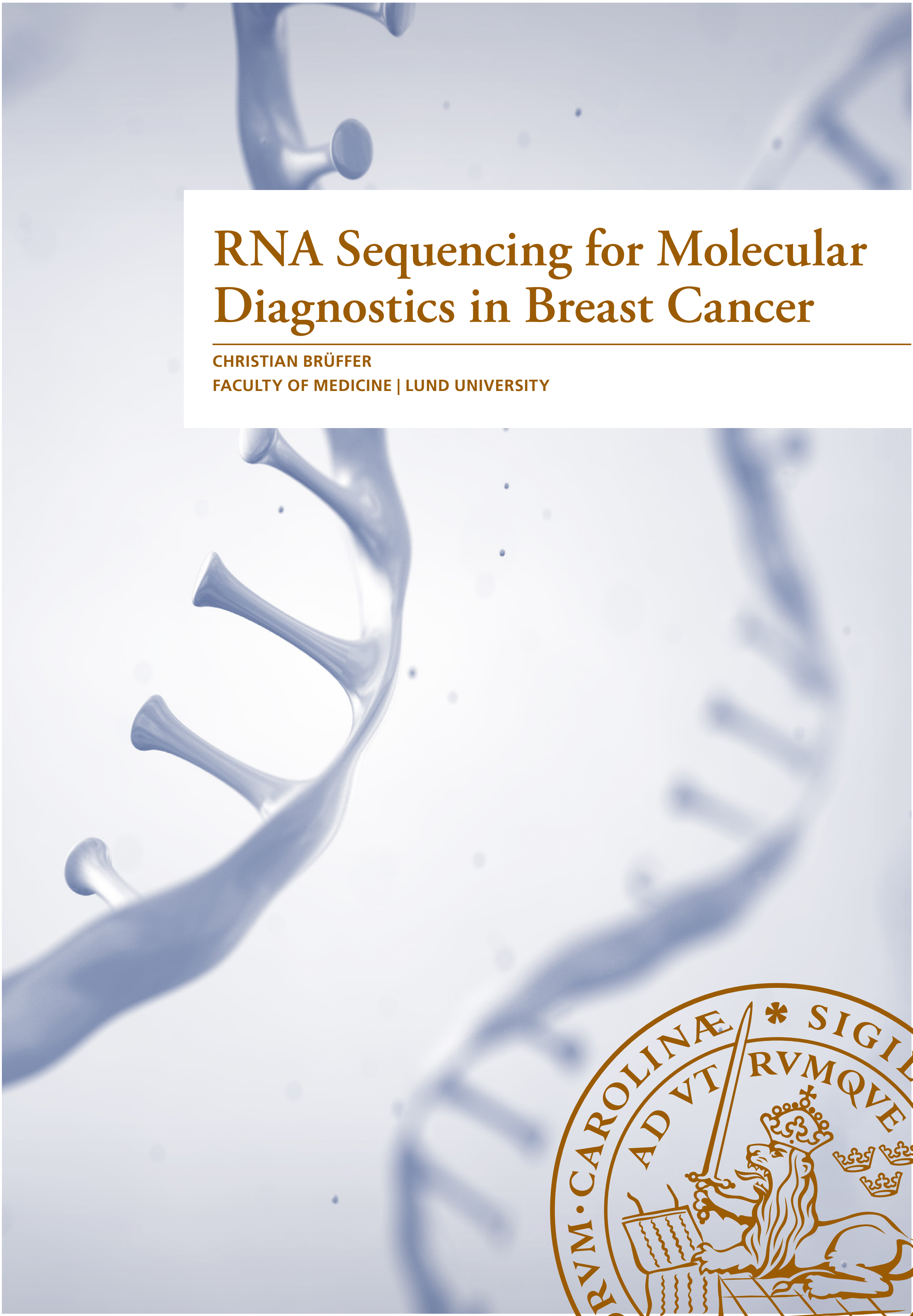
# RNA Sequencing for Molecular Diagnostics in Breast Cancer

Brueffer, Christian



# RNA Sequencing for Molecular Diagnostics in Breast Cancer

CHRISTIAN BRÜFFER
FACULTY OF MEDICINE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2021:2
ISBN 978-91-8021-008-9
ISSN 1652-8220

9 789180 210089

# RNA Sequencing for Molecular Diagnostics in Breast Cancer

# RNA Sequencing for Molecular Diagnostics in Breast Cancer

Christian Brüffer

LUND
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Medicine, Lund University, Sweden.

To be defended in Room E24, Medicon Village Building 404,
Lund on Wednesday the 13th of January 2021 at 13:00.

*Faculty opponent*

Dr. Aleix Prat, MD PhD

Hospital Clínic de Barcelona
Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)
University of Barcelona

Barcelona, Spain

**Organization**
LUND UNIVERSITY
Faculty of Medicine
Department of Clinical Sciences, Lund
Division of Oncology

**Document name** DOCTORAL DISSERTATION

**Date of issue** 2021-01-13

Author(s) Christian Brüffer

Sponsoring organization

**Title and subtitle** RNA Sequencing for Molecular Diagnostics in Breast Cancer

**Abstract**

Breast cancer is the most common type of cancer in women and, in Sweden, is the most deadly second only to lung cancer. While treatment and diagnostic options have improved in the past decades and short- to mid-term survival is good, long-term survival is much poorer. On the other hand, many women are likely cured by surgery and radiotherapy alone, but receive unnecessary adjuvant treatment leading to undesirable health-related and economic side-effects. Reliably differentiating high-risk from low-risk patients to provide optimal treatment remains a challenge.

The Sweden Cancerome Analysis Network–Breast (SCAN-B) project was initiated in 2009 and aims to improve breast cancer outcomes by developing new diagnostics and treatment-predictive tests. Within SCAN-B, tumor material and blood are being biobanked and the transcriptomes of many thousands of breast tumors are being analyzed using RNA sequencing (RNA-seq). The resulting sample collection and dataset provide an unprecedented resource for research, and the information therein may harbor ways to improve prognosis and to predict tumor susceptibility or resistance to therapies.

In the four original studies included in this thesis we explored the use of RNA-seq as a diagnostic tool within breast cancer. In study I we described the SCAN-B processes and protocols, and analyzed early data to show the feasibility of using RNA-seq as a diagnostic platform. We showed that the patient population enrolled in SCAN-B largely reflects the characteristics of the total breast cancer patient population and benchmarked RNA-seq against prior techniques. In study II we diagnosed problems in commonly used RNA-seq alignment software and described the development of a software tool to correct the problems and improve data usability. Study III focused on diagnostics for determining the status of the important breast cancer biomarkers ER, PgR, HER2, Ki67, and Nottingham histological grade. We assessed the reproducibility of histopathology in measuring these biomarkers, and developed new ways of predicting their status using RNA-seq-based gene expression. We showed that expression-based biomarkers add value to histopathology by improving prognostic possibilities. In study IV we focused on the prospects of using RNA-seq to detect mutations. We developed a new computational method to profile mutations and used it to describe the mutational landscape of thousands of patient tumors and its impact on patient survival. In particular, we identified mutations in a subset of patients that are known to confer resistance to standard treatments.

The hope is that, together, the diagnostic results made possible by the studies herein may one day enable oncologists to adapt treatment plans accordingly and improve patient quality of life and outcomes.

**Key words** breast cancer, RNA-seq, diagnostics, precision medicine, biomarker, gene expression, mutation, SCAN-B

Classification system and/or index terms (if any)

Supplementary bibliographical information

**Language** English

**ISSN** and key title 1652-8220
Lund University, Faculty of Medicine Doctoral Dissertation Series 2021:2

**ISBN** 978-91-8021-008-9

Recipient’s notes

**Number of pages** 110

Price

Security classification

Signature

Date 2020-12-03

# RNA Sequencing for Molecular Diagnostics in Breast Cancer

Christian Brüffer

LUND
UNIVERSITY

Department of Clinical Sciences, Lund
Faculty of Medicine, Lund University

The LaTeX sources and plot scripts for this thesis are publicly available to enable re-use:
LaTeX sources: `https://github.com/cbrueffer/phd_thesis`
Plot scripts: `https://github.com/cbrueffer/phd_thesis_plots`
SCAN-B map: `https://github.com/cbrueffer/scanb_map`

Lund University, Faculty of Medicine Doctoral Dissertation Series 2021:2
ISBN: 978-91-8021-008-9
ISSN: 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

*A tremendous feeling of peace came over him. He knew that at last, for once and for ever,*
*it was now all, finally, over.*

— Douglas Adams, The Hitchhiker's Guide to the Galaxy

# Contents

## II Original Studies

# List of Original Studies

This thesis is based on the following original studies, which are referred to in the text by their Roman numerals:

I **The Sweden Cancerome Analysis Network-Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine**

Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, **Brueffer C**, Tang MHE, Reuterswärd C, Schulz R, Karlsson A, Ehinger A, Malina J, Manjer J, Malmberg M, Larsson C, Rydén L, Loman N, Borg Å

*Genome Medicine, 2015. 7(1):20*

II **TopHat-Recondition: A post-processor for TopHat unmapped reads**

**Brueffer C** and Saal LH

*BMC Bioinformatics, 2016. 17(1):199*

III **Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network–Breast Initiative**

**Brueffer C***, Vallon-Christersson J*, Grabau D, Ehinger A, Häkkinen J, Hegardt C, Malina J, Chen Y, Bendahl PO, Manjer J, Malmberg M, Larsson C, Loman N, Rydén L, Borg Å, Saal LH

*JCO Precision Oncology, 2018. 2:1–18*

IV **The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome**

**Brueffer C**, Gladchuk S, Winter C, Vallon-Christersson J, Hegardt C, Häkkinen J, George AM, Chen Y, Ehinger A, Larsson C, Loman N, Malmberg M, Rydén L, Borg Å, Saal LH

*EMBO Molecular Medicine, 2020. 12(10):e12118*

*Authors contributed equally to this work.

# Author Contributions

My contributions to the studies included in this thesis were as follows:

I **The Sweden Cancerome Analysis Network-Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine**

I contributed the software described in study II as well as input to the development of the SCAN-B computational pipeline, performed subtyping, compared microarray and RNA-seq based expression and intrinsic subtypes, contributed to data analysis, deposited the data in the NCBI Gene Expression Omnibus (GEO), and contributed to writing the manuscript.

II **TopHat-Recondition: A post-processor for TopHat unmapped reads**

I diagnosed the problems in TopHat/TopHat2, designed and developed the software TopHat-Recondition, and drafted and revised the manuscript.

III **Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network–Breast Initiative**

I evaluated different machine learning approaches on training data, trained and evaluated the final classifiers, performed classification and survival analysis in the 3,273 patient validation cohort, deposited the data in NCBI GEO, and drafted and revised the manuscript.

IV **The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome**

I participated in study design, implemented the DNA/RNA mutation calling pipeline, performed the mutation calling, and co-supervised a masters student who worked on variant filtering. I performed all downstream analysis of the mutations, performed the survival analysis, developed the SCAN-B MutationExplorer web application, and drafted and revised the manuscript.

# Additional Publications and Preprints

- **precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map regions**

  Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, Jáspez D, Lorenzo-Salazar JM, Muñoz-Barrera A, Rubio-Rodríguez LA, Flores C, Kyriakidis K, Malousi A, Shafin K, Pesout T, Jain M, Paten B, Chang PC, Kolesnikov A, Nattestad M, Baid G, Goel S, Yang H, Carroll A, Eveleigh R, Bourgey M, Bourque G, Li G, MA C, Tang L, DU Y, Zhang S, Morata J, Tonda R, Parra G, Trotta JR, **Brueffer C**, *et al.*

  *bioRxiv, 2020 (preprint)*

- **Features of increased malignancy in eosinophilic clear cell renal cell carcinoma**

  Nilsson H, Lindgren D, Axelson H, **Brueffer C**, Saal LH, Lundgren J, Johansson ME.

  *The Journal of Pathology, 2020. 252(4):384–397*

- **A crowdsourced set of curated structural variants for the human genome**

  Chapman LM, Spies N, Pai P, Lim CS, Carroll A, Narzisi G, Watson C, Proukakis C, Clarke W, Nariai N, Dawson E, Jones G, Blankenberg D, **Brueffer C**, Xiao C, Kolora SRR, Alexander N, Wolujewicz P, Ahmed A, Smith G, Shehreen S, Wenger AM, Salit M, Zook J.

  *PLoS Computational Biology, 2020. 16(6):e1007933*

- **Detection of circulating tumor cells and circulating tumor DNA before and after mammographic breast compression in a cohort of breast cancer patients scheduled for neoadjuvant treatment**

  Förnvik D, Aaltonen KE, Chen Y, George AM, **Brueffer C**, Rigo R, Loman N, Saal LH, Rydén L.

  *Breast Cancer Research and Treatment, 2019. 177(2):447–445*

- **Bioconda: sustainable and comprehensive software distribution for the life sciences**

  Grüning B*, Dale R*, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Caprez A, Batut B, Haudgaard M, Cokelaer T, Beauchamp KA, Pedersen BS, Hoogstrate Y, Ryan D, Bretaudeau A, Le Corguillé G, **Brueffer C** *et al.*

  *Nature Methods, 2018. 15(7):475–476*

*Authors contributed equally to this work.

- **Contralateral breast cancer can represent a metastatic spread of the first primary tumor: determination of clonal relationship between contralateral breast cancers using next-generation whole genome sequencing**

  Alkner S*, Tang MHE*, **Brueffer C**, Dahlgren M, Chen Y, Olsson E, Winter C, Baker S, Ehinger A, Rydén L, Saal LH, Fernö M, Gruvberger-Saal SK.

  *Breast Cancer Research, 2015. 17:102*

- **Remarkable similarities of chromosomal rearrangements between primary human breast cancers and matched distant metastases as revealed by whole-genome sequencing**

  Tang MHE*, Dahlgren M*, **Brueffer C**, Tjitrowirjo T, Winter C, Chen Y, Olsson E, Wang K, Törngren T, Sjöström M, Grabau D, Bendahl PO, Rydén L, Niméus E, Saal LH, Borg Å, Gruvberger-Saal SK.

  *Oncotarget, 2015. 6(35):37169–37184*

*Authors contributed equally to this work.

# Abstract

Breast cancer is the most common type of cancer in women and, in Sweden, is the most deadly second only to lung cancer. While treatment and diagnostic options have improved in the past decades and short- to mid-term survival is good, long-term survival is much poorer. On the other hand, many women are likely cured by surgery and radiotherapy alone, but receive unnecessary adjuvant treatment leading to undesirable health-related and economic side-effects. Reliably differentiating high-risk from low-risk patients to provide optimal treatment remains a challenge.

The Sweden Cancerome Analysis Network–Breast (SCAN-B) project was initiated in 2009 and aims to improve breast cancer outcomes by developing new diagnostics and treatment-predictive tests. Within SCAN-B, tumor material and blood are being biobanked and the transcriptomes of many thousands of breast tumors are being analyzed using RNA sequencing (RNA-seq). The resulting sample collection and dataset provide an unprecedented resource for research, and the information therein may harbor ways to improve prognosis and to predict tumor susceptibility or resistance to therapies.

In the four original studies included in this thesis we explored the use of RNA-seq as a diagnostic tool within breast cancer. In study i we described the SCAN-B processes and protocols, and analyzed early data to show the feasibility of using RNA-seq as a diagnostic platform. We showed that the patient population enrolled in SCAN-B largely reflects the characteristics of the total breast cancer patient population and benchmarked RNA-seq against prior techniques. In study ii we diagnosed problems in commonly used RNA-seq alignment software and described the development of a software tool to correct the problems and improve data usability. Study iii focused on diagnostics for determining the status of the important breast cancer biomarkers ER, PgR, HER2, Ki67, and Nottingham histological grade. We assessed the reproducibility of histopathology in measuring these biomarkers, and developed new ways of predicting their status using RNA-seq-based gene expression. We showed that expression-based biomarkers add value to histopathology by improving prognostic possibilities. In study iv we focused on the prospects of using RNA-seq to detect mutations. We developed a new computational method to profile mutations and used it to describe the mutational landscape of thousands of patient tumors and its impact on patient survival. In particular, we identified mutations in a subset of patients that are known to confer resistance to standard treatments.

The hope is that, together, the diagnostic results made possible by the studies herein may one day enable oncologists to adapt treatment plans accordingly and improve patient quality of life and outcomes.

# Popular summary

Breast cancer is the most common type of cancer in women and, in Sweden, is the most deadly second only to lung cancer. In the western world, approximately 1 in 8 women will be diagnosed with breast cancer in their lifetime, largely fueled by lifestyle and dietary choices. Like all cancers, breast cancer is caused by alterations in the genome of normal cells that lead them to grow uncontrollably. Diagnostic and treatment options have expanded in the past decades, with the introduction of endocrine and anti-HER2 therapies. While this has lead to good short-term to mid-term survival of patients, long-term survival is a lot poorer. On the other hand, many women are likely cured by surgery and radiotherapy alone, but are being "overtreated", leading to unnecessary health-related and economic side-effects. Reliably differentiating patients at high risk of disease relapse from those with low risk remains a major challenge.

The first sequencing of a human genome in 2001 has set in motion an unprecedented amount of knowledge generation and technology development in biology and medicine. Through the advent of high-throughput sequencing technologies that transform the genetic material of DNA and RNA into large datasets, biology and medicine are becoming increasingly reliant on the field of bioinformatics which provides the computational knowledge to analyze these datasets. The resulting insights have allowed us to better understand widespread and complex diseases such as cancer. Our increased understanding holds the promise for a future where precision medicine is reality, and a patient receives treatments that target the specific weaknesses of their tumor. However, translating the improved understanding of tumors into meaningful clinical interventions remains a challenge and requires the analysis of large, well characterized patient cohorts.

The Sweden Cancerome Analysis Network–Breast (SCAN-B) project was initiated in 2009 and aims to improve breast cancer outcomes by developing new diagnostics and treatment-predictive tests. Within the nine participating SCAN-B hospitals the biological material from many thousands of breast cancer patients is being collected and analyzed using RNA sequencing (RNA-seq). This technique probes the cancer transcriptome, the complete picture of all genes turned on and off in a tumor, and enables the precise measurement of gene activity (expression) and gene alterations (mutations) in patient tumors. This information, when trained on patient samples with treatment and outcome information, can then be used to predict a new patient's prognosis and may signal susceptibility or resistance to specific therapies – which is the goal of precision medicine.

In the four original studies included in this thesis we explored the use of RNA-seq as a diagnostic tool within breast cancer. In study I we described the SCAN-B processes and protocols, and analyzed early data to show the feasibility of using RNA-seq as a diagnostic platform. We showed that the patient population enrolled in SCAN-B largely reflects the characteristics of the total breast cancer patient population and benchmarked RNA-seq against previous techniques. In study II we diagnosed problems in commonly used RNA-seq analysis software and described the development of a software tool to correct these problems. Study III focused on diagnostics for determining the status of important breast cancer biomarkers. We assessed the reproducibility of the currently used methods to measure these biomarkers, and developed new ways of predicting their status using gene expression as determined using RNA-seq. We showed that these gene expression-based biomarkers add value to the currently used techniques by improving prognostic possibilities. In study IV we focused on the prospects of using RNA-seq to determine gene mutations. We developed a new computational method to profile mutations and used it to describe the mutational landscape of thousands of patient tumors and its impact on patient survival. In particular we were able to identify mutations in a subset of patients that are known to confer resistance to standard treatments. Providing this information to the clinic may enable oncologists to adapt treatment plans accordingly.

The diagnostic tools described in this thesis are being evaluated, improved, and validated further, and will hopefully benefit patients in SCAN-B-participating hospitals in the future.

# Populärwissenschaftliche Zusammenfassung

Brustkrebs ist die häufigste Krebsart bei Frauen und in Schweden nach Lungenkrebs die Krebsart mit den meisten Todesfällen. Bedingt durch den Lebenswandel und Ernährungsgewohnheiten erkrankt in der westlichen Welt etwa jede achte Frau in ihrem Leben an Brustkrebs. Wie alle Krebsarten wird Brustkrebs durch Veränderungen im Genom von normalen Körperzellen hervorgerufen, die dazu führen, dass sich die Zellen unkontrolliert vermehren. Behandlungs- und Diagnostikmethoden haben sich in den letzten Jahrzehnten verbessert, vor allem durch die Einführung von Hormon- und Anti-HER2-Therapien. Während dies zu guten kurz- bis mittelfristigen Überlebenschancen geführt hat, sind die langfristigen Überlebenschancen deutlich geringer. Andererseits werden viele Frauen mit hoher Wahrscheinlichkeit bereits durch die operative Entfernung des Tumors mit anschließender Bestrahlung geheilt. Diese werden dann allerdings "übertherapiert", was zu unerwünschten gesundheitlichen und finanziellen Nebenwirkungen führt. Die verlässliche Unterscheidung von Patientinnen und Patienten mit einem hohen Risiko der Rückerkrankung von solchen mit einem niedrigen Risiko ist immer noch eine große Herausforderung.

