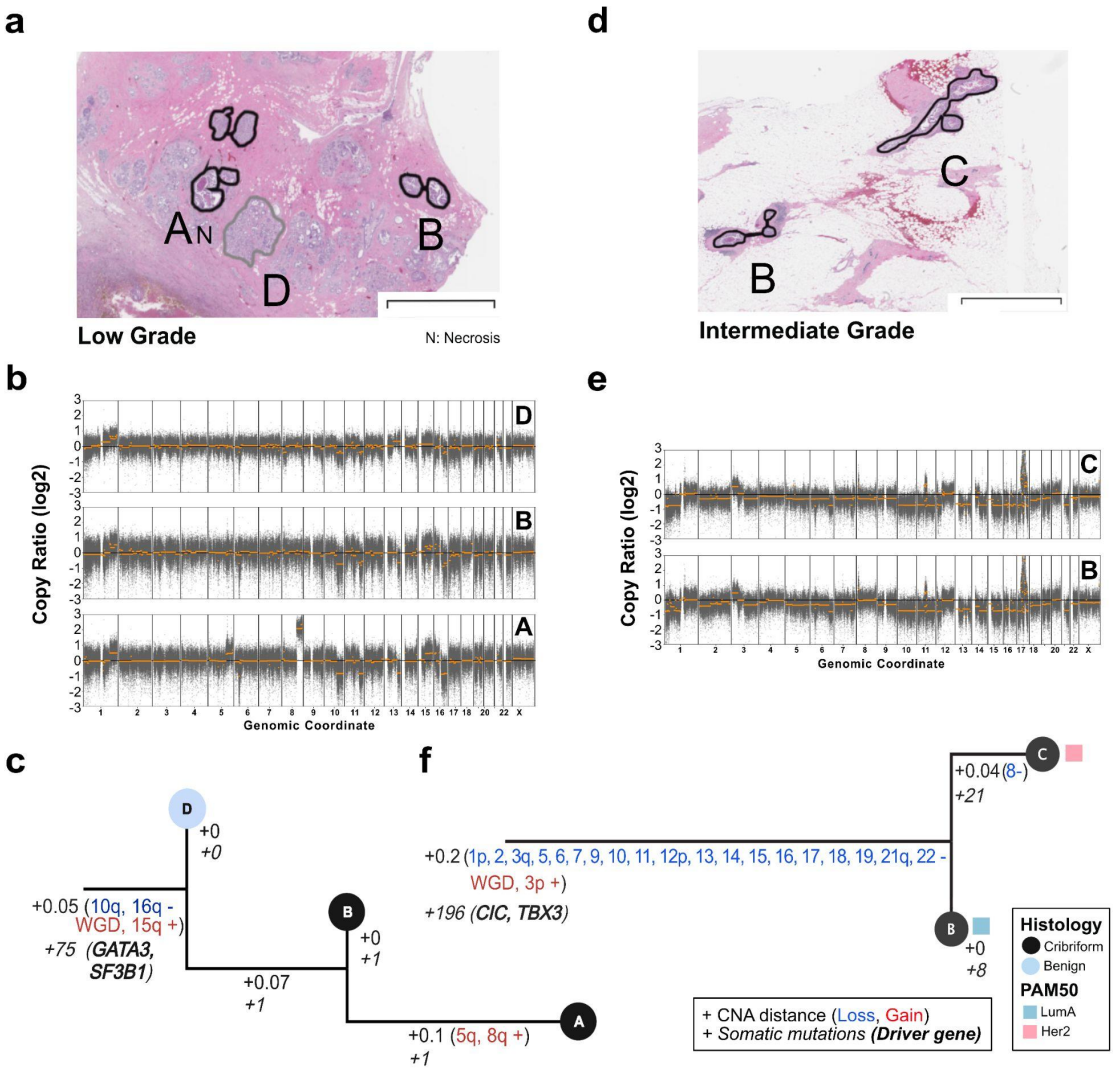


Figure 2.3. Clonal relationships of multi-region DCIS.
Multi-region phylogenetic reconstruction using both CNA and somatic mutations for MCL76_061_16200 **(a-c)** and MCL76_077_15300 **(d-f)**. For each case, the spatial annotation of the microdissected regions on the H&E images (a,d), corresponding copy number profiles (b,e) and phylogenetic trees (c,f) are displayed. Copy number profile plots show bins (grey dots) and segment (orange) $\log_2$ copy number ratio (y-axis). The phylogenetic tree leaves (single dissected region) are colored according to histological type and the branches (hamming distances based on CNA segments) annotated with corresponding specific somatic alterations or their total number (CNA: regular, genes: italic font). The tree root corresponds to an inferred normal diploid ancestor. PAM50 subtype of the region is indicated when available. Annotations and trees are available for 10 additional samples in Supplementary Figure 2.4. The scale bars in panel **a** and **b** correspond to a size of 3mm.

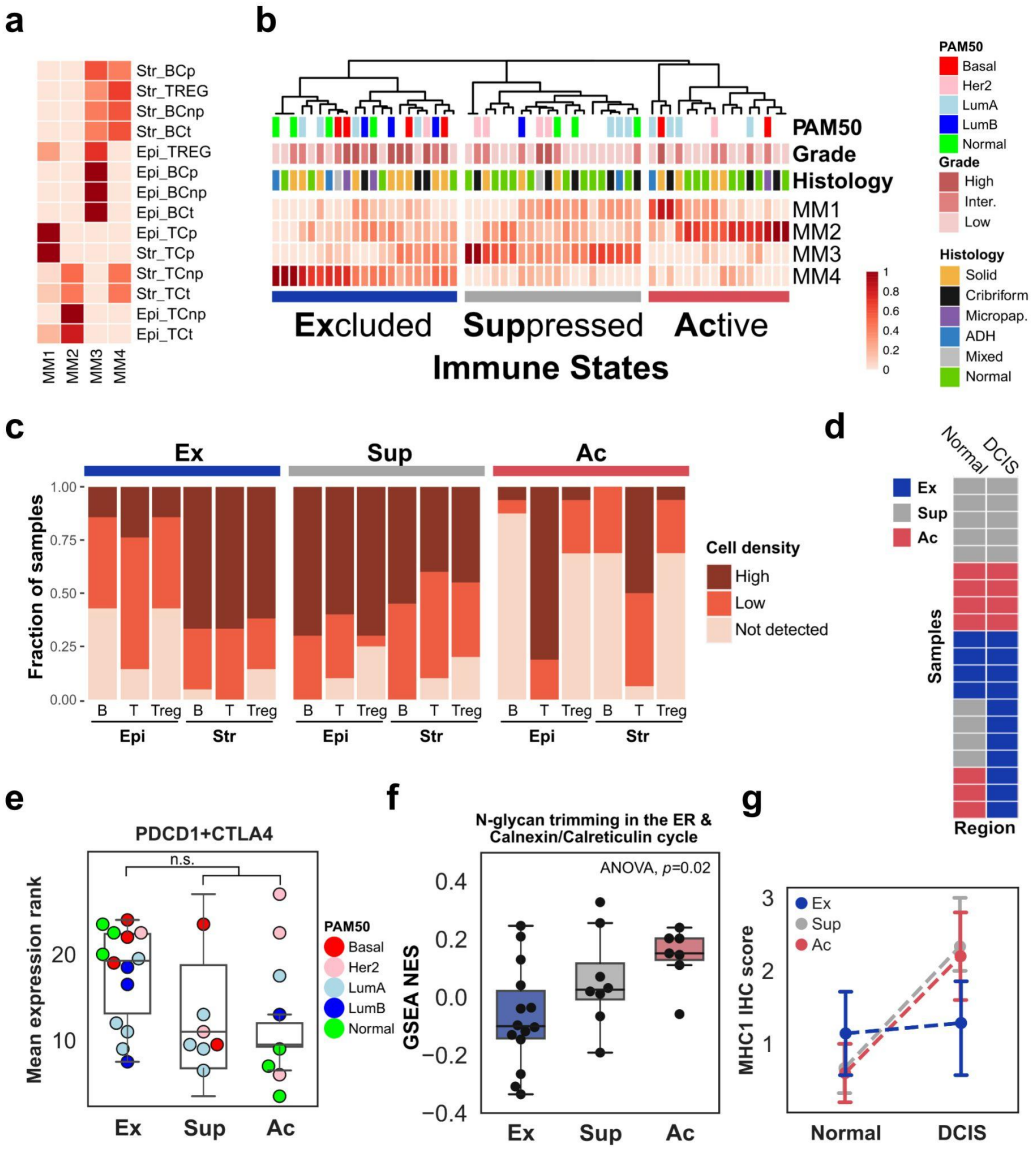


Figure 2.4. Characterization of the immune landscape.
Decomposition of immune cell density scores by non-negative matrix factorization (NMF) into **(a)** W-matrix which shows the composition of the Meta-Markers (columns MM1-4) according to the densities scores (red scale) of each cell type (BC: B-cells, TC: T-cells, TREG: regulatory T-cells), proliferative state (p: Ki67+, np: KI67-, t: total) and regional location (Epi: Epithelium, Str: Stroma) and **(b)** H-matrix which classifies normal and DCIS regions into 3 immune states according to Meta-Markers. **(c)** Fraction of stromal and epithelial regions from samples in each immune-state with high, low or no T-cells (T), B-cells (B) and regulatory T-cells (Treg) densities. **(d)** Immune-state comparison in 20 samples (rows) with matching normal (left column) and DCIS (right column) regions. **(e)** Expression of immune checkpoint receptors genes, PDCD1 and CTLA-4 in each immune state. **(f)** GSEA normalized enrichment score (NES) for a Reactome gene set across immune states. **(g)** Distribution of the expression of the MHC-I complex scored by immunohistochemical staining in DCIS and normal adjacent regions. The median scores of the adjacent and DCIS region in each immune state are connected with a dotted line.

2.7 Tables

Table 2.1. Clinical and pathological features of the patient and specimen studied.

Patient ID	Block ID	Age at Index	Size (cm)	Laterality	Grade	Architecture	ER	HER2[1]	N. Regions	Diagnosis Order[2]
MCL76_044	12800	56	0.9	Left	Low	Cribriform	+	-	1	Index
MCL76_049	18100	50	6	Left	Low	Cribriform	+	-	1	Index
MCL76_060	16100	47	17	Left	Low	Cribriform	+	-	3	Index
MCL76_061	16200	34	8	Left	Low	Cribriform	+	-	4	Index
MCL76_064	15200	78	0.4	Left	Low	Cribriform	+	-	3	Recur. (+18 mos.)
MCL76_066	14400	70	1.1	Right	Low	Cribriform	+	-	3	Index
	16500	70	0.3	Right	Low	ADH	+	-	1	Recur. (+14 mos.)
MCL76_076	15700	45	4.1	Right	Low	Cribriform	+	-	5	Index
MCL76_078	15500	68	1.4	Left	Low	Solid	+	-	3	Index
MCL76_080	15800	59	3.7	Left	Low	ADH	+	-	3	Index
MCL78_020	10001	59	0.3	Right	Low	Cribriform	+	+	1	Index
MCL76_012	11600	50	3.6	Right	Inter.	Solid	+	-	2	Index
MCL76_048	13100	51	3.8	Right	Inter.	Cribriform	+	-	3	Index
MCL76_064	14600	78	NA	Left	Inter.	Solid	+	NA	1	Index
MCL76_067	16600	54	6	Left	Inter.	Solid	+	+	3	Index
MCL76_070	16400	69	8	Right	Inter.	Solid	+	-	3	Index
MCL76_071	14800	68	5.8	Right	Inter.	Micropapillary	-	-	3	Index
MCL76_074	14700	45	14	Right	Inter.	Cribriform	+	-	3	Index
MCL76_077	15300	70	1.2	Left	Inter.	Cribriform	-	+	2	Index
MCL76_079	15400	62	3.4	Right	Inter.	Cribriform	-	+	3	Index
MCL78_001	10001	50	2.5	Right	Inter.	Cribriform	NA	-	1	Index
MCL78_002	10001	48	2	Left	Inter.	Solid	+	-	1	Index
MCL78_006	10001	75	4	Left	Inter.	Cribriform	+	+	1	Index
MCL78_007	10001	43	1.6	Right	Inter.	Cribriform	+	+	1	Index
MCL78_008	10001	66	1.5	Right	Inter.	Solid	+	+	1	Index
MCL78_009	10001	78	2.4	Right	Inter.	Solid	+	-	1	Index
MCL78_010	10001	67	1.1	Left	Inter.	Cribriform	-	+	1	Index
MCL78_011	10001	59	0.6	Right	Inter.	Solid	+	+	1	Index
MCL78_013	10001	63	2.2	Right	Inter.	Mixed	-	+	1	Index
MCL78_016	10001	65	4.5	Left	Inter.	Solid	-	+	1	Index
MCL78_017	10001	52	2.5	Left	Inter.	Solid	+	+	1	Index
MCL78_018	10001	65	2	Right	Inter.	Mixed	+	equ	2	Index
MCL76_007	11000	78	3.5	Left	High	Solid	-	-	3	Index
	11100	78	2.6	Right	High	Solid	-	-	4	Recur. (+39 mos.)
MCL76_016	11800	35	5	Left	High	Mixed	+	+	4	Index
MCL76_025	16800	75	1.2	Right	High	Solid	-	NA	2	Index
MCL76_068	14900	59	9.5	Left	High	Cribriform	-	+	3	Index
MCL78_003	10001	43	5	Left	High	Solid	+	equ	1	Index
MCL78_005	10001	81	0.5	Right	High	Solid	+	-	1	Index
MCL78_012	10001	54	4	Right	High	Solid	-	+	1	Index
	10014	54	3	Right	High	Solid	-	NA	1	Synchronous
MCL78_015	10001	57	0.5	Left	High	Micropapillary	+	+	1	Index
MCL78_019	10001	57	1.9	Left	High	Solid	-	equ	1	Index

1. Inferred from ERBB2 copy number and expression (Figure S1), equ=equivocal
2. Recur.=Recurrence, mos.=months. All recurrence DCIS were in different quadrants than the index.

Table 2.2. Frequency of *PIK3CA*, *TP53* and *GATA3* driver mutations in previously reported DCIS studies and pure DCIS in this study.

Gene	Pang et al. 2017 (N=20)	Lin et al. 2019 (N=65)	Nagasawa et al. 2021 (N=72)	Pareja et al. 2020 (N=7)	This study					
					All[1]	Grade		Histology		
						Low	Inter.-High	Cribriform	Solid	Other
PIK3CA	55%	40%	50%	0%	43% (10/23)	29% (2/7)	50% (8/16)	55% (6/11)	40% (4/10)	0% (0/2)
TP53	30%	13.8%	21%	14.30%	31.3% (5/16)	0% (0/4)	41.7% (5/12)	33.3% (3/9)	40% (2/5)	0% (0/2)
GATA3	45%	13.8%	56%	28.60%	20% (3/15)	75% (3/4)	0% (0/11)	33.3% (2/6)	0% (0/9)	50% (1/2)

1. The denominator represents samples with at least 20x coverage across the targeted regions

2.8 Supplemental Data, Tables and Figures

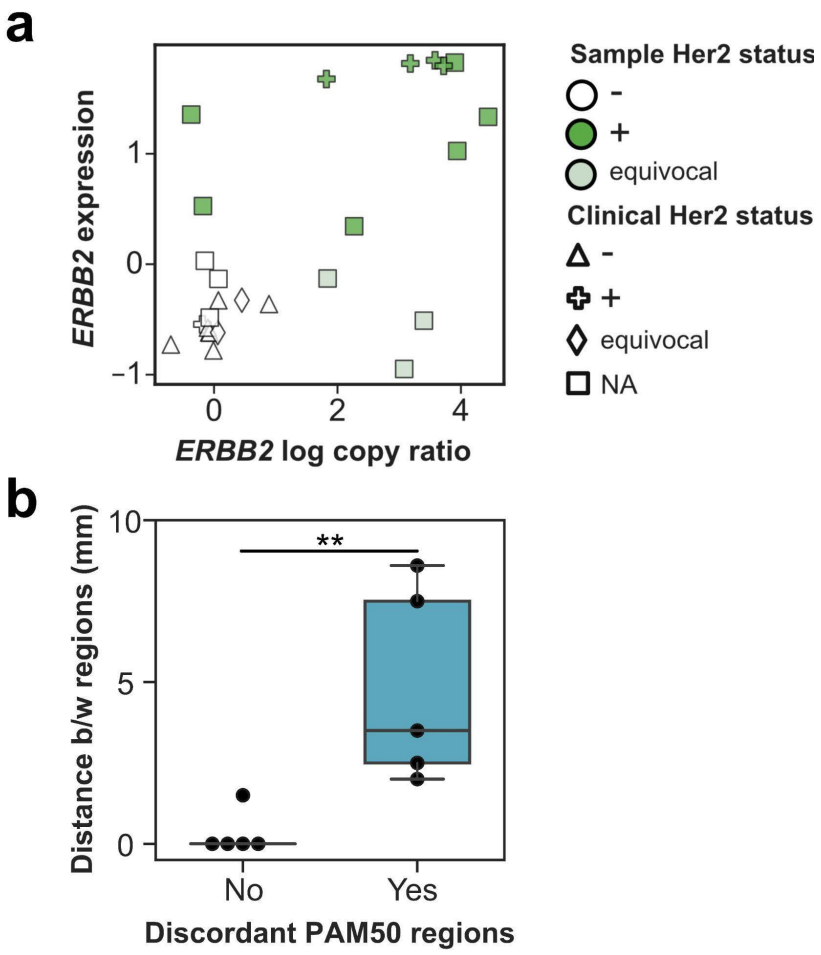


Figure S2.1. Pure DCIS characterization.
(a) Estimation of Her2 status. The DNA-based log2 copy number ratio (x-axis) and RNA-based expression level (y-axis) of *ERBB2* gene are displayed for 26 samples with both data available. **(b)** PAM50 discordance between regions in relationship with spatial distance between regions, **p<0.01, Mann-Whitney U test.

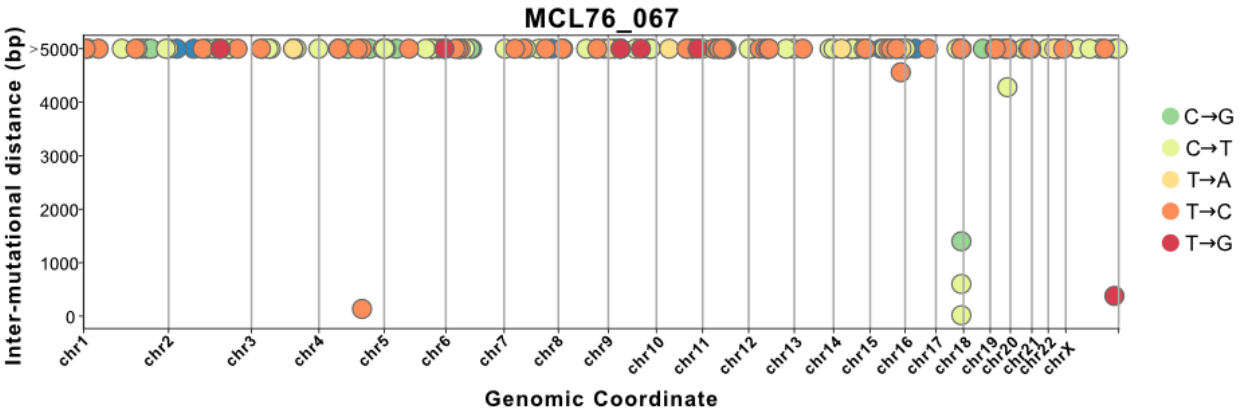


Figure S2.2. Likely kataegis event in MCL76_067_16600 in chromosome 17.
Along the genome coordinate (x-axis), the relative distance between proximal mutations (y-axis) as well as their substitution type (colors) are indicated.

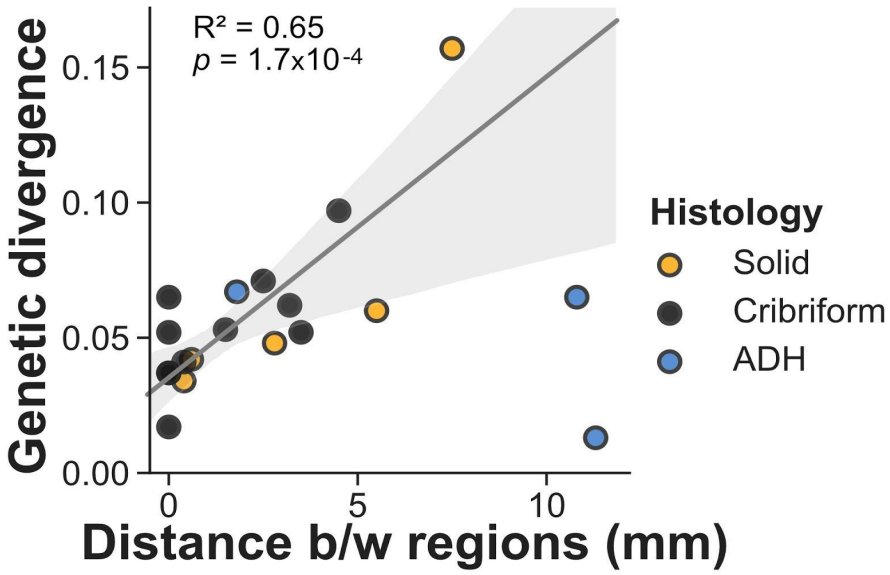


Figure S2.3. Genetic divergence in multi-region DCIS.
CNA-based genetic divergence (y-axis) of each pair of histologically concordant regions (dot) as a function of the minimum physical distance between them (x-axis), colored by their histology. Linear regression line fit of DCIS samples shown in solid dark gray line, with 95% confidence interval estimate based on bootstrapping in light gray.

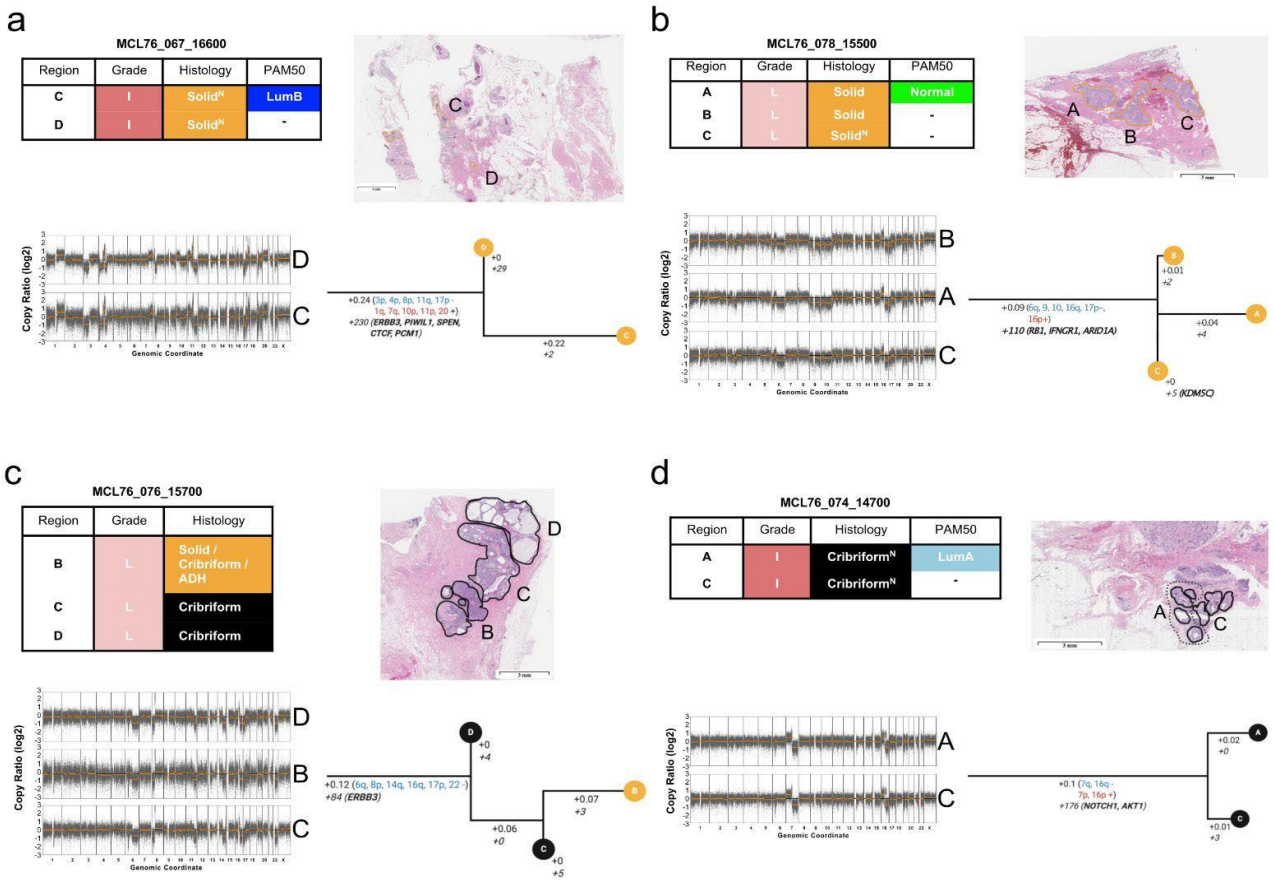


Figure S2.4. Phylogenetic trees for multi-region DCIS samples.
(a-j) Clonal reconstruction using CNA and somatic mutations in 25 related regions across 10 samples. In each panel, *top left:* a table describing the name, nuclear grade and histological architecture of each region in a sample is shown, (necrosis is indicated with N); *top right:* shows an H&E image of the sample with dissected regions drawn on the image; *bottom left*: Copy number profiles for each region in the sample, genomic coordinates are indicated on the X-axis and the $\log_2$ copy ratio on the Y-axis. Bins are indicated in dark-grey and segments in orange; *bottom right*: Phylogenetic tree for the sample with leaf nodes indicating a single dissected region colored by histology. The tree is rooted to a normal diploid ancestor. Branch lengths are hamming distances based on CNA segments. Branches are labeled with 1) CNA-based branch length, and, when available, 2) arm-level CNA losses (blue) and gains (red) and 3) somatic mutation number (italic) with mutations in breast cancer driver genes indicated (black: high coverage, grey: low coverage). CNA smaller than arms, or on driver genes are not displayed.