Die erstmalige Sequenzierung eines menschlichen Genoms im Jahr 2001 hat eine beispiellose Wissens- und Technologieentwicklung in den Bereichen Biologie und Medizin in Gang gesetzt. Durch die Einführung von Hochdurchsatz-Sequenzierungstechnologien, die die biologischen Materialien DNA und RNA in große Datenmengen umsetzen, sind Biologie und Medizin zunehmend auf das Feld der Bioinformatik angewiesen, das die nötigen Kenntnisse bereitstellt, um diese Datenmengen rechnergestützt zu analysieren. Die dadurch entstehenden Erkenntnisse haben es uns erlaubt, weit verbreitete und komplexe Krankheiten wie Krebs besser zu verstehen. Dieses verbesserte Verständnis bringt die Möglichkeit der Präzisionsmedizin näher, bei der ein Patient eine Behandlung bekommt, die maßgeschneidert die Schwächen des jeweiligen Tumors ausnutzt. Das erweiterte Wissen in wirksame Interventionen umzusetzen ist jedoch eine Herausforderung und erfordert die Verfügbarkeit und Analyse von großen und gut charakterisierten Patientenkohorten.

Das Sweden Cancerome Analysis Network–Breast (SCAN-B) Projekt wurde im Jahr 2009 in Schweden ins Leben gerufen und zielt darauf ab, die Überlebenschancen von Brustkrebspatienten durch die Entwicklung von neuen Diagnostik- und Therapieerfolg-Vorhersagemöglichkeiten zu verbessern. In den neun teilnehmenden Kliniken wird das biologische Material von tausenden Brustkrebspatienten gesammelt und mittels RNA-Sequenzierung (RNA-seq) analysiert. Diese Methode untersucht das Transkriptom von Krebszellen, also die Gesamtheit der Boten-RNA (mRNA) eines Tumors, die anzeigt, welche Gene ein- und ausgeschaltet sind. Dies ermöglicht die präzise Messung der Genaktivität (Expression) und von Genveränderungen (Mutationen) in Tumoren. Zusammen mit Überlebensdaten der

Patienten können diese Informationen dann dazu genutzt werden, Modelle zu entwickeln (zu "trainieren"), die präzisere Prognosen für zukünftige Patienten liefern, und vorhersagen könnten, ob ein Tumor anfällig für, oder resistent gegen bestimmte Therapien ist – das letztendliche Ziel der Präzisionsmedizin.

In den vier Studien, die im Zuge dieser Doktorarbeit durchgeführt wurden und hier diskutiert werden, wollten wir die Möglichkeiten der RNA-seq als Mittel für die Brustkrebsdiagnostik erforschen. In Studie I haben wir die Prozesse und Protokolle des SCAN-B Projektes beschrieben und erste in SCAN-B generierte Daten analysiert, um die Möglichkeiten der RNA-seq als diagnostisches Mittel aufzuzeigen. Wir konnten außerdem zeigen, dass die Patientenpopulation in SCAN-B größtenteils die Eigenschaften aller Brustkrebspatienten im Studiengebiet widerspiegelt, und haben die RNA-seq mit vorherigen Methoden zur Transkriptomanalyse verglichen. In Studie II haben wir Probleme in häufig genutzter Software zur Analyse von RNA-seq-Daten aufgezeigt, und die Entwicklung eines Softwarewerkzeugs beschrieben, das diese Probleme behebt. In Studie III haben wir uns auf die Bestimmung wichtiger Brustkrebsbiomarker fokussiert. Wir haben die Reproduzierbarkeit der momentan genutzten Labormethoden evaluiert und neue Methoden entwickelt, um den Wert dieser Biomarker mittels Genexpression zu bestimmen. Wir konnten zeigen, dass diese genexpressions-basierten Biomarker den momentan genutzten Methoden wertvolle Zusatzinformationen hinzufügen die die Prognosemöglichkeiten dieser Methoden verbessern. In Studie IV haben wir die Möglichkeiten eruiert, Genmutationen auf der Basis von RNA-seq zu bestimmen. Dazu haben wir eine rechnergestützte Methode zur Mutationsbestimmung entwickelt. Diese haben wir angewandt, um die Gesamtheit der Mutationen in den Tumoren tausender Patienten zu beschreiben und deren Einfluss auf die Überlebenschancen der Patienten zu analysieren. Insbesondere konnten wir in einigen Tumoren Mutationen entdecken, von denen bekannt ist, dass sie Resistenz gegen Standardtherapien verleihen. Diese Informationen könnten es den behandelnden Onkologen in Zukunft erlauben, Therapiepläne frühzeitig entsprechend anzupassen.

Die in dieser Doktorarbeit beschriebenen diagnostischen Möglichkeiten werden gegenwärtig weiter ausgewertet, verbessert und validiert. In Zukunft werden sie hoffentlich allen Patienten zugutekommen, die in SCAN-B Kliniken behandelt werden.

# Abbreviations

| | |
|---|---|
| ABiM | All Breast Cancers in Malmö study |
| AIMS | Absolute Intrinsic Molecular Subtypes |
| AJCC | American Joint Committee on Cancer |
| ASR | Age-standardized incidence rate |
| bp | base pair |
| BAC | bacterial artificial chromosome |
| BAM | Binary alignment/map file format |
| BCS | breast-conserving surgery |
| ctDNA | Circulating tumor DNA |
| CNV | Copy-number variant |
| CTC | Circularing tumor cell |
| DCIS | Ductal carcinoma *in situ* |
| dNTP | deoxyribonucleotide triphosphate; A, T, G, or C |
| ER | Estrogen receptor |
| FDA | Food and Drug Administration |
| ESMO | European Society for Medical Oncology |
| FPKM | Fragments per kilobase of exon per million mapped reads |
| GEO | Gene expression omnibus |
| HER2 | Human epidermal growth factor receptor 2 |
| HoR | Hormone receptor (ER and/or PgR) |
| HR | Hazard ratio |
| HTS | High-throughput sequencing, also called next-generation sequencing, deep sequencing, or massively parallel sequencing |
| indel | Short insertion or deletion |
| IDC | Invasive ductal carcinoma |
| IHC | Immunohistochemistry |
| ILC | Invasive lobular carcinoma |
| KM | Kaplan-Meier |
| LoH | Loss of Heterozygosity |
| mRNA | messenger RNA |
| MAF | Mutant allele frequency |

| | |
|---|---|
| Mb | Megabase |
| MRD | Minimal residual disease |
| NCBI | National Center for Biotechnology Information |
| NHG | Nottingham histological grade |
| NMD | Nonsense-mediated decay |
| NMF | Non-negative matrix factorization |
| OS | Overall survival |
| PAM | Prediction Analysis of Microarrays |
| PAM50 | Prediction Analysis of Microarrays 50 gene signature |
| PARP | Poly (ADP-ribose) polymerase |
| PCR | Polymerase chain-reaction |
| PgR | Progesterone receptor |
| RNA-seq | Illumina short-read cDNA sequencing |
| RPKM | Reads per kilobase of exon per million mapped reads |
| SCAN-B | Sweden Cancerome Analysis Network–Breast |
| SERD | Selective estrogen receptor degrader |
| SNP | Single nucleotide polymorphism |
| SNV | Single nucleotide variant |
| SSP | Single sample predictor |
| SV | Structural variant |
| TCGA | The Cancer Genome Atlas |
| TKI | Tyrosine kinase inhibitor |
| TMB | Tumor mutational burden |
| TNBC | Triple-negative breast cancer |
| TNM | TNM (tumor, node, metastasis) staging system |
| TPM | Transcripts per million reads |
| TRK | Tyrosine receptor kinase |
| UICC | Union for International Cancer Control |
| VAF | Variant allele frequency |
| WES | Whole exome sequencing |
| WGS | Whole genome sequencing |

# List of Figures

# List of Tables

# Part I

# Research Context

# 1 Introduction

*Everything starts somewhere, although many physicists disagree.*

— Terry Pratchett, Hogfather

## 1.1 Cancer

Cancer is a disease that has long plagued humans, animals [1] – including dinosaurs [2] – and, to a certain extent, even plants [3, 4]. Evidence of tumors has been found in Neanderthals [5], while the earliest records of tumors in humans come from ancient Egypt, both via evidence from mummies/skeletons [6–8] and descriptions of various tumor types in the Edwin Smith Papyrus – an ancient medical text. The abundance of evidence for tumors across domains of life and human civilizations suggests that cancer is an unavoidable consequence of evolution [9]. However, the risk for developing cancer is modulated by factors such as lifestyle and increasing life expectancy across the globe (see Section 1.4.1).

Historically, cancer has been attributed to many different causes [10]. For example, the ancient Greeks thought it was a product of the four "humors" (black bile, yellow bile, phlegm, and blood) becoming unbalanced. Theodor Boveri in 1902 was the first to suggest cancer developing from mitotic origins affecting the chromosomes [11]. While our understanding of cancer biology steadily increased since then, for example through landmark discoveries such as the genes *BRCA1* [12, 13] and *BRCA2* [14, 15] and their relation to breast cancer susceptibility, the release of the first human genome draft sequence in 2001 [16, 17] has marked a turning point in our understanding of cancer and its underpinnings.

Generally, cancers can be differentiated into carcinomas (solid tumors of epithelial origin), sarcomas (solid tumors originating in supportive and connective tissue), myelomas (originating in plasma cells of the bone marrow), leukemias (originating in the bone marrow), lymphomas (originating in the lymphatic system), and mixed types [18]. All cancers share certain traits, summarized by Hanahan and Weinberg as a list of disease-defining hallmarks of cancer in 2000 [19], and in an updated form in 2011 [20]. The hallmarks are summarized in Figure 1.1 and describe the ways tumors overcome the inherent cellular control mechanisms, grow their own blood vessels, escape the host immune system, and achieve invasion. The genomic changes leading to these hallmarks can either be activating, for example causing an activation of cell growth and differentiation, or deactivating, for example inhibiting mechanisms involved in cellular regulation and damage repair. Activating mutations affect oncogenes such as *MYC* and *PIK3CA* that have the potential to induce tumor growth, while deactivating mutations affect tumor suppressor genes such as *TP53*

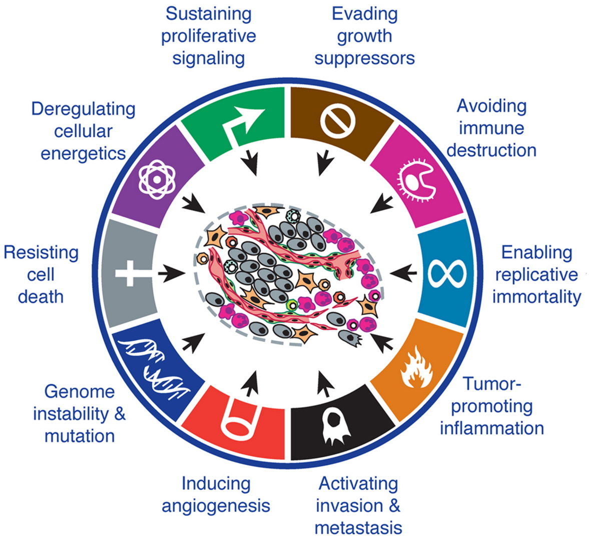

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

**Figure 1.1.** The hallmarks of cancer.

Source: Hanahan & Weinberg [20]. Reproduced with permission from Elsevier.

and *PTEN* that act as moderating breaks on cellular processes.

## 1.2 The Cancer Genome

Cancer arises from genomic mutations that can occur years to decades before diagnosis [21], or may even be inherited and present at birth. Mutations can arise spontaneously, for example due to errors during mitosis, tautomeric base pairing [22, 23], or through outside damaging influence such as carcinogens. These mutations can then accumulate, for example through DNA proofreading mistakes caused by defective DNA polymerases resulting from previously acquired mutations [24].

Mutations in cancer are generally divided into driver mutations that actively promote tumor growth and are therefore positively selected for, and passenger mutations that happen as byproducts due to the unstable nature of the tumor genome, for example due to impaired DNA repair mechanisms [25]. These mutations and the genes harboring them are being catalogued by the IntOGen project and others [26–29]. The general model is that few mutations are drivers and the majority of mutations are passengers, although this simplistic view is being challenged [30].

The emergence of sensitive detection methods has allowed us to better understand tumorigenesis by investigating somatic mutations in normal tissues [31, 32]. Studies in normal cells from skin [33, 34], endometrium [35], esophagus [36], colon [37], bladder [38], breast [39], and urethra [40] tissue have shown a variety of somatic mutations and positive selection for them [33, 36, 37]. *TP53* mutations in particular have been found to be clonally selected over the course of a human lifetime [41]. In general, somatic mutations accumulate with age in normal tissues [42], but even the presence of driver mutations does not necessarily lead to carcinogenesis [43].

The different types of mutations that characterize the cancer genome, as well as the grouping of these mutations into signatures and mutational burden are detailed in the following sections.

### 1.2.1 Single Nucleotide Variants

Single nucleotide variants are the most common type of mutation in cancer. The possible nucleotide substitutions can be reduced to the six substitution types `C>A`, `C>G`, `C>T`, `T>A`, `T>C`, and `T>G`. Transitions (`C>T` and `T>C`) are generally more common than transversions (`C>A`, `C>G`, `T>A`, and `T>G`), since substitutions between purines (`A` and `G`) and between pyrimidines (`C` and `T`) are sterically more likely than those between purines and pyrimidines. Depending on whether or not SNVs lie in a region of the genome coding for protein sequence, they are classed as coding or non-coding (Figure 1.2). Coding SNVs are further stratified into synonymous and non-synonymous variants depending on whether or not they change the amino acid sequence of a protein. Comprehensive classifications, such as the Sequence Ontology controlled vocabulary [44], further stratify non-coding, synonymous, and non-synonymous variants into multiple subclasses based on their predicted impact. Simplified versions are commonly being used for classification, such as the one we used in study IV to classify non-synonymous variants into missense variants (for those mutations that lead to a different amino acid being incorporated into the protein sequence) and nonsense variants (for mutations that induce/remove start or stop codons). Non-coding and synonymous SNVs are not stratified further. The mutation classes differ in the severity of their functional impact, where nonsense mutations that lead to a premature stop codon and loss of the downstream protein are most severe. In cancer these mutations often affect tumor suppressor genes such as *TP53*.

While non-synonymous mutations have long been in the spotlight of research, non-coding and synonymous variants have been understudied. However, increasing evidence suggests that both have measurable impact on oncogenesis. Non-coding variants have been found to act as drivers across cancer types [45]. Synonymous mutations, which have been thought to be silent, may play an important role both in the normal genome [46] and in cancer [47, 48]. While not directly altering protein amino acid sequences, they can affect splicing and expression regulation and may exert a driving effect in this way.

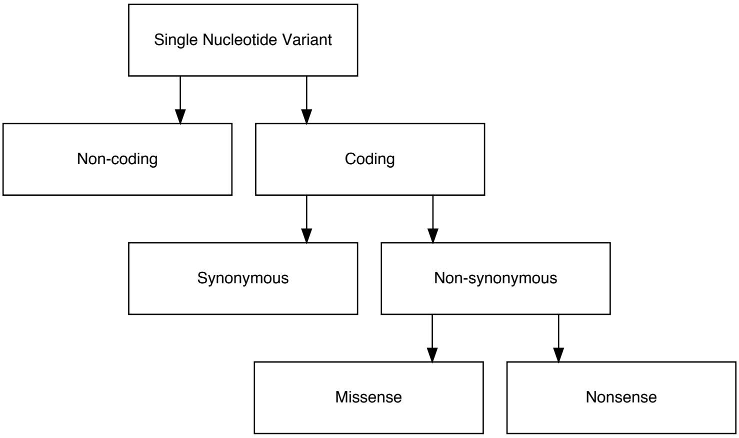


**Figure 1.2.** Classification of single-nucleotide variants (SNVs).

### 1.2.2 Short Insertions and Deletions

Short insertions and deletions (indels) are small, ≤50bp, genomic alterations. If the number of inserted or deleted bases is divisible by three (the length of a codon) the indel is in-frame, otherwise it is classified as frame-shift since it changes the reading frame. Frame-shift indels are common cancer mutations, particularly in tumor suppressor genes, where they disrupt transcription by inducing premature stop codons. By comparison, in-frame indels are generally less disruptive but still lead to protein alterations that may affect normal function.

### 1.2.3 Structural Variants, Copy Number Variants, and Gene Fusions

Structural variants (SV) are genomic changes that rearrange the sequence of one or two chromosomes and have a size of >50bp [49]. Rearrangements can occur within one chromosome (intra-chromosomal rearrangements) or between two chromosomes (intra-chromosomal rearrangements). Unbalanced SVs affect copy-number relative to the reference genome, meaning gain or loss of genetic loci, and are referred to as copy-number variants (CNVs). CNVs can be insertions, deletions, or duplications [50]. These are common in cancer, where they can lead to overexpression of oncogenes such as *ERBB2* due to increased gene dosage which then drives tumor-growth. Compared to CNVs, simple inversions and translocations are copy-number neutral, although translocations are often complex and associated to copy number changes.

Gene fusions are consequences of SVs, where one or both break ends of an SV lie in a genic region, resulting in a new in-frame gene configuration. Gene fusions are common in many cancers and can be important driver mutations. The best known example is the *BCR-ABL1* fusion gene resulting from a translocation between chromosomes 22 and 9, that is common in chronic myelogenous leukemia (CML) and acute lymphobastic leukemia (ALL).

In contrast to SNVs that develop continuously during the lifetime of a tumor, many SVs largely occur early in tumor development during the "telomere crisis" [51, 52]. Individual SVs can be part of complex structural events such as chromothripsis, which describes a single catastrophic chromosomal shattering event followed by incorrect DNA repair [53]. Since then, other recurring complex events have been described, each having their own signature of structural events [54–56]. Due to their early occurrence, SVs are ideal biomarkers as many tumor clones will share them. This can be exploited in early detection of disease recurrence [57].

### 1.2.4 Epigenetics

Epigenetic changes are those that do not involve alteration of the DNA nucleotide sequence and play a major role in tumor development [58]. Several types of epigenetic alterations exist, including promoter hyper- and hypomethylation and histone modifications. Promoter hypermethylation has a major influence on transcription dynamics through its ability to silence genes, while hypomethylation has the opposite effect and can lead to increased transcription. Examples in cancer are *BRCA1* and *PTEN* hypermethylation, where transcriptional silencing leads to loss of protein expression, contributing to oncogenesis. Histone modifications are addition or loss of functional groups from histone proteins, performed by certain enzymes. Histones are a principal determinant of chromatin openness and transcription, and alteration of modifications can adversely affect transcription of genes wound around an affected histone.

### 1.2.5 Mutational Signatures

The mutational processes that shape the tumor genome often generate tell-tale "signatures" of mutation type combinations in the genome. Alexandrov *et al* [59] first employed non-negative matrix factorization (NMF) to describe a variety of signatures covering SNVs, their immediate neighbor bases ("sequence context"), and indels across 30 cancer types. They could associate 11 signatures with specific causes, such as overactivity of members of the APOBEC family of cytidine deaminases [60], or exposure to ultraviolet light. Since then, the original signatures have been refined and dozens of other signatures, including those derived from SVs and CNVs, have been described [61–63]. Importantly, mutational signatures caused by environmental mutagens [64] and chemotherapies [65] have been catalogued and may shed further light on these factors.