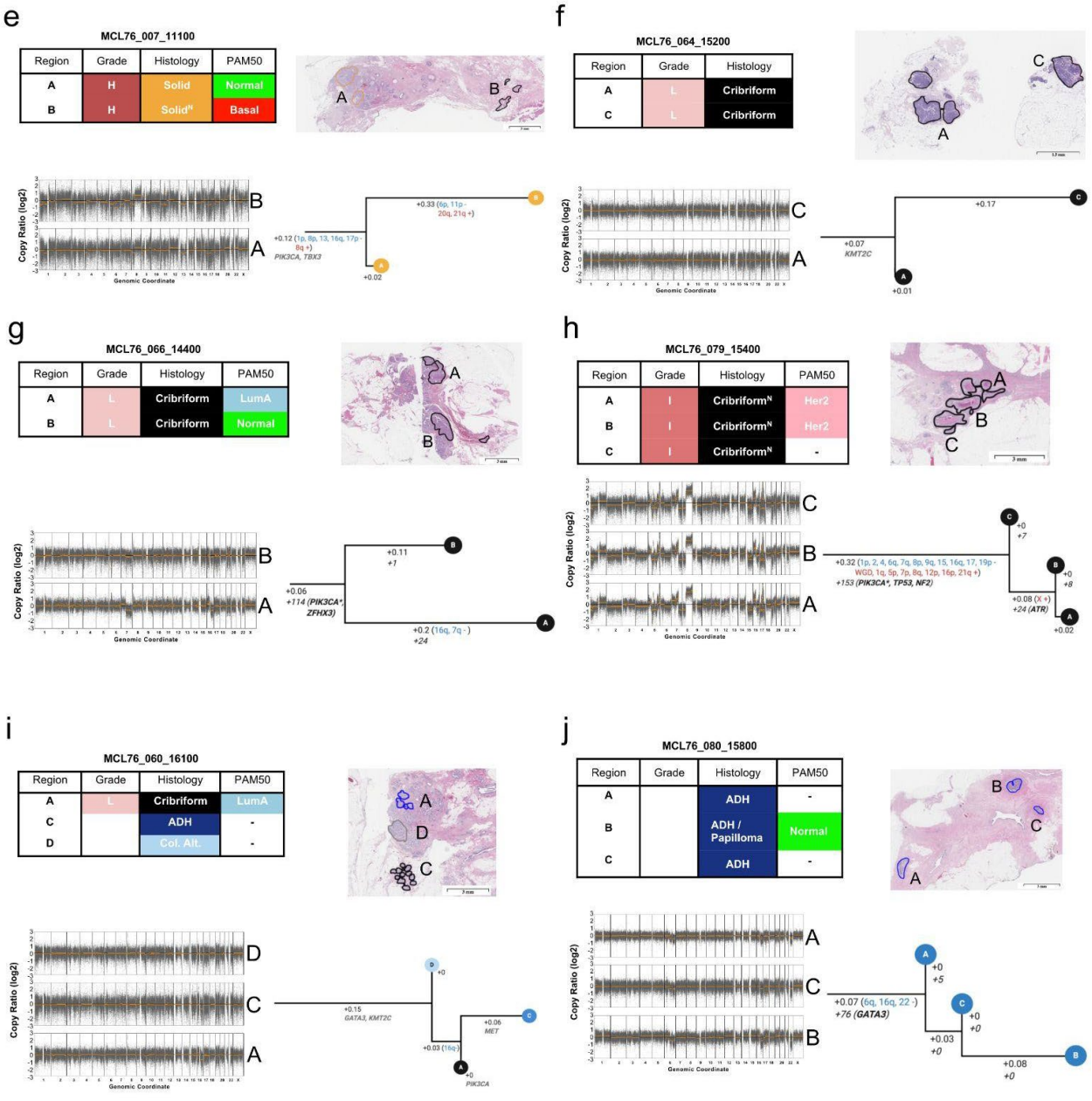


Figure S2.4. Phylogenetic trees for multi-region DCIS samples. *(cont.)*

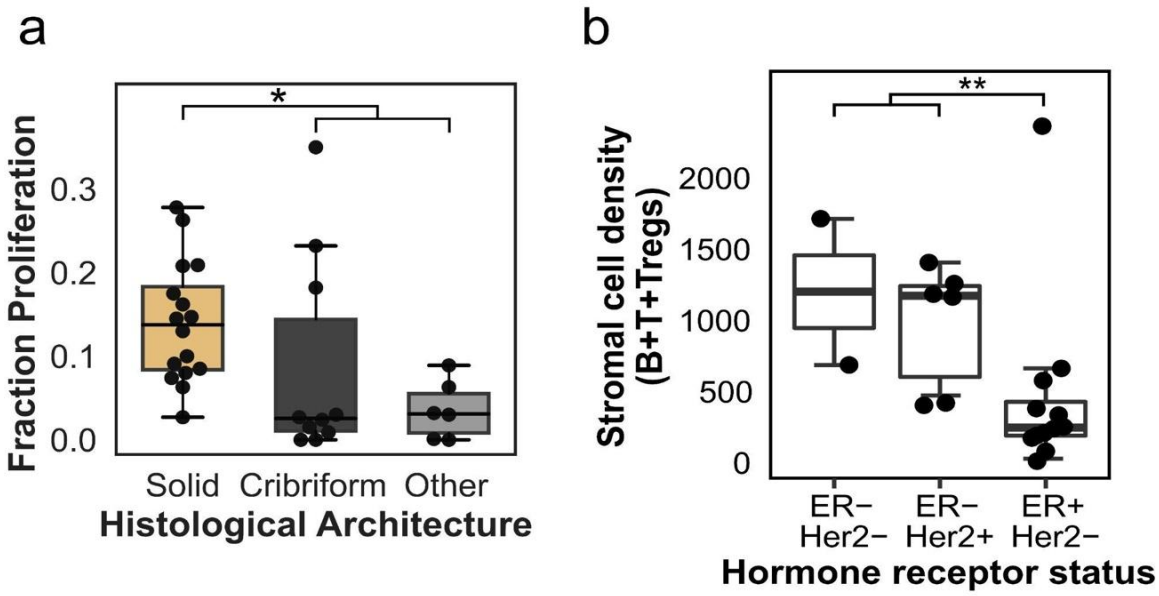


Figure S2.5. Relationships between DCIS mIHC-derived and histological / hormonal features.

(a) Estimates of the fraction of Ki67+ cytokeratin positive cells in the epithelium of the mIHC images according to each histological architecture. *p<0.05, Mann-Whitney U test. **(b)** Stromal immune cell density differences between DCIS subtypes. B=B-cells, T=T-cells and Tregs=Regulatory T-cells. **p<0.01, Mann-Whitney U test.

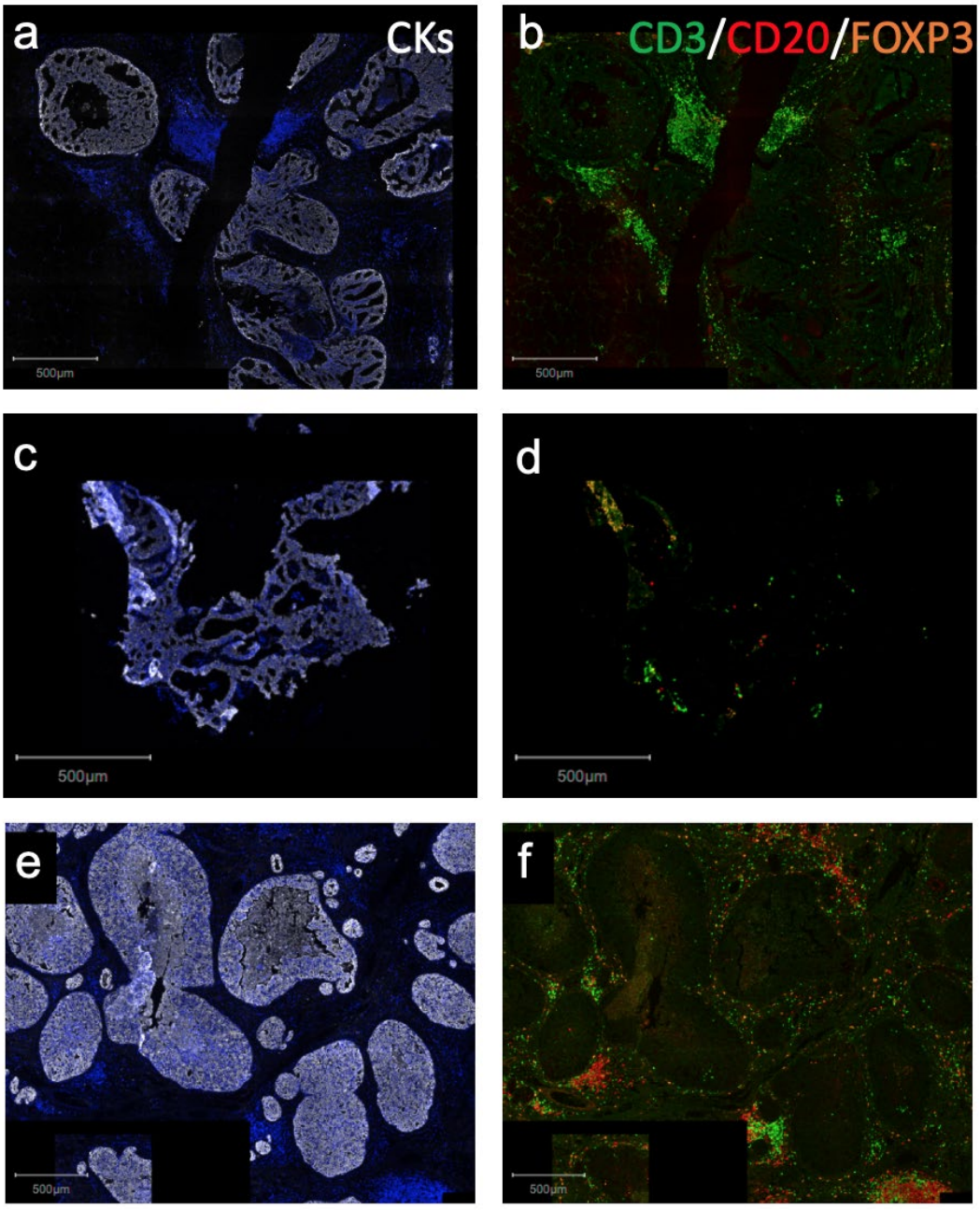


Figure S2.6. Multiplex immuno-fluorescent images representative of the three immune-states.
The pan-cytokeratin (white) and nuclear (DAPI - blue) staining **(a,c,e)** and the matching CD3 (T-cells, green), CD20 (B-cells, red) and FOXP3 (T-reg, orange) stainings **(b,d,f)** are shown for specimen representative of the Active (a,b, MCL78_013_10001), Suppressed (c,d, MCL76_049_18100) and Excluded (e,f,MCL76_074_14700) states.

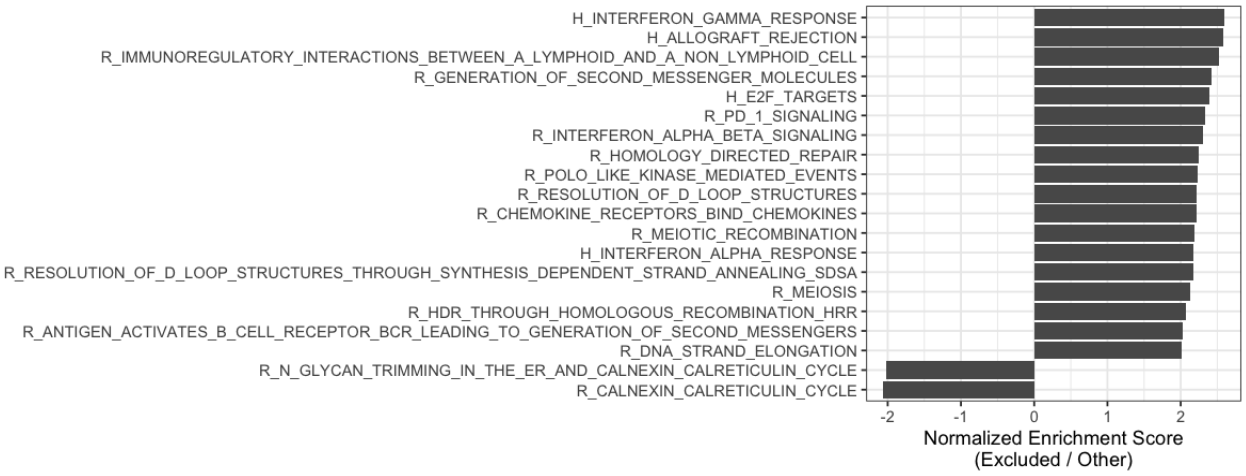


Figure S2.7. Gene sets significantly deregulated in the epithelium of regions in the Excluded immune state.
All Hallmark (H) and Reactome (R) genesets were tested. Genesets with an absolute normalized enrichment score greater than 2 and with FDR less than 0.05 are represented.

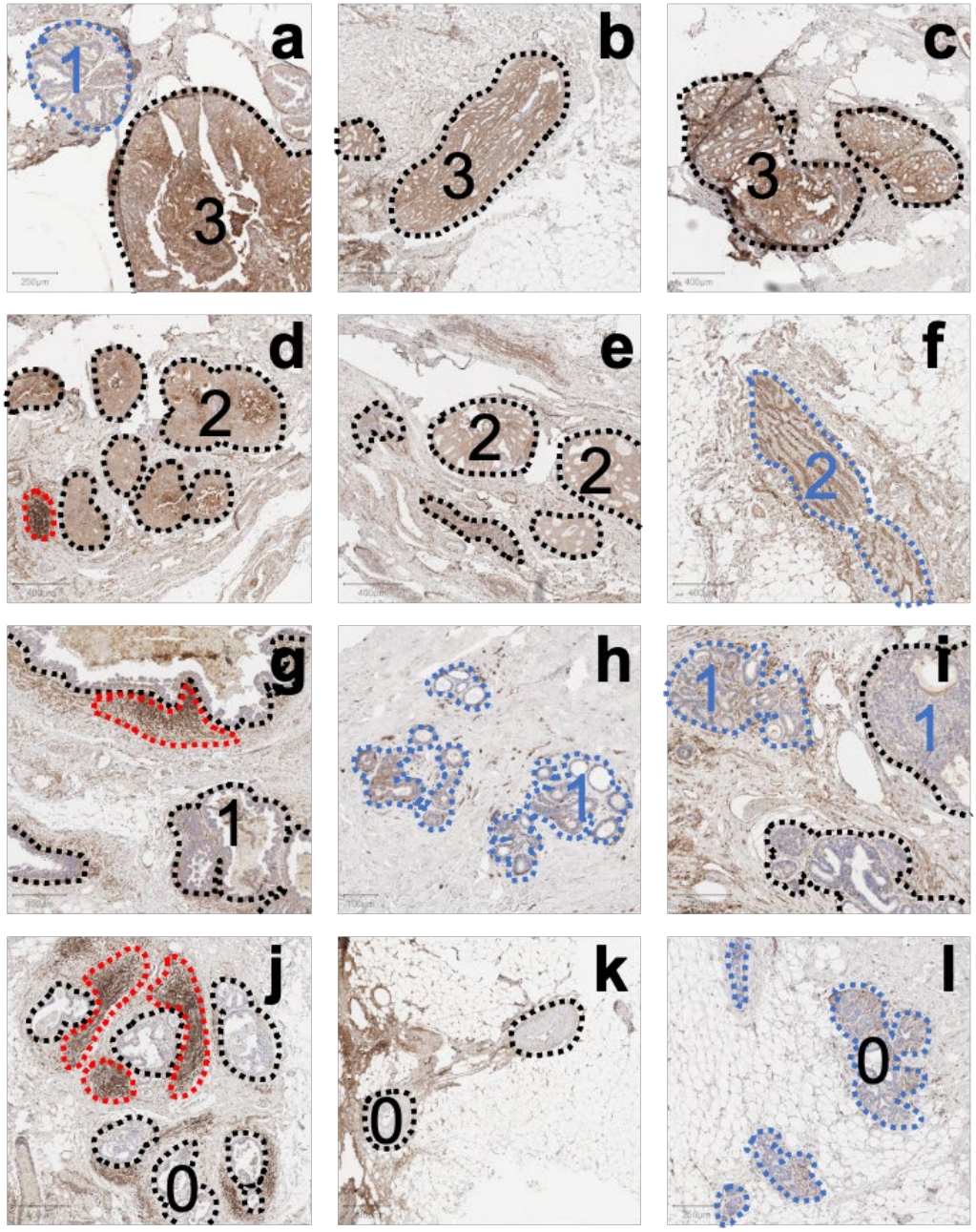


Figure S2.8. MHC1 immunostaining scoring.
Selected areas of DCIS (black) or normal ducts (blue) are indicated together with their associated expression score from 3 (panel **a-c**), 2 (**d-f**), 1 (**g-i**), 0 (**j-l**). Additional MHC1 high lymphocytes areas are indicated in red. Scale bar indicated in bottom left.

2.9 Author Contributions

J.L.S., G.S.S., G.L.H., L.J.E., A.D.B., and O.H. designed the study, F.H., A.D.B., D.L.W., B.L.S., and T.O.K. selected and annotated the specimen. J.S., K.J., H.Y., H.M., M.F.E., and J.G. generated the data, O.H., D.N., A.O., and H.M. analyzed the data, O.H. and D.N. wrote the manuscript. All authors reviewed and approved the manuscript.

2.10 Acknowledgements

We are grateful to Drs. Michael Campbell, Christina Yau, and Kathleen Curtius for helpful conversations, Drs. Alfredo Molinolo, Oluwole Fadare, Sharmeela Kaushal, Valeria Estrada, and Mrs. Kimberly Mcintyre for their support and assistance in the tissue collection, preparation, and dissection, Mrs. Eliza Jeong, Marcy Andersen, and Nicole Lee for their assistance retrieving clinical information. We thank the technical assistance of the Vermont Integrative Genomics Resource Massively Parallel Sequencing Facility with the combined support of the University of Vermont Cancer Center, Lake Champlain Cancer Research Organization, UVM College of Agriculture and Life Sciences, and the UVM Larner College of Medicine. We acknowledge the work of all working groups from the NCI Consortium for the Molecular Characterization of Screen Detected Lesions (MCL), in particular Christopher Amos, Daniel Crichton, Heather Kincaid, Kirsten Anton, Luca Cinquini, and David Liu for their assistance in the data management and sharing. Figure 2.1 was created with BioRender.com and printed with permission. This work is supported by funding from the National Institute of Health (U01CA196406, U01CA196406-03S1, U01CA196383, R01DE026644, T32GM008806, T15LM011271), the National Cancer Institute (P30CA023100), and the California Tobacco-Related Disease Research Program pre-doctoral fellowship to DN (28DT-0011). The funding bodies had no role in the design of the study; collection, analysis, and interpretation of data; or in the writing of the manuscript.

Chapter 2, in full, is a reformatted presentation of the material as it appears as “The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ.” by Daniela Nachmanson, Adam Officer, Hidetoshi Mori, Jonathan Gordon, Mark F. Evans, Joseph Steward, Huazhen Yao, Thomas O’Keefe, Farnaz Hasteh, Gary S. Stein, Kristen Jepsen, Donald L. Weaver, Gillian L. Hirst, Brian L. Sprague, Laura J. Esserman, Alexander D.

Borowsky, Janet L. Stein, Olivier Harismendy. The dissertation author was one of the primary investigators and author of this material.

2.11 References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.* **70,** 7–30 (2020).

2. Bleyer, A. & Welch, H. G. Effect of three decades of screening mammography on breast-cancer incidence. *N. Engl. J. Med.* **367,** 1998–2005 (2012).

3. Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. *Lancet* **380,** 1778–1786 (2012).

4. Sprague, B. L., Vacek, P. M., Herschorn, S. D., James, T. A., Geller, B. M., Trentham-Dietz, A., Stein, J. L. & Weaver, D. L. Time-varying risks of second events following a DCIS diagnosis in the population-based Vermont DCIS cohort. *Breast Cancer Res. Treat.* **174,** 227–235 (2019).

5. Gorringe, K. L. & Fox, S. B. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. *Front. Oncol.* **7,** 248 (2017).

6. Pang, J.-M. B., Savas, P., Fellowes, A. P., Mir Arnau, G., Kader, T., Vedururu, R., Hewitt, C., Takano, E. A., Byrne, D. J., Choong, D. Y., Millar, E. K., Lee, C. S., O'Toole, S. A., Lakhani, S. R., Cummings, M. C., Mann, G. B., Campbell, I. G., Dobrovic, A., Loi, S., Gorringe, K. L. & Fox, S. B. Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. *Mod. Pathol.* **30,** 952–963 (2017).

7. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J. & Forbes, S. A. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47,** D941–D947 (2019).

8. Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M. & Bernard, P. S. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (2009).

9. Lin, C.-Y., Vennam, S., Purington, N., Lin, E., Varma, S., Han, S., Desa, M., Seto, T., Wang, N. J., Stehr, H., Troxell, M. L., Kurian, A. W. & West, R. B. Genomic landscape of ductal carcinoma in situ and association with progression. *Breast Cancer Res. Treat.* **178,** 307–316 (2019).

10. Nagasawa, S., Kuze, Y., Maeda, I., Kojima, Y., Motoyoshi, A., Onishi, T., Iwatani, T., Yokoe, T., Koike, J., Chosokabe, M., Kubota, M., Seino, H., Suzuki, A., Seki, M., Tsuchihara, K., Inoue, E., Tsugawa, K., Ohta, T. & Suzuki, Y. Genomic profiling reveals heterogeneous populations of ductal carcinoma in situ of the breast. *Commun Biol* **4,** 438 (2021).

11. Pareja, F., Brown, D. N., Lee, J. Y., Da Cruz Paula, A., Selenica, P., Bi, R., Geyer, F. C.,

Gazzo, A., da Silva, E. M., Vahdatinia, M., Stylianou, A. A., Ferrando, L., Wen, H. Y., Hicks, J. B., Weigelt, B. & Reis-Filho, J. S. Whole-Exome Sequencing Analysis of the Progression from Non-Low-Grade Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. *Clin. Cancer Res.* **26,** 3682–3693 (2020).

12. Abba, M. C., Gong, T., Lu, Y., Lee, J., Zhong, Y., Lacunza, E., Butti, M., Takata, Y., Gaddis, S., Shen, J., Estecio, M. R., Sahin, A. A. & Aldaz, C. M. A Molecular Portrait of High-Grade Ductal Carcinoma In Situ. *Cancer Res.* **75,** 3980–3990 (2015).

13. Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., Casasent, T., Meric-Bernstam, F., Edgerton, M. E. & Navin, N. E. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172,** 205–217.e12 (2018).

14. Gerdes, M. J., Gökmen-Polar, Y., Sui, Y., Pang, A. S., LaPlante, N., Harris, A. L., Tan, P.-H., Ginty, F. & Badve, S. S. Single-cell heterogeneity in ductal carcinoma in situ of breast. *Mod. Pathol.* **31,** 406–417 (2018).

15. Pruneri, G., Lazzeroni, M., Bagnardi, V., Tiburzio, G. B., Rotmensz, N., DeCensi, A., Guerrieri-Gonzaga, A., Vingiani, A., Curigliano, G., Zurrida, S., Bassi, F., Salgado, R., Van den Eynden, G., Loi, S., Denkert, C., Bonanni, B. & Viale, G. The prevalence and clinical relevance of tumor-infiltrating lymphocytes (TILs) in ductal carcinoma in situ of the breast. *Ann. Oncol.* **28,** 321–328 (2017).

16. Campbell, M. J., Baehner, F., O'Meara, T., Ojukwu, E., Han, B., Mukhtar, R., Tandon, V., Endicott, M., Zhu, Z., Wong, J., Krings, G., Au, A., Gray, J. W. & Esserman, L. Characterizing the immune microenvironment in high-risk ductal carcinoma in situ of the breast. *Breast Cancer Res. Treat.* **161,** 17–28 (2017).

17. Trinh, A., Gil Del Alcazar, C. R., Shukla, S. A., Chin, K., Chang, Y. H., Thibault, G., Eng, J., Jovanović, B., Aldaz, C. M., Park, S. Y., Jeong, J., Wu, C., Gray, J. & Polyak, K. Genomic Alterations during the In Situ to Invasive Ductal Breast Carcinoma Transition Shaped by the Immune System. *Mol. Cancer Res.* **19,** 623–635 (2021).

18. Lesurf, R., Aure, M. R., Mørk, H. H., Vitelli, V., Oslo Breast Cancer Research Consortium (OSBREAC), Lundgren, S., Børresen-Dale, A.-L., Kristensen, V., Wärnberg, F., Hallett, M. & Sørlie, T. Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Rep.* **16,** 1166–1179 (2016).

19. Gil Del Alcazar, C. R., Huh, S. J., Ekram, M. B., Trinh, A., Liu, L. L., Beca, F., Zi, X., Kwak, M., Bergholtz, H., Su, Y., Ding, L., Russnes, H. G., Richardson, A. L., Babski, K., Min Hui Kim, E., McDonnell, C. H., 3rd, Wagner, J., Rowberry, R., Freeman, G. J., Dillon, D., Sorlie, T., Coussens, L. M., Garber, J. E., Fan, R., Bobolis, K., Allred, D. C., Jeong, J., Park, S. Y., Michor, F. & Polyak, K. Immune Escape in Breast Cancer During In Situ to Invasive Carcinoma Transition. *Cancer Discov.* **7,** 1098–1115 (2017).

20. Allen, M. D., Marshall, J. F. & Jones, J. L. αvβ6 Expression in myoepithelial cells: a novel marker for predicting DCIS progression with therapeutic potential. *Cancer Res.* **74,** 5942–5947 (2014).

21. Delort, L., Cholet, J., Decombat, C., Vermerie, M., Dumontet, C., Castelli, F. A., Fenaille, F., Auxenfans, C., Rossary, A. & Caldefie-Chezet, F. The Adipose Microenvironment Dysregulates the Mammary Myoepithelial Cells and Could Participate to the Progression of Breast Cancer. *Front Cell Dev Biol* **8,** 571948 (2020).

22. Allinen, M., Beroukhim, R., Cai, L., Brennan, C., Lahti-Domenici, J., Huang, H., Porter, D., Hu, M., Chin, L., Richardson, A., Schnitt, S., Sellers, W. R. & Polyak, K. Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* **6,** 17–32 (2004).

23. Hu, M., Yao, J., Carroll, D. K., Weremowicz, S., Chen, H., Carrasco, D., Richardson, A., Violette, S., Nikolskaya, T., Nikolsky, Y., Bauerlein, E. L., Hahn, W. C., Gelman, R. S., Allred, C., Bissell, M. J., Schnitt, S. & Polyak, K. Regulation of in situ to invasive breast carcinoma transition. *Cancer Cell* **13,** 394–406 (2008).

24. Unsworth, A., Anderson, R. & Britt, K. Stromal fibroblasts and the immune microenvironment: partners in mammary gland biology and pathology? *J. Mammary Gland Biol. Neoplasia* **19,** 169–182 (2014).