### 1.2.6 Tumor Mutational Burden

Tumor mutational burden (TMB) is a measure for the overall number of mutations in a tumor, typically normalized by megabase (Mb) of sequence. It has been proposed as a biomarker that may be useful for indicating sensitivity to immunotherapies [66]. For as-yet incompletely understood reasons, these therapies show heterogeneous response and currently no biomarker is available to reliably predict treatment outcome. TMB is believed to be a surrogate for neoepitope formation, where body-foreign immunogenic peptides are expressed by the tumor. TMB is not without controversy, as many questions around it remain unsolved. They start with how to define TMB, since the number of detected tumor mutations is a function of sequencing experiment setup. Whole genome sequencing (WGS) or whole exome sequencing (WES) will uncover more mutations than a panel targeting few genes, not even considering RNA sequencing (RNA-seq) based TMB which we investigated in study IV. Another factor is sequencing depth, where sequencing deeper will result in more mutations than sequencing shallow. TMB also varies by tumor site and subtype [67], possibly necessitating different TMB cutoffs to stratify tumors into TMB-low and TMB-high. Efforts to harmonize TMB determination in certain settings and to account for some of these questions are ongoing [68].

In 2020 the U.S. Food and Drug Administration (FDA) granted approval for pembrolizumab in TMB-high solid tumors, where the TMB cutoff was defined as ≥10 mut/Mb. This is the first FDA drug approval that allows TMB as a biomarker and, given the questions around TMB, this decision was highly controversial with voices both for [69] and against [70]. Adding to the controversy, a reanalysis of public clinical study datasets suggests that TMB is in fact not a good marker of response to immune checkpoint blockage [71], but that the supposed signal was a statistical artifact. It has been proposed that it may not the overall mutational burden, but only indels that trigger mRNA nonsense mediated decay that signal response to immunotherapy [72, 73].

## 1.3 The Cancer Transcriptome

While the genome provides cellular blueprints, the transcriptome represents the dynamic state of the cell. Compared to the genome, the transcriptome is underexplored, perhaps partly due to its inherent complexity. It encompasses the entirety of cellular transcripts (RNAs), the most important and basic element of which is messenger RNA (mRNA). Through transcription from a single gene precursor mRNA is produced, which, through alternative splicing and alternative polyadenylation [74, 75], may be processed into a variety of mature mRNA isoforms. Adding to this, a variety of non-coding RNAs exist, such as transfer RNA (tRNA), microRNA (miRNA), Piwi-interacting RNA (piRNA), vault RNA (vtRNA), and others. These do not code for proteins, but may have functional interactions with each other, with DNA, with mRNA, or with proteins, leading to a complex and dynamic interaction network that is difficult to grasp. Another level of complexity is added by the epitranscriptome, a collection of more than 170 types of RNA editing and modifications, such as deamination of adenosine to inosine (A-to-I editing), methylation of adenosine to $N^6$-methyladenosine ($m^6A$ modification), or pseudouridine ($\psi$), that can modulate gene expression levels, protein translation, and localization [76–79]. Lastly, cellular processes, such as nonsense-mediated decay, impact gene expression levels. This may happen by removing mRNAs that contain premature stop codons, for example induced by transcription errors or small DNA indels.

Compared to the normal transcriptome, the cancer transcriptome is dysregulated due to changes in transcriptomic processes that alter the delicate and complex balance of the transcriptome. Indeed, all known transcriptomic features and processes have been implicated in tumor development when dysregulated, such as gene expression [80, 81], alternative splicing [82–84] and intron retention [85], and alternative polyadenylation [86], non-coding RNAs [45, 87, 88], RNA editing and modifications [89–91], and transcriptomic pathways [72, 92].

The properties of the transcriptome as mediator between DNA and proteome make it an interesting target for diagnostics. It contains information currently diagnostically exploited on the DNA level, provides a wealth of information that can only be probed on the transcriptome level, and through mRNA expression and modifications has direct impact on the proteome.

## 1.4 Breast Cancer

Breast cancer is the most common form of cancer in women. It mostly originates in the duct tissue (~80%, ductal carcinoma) and lobules (~20%, lobular carcinoma) of the breast [93], depicted in Figure 1.3. It is inherently heterogeneous, with multiple subtypes that have distinct genetic, phenotypic, and clinical presentations that translate into differing prognosis, risk profiles, and susceptibility to treatments. Although the disease can occur in both women and men, approximately 99% of patients are female [94]. While there are many commonalities in the disease between women and men, considerable differences exist in terms of genetics and clinical characteristics [95–97]. This thesis focuses exclusively on breast cancer in women, and the term “breast cancer” in this thesis will from here on only refer to the disease affecting women. Compared to other cancer types, considerable progress has been made in breast cancer diagnosis, treatment, and subsequent patient survival in the last four decades [98].

### 1.4.1 Incidence and Mortality

Breast cancer is the most common kind of cancer worldwide accounting for nearly 2.1 million newly diagnosed cases and nearly 630,000 deaths in 2018 [99]. This is 11.6% of all new cancer cases and 24.2% of cases in women.

There are substantial regional differences in global breast cancer incidence, visualized us-

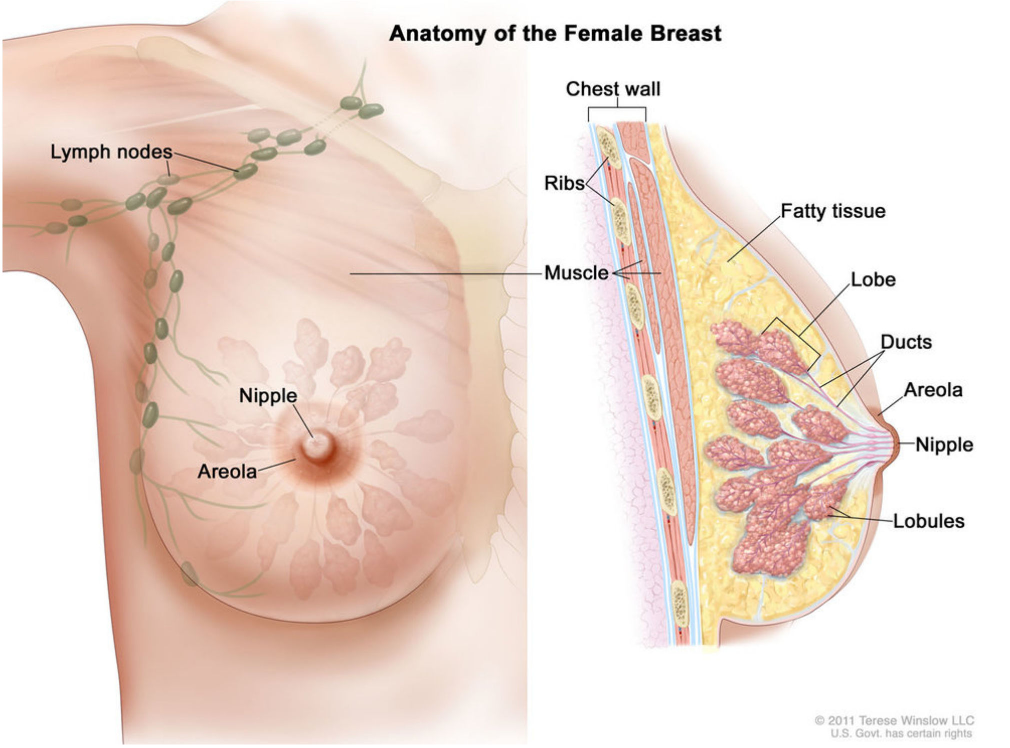


**Figure 1.3.** Anatomy of the female breast. Highlighted are the lymph nodes, nipple, areola, muscles, chest wall, ribs, fatty tissue, as well as lobules and ducts.

For the National Cancer Institute © 2011 Terese Winslow LLC, U.S. Government has certain rights. Reproduced with permission from the copyright holder.

ing data from the World Health Organization for 2018 in Figure 1.4. Incidence is age-standardized to account for the varying age structure between populations. Western societies have the highest incidence, largely influenced by lifestyle and dietary choices.

In 2018, 30,511 women in Sweden were diagnosed with cancer, of which 7,558 women were diagnosed with breast cancer and 1,391 women succumbed to the disease. This makes breast cancer the second most deadly type of cancer in Sweden behind lung cancer [100]. Despite the large number of total deaths, patient survival is generally very good in the short (98% 1-year survival) to mid-term (88.5% 5-year survival) compared to other types of cancer. However, 5-year survival cannot be considered a cure, and survival rates significantly decline in the long and very-long term (60% 15-year survival, 50% 20-year survival), as patients experience recurrence of their disease [101, 102].

Breast cancer is the most common type of cancer in women and, in Sweden, is the most deadly second only to lung cancer.

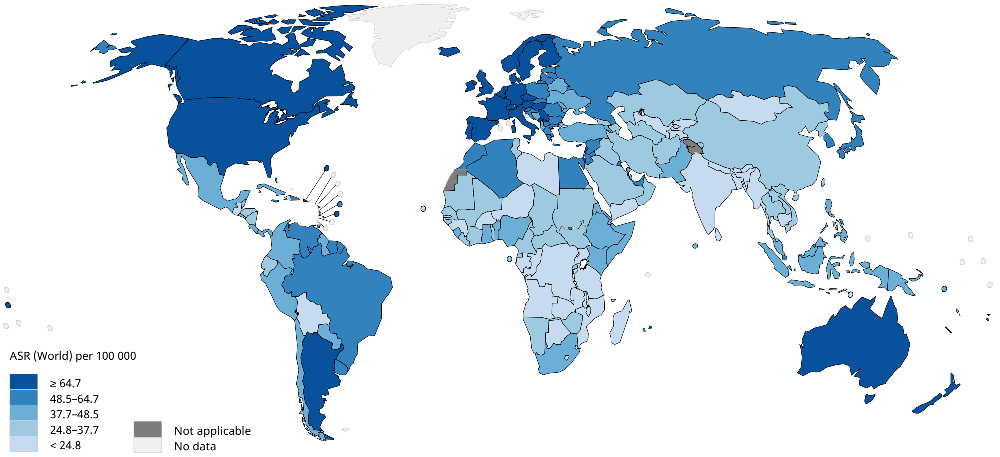


**Figure 1.4.** Global estimated age-standardized incidence rate (ASR) for breast cancer per 100,000 women for the year 2018.

Source: World Health Organization Global Cancer Observatory (`https://gco.iarc.fr`)

### 1.4.2 Risk Factors

A diverse range of factors have been identified that increase women's life-time risk of developing breast cancer. Age is the most important risk factor, as mutations accumulate in normal cells over time. In 2018 in Sweden, only 4% of invasive breast cancers were diagnosed in women under the age of 40 [103]. Breast cancer risk, particularly in post-menopausal women, is modulated by factors that alter endogenous sex hormone levels. High baseline hormone levels, oral contraceptives, early menarche, late menopause, and most hormonal replacement therapies during menopause increase breast cancer risk [104–106]. Additionally, reproductive aspects such as parity, age at first childbirth, the number of children, and breast feeding have complex effects on breast cancer risk [107].

A variety of dietary and lifestyle factors have been found to increase breast cancer risk: consumption of alcohol [108, 109] and processed meat [110, 111], as well as active and passive exposure to tobacco smoke [112]. Obesity and high body fat content, both measured as body-mass index (BMI) and in a BMI-independent way [113–115], as well as lack of exercise [116, 117] are associated with higher risk. Lastly, exposure to environmental factors such as ionizing radiation, including X-radiation and gamma radiation, elevates risk.

A particularly important risk factor is a family history of cancer as approximately 5%–10% of breast cancers are hereditary. The mechanism of action is thought to be Knudson's two-hit hypothesis [118], whereby patients have inherited a damaged copy of a risk gene from their parents (first hit), and the second copy is damaged during the person's lifetime leading to loss of heterozygosity (LoH), for example by exposure to environmental carcinogens (second hit). Approximately 25% of all hereditary cases can be explained by high-risk variants in the *BRCA1* and *BRCA2* genes [119]. Rare germline mutations in other high-penetrance genes cause specific forms of breast cancer, the most prominent being *PTEN* hamartoma tumor syndrome caused by *PTEN* variants, and Li-Fraumeni syndrome caused by *TP53* variants. The remaining cases can be partly attributed to variants in medium to low risk genes including *CHEK2*, *PALB2*, *RAD50*, *ATM*, and *BARD1*. However, a proportion of cases cannot be explained by the risk genes known to date. In Sweden, breast cancer risk variants are being explored through initiatives such as the SWEA study, and efforts to identify unknown risk variant carriers through studies such as BRCAsearch [120]. In all hereditary cases genetic counseling is imperative to guide possible prophylactic measures such as mastectomy and/or oophorectomy, and to determine whether the patient's relatives may carry the risk alleles.

### 1.4.3 Diagnosis

Breast cancer is most often detected either through early detection techniques such as mammographic screening, or self-examination of the breasts by the patient. While mammographic screening has led to early detection of many breast cancers [121], it is not without controversy as it can also lead to overdiagnosis [122]. It is predicted that a significant number of detected lesions may never become invasive during the patient's lifetime, however we currently lack the tools to detect which ones. On the other hand, current screening methods can miss lesions, for example due to lobular phenotype of the lesion [123], or due to high breast density [124].

To guide treatment decisions, tumor biopsy and surgery samples are evaluated using histopathological and/or genomic methods and classified by their morphological, clinicopathological, and genomic features. The most important classification schemes are described in the following sections.

### 1.4.4 Classification

Several systems exist to class tumors into prognostic and treatment-predictive subgroups. These include systems based on histopathology such as Nottingham histological grade (NHG) and TNM stage, and molecular methods based on gene expression signatures.

#### Histopathology

Between 15% and 30% of breast tumors are *in situ* carcinomas; that is, the tumor cell growths have not broken through the basement membrane layer. These are often detected using screening programs and consequently the exact percentage of *in situ* tumors depends on the prevalence of screening in the population. Based on the site of origin one can differentiate ductal carcinoma *in situ* (DCIS, ~80%) and lobular carcinoma *in situ*

**Table 1.1.** Nottingham histological grade scoring and interpretation.

| Score | Grade | Interpretation |
|---|---|---|
| 3–5 | 1 | well differentiated |
| 6–7 | 2 | moderately differentiated |
| 8–9 | 3 | poorly differentiated |

Source: Elston & Ellis [127]

(LCIS, ~20%) [125]. Most *in situ* carcinomas are benign, but some harbor malignant potential and may or may not become invasive if left untreated. One of the major challenges is improving diagnostics to enable this distinction.

Invasive carcinomas constitute between 70% and 85% of all breast cancers. The majority of these are invasive ductal carcinomas (IDC, ~79%) of not otherwise specified (NOS) type, followed by invasive lobular carcinomas (ILC, ~10%). The remaining cases can be further stratified based on cytological features into tubular (~2%), medullary (~5%), mucinous (~2%), papillary (1%-2%), and cribriform (0.8%-3.5%) cancer [126].

### Grade

Nottingham histological grade according to the Elston and Ellis modified Scarff-Bloom-Richardson system (NHG) is a morphological marker that describes how closely tumor cells resemble normal breast epithelial cells [127]. Generally with increasing grade, resemblance to normal cells decreases and tumor aggressiveness is thought to increase. NHG is a compound score consisting of the three morphologic components tubular differentiation, number of mitoses, and nuclear pleomorphism. The component-scores are determined individually for a tumor, added together, and categorized according to Table 1.1. NHG is a strong prognostic factor in breast cancer [128], however it has long had reproducibility problems [129] which we also observed in study III.

### Stage

Pathologic stage describes how advanced a cancer is. The TNM system is the most widely used staging system in breast cancer. It was originally proposed by Denoix in 1946 [130] and today is maintained by the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC). The TNM system classifies cancer by the size of the tumor (T), the number of lymph nodes containing tumor cells (N), and metastatic spread (M). Each of these categories has subcategories, such as T1–T4 for increasing tumor size, that describe the extent of disease progression. In the simplest use, the stage group is then determined using only the T, N, and M subcategories according to Table 1.2. Stage grouping can be made more fine grained by incorporating additional in-

**Table 1.2.** Pathologic stage as defined by the 8th edition of the AJCC TNM system description using only the mandatory parameters T, N, and M.

| Stage | TNM Categories | Interpretation |
|---|---|---|
| 0 | Tis N0 M0 | pre-invasive stage |
| I | T1 N0 M0 | low stage |
| | T0 N1mi M0 | |
| | T1 N1mi M0 | |
| II | T0 N1 M0 | intermediate stage |
| | T1 N1 M0 | |
| | T2 N0 M0 | |
| | T2 N1 M0 | |
| | T3 N0 M0 | |
| III | T0 N2 M0 | high stage |
| | T1 N2 M0 | |
| | T2 N2 M0 | |
| | T3 N1 M0 | |
| | T3 N2 M0 | |
| | T4 N0 M0 | |
| | T4 N1 M0 | |
| | T4 N2 M0 | |
| | Any T N3 M0 | |
| IV | Any T Any N M1 | metastatic stage |

Source: AJCC Cancer Staging Manual 8th Ed. [131]

formation such as prefix modifiers describing the information source and may be modified by NHG, histological receptor status, and the score of the Oncotype DX genomic assay (see Section 1.4.4).

### Receptor Status

The expression status of the receptor proteins estrogen receptor (ER), progesterone receptor (PR or PgR), and human epidermal growth factor receptor 2 (HER2) is routinely determined using immunohistochemistry (IHC) for breast tumors and is of prime importance for prognosis and treatment (see Section 1.4.5). Tumor slides are stained for these receptors using antibodies. Stained cells are counted or estimated versus non-stained cells, resulting in a stained cell percentage. Receptor status is dichotomized into positive/negative status based on a cutoff. In Sweden for ER/PgR, a cutoff of 10% stained cells is used, while internationally a cutoff of 1% is common. For HER2, an additional *ERBB2* gene copy-number analysis using fluorescence or silver *in situ* hybridization (FISH or SISH) is recommended if the HER2 IHC result is inconclusive. Recently, a new subgroup of HER2-low has been proposed to mark tumors with low HER2 protein expression and no *ERBB2* gene amplification that would traditionally be called HER2- [132]. Increasing evidence suggests that a subset of these tumors may benefit from HER2 targeting agents.

By combining ER, PgR, and HER2 status, tumors can be categorized into clinical subgroups, whereby ER and PgR may be summarized into hormone receptor (HoR[1]) status. Patients with HoR+ tumors have a better survival rate than those with HoR- tumors [133]. This includes the HoR+/HER2- group, which constitutes the largest subgroup with 68% of cases in the U.S. between 2013 and 2017 [134], and generally has the best prognosis [135] followed by HoR+/HER2+ tumors (~10%). Compared to these HoR+ groups, survival of patients with HoR-/HER2+ (~4%) is significantly worse. Triple-negative breast cancer (TNBC, ~10%) lacks expression of all three receptors, and thus offers no molecular targets for the most common targeted agents. Consequently it has the worst prognosis, with chemotherapy being the only treatment option.

### Intrinsic Subtypes

In addition to classing tumors by morphology and histology, they can be stratified by their intrinsic subtype. These define distinct groups of tumors with similar gene expression patterns and clinical characteristics. Molecular subtypes were first discovered by Perou and Sørlie *et al* [80] who performed unsupervised hierarchical clustering on the global gene expression profiles of normal tissues and breast tumor tissues from 42 patients. The subtypes were quickly found to be prognostic [136]. The originally reported subtypes Luminal-like, Basal-like, HER2-enriched, and Normal-like were later refined by differentiating the Luminal-like group into Luminal A-like and Luminal B-like tumors [137]. More recently the Claudin-low subtype has been defined [138], although its status as a true intrinsic subtype has been disputed [139]. The subtypes have been reproduced numerous times across technology platforms [140, 141] and in metastatic tumors [142–145]. They also exhibit distinct methylation patterns [146].

The Luminal- and Basal-like subtypes were originally named due to the similarity of their gene expression patterns to normal luminal and basal epithelial cells. In the Luminal-like case this is a gene expression signature reflecting estrogen receptor activation. The Luminal A-like subtype is characterized by a normal HER2 expression profile and low activity of proliferation genes, while Luminal B-like tumors show elevated proliferation and can have *ERBB2* overexpression. The Basal-like subtype is characterized by a gene expression signature including activation of basal keratins, integrin-$\beta 4$, and laminin. The HER2-enriched subtype features a signature of *ERBB2* overactivation [80]. Samples of Normal-like subtype typically cluster together with true normal breast tissue samples. The existence of Normal-like as a true intrinsic subtype has been questioned as it is possibly a

[1] A more common abbreviation for hormone receptor is HR, however this abbreviation is also commonly used for the Hazard Ratio. We therefore opted for abbreviating hormone receptor as HoR in studies III, IV, and in this thesis.

technical artifact caused by samples with low tumor cell content [147–149]. It is therefore sometimes omitted from analysis.