25. Sinha, V. C. & Piwnica-Worms, H. Intratumoral Heterogeneity in Ductal Carcinoma In Situ: Chaos and Consequence. *J. Mammary Gland Biol. Neoplasia* **23,** 191–205 (2018).

26. Nachmanson, D., Steward, J., Yao, H., Officer, A., Jeong, E., O'Keefe, T. J., Hasteh, F., Jepsen, K., Hirst, G. L., Esserman, L. J., Borowsky, A. D. & Harismendy, O. Mutational profiling of micro-dissected pre-malignant lesions from archived specimens. *BMC Med. Genomics* **13,** 173 (2020).

27. Foley, J. W., Zhu, C., Jolivet, P., Zhu, S. X., Lu, P., Meaney, M. J. & West, R. B. Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome Res.* **29,** 1816–1825 (2019).

28. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H. & Campbell, P. J. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).

29. Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R. J., Abascal, F., Hall, M. W. J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., Fitzgerald, R. C., Handford, P. A., Campbell, P. J., Saeb-Parsy, K. & Jones, P. H. Somatic mutant clones colonize the human esophagus with age. *Science* **362,** 911–917 (2018).

30. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

31. Bielski, C. M., Zehir, A., Penson, A. V., Donoghue, M. T. A., Chatila, W., Armenia, J., Chang, M. T., Schram, A. M., Jonsson, P., Bandlamudi, C., Razavi, P., Iyer, G., Robson, M. E., Stadler, Z. K., Schultz, N., Baselga, J., Solit, D. B., Hyman, D. M., Berger, M. F. & Taylor, B. S. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50,** 1189–1195 (2018).

32. D'Antonio, M., Tamayo, P., Mesirov, J. P. & Frazer, K. A. Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels. *Cell Reports* **16,** 672–683 (2016).

33. Afzaljavan, F., Sadr, A. S., Savas, S. & Pasdar, A. GATA3 somatic mutations are associated with clinicopathological features and expression profile in TCGA breast cancer patients. *Sci. Rep.* **11,** 1679 (2021).

34. Emmanuel, N., Lofgren, K. A., Peterson, E. A., Meier, D. R., Jung, E. H. & Kenny, P. A. Mutant GATA3 Actively Promotes the Growth of Normal and Malignant Mammary Cells. *Anticancer Res.* **38,** 4435–4441 (2018).

35. Kader, T., Hill, P., Zethoven, M., Goode, D. L., Elder, K., Thio, N., Doyle, M., Semple, T., Sufyan, W., Byrne, D. J., Pang, J.-M. B., Murugasu, A., Miligy, I. M., Green, A. R., Rakha, E. A., Fox, S. B., Mann, G. B., Campbell, I. G. & Gorringe, K. L. Atypical ductal hyperplasia is a multipotent precursor of breast carcinoma. *J. Pathol.* **248,** 326–338 (2019).

36. Kader, T., Elder, K., Zethoven, M., Semple, T., Hill, P., Goode, D. L., Thio, N., Cheasley, D., Rowley, S. M., Byrne, D. J., Pang, J.-M., Miligy, I. M., Green, A. R., Rakha, E. A., Fox, S. B., Mann, G. B., Campbell, I. G. & Gorringe, K. L. The genetic architecture of breast papillary lesions as a predictor of progression to carcinoma. *NPJ Breast Cancer* **6,** 9 (2020).

37. Cai, Y., Crowther, J., Pastor, T., Abbasi Asbagh, L., Baietti, M. F., De Troyer, M., Vazquez, I., Talebi, A., Renzi, F., Dehairs, J., Swinnen, J. V. & Sablina, A. A. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. *Cancer Cell* **29,** 751–766 (2016).

38. Thompson, E., Taube, J. M., Elwood, H., Sharma, R., Meeker, A., Warzecha, H. N., Argani, P., Cimino-Mathews, A. & Emens, L. A. The immune microenvironment of breast ductal carcinoma in situ. *Mod. Pathol.* **29,** 249–258 (2016).

39. Danforth, D. N., Jr. Genomic Changes in Normal Breast Tissue in Women at Normal Risk or at High Risk for Breast Cancer. *Breast Cancer* **10,** 109–146 (2016).

40. Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–1120 (2013).

41. Risom, T., Glass, D. R., Liu, C. C., Rivero-Gutiérrez, B., Baranski, A., McCaffrey, E. F., Greenwald, N. F., Kagel, A., Strand, S. H., Varma, S., Kong, A., Keren, L., Srivastava, S., Zhu, C., Khair, Z., Veis, D. J., Deschryver, K., Vennam, S., Maley, C., Shelley Hwang, E., Marks, J. R., Bendall, S. C., Colditz, G. A., West, R. B. & Angelo, M. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *bioRxiv* 2021.01.05.425362 (2021). doi:10.1101/2021.01.05.425362

42. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A. & Bernstein, B. E. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).

43. Roden, D. L., Baker, L. A., Elsworth, B., Chan, C.-L., Harvey, K., Deng, N., Wu, S. Z., Cazet, A., Nair, R. & Swarbrick, A. Single cell transcriptomics reveals molecular subtype and functional heterogeneity in models of breast cancer. *bioRxiv* 282079 (2018). doi:10.1101/282079

44. Allred, D. C., Wu, Y., Mao, S., Nagtegaal, I. D., Lee, S., Perou, C. M., Mohsin, S. K., O'Connell, P., Tsimelzon, A. & Medina, D. Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. *Clin. Cancer Res.* **14,** 370–378 (2008).

45. Sun, R., Hu, Z. & Curtis, C. Big Bang Tumor Growth and Clonal Evolution. *Cold Spring Harb. Perspect. Med.* **8,** (2018).

46. Polyak, K. Is breast tumor progression really linear? *Clin. Cancer Res.* **14,** 339–341 (2008).

47. Zeng, Z., Vo, A., Li, X., Shidfar, A., Saldana, P., Blanco, L., Xuei, X., Luo, Y., Khan, S. A. & Clare, S. E. Somatic genetic aberrations in benign breast disease and the risk of subsequent breast cancer. *NPJ Breast Cancer* **6,** 24 (2020).

48. Silverstein, M. J. The University of Southern California/Van Nuys prognostic index for ductal carcinoma in situ of the breast. *Am. J. Surg.* **186,** 337–343 (2003).

49. Mannu, G. S., Wang, Z., Broggio, J., Charman, J., Cheung, S., Kearins, O., Dodwell, D. & Darby, S. C. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988-2014: population based observational cohort study. *BMJ* **369,** m1570 (2020).

50. Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. & Mulvihill, J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81,** 1879–1886 (1989).

51. Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F. & Antoniou, A. C. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21,** 1708–1718 (2019).

52. Kos, Z., Roblin, E., Kim, R. S., Michiels, S., Gallas, B. D., Chen, W., van de Vijver, K. K., Goel, S., Adams, S., Demaria, S., Viale, G., Nielsen, T. O., Badve, S. S., Symmans, W. F., Sotiriou, C., Rimm, D. L., Hewitt, S., Denkert, C., Loibl, S., Luen, S. J., Bartlett, J. M. S., Savas, P., Pruneri, G., Dillon, D. A., Cheang, M. C. U., Tutt, A., Hall, J. A., Kok, M., Horlings, H. M., Madabhushi, A., van der Laak, J., Ciompi, F., Laenkholm, A.-V., Bellolio, E., Gruosso, T., Fox, S. B., Araya, J. C., Floris, G., Hudeček, J., Voorwerk, L., Beck, A. H., Kerner, J., Larsimont, D., Declercq, S., Van den Eynden, G., Pusztai, L., Ehinger, A., Yang, W., AbdulJabbar, K., Yuan, Y., Singh, R., Hiley, C., Bakir, M. A., Lazar, A. J., Naber, S., Wienert, S., Castillo, M., Curigliano, G., Dieci, M.-V., André, F., Swanton, C., Reis-Filho, J., Sparano, J., Balslev, E., Chen, I.-C., Stovgaard, E. I. S., Pogue-Geile, K., Blenman, K. R. M., Penault-Llorca, F., Schnitt, S., Lakhani, S. R., Vincent-Salomon, A., Rojo, F., Braybrooke, J. P., Hanna, M. G., Soler-Monsó, M. T., Bethmann, D., Castaneda, C. A., Willard-Gallo, K.,

Sharma, A., Lien, H.-C., Fineberg, S., Thagaard, J., Comerma, L., Gonzalez-Ericsson, P., Brogi, E., Loi, S., Saltz, J., Klaushen, F., Cooper, L., Amgad, M., Moore, D. A., Salgado, R. & International Immuno-Oncology Biomarker Working Group. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer* **6,** 17 (2020).

53. Hendry, S., Salgado, R., Gevaert, T., Russell, P. A., John, T., Thapa, B., Christie, M., van de Vijver, K., Estrada, M. V., Gonzalez-Ericsson, P. I., Sanders, M., Solomon, B., Solinas, C., Van den Eynden, G. G. G. M., Allory, Y., Preusser, M., Hainfellner, J., Pruneri, G., Vingiani, A., Demaria, S., Symmans, F., Nuciforo, P., Comerma, L., Thompson, E. A., Lakhani, S., Kim, S.-R., Schnitt, S., Colpaert, C., Sotiriou, C., Scherer, S. J., Ignatiadis, M., Badve, S., Pierce, R. H., Viale, G., Sirtaine, N., Penault-Llorca, F., Sugie, T., Fineberg, S., Paik, S., Srinivasan, A., Richardson, A., Wang, Y., Chmielik, E., Brock, J., Johnson, D. B., Balko, J., Wienert, S., Bossuyt, V., Michiels, S., Ternes, N., Burchardi, N., Luen, S. J., Savas, P., Klauschen, F., Watson, P. H., Nelson, B. H., Criscitiello, C., O'Toole, S., Larsimont, D., de Wind, R., Curigliano, G., André, F., Lacroix-Triki, M., van de Vijver, M., Rojo, F., Floris, G., Bedri, S., Sparano, J., Rimm, D., Nielsen, T., Kos, Z., Hewitt, S., Singh, B., Farshid, G., Loibl, S., Allison, K. H., Tung, N., Adams, S., Willard-Gallo, K., Horlings, H. M., Gandhi, L., Moreira, A., Hirsch, F., Dieci, M. V., Urbanowicz, M., Brcic, I., Korski, K., Gaire, F., Koeppen, H., Lo, A., Giltnane, J., Rebelatto, M. C., Steele, K. E., Zha, J., Emancipator, K., Juco, J. W., Denkert, C., Reis-Filho, J., Loi, S. & Fox, S. B. Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma In Situ, Metastatic Tumor Deposits and Areas for Further Research. *Adv. Anat. Pathol.* **24,** 235–251 (2017).

54. Cornel, A. M., Mimpen, I. L. & Nierkens, S. MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy. *Cancers* **12,** (2020).

55. Garrido, M. A., Rodriguez, T., Zinchenko, S., Maleno, I., Ruiz-Cabello, F., Concha, Á., Olea, N., Garrido, F. & Aptsiauri, N. HLA class I alterations in breast carcinoma are associated with a high frequency of the loss of heterozygosity at chromosomes 6 and 15. *Immunogenetics* **70,** 647–659 (2018).

56. Campbell, M. J., McCune, E., Bolen, J., VandenBerg, S., Chien, J., Wong, J. & Esserman, L. Abstract 961: Intralesional injection of anti-PD-1 (pembrolizumab) results in increased T cell infiltrate in high risk DCIS. *Cancer Res.* **78,** 961–961 (2018).

57. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

58. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

59. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data.

Cell **177,** 1888–1902.e21 (2019).

60. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20,** 296 (2019).

61. Gendoo, D. M. A., Ratanasirigulchai, N., Schröder, M. S., Paré, L., Parker, J. S., Prat, A. & Haibe-Kains, B. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32,** 1097–1099 (2016).

62. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

63. Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T. & Hahn, W. C. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462,** 108–112 (2009).

64. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. & D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46,** D649–D655 (2018).

65. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. & Mesirov, J. P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).

66. Didion, J. P., Martin, M. & Collins, F. S. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5,** e3720 (2017).

67. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013). at <http://arxiv.org/abs/1303.3997>

68. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9,** 13 (2014).

69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

70. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J. C. & Dry, J. R. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44,** e108 (2016).

71. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK,*

and WDL in Terra. ('O'Reilly Media, Inc.', 2020).

72. Guimera, R. V. bcbio-nextgen: Automated, distributed next-gen sequencing pipeline. *EMBnet.journal* **17,** 30 (2011).

73. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6,** 80–92 (2012).

74. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

75. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G. & Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

76. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L. & Maglott, D. R. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46,** D1062–D1067 (2018).

77. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Genome Aggregation Database Consortium, Neale, B. M., Daly, M. J. & MacArthur, D. G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (2020).

78. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12,** e1004873 (2016).

79. Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L. & Kristensen, V. N. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 16910–16915 (2010).

80. Nilsen, G., Liestøl, K., Van Loo, P., Moen Vollan, H. K., Eide, M. B., Rueda, O. M., Chin, S.-F., Russell, R., Baumbusch, L. O., Caldas, C., Børresen-Dale, A.-L. & Lingjaerde, O. C. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13,** 591 (2012).

81. Islam, S. M. A., Ashiqul Islam, S. M., Wu, Y., Díaz-Gay, M., Bergstrom, E. N., He, Y., Barnes, M., Vella, M., Wang, J., Teague, J. W., Clapham, P., Moody, S., Senkin, S., Li, Y. R., Riva, L., Zhang, T., Gruber, A. J., Vangara, R., Steele, C. D., Otlu, B., Khandekar, A., Abbasi, A., Humphreys, L., Syulyukina, N., Brady, S. W., Alexandrov, B. S., Pillay, N., Zhang, J., Adams, D. J., Marticorena, I., Wedge, D. C., Landi, M. T., Brennan, P., Stratton, M. R., Rozen, S. G. & Alexandrov, L. B. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. doi:10.1101/2020.12.13.422570

82. Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, Getz, G., Rozen, S. G., Stratton, M. R. & PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101 (2020).

83. Kim, S., Kim, S., Kim, J., Kim, B., Kim, S. I., Kim, M. A., Kwon, S. & Song, Y. S. Evaluating Tumor Evolution via Genomic Profiling of Individual Tumor Spheroids in a Malignant Ascites from a Patient with Ovarian Cancer Using a Laser-aided Cell Isolation Technique. doi:10.1101/282277

84. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).

85. Reiter, J. G., Makohon-Moore, A. P., Gerold, J. M., Bozic, I., Chatterjee, K., Iacobuzio-Donahue, C. A., Vogelstein, B. & Nowak, M. A. Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8,** 14114 (2017).

86. Mori, H., Bolen, J., Schuetter, L., Massion, P., Hoyt, C. C., VandenBerg, S., Esserman, L., Borowsky, A. D. & Campbell, M. J. Characterizing the Tumor Immune Microenvironment with Tyramide-Based Multiplex Immunofluorescence. *J. Mammary Gland Biol. Neoplasia* **25,** 417–432 (2020).

87. Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J. A., Salto-Tellez, M. & Hamilton, P. W. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7,** 16878 (2017).

CHAPTER 3: Accurate genome-wide germline profiling from decade-old archival tissue DNA reveals the contribution of common variants to precancer disease outcome.

3.1 Abstract

Inherited variants have been shown to contribute to cancer risk, disease progression, and response to treatment. Such studies are, however, arduous to conduct, requiring large sample sizes, cohorts or families, and more importantly, a long follow-up to measure a relevant outcome such as disease onset or progression. Unless collected for a dedicated study, germline DNA from blood or saliva are typically not available retrospectively, in contrast to surgical tissue specimens which are systematically archived. We evaluated the feasibility of using DNA extracted from low amounts of fixed-formalin paraffin-embedded (FFPE) tumor tissue to obtain accurate germline genetic profiles. Using matching blood and archival tissue DNA from 10 individuals, we benchmarked low-coverage whole-genome sequencing (lc-WGS) combined with genotype imputation and measured genome-wide concordance of genotypes, polygenic risk scores (PRS), and HLA haplotypes. Concordance between blood and tissue was high (r^2>0.94) for common genome-wide single nucleotide polymorphisms (SNPs) and across 22 disease-related PRS (mean r=0.93). HLA haplotypes imputed from tissue DNA were 96.7% (Class I genes) and 82.5% (Class II genes) concordant with deep targeted sequencing of HLA from blood DNA. Using the validated methodology, we estimated breast cancer PRS in 36 patients diagnosed with breast ductal carcinoma in situ (11.7 years median follow-up time) including 22 who were diagnosed with breast cancer subsequent event (BSCE). PRS was significantly associated with BCSE (HR=2.5, 95%CI: 1.4–4.5) and the top decile patients were modeled to have a 24% chance of BCSE at 10 years,

hence suggesting the addition of PRS could improve prognostic models which are currently inadequate. The abundance and broad availability of archival tissue specimens in oncology clinics, paired with the effectiveness of germline profiling using lc-WGS and imputation, represents an alternative cost and resource-effective alternative in the design of long-term disease progression studies.

3.2 Introduction

The study of the contribution of germline genetic variation to disease risk or treatment outcome typically requires blood or saliva samples as a source of constitutive DNA. Depending on the phenotype studied, such samples may not be banked and readily available. Samples may have to be prospectively collected, which hinders studies requiring long-term follow-up or obtained after contacting potential subjects of interest, which can be logistically and ethically challenging or impossible if a patient has relocated or died. In cancer research, there is a growing interest in directly profiling tumor tissue to obtain germline measures such as ancestry, polygenic risk, and HLA-typing [1]. Array-based genotyping followed by imputation from a reference population has been a standard method to genotype genome-wide SNPs in the human genome, but its compatibility with DNA obtained from archival tissue specimens remains to be established [2,3]. The approach can be challenging when the amount of tissue available for research is limited, which is often the case with surgical excisions of premalignant lesions or with most needle biopsies.

Recently low-coverage whole-genome sequencing (lc-WGS) has emerged as an attractive alternative to single nucleotide polymorphism (SNP) array by offering higher throughput at a reduced cost, reduced DNA input, and improved genotyping accuracy [4–6]. In fact, recent studies have shown the feasibility of using frozen tissue for germline profiling by imputing genotypes from off-target reads repurposed from tumor-targeted panel sequencing data, effectively equivalent to ultra-low coverage (less than 0.1x) whole-genome sequencing [1]. It is therefore likely that, in the absence of available targeted sequencing data, lc-WGS can be performed with DNA of lower quality and quantity to enable the imputation of germline variants from archival tissue specimens.

If accurate, such an approach could have important implications for the study of the contribution of inherited risk factors to the progression of pre-malignant disease. For many cancer types, the widespread adoption of cancer screening has led to an increase in the detection of pre-malignant lesions. Despite such efforts, screening has had limited impact on overall survival [7]. Clinical guidelines vary widely from watchful waiting or biopsy as for prostatic intraepithelial neoplasia to surgery and adjuvant treatment as for ductal carcinoma in situ (DCIS) of the breast [8,9]. In absence of reliable progression risk biomarkers and models, these interventions may have deleterious consequences at the two clinical extremes: delay in life-saving treatment or complications from overtreatment. DCIS is the most common breast cancer-related diagnosis, comprising ~20% of annual cases in the U.S. [10]. In breast disease, factors that impact the risk of breast cancer subsequent event (BCSE), defined as an in situ or invasive breast cancer neoplasm developed at least 6 months after treatment of a DCIS diagnosis, include age, size, grade of the lesion, hormone receptor status, and molecular profile. Their combined effect in risk models such as the University of Southern California / Van Nuys Prognostic Index has not resulted in any reliable BCSE risk prediction model and additional, more in-depth molecular and histological characterization is needed [11–16].

Given the independence between DCIS and associated BCSE in upwards of 20% of cases as evidenced by molecular studies comparing genomic profiles of initial DCIS and subsequent ipsilateral BCSE, systemic risk factors need to be considered in addition to those related to the index lesion [17]. While penetrant germline pathogenic variants exist and represent strong risk factors in breast cancer susceptibility genes such as *BRCA1, BRCA2, CHEK2, PALB2*, and *PMS2*, they are only present in 1.5% of all women [18]. Meanwhile, population-based genome-wide association studies (GWAS), have identified multiple common variants associated with lifetime risk of

invasive breast cancer (IBC) [2,19]. The same SNPs have also been associated with risk of DCIS demonstrating the shared genetic susceptibility for IBC and DCIS [20]. It is however unclear if these SNPs are also associated with DCIS progression. Polygenic risk scores (PRS) derived from the allelic burden of risk-associated SNPs are now being added to common breast cancer risk models, significantly improving their performance, with individuals in the top percentile having a 3-5 fold increase in lifetime risk relative to women with risk scores in the middle quintile of those studied [21,22]. It is thus possible that DCIS patients with elevated breast cancer PRS are also at higher risk of BCSE and the addition of PRS could improve DCIS prognostic models akin to lifetime breast cancer risk models. Since BCSE can occur years after the initial DCIS diagnosis and is uncommon - observed in 10 to 25% of patients after 10 years, depending on treatment and known risk factors - a retrospective study is much more feasible for the purposes of validation [23,24]. Formalin-fixed paraffin-embedded (FFPE) tissue (referred to as archival tissue) from the DCIS biopsy or resection are therefore the only source of genetic material available and their validity for genome-wide genotyping of germline variants would be critical to the feasibility of such study.

Here we evaluate the validity of repurposing archival tissue specimens for germline genetic studies. We performed lc-WGS and imputed genotypes for 10 pairs of matching blood and tumor tissue samples to benchmark the accuracy for calling genome-wide genotypes, HLA haplotypes, and for implementing PRS. The reported results indicate the high accuracy of germline genotypes and haplotypes obtained from archival tissue DNA. Using this methodology we estimate breast cancer PRS in 36 DCIS patients and demonstrate its association with BCSE.

3.3 Results

Concordance of lc-WGS imputed genotypes between blood DNA and FFPE tissue DNA.

In order to establish the analytical validity of FFPE tissue DNA for germline genotyping and genotype imputation from lc-WGS, we selected 10 subjects including from European, African, and Asian ancestries with matching FFPE tissue and whole blood. The archival tissue blocks were between 3 and 9 years old and yielded between 5 and 176 ng of DNA, which was then prepared for sequencing with a low-input protocol (see Methods). Mean coverage depth was 0.92x (range 0.68-1.41x) and 0.7x (range 0.44-0.97) for blood and tissue, respectively. Genotypes were imputed using a Gibbs sampling method specifically designed for lc-WGS, which leverages haplotype reference panel information (1000G 30x NYGC reference panel - N=3,202 individuals; see Methods) [5,25]. Overall genotypes were imputed for 61,715,567 SNPs in each of the 20 samples, of which 43,274,690 (70.1%) were considered high quality (Impute INFO score >0.80) [26]. Genotype concordance between blood and tissue increased with the minor allele frequency (MAF) of the variant in the global population. For SNPs with MAF of 0.1 or more, the aggregate r^2 was greater than 91% for all SNPs, and greater than 94% for high-quality SNPs (Figure 3.1a). The concordance between blood and tissue was not lower for the two individuals who were non-white (Figure S3.1). In contrast, the concordance was lower (87% at SNPs with MAF greater or equal to 0.1) when the sequencing coverage depth of the tissue DNA was lower (Figure S3.2). Overall, the strongest discordance between blood and tissue was observed for SNPs at MAF lower than 0.01 which are typically imputed with decreased accuracy irrespective of the sample type [27].

The presence of somatic mutations and copy number alterations (CNA) in DNA from malignant cells has the potential to decrease local imputation accuracy. In particular, CNA may play a larger role than somatic mutations, as recently reported [1]. We estimated the effect of CNA

status on the genotype concordance between blood and tissue across SNPs located in DNA regions that are copy neutral, in a copy gain, or in a copy loss. The studied DNA samples had, on average, 15% of the genome (range 0 to 65%) involved in CNA while no CNA was detected in the blood (see Methods, Table S3.1). Common SNPs (MAF≥0.1) located in copy neutral or copy gain regions had a remarkable blood-tissue genotype concordance r^2 higher than 95%, while those in regions of copy number loss showed lower concordance r^2 of 83% (Figure 3.1b). The decreased imputation accuracy in areas of copy number loss can likely be explained by the decrease in allele-specific coverage depth, resulting in missed heterozygotes or a sparser scaffold for imputation.

We conclude that tissue-derived genome-wide genotypes faithfully represent germline profiles obtained from blood, especially at SNPs frequent in the population (MAF≥0.1). Discrepancies between tissue and blood can be explained by decreased coverage depth caused by technical (insufficient sequencing) or genetic (copy number loss) limitations and mainly affecting rare SNPs (MAF<0.01). These lower frequency SNPs are less likely to reach statistical significance in GWAS studies unless they have extreme effect size and therefore are rarely incorporated into PRS models. Taken together, the results suggest the feasibility of using archival tissues as a source of constitutive DNA in genetic studies relying on common SNPs.

Concordance of tissue-derived PRS.

We next sought to further validate the performance of tissue-based genome-wide genotyping to accurately estimate PRS in individuals. Germline variants can be used to estimate disease risk in individuals by summing the effects of previously identified risk alleles carried by an individual into a personalized PRS. The clinical utility of PRS is currently being evaluated in multiple settings, including breast cancer screening and surveillance, where elevated PRS can be included in lifetime risk models [28]. The ability to accurately estimate PRS retrospectively, using

archival tissue DNA, would greatly improve the ability to conduct large retrospective studies with long-term outcomes. We investigated multiple PRS derived from GWAS of susceptibility to 16 cancer types, and 6 non-cancer phenotypes [29–31]. We computed a tissue and blood-derived PRS for 10 individuals (see Methods) using the imputed genotypes from lc-WGS sequencing data described above. Overall 93% (2,744 of 2,962) of PRS single-nucleotide variant sites were successfully imputed, 84% of which were high quality (Table S3.2). In each of the 16 cancer types, the tissue-derived PRS closely matched the blood-derived PRS, evidenced by high correlation coefficients ($r \geq 0.9$) in 12/16 of the PRS (Figure 3.2a). We saw similar results when evaluating PRS for non-cancer phenotypes, with 4/6 being highly correlated ($r \geq 0.9$) (Figure 3.2b). Differences between PRS in blood and tissue were associated with decreased tumor genome coverage ($r=-0.26$, $p=0.02$), but not with copy number loss (Figure S3.3). Overall we report that archival tissue DNA profiled with lc-WGS resulted in a reliable PRS estimate in an individual and preserved relative ranks in a cohort enabling studies such as the use case presented below.

Contribution of breast polygenic risk scores to DCIS prognosis.