Since expression profiling remains a non-standard diagnostic tool, surrogate intrinsic subtypes can be derived from traditional clinicopathological biomarkers in combination with Ki67 protein status as a surrogate marker for proliferation, and NHG using the St. Gallen classification schema [150]. NHG may also be useful in refining the classification, particularly for differentiating between Luminal A-like and B-like tumors [151, 152]. However, concordance between expression-based subtypes and surrogate subtypes is generally poor [153–155], and thus the surrogate classification remains an imperfect stopgap solution until expression profiling is integrated into the clinical routine.

In addition to aiding our understanding of breast cancer biology, the introduction of the St. Gallen surrogate subtypes is a testament to the importance and potential clinical impact of the intrinsic subtypes. In particular the intrinsic subtypes are useful in refining the traditional clinicopathological grouping by receptor status, where the groups HoR+/HER2-, HoR+/HER2+, HoR-/HER2+, and HoR-/HER2- show heterogeneous compositions of molecular subtypes [148] with prognostic and treatment-predictive implications [154, 156–159].

### Gene Expression Signatures

Gene expression signatures provide a dimension to breast cancer classification beyond traditional clinicopathological biomarkers. Based on the expression of a defined number of genes, they capture the transient state of a tumor and are used to define phenotypes such as the intrinsic subtypes and biomarker status, and to predict risk. While a plethora of multi-gene signatures have been developed in the research setting to date, these signatures have shown little gene overlap [160]. Wirapati *et al* performed an early meta-analysis of nine expression signatures across 2,833 tumors and found concordance in terms of signature gene function [161]. Their findings were later reproduced and extended by Huang *et al* [162]. More recently, within the SCAN-B study (see Section 1.9), 19 gene signatures for subtyping and risk prediction were benchmarked across a large population-based tumor series [163] and found to provide additional prognostic value over traditional clinicopathological classifications in ER+/HER2- disease. However, signatures did not provide further risk stratification in the patient subgroups with ER-/HER2+ and TNBC disease that have particularly bad prognosis.

The clinical implications of risk prediction signatures have been reviewed several times [164–166], highlighting in particular those signatures that have been commercialized and/or validated in large patient cohorts. The most widely used signatures are the 21 gene signature, commercialized as the Oncotype DX assay (Genomic Health) [167–169], the 70 gene signature commercialized as MammaPrint (Agendia) [170], the PAM50 Risk of Recurrence (RoR) score (which excludes the Normal-like subtype), commercialized as the

Prosigna Breast Cancer Prognostic Gene Signature Assay (NanoString Technologies) [148, 171], and EndoPredict (Myriad Genomics). They share approval as risk prediction signatures for early breast cancers that are at risk of developing distant metastases, and thus may be utilized to decide upon adjuvant therapy. Several clinical trials are in progress to validate this potential, including MINDACT (ClinicalTrials.gov identifier NCT00433589), TAILORx (ClinicalTrials.gov identifier NCT00310180), and RxPONDER (ClinicalTrials.gov identifier NCT01272037)) [172–174]. Early results indicate that gene expression profiling tests can indeed identify low risk patients that may be spared unnecessary treatment [175–177].

Other signatures try to reproduce standard histopathological biomarkers such as receptor status [178–184] and NHG [185–187]. Genomic grade signatures classify tumors into low or high grade, thus clarifying the intermediate NHG class grade 2. Commercial variants of this concept are MapQuant DX (Ipsogen) and Breast Cancer Index (Biotheranostics).

Most commercial gene expression signatures presented here are FDA-approved and were explicitly endorsed by the 2017 St. Gallen conference consensus panel as tools for guiding treatment with adjuvant chemotherapy in node-negative tumors, including MammaPrint, PAM50 RoR, EndoPredict, and Breast Cancer Index.

In Sweden, these tests are not widely used as they are expensive and the cost-benefit ratio has not yet been fully established. Instead, traditional clinicopathological variables and surrogate subtypes are being used for prognostication and definition of treatment regimens. However, increasingly guidelines now do recommend the use of a gene expression risk stratification test, and use of such tests is anticipated to increase dramatically in Sweden in the near future.

### 1.4.5 Treatment

The goal of primary breast cancer treatment is to remove all remnants of the tumor and prevent it from relapsing. The strategy currently recommended by the European Society for Medical Oncology (ESMO) is outlined in Figure 1.5. Primary treatment is, in virtually all cases, surgical removal of the tumor, if possible using breast-conserving surgery (BCS). To prevent relapse in the BCS case, additional radiotherapy is crucial to eradicate possible leftover tumor deposits. For tumors that are too large for BCS, neoadjuvant therapy may be attempted to shrink the tumor to a size where BCS is feasible; otherwise, mastectomy is performed. To support primary treatment and reduce the risk of recurrence, adjuvant treatment is recommended.

#### Neoadjuvant Treatment

Neoadjuvant therapies are administered before the primary treatment. They may be used to shrink a tumor down to a size that makes it feasible to perform surgery at all, or less invasive

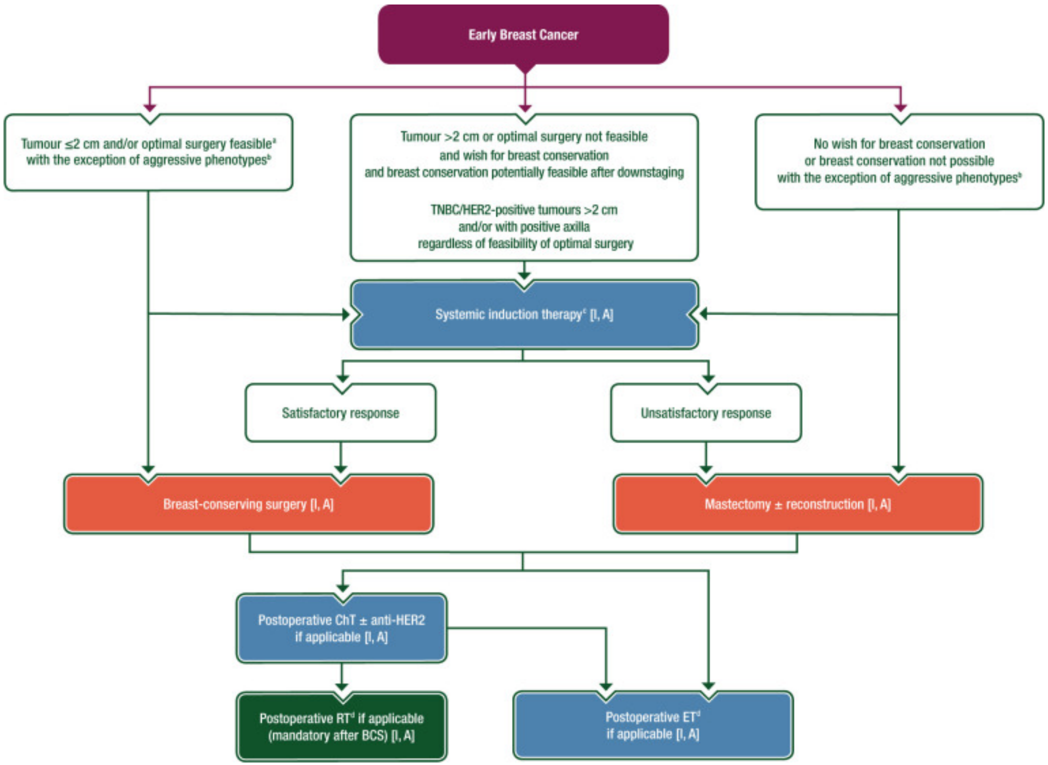

Early Breast Cancer
Tumour ≤2 cm and/or optimal surgery feasible^a with the exception of aggressive phenotypes^b
Tumour >2 cm or optimal surgery not feasible and wish for breast conservation and breast conservation potentially feasible after downstaging
TNBC/HER2-positive tumours >2 cm and/or with positive axilla regardless of feasibility of optimal surgery
No wish for breast conservation or breast conservation not possible with the exception of aggressive phenotypes^b
Systemic induction therapy^c [I, A]
Satisfactory response
Unsatisfactory response
Breast-conserving surgery [I, A]
Mastectomy ± reconstruction [I, A]
Postoperative ChT ± anti-HER2 if applicable [I, A]
Postoperative RT^d if applicable (mandatory after BCS) [I, A]
Postoperative ET^d if applicable [I, A]

**Figure 1.5.** Algorithm for primary treatment of early breast cancer by the European Society for Medical Oncology (ESMO).

**a** Biology that requires ChT (TNBC, HER2-positive, luminal B-like), to assess response and prognosis and eventually decide on postoperative therapies, should preferentially receive preoperative ChT.

**b** Aggressive phenotypes: TNBC or HER2-positive breast cancer.

**c** If ChT is planned, it should all be given as neoadjuvant.

**d** Concomitant postoperative RT, postoperative ET and anti-HER2 therapy

Abbreviations: BCS – breast-conserving surgery; ChT – chemotherapy; ET – endocrine therapy; HER2 – epidermal growth factor receptor 2; RT – radiotherapy; TNBC – triple-negative breast cancer.

Figure and descriptions a–d reprinted from Cardoso *et al* [188] with permission from Elsevier.

surgery. The neoadjuvant period also provides a window of opportunity for testing how the tumor reacts to therapies before it is surgically removed, potentially proving guidance for adjuvant treatment. This concept is being utilized in clinical trials such as the I-SPY study [189] (ClinicalTrials.gov identifier NCT01042379).

### Adjuvant Treatment

Adjuvant treatments are administered after and in support of primary treatment and are guided by biomarkers and clinicopathological features. The current ESMO adjuvant treatment recommendations are outlined in Figure 1.6 [188]. Tumors determined to be HoR+ are treated with targeted anti-hormonal agents such as tamoxifen, or aromatase inhibitors that block estrogen synthesis, such as letrozole. Those showing overexpression of the HER2 receptor protein receive targeted anti-HER2 treatment, for example the monoclonal antibody trastuzumab. All patients with the exception of those with ER+/HER2- (Luminal A-like) disease receive additional chemotherapy, typically anthracycline and taxane. For triple-negative tumors chemotherapy is currently the only first-line treatment option.

### Advanced Disease

Despite primary and adjuvant treatment, tumors can advance or recur even after many years of a patient being disease-free. Often these tumors have developed resistance to standard treatments used in early disease, so different therapies are needed. Recent years have seen several treatment innovations in this area, mostly in combination with other therapies.

In hormone therapy resistant ER+/HER2- breast cancer, CDK4/6 inhibitors are showing promise, for example when combined with selective estrogen receptor degraders (SERDs) such as fulvestrant [190]. In hospitals participating in the SCAN-B study (see section 1.9), this type of drug has been used since spring 2017.

Several targeted drugs are being approved or showing strong promise in tumors with specific gene mutations. Approximately one-third of breast tumors have mutations in the *PIK3CA* oncogene. While targeting this gene effectively has long been difficult, recent advances have led to the approval of the PI3K inhibitor alpelisib (Novartis) [191] in *PIK3CA*-mutant HoR+/HER2- disease in combination with fulvestrant [192]. Metastatic tumors often develop therapy resistance to standard treatments such as anti-HER2 drugs prescribed in early disease. Tyrosine kinase inhibitors (TKIs) such as the recently FDA-approved tucatinib can overcome this resistance. Additionally, Poly (ADP-ribose) polymerase (PARP) inhibitors have shown promise in DNA-repair-deficient tumors such as those with *BRCA1* or *BRCA2* mutations.

Another emerging class of drug are antibody-drug conjugates, where traditional antibodies such as trastuzumab are conjugated with a cytotoxic agent. This combination leads to a more targeted release of the traditionally systemically-working cytotoxin by guidance to

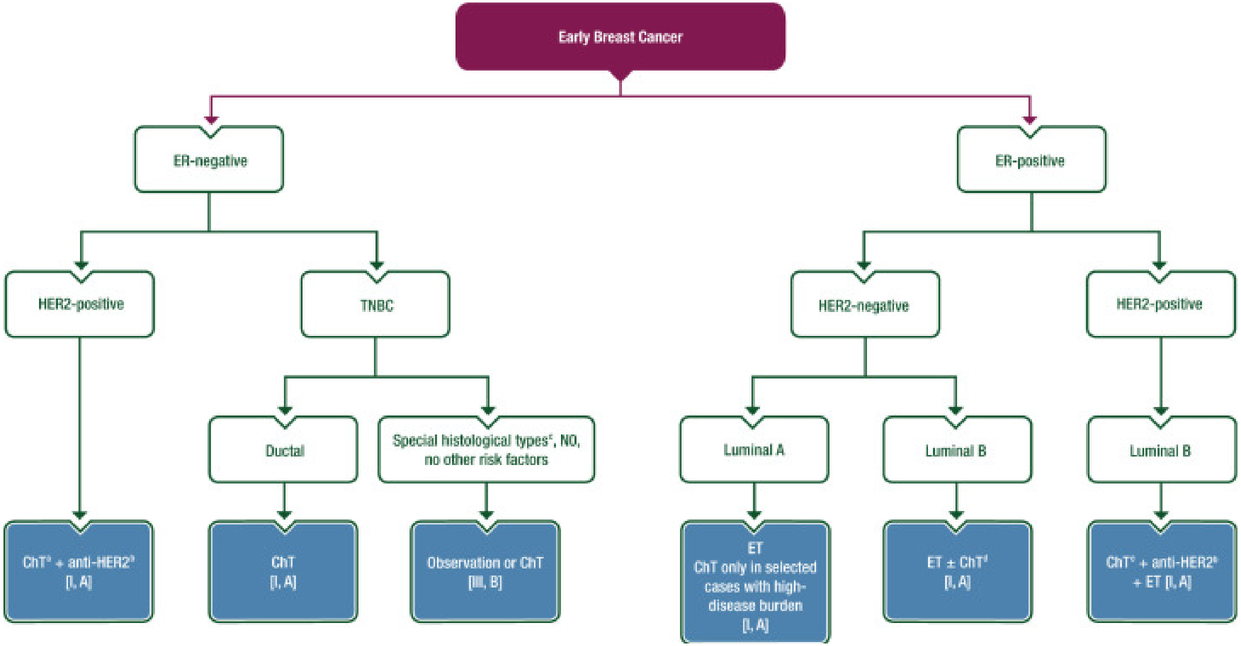

Early Breast Cancer
ER-negative
ER-positive
HER2-positive
TNBC
HER2-negative
HER2-positive
Ductal
Special histological types[c], N0, no other risk factors
Luminal A
Luminal B
Luminal B
ChT[a] + anti-HER2[b] [I, A]
ChT [I, A]
Observation or ChT [III, B]
ET ChT only in selected cases with high-disease burden [I, A]
ET ± ChT[d] [I, A]
ChT[e] + anti-HER2[b] + ET [I, A]

**Figure 1.6.** Algorithm for adjuvant treatment of early breast cancer by the European Society for Medical Oncology (ESMO).
**a** With possible exception of selected cases with very low risk T1abN0.
**b** Anti-HER2: trastuzumab ± pertuzumab.
**c** Adenoid cystic or apocrine, secretory carcinoma, low-grade metaplastic carcinoma.
**d** Depending on level of ER and PgR expression, proliferation, genomically assessed risk, tumor burden and/or patient preference.
**e** Except for very low-risk patients T1abN0 for whom ET/anti-HER2 therapy alone can be considered.
Abbreviations: ChT – chemotherapy; ER – estrogen receptor; ET – endocrine therapy; HER2 – epidermal growth factor receptor 2; N0 – node negative; PgR – progesterone receptor; TNBC – triple-negative breast cancer.

Figure and descriptions a–e reprinted from Cardoso *et al* [188] with permission from Elsevier.

tumor cells via the antibody. Examples of these drugs include trastuzumab emtansine and trastuzumab deruztecan approved for treatment of advanced HER2+ tumors, and sacituzumab govitecan in advanced TNBC.

Lastly, immunotherapies have shown promise in many types of difficult to treat tumors. In breast cancer, PD-1/PD-L1 inhibitors such as atezolizumab and pembrolizumab are being used to treat metastatic tumors and triple-negative tumors [193]. Beyond PD-1 checkpoint blockage approaches, tumor infiltrating lymphocytes have shown promise in select cases [194], and other options have been reviewed by Chrétien *et al* [195]. However, immunotherapies come with their own risks and are prone to causing a wide range of immune-related adverse events such as cytokine storms [196] and elevated risk for developing secondary cancers [197, 198]. Additionally it remains unclear which patients benefit from immunotherapy and there is a lack of biomarkers predictive of treatment success.

### 1.4.6 Molecular Landscape

Breast cancers are driven by unique genomic and transcriptomic properties. Large-scale high-throughput sequencing initiatives including The Cancer Genome Atlas (TCGA), the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), the International Cancer Genome Consortium (ICGC) [199], and others have thoroughly mapped the genomic landscape of both early and advanced breast cancer in the past decade [200–206]. Others have explored the clonal evolution of breast tumors [207–209].

On average, breast tumors have a low to medium mutation burden compared to other cancer types such as melanoma [59, 67, 206, 210]. The landscape in early breast cancer is dominated by mutations in the genes *PIK3CA* and *TP53*. The oncogene *PIK3CA* is mutated in ~35% of breast tumors, and features a wide spectrum of missense mutations that lead to overactivation of growth signalling via the PI3K-AKT-mTOR pathway. This includes the most common mutation in breast cancer, *PIK3CA* H1047R [202, 211]. The tumor suppressor gene *TP53* on the other hand is frequently disrupted by dominant-negative point mutations, frame-shift indels, and nonsense mutations that trigger nonsense-mediated decay of the incomplete *TP53* transcripts.

Copy-number alterations play a major role in breast cancer. Particularly the locus 17q11 around the oncogene *ERBB2* is frequently amplified, which has lead to the adoption of HER2 testing to guide treatment with anti-HER2 therapy. Further recurrently amplified loci are 8q11, 8q13, and 8q24 on chromosome 8, as well as 17q23 and 20q13, involving genes such as *MYC* [202]. Other loci are frequently disrupted or lost entirely, such as those involving *PTEN* [212] and *RB1* [213]. Other SVs are common [214] and genomic SV hotspots exist [215], but recurrent SVs and gene fusions are rare [215–218]. However, expressed fusions do appear to negatively impact patient survival in advanced HoR+ breast cancer [219], and fusions deregulating miRNAs and snoRNAs have been reported [220, 221].

Other common occurrences in early breast cancer are epigenetic marks such as methylation. Promoter hypermethylation is common and leads to reduced or lost expression of important tumor suppressor genes such as *BRCA1* [222, 223] and *PTEN* [224]. Conversely, hypomethylation leads to increased expression and has for example been shown in certain Basal-like tumors in the loci containing the genes *MIA*, *KRT17*, and *KRT5* [225].

The overall processes leading to these genomic alterations leave a mark in the genome in the form of mutational signatures that have been thoroughly described in previous studies [54, 59, 226, 227]. All alterations impact the transcriptomic landscape of breast cancer in specific ways and distinct expression signatures have been associated with several of them, such as gross *PTEN* structural aberrations [228].

The alteration landscape of invasive tumors varies by histological and molecular subtype [229]. Ductal carcinomas are characterized by SNVs and indels in *TP53*, *GATA3* and *MAP3K1* and CNVs involving the oncogenes *ERBB2* and *MYC*, and others. Lobular tumors are defined by nonsense SNVs and frame-shift indels involving *CDH1*, leading to loss of mRNA expression and the characteristic loss of E-cadherin protein. Alterations in *PIK3CA* and *PTEN* are also associated with the lobular subtype. Tumors of HER2-enriched and Basal-like subtype have a higher SNV and indel load than Luminal-like tumors [202], and harbor more SVs [215, 218], highlighting the genomic instability inherent to these subtypes. All molecular subtypes are also associated with distinct DNA methylation patterns [146].

The patterns of genomic alterations shift during the evolution from early to advanced breast cancer, partly due to selection pressure from adjuvant treatment. Compared to primary tumors, metastases have a higher TMB, an elevated frequency of resistance mutations such as in *ESR1* [230] as well as shifted mutational signatures that are associated with adjuvant treatment [231]. Some metastases switch molecular subtypes compared to the primary tumor they derived from, particularly from Luminal A-like towards more aggressive subtypes [145, 232, 233].