We next demonstrated the utility of lc-WGS to investigate the contribution of breast cancer PRS to predict breast cancer subsequent events (BCSE - in situ or invasive, irrespective of laterality) after a DCIS diagnosis using a retrospective study design (Figure 3.3a). We assembled a cohort of patients diagnosed with pure DCIS (N=25 cases) who were then diagnosed with a BCSE at least 6 months after the DCIS diagnosis. We then complemented this cohort with a set of patients (N=25 controls) diagnosed with pure DCIS who did not develop a BCSE for at least 5 years. A median of 51.2 ng (range 6.6-300) of DNA was extracted from the primary DCIS FFPE specimen archived between 6 and 25 years (Table S3.3). The extracted DNA was sequenced to an average coverage depth of 0.89x (range 0.2-1.8x) (see Methods). Fourteen out of 50 (28%) samples

yielded insufficient coverage (N=5) or had evidence of contamination with another patient (N=9) and were excluded, leaving 22 cases and 14 controls for analysis (Table 3.1). The median time to BCSE was 6.2 years (min: 1.4, max: 10.9), and patients without BCSE had a median time to follow-up of 11.7 years (min: 6.7, max: 19.6). Cases and controls were approximately matched for age, ancestry, DCIS size, grade, and ER status (Table S3.4). We then performed imputation as described earlier which resulted in high-quality genotypes at a total of 27,605,021 SNP loci.

In order to evaluate the relationship between breast cancer PRS and DCIS prognosis, we curated 6 previously established breast cancer PRS, measuring risk for both overall and ER+ breast cancer, consisting of 859 total and 674 unique sites (see Methods, Table S3.2) [22,29,32–34]. We computed PRS for each patient, and compared groups with and without BCSE (Figure 3.3b) (see Methods). Patients with BCSE showed near significant elevated values across all 6 PRS (mean 1.4x fold increase, minimum p=0.06). We next measured the prognostic value of PRS in a multivariate Cox proportional hazard model to account for other risk factors previously associated with DCIS progression such as age, DCIS size, histological grade, and ancestry. We found that three of the breast cancer PRS had a significant ($q<0.01$) impact on BCSE risk, with the most impactful overall and ER+ breast cancer PRS hazard ratios of 2.5 (95%CI: 1.4–4.5, q=0.008) and 2.01 (1.3–3.1, q=0.007) respectively (Figure 3.3c, Figure S3.4). Adding PRS to the model improved the discrimination between patients with and without BCSE raising the mean C-index from 0.66 to 0.71 (Figure 3.3d). In contrast, none of the six non-cancer PRS contributed significantly to the BCSE prognosis, indicating that the effects observed are likely specific to the underlying genetic risk specific to breast cancer (Figure 3.3e). We estimate that 10 years post-DCIS diagnosis, approximately 24% of patients with the highest decile of breast PRS will have a BCSE, as opposed to approximately 3% of patients with PRS in the lowest decile (Figure 3.3f).

Even with this limited dataset, there is a suggestive contribution of pre-established breast cancer PRS in DCIS prognosis, though this will require validation in a larger independent cohort. Independent of its possible clinical significance, and acknowledging the need for additional validation of the results, the presented use case demonstrates the feasibility of using DNA from tissues archived for decades to associate germline genetic factors with long-term patient outcomes and gain new insight into disease etiology and progression.

Imputation of HLA-gene alleles from lc-WGS.

In addition to SNP genotyping, we next investigated whether lc-WGS of archival tissue could be used to determine the haplotypes of the various HLA genes. HLA genes are some of the most polymorphic genes in the human genome and the major histocompatibility complex plays a critical role in antigen presentation to the immune system, particularly in tumorigenesis [35–37]. Using samples collected from 14 patients, including 10 patients with both blood and tissue DNA available, we compared HLA-alleles imputed from genome-wide genotypes obtained from lc-WGS against the results of clinical-grade deep targeted sequencing of the HLA locus from matching blood DNA samples (referred to as gold standard - see Methods). Alleles for Class I (HLA-A, B, C genes) and Class II (DRB1, DQB1 genes) were imputed using QUILT-HLA against the 1000G reference panel [6]. Overall 4 field allele calls from blood DNA were 92.8% (78/84) and 80.4% (45/56) concordant with the gold standard for Class I and Class II genes respectively (Figure 3.4a). At a lower 2 field resolution, the concordance was 97% for Class I and 91% for Class II (Figure S3.5). The decreased accuracy for HLA Class II, particularly for DRB1 likely reflects the increased diversity of these loci in comparison to Class I as well as the presence of pseudo-genes which may introduce ambiguity in the alignment of short sequence reads [38].

In order to evaluate the effect of DNA source on HLA-typing accuracy, we compared tissue-derived HLA types to the gold standard. We found 4 field allele calls from tissue were 96.7% (58/60) and 82.5% (33/40) concordant with the gold standard blood HLA-typing, for Class I and Class II respectively (Figure 3.4b). In 49/50 comparisons between blood and tissue, tissue-derived samples provided as accurate calls, suggesting that the DNA source did not have an impact on imputation quality (Figure 3.4c). Overall, HLA-types that did not match the gold standard had worse imputation quality as reflected by their lower posterior probabilities (Figure 3.4d). The high accuracy of HLA-typing from lc-WGS as well as the consistent results between blood and tissue-based DNA demonstrates that remarkably, imputed HLA-types from lc-WGS on archival tissue are comparable against deep targeted HLA sequencing on blood, with a fraction of the required DNA input and a streamlined protocol.

3.4 Discussion

Here we rethink the traditional design of germline genetic studies by answering the question, when typical DNA sources such as blood, saliva or urine are unavailable, can we extract the same information from archival tissue specimens? Often collected for histological examination and diagnosis and then stored indefinitely, these samples offer an abundant source of genetic material from patients with potentially long clinical follow-up. By using lc-WGS and recent advances in genotype imputation, we compared the concordance of germline genotypes obtained from blood DNA and archival tissue DNA in 10 different individuals. Archival tissue faithfully represented the germline profile of common SNPs obtained from blood both at the genome-wide level and across well-established PRS. Beyond concordance at the SNP-level, we also demonstrated accurate genotyping at highly polymorphic HLA alleles. To our knowledge, we present the first evidence that HLA-typing using lc-WGS from archival tissue is as accurate as true clinical-grade HLA-typing. Our results support the future utilization of archival tissue to construct large retrospective studies to characterize the role of germline variants in disease etiology, progression, and treatment.

The use of archival tissue as a source of constitutive DNA will enable a wealth of retrospective studies by repurposing specimens archived by most clinical sites to help address the genetic underpinnings of disease with long-outcome, such as the progression of pre-malignant lesions as presented here. Such studies would either require long follow up after the initial sample collection, or a massive and costly effort to retrospectively collect blood or saliva samples. In contrast, provided the subjects have been offered diagnostic biopsies, or surgical treatment, the course of their clinical care or study participation, their left-over specimen can be used to enable post-hoc genetic analysis. Of course, such studies would require approval of the Institutional

Review Boards (IRB) and, since 2015, informed consent needs to be explicit about the use of specimens and data for genetic research and the risk for privacy it entails [39]. Commonly, IRBs waive the requirement for consent from patients deceased or lost to follow-up, however, such data needs to be distributed with caution and typically protected by a Data Access Policies the researcher has to comply with. As such, while our approach can enable large retrospective genetic studies where informed consent may be waived, the eligibility of each patient, and the overall data sharing policy need to be carefully considered.

Our report includes the application of the approach to interrogate the contribution of genetic factors to breast DCIS progression. The relatively good outcome of the disease poorly justified a thorough collection of risk variables, especially those related to inherited risk. However, overtreatment of DCIS, and its harms, is being increasingly acknowledged and systematic reviews of clinicopathological factors have not resulted in reliable models of progression [11,12,40]. Most epidemiological studies need to be large due to the slow progression and rarity of poor outcomes and rely exclusively on medical chart review [24,41,42]. As such, additional factors that are hard or impossible to collect from the charts such as mammography or magnetic resonance imaging, digital pathology, or germline inherited factors have not been as thoroughly and systematically investigated. We made the narrow hypothesis that lifetime breast cancer susceptibility - which can be seen as progression from normal to malignant epithelium - and progression of DCIS share the same genetic risk factors. We tested this hypothesis by measuring breast PRS in a small cohort of carefully selected DCIS subjects using our approach. Given the effect size of PRS contribution to breast cancer (HR=1.61), we anticipated that a balanced cohort of 36 patients would be sufficient to measure an effect size of HR=1.6 or greater representative of the contribution of other risk factors to DCIS progression [22,43,44]. Thanks to the accurate PRS estimate obtained from left-over

surgical specimens, we were able to see that germline variation likely contributed significantly to the DCIS progression to an extent similar or greater to previously investigated risk factors such as grade, age, and Her2 overexpression [40]. Such findings would clearly need to be validated in a larger cohort, where a more comprehensive set of covariates would be accounted for, including treatment. Subsequent larger studies would also be important to evaluate competing risk models for subsequent in situ versus invasive disease, or laterality of the event, where PRS may contribute more in particular contexts. The modest cost and relative experimental simplicity of our approach, accompanied by a state-of-the-art imputation strategy can likely be scaled up provided diagnostic sections or left-over specimens can be found. Several large DCIS cohorts are generating mutational profiles, including some with lc-WGS and associated with clinical outcomes, which would be particularly suitable for validation in the future [17,45].

In the study of malignant progression as well as the onset and progression of multiple other diseases, the overactivity or inactivity of the immune system represents a key factor. A large contribution of variation in immune traits is inherited and yet the role of this contribution in disease progression is poorly understood [46,47]. In particular, the genetic diversity of the MHC, one of the most polymorphic regions of the genome, is a real challenge to study the role of the adaptive immune system. In the context of tumorigenesis, the failure of the major histocompatibility complex (MHC) to present antigens to the immune system is being increasingly recognized as contributing to cancer immune evasion and failure to respond to immune checkpoint inhibitors [48–50]. The determination of the HLA haplotypes, encoding the MHC is typically limited to the setting of organ or bone marrow transplants and not typically performed in other epidemiological studies. Recent reports however show the importance of the HLA-type in understanding the exposed mutanome, and its consideration can have important predictive value in the context of

immunotherapies [35,36,51]. But with a lack of systemic HLA-typing or absence of genetic material to do so, such studies are hard to replicate or scale-up. To address this, we demonstrated that we can assign 4 field alleles to HLA-A, B, C, and DRB1, DQB1 genes by reference informed imputation of lc-WGS data [6]. These imputed HLA-types had comparable accuracy to deep targeted sequencing of the HLA locus with a fraction of the required DNA input (5 vs 40,000 ng) and with a simplified protocol (no need for targeted capture). The improvement in both sample requirement and throughput to HLA-typing supports the evaluation in lc-WGS with imputation in replacing current clinical standard tests.

While offering many benefits, there are still some limitations to lc-WGS paired with imputation for germline profiling of archival tissue. Similar to previous reports benchmarking lc-WGS imputation, error increases with decreasing minor allele frequency [5,6]. This would preclude the use of this strategy for the identification of rare variants of high penetrance associated with familiar risk (*BRCA*, Lynch, or Li-Fraumeni syndromes). Similarly, genotypes from rare risk-associated SNPs or HLA-types only found in small populations would be more likely missed by this approach. In the future, the availability of even larger and more diverse reference populations may help mitigate this effect. For the purposes of this study we utilized the unrestricted 1000G reference panel (N=3,202 haplotypes), however larger extensive, though restricted, panels such as Haplotype Reference Consortium (HRC) (N=64,976) or TopMed (N=53,831) exist [25,52,53]. Low coverage depth represents an additional limitation of our approach. While a restricted number of reads sequenced from a WGS library can result in decreased imputation accuracy, another source of tumor-specific decreased coverage is somatic copy number alterations (CNA). We observed that regions in a copy number loss resulted in decreased imputation accuracy. Similar observations were recently reported in a study performing germline imputation from discarded reads from

targeted-sequencing tumor-derived tissue [1]. Here the choice of the tissue source, or the possibility to dissect normal histological regions, can help mitigate these effects. Indeed the use of adjacent normal tissue, pre-malignant or low-grade lesions or even lymphocytic aggregates, or lymph node specimens would enrich for diploid cells resulting in fewer inaccurate genomic regions. In contrast, imputation in high-grade lesions or invasive tumors with prominent aneuploidy needs to be carefully considered and may be mitigated in the largest dataset where available CNA profiles could be used as prior information in the imputation strategy.

In conclusion, our study demonstrates that archival tumor tissue is an appropriate DNA source to measure germline genetic variation in lieu of normal tissue or blood. By shallow sequencing of the genome, and imputing missing sequences using haplotypes from thousands of individuals, the resulting genotypes, particularly for common SNPs and HLA alleles between blood and archival tissue were quite comparable. Especially in the study of slow progressing or rare diseases which may have been logistically unrealistic due to a long time to events and large sample numbers required, this framework has the potential to enable very large retrospective genetic studies, driving both basic research and translational discoveries.

3.5 Materials and Methods

Patient selection.

For the tissue-blood benchmarking study, a total of N=14 Lung adenocarcinoma cancer patients with available tumor tissue and matching buffy coat in N=10 were selected from the Moores Cancer Center Tissue and Technology Shared Resource (BTTSR).

For the DCIS PRS study, a total of 50 patients were originally selected from the UC San Diego ATHENA DCIS registry - a retrospective registry approved by the UCSD and UCSF IRB. Case patients with BCSE were first selected on the basis of time to BCSE, surgery type, care location, and availability of archival tissue blocks. Control patients were then selected from patients without BCSE, with long follow-up time and matching cases for risk factors including age at DCIS, ancestry, DCIS grade, DCIS size, treatment type, ER, and Her2 status when available (Table 3.1).

Sample Preparation.

Blood DNA was extracted from 50 µL of buffy-coat using DNAeasy blood and tissue kit (Qiagen). Tissue blocks were sectioned in 5 µm scrolls and 3 to 5 scrolls were used to extract DNA with Covaris FFPE truXTRAC FFPE tNA kit using M220 Covaris Focused UltraSonicator (Covaris). DNA was quantified using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific).

Low-coverage whole genome sequencing (lc-WGS).

Between 5-300ng of DNA was used as input for the library preparation using NEB Ultra II FS library preparation kit (New England Biolabs), which combines enzymatic fragmentation with end-repair and A-tailing in the same tube. Ligated and purified libraries were amplified using KAPA HiFi HotStart Real-time PCR 2X Master Mix (KAPA Biosystems). Samples were amplified with 5 µL of KAPA P5 and KAPA P7 primers. The reactions were denatured for

45 seconds (sec) at 98 °C and amplified 13–15 cycles for 15 sec at 98 °C, for 30 sec at 65 °C, and for 30 sec at 72 °C, followed by final extension for 1 min at 72 °C. Samples were amplified until they reached Fluorescent Standard 3, cycles being dependent on input DNA quantity and quality. PCR reactions were then purified using 1x AMPure XP bead clean-up and eluted into 20 µL of nuclease-free water. The amplified and purified libraries were analyzed using the Agilent 4200 Tapestation (D1000 ScreenTape) and quantified by fluorescence (Qubit dsDNA HS assay). Sample libraries with distinct indices were pooled in equimolar amounts, then sequenced to a target coverage of 0.5x, using paired-end 2x100bp reads on a NovaSeq 6000 (Illumina).

Sequencing read processing and sample quality control.

Sequencing libraries were deconvoluted using bcl2fastq [54]. Adapter sequences were trimmed from the raw fastq files using atropos (v1.1.31) [55]. The trimmed reads were then aligned to GRCh38 using bwa-mem (v0.7.17) [56]. Duplicate reads were then marked using biobambam (v2.0.87)[57]. Overall genome-wide coverage was measured using mosdepth (v0.2.6), and contamination was measured using verifyBamID2 (v1.0.6) [58,59]. For the DCIS, samples with less than 0.45x coverage or were estimated to be >5% contaminated were removed from downstream analyses.

Imputation of genotypes from lc-WGS.

Genome-wide genotypes were imputed using lc-WGS specific method GLIMPSE (v1.1.1) with the hg38 version of 1000G 30x NYGC reference panel (N=3,202 individuals) [5,25] Phasing and imputation were performed directly on BAM files in individual chunks of each chromosome using “GLIMPSE_phase”, and then the imputed variants were subsequently ligated together for each chromosome using “GLIMPSE_ligate”. We note that short insertions and deletions were

excluded from any analysis as these are currently unreliable from lc-WGS and not currently imputed by the strategy implemented [1].

Measuring imputation concordance.

Imputation concordance between samples was summarized using squared Pearson correlation values obtained from the bcftools "stats" function (v1.9), which captures the correlation between allele dosages of variants in each minor allele frequency (MAF) bin [60]. Variants across all the autosomes were used in genome-wide benchmarking performance and all chromosomes for PRS evaluations.

Copy number analysis.

Copy number alterations (CNAs) were called using CNVkit (v0.9.9) in "wgs" mode, average bin size was set at 100,000 bp [61]. A set of unrelated normal tissues sequenced with the same protocol were used to generate a panel of normals used during CNA calling. Any bins with a $\log_2$ copy ratio lower than -15, were considered artifacts and removed. Breakpoints between copy number segments were determined using the circular binary segmentation algorithm ($p < 10{-}4$). Copy number genomic burden was computed as the sum of sizes of segments in a gain ($\log_2(\text{ratio}) > 0.3$) or loss ($\log_2(\text{ratio}) < -0.3$) over the sum of the sizes of all segments.

Clinical standard HLA genotyping.

Reference HLA genotyping was performed on approximately 40 μg genomic DNA extracted from buffy-coat aliquots. Samples were prepared using targeted hybrid-capture with AlloSeq Tx17 reagents (CareDx). Samples were pooled and sequenced in 2x150 bp read-length on iSeq 100 instruments (Illumina). Sequence data was analyzed using Assign (v1.0.2) software (CareDx) and IMGT-HLA reference database (v3.43.0.1) [62].

Measuring PRS.

Polygenic risk scores (PRS) were computed using the following equation:

$$\text{PRS} = \sum_{i=1}^{n} \beta_i x_i \quad \text{Equation 1.}$$

Equation 1. PRS is computed as a function of β_i which is the per-allele log odds ratio, or beta coefficient for the risk SNP allele i, and x_i is the dosage of the risk allele i {0,1,2}, and n is the total number of SNPs composing the PRS. PRS scores were then scaled using z-score transformation. PRS sites and effect weights were all obtained from the Polygenic Score (PGS) Catalog [63]. The catalog numbers and descriptions of each PRS are listed in Table S3.2.

Cox proportional hazard model construction for breast PRS.

Cox proportional hazard models were constructed non-parametrically, using Breslow's method with robust estimates in lifelines survival analysis package in Python [64]. A separate model was constructed for each of the six evaluated breast PRS, in order to measure the effect on risk of BCSE, by PRS, DCIS nuclear grade, age of the patient at diagnosis, the size of the lesion, and whether the ancestry of the individual was European or not. Each covariate was tested for violation of the proportional hazards assumption. The 5 samples missing lesion size, were excluded from the model. In the 6 DCIS samples missing grade, grade was assigned on the basis of the tertiles of copy number burden distribution observed in the cohort since grade and copy number burden are highly correlated [14].

Multiple hypothesis correction for non-independent PRS.

In order to perform multiple hypothesis correction on multiple non-independent PRS, such as the breast PRS, we implemented the Li and Ji method in R package *meff* to estimate for the effective number of tests performed [65,66]. The effective number of tests was then used to generate Bonferroni corrected p-values, labeled as q-values.

3.6 Figures

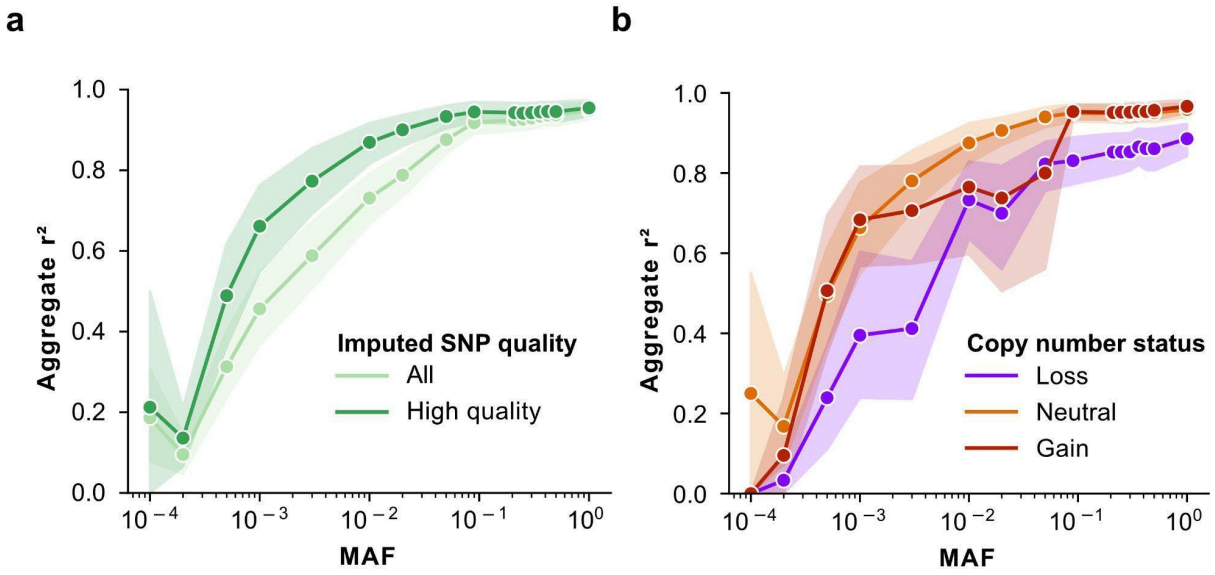


Figure 3.1. Assessment of genome-wide concordance of lc-WGS imputed genotypes in tissue versus blood of N=10 patients.
(a,b) Genome-wide concordance (Pearson correlation coefficient squared - y-axis) of allele dosages across all genotyped SNPs between blood and tissue as a function of their minor allele frequency (MAF, x-axis). Concordance was calculated for each individual and each filtering category including genotype imputation quality (a) with all genotypes shown in light green and high-quality genotypes (INFO>80) in dark green, and copy number status of high-quality genotypes in tissue (b), from SNPs located in a region that was copy neutral (orange), gain (red) or loss (blue). For any given bin corresponding to a patient, MAF and filtering category had to have a minimum of 1,000 SNPs to be included. Error estimates from 95% confidence intervals computed from 1,000 bootstrapping iterations are indicated as shaded areas.

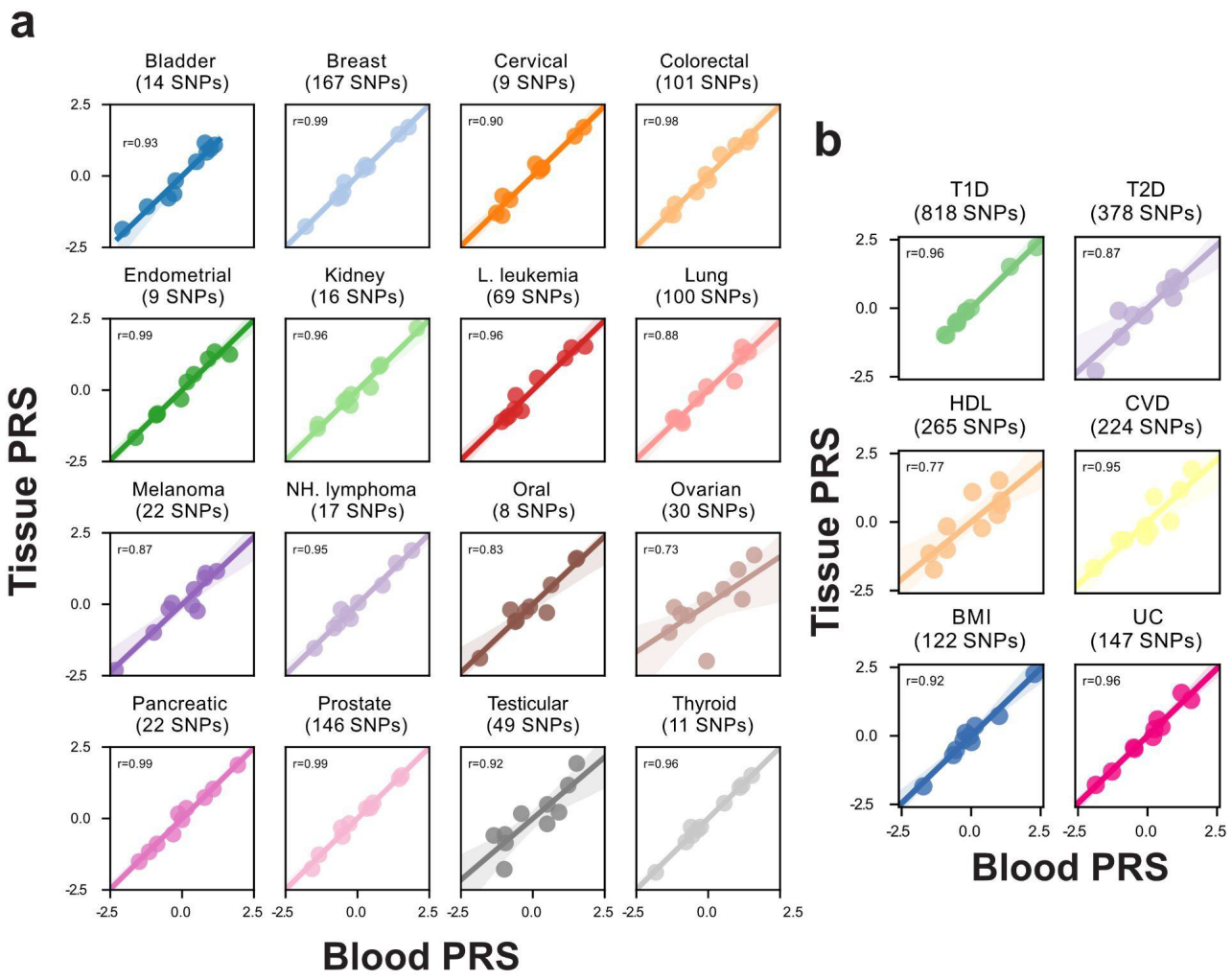
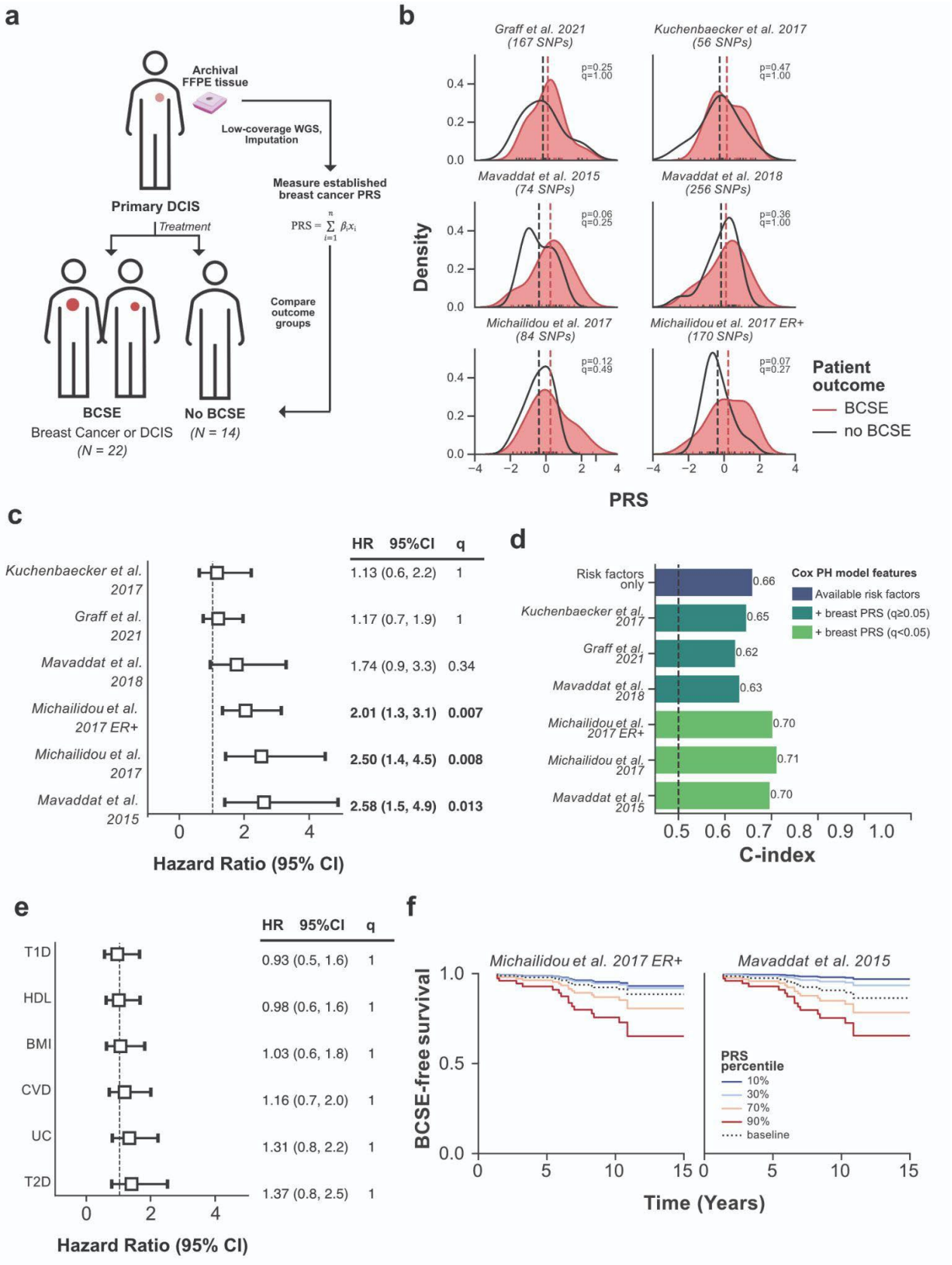


Figure 3.2. Blood versus tissue-derived PRS.
(a) Cancer and **(b)** non-cancer PRS computed from imputed genotypes from lc-WGS of blood (x-axis) and tissue (y-axis) of the same patient. Spearman correlation coefficient, r, was measured between blood and tissue PRS values across N=10 patients, for each normalized PRS. T1D: Type 1 diabetes, T2D: Type 2 diabetes, HDL: High-density lipoprotein, CVD: Cardiovascular disease, BMI: Body mass index, UC: Ulcerative colitis.