## 1.5 The Human Genome and High-Throughput Transcriptome Profiling

The sequencing of the human genome and the release of the first draft sequences in 2001 were monumental efforts that fundamentally changed our view on biology and our approach to biomedical research [234]. With the evolving push towards personalized and precision medicine, the currently used human reference genome starts to show its limitations [235] and a variety of solutions have been proposed. These range from using multiple reference genomes, for example per-population reference genomes [236], to constructing pan-genomes [237, 238], to moving on from the currently predominant linear genome representations towards a graph-based genome representation that it better suited to represent complex genomic diversity [239].

### 1.5.1 The Human Reference Genome

Virtually all work in human cancer genomics is currently performed relative to the human genome reference sequence [16, 17]. The sequence was the result of the herculean effort by the Human Genome Project (HGP) headed by Francis Collins, as well as the company Celera Genomics headed by Craig Venter. The HGP approach was to tile along the human genome sequence using bacterial artificial chromosomes (BACs), thus sequencing the genome step by step. Celera's approach was to turn sequencing into a computational problem. The genome was shotgun sequenced by cutting DNA into short oligonucleotide pieces, sequencing them, and computationally re-assembling them into contigs. Being competitors for a long time, ultimately the two groups combined their efforts into the human reference genome. The "finished" sequence of the human genome was published in 2004 [240].

Two facts about the human genome reference are important to consider when using it for analyses. First, its sequence does not represent the sequence of one specific human being, although approximately 70% of the original human genome reference sequence originated from one person. Thus it is an amalgamation of sequences derived from different human beings [241, 242]. Second, the 2001 and 2004 genomes were by no means complete. The genome contains vast stretches of DNA that are inherently difficult to sequence, such as centromeres, telomeres, and other highly repetitive regions. Other regions have been sequenced, but so far could not be correctly placed, and are distributed as additional contigs ("patches") in some versions of the reference assembly. Since the release of the "finished" sequence in 2004, the reference assembly has been steadily improved [243], culminating in the current GRCh38 assembly [244]. However, to date the sequence remains incomplete.

Another consideration is the lack of diversity in the reference genome. Global genetic diversity has been explored in a variety of large-scale studies [245–247]. The difference between a typical human genome and the human reference genome has been estimated to be 4.1-5 million sites [246], which is likely an underestimation [248]. This was further illustrated by Sherman *et al* [249] who constructed an African pan-genome and found that it contained approximately 296.5Mb of sequence that has no representation in the human reference genome, resulting in these sites potentially being ignored in analyses relative to the reference genome. Similar results, albeit on a smaller scale, were found in the Icelandic population [250]. Consequently, lack of diversity may pose problems for precision medicine in populations with underrepresented genomic information.

Taken together, these limitations have long posed problems to analysis and clinical translation of sequencing [235, 237], such as reference bias and underdetection of potentially disease-relevant genes [251], and they are being addressed in various ways. The emergence of improved technologies such as long (>10,000bp) read sequencing promises to fill the gaps in the human genome sequence in the near future. For example, only recently the map of the Y chromosome centromere was generated [252], and the complete structures and sequences of chromosomes 8 and X were described [253, 254]. Further, it allows to detect more human diversity in form of SVs that were not possible to resolve using short-read sequencing data [248]. The lack of genome diversity has made clear that one single reference genome is not enough for future research and clinical purposes [237]. Efforts are ongoing in many countries such as Sweden [255–257], Denmark [258], and Japan [259, 260] to develop country-specific and even region-specific resources, for example reference genomes and variation databases. Adoption of graph genomes that are directly able to incorporate genetic variation is another possibility and an active area of research [239].

### 1.5.2 High-throughput Sequencing

Sequencing technology has improved remarkably since the first RNases were sequenced in the 1960s [261, 262]. High-throughput short-read sequencing (HTS) – also called high-throughput sequencing, next-generation sequencing (NGS), deep sequencing, second-generation sequencing, or massively parallel sequencing – refers to the repeated parallel sequencing of short (<1,000bp) DNA or RNA fragments. Since its introduction in the 2000s it has transformed biomedical research and our understanding of disease biology. This technique can result in up to thousands of sequence determinations of the same genomic locus and has been adopted into a plethora of sequencing methodologies, the most common ones being whole-genome sequencing (WGS), whole-exome sequencing (WES), targeted-capture sequencing, RNA sequencing (RNA-seq), and methylation analysis using bisulphite sequencing or TET-assisted pyridine borane sequencing (TAPS) [263]. Compared to earlier profiling technologies such as microarrays, HTS has higher resolution down to single nucleotides, can measure previously unknown (*de novo*) sequences, and provides greater dynamic range.

A specific variant of HTS – sequencing by synthesis – was originally developed by the company Solexa, and later bought by Illumina. Also referred to as Illumina sequencing, it is currently the most commonly used sequencing technology with an estimated worldwide instrument market share of 80% [264]. All studies included in this thesis make use of data generated using Illumina RNA-seq, and the technique is described in detail in Section 3.3.

While Illumina sequencing is currently the most commonly used technology, in the last few years third-generation sequencing or long-read sequencing has gained traction. These technologies, most prominently developed by the companies Oxford Nanopore and Pacific Biosciences, enable read lengths of tens of thousands of bases (Pacific Biosciences) to more than one million bases (Oxford Nanopore). These characteristics open up new opportunities to fill the gaps in reference genomes (see Section 1.5.1), improve SV detection [265], and variant calling in traditionally difficult to handle genome regions, as demonstrated in the precisionFDA Truth Challenge V2 [266].

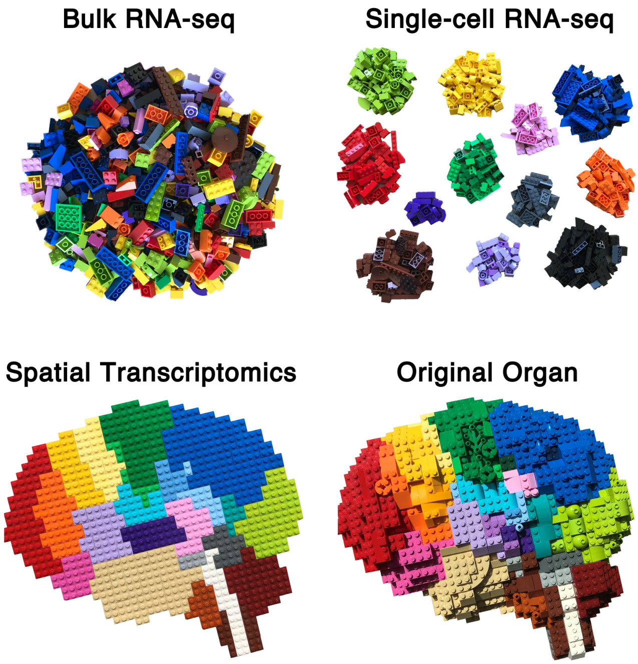


**Figure 1.7.** Visualization of the transcriptome profiling techniques bulk RNA-seq, single-cell RNA-seq, and spatial transcriptomics, as well as the original tissue donor organ using toy bricks. Each brick represents one cell, and color coding depicts cells with similar expression patterns and thus similar phenotype.

Image credit: Bo Xia (`https://twitter.com/BoXia7`)

### 1.5.3 Transcriptome Profiling

Different techniques are available for probing the cancer transcriptome. Expression microarrays became available in the early 2000s and could be used to measure the expression of known genes and isoforms. Around 2010, high-throughput short-read sequencing (RNA-seq) started to evolve [267] and has since effectively replaced microarrays as the principal method for transcriptome profiling. Unlike expression microarrays, RNA-seq is not restricted to known sequences and provides single base-pair resolution. The three most important steps in RNA-seq evolution are visualized in Figure 1.7. Bulk RNA-seq was the first available technique and provides an average readout across the input material. Single-cell RNA-seq became available in the early 2010s and gives a readout on the level of individual cells. Spatial transcriptomics was developed in 2016 and enhances single-cell RNA-seq by enabling spatial resolution of mRNAs in individual tissue sections [268]. The focus of this thesis is bulk RNA-seq, which will be referred to as RNA-seq from here on.

In recent years sequencing has been making inroads into the clinic and it is being used to stratify patients into relevant clinical subtypes, identify treatment-predictive or prognostic genomic aberrations, and to track minimal residual disease. In many modalities the focus has been on introducing DNA-based sequencing (DNA-seq) into the clinic, mostly in order to determine genomic driver alterations. RNA sequencing has received less focus outside of the research community, although some clinical applications in mendelian diseases [269–271], myeloproliferative neoplasms [272], childhood cancers [273] and others have emerged. In many cases RNA-seq accompanies DNA-seq in multi-omics approaches [274], and is typically used for subtyping and gene expression signatures. However, the capabilities of RNA-seq beyond this use case are now being recognized, as evident from recent reviews that have highlighted the growing importance and capabilities of RNA-seq, both from a technical and clinical view [275–278].

RNA-seq offers many advantages over previous methods. It has a greater dynamic range and reproducibility, and can detect *de novo* transcripts such as fusion genes in addition to quantifying known transcripts [267]. In addition to isoform and gene expression it offers single-base resolution, which unlocks a range of applications, for example the possibility to detect sequence variants [279–286], coarse copy-number aberrations [287–290], and structural variants [291, 292].

Through its sweet-spot between the genome and the proteome, transcriptome profiling using RNA-seq may be a powerful first-line clinical diagnostic tool. By enabling profiling of expression and genomic alterations simultaneously and within an actionable timeframe from surgery, a variety of gene expression signatures, for example for treatment response prediction, can be applied and drug susceptibility and resistance mutations can be evaluated.

## 1.6 Bioinformatics

While computational methods have been used in biochemistry since the 1960s, for example through the work of Margaret Dayhoff [293], the initial sequencing of the human genome and the advent of high-throughput molecular techniques such as HTS have transformed biology into a data-driven subject that requires computational knowledge. The term bioinformatics was originally coined in the 1970s by Paulien Hogeweg and Ben Hesper to describe "the study of informatic processes in biotic systems" [294]. Since then the term has evolved to describe a vibrant interdisciplinary field that develops, curates, and applies computational methods to transform data into biological and clinical insights. Bioinformatics encompasses a wide range of subject areas, including structural bioinformatics and proteomics, HTS, sequence analysis, and biological networks. An integral part of the field's culture has been the embrace of open source software development and permissible licenses for code and, increasingly, data [295–297]. In spite of its importance for the life sciences, the field still struggles with acceptance in the academic and medical realm, including lack of funding for development and maintenance of even critical methods [298], as well as lack of recognition and career options [299–301].

Bioinformatics is a crucial part of HTS, as the sequencing process transforms a traditionally wet-lab problem into a computational problem. Data processing is performed using computational pipelines or workflows, i.e. chains of different methods that work in concert to transform the data into the desired outcome or insight. To gain a better understanding of breast cancer and develop clinically meaningful diagnostic tools, the development of new computational methods and workflows is paramount.

## 1.7 Major Challenges

Survival of breast cancer patients has improved in recent years, however many challenges remain. Screening programs have enabled the early detection of lesions and thus either the chance to remove them before potentially becoming malignant, or if already malignant, to prevent cancer from spreading. However, this has resulted in increased detection of *in situ* lesions that would perhaps never develop into invasive breast cancer. Distinguishing harmless lesions from those that will become malignant is currently not reliably possible. As a consequence a significant fraction of women are likely cured by surgery and radiotherapy alone, or may only need comparably mild adjuvant therapy, but are being overtreated and thus suffer from unnecessary side effects, including long-term effects such as developing secondary cancer [302]. In addition to its physical and psycho-social effects, overtreatment also poses a significant economic burden on healthcare systems [303] and the patients themselves. Much effort is being put into finding ways to downstage low-risk tumors to spare treatment, such as the TAILORx clinical trial and others [304]. On the other end of the spectrum, patients who have lived disease-free for 15 years or more may still develop disease recurrence and ultimately succumb [101]. Distinguishing patients who will do well from those that will not is a major task for the future. Current diagnostic tools are imperfect, and in addition to the cases mentioned above, they also falsely identify a small proportion of cases as low risk, when in fact they are high-risk and could perhaps benefit from more or different treatments. We partly address this challenge in study III.

While treatment options have broadened in the last decades, resistance to drugs coupled with tumor heterogeneity continues to be a major challenge. External stimuli, such as treatments, provide selection pressure on the tumor and drive the evolution of resistant clones. Clones that harbor or develop a resistance mutation can thrive while competing clones succumb [305]. Examples of clinically important resistance mutations are endocrine-resistance causing mutations in *ESR1* [230], *ERBB2* [306, 307], genes of the *FGF* and *FGFR* families [308], as well as activating *ESR1* mutations and *PTEN* loss of function mutations that lead to alpelisib resistance [309]. While resistance mutations are particularly prevalent in metastatic tumors [307, 310], they can already occur in treatment-naïve primary tumors [311], as we also show in study IV.

A major unmet need is effective treatments for the patient population with triple-negative tumors. While for other tumors targeted treatment options such as anti-hormonal or anti-HER2 therapies are available, TNBC tumors currently lack viable molecular targets and have poorer survival. In recent years several subtypes of TNBC have been identified [312], and alternative therapies such as PARP inhibitors and immune checkpoint inhibitors [313] are in clinical trials and showing promise, possibly also in combination [314]. Many of these tumors harbor germline or somatic *BRCA1*/*BRCA2* mutations, or exhibit "BRCAness", meaning they exhibit homologous repair deficiency but are not BRCA-mutated [315]. These tumors can be detected using mutational and copy-number signatures such as HRDetect [316], and may likewise benefit from therapies such as PARP inhibitors.

The primary cause for breast cancer death is relapse of the disease in form of metastases. Detecting relapse is currently routinely being done using imaging techniques, either at regular checkup-intervals, or prompted by patient symptoms such as headache or bone pain. Since the early 2010s liquid biopsy approaches in form of circulating tumor cells (CTC) and circulating tumor DNA (ctDNA) have shown promise in tackling this problem [57, 317]. Both rely on the fact that tumors shed genetic material into bodily fluids such as blood where it can be detected in patient plasma from a simple blood draw. Prospective studies will be needed to bring these technologies into routine use for monitoring of minimum residual disease and treatment response.

## 1.8 Precision Medicine

Precision medicine – also called personalized medicine – is an approach to tailor treatments to the specific genetic, environmental and lifestyle conditions of a patient. In cancer this means in particular determining and taking into account the genomic traits of the tumor and targeting its specific aberrations, as well as adjusting the therapy choices, doses, and durations depending on the clinical follow-up and predictive laboratory tests. The field is moving from designing drugs by tumor site to targeting specific genomic aberrations across different cancer types. For example a recent clinical trial of the HER tyrosine receptor kinase (TRK) inhibitor neratinib enrolled patients based on mutations in the *ERBB2* and *ERBB3* genes, independent of the tumor site [318]. Another trial evaluated the efficacy of the TRK inhibitor larotrectinib across cancers with TRK fusions [319]. A milestone for precision medicine occurred in the year 2017 when the drug pembrolizumab achieved approval by the FDA for treatment of solid tumors with microsatellite instability or DNA mismatch repair deficiency, independent of tumor type. This marked the first time a drug was approved based on genomic biomarkers alone instead of histopathology [320].

While precision medicine has many proponents, it is not without controversy [321]. Critics have noted that thus far few tangible success stories exist, despite high promises and expectations, and a great deal of money that has been spent in this area by funding agencies, nonprofit organizations, and corporations. Further, genome-driven oncology does not currently benefit the majority of U.S. patients [322]. Even for the ones it does benefit, the number of patients with measurable survival benefits varies [321].

Besides matters of practical implementation and the arguments for and against precision medicine, it is important to keep its economic side in mind. Health care systems need to balance patient care with economic cost, which is a challenge in western societies due to rising cancer incidence and notoriously expensive cancer drugs [323, 324]. Genomics-guided diagnostics have the potential to drive down costs in the future by optimizing drug allocation and the ability to perform a multitude of computational tests and signatures based on data from a single laboratory test. Whether these hopes actually become reality needs to be determined by detailed economic studies, however preliminary studies indicate that sequencing benefits patients and is cost effective [325, 326].

As in most cases, the current reality about precision medicine lies between the extreme positions of its proponents and opponents. While the expectations on precision medicine were overly optimistic, particularly after the release of the first draft sequences of the human genome (see Section 1.5.1) and the early days of HTS, the reality is that precision medicine is moving into the clinics. In certain modalities such as advanced non-small cell lung cancer (NSCLC), diagnostics technologies such as qPCR, dPCR, and targeted sequencing are routinely being used to test for the presence of biomarkers such as the *EGFR* L858R mutation, which signals susceptibility to TKIs such as afatinib and crizotinib, and *EGFR* T790M which confers resistance to these TKIs, but which can be overcome with other drugs such as osimertinib. While not all cancer patients currently benefit from genomic technologies and targeted treatments, these numbers will increase, as they already have between 2006 and 2016 [322]. An example of precision medicine in real-life is the National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH) study (ClinicalTrials.gov identifier NCT02465060) [327] study, where patients are guided to therapies based on their genomic profiles across several tumor types.

## 1.9 The Sweden Cancerome Analysis Network – Breast (SCAN-B) Initiative

The Sweden Cancerome Analysis Network – Breast (SCAN-B) Initiative (ClinicalTrials.gov identifier NCT02306096) is a precision medicine initiative started in 2009 by Prof. Åke Borg as a joint effort of researchers, physicians, nurses and other health-care specialists to improve diagnostics, treatment, survival, and quality of life for breast cancer patients. The initiative is described in detail in study I and by Rydén *et al* [328]. The SCAN-B study initially started enrolling patients in 2010 at seven participating hospitals in Malmö,

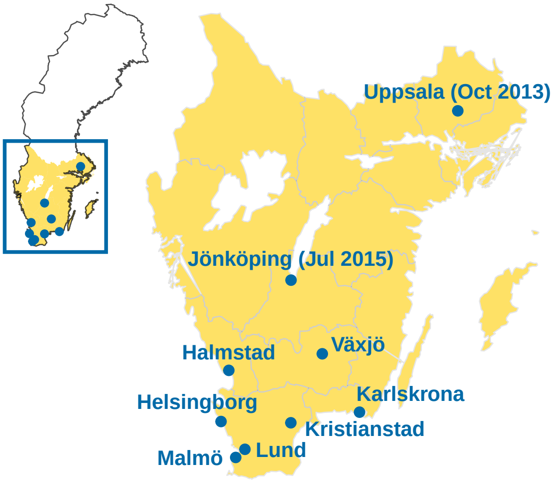


**Figure 1.8.** Map of Sweden with the sites participating in the SCAN-B initiative marked: Malmö, Lund, Kristianstad, Helsingborg, Karlskrona, Halmstad, Växjö, Uppsala, and Jönköping.

Lund, Kristianstad, Helsingborg, Karlskrona, Halmstad, and Växjö, all located in the South Sweden healthcare region. Since then sites in Uppsala (2013) and Jönköping (2015) have joined the effort (Figure 1.8), and there is a standing open invitation to hospitals in the Nordics to join.

The goals of the initiative are threefold: to introduce gene expression and genomic tumor profiling into the clinical routine for breast cancer; to improve tumor classification, diagnosis, prognostication and prediction of treatment effects; and to make improvements accessible to patients through implementation within the healthcare system, clinical trials, and cooperation with the drug and biotechnology industry.

All patients with breast cancer at participating sites are eligible to enroll in SCAN-B, which started in August 2010 with the main SCAN-B study enrolling patients with primary breast cancer. In addition, since January 2019 patients with metastatic breast cancer are eligible to enroll in the SCAN-B-rec sub-study (ClinicalTrials.gov identifier NCT03758976). Each participating patient gives written informed consent, and donates a piece of their tumor as well as a pre-operative blood sample. After the surgery further blood samples are taken at defined follow-up time points. All samples are sent to the Division of Oncology at Lund University for central analysis and biobank storage. Currently the analysis process consists of performing mRNA sequencing (RNA-seq) of the tumor samples, typically within one week of surgery. This short time-span from surgery to data is critical for the eventual translation of biomarkers in a clinically-actionable manner.

Compared to earlier studies and patient cohorts, SCAN-B represents a significant advance.