Figure 3.3. Breast cancer polygenic risk score in DCIS patients with and without a breast cancer subsequent event.
(a) Schematic overview of the study design. Treatment consisted of surgery and adjuvant radiation or endocrine therapy. **(b)** Comparison of breast cancer PRS score distribution between patients with (red) or without (black) a breast cancer subsequent event (BCSE). Dashed vertical lines represent mean normalized PRS values for each respective group. Groups were compared with two-sided Mann-Whitney U test, and FDR corrected q-values were computed using Bonferroni corrected p-values for the effective number of tests. Distributions were generated using kernel density estimates of histograms. **(c)** Forest plot representation of hazard ratios (square) and 95% confidence intervals (error-bars), for each tested breast cancer PRS, obtained from a Cox Proportional-Hazard model accounting for DCIS size, grade, and age, the ancestry of the patient (Figure S4). The dotted line represents a log hazard ratio of 1, or having no effect on the outcome. The q-values represent Bonferroni corrected p-values for the effective number of tests. Significant hazard ratios ($q<0.05$) are indicated in bold text. **(d)** Evaluation of discrimination of Cox proportional hazard model for BCSE vs no BCSE outcome using Harrel's C-index (y-axis) for models only using available risk factors versus available risk factors and breast cancer PRS, colored by the significance of hazard ratios for breast PRS ($q<0.05$, light green). **(e)** Same as (c) but for non-cancer PRS. **(f)** Cox proportional hazard estimate of breast cancer subsequent event (BCSE) - free survival for two overall and ER+ breast cancer PRS over time in years. Curves are obtained by varying PRS (solid colored lines from blue as lowest and red as highest PRS percentile), as compared to each model baseline (dashed line) while keeping all other covariates the same. Each case and control was weighted by the epidemiological incidence of BCSE treated with surgery and endocrine therapy (15% at 10 years) [24].


a
Archival FFPE tissue
Low-coverage WGS, Imputation
Measure established breast cancer PRS
$PRS = \sum_{i=1}^{n} \beta_i x_i$
Primary DCIS
Treatment
Compare outcome groups
BCSE
Breast Cancer or DCIS
(N = 22)
No BCSE
(N = 14)
b
Graff et al. 2021 (167 SNPs)
p=0.25 q=1.00
Kuchenbaecker et al. 2017 (56 SNPs)
p=0.47 q=1.00
Mavaddat et al. 2015 (74 SNPs)
p=0.06 q=0.25
Mavaddat et al. 2018 (256 SNPs)
p=0.36 q=1.00
Michailidou et al. 2017 (84 SNPs)
p=0.12 q=0.49
Michailidou et al. 2017 ER+ (170 SNPs)
p=0.07 q=0.27
Density
PRS
Patient outcome
BCSE
no BCSE
c
HR 95%CI q
Kuchenbaecker et al. 2017
1.13 (0.6, 2.2) 1
Graff et al. 2021
1.17 (0.7, 1.9) 1
Mavaddat et al. 2018
1.74 (0.9, 3.3) 0.34
Michailidou et al. 2017 ER+
2.01 (1.3, 3.1) 0.007
Michailidou et al. 2017
2.50 (1.4, 4.5) 0.008
Mavaddat et al. 2015
2.58 (1.5, 4.9) 0.013
Hazard Ratio (95% CI)
d
Risk factors only
Kuchenbaecker et al. 2017
Graff et al. 2021
Mavaddat et al. 2018
Michailidou et al. 2017 ER+
Michailidou et al. 2017
Mavaddat et al. 2015
0.66
0.65
0.62
0.63
0.70
0.71
0.70
Cox PH model features
Available risk factors
+ breast PRS (q≥0.05)
+ breast PRS (q<0.05)
C-index
e
HR 95%CI q
T1D
0.93 (0.5, 1.6) 1
HDL
0.98 (0.6, 1.6) 1
BMI
1.03 (0.6, 1.8) 1
CVD
1.16 (0.7, 2.0) 1
UC
1.31 (0.8, 2.2) 1
T2D
1.37 (0.8, 2.5) 1
Hazard Ratio (95% CI)
f
Michailidou et al. 2017 ER+
Mavaddat et al. 2015
BCSE-free survival
PRS percentile
10%
30%
70%
90%
baseline
Time (Years)

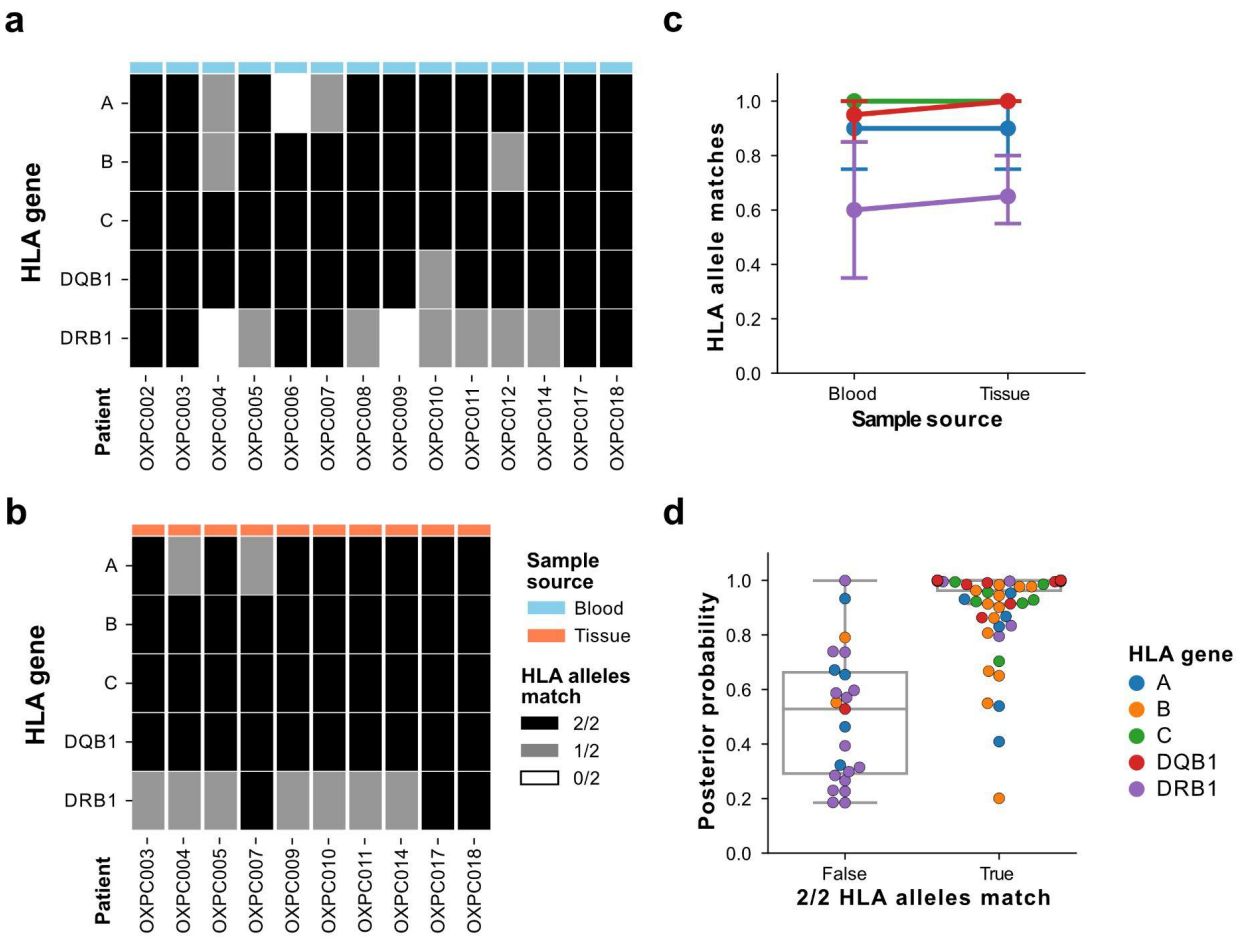


Figure 3.4. Assessment of 4 field HLA-typing accuracy from lc-WGS.
(a,b) Number of concordant HLA alleles (0: white, 1: grey:, 2: black) between haplotypes from the clinical gold standard and those imputed using QUILT-HLA for class I (A, B, C) and class II (DQB1 and DRB1) HLA genes (rows) using (a) blood DNA of 14 patients or (b) tissue DNA of 10 patients (columns). (c) Fraction of HLA alleles correctly imputed (y-axis), versus the sample source of the DNA (x-axis), colored by the HLA gene. (d) Imputation posterior probability from QUILT-HLA for each HLA gene (color) and sample (dot), compared between samples with perfect HLA-gene concordance (both alleles match) versus those with errors.

3.7 Tables

Table 3.1. Clinical characteristics of the DCIS cohort.

Patient	BCSE?	Race	Ethnicity	Age at diagnosis	Type of surgery	Pathologic size (cm)	Nuclear grade	ER Status (0/1)	Days to BCSE or last followup
OXPA003	Yes	White	Non-Hispanic	46	Lumpectomy	0.8	1	+	2554
OXPA028	Yes	White	Non-Hispanic	51	Mastectomy	0.5	1	+	1012
OXPA033	Yes	White	Non-Hispanic	79	Lumpectomy	NA	1	+	1024
OXPA036	Yes	White	Non-Hispanic	58	Lumpectomy	NA	1	+	2296
OXPA161	Yes	White	Non-Hispanic	62	Mastectomy	0.4	1	+	3752
OXPA166	Yes	White	Non-Hispanic	57	Lumpectomy	0.7	1	+	7037
OXPA020	Yes	White	Non-Hispanic	57	Lumpectomy	NA	1	NA	5967
OXPA021	Yes	White	Non-Hispanic	50	Lumpectomy	0.9	1	NA	2398
OXPA527	Yes	White	Hispanic	42	Lumpectomy	0.7	1	NA	2203
OXPA002	Yes	Asian	Non-Hispanic	44	Lumpectomy	1.4	2	+	2492
OXPA006	Yes	White	Non-Hispanic	66	Lumpectomy	1	2	+	3022
OXPA032	Yes	Asian	Non-Hispanic	42	Lumpectomy	2.5	2	+	502
OXPA044	Yes	White	Non-Hispanic	56	Lumpectomy	0.9	2	+	2147
OXPA064	Yes	White	Non-Hispanic	78	Lumpectomy	0.4	2	+	553
OXPA179	Yes	White	Hispanic	79	Lumpectomy	1.2	2	NA	3077
OXPA147	Yes	White	Non-Hispanic	65	Lumpectomy	1.1	3	+	1981
OXPA185	Yes	White	Non-Hispanic	49	Lumpectomy	0.3	3	+	497
OXPA150	Yes	White	Non-Hispanic	60	Lumpectomy	0.4	NA	+	2402
OXPA153	Yes	White	Non-Hispanic	68	Lumpectomy	0.5	NA	+	3970
OXPA246	Yes	White	Non-Hispanic	79	Lumpectomy	0.5	NA	+	1179
OXPA267	Yes	White	Non-Hispanic	63	Mastectomy	NA	NA	+	3083
OXPA151	Yes	White	Non-Hispanic	72	Lumpectomy	NA	NA	NA	1029
OXPA644	No	White	Non-Hispanic	56	Lumpectomy	1.1	1	-	2456
OXPA347	No	White	Non-Hispanic	83	Lumpectomy	5	1	+	3894
OXPA508	No	White	Non-Hispanic	54	Lumpectomy	0.9	1	+	2570
OXPA172	No	White	Non-Hispanic	72	Lumpectomy	1.8	1	NA	6903
OXPA295	No	White	Non-Hispanic	56	Lumpectomy	1.2	1	NA	4903
OXPA092	No	White	Non-Hispanic	55	Lumpectomy	0.5	2	+	3822
OXPA156	No	White	Hispanic	42	Lumpectomy	0.5	2	+	7170
OXPA392	No	White	Non-Hispanic	58	Lumpectomy	0.6	2	+	3709
OXPA445	No	White	Non-Hispanic	57	Lumpectomy	1.3	2	+	3594
OXPA501	No	Asian	Non-Hispanic	43	Lumpectomy	1.2	2	+	3044
OXPA530	No	White	Hispanic	65	Lumpectomy	1.5	2	+	3167
OXPA540	No	Asian	Non-Hispanic	45	Lumpectomy	1.8	2	+	2941
OXPA182	No	White	Non-Hispanic	55	Lumpectomy	0.5	3	NA	4543
OXPA146	No	White	Non-Hispanic	48	Lumpectomy	1	NA	+	7067

3.8 Supplemental Data, Tables and Figures

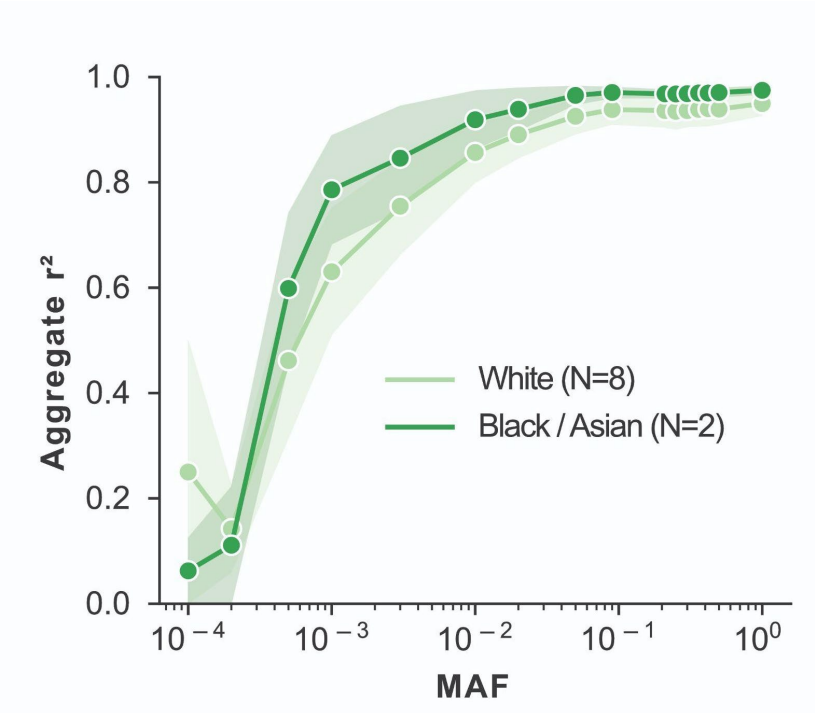


Figure S3.1. Effect of patient ancestry on lc-WGS imputation concordance between blood and tissue.
Comparison of concordance between blood and tissue-based on ancestry background of the patient, with White ancestry in light green and Black or Asian ancestry in dark green. Pearson correlation squared (r^2) is for all aggregated SNPs within a MAF bin. When available, 95% confidence intervals are shaded around the line based on 1000 bootstrap iterations.

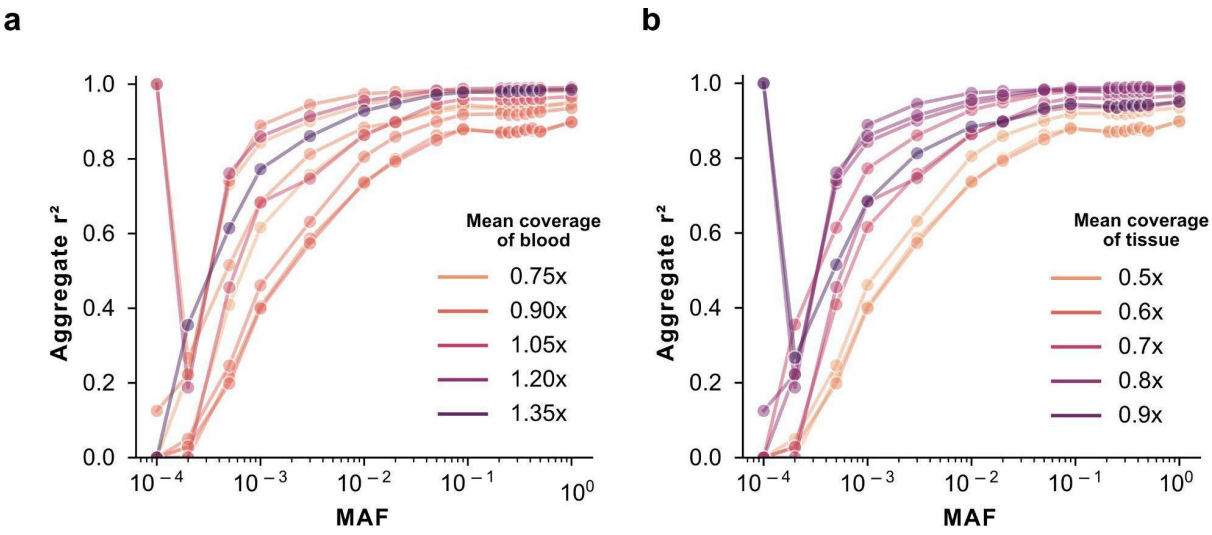


Figure S3.2. Effect of coverage-related features on lc-WGS imputation concordance between blood and tissue.
(a,b) Comparison of concordance as measured by squared Pearson correlation (y-axis) between blood and tissue as a function of MAF (x-axis) based on mean sequencing genome coverage depth of blood (a) or tissue (b).

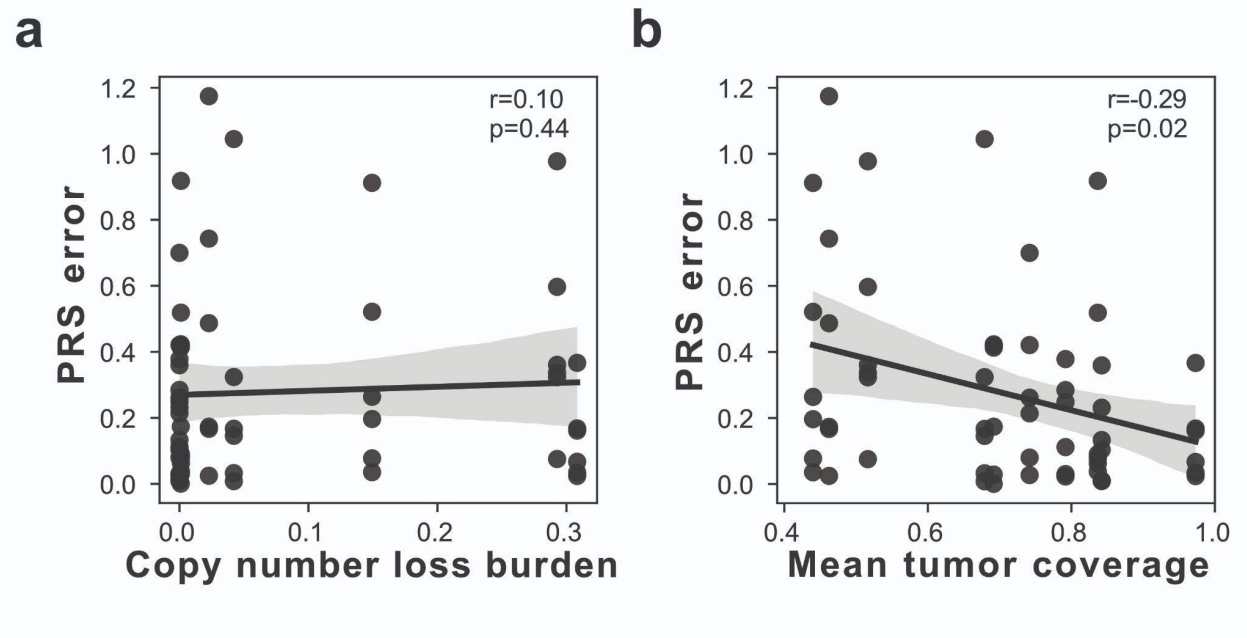


Figure S3.3. Effect of coverage-related features on the error in non-cancer PRS calculation. (a-b) PRS error (y-axis), as measured by the absolute difference between blood and tissue PRS across all non-cancer PRS, as a function of (a) fraction of genome in a copy number loss or (b) mean tissue/tumor genome coverage (x-axis). The 95% confidence intervals are shaded around the line based on 1000 bootstrap iterations. Spearman correlation coefficient, r, and corresponding p-value are indicated as text in the upper right.

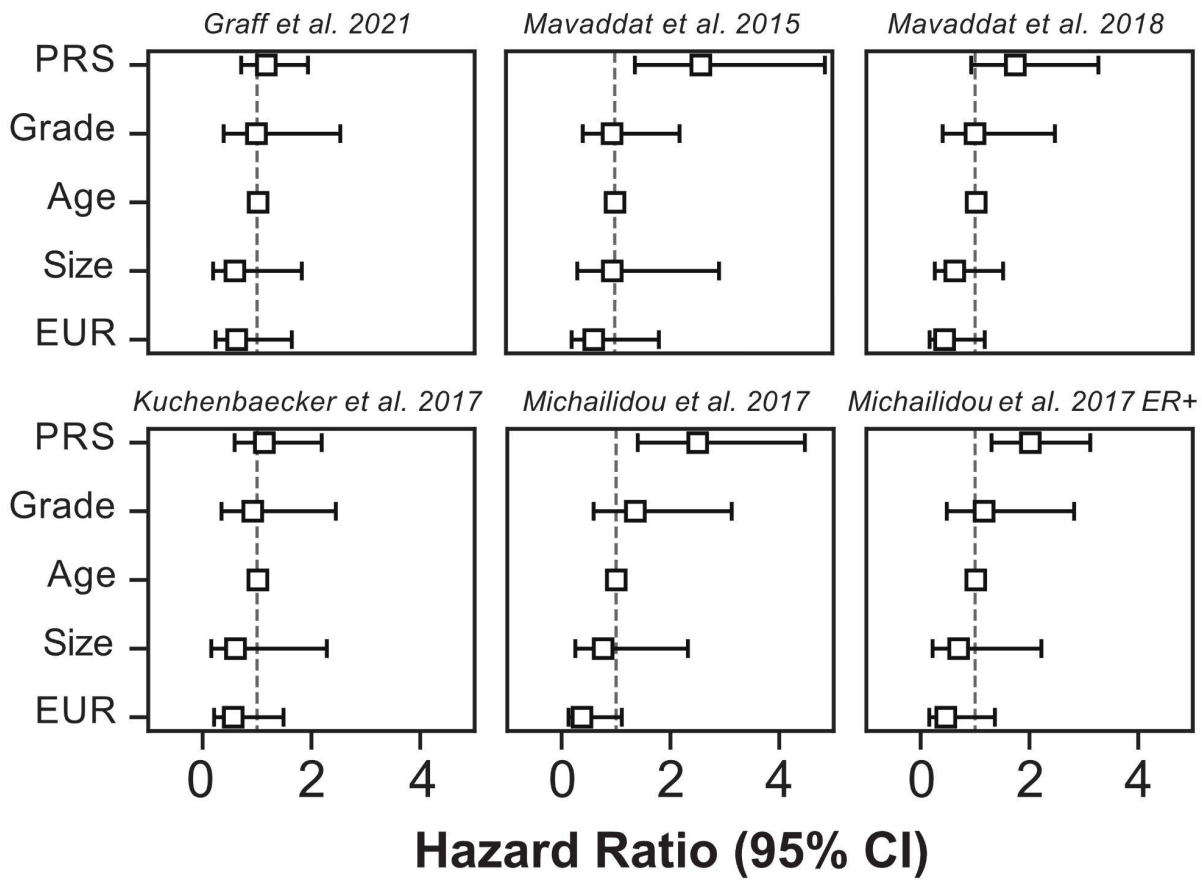


Figure S3.4. Cox proportional hazard models measuring BCSE outcome in DCIS patients for 6 breast cancer PRS.