Patients have been treated with modern nationally standardized care regimens, such as anti-hormonal therapies and anti-HER2 therapies, and have been enrolled prospectively across a wide geography with nearly all new patients diagnosed being consented for SCAN-B. Older cohorts are often not representative since these treatments were not available at the time, or the cohorts are heterogeneous or not population-based, complicating comparisons with today's patients and thus seriously hindering their use in biomarker development. Its population-based nature and real-world conditions ensures that conclusions drawn from SCAN-B-based studies are representative and generalizable for the wider population. The importance of this was highlighted by Xie *et al* [329] who analyzed 70 widely used public breast cancer gene expression datasets and found that they do not reflect the disease at a population level. Instead, high grade and ER- tumors are over-represented, potentially leading to biased conclusions. As of December 2020, more than 16,000 patients have consented to be part of SCAN-B, translating to approximately 85% of eligible patients, and more than 13,500 RNA-seq libraries (including replicates) have been sequenced.

In addition to studies I–IV included in this thesis and many ongoing projects, the uses of the SCAN-B cohort thus far have covered research into many areas such as triple-negative [205] and *BRCA1*-abnormal tumors [223], benchmarking of gene expression signatures [163], investigation of gene fusions [220], molecular subtyping [155, 330] and psychological resilience [331], development of predictors of lymph-node metastasis [332], and comparisons of circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) for liquid biopsies [333].

# 2 Aims

*“They’re pretty high mountains,” said Azhural, his voice now edged with doubt. “Slopes go up, slopes go down,” said M’bu gnomically. “That’s true,” said Azhural. “Like, on average, it’s flat all the way.”*

— Terry Pratchett, Moving Pictures

RNA-seq is a versatile yet underused technique for cancer diagnostics. The overarching aim of this thesis was to evaluate, explore, and improve the usability of RNA-seq as a clinical diagnostics tool within the SCAN-B project and breast cancer diagnostics.

The specific aims of the four studies included in this thesis were as follows:

I To describe the SCAN-B study and its protocols and computational RNA-seq pipeline, provide an early evaluation of the enrolled patient cohort, evaluate the generated RNA-seq data by comparison of RNA-seq and expression microarrays performed on the same samples, and prototype variant calling.

II To describe format validity problems in the widely-used RNA-seq alignment software packages TopHat and TopHat2 and develop a software tool to correct the problems.

III To assess variation within standard clinical histopathology, explore classification of clinically important biomarkers from RNA-seq-based gene expression profiles, and validate the classifications on overall patient survival in an independent cohort.

IV To develop a computational pipeline for detection of somatic SNVs and indels from tumor-only RNA-seq data, and explore the mutational landscape of a large real-world primary breast cancer cohort in relation to patient overall survival.

# 3 | Methods

*Sometimes it's better to light a flamethrower than curse the darkness.*
— Terry Pratchett, Men at Arms

## 3.1 Patients, Samples, and Ethics

All patients who contributed tumor material to the studies in this thesis were enrolled in the SCAN-B study, or the precursor study All Breast Cancer in Malmö (ABiM). As part of the enrollment process they were informed about the study by trained medical professionals, and patients provided written informed consent. Studies i, iii, and iv were approved by the Lund ethics review board and performed in accordance with the Declaration of Helsinki. Study ii did not include patient material or data and thus no ethics permissions were required.

The enrollment, biospecimen sampling, and analysis processes of the SCAN-B study are described in detail in study i. Importantly, surgical tumor specimen are kept in RNAlater (Ambion) preservative after the routine pathology assessment, ensuring high quality RNA for later sequencing. Only remaining tumor material after the pathological assessment is included in SCAN-B, ensuring that SCAN-B enrollment is not a detriment to routine clinical care. ABiM samples were collected at surgery and stored fresh-frozen.

**Table 3.1.** Patient datasets and experimental setups used in studies i, iii, and iv.

| Source | Material | Experimental Setup | Patients | Samples | Study |
|---|---|---|---|---|---|
| SCAN-B | Tumor | RNA-seq / Microarray | 49 | 49 | i |
| SCAN-B | Tumor | RNA-seq | 3,273 | 3,273 | iii |
| ABiM | Tumor / Normal | RNA-seq / Targeted DNA-seq | 273 | 275 | iv |
| SCAN-B | Tumor | RNA-seq | 3,217 | 3,217 | iv |

The patient cohorts and experimental setups used in this thesis are described in Table 3.1. Study i included 49 tumors that were analyzed using RNA-seq and expression microarrays. Study iii included 3,273 tumors, which were selected according to the flow diagram in Figure 3.1 to include all invasive, non-metastatic, unilateral breast tumors. For study iv we used a cohort of 275 tumors from 273 patients (two patients with bilateral disease) assembled from the ABiM study, and re-used the cohort assembled for study iii. Applying additional quality checks reduced the number of patient tumors in the latter cohort to 3,217.

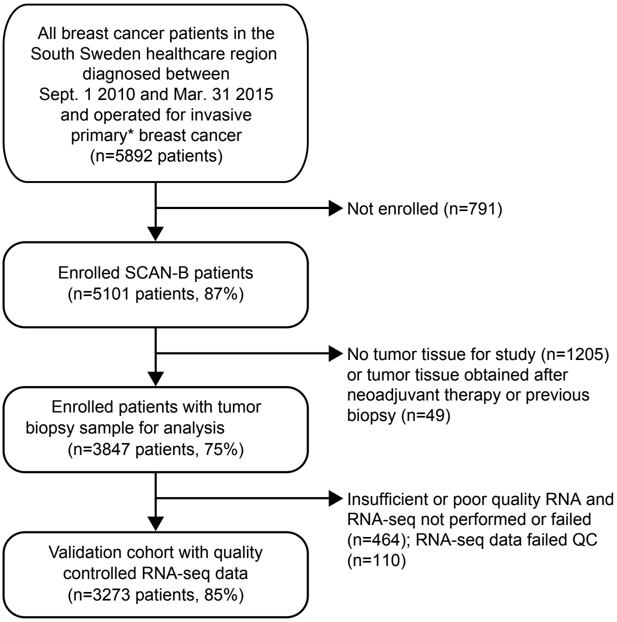

All breast cancer patients in the South Sweden healthcare region diagnosed between Sept. 1 2010 and Mar. 31 2015 and operated for invasive primary* breast cancer (n=5892 patients)
Not enrolled (n=791)
Enrolled SCAN-B patients (n=5101 patients, 87%)
No tumor tissue for study (n=1205) or tumor tissue obtained after neoadjuvant therapy or previous biopsy (n=49)
Enrolled patients with tumor biopsy sample for analysis (n=3847 patients, 75%)
Insufficient or poor quality RNA and RNA-seq not performed or failed (n=464); RNA-seq data failed QC (n=110)
Validation cohort with quality controlled RNA-seq data (n=3273 patients, 85%)

**Figure 3.1.** Patient cohort diagram for study III. * Non-metastatic primary unilateral breast cancer, which excluded patients with a diagnosis of synchronous (<3 months) contralateral invasive breast cancer.

Source: Adapted from Study III, Supplementary Figure A1 (CC-BY 4.0)

## 3.2 DNA Microarrays

DNA microarrays were first developed in the mid 1990s [334] and were the dominant tool to measure RNA expression, SNPs, methylation, and other markers in the 2000s. They allowed the expression of large numbers of genes to be simultaneously measured for the first time. While RNA-seq has since become the preferred tool for expression profiling, microarrays are still widely used. Microarrays are chips with tens of thousands to hundreds of thousands of short oligonucleotide probes attached to a solid surface. The working principle is visualized in Figure 3.2. Each probe is complementary to a part of the target DNA molecule or "feature" to be measured. For gene expression profiling, the input mRNA is reverse-transcribed into cDNA and labelled with a fluorescent dye. The cDNA sample solution is then flooded over the chip, allowing the cDNA to hybridize to the probes, while unhybridized cDNA is washed away. Hybridization is quantified using a scanner that excites the labelled DNA using a laser and measures the resulting fluorescence. The resulting data needs to be analyzed while accounting for technical factors.

Microarrays have several limitations. First, they can only detect previously known se-

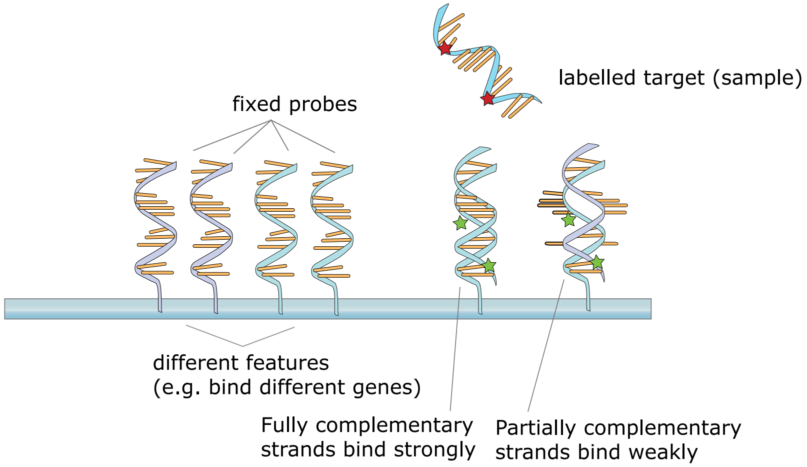


**Figure 3.2.** Microarray working principle. Fluorescently labelled target sequences hybridize to complementary probes on the microarray surface. Probe-bound sequences produce a fluorescence signal when laser-excited (green stars), while unbound sequences are washed away and thus do not produce signals (red stars).

Source: `https://commons.wikimedia.org` (Public Domain)

quences, since complementary probes need to be present on the chip for detection. In cancer, this means *de novo* transcripts such as novel gene fusions, mutated transcripts, or unknown isoforms resulting from, for example, aberrant splicing cannot be detected. While microarrays can measure expression, they cannot resolve the more intricate features of the transcriptome such as RNA modifications, or sequence changes such as those caused by somatic mutations. Technical problems such as cross-hybridization and hybridization failure lead to high levels of background noise and missing values in the resulting data [335, 336]. High background noise and signal saturation limit the dynamic range so very low or very high expression cannot be accurately reflected. Due to variability of microarray platforms, comparison of microarray datasets from different platforms is inherently difficult and requires extensive normalization (techniques reviewed by Walsh *et al* [337]. Like many other high-throughput techniques, including RNA-seq, microarrays are prone to batch effects that influence interpretation and may have to be corrected [338, 339].

These technical problems pose challenges to the data analysis of microarray experiments. To compensate, microarrays typically contain control probes to estimate background noise. Using these, the estimated noise can be removed from all other measurements. Another common problem is missing values, since many downstream algorithms require complete data. This has led to the development of a variety of imputation methods to "predict" missing values from other samples, reviewed by Aittokallio [340]. If a gene shows missing expression values in too many samples it is typically prudent to exclude it from further analysis, since incorrect imputation may lead to false conclusions.

In study I we compared expression data resulting from the SCAN-B RNA-seq pipeline to Illumina Human HT12 v4 BeadChip microarrays covering 47,231 probes across the human transcriptome [341]. While expression levels were comparable between the platforms, the comparison highlighted the superiority of RNA-seq compared to microarrays in terms of dynamic range and reproducibility. We also compared molecular subtyping based on three gene lists and found high concordance between the two technologies. However, this task highlighted another issue, in that probe annotations may be wrong or incomplete, but are required to map probes to their target mRNAs. For example in study I we were able to match most but not all probes included in the Sørlie [137] and Hu [140] subtyping lists with our RNA-seq data.

## 3.3 High-Throughput Sequencing

High-throughput sequencing has revolutionized molecular biology. While third-generation technologies such as Pacific Biosciences single molecule real-time sequencing (SMRT) and Oxford Nanopore are gaining popularity, Illumina sequencing is by far the most commonly used [342]. Depending on the setup it can be used for DNA sequencing (DNA-seq) or RNA-seq. Common DNA-seq experimental setups are WGS, WES, or targeted sequencing hybrid capture-based panels. The difference between these setups is which exact part of the genome is being sequenced. As part of study IV, we used a custom hybrid-capture sequencing panel, while the remaining sequencing data in studies I, III, and IV is based on whole mRNA RNA-seq.

Illumina sequencing machines use a method based on sequencing by synthesis (SBS) and reversible terminators. The method was invented by Shankar Balasubramanian and David Klenerman, and first commercialized by the company Solexa, which Illumina acquired in 2007. A schematic of Illumina sequencing is shown in Figure 3.3. The sequencing process is performed in a "flow cell" – a compartment containing lanes through which reagents can flow in from one side, react with the surface, and are flushed out the other side. Depending on the sequencing machine model one or more flow cells can operate in parallel, with each flow cell containing multiple lanes. During sample preparation, sequencing adapters are ligated to the template cDNA molecules to be interrogated. These adapters are immobilized onto the flow cell surface using complementary bait oligonucleotides. Each immobilized template sequence is multiplied into a cluster of ~1,000 sequences using bridge amplification (Figure 3.3A). On the flow cell surface(s), many millions of clusters are generated which can then be sequenced simultaneously and in parallel. Depending on the setup, sequencing is performed in single-end or paired-end mode. In single-end mode, each molecule in a cluster is only sequenced from one end, resulting in a single read. In paired-end more, each molecule is sequenced from both ends, resulting in two reads. This provides additional in-

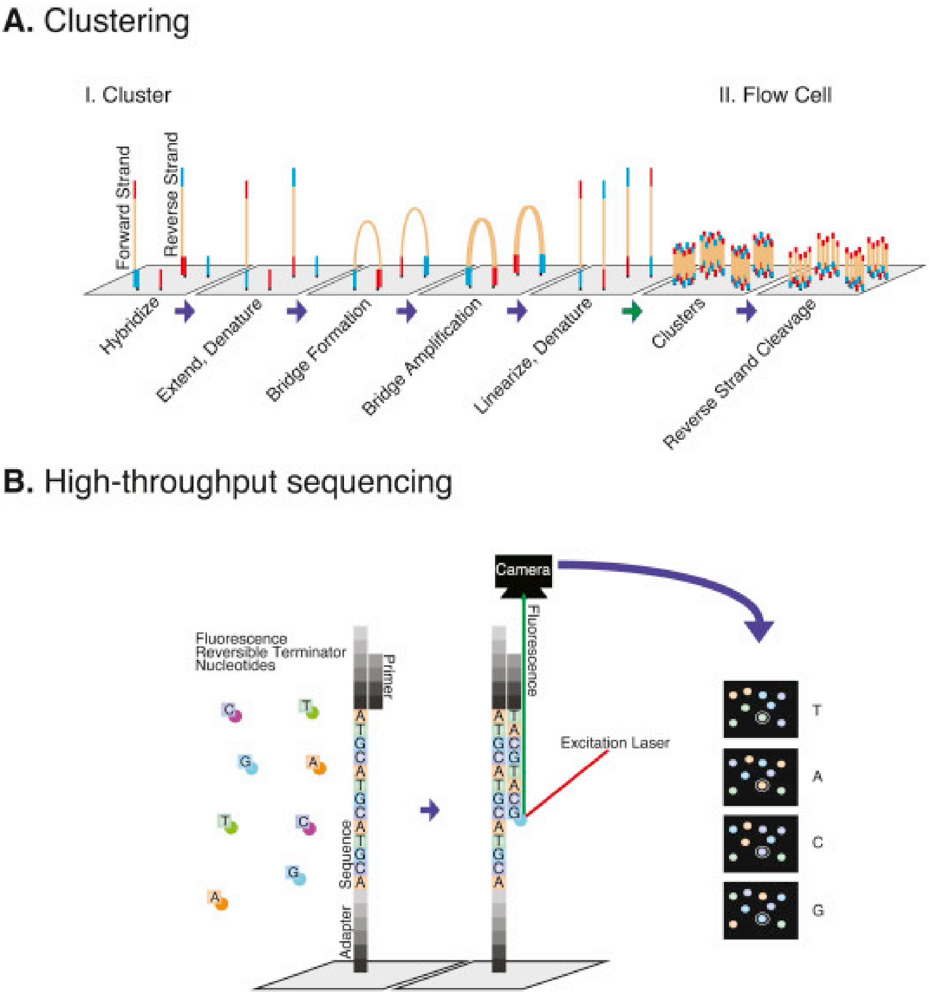


**Figure 3.3.** Illumina sequencing working principle using four distinct colors for the nucleotides `A`, `T`, `G`, `C`. **A.** The path from hybridization of template reads to the flow cell to cluster generation using bridge amplification. **B.** High-throughput sequencing using sequencing by synthesis with reversible terminators.

Source: Chaitankar *et al* [343]; reprinted with permission from Elsevier (panel C. not shown).

formation to subsequent *in silico* analysis, as the expected distance between the paired reads, as well as their orientation is known. The sequencing process itself is performed in cycles, where each cycle starts by flooding the lanes with millions of deoxyribonucleotide triphosphates (dNTPs: `A`, `T`, `G`, `C`) that contain a reversible terminator that blocks polymerase activity (Figure 3.3B). Each terminator is labelled with one of two or four (depending on the instrument) fluorophores that emit a different color when laser-excited. One dNTP

**Table 3.2.** Phred base qualities for defined base call accuracies.

| Phred Quality Score ($Q$) | Probability of Incorrect Base Call ($P$) | Base Call Accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

Source: Illumina Technical Note: Quality Scores for Next-Generation Sequencing [346].

is incorporated into each template strand on the flow cell. After that, two or four flow cell images are recorded, corresponding to the number of colors used and the terminators are removed so the next cycle can begin. Due to the presence of terminators, each cycle can only incorporate one base. The number of sequencing cycles is user-configurable, but is typically between 50 and 150 cycles, depending on the sequencing instrument. The per-cycle images are analyzed and converted into base calls by on-instrument software. In addition to the base calls themselves, Illumina sequencers assign a quality value to each call. This value is determined according to the Phred model shown in Equation 3.1 [344, 345], where $P$ is the probability of calling a wrong base. A Phred score of 30 therefore equals a base call accuracy of 99.9% (Table 3.2). Base quality values are crucial for downstream analysis, and in a typical Illumina sequencing run the vast majority of bases have a Phred quality of ≥30.

$$Q = -10 \cdot \log_{10}(P) \tag{3.1}$$

Multiple sample libraries can be sequencing in the same flow cell using multiplexing or pooling. During the library preparation DNA molecules are tagged using sample-specific barcode sequences. Multiple libraries are then pooled and sequenced concurrently. After sequencing the reads can be demultiplexed *in silico* into sample-specific sequence files using their sample barcodes.

All sequencing data used in this thesis was generated using Illumina HiSeq 2000 and NextSeq 500 instruments. These instruments differ in their workings in that the HiSeq 2000 represents each of the nucleotides `A`, `C`, `G`, and `T` with a distinct fluorophore emitting a different color when laser-excited. The NextSeq uses a simplified system based on two colors where `C` (red) or `T` (green) are labelled with dedicated colors, `A` is labelled with both colors, and `G` is unlabelled. This system causes a reduction of the number of images that need to be taken during each sequencing cycle to two, down from four with the four-color system. While this simplified system has led to a decrease of sequencing price, it is more prone to over-calling `G` bases, since a genuine base call cannot be distinguished from a situation where no signal is detected due to technical error, such as cluster degradation [347].

## 3.4 RNA Sequencing

RNA sequencing using high-throughput short-read sequencing (RNA-seq) has emerged as the leading methodology for transcriptome profiling [267]. In recent years, it has effectively replaced microarrays as the principal method for transcriptome profiling since it offers many advantages over previous methods. RNA-seq has a greater dynamic range and reproducibility, detection of *de novo* transcripts such as fusion genes in addition to quantifying known transcripts, as well single-base resolution. These capabilities enable a multitude of applications, such as the possibility to detect fusion genes and calling sequence variants [279–282, 284–286], coarse copy-number aberrations [287–290], and structural variants [291, 292], and the analysis of splicing and isoform switching [348].

While the name suggests direct sequencing of RNA molecules, it is instead typically performed by sequencing cDNA resulting from RNA reverse-transcription. While direct Illumina short-read RNA sequencing is possible in principle, it has never matured and consequently is essentially unused [349, 350]. More recent sequencing technologies such as nanopore sequencing have been used to directly sequence RNA [351–353], and even detect RNA modifications [354, 355].

A general overview of the RNA-seq workflow from sample to result is depicted in Figure 3.4. To prepare a sample for sequencing, input RNA has to be transformed into a sequencing library. All RNA-seq of SCAN-B samples included in this study were sequenced using a customized version of the stranded dUTP protocol [356]. Selection of this protocol was made based the results of a comparison of stranded protocols from the literature [357] and an in-house comparison of the Parkhomchuk second strand dUTP approach, Illumina directional RNA ligation and the Epicentre ScriptSeq protocols. Newer libraries within the SCAN-B RNA-seq workflow have shifted to newer protocols and today use the off-the-shelf Illumina TruSeq protocol.

RNA-seq libraries can be prepared in a variety of ways. Two important properties are whether or not the library preserves information about which strand a transcript originated from (strand specificity), and whether single-end of paired-end sequencing is performed, i.e. whether a template molecule is only sequenced from one end, or both ends.

### 3.4.1 Library Preparation and Sequencing

The RNA-seq data used in studies i, iii, and iv are based on libraries originating from a customized version of the strand-specific dUTP protocol by Parkhomchuk *et al* [356], that is described in detail in study i. While RNA-seq library preparation protocols have many commonalities with DNA-seq protocols, specific steps are included to ensure preservation of RNA properties. Most importantly, care has to be taken to preserve strandedness.

The customized dUTP protocol used for all SCAN-B samples included in the studies within

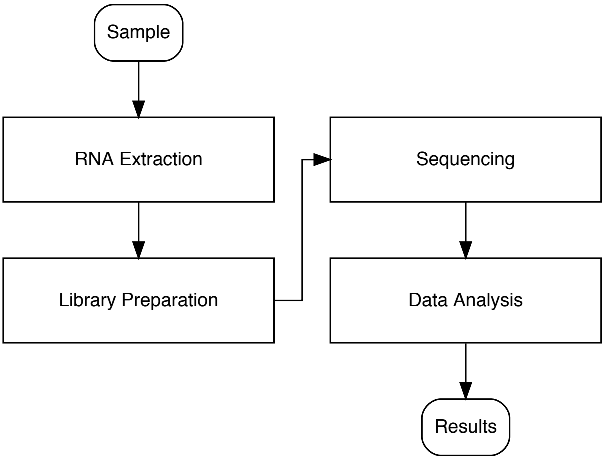


**Figure 3.4.** High-level view of the RNA-seq workflow.

this thesis is described in detail in study 1, and the steps are summarized in Figure 3.5. In brief, starting from $1\mu g$ total RNA, mRNA is purified using poly-DT DynaBeads (Thermo Fisher Scientific), and subjected to Zinc-mediated fragmentation (Ambion). The resulting approximately 240bp fragments are isolated using Zymo spin columns (Zymo Research). Using the fragmented mRNA as input, first-strand cDNA synthesis is performed by adding random hexamer primers, reverse transcriptase, and dNTPs. Following cleanup of excess reagents, second-strand synthesis is initiated by adding polymerase and dNTPs with dUTP instead of dTTP. Resulting double-stranded cDNA is isolated using Zymo spin columns (Zymo Research), followed by 5'/3' end-repair and ligation of Illumina TruSeq sequencing adapters (Illumina), including sample barcodes. Size-selection is then performed to remove excess free adapters. To preserve strandedness, the dUTP-containing second cDNA strand is digested using uracil-DNA glycolase (UDG).

An important parameter in a sequencing setup is the average depth of coverage or number of reads to target for each samples. In a DNA-seq experiment, sequencing reads are distributed approximately uniformly along the targeted area of genome. An example target depth is 30X, meaning on average each targeted base is covered by 30 reads. Deviations from the uniform coverage assumption occur due to biases during library preparation and sequencing, for example caused by GC-rich regions [358, 359]. RNA-seq differs from DNA-seq in that reads are distributed approximately proportional to their expression level in the input sample, meaning that the average sequencing depth across an RNA-seq dataset is not a useful metric. Instead, the total number of sequencing reads is used to express how deep one has sequenced.

The product of the library preparation process is a library of adapter-ligated double-stranded cDNA that is ready to be loaded onto a sequencer and multiple libraries are pooled. Within

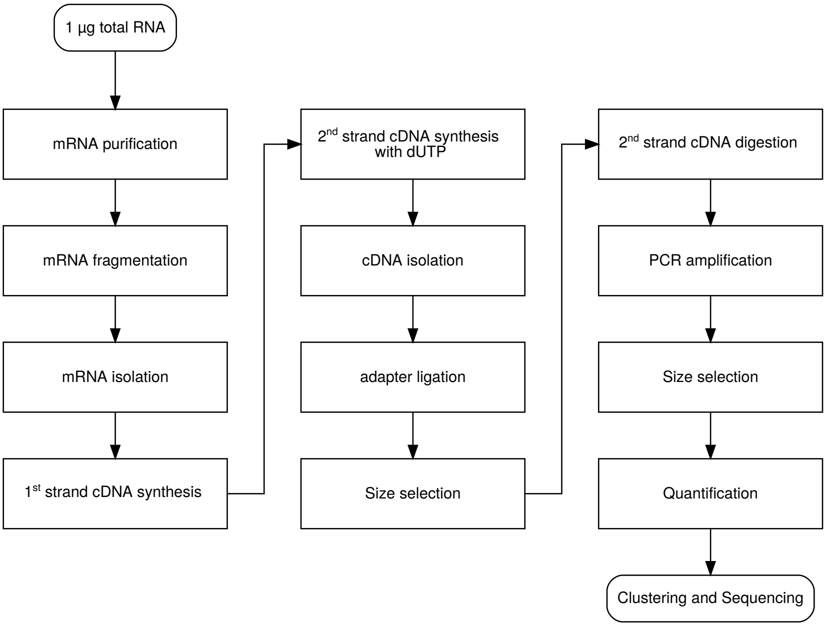


**Figure 3.5.** Simplified workflow of the dUTP library preparation protocol.

SCAN-B, on the HiSeq 2000 instrument library pools were sequenced across two flow cells in two lanes per flow cell to insure against technical failures and to reduce technical bias. The newer NextSeq 500 instrument loads one flow cell containing 4 lanes, and the library pool is automatically sequenced across all lanes. All RNA-seq samples included in studies I, III, and IV were prepared using the dUTP protocol, and sequenced in paired-end mode on Illumina HiSeq 2000 or NextSeq 500 sequencers with a sequencing target of approximately 30 million read-pairs per sample.

### 3.4.2 Computational RNA-seq Analysis

Data processing and analysis is a key step in the RNA-seq workflow. The SCAN-B RNA-seq processing pipeline that studies I, III, and IV relied on is implemented within the BASE laboratory information management system [360, 361] through the extension package Reggie [362]. The pipeline follows the general computational RNA-seq workflow outlined in Figure 3.6, and is described in detail in studies I and III, as well as by Häkkinen *et al* [362]. It will be summarized here, and discussed in more detail in the following sections.

In brief, the SCAN-B computational pipeline used in studies I, III, and IV consisted of the following steps. Base-calling was performed using Illumina's on-instrument software. After sample demultiplexing, reads were trimmed to remove adapters and low quality 5'/3' bases using Trimmomatic [363]. Reads that aligned to the PhiX phage genome, ribosomal

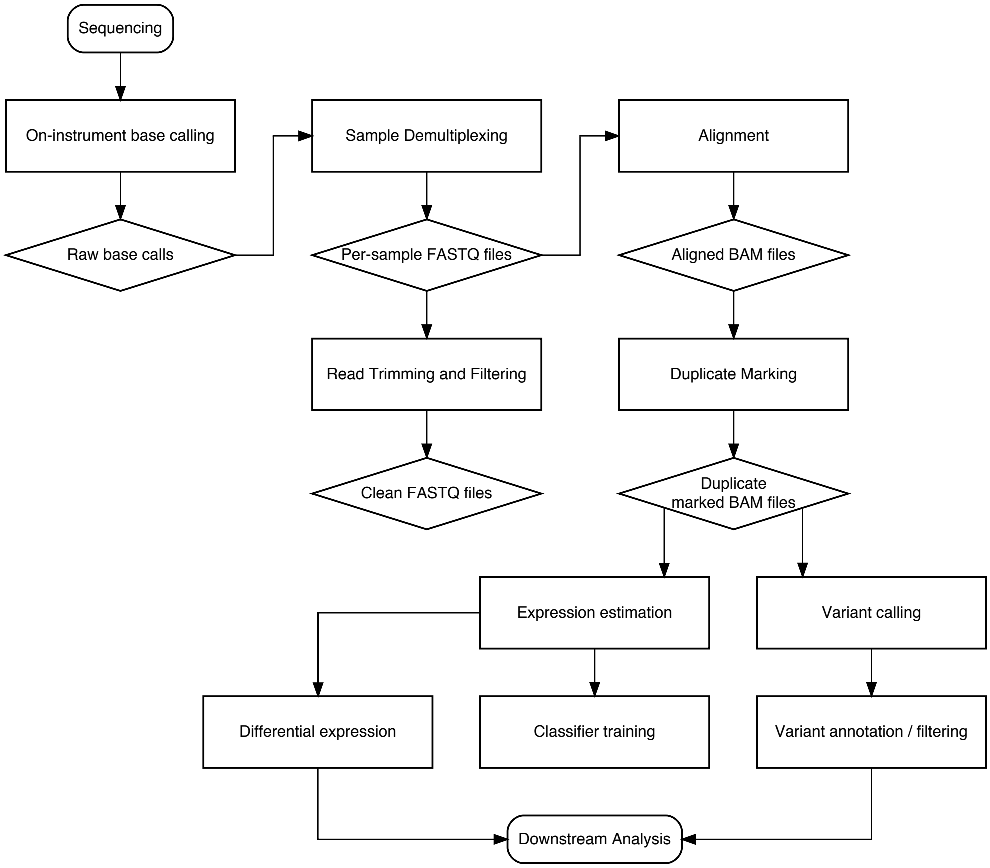


**Figure 3.6.** General computational RNA-seq workflow.

DNA/RNA, or the UCSC RepeatMasker track [364] using Bowtie2 [365] were removed. Bowtie 2 alignments were also used to estimate the fragment size distribution of the remaining reads. Using this information, reads were aligned to the GRCh37/hg19 (study I) or the GRCh38/hg38 (studies III and IV) version of the human reference assembly and the UCSC knownGenes transcriptome model using TopHat2 [366]. Unmapped reads were corrected using TopHat-Recondition as discussed in paper II. Using the aligned reads, transcript expression was estimated using Cufflinks [367, 368] and summed on the gene level.

More recently, SCAN-B samples are being processed with a pipeline that has been updated to use HISAT2 [369] for alignment, and StringTie [370] for expression estimation. These updates provide considerable improvements to both run-time and resource use. In addition, StringTie provides the ability to output read counts, which is the recommended input for differential expression profiling according to best practices [371].

The variant calling pipeline developed as part of study IV was based on demultiplexed

FASTQ files from the SCAN-B pipeline, but deviated from it from this point. Alignment was based on HISAT2 using a version of the GRCh38 reference assembly that included alternative sequences and decoys to optimize alignment and reduce artifactual variants caused by alignment problems. The pipeline was implemented within the bcbio-nextgen [372] framework.

The individual steps involved in RNA-seq computational analysis are discussed in the following sections.

### Demultiplexing and FASTQ Conversion

Illumina sequencers convert fluorescence signals from read clusters into nucleotide base calls using Illumina's on-instrument CASAVA software. Base calls are stored in the Illumina-own BCL format. To convert BCL to the widely used FASTQ format, and to separate the pooled samples for further analysis, sequencing reads are demultiplexed into sample-specific FASTQ files. Two widely used software solutions for this task are the Illumina-own bcl2fastq and IlluminaBasecallsToFastq from the Picard suite [373], which is used to demultiplex SCAN-B sequencing runs.

### Read Trimming and Filtering

Raw sequencing reads may be very short, contain adapter sequences, and/or low quality 5' and 3' bases, all of which may complicate subsequent analysis and in particular read alignment. Adapter contamination occurs when the cDNA template being sequenced is shorter than the requested read length and thus sequencing continues into the adapter. Low quality bases occur, for example, at the 5' ends due quality model calibration, and at the 3' end due to imperfect sequencing. Each cluster on the flow cell consists of ~1,000 individual cDNA templates. After many sequencing cycles, synthesizing the individual reads can get out of sync causing lower confidence base calling.

Many computational pipelines, including SCAN-B, trim these potentially problematic read features to improve downstream analysis. Whether and how much trimming improves downstream analysis however, and which parameters are optimal, is often unclear and systematic evaluations of these questions are rare. An early study concluded that it is beneficial for germline variant detection, but not necessarily for expression profiling with modern alignment software such as TopHat2 [374]. Similarly, a recent study concluded that trimming is not necessary for expression profiling, since modern aligners such as HISAT2 support soft-clipping [375], which allows for low-quality read ends to remain in place in an alignment dataset, leaving it to downstream processing software to ignore or consider them. If trimming is performed, the choice of parameters that guide trimming aggressiveness have a substantial impact on analysis quality [376].

Sequencing datasets frequently contain unwanted reads. For example the PhiX lambda phage DNA is frequently spiked-in as a control, and RNA samples contain a large amount of ribosomal RNA (rRNA). Although RNA-seq library preparation protocols typically include rRNA depletion, or as within the SCAN-B protocol, selection of poly(A)-tailed RNA, these procedures are imperfect and rRNA may still be sequenced. Removing these sequences saves computational time and space, removes a potential of analysis errors, and improves expression estimation. Technically this is accomplished by aligning all reads against the unwanted sequences, and only selecting reads that do not align to these sequences for future analysis.

### Alignment

During read alignment, also called mapping, individual reads are placed into the correct position along a reference genome. For RNA-seq this is typically done with the help of a transcriptome annotation that provides information about splice junctions and transcript isoforms. Compared to aligners written for DNA, RNA-seq aligners are splicing-aware and can take this extra information into account during alignment. Aligners that fall into this category include TopHat [377], TopHat2 [366], HISAT [378], HISAT2 [369], and STAR [379]. In studies I and III we used TopHat2 in combination with the post-processor TopHat-Recondition described in study II to correct problems in unaligned reads. In study IV we used HISAT2, the successor of TopHat2.

### Duplicate Marking

Duplicate reads most often occur as a product of PCR during the library preparation process (PCR duplicates) or due to the sequencer detecting the same template cluster multiple times (optical duplicates), but they can also occur naturally as true duplicates. Marking duplicate reads allows downstream analyses to ignore them, since in many cases they do not add additional information, but can bias analyses such as variant calling. Several software tools for duplicate marking exist [380–384] and virtually all of them follow the approach implemented by the MarkDuplicates tool from the Picard suite [373]. It works by comparing the 5' coordinates and sequences of single reads or read-pairs. Matching reads/read-pairs are ranked by the sum of their base qualities, and all but the highest scoring read/read-pair are flagged as duplicate. While the tools work similarly at the core, they differ in implementation, which influences performance and functionality. For example, biobambam supports steaming, meaning it can operate in Unix pipes, while Picard MarkDuplicates does not [383].

Whether or not marking duplicates in RNA-seq data is appropriate is unclear, but may depend on the specific use case. The chance that reads with the same start/end coordinates arise from distinct molecules is typically higher in an RNA-seq experiment compared to WES/WGS due to lower library complexity. On the other hand, library preparation involving PCR results in many reads that originate from the same molecule. Parekh *et al*

[385] found the impact of duplicate marking on RNA-seq differential expression in the best case improved metrics only mildly, and could otherwise even be detrimental. On the other hand, Quinn *et al* [386] found duplicate marking beneficial in RNA-seq SNP detection. Not marking duplicates in this case can severely overestimate variant allele frequencies (VAFs), potentially leading to false positive calls (see Section 3.4.2). Thus, whether or not to mark duplicates is a judgement call that has to be made with taking the experimental setup, for example the number of PCR cycles, and the analysis endpoint in mind.

The need for duplicate marking can be avoided or reduced by using PCR-free library preparation protocols, or by using unique molecular identifiers (UMIs) [387], where during library preparation each input molecule is tagged with an individual barcode. After sequencing, reads can then be regarded as duplicated if they share barcodes and map to the same genome coordinates [388–390]. This approach is common in single-cell RNA-seq [391] and HTS approaches for sensitive variant detection.

The standard SCAN-B computational pipeline as used in studies i, iii, and iv performs expression profiling as analysis endpoint and does implement duplicate marking, although the Cufflinks expression estimation software does not exclude duplicate reads from analysis. The variant calling pipeline developed in study iv is distinct from the standard pipeline, but does mark duplicates as well, for the reasons outlined above.

### Expression Profiling

Expression estimation on the transcript and gene level is the most common use-case for RNA-seq. Traditionally, the input for this type of analysis are sequencing reads that have been aligned to a reference genome, although methods based on pseudo-alignment have become available since [392, 393]. With the help of a transcript annotation that describes introns and exons, the number of reads can be counted per transcript. Raw counts are biased by transcript length and number of reads per sample so counts need to be normalized to enable within-sample and between-sample comparison. Different methods for normalization are available, all with their own biases and drawbacks [371, 394–396]. Starting from raw read counts, the measures reads/fragments per kilobase of exon model per million mapped reads (RPKM/FPKM, for single-end/paired-end data respectively) were introduced as a measure for expression that is within-sample normalized for library size and transcript length [397]. This makes it difficult to compare RPKM/FPKM measurements between samples [398]. A later measure is transcripts per million reads (TPM), which accounts for the same factors, but reverses the order of normalization operations to enable better comparability between samples. Contemporary methods for differential expression analysis largely require raw counts, since they account for transcript length and library size themselves [371].

In studies i, iii, and iv we used expression estimated in FPKM as generated by Cufflinks [367, 368]. To reduce skewing of the data and ease fold-change calculations and comparisons we further transformed the values using $\log_2(FPKM + C)$, with $C = 0.1$. The addition of a constant is needed to avoid zeros, since $\log_2(0)$ is undefined.

### Variant Calling

Variant calling (detection) of SNVs and indels from HTS data is important to identify somatic cancer mutations. Since HTS has become available, dozens of methods (callers) have been developed (partly reviewed and/or benchmarked in [399–407]). An important distinction is whether a method has been developed for germline calling, somatic calling, or both. Somatic variant calling, particularly in cancer genomes, poses additional challenges due to aneuploidy and potential artifacts due to challenging read alignment. The gold standard for somatic variant detection is calling using data from matched tumor and normal samples from the same patient. This setup allows the caller to reliably differentiate between somatic and germline variants. While DNA-seq can in principle recover variants genome-wide, RNA-seq variant calling is limited to the expressed parts of the genome. Even if a variant is expressed it may be missed, due to transcriptional processes such as NMD removing the mutated transcripts from view before they can be captured for sequencing. On the other hand, RNA-seq provides exceptional sequence coverage in highly expressed genes, which may be used to detect low VAF variants that may be missed by DNA-seq.

A wide variety of somatic variant callers is available, with VarScan [408], VarDict/VarDict-Java [409], and MuTect2 [410] being among the most widely used. While variant calling from DNA is the most common setting and not all somatic variant callers have been tested or are recommended for RNA-seq, several approaches for RNA-seq mutation calling, mostly in combination with matched tumor and/or normal DNA, have been developed [279–286]. Combined variant calling of tumor RNA and tumor DNA makes it possible to discern true somatic mutations from transcriptomic effects such as RNA editing, while the use of normal DNA has already been discussed.

Often, reference datasets such as those generated by the Genome in a Bottle consortium are used for benchmarking and optimizing variant calling pipelines [411, 412]. These are currently focused on DNA-seq and comparably well characterized tumor RNA-seq datasets are missing.

For detection of SNVs and indels we used VarScan [408] in study I, while in study IV we used VarDict-Java, a reimplementation of VarDict [409] in the Java language. We switched from VarScan to VarDict-Java based on its streamlined workflow that performs realignment around indels, does not require sequence pileups as input, and generates a wealth of sequence-based annotations, as well as internal benchmarks that were later affirmed by external benchmarks [404, 405, 413]. MuTect2 [410] only became available during the course of study IV, and has since been used for RNA-seq data [414, 415]. Quaglieri *et al* [413] determined that 30–40 million read-pairs are necessary to detect the majority of known recurrent mutations in the tested acute myeloid leukemia TCGA samples, which matches the sequencing target of approximately 30 million read-pairs per sample used by SCAN-B.