Forest plot representation of hazard ratios (square) and 95% confidence intervals (error-bars), for each normalized breast cancer PRS and covariates for DCIS BCSE risk including DCIS nuclear grade (Grade), age of the patient at diagnosis (Age), the size of the DCIS lesion (Size), and whether the ancestry of the individual was European (EUR). The dotted line represents a hazard ratio of 1, indicating no effect on BCSE risk, >1 indicating increased, and <1 indicating decreased risk.

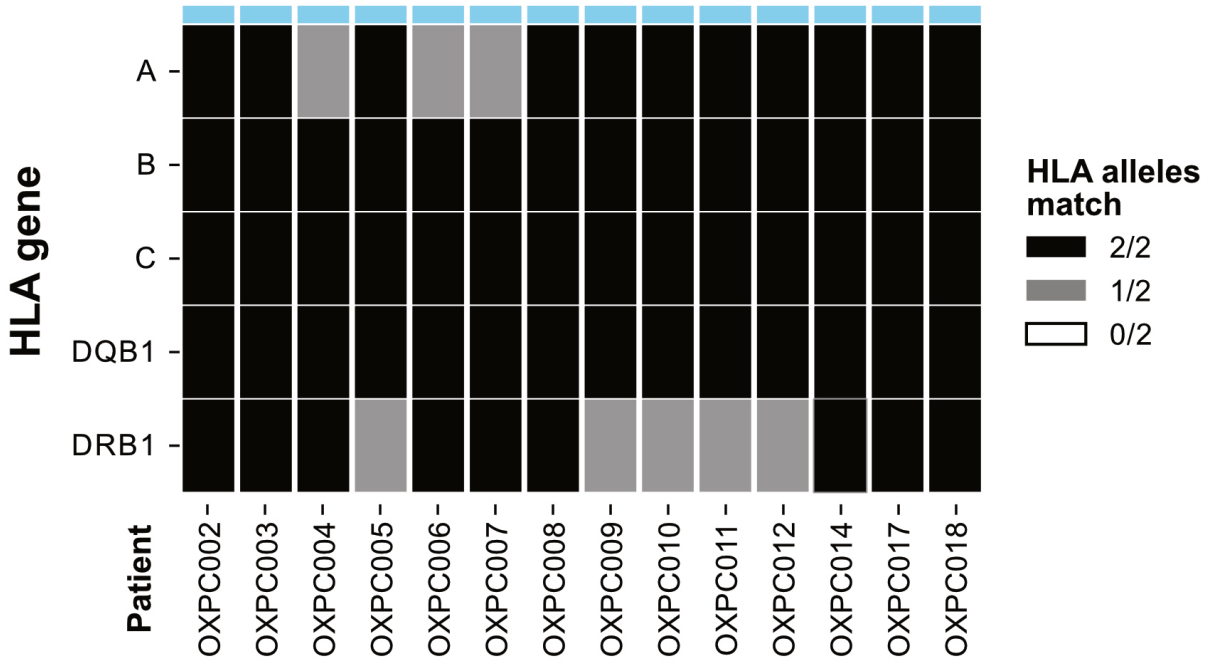


Figure S3.5. Assessment of 2 field HLA-typing accuracy from lc-WGS.
Number of concordant HLA alleles (0: white, 1: grey:, 2: black) between haplotypes from the clinical gold standard and those imputed using QUILT-HLA for class I (A, B, C) and class II (DQB1 and DRB1) HLA genes (rows) using blood DNA of 14 patients.

Table S3.1. Description of the studied samples.

Patient	DNA source	Ancestry	Input DNA (ng)	Coverage	CNA burden	Age of block (yr)[1]	Analysis[2]
OXPC002	Blood	White	68.4	1.11	0.00	-	H
OXPC003	Blood	White	31.8	1.41	0.00	-	GPH
OXPC003	Tissue	White	28.4	0.69	0.00	7	GPH
OXPC004	Blood	White	32.1	0.84	0.00	-	GPH
OXPC004	Tissue	White	5.5	0.44	0.15	7	GPH
OXPC005	Blood	White	53.1	0.79	0.00	5	GPH
OXPC005	Tissue	White	21.8	0.97	0.48	-	GPH
OXPC006	Blood	White	48.9	0.95	0.00	-	H
OXPC007	Blood	White	54.6	0.87	0.00	-	GPH
OXPC007	Tissue	White	35.1	0.46	0.08	6	GPH
OXPC008	Blood	White	45.3	0.89	0.00	-	H
OXPC009	Blood	White	12.2	1.10	0.00	-	GPH
OXPC009	Tissue	White	11.0	0.84	0.01	5	GPH
OXPC010	Blood	Black or African American	16.9	0.85	0.00	-	GPH
OXPC010	Tissue	Black or African American	176.4	0.84	0.01	5	GPH
OXPC011	Blood	White	53.4	0.89	0.00	-	GPH
OXPC011	Tissue	White	20.4	0.52	0.65	5	GPH
OXPC012	Blood	White	66.0	0.71	0.00	-	H
OXPC014	Blood	White	12.1	0.68	0.00	-	GPH
OXPC014	Tissue	White	15.0	0.68	0.11	3	GPH
OXPC017	Blood	White	24.2	0.71	0.00	-	GPH
OXPC017	Tissue	White	30.6	0.79	0.01	3	GPH
OXPC018	Blood	Asian	14.8	1.01	0.00	-	GPH
OXPC018	Tissue	Asian	25.3	0.74	0.00	8	GPH

1. Age of block inferred as the difference in years from DNA extraction (2021) to the year of the diagnosis.
2. Analysis type sample was used in, which was one of:
   GPH: Genome wide, PRS and HLA
   H: HLA only (for samples without both tissue and blood)

Table S3.2. PRS description.

Phenotype	# Variants	# SNPs	# SNPs imputed	# SNPs high-quality	PGSID	Publication source
Bladder cancer	15	15	14	13	PGS000071	https://doi.org/10.1038/s41467-021-21288-z
Breast cancer *(Graff et al., 2021)*	187	172	167	153	PGS000072	https://doi.org/10.1038/s41467-021-21288-z
Cervical cancer	10	9	9	9	PGS000073	https://doi.org/10.1038/s41467-021-21288-z
Colorectal cancer	103	103	101	90	PGS000074	https://doi.org/10.1038/s41467-021-21288-z
Endometrial cancer	9	9	9	9	PGS000075	https://doi.org/10.1038/s41467-021-21288-z
Kidney cancer	19	19	16	16	PGS000076	https://doi.org/10.1038/s41467-021-21288-z
Lymphocytic leukemia	75	75	69	65	PGS000077	https://doi.org/10.1038/s41467-021-21288-z
Lung cancer	109	109	100	85	PGS000078	https://doi.org/10.1038/s41467-021-21288-z
Melanoma	24	24	22	21	PGS000079	https://doi.org/10.1038/s41467-021-21288-z
Non-Hodgkin's lymphoma	19	18	17	16	PGS000080	https://doi.org/10.1038/s41467-021-21288-z
Oral cavity and pharyngeal cancers	14	13	8	8	PGS000081	https://doi.org/10.1038/s41467-021-21288-z
Ovarian cancer	36	32	30	27	PGS000082	https://doi.org/10.1038/s41467-021-21288-z
Pancreatic cancer	22	22	22	18	PGS000083	https://doi.org/10.1038/s41467-021-21288-z
Prostate cancer	161	152	146	136	PGS000084	https://doi.org/10.1038/s41467-021-21288-z
Testicular cancer	52	52	49	45	PGS000086	https://doi.org/10.1038/s41467-021-21288-z
Thyroid cancer	12	11	11	10	PGS000087	https://doi.org/10.1038/s41467-021-21288-z
Type 1 Diabetes (T1D)	825	825	818	747	PGS001817	https://doi.org/10.1016/j.ajhg.2021.11.008
Type 2 Diabetes (T2D)	384	384	378	321	PGS000832	https://doi.org/10.1038/s41588-021-00948-2
High lipoprotein density (HLD)	303	302	265	241	PGS000845	https://doi.org/10.1038/s41588-021-00948-2
Body mass index (BMI)	122	122	122	117	PGS000841	https://doi.org/10.1038/s41588-021-00948-2
Cardiovascular disease (CVD)	330	329	224	212	PGS000863	https://doi.org/10.1038/s41588-021-00948-2
Ulcerative colitis (UC)	179	165	147	132	PGS001306	https://doi.org/10.1101/2021.09.02.21262942
Breast cancer (Kuchenbaecker et al., 2017)	88	86	56	51	PGS000045	https://doi.org/10.1093/jnci/djw302
Breast cancer (Michailidou et al., 2017)	85	85	84	65	PGS000538	https://doi.org/10.1038/nature24284
Breast cancer ER+ (Michailidou et al., 2017)	174	174	170	125	PGS000530	https://doi.org/10.1038/nature24284
Breast cancer (Mavaddat et al., 2015)	77	77	74	67	PGS000001	https://doi.org/10.1093/jnci/djv036
Breast cancer (Mavaddat et al., 2018)	313	265	256	230	PGS000004	https://doi.org/10.1016/j.ajhg.2018.11.002

Table S3.3. DCIS cohort technical characteristic description.

Patient	Input DNA (ng)	Mean coverage	CNA burden	Age of block (yr)[1]	Passed QC?
OXPA002	68.4	0.93	0.05	18	Yes
OXPA003	19.6	1.07	0.04	21	Yes
OXPA006	106.8	0.96	0.06	14	Yes
OXPA020	69.6	1.07	0.02	22	Yes
OXPA021	99	1.11	0.03	21	Yes
OXPA028	29.64	0.84	0.05	16	Yes
OXPA032	38.4	1.09	0.04	15	Yes
OXPA033	300	0.96	0.08	15	Yes
OXPA036	12.7	0.88	0.07	14	Yes
OXPA044	65.4	1.10	0.03	11	Yes
OXPA064	15	1.12	0.04	7	Yes
OXPA092	10.3	0.49	0.02	14	Yes
OXPA146	24.06	0.81	0.08	25	Yes
OXPA147	300	0.74	0.04	25	Yes
OXPA150	30.9	0.92	0.04	24	Yes
OXPA151	64.8	0.79	0.37	24	Yes
OXPA153	201	0.96	0.05	24	Yes
OXPA156	19.9	0.83	0.03	24	Yes
OXPA161	15.4	0.95	0.04	24	Yes
OXPA166	300	0.91	0.04	24	Yes
OXPA172	28.02	1.16	0.03	23	Yes
OXPA179	102	0.79	0.04	23	Yes
OXPA182	31.2	0.83	0.28	23	Yes
OXPA185	52.2	0.89	0.04	23	Yes
OXPA246	109.2	1.16	0.04	20	Yes
OXPA267	204	1.01	0.07	19	Yes
OXPA295	25	1.65	0.02	18	Yes
OXPA347	40.2	1.79	0.05	16	Yes
OXPA392	6.6	1.36	0.04	15	Yes
OXPA445	7.5	1.12	0.02	14	Yes
OXPA501	119.4	0.93	0.04	13	Yes
OXPA508	25.5	0.67	0.05	13	Yes
OXPA527	48.9	0.68	0.10	13	Yes
OXPA530	14.4	0.74	0.30	13	Yes
OXPA540	72.6	0.96	0.05	12	Yes
OXPA644	30.9	0.59	0.04	10	Yes
OXPA007	82	0.15	N/A	11	No (Low cov.)
OXPA035	17.5	0.31	N/A	14	No (Low cov.)
OXPA619	35.2	0.31	N/A	11	No (Low cov.)
OXPA066	300	0.34	N/A	7	No (Low cov.)
OXPB024	53.7	0.47	N/A	6	No (Low cov.)
OXPA025	300	0.49	N/A	17	No (Contam.)
OXPA574	50.1	0.80	N/A	12	No (Contam.)
OXPA005	18.6	0.81	N/A	15	No (Contam.)
OXPA269	11.9	0.91	N/A	19	No (Contam.)
OXPA165	81.6	0.92	N/A	24	No (Contam.)
OXPA040	300	0.93	N/A	13	No (Contam.)
OXPA169	300	0.97	N/A	24	No (Contam.)
OXPB009	224.4	0.99	N/A	7	No (Contam.)
OXPA029	170.4	1.04	N/A	16	No (Contam.)

1. Age of block inferred as the difference in years from DNA extraction (2021) to the year of the diagnosis.

Table S3.4. DCIS cohort covariate association with patient outcome.

Clinical feature		No BCSE[2] (N=14)	BCSE (N=22)	Significance[3]
Grade[1]	Low	36% (5)	41% (9)	*p=0.59*
	Intermediate	50% (7)	27% (6)	
	High	7% (1)	9% (2)	
ER status[1]	+	71% (10)	77% (17)	*p=0.39*
	-	7% (1)	0% (0)	
Ethnicity[1]	Hispanic	14% (2)	9% (2)	*p=0.63*
	Non-Hispanic	86% (12)	91% (20)	
Race	Asian	14% (2)	9% (2)	*p=0.63*
	White	86% (12)	91% (20)	
Pathologic size (cm)[1]		1.35	0.85	*p=0.05*
Age at diagnosis (yrs)		56.3	60.1	*p=0.27*

1. Missing values not represented here, but can be found in Table S3.2.
2. BCSE: Breast cancer subsequent event.
3. P-values were computed using Fisher Exact test for ER status, Race, Ethnicity and Chi-square test for Grade. For continuous features, size and age, Mann-Whitney U test was used to compare groups.

3.9 Author contributions

D.N., L.B., M.P. performed the analysis, J.S., C.C., G.P.M., D.N. generated the data, N.Q.L., T.J.O., G.Y.L., F.H. collected the specimen and reviewed the clinical data. O.H., H.C. directed the study. O.H, D.N. wrote the manuscript. All authors reviewed and approved the manuscript.

3.10 Acknowledgements

We thank Adam Maihofer for his statistical advice, Sharmeela Kaushal, Valeria Estrada, Kimberly Mcintyre, and the staff of the Moores Cancer Center Biorepository and Tissue Technology Shared Resources for the samples collection and processing. We are grateful to Kristen Jepsen and Huazhen Yao from the IGM genomics center for their technical expertise and sample genotyping and sequencing. We also kindly thank CareDx for providing the reagents for the targeted HLA sequencing.

Chapter 3, in full, is a reformatted presentation of the material currently under review titled, "Accurate genome-wide germline profiling from decade-old archival tissue DNA reveals the contribution of common variants to precancer disease outcome" by Daniela Nachmanson, Meghana Pagadala, Joseph Steward, Callie Cheung, Lauryn Keeler Bruce, Nicole Q. Lee, Thomas J. O'Keefe, Grace Y. Lin, Farnaz Hasteh, Gerald P. Morris, Hannah Carter, Olivier Harismendy. The dissertation author was the primary investigator and author of this material.

3.11 References

1. Gusev, A., Groha, S., Taraszka, K., Semenov, Y. R. & Zaitlen, N. Constructing germline research cohorts from the discarded reads of clinical tumor sequences. *Genome Med.* **13,** 179 (2021).

2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).

3. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P. & Marchini, J. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (2018).

4. Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., Sullivan, P. F., Bergen, S., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S. M., Haas, D. W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P. I. W., Reich, D. & Price, A. L. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44,** 631–635 (2012).

5. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53,** 120–126 (2021).

6. Davies, R. W., Kucka, M., Su, D., Shi, S., Flanagan, M., Cunniff, C. M., Chan, Y. F. & Myers, S. Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* **53,** 1104–1111 (2021).

7. Bleyer, A. & Welch, H. G. Effect of three decades of screening mammography on breast-cancer incidence. *N. Engl. J. Med.* **367,** 1998–2005 (2012).

8. Clouston, D. & Bolton, D. In situ and intraductal epithelial proliferations of prostate: definitions and treatment implications. Part 1: Prostatic intraepithelial neoplasia. *BJU Int.* **109 Suppl 3,** 22–26 (2012).

9. Cuzick, J., Sestak, I., Pinder, S. E., Ellis, I. O., Forsyth, S., Bundred, N. J., Forbes, J. F., Bishop, H., Fentiman, I. S. & George, W. D. Effect of tamoxifen and radiotherapy in women with locally excised ductal carcinoma in situ: long-term results from the UK/ANZ DCIS trial. *Lancet Oncol.* **12,** 21–29 (2011).

10. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72,** 7–33 (2022).

11. Benson, J. R. & Wishart, G. C. Predictors of recurrence for ductal carcinoma in situ after breast-conserving surgery. *Lancet Oncol.* **14,** e348–57 (2013).

12. Silverstein, M. J. The University of Southern California/Van Nuys prognostic index for ductal carcinoma in situ of the breast. *Am. J. Surg.* **186,** 337–343 (2003).

13. Nachmanson, D., Steward, J., Yao, H., Officer, A., Jeong, E., O'Keefe, T. J., Hasteh, F., Jepsen, K., Hirst, G. L., Esserman, L. J., Borowsky, A. D. & Harismendy, O. Mutational profiling of micro-dissected pre-malignant lesions from archived specimens. *BMC Med. Genomics* **13,** 173 (2020).

14. Nachmanson, D., Officer, A., Mori, H., Gordon, J., Evans, M. F., Steward, J., Yao, H., O'Keefe, T., Hasteh, F., Stein, G. S., Jepsen, K., Weaver, D. L., Hirst, G. L., Sprague, B. L., Esserman, L. J., Borowsky, A. D., Stein, J. L. & Harismendy, O. The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. *NPJ Breast Cancer* **8,** 6 (2022).

15. O'Keefe, T. J., Blair, S. L., Hosseini, A., Harismendy, O. & Wallace, A. M. HER2-Overexpressing Ductal Carcinoma In Situ Associated with Increased Risk of Ipsilateral Invasive Recurrence, Receptor Discordance with Recurrence. *Cancer Prev. Res.* **13,** 761–772 (2020).

16. Zhou, W., Jirström, K., Amini, R.-M., Fjällskog, M.-L., Sollie, T., Lindman, H., Sørlie, T., Blomqvist, C. & Wärnberg, F. Molecular subtypes in ductal carcinoma in situ of the breast and their relation to prognosis: a population-based cohort study. *BMC Cancer* **13,** 512 (2013).

17. Lips, E. H., Kumar, T., Megalios, A., Visser, L. L., Sheinman, M., Fortunato, A., Shah, V., Hoogstraat, M., Sei, E., Mallo, D. & Others. Genomic profiling defines variable clonal relatedness between invasive breast cancer and primary ductal carcinoma in situ. *medRxiv* (2021). at <https://www.medrxiv.org/content/10.1101/2021.03.22.21253209v1.abstract>

18. Boddicker, N. J., Hu, C., Weitzel, J. N., Kraft, P., Nathanson, K. L., Goldgar, D. E., Na, J., Huang, H., Gnanaolivu, R. D., Larson, N., Yussuf, A., Yao, S., Vachon, C. M., Trentham-Dietz, A., Teras, L., Taylor, J. A., Scott, C. E., Sandler, D. P., Pesaran, T., Patel, A. V., Palmer, J. R., Ong, I. M., Olson, J. E., O'Brien, K., Neuhausen, S., Martinez, E., Ma, H., Lindstrom, S., Le Marchand, L., Kooperberg, C., Karam, R., Hunter, D. J., Hodge, J. M., Haiman, C., Gaudet, M. M., Gao, C., LaDuca, H., Lacey, J. V., Dolinsky, J. S., Chao, E., Carter, B. D., Burnside, E. S., Bertrand, K. A., Bernstein, L., Auer, P. W., Ambrosone, C., Yadav, S., Hart, S. N., Polley, E. C., Domchek, S. M. & Couch, F. J. Risk of Late-Onset Breast Cancer in Genetically Predisposed Women. *J. Clin. Oncol.* **39,** 3430–3440 (2021).

19. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., Perkins, B. J., Czene, K., Eriksson, M., Darabi, H., Brand, J. S., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nielsen, S. F., Rahman, N., Turnbull, C., BOCS, Fletcher, O., Peto, J., Gibson, L., dos-Santos-Silva, I., Chang-Claude, J., Flesch-Janys, D., Rudolph, A., Eilber, U., Behrens, S., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Khan, S., Aaltonen, K., Ahsan, H., Kibriya, M. G., Whittemore, A. S., John, E. M., Malone, K. E., Gammon, M. D., Santella, R. M., Ursin, G., Makalic, E., Schmidt, D. F., Casey, G., Hunter, D. J., Gapstur, S. M., Gaudet, M. M., Diver, W. R., Haiman, C. A., Schumacher, F., Henderson, B. E., Le Marchand, L., Berg, C. D., Chanock, S. J., Figueroa, J., Hoover, R. N., Lambrechts, D., Neven, P., Wildiers, H., van Limbergen, E., Schmidt, M. K., Broeks, A., Verhoef, S., Cornelissen, S., Couch, F. J., Olson, J. E., Hallberg, E., Vachon, C., Waisfisz, Q., Meijers-Heijboer, H., Adank, M. A., van der Luijt, R. B., Li, J., Liu, J., Humphreys, K., Kang, D., Choi, J.-Y., Park, S. K., Yoo, K.-Y.,

Matsuo, K., Ito, H., Iwata, H., Tajima, K., Guénel, P., Truong, T., Mulot, C., Sanchez, M., Burwinkel, B., Marme, F., Surowy, H., Sohn, C., Wu, A. H., Tseng, C.-C., Van Den Berg, D., Stram, D. O., González-Neira, A., Benitez, J., Zamora, M. P., Perez, J. I. A., Shu, X.-O., Lu, W., Gao, Y.-T., Cai, H., Cox, A., Cross, S. S., Reed, M. W. R., Andrulis, I. L., Knight, J. A., Glendon, G., Mulligan, A. M., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., kConFab Investigators, AOCS Group, Lindblom, A., Margolin, S., Teo, S. H., Yip, C. H., Taib, N. A. M., Tan, G.-H., Hooning, M. J., Hollestelle, A., Martens, J. W. M., Collée, J. M., Blot, W., Signorello, L. B., Cai, Q., Hopper, J. L., Southey, M. C., Tsimiklis, H., Apicella, C., Shen, C.-Y., Hsiung, C.-N., Wu, P.-E., Hou, M.-F., Kristensen, V. N., Nord, S., Alnaes, G. I. G., NBCS, Giles, G. G., Milne, R. L., McLean, C., Canzian, F., Trichopoulos, D., Peeters, P., Lund, E., Sund, M., Khaw, K.-T., Gunter, M. J., Palli, D., Mortensen, L. M., Dossus, L., Huerta, J.-M., Meindl, A., Schmutzler, R. K., Sutter, C., Yang, R., Muir, K., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Hartman, M., Miao, H., Chia, K. S., Chan, C. W., Fasching, P. A., Hein, A., Beckmann, M. W., Haeberle, L., Brenner, H., Dieffenbach, A. K., Arndt, V., Stegmaier, C., Ashworth, A., Orr, N., Schoemaker, M. J., Swerdlow, A. J., Brinton, L., Garcia-Closas, M., Zheng, W., Halverson, S. L., Shrubsole, M., Long, J., Goldberg, M. S., Labrèche, F., Dumont, M., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Brauch, H., Hamann, U., Brüning, T., GENICA Network, Radice, P., Peterlongo, P., Manoukian, S., Bernard, L., Bogdanova, N. V., Dörk, T., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J. M., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., Van Asperen, C. J., Jakubowska, A., Lubinski, J., Jaworska, K., Huzarski, T., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Slager, S., Toland, A. E., Ambrosone, C. B., Yannoukakos, D., Kabisch, M., Torres, D., Neuhausen, S. L., Anton-Culver, H., Luccarini, C., Baynes, C., Ahmed, S., Healey, C. S., Tessier, D. C., Vincent, D., Bacot, F., Pita, G., Alonso, M. R., Álvarez, N., Herrero, D., Simard, J., Pharoah, P. P. D. P., Kraft, P., Dunning, A. M., Chenevix-Trench, G., Hall, P. & Easton, D. F. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47,** 373–380 (2015).

20. Petridis, C., Brook, M. N., Shah, V., Kohut, K., Gorman, P., Caneppele, M., Levi, D., Papouli, E., Orr, N., Cox, A., Cross, S. S., Dos-Santos-Silva, I., Peto, J., Swerdlow, A., Schoemaker, M. J., Bolla, M. K., Wang, Q., Dennis, J., Michailidou, K., Benitez, J., González-Neira, A., Tessier, D. C., Vincent, D., Li, J., Figueroa, J., Kristensen, V., Borresen-Dale, A.-L., Soucy, P., Simard, J., Milne, R. L., Giles, G. G., Margolin, S., Lindblom, A., Brüning, T., Brauch, H., Southey, M. C., Hopper, J. L., Dörk, T., Bogdanova, N. V., Kabisch, M., Hamann, U., Schmutzler, R. K., Meindl, A., Brenner, H., Arndt, V., Winqvist, R., Pylkäs, K., Fasching, P. A., Beckmann, M. W., Lubinski, J., Jakubowska, A., Mulligan, A. M., Andrulis, I. L., Tollenaar, R. A. E. M., Devilee, P., Le Marchand, L., Haiman, C. A., Mannermaa, A., Kosma, V.-M., Radice, P., Peterlongo, P., Marme, F., Burwinkel, B., van Deurzen, C. H. M., Hollestelle, A., Miller, N., Kerin, M. J., Lambrechts, D., Floris, G., Wesseling, J., Flyger, H., Bojesen, S. E., Yao, S., Ambrosone, C. B., Chenevix-Trench, G., Truong, T., Guénel, P., Rudolph, A., Chang-Claude, J., Nevanlinna, H., Blomqvist, C., Czene, K., Brand, J. S., Olson, J. E., Couch, F. J., Dunning, A. M., Hall, P., Easton, D. F., Pharoah, P. D. P., Pinder, S. E., Schmidt, M. K., Tomlinson, I., Roylance, R., García-Closas, M. & Sawyer, E. J. Genetic predisposition to ductal carcinoma in situ of the breast. *Breast Cancer Res.* **18,** 22 (2016).

21. Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F. & Antoniou, A. C. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21,** 1708–1718 (2019).

22. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., Wang, Q., Bolla, M. K., Yang, X., Adank, M. A., Ahearn, T., Aittomäki, K., Allen, J., Andrulis, I. L., Anton-Culver, H., Antonenkova, N. N., Arndt, V., Aronson, K. J., Auer, P. L., Auvinen, P., Barrdahl, M., Beane Freeman, L. E., Beckmann, M. W., Behrens, S., Benitez, J., Bermisheva, M., Bernstein, L., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Bonanni, B., Børresen-Dale, A.-L., Brauch, H., Bremer, M., Brenner, H., Brentnall, A., Brock, I. W., Brooks-Wilson, A., Brucker, S. Y., Brüning, T., Burwinkel, B., Campa, D., Carter, B. D., Castelao, J. E., Chanock, S. J., Chlebowski, R., Christiansen, H., Clarke, C. L., Collée, J. M., Cordina-Duverger, E., Cornelissen, S., Couch, F. J., Cox, A., Cross, S. S., Czene, K., Daly, M. B., Devilee, P., Dörk, T., Dos-Santos-Silva, I., Dumont, M., Durcan, L., Dwek, M., Eccles, D. M., Ekici, A. B., Eliassen, A. H., Ellberg, C., Engel, C., Eriksson, M., Evans, D. G., Fasching, P. A., Figueroa, J., Fletcher, O., Flyger, H., Försti, A., Fritschi, L., Gabrielson, M., Gago-Dominguez, M., Gapstur, S. M., García-Sáenz, J. A., Gaudet, M. M., Georgoulias, V., Giles, G. G., Gilyazova, I. R., Glendon, G., Goldberg, M. S., Goldgar, D. E., González-Neira, A., Grenaker Alnæs, G. I., Grip, M., Gronwald, J., Grundy, A., Guénel, P., Haeberle, L., Hahnen, E., Haiman, C. A., Håkansson, N., Hamann, U., Hankinson, S. E., Harkness, E. F., Hart, S. N., He, W., Hein, A., Heyworth, J., Hillemanns, P., Hollestelle, A., Hooning, M. J., Hoover, R. N., Hopper, J. L., Howell, A., Huang, G., Humphreys, K., Hunter, D. J., Jakimovska, M., Jakubowska, A., Janni, W., John, E. M., Johnson, N., Jones, M. E., Jukkola-Vuorinen, A., Jung, A., Kaaks, R., Kaczmarek, K., Kataja, V., Keeman, R., Kerin, M. J., Khusnutdinova, E., Kiiski, J. I., Knight, J. A., Ko, Y.-D., Kosma, V.-M., Koutros, S., Kristensen, V. N., Krüger, U., Kühl, T., Lambrechts, D., Le Marchand, L., Lee, E., Lejbkowicz, F., Lilyquist, J., Lindblom, A., Lindström, S., Lissowska, J., Lo, W.-Y., Loibl, S., Long, J., Lubiński, J., Lux, M. P., MacInnis, R. J., Maishman, T., Makalic, E., Maleva Kostovska, I., Mannermaa, A., Manoukian, S., Margolin, S., Martens, J. W. M., Martinez, M. E., Mavroudis, D., McLean, C., Meindl, A., Menon, U., Middha, P., Miller, N., Moreno, F., Mulligan, A. M., Mulot, C., Muñoz-Garzon, V. M., Neuhausen, S. L., Nevanlinna, H., Neven, P., Newman, W. G., Nielsen, S. F., Nordestgaard, B. G., Norman, A., Offit, K., Olson, J. E., Olsson, H., Orr, N., Pankratz, V. S., Park-Simon, T.-W., Perez, J. I. A., Pérez-Barrios, C., Peterlongo, P., Peto, J., Pinchev, M., Plaseska-Karanfilska, D., Polley, E. C., Prentice, R., Presneau, N., Prokofyeva, D., Purrington, K., Pylkäs, K., Rack, B., Radice, P., Rau-Murthy, R., Rennert, G., Rennert, H. S., Rhenius, V., Robson, M., Romero, A., Ruddy, K. J., Ruebner, M., Saloustros, E., Sandler, D. P., Sawyer, E. J., Schmidt, D. F., Schmutzler, R. K., Schneeweiss, A., Schoemaker, M. J., Schumacher, F., Schürmann, P., Schwentner, L., Scott, C., Scott, R. J., Seynaeve, C., Shah, M., Sherman, M. E., Shrubsole, M. J., Shu, X.-O., Slager, S., Smeets, A., Sohn, C., Soucy, P., Southey, M. C., Spinelli, J. J., Stegmaier, C., Stone, J., Swerdlow, A. J., Tamimi, R. M., Tapper, W. J., Taylor, J. A., Terry, M. B., Thöne, K., Tollenaar, R. A. E. M., Tomlinson, I., Truong, T., Tzardi, M., Ulmer, H.-U., Untch, M., Vachon, C. M., van Veen, E. M., Vijai, J., Weinberg, C. R., Wendt, C., Whittemore, A. S., Wildiers, H., Willett, W., Winqvist, R., Wolk, A., Yang, X. R., Yannoukakos, D., Zhang, Y., Zheng, W., Ziogas, A., ABCTB Investigators, kConFab/AOCS Investigators, NBCS

Collaborators, Dunning, A. M., Thompson, D. J., Chenevix-Trench, G., Chang-Claude, J., Schmidt, M. K., Hall, P., Milne, R. L., Pharoah, P. D. P., Antoniou, A. C., Chatterjee, N., Kraft, P., García-Closas, M., Simard, J. & Easton, D. F. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104,** 21–34 (2019).

23. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Correa, C., McGale, P., Taylor, C., Wang, Y., Clarke, M., Davies, C., Peto, R., Bijker, N., Solin, L. & Darby, S. Overview of the randomized trials of radiotherapy in ductal carcinoma in situ of the breast. *J. Natl. Cancer Inst. Monogr.* **2010,** 162–177 (2010).

24. Sprague, B. L., Vacek, P. M., Herschorn, S. D., James, T. A., Geller, B. M., Trentham-Dietz, A., Stein, J. L. & Weaver, D. L. Time-varying risks of second events following a DCIS diagnosis in the population-based Vermont DCIS cohort. *Breast Cancer Res. Treat.* **174,** 227–235 (2019).

25. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy-Gallego, E., The Human Genome Structural Variation Consortium, Flicek, P., Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., Narzisi, G. & Zody, M. C. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* 2021.02.06.430068 (2021). doi:10.1101/2021.02.06.430068

26. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).

27. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49,** 1304–1310 (2017).

28. Esserman, L., Eklund, M., Veer, L. V., Shieh, Y., Tice, J., Ziv, E., Blanco, A., Kaplan, C., Hiatt, R., Fiscalini, A. S., Yau, C., Scheuner, M., Naeim, A., Wenger, N., Lee, V., Heditsian, D., Brain, S., Parker, B. A., LaCroix, A. Z., Madlensky, L., Hogarth, M., Borowsky, A., Anton-Culver, H., Kaster, A., Olopade, O. I., Sheth, D., Garcia, A., Lancaster, R. & Plaza, M. The WISDOM study: a new approach to screening can and should be tested. *Breast Cancer Res. Treat.* **189,** 593–598 (2021).

29. Graff, R. E., Cavazos, T. B., Thai, K. K., Kachuri, L., Rashkin, S. R., Hoffman, J. D., Alexeeff, S. E., Blatchins, M., Meyers, T. J., Leong, L., Tai, C. G., Emami, N. C., Corley, D. A., Kushi, L. H., Ziv, E., Van Den Eeden, S. K., Jorgenson, E., Hoffmann, T. J., Habel, L. A., Witte, J. S. & Sakoda, L. C. Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat. Commun.* **12,** 970 (2021).

30. Tanigawa, Y., Qian, J., Venkataraman, G., Justesen, J. M., Li, R., Tibshirani, R., Hastie, T. & Rivas, M. A. Significant sparse polygenic risk scores across 813 traits in UK Biobank. *bioRxiv* (2021). doi:10.1101/2021.09.02.21262942

31. Mansour Aly, D., Dwivedi, O. P., Prasad, R. B., Käräjämäki, A., Hjort, R., Thangam, M., Åkerlund, M., Mahajan, A., Udler, M. S., Florez, J. C., McCarthy, M. I., Regeneron Genetics Center, Brosnan, J., Melander, O., Carlsson, S., Hansson, O., Tuomi, T., Groop, L. & Ahlqvist, E. Genome-wide association analyses highlight etiological differences underlying

newly defined subtypes of diabetes. *Nat. Genet.* **53,** 1534–1542 (2021).

32. Mavaddat, N., Pharoah, P. D. P., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., Luben, R., Brown, J., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Czene, K., Darabi, H., Eriksson, M., Peto, J., Dos-Santos-Silva, I., Dudbridge, F., Johnson, N., Schmidt, M. K., Broeks, A., Verhoef, S., Rutgers, E. J., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M. J., Figueroa, J., Chanock, S. J., Brinton, L., Lissowska, J., Couch, F. J., Olson, J. E., Vachon, C., Pankratz, V. S., Lambrechts, D., Wildiers, H., Van Ongeval, C., van Limbergen, E., Kristensen, V., Grenaker Alnæs, G., Nord, S., Borresen-Dale, A.-L., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Fasching, P. A., Haeberle, L., Ekici, A. B., Beckmann, M. W., Burwinkel, B., Marme, F., Schneeweiss, A., Sohn, C., Trentham-Dietz, A., Newcomb, P., Titus, L., Egan, K. M., Hunter, D. J., Lindstrom, S., Tamimi, R. M., Kraft, P., Rahman, N., Turnbull, C., Renwick, A., Seal, S., Li, J., Liu, J., Humphreys, K., Benitez, J., Pilar Zamora, M., Arias Perez, J. I., Menéndez, P., Jakubowska, A., Lubinski, J., Jaworska-Bieniek, K., Durda, K., Bogdanova, N. V., Antonenkova, N. N., Dörk, T., Anton-Culver, H., Neuhausen, S. L., Ziogas, A., Bernstein, L., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., van Asperen, C. J., Cox, A., Cross, S. S., Reed, M. W. R., Khusnutdinova, E., Bermisheva, M., Prokofyeva, D., Takhirova, Z., Meindl, A., Schmutzler, R. K., Sutter, C., Yang, R., Schürmann, P., Bremer, M., Christiansen, H., Park-Simon, T.-W., Hillemanns, P., Guénel, P., Truong, T., Menegaux, F., Sanchez, M., Radice, P., Peterlongo, P., Manoukian, S., Pensotti, V., Hopper, J. L., Tsimiklis, H., Apicella, C., Southey, M. C., Brauch, H., Brüning, T., Ko, Y.-D., Sigurdson, A. J., Doody, M. M., Hamann, U., Torres, D., Ulmer, H.-U., Försti, A., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Andrulis, I. L., Knight, J. A., Glendon, G., Marie Mulligan, A., Chenevix-Trench, G., Balleine, R., Giles, G. G., Milne, R. L., McLean, C., Lindblom, A., Margolin, S., Haiman, C. A., Henderson, B. E., Schumacher, F., Le Marchand, L., Eilber, U., Wang-Gohrke, S., Hooning, M. J., Hollestelle, A., van den Ouweland, A. M. W., Koppert, L. B., Carpenter, J., Clarke, C., Scott, R., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J. M., Brenner, H., Arndt, V., Stegmaier, C., Karina Dieffenbach, A., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Offit, K., Vijai, J., Robson, M., Rau-Murthy, R., Dwek, M., Swann, R., Annie Perkins, K., Goldberg, M. S., Labrèche, F., Dumont, M., Eccles, D. M., Tapper, W. J., Rafiq, S., John, E. M., Whittemore, A. S., Slager, S., Yannoukakos, D., Toland, A. E., Yao, S., Zheng, W., Halverson, S. L., González-Neira, A., Pita, G., Rosario Alonso, M., Álvarez, N., Herrero, D., Tessier, D. C., Vincent, D., Bacot, F., Luccarini, C., Baynes, C., Ahmed, S., Maranian, M., Healey, C. S., Simard, J., Hall, P., Easton, D. F. & Garcia-Closas, M. Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* **107,** (2015).

33. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., Bolla, M. K., Wang, Q., Tyrer, J., Dicks, E., Lee, A., Wang, Z., Allen, J., Keeman, R., Eilber, U., French, J. D., Qing Chen, X., Fachal, L., McCue, K., McCart Reed, A. E., Ghoussaini, M., Carroll, J. S., Jiang, X., Finucane, H., Adams, M., Adank, M. A., Ahsan, H., Aittomäki, K., Anton-Culver, H., Antonenkova, N. N., Arndt, V., Aronson, K. J., Arun, B., Auer, P. L., Bacot, F., Barrdahl, M., Baynes, C., Beckmann, M. W., Behrens, S., Benitez, J., Bermisheva, M., Bernstein, L., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Bonanni, B., Børresen-Dale, A.-L., Brand, J. S., Brauch, H., Brennan, P.,

Brenner, H., Brinton, L., Broberg, P., Brock, I. W., Broeks, A., Brooks-Wilson, A., Brucker, S. Y., Brüning, T., Burwinkel, B., Butterbach, K., Cai, Q., Cai, H., Caldés, T., Canzian, F., Carracedo, A., Carter, B. D., Castelao, J. E., Chan, T. L., David Cheng, T.-Y., Seng Chia, K., Choi, J.-Y., Christiansen, H., Clarke, C. L., NBCS Collaborators, Collée, M., Conroy, D. M., Cordina-Duverger, E., Cornelissen, S., Cox, D. G., Cox, A., Cross, S. S., Cunningham, J. M., Czene, K., Daly, M. B., Devilee, P., Doheny, K. F., Dörk, T., Dos-Santos-Silva, I., Dumont, M., Durcan, L., Dwek, M., Eccles, D. M., Ekici, A. B., Eliassen, A. H., Ellberg, C., Elvira, M., Engel, C., Eriksson, M., Fasching, P. A., Figueroa, J., Flesch-Janys, D., Fletcher, O., Flyger, H., Fritschi, L., Gaborieau, V., Gabrielson, M., Gago-Dominguez, M., Gao, Y.-T., Gapstur, S. M., García-Sáenz, J. A., Gaudet, M. M., Georgoulias, V., Giles, G. G., Glendon, G., Goldberg, M. S., Goldgar, D. E., González-Neira, A., Grenaker Alnæs, G. I., Grip, M., Gronwald, J., Grundy, A., Guénel, P., Haeberle, L., Hahnen, E., Haiman, C. A., Håkansson, N., Hamann, U., Hamel, N., Hankinson, S., Harrington, P., Hart, S. N., Hartikainen, J. M., Hartman, M., Hein, A., Heyworth, J., Hicks, B., Hillemanns, P., Ho, D. N., Hollestelle, A., Hooning, M. J., Hoover, R. N., Hopper, J. L., Hou, M.-F., Hsiung, C.-N., Huang, G., Humphreys, K., Ishiguro, J., Ito, H., Iwasaki, M., Iwata, H., Jakubowska, A., Janni, W., John, E. M., Johnson, N., Jones, K., Jones, M., Jukkola-Vuorinen, A., Kaaks, R., Kabisch, M., Kaczmarek, K., Kang, D., Kasuga, Y., Kerin, M. J., Khan, S., Khusnutdinova, E., Kiiski, J. I., Kim, S.-W., Knight, J. A., Kosma, V.-M., Kristensen, V. N., Krüger, U., Kwong, A., Lambrechts, D., Le Marchand, L., Lee, E., Lee, M. H., Lee, J. W., Neng Lee, C., Lejbkowicz, F., Li, J., Lilyquist, J., Lindblom, A., Lissowska, J., Lo, W.-Y., Loibl, S., Long, J., Lophatananon, A., Lubinski, J., Luccarini, C., Lux, M. P., Ma, E. S. K., MacInnis, R. J., Maishman, T., Makalic, E., Malone, K. E., Kostovska, I. M., Mannermaa, A., Manoukian, S., Manson, J. E., Margolin, S., Mariapun, S., Martinez, M. E., Matsuo, K., Mavroudis, D., McKay, J., McLean, C., Meijers-Heijboer, H., Meindl, A., Menéndez, P., Menon, U., Meyer, J., Miao, H., Miller, N., Taib, N. A. M., Muir, K., Mulligan, A. M., Mulot, C., Neuhausen, S. L., Nevanlinna, H., Neven, P., Nielsen, S. F., Noh, D.-Y., Nordestgaard, B. G., Norman, A., Olopade, O. I., Olson, J. E., Olsson, H., Olswold, C., Orr, N., Pankratz, V. S., Park, S. K., Park-Simon, T.-W., Lloyd, R., Perez, J. I. A., Peterlongo, P., Peto, J., Phillips, K.-A., Pinchev, M., Plaseska-Karanfilska, D., Prentice, R., Presneau, N., Prokofyeva, D., Pugh, E., Pylkäs, K., Rack, B., Radice, P., Rahman, N., Rennert, G., Rennert, H. S., Rhenius, V., Romero, A., Romm, J., Ruddy, K. J., Rüdiger, T., Rudolph, A., Ruebner, M., Rutgers, E. J. T., Saloustros, E., Sandler, D. P., Sangrajrang, S., Sawyer, E. J., Schmidt, D. F., Schmutzler, R. K., Schneeweiss, A., Schoemaker, M. J., Schumacher, F., Schürmann, P., Scott, R. J., Scott, C., Seal, S., Seynaeve, C., Shah, M., Sharma, P., Shen, C.-Y., Sheng, G., Sherman, M. E., Shrubsole, M. J., Shu, X.-O., Smeets, A., Sohn, C., Southey, M. C., Spinelli, J. J., Stegmaier, C., Stewart-Brown, S., Stone, J., Stram, D. O., Surowy, H., Swerdlow, A., Tamimi, R., Taylor, J. A., Tengström, M., Teo, S. H., Beth Terry, M., Tessier, D. C., Thanasitthichai, S., Thöne, K., Tollenaar, R. A. E. M., Tomlinson, I., Tong, L., Torres, D., Truong, T., Tseng, C.-C., Tsugane, S., Ulmer, H.-U., Ursin, G., Untch, M., Vachon, C., van Asperen, C. J., Van Den Berg, D., van den Ouweland, A. M. W., van der Kolk, L., van der Luijt, R. B., Vincent, D., Vollenweider, J., Waisfisz, Q., Wang-Gohrke, S., Weinberg, C. R., Wendt, C., Whittemore, A. S., Wildiers, H., Willett, W., Winqvist, R., Wolk, A., Wu, A. H., Xia, L., Yamaji, T., Yang, X. R., Har Yip, C., Yoo, K.-Y., Yu, J.-C., Zheng, W., Zheng, Y., Zhu, B., Ziogas, A., Ziv, E., ABCTB Investigators, ConFab/AOCS Investigators, Lakhani, S. R., Antoniou, A. C., Droit, A., Andrulis, I. L., Amos, C. I., Couch, F. J., Pharoah, P. D. P., Chang-Claude, J., Hall, P.,

Hunter, D. J., Milne, R. L., García-Closas, M., Schmidt, M. K., Chanock, S. J., Dunning, A. M., Edwards, S. L., Bader, G. D., Chenevix-Trench, G., Simard, J., Kraft, P. & Easton, D. F. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551,** 92–94 (2017).

34. Kuchenbaecker, K. B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., Domchek, S. M., Robson, M., Spurdle, A. B., Ramus, S. J., Mavaddat, N., Terry, M. B., Neuhausen, S. L., Schmutzler, R. K., Simard, J., Pharoah, P. D. P., Offit, K., Couch, F. J., Chenevix-Trench, G., Easton, D. F. & Antoniou, A. C. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *J. Natl. Cancer Inst.* **109,** (2017).

35. Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M. J., van de Haar, J., Engin, H. B., de Prisco, N., Ideker, T., Hildebrand, W. H., Font-Burgada, J. & Carter, H. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **171,** 1272–1283.e15 (2017).

36. Marty Pyke, R., Thompson, W. K., Salem, R. M., Font-Burgada, J., Zanetti, M. & Carter, H. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell* **175,** 1991 (2018).

37. McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., Birkbak, N. J., Veeriah, S., Van Loo, P., Herrero, J., Swanton, C. & TRACERx Consortium. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171,** 1259–1271.e11 (2017).

38. Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., Mosley, J. D., Mallal, S., Denny, J. C., Phillips, E. J. & Roden, D. M. Comparison of HLA allelic imputation programs. *PLoS One* **12,** e0172444 (2017).

39. Wheeland, D. G. Final NIH genomic data sharing policy. *Fed. Regist.* **79,** 51345–51354 (2014).

40. Gorringe, K. L. & Fox, S. B. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. *Front. Oncol.* **7,** 248 (2017).

41. Esserman, L. J. & WISDOM Study and Athena Investigators. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. *NPJ Breast Cancer* **3,** 34 (2017).

42. Alaeikhanehshir, S., Engelhardt, E. G., van Duijnhoven, F. H., van Seijen, M., Bhairosing, P. A., Pinto, D., Collyar, D., Sawyer, E., Hwang, S. E., Thompson, A. M., Wesseling, J., Lips, E. H., Schmidt, M. K. & PRECISION. The impact of patient characteristics and lifestyle factors on the risk of an ipsilateral event after a primary DCIS: A systematic review. *Breast* **50,** 95–103 (2020).

43. Cheung, S., Booth, M. E., Kearins, O. & Dodwell, D. Risk of subsequent invasive breast cancer after a diagnosis of ductal carcinoma in situ (DCIS). *Breast* **23,** 807–811 (2014).

44. Curigliano, G., Disalvatore, D., Esposito, A., Pruneri, G., Lazzeroni, M., Guerrieri-Gonzaga, A., Luini, A., Orecchia, R., Goldhirsch, A., Rotmensz, N., Bonanni, B. & Viale, G. Risk of subsequentin situ and invasive breast cancer in human epidermal growth factor

receptor 2-positive ductal carcinomain situ. *Ann. Oncol.* **26,** 682–687 (2015).

45. Strand, S. H., Rivero-Gutiérrez, B., Houlahan, K. E., Seoane, J. A., King, L., Risom, T., Simpson, L. A., Vennam, S., Khan, A., Cisneros, L., Hardman, T., Harmon, B., Couch, F., Gallagher, K., Kilgore, M., Wei, S., DeMichele, A., King, T., McAuliffe, P. F., Nangia, J., Lee, J., Tseng, J., Storniolo, A. M., Thompson, A., Gupta, G., Burns, R., Veis, D. J., DeSchryver, K., Zhu, C., Matusiak, M., Wang, J., Zhu, S. X., Tappenden, J., Ding, D. Y., Zhang, D., Luo, J., Jiang, S., Varma, S., Anderson, L., Straub, C., Srivastava, S., Curtis, C., Tibshirani, R., Angelo, R. M., Hall, A., Owzar, K., Polyak, K., Maley, C., Marks, J. R., Colditz, G. A., Shelley Hwang, E. & West, R. B. DCIS genomic signatures define biology and correlate with clinical outcome: a Human Tumor Atlas Network (HTAN) analysis of TBCRC 038 and RAHBT cohorts. *bioRxiv* 2021.06.16.448585 (2021). doi:10.1101/2021.06.16.448585

46. Mangino, M., Roederer, M., Beddall, M. H., Nestle, F. O. & Spector, T. D. Innate and adaptive immune traits are differentially affected by genetic and environmental factors. *Nat. Commun.* **8,** 13850 (2017).

47. Orrù, V., Steri, M., Sole, G., Sidore, C., Virdis, F., Dei, M., Lai, S., Zoledziewska, M., Busonero, F., Mulas, A., Floris, M., Mentzen, W. I., Urru, S. A. M., Olla, S., Marongiu, M., Piras, M. G., Lobina, M., Maschio, A., Pitzalis, M., Urru, M. F., Marcelli, M., Cusano, R., Deidda, F., Serra, V., Oppo, M., Pilu, R., Reinier, F., Berutti, R., Pireddu, L., Zara, I., Porcu, E., Kwong, A., Brennan, C., Tarrier, B., Lyons, R., Kang, H. M., Uzzau, S., Atzeni, R., Valentini, M., Firinu, D., Leoni, L., Rotta, G., Naitza, S., Angius, A., Congia, M., Whalen, M. B., Jones, C. M., Schlessinger, D., Abecasis, G. R., Fiorillo, E., Sanna, S. & Cucca, F. Genetic variants regulating immune cell levels in health and disease. *Cell* **155,** 242–256 (2013).

48. Garrido, F. *MHC Class-I Loss and Cancer Immune Escape*. (Springer International Publishing, 2020).

49. Sade-Feldman, M., Jiao, Y. J., Chen, J. H., Rooney, M. S., Barzily-Rokni, M., Eliane, J.-P., Bjorgaard, S. L., Hammond, M. R., Vitzthum, H., Blackmon, S. M., Frederick, D. T., Hazar-Rethinam, M., Nadres, B. A., Van Seventer, E. E., Shukla, S. A., Yizhak, K., Ray, J. P., Rosebrock, D., Livitz, D., Adalsteinsson, V., Getz, G., Duncan, L. M., Li, B., Corcoran, R. B., Lawrence, D. P., Stemmer-Rachamimov, A., Boland, G. M., Landau, D. A., Flaherty, K. T., Sullivan, R. J. & Hacohen, N. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8,** 1136 (2017).

50. Zaretsky, J. M., Garcia-Diaz, A., Shin, D. S., Escuin-Ordinas, H., Hugo, W., Hu-Lieskovan, S., Torrejon, D. Y., Abril-Rodriguez, G., Sandoval, S., Barthly, L., Saco, J., Homet Moreno, B., Mezzadra, R., Chmielowski, B., Ruchalski, K., Shintaku, I. P., Sanchez, P. J., Puig-Saus, C., Cherry, G., Seja, E., Kong, X., Pang, J., Berent-Maoz, B., Comin-Anduix, B., Graeber, T. G., Tumeh, P. C., Schumacher, T. N. M., Lo, R. S. & Ribas, A. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375,** 819–829 (2016).

51. Goodman, A. M., Castro, A., Pyke, R. M., Okamura, R., Kato, S., Riviere, P., Frampton,

G., Sokol, E., Zhang, X., Ball, E. D., Carter, H. & Kurzrock, R. MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome Med.* **12,** 45 (2020).

52. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., van Duijn, C. M., Gillies, C. E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J. C., Boomsma, D., Branham, K., Breen, G., Brummett, C. M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F. S., Corbin, L. J., Smith, G. D., Dedoussis, G., Dorr, M., Farmaki, A.-E., Ferrucci, L., Forer, L., Fraser, R. M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O. L., Hveem, K., Kretzler, M., Lee, J. C., McGue, M., Meitinger, T., Melzer, D., Min, J. L., Mohlke, K. L., Vincent, J. B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J. B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P. E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van den Berg, L. H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M. G., Wilson, J. F., Frayling, T., de Bakker, P. I. W., Swertz, M. A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C. A., Myers, R. M., Boehnke, M., McCarthy, M. I., Durbin, R. & Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).

53. Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-B., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., Shetty, A. C., Blackwell, T. W., Smith, A. V., Wong, Q., Liu, X., Conomos, M. P., Bobo, D. M., Aguet, F., Albert, C., Alonso, A., Ardlie, K. G., Arking, D. E., Aslibekyan, S., Auer, P. L., Barnard, J., Barr, R. G., Barwick, L., Becker, L. C., Beer, R. L., Benjamin, E. J., Bielak, L. F., Blangero, J., Boehnke, M., Bowden, D. W., Brody, J. A., Burchard, E. G., Cade, B. E., Casella, J. F., Chalazan, B., Chasman, D. I., Chen, Y.-D. I., Cho, M. H., Choi, S. H., Chung, M. K., Clish, C. B., Correa, A., Curran, J. E., Custer, B., Darbar, D., Daya, M., de Andrade, M., DeMeo, D. L., Dutcher, S. K., Ellinor, P. T., Emery, L. S., Eng, C., Fatkin, D., Fingerlin, T., Forer, L., Fornage, M., Franceschini, N., Fuchsberger, C., Fullerton, S. M., Germer, S., Gladwin, M. T., Gottlieb, D. J., Guo, X., Hall, M. E., He, J., Heard-Costa, N. L., Heckbert, S. R., Irvin, M. R., Johnsen, J. M., Johnson, A. D., Kaplan, R., Kardia, S. L. R., Kelly, T., Kelly, S., Kenny, E. E., Kiel, D. P., Klemmer, R., Konkle, B. A., Kooperberg, C., Köttgen, A., Lange, L. A., Lasky-Su, J., Levy, D., Lin, X., Lin, K.-H., Liu, C., Loos, R. J. F., Garman, L., Gerszten, R., Lubitz, S. A., Lunetta, K. L., Mak, A. C. Y., Manichaikul, A., Manning, A. K., Mathias, R. A., McManus, D. D., McGarvey, S. T., Meigs, J. B., Meyers, D. A., Mikulla, J. L., Minear, M. A., Mitchell, B. D., Mohanty, S., Montasser, M. E., Montgomery, C., Morrison, A. C., Murabito, J. M., Natale, A., Natarajan, P., Nelson, S. C., North, K. E., O'Connell, J. R., Palmer, N. D., Pankratz, N., Peloso, G. M., Peyser, P. A., Pleiness, J., Post, W. S., Psaty, B. M., Rao, D. C., Redline, S., Reiner, A. P., Roden, D., Rotter, J. I., Ruczinski, I., Sarnowski, C., Schoenherr, S., Schwartz, D. A., Seo, J.-S., Seshadri, S., Sheehan, V. A., Sheu, W. H., Shoemaker, M. B., Smith, N. L., Smith, J. A., Sotoodehnia, N., Stilp, A. M., Tang, W., Taylor, K. D., Telen, M., Thornton, T. A., Tracy, R. P., Van Den Berg, D. J., Vasan, R. S., Viaud-Martinez, K. A., Vrieze, S., Weeks, D. E., Weir, B. S., Weiss, S. T., Weng, L.-

C., Willer, C. J., Zhang, Y., Zhao, X., Arnett, D. K., Ashley-Koch, A. E., Barnes, K. C., Boerwinkle, E., Gabriel, S., Gibbs, R., Rice, K. M., Rich, S. S., Silverman, E. K., Qasba, P., Gan, W., NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Papanicolaou, G. J., Nickerson, D. A., Browning, S. R., Zody, M. C., Zöllner, S., Wilson, J. G., Cupples, L. A., Laurie, C. C., Jaquish, C. E., Hernandez, R. D., O'Connor, T. D. & Abecasis, G. R. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590,** 290–299 (2021).

54. bcl2fastq. at <https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html>

55. Didion, J. P., Martin, M. & Collins, F. S. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5,** e3720 (2017).

56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013). at <http://arxiv.org/abs/1303.3997>

57. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9,** 13 (2014).

58. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34,** 867–868 (2018).

59. Zhang, F., Flickinger, M., Taliun, S. A. G., InPSYght Psychiatric Genetics Consortium, Abecasis, G. R., Scott, L. J., McCaroll, S. A., Pato, C. N., Boehnke, M. & Kang, H. M. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res.* **30,** 185–194 (2020).

60. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Others. Twelve years of SAMtools and BCFtools. Gigascience 10: giab008. (2021).

61. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12,** e1004873 (2016).

62. Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P. & Marsh, S. G. E. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43,** D423–31 (2015).

63. Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., Danesh, J., MacArthur, J. A. L. & Inouye, M. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53,** 420–425 (2021).

64. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4,** 1317 (2019).

65. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95,** 221–227 (2005).

66. Cinar, O. & Viechtbauer, W. PoolR: Package for pooling the results from (dependent) tests. (2016).

EPILOGUE

Conclusion

Over the next decade, as attention shifts to early cancer detection and prevention, one of the key challenges will be to elucidate the biology and trajectory of precancer [1]. By understanding the mechanisms of precancer development and progression to malignancy, we can identify better therapies for cancer prevention, identify individuals at high risk of progression and gain a detailed understanding of cancer evolution. As precancer progression can happen over the span of a decade, and biopsies are small and often formalin-fixed and paraffin-embedded (FFPE), precancer has historically been logistically difficult to study. In this dissertation, I demonstrated the possibility to perform comprehensive profiling of both acquired and inherited genetic alterations from some of the most challenging, but abundant, clinical specimens. The ability to sequence LCM FFPE tissue contributed to the multi-modal profiling of a cohort of DCIS and the construction of a breast precancer atlas. Likewise, in establishing a germline variant calling workflow from low-coverage genome sequencing of only archival tissue, we were able to assess the contribution of inherited variation to breast precancer outcome.

In Chapter 1, I illustrated that we can obtain reliable somatic mutation calls from poor quality and as little DNA as 3 ng or the equivalent of ~500 cells [1]. Using DNA from mirrored frozen and FFPE tissue I benchmarked somatic mutation calls and provided estimates for sensitivity and precision. Further, to reduce sample sequencing failure, I used pre-sequencing metrics from 166 FFPE library preparations to construct a model to predict exome coverage which we integrated into a web tool called PROjeCt ExomE Depth (PROCEED). We then applied this work to the somatic mutation profiling of ductal carcinoma in situ (DCIS).

In Chapter 2, I presented a multi-modal characterization of a total of 85 dissected regions from 39 patients with breast precancer in the absence of any invasive disease [2]. This cohort featured DCIS with varying nuclear grades, including low-grade, receptor status, and histological architectures representative of what would be observed in the clinic. We performed whole-exome sequencing enabled by the efforts of Chapter 1, RNA sequencing, histopathological analysis, and mIHC for immune cell abundances on multiple regions of the same tissue slide. We found that even between regions within millimeters of one another, there is abundant genetic, transcriptional, and immune microenvironment diversity, which is often correlated to one another. Further, we identified novel histological and molecular associations and an immune epithelium exclusion phenotype with evidence of immune evasion features. This work represents an important landscape for pure DCIS across a spectrum of diverse presentations.

Lastly, in Chapter 3, we reconsider whether we can use only archival tissue instead of traditional sources of germline DNA (i.e. blood, saliva, urine), to perform a germline genetic study. We performed low-coverage whole-genome sequencing (lc-WGS) followed by reference haplotype-based imputation to obtain patient genotypes directly from archival tissue. By comparing genotypes from blood versus archival tissue of the same patient I found that archival tissue faithfully represented the germline profile of common SNPs obtained from blood both at the genome-wide level, across well-established PRS, and even HLA-types. I then used this framework to test whether lifetime breast cancer susceptibility, as captured by established breast cancer PRS, and progression of DCIS share the same genetic risk factors. In a case-control group of selected patients (N=36), we did indeed find that increased breast PRS could be used as a prognostic marker for a subsequent DCIS or progression outcome. This work as a whole supports the future utilization

of archival tissue to construct large retrospective studies to characterize the role of germline variants in disease etiology, progression, and treatment.

By optimizing and measuring the feasibility of both somatic and germline profiling of the genetic profiles of precancer lesions, I was able to apply these methods to map DCIS and identify genetic prognostic markers of DCIS outcome. This work identified novel findings in the molecular features of a diverse array of DCIS, and overall helped advance our understanding of early breast tumorigenesis.

Limitations and Future Directions

Though this work represents important steps forward in the accessibility of clinical samples to molecular characterization, mapping of ductal carcinoma in situ molecular (DCIS) features, and identification of high-risk adverse outcome biomarkers, there are still many avenues of study to continue down and important limitations to keep in mind.

Both in Chapters 1 and 3, we utilized archival tissues and showed the validity of the nucleic material for somatic and germline profiling respectively. In both utilizations, we still experienced sample failure, either due to insufficient coverage or the presence of contaminating DNA. We mitigated some exome sample failure with our trained model PROCEED, to help predict which samples would likely not produce sufficient exome coverage, and a similar approach could help with the lc–WGS data as well. Samples with very little DNA are particularly susceptible to DNA contamination which is hard to detect until samples are sequenced and thus represents a source of sample failure that is harder to address. However, optimized archival tissue block processing, including the use of PCR-free lab stations and careful replacement of tissue sectioning blades between every tissue block, could offer some help. Lastly, although the blunt-end ligation strategy offered improved data recovery, the multi-step ligations it required generated a slightly tedious

protocol. Simultaneously to our efforts to perform DNA sequencing of low-abundance DNA from LCM archival clinical samples in Chapter 1, a similar method was being optimized at the Wellcome Sanger Institute [4]. This method utilizes a library preparation protocol that includes enzymatic fragmentation, end-repair, and adapter ligation performed in the same tube before PCR amplification, offering a simplified workflow. Both the enzymatic and blunt-end ligation-based library preparations offer increased coverage over standard sonication fragmentation and A-tail ligation-based protocols for low-input DNA. While in Chapter 2 we used the blunt-end ligation-based library preparation, in Chapter 3, the DNA was prepared using the enzymatic approach for the simplified workflow.

Although we demonstrated that we were able to call somatic mutations with limited additional artifacts there are still complications with variant calling from tumor-only FFPE samples. In the Discussion of Chapter 1, we detail the challenges of germline removal. As a brief summary, blood which is typically used to distinguish germline variants from somatic mutations is typically unavailable retrospectively. Sequencing of adjacent normal tissue can be technically challenging and suboptimal as it often harbors somatic mutations due to "field effects" evidenced in many tissue types, including breast [5–9]. Further, in Chapter 1 we observed that FFPE artifacts in our mirrored FFPE and frozen sample were pervasive and present even at variant allelic fractions (VAF) above 10%. To mitigate both the effects of likely residual germline variants and potential FFPE-based artifacts in the genomic profiling of DCIS in Chapter 2, we sequenced an additional panel of (N=18) unrelated normal DNA, processed using the exact same protocol and analysis [4]. Further, to proceed with caution, we focused our DCIS genomics results in Chapter 2 on specific driver alterations, or multi-region comparisons, where the phylogenetic trees were also confirmed based on CNA, which would be minimally or not at all affected by FFPE artifacts or germline

contamination. Conclusions for the DCIS that relied on measuring mutational burden, were presented with the caveat that rare or private residual germline variants may exist among the DCIS somatic mutation calls. The implementation of the stringent threshold for mutations, especially in terms of VAF, limits the exploration of genetic diversity in precancers, where lower-frequency subclones may be hard to distinguish from artifacts. There are several ways to improve both of these issues in the future. More extensive panels of sequenced normal tissue, matched for patient ethnicity, protocol, and analysis type would continue to improve germline and artifact removal in precancer. Additionally, algorithmic improvements in removing FFPE or DNA damage artifacts that leverage machine learning or mutational signature-based systems will likely support improved accuracy exome- or genome-wide somatic profiling from FFPE tissues in the future [10–12]. In cancer, intratumoral genomic heterogeneity has important implications for disease progression and treatment outcomes and is considered the "fuel for clonal evolution" [13,14]. In DCIS, genetic intra-lesion heterogeneity has already been shown to be as extensive as IBC, but the relationship between this genetic heterogeneity and long-term disease outcome is yet to be understood [15,16]. Improvements to the low-frequency mutation detection in archival tissue-derived DNA will be critical to addressing these questions.

In Chapter 2, I presented a multi-modal characterization of a total of 85 dissected regions from 43 patients with breast precancer. Despite our success in profiling incredibly challenging specimens, we did not assess our findings in the context of the inherited genetic landscape. There is an increasing appreciation for the interplay between the germline background and acquired somatic alterations and how that impacts cancer risk, and tumorigenesis [17]. In breast, a cancer type with a particularly strong genetic contribution, future work integrating germline variants and somatic mutations in DCIS would be particularly valuable [18,19]. Further, larger cohorts which can

be sufficiently powered to link molecular features with outcomes will be critical for the field. During the time that we released this work, two other groups published pre-prints of their breast cancer atlases concurrently, including an atlas led by the Human Tumor Atlas Network (HTAN) and the PREvent ductal Carcinoma In Situ Invasive Overtreatment Now (PRECISION) consortium [20,21]. Each atlas presents a different angle on DCIS and uses varying methodologies that have both benefits and drawbacks i.e. low-coverage whole-genome sequencing for genomic profiling, single-cell genomics, and transcriptomics, or multiplexed ion beam imaging for immune microenvironment characterization. Future atlases make also incorporate spatial-omics directly from FFPE tissue as it becomes higher throughput and has increased resolution, this may be particularly pertinent for making spatially resolved molecular observations without tissue dissection [22]. These additional atlases also provide excellent independent cohorts for validation amongst one another for compatible data-type. This is particularly pertinent to Chapter 3, where our observation that elevated breast PRS is associated with adverse DCIS outcomes, will be able to be validated in these upcoming studies. Important further work will include DCIS atlases incorporating inherited germline backgrounds, to provide a more comprehensive description and characterization of DCIS.

Lastly, although we applied improvements in somatic and germline profiling to breast precancer, these methodologies utilizing archival tissue can be applied to any precancer type where most samples are archived. The breast precancer atlas was in part generated in tandem with precancer profiling of prostate, lung, pancreas, and melanoma at other institutions as a pan-precancer effort coordinated by the National Cancer Institute (NCI) funded Molecular and Cellular Characterization of Screen-Detected Lesions Consortium [23]. Integration of data from other precancer types will help distinguish universal patterns of progression of premalignancy, versus

those which are tissue-specific. The field of early cancer detection is rapidly evolving with new approaches, including liquid biopsies for cell-free tumor DNA, which will likely soon enable the detection of precancer across an increased range of tissues. For improved detection to reduce cancer morbidity, understanding early carcinogenesis across all tissues will be critical for developing and tailoring chemopreventative interventions and will likely represent the next major forefront of cancer prevention.

References

1. Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R. C., Gambhir, S. S., Kuhn, P., Rebbeck, T. R. & Balasubramanian, S. Early detection of cancer. *Science* **375,** eaay9040 (2022).

2. Nachmanson, D., Steward, J., Yao, H., Officer, A., Jeong, E., O'Keefe, T. J., Hasteh, F., Jepsen, K., Hirst, G. L., Esserman, L. J., Borowsky, A. D. & Harismendy, O. Mutational profiling of micro-dissected pre-malignant lesions from archived specimens. *BMC Med. Genomics* **13,** 173 (2020).

3. Nachmanson, D., Officer, A., Mori, H., Gordon, J., Evans, M. F., Steward, J., Yao, H., O'Keefe, T., Hasteh, F., Stein, G. S., Jepsen, K., Weaver, D. L., Hirst, G. L., Sprague, B. L., Esserman, L. J., Borowsky, A. D., Stein, J. L. & Harismendy, O. The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. *NPJ Breast Cancer* **8,** 6 (2022).

4. Ellis, P., Moore, L., Sanders, M. A., Butler, T. M., Brunner, S. F., Lee-Six, H., Osborne, R., Farr, B., Coorens, T. H. H., Lawson, A. R. J. & Others. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16,** 841–871 (2021).

5. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H. & Campbell, P. J. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).

6. Martincorena, I., Fowler, J. C., Wabik, A., Lawson, A. R. J., Abascal, F., Hall, M. W. J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M. R., Fitzgerald, R. C., Handford, P. A., Campbell, P. J., Saeb-Parsy, K. & Jones, P. H. Somatic mutant clones colonize the human esophagus with age. *Science* **362,** 911–917 (2018).

7. Salk, J. J., Loubet-Senear, K., Maritschnegg, E., Valentine, C. C., Williams, L. N., Higgins, J. E., Horvat, R., Vanderstichele, A., Nachmanson, D., Baker, K. T., Emond, M. J., Loter, E., Tretiakova, M., Soussi, T., Loeb, L. A., Zeillinger, R., Speiser, P. & Risques, R. A. Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Rep.* **28,** 132–144.e3 (2019).

8. Danforth, D. N., Jr. Genomic Changes in Normal Breast Tissue in Women at Normal Risk or at High Risk for Breast Cancer. *Breast Cancer* **10,** 109–146 (2016).

9. Curtius, K., Wright, N. A. & Graham, T. A. An evolutionary perspective on field cancerization. *Nat. Rev. Cancer* **18,** 19–32 (2018).

10. Yost, S. E., Smith, E. N., Schwab, R. B., Bao, L., Jung, H., Wang, X., Voest, E., Pierce, J. P., Messer, K., Parker, B. A., Harismendy, O. & Frazer, K. A. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40,** e107 (2012).

11. Kim, H., Lee, A. J., Lee, J., Chun, H., Ju, Y. S. & Hong, D. FIREVAT: finding reliable variants without artifacts in human cancer samples using etiologically relevant mutational signatures. *Genome Med.* **11,** 81 (2019).

12. Guo, Q., Lakatos, E., Al Bakir, I., Curtius, K., Graham, T. A. & Mustonen, V. The mutational signatures of formalin fixation on the human genome. *bioRxiv* 2021.03.11.434918 (2021). doi:10.1101/2021.03.11.434918

13. Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P. & Maley, C. C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22,** 105–113 (2015).

14. Maley, C. C., Aktipis, A., Graham, T. A., Sottoriva, A., Boddy, A. M., Janiszewska, M., Silva, A. S., Gerlinger, M., Yuan, Y., Pienta, K. J., Anderson, K. S., Gatenby, R., Swanton, C., Posada, D., Wu, C.-I., Schiffman, J. D., Hwang, E. S., Polyak, K., Anderson, A. R. A., Brown, J. S., Greaves, M. & Shibata, D. Classifying the evolutionary and ecological features of neoplasms. *Nat. Rev. Cancer* **17,** 605–619 (2017).

15. Sinha, V. C. & Piwnica-Worms, H. Intratumoral Heterogeneity in Ductal Carcinoma In Situ: Chaos and Consequence. *J. Mammary Gland Biol. Neoplasia* **23,** 191–205 (2018).

16. Pareja, F., Brown, D. N., Lee, J. Y., Da Cruz Paula, A., Selenica, P., Bi, R., Geyer, F. C., Gazzo, A., da Silva, E. M., Vahdatinia, M., Stylianou, A. A., Ferrando, L., Wen, H. Y., Hicks, J. B., Weigelt, B. & Reis-Filho, J. S. Whole-Exome Sequencing Analysis of the Progression from Non-Low-Grade Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. *Clin. Cancer Res.* **26,** 3682–3693 (2020).

17. Srinivasan, P., Bandlamudi, C., Jonsson, P., Kemel, Y., Chavan, S. S., Richards, A. L., Penson, A. V., Bielski, C. M., Fong, C., Syed, A., Jayakumaran, G., Prasad, M., Hwee, J., Sumer, S. O., de Bruijn, I., Li, X., Gao, J., Schultz, N., Cambria, R., Galle, J., Mukherjee, S., Vijai, J., Cadoo, K. A., Carlo, M. I., Walsh, M. F., Mandelker, D., Ceyhan-Birsoy, O., Shia, J., Zehir, A., Ladanyi, M., Hyman, D. M., Zhang, L., Offit, K., Robson, M. E., Solit, D. B., Stadler, Z. K., Berger, M. F. & Taylor, B. S. The context-specific role of germline pathogenicity in tumorigenesis. *Nat. Genet.* **53,** 1577–1585 (2021).

18. Mavaddat, N., Pharoah, P. D. P., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., Luben, R., Brown, J., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Czene, K., Darabi, H., Eriksson, M., Peto, J., Dos-Santos-Silva, I., Dudbridge, F., Johnson, N., Schmidt, M. K., Broeks, A., Verhoef, S., Rutgers, E. J., Swerdlow, A., Ashworth, A., Orr, N., Schoemaker, M. J., Figueroa, J., Chanock, S. J., Brinton, L., Lissowska, J., Couch, F. J., Olson, J. E., Vachon, C., Pankratz, V. S., Lambrechts, D., Wildiers, H., Van Ongeval, C., van Limbergen, E., Kristensen, V., Grenaker Alnæs, G., Nord, S., Borresen-Dale, A.-L., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Fasching, P. A., Haeberle, L., Ekici, A. B., Beckmann, M. W., Burwinkel, B., Marme, F., Schneeweiss, A., Sohn, C., Trentham-Dietz, A., Newcomb, P., Titus, L., Egan, K. M., Hunter, D. J., Lindstrom, S., Tamimi, R. M., Kraft, P., Rahman, N., Turnbull, C., Renwick, A., Seal, S., Li, J., Liu, J., Humphreys, K., Benitez, J., Pilar Zamora, M., Arias Perez, J. I., Menéndez,

P., Jakubowska, A., Lubinski, J., Jaworska-Bieniek, K., Durda, K., Bogdanova, N. V., Antonenkova, N. N., Dörk, T., Anton-Culver, H., Neuhausen, S. L., Ziogas, A., Bernstein, L., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., van Asperen, C. J., Cox, A., Cross, S. S., Reed, M. W. R., Khusnutdinova, E., Bermisheva, M., Prokofyeva, D., Takhirova, Z., Meindl, A., Schmutzler, R. K., Sutter, C., Yang, R., Schürmann, P., Bremer, M., Christiansen, H., Park-Simon, T.-W., Hillemanns, P., Guénel, P., Truong, T., Menegaux, F., Sanchez, M., Radice, P., Peterlongo, P., Manoukian, S., Pensotti, V., Hopper, J. L., Tsimiklis, H., Apicella, C., Southey, M. C., Brauch, H., Brüning, T., Ko, Y.-D., Sigurdson, A. J., Doody, M. M., Hamann, U., Torres, D., Ulmer, H.-U., Försti, A., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Andrulis, I. L., Knight, J. A., Glendon, G., Marie Mulligan, A., Chenevix-Trench, G., Balleine, R., Giles, G. G., Milne, R. L., McLean, C., Lindblom, A., Margolin, S., Haiman, C. A., Henderson, B. E., Schumacher, F., Le Marchand, L., Eilber, U., Wang-Gohrke, S., Hooning, M. J., Hollestelle, A., van den Ouweland, A. M. W., Koppert, L. B., Carpenter, J., Clarke, C., Scott, R., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J. M., Brenner, H., Arndt, V., Stegmaier, C., Karina Dieffenbach, A., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Offit, K., Vijai, J., Robson, M., Rau-Murthy, R., Dwek, M., Swann, R., Annie Perkins, K., Goldberg, M. S., Labrèche, F., Dumont, M., Eccles, D. M., Tapper, W. J., Rafiq, S., John, E. M., Whittemore, A. S., Slager, S., Yannoukakos, D., Toland, A. E., Yao, S., Zheng, W., Halverson, S. L., González-Neira, A., Pita, G., Rosario Alonso, M., Álvarez, N., Herrero, D., Tessier, D. C., Vincent, D., Bacot, F., Luccarini, C., Baynes, C., Ahmed, S., Maranian, M., Healey, C. S., Simard, J., Hall, P., Easton, D. F. & Garcia-Closas, M. Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* **107,** (2015).

19. Karaayvaz-Yildirim, M., Silberman, R. E., Langenbucher, A., Saladi, S. V., Ross, K. N., Zarcaro, E., Desmond, A., Yildirim, M., Vivekanandan, V., Ravichandran, H., Mylavagnanam, R., Specht, M. C., Ramaswamy, S., Lawrence, M., Amon, A. & Ellisen, L. W. Aneuploidy and a deregulated DNA damage response suggest haploinsufficiency in breast tissues of BRCA2 mutation carriers. *Sci Adv* **6,** eaay2611 (2020).

20. Strand, S. H., Rivero-Gutiérrez, B., Houlahan, K. E., Seoane, J. A., King, L., Risom, T., Simpson, L. A., Vennam, S., Khan, A., Cisneros, L., Hardman, T., Harmon, B., Couch, F., Gallagher, K., Kilgore, M., Wei, S., DeMichele, A., King, T., McAuliffe, P. F., Nangia, J., Lee, J., Tseng, J., Storniolo, A. M., Thompson, A., Gupta, G., Burns, R., Veis, D. J., DeSchryver, K., Zhu, C., Matusiak, M., Wang, J., Zhu, S. X., Tappenden, J., Ding, D. Y., Zhang, D., Luo, J., Jiang, S., Varma, S., Anderson, L., Straub, C., Srivastava, S., Curtis, C., Tibshirani, R., Angelo, R. M., Hall, A., Owzar, K., Polyak, K., Maley, C., Marks, J. R., Colditz, G. A., Shelley Hwang, E. & West, R. B. DCIS genomic signatures define biology and correlate with clinical outcome: a Human Tumor Atlas Network (HTAN) analysis of TBCRC 038 and RAHBT cohorts. *bioRxiv* 2021.06.16.448585 (2021). doi:10.1101/2021.06.16.448585

21. Lips, E. H., Kumar, T., Megalios, A., Visser, L. L., Sheinman, M., Fortunato, A., Shah, V., Hoogstraat, M., Sei, E., Mallo, D. & Others. Genomic profiling defines variable clonal relatedness between invasive breast cancer and primary ductal carcinoma in situ. *medRxiv* (2021). at <https://www.medrxiv.org/content/10.1101/2021.03.22.21253209v1.abstract>

22. Gracia Villacampa, E., Larsson, L., Mirzazadeh, R., Kvastad, L., Andersson, A.,

Mollbrink, A., Kokaraki, G., Monteil, V., Schultz, N., Appelberg, K. S., Montserrat, N., Zhang, H., Penninger, J. M., Miesbach, W., Mirazimi, A., Carlson, J. & Lundeberg, J. Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genomics* **1,** 100065 (2021).

23. Institute, N. C. & National Cancer Institute. Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions. *Definitions* (2020). doi:10.32388/lp2r3b