### *Error Sources*

A major challenge in variant calling is the differentiation between true somatic variants and false positive calls due to technical artifacts or germline variants, particular when calling using tumor-only data. Technical artifacts can occur at different levels before, during, and after the sequencing process (reviewed in [416]). Before sequencing, prolonged time between surgery and sample preservation by flash-freezing or RNAlater may lead to DNA/RNA degradation. During library preparation, several processes may induce artifacts, most importantly PCR amplification of DNA leading to misincorporations and chimeras [417–419]. The sequencing process itself is imperfect and can lead to wrong base calls, particularly in challenging regions such as repeats, homopolymer stretches, and GC-rich regions [358, 359, 420, 421]. Even after sequencing, alignment artifacts may lead to false positive variant calls.

### *Annotation*

The wide range of error sources necessitates extensive filtering to remove false-positive calls. To enable better filtering and aid in downstream analysis, variant calls need to be annotated with additional information that allows them to be evaluated in the genomic and clinical context. Relevant information includes locational context, such as nearby genes or location in an intron or exon, population frequency through databases such as dbSNP [422] and the Genome Aggregation Database (gnomAD) [423]. Clinical information, such as cancer driver status and whether presence of the variant signals susceptibility or resistance to drugs, for example through database such as the Catalogue of Somatic Mutations in Cancer [424, 425] or CIViC [426], is particularly important in cancer. An additional layer is the predicted functional impact which can be obtained from tools such as PolyPhen [427], SnpEff [428], and the Ensembl Variant Effect Predictor (VEP) [429]. In study I we used ANNOVAR [430] for annotation, while in study IV we used vcfanno [431] due to its speed and flexibility.

### *Filtering*

Using variant annotations added by the variant caller and during the annotation process, variants can be filtered using various criteria. The exact settings may vary depending on the specifics of the downstream analysis. For tumor-only somatic variant calling, filtering germline variants is essential. To compensate for the lack of a matched normal sample this can be partly addressed using global germline variant databases such as dbSNP [422] and gnomAD [423], and national resources such SweGen [432]. For filtering of technical artifacts a variety of variables are important, for example base quality, proximity to low complexity areas such as repeats or homopolymers, and GC content. Since valid and clinically important variants may be filtered out, variants may be "rescued" by evaluating their presence in databases such as COSMIC [424] and CIViC [426]. Even after filtering, manual refinement and curation of variant calls is often necessary to establish clinical relevance. This process is not currently standardized, however standard operating procedures have been proposed [433, 434].

Technical background noise makes it challenging to detect variants below 1% VAF, although sequencing approaches exist that go below this [435, 436]. Competing technologies such as digital PCR (dPCR) allow detection of single or up to few specific variants as low as 0.001% VAF using assay technologies such as IBSAFE/SAGAsafe (SAGA Diagnostics) [333, 437–439] with lower cost and faster turnaround time than HTS.

### Quality Control and Sequencing Metrics

High-throughput sequencing is a complicated process comprised of many individual steps, ranging from the initial nucleic acid extraction from a sample, over the preparation of sequencing libraries, to the sequencing process itself. Each step may induce errors and biases, which makes quality control at various points in the process a necessity [440].

Basic quality control can be performed after demultiplexing to check whether base- and read-level metrics conform to expectations. Important metrics include the number of reads per sample barcode, average base quality and read length, GC and unique kmer content, and sequencing adapter contamination. These metrics are partly generated by demultiplexing software such as IlluminaBasecallsToFastq from the Picard suite [373], the popular software FastQC [441], and others. If quality problems are detected at this stage, further analysis can be avoided and the library can be re-sequencing, or a new library can be prepared. More advanced metrics can be examined after sequence alignment and duplicate marking. Interesting metrics at this level include the percentage of uniquely aligned reads, the percentage of duplicated reads, average insert size, and for paired-end sequenced libraries the percentage of properly paired reads, defined as read-pairs with both reads aligned within the expected distance and orientation. A variety of software packages for QC at this stage exist, including RNA-SeQC and Qualimap [442, 443]. A comprehensive quality analysis of early SCAN-B RNA-seq datasets has been performed previously [444].

An important confounding problem in RNA-seq are batch effects, where technical factors add variation and thus have systematic impact on the results. In other words, they describe a setting where variation between samples can be better explained by technical factors than by true biological variation. This is a problem in particular for quantitative analyses such as expression estimation. Several approaches have been developed for detecting and correcting batch effects in high-throughput experiments in general, and expression data in particular [338, 445–448]. Within SCAN-B, laboratory and sequencing processes have been optimized to minimize batch effects, and the *swamp* R package [446] is used to correlate expression data with a set of clinical variables, as well as technical variables such as dates of library preparation and sequencing.

## 3.5 DNA Sequencing

DNA sequencing in various forms is the most commonly used form of sequencing. Although this study focuses on RNA-seq, we used DNA-seq to optimize properties of our RNA-seq analyses. In study IV we used 275 samples from 273 patients that were sequenced using a custom targeted panel of 1,697 genes and 1,047 miRNAs (Agilent SureSelect) [449].

The computational analysis pipeline is very similar to that discussed for RNA-seq in Section 3.4, but simpler in many ways, and excludes expression-related analyses that are specific to RNA-seq. Sequence alignment does not need to account for a transcriptome model including splice sites, making the alignment process easier. While variant calling from DNA-seq data is not a solved problem and considerable variation between approaches still exists [450–452], it is easier compared to RNA-seq calling, since the complexity of the transcriptome does not contribute to false positive calls.

## 3.6 Molecular Subtype Inference

Since the description of the intrinsic molecular subtypes Luminal A-like, Luminal B-like, HER2-enriched, Basal-like and Normal-like [80, 137], multiple gene signatures have been developed for classifying tumors [137, 140, 148, 453, 454]. The PAM50 method [148] has become the *de facto* standard for classifying tumor expression profiles into intrinsic molecular subtypes. It was named after the Prediction Analysis of Microarrays (PAM) [455] method that was used to determine the subtype centroids, and the list of 50 genes the signature is based on. More recent and refined subtyping schemes such as IntClust [456] comprising ten subtypes derived from gene expression and CNVs have not gained much traction yet.

Inference of molecular subtypes is performed using single sample predictors (SSPs), defined by Perou *et al* as "... any predictor where the algorithm and any parameter values are exclusively determined from a training set, and test cases are assessed independently" [457]. Widespread clinical translation of subtyping SSPs has been hampered by several factors, summarized by Staaf and Ringnér [458]. Notably, different methodologies were found to be only moderately concordant on a cohort level, although they stratified prognosis similar to one another. There was a lack of robustness on a single sample level, where subtype inferences with different SSPs often disagree [459, 460]. The causes for this include inexact subtype definitions exemplified by the progression from the initially suggested classification gene list [136] to the now widely used PAM50 signature, to unclear descriptions of methodology, all taken together leading to reproducibility problems [459–462].

An additional problem is that previously developed SSPs, including PAM50, rely on normalization methods such as gene centering against a large heterogeneous set of samples for robust classification [461, 463]. Centering scales the expression of each gene across samples in the dataset so that the mean or median expression is 0. This compensates for different expression scales caused by technical platforms different from that originally used to train the SSP, as well as batch effects between datasets. This step adds an implicit dependency on other samples and contradicts the definition of an SSP as given previously.

To improve the robustness of subtype classification, a variety of different approaches have been proposed [454, 464–468]. Particularly the Absolute Intrinsic Molecular Subtypes (AIMS) approach suggested by Paquet and Hallett [454] garnered interest, as it promises true single sample prediction. During training, AIMS generates a number of binary rules in the form "`ESR1 < FOXC1`" that capture the relative within-sample expression between two genes in a specific subtype, and thus obviating additional normalization.

We performed inference of molecular subtypes in studies I, III, and IV. Study I was still at an early stage in the SCAN-B, were subtyping had not yet been implemented on a SCAN-B-wide level. For this study we used PAM50 subtyping implemented in the *genefu* R package [469]. In time for studies III and IV we refined our subtyping procedure by following the approach by Parker *et al* [148] using a SCAN-B-internal reference dataset that matches the original Parker *et al* cohort in terms of clinical characteristics. Before subtyping tumors, gene expression of the PAM50 genes for each tumor was centered to the reference dataset in order to normalize it to the original training cohort.

## 3.7 Histopathology

Histopathology is the mainstay of current cancer diagnostics. In breast cancer, immunohistochemistry (IHC) is routinely used to determine the status of the biomarkers ER, PR, HER2, Ki67, and NHG. It works by staining tissue slides with protein-specific antibodies, and counting or estimating the percentage of stained cells in a representative section of the slide. Using a cutoff value, the percentage is categorized, typically into the "low" and "high" categories. Generally, this technique is prone to reproducibility problems, both from a technical and human perspective. On the technical side, dozens of factor can influence good IHC results, such as choice of antibody and method of staining [470]. On the human side, the same IHC staining may be interpreted differently by two pathologists, or even by the same pathologist at different times. Considerable effort has been spent to standardize IHC for routine breast cancer biomarkers in terms of antibodies, procedures, and interpretation. This has resulted in very high concordance for ER and PgR. While HER2 reproducibility is good with IHC alone, additional gene copy-number testing using *in situ* methods such as FISH or SISH is recommended in borderline cases to verify *ERBB2* amplification [471]. Reproducibility for NHG is considerably lower, mostly owing to the existence of the intermediate category grade 2. Ki67 only recently entered the clinical guidelines, and standardization efforts are still ongoing [472]. While concordance is improving, it does not reach the high standards of ER, PgR, and HER2 yet.

One way to improve reproducibility is digital pathology, where the IHC slides are scanned and scoring is performed by machine learning based pattern recognition [473–476]. However, image recognition only addresses the variation caused by the reader; the technical problem of staining variation remains. Other approaches have evaluated the possibility of using gene expression based to determine biomarker status [178–184], however they are still not widely used. An important consideration for gene expression based approaches is that gene expression and protein expression are not always well correlated. Processes such as NMD can remove mRNA transcripts before translation, and epitranscriptomic modifications may impact translation rates. The mechanisms of mRNA and protein correlation have been reviewed by Buccitelli and Selbach [477].

In study III we evaluated the concordance between stains and between readers by performing a multiple-stain, multiple-reader evaluation in a cohort of 405 breast tumors. We also proposed gene expression based approaches of determining the status of ER, PgR, HER2, Ki67, and NHG, and validated classifiers for all five biomarkers in a large population-based SCAN-B cohort.

## 3.8 Machine Learning

Machine learning refers to the process of teaching a computer to perform a task based on prior learning. The field encompasses a wide range of techniques. In genomics, the two most commonly used approaches are supervised and unsupervised learning [478]. This nomenclature refers to whether the actual classes or "labels" of the input data are available to the algorithm (supervised) or not (unsupervised). Supervised methods such as Random Forests and Support Vector Machines are often used for sample classification, while unsupervised methods such as k-means and hierarchical clustering structure data solely using the intrinsic properties of the dataset and are thus well suited for exploratory analysis. Artificial neural networks with many layers ("deep learning") can be either supervised or unsupervised, depending on how they are used [479].

Hundreds of learning algorithms have been developed, complicating selection and evaluation. According to the "no free lunch" theorem [480], there is no one algorithm that is clearly superior in all use cases, and an approach that works well in one problem domain may work poorly in a different one. In general it is unclear whether machine learning methods are generally superior to simpler statistical approaches such as logistic regression [481].

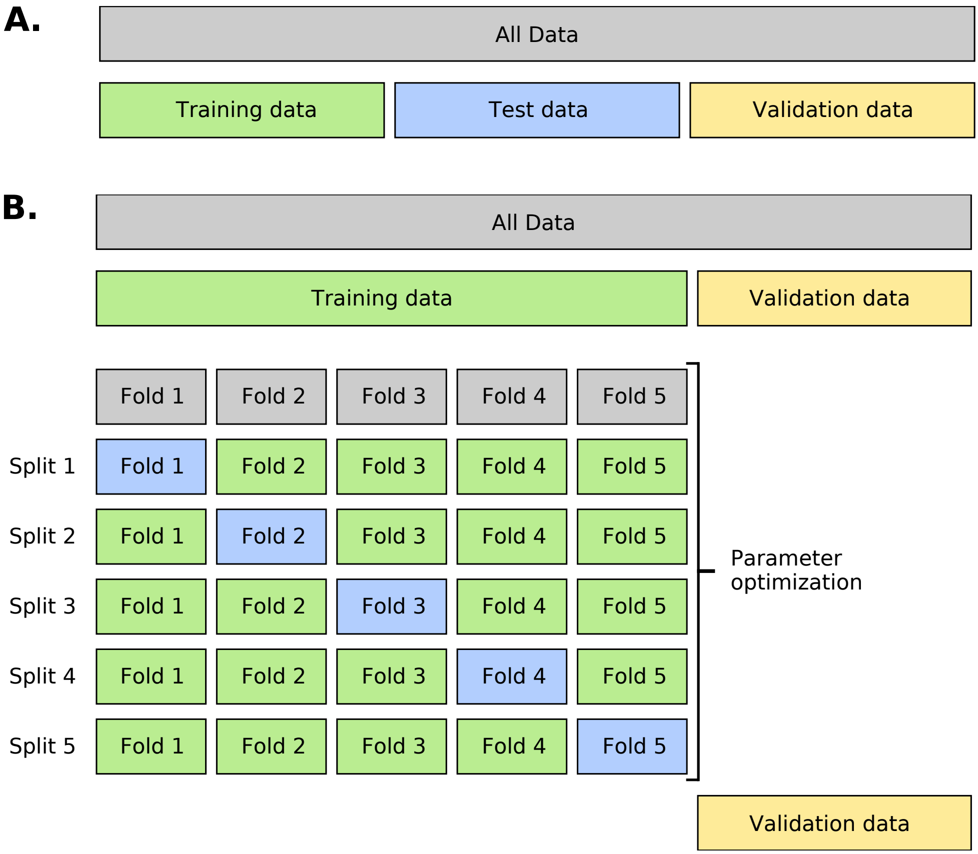

A.
All Data
Training data
Test data
Validation data
B.
All Data
Training data
Validation data
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Parameter optimization
Validation data

**Figure 3.7.** Data splitting for model training and validation. **A.** Standard data split into training, test, and validation datasets. **B.** Cross-validation data split into training and validation datasets.

Machine learning in biology and particularly in genomics typically has to deal with the "curse of dimensionality", where the number of features (variables) is much larger than the number of observations (samples). Thus, a major consideration in supervised machine learning is bias due to overfitting which can reduce the generalizability of a model. Overfitting can be counteracted in multiple additive ways. The best but often most impractical way is increasing the number of samples. Generally, simpler models are less likely to suffer from overfitting than complex models. Ways to achieve simpler models are removing features or adding regularization which penalizes complex models. Splitting data appropriately for training and testing is another way to avoid overfitting. Data is typically split into three subsets: training, test, and validation[1] (Figure 3.7A). A model is fitted in the training set and evaluated in the test set. Based on the evaluation the model features and

[1] There are differences in the nomenclature for test and validation in the literature, causing the two terms to be used interchangeably.

parameters can be adapted before being evaluated again. This training loop can be repeated as many times as needed, although this may, again, contribute to overfitting. Alternatively, particularly when the available dataset is small, cross-validation can be used (Figure 3.7B). During cross-validation, the sample set $N$ is split into $M$ equal partitions. $M - 1$ parts are used for training and the $M^{\text{th}}$ part is used as a test set. The procedure is repeated until every partition has been used as test set.

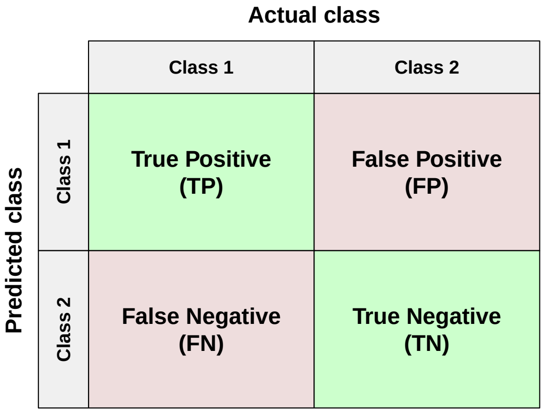
Importantly, information leaks from the validation dataset to the model must be avoided. These can happen for example when testing a model against the validation dataset and adjusting the model based on the results. In this case the validation dataset loses its independence and further tests of the adjusted model are invalid, as the performance measures obtained from it would not be generalizable.

In studies I, III, and IV we used a variety of unsupervised and supervised methods. Study I and IV used hierarchical clustering to group samples based on gene expression and pathway mutations. In study III we used the supervised PAM method based on nearest shrunken centroids [455] implemented in the *caret* and *pamr* R packages [482, 483] to develop classification models for the breast cancer biomarkers ER, PgR, HER2, Ki67, and NHG based on gene expression data. A centroid here is the mean expression value of all included genes. The method was initially developed for microarray-based transcription profiling and thus can handle highly dimensional data. It works by calculating a standardized centroid for each class in the training dataset by dividing the per-class centroids by the respective within-class standard deviation, and afterwards "shrinking" the per-class centroids towards the overall centroid by a user-configurable threshold parameter. The shrinkage step reduces the effect of noise and eliminates non-informative genes [455]. To classify a new sample, the standardized centroid for this sample is calculated, and the distance to each per-class centroid is determined. The sample is then assigned the class with the shortest distance of its per-class centroid to the sample centroid. In our study we used repeated cross-validation to optimize the shrinkage parameter, train classification models, and estimate the variation across multiple different cross-validation splits. We validated the resulting classifiers in a large cohort of breast cancer samples. We strictly adhered to keeping the validation set independent and only used it to evaluate the final models.

While we developed RNA-seq variant filters in study IV using a training and validation set, we did not aim to keep the validation set independent. Due to the complexity of the problem we instead explicitly used information from the validation set to improve the filters. As such the filters may be overfitted to SCAN-B data and not generalizable to other datasets.

### 3.8.1 Classifier Performance Metrics

The performance of a binary classification model can be expressed through a confusion matrix (Figure 3.8) which describes true positives (TP), true negatives (TN), false positives (FP,



| Predicted class | | Actual class: Class 1 | Actual class: Class 2 |
|---|---|---|---|
| | Class 1 | True Positive (TP) | False Positive (FP) |
| | Class 2 | False Negative (FN) | True Negative (TN) |

**Figure 3.8.** Confusion matrix for two-class problems, consisting of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Type I errors), false negatives (FN, Type II errors). Based on these a variety of model performance metrics have been developed. Simple measures such as accuracy (Equation 3.2) assess raw performance, but do not account for the fact that a classification model can "guess" right just by chance. More sophisticated measures such as Kappa (Equation 3.4) and Matthews Correlation Coefficient (MCC; Equation 3.5) take this chance effect into account. These metrics are commonly interpreted according to a scheme proposed by Viera and Garrett [484], outlined in Table 3.3, that adds intuition to the pure numbers.

A common way to visualize the performance of a classifier using different thresholds is plotting the receiver operating characteristic (ROC) curve using the metrics sensitivity and $1-$specificity. Other graphical methods that are thought to perform better than ROC for unbalanced datasets include precision/recall curves and MCC-F1 curves [485]. One can then select the threshold that maximizes the area under the curve (AUC).

Which metric to use for classifier evaluation during training includes the question which property of a classifier to prioritize. In study III we used balanced accuracy (Equation 3.3) [486] during training for this task, since it strikes a balance between sensitivity and specificity, and works well in unbalanced datasets. For performance evaluation in the validation dataset we used accuracy (Equation 3.2), MCC (Equation 3.5), Kappa (Equation 3.4), and positive/negative predictive value (Equations 3.6 and 3.7).

Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

**Table 3.3.** Common interpretation of the Kappa and MCC statistics according to Viera and Garrett.

| Kappa / MCC | Agreement |
|---|---|
| $\leq 0$ | Less than chance |
| 0.01–0.20 | Slight |
| 0.21–0.40 | Fair |
| 0.41–0.60 | Moderate |
| 0.61–0.80 | Substantial |
| 0.81–0.99 | Almost perfect |

Source: Viera and Garrett [484]

Balanced Accuracy (BACC):

$$BACC = \frac{1}{2}(\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}) \tag{3.3}$$

Cohen and Fleiss' Kappa:

$$Kappa = \frac{(p_o - p_c)}{(1 - p_c)} \tag{3.4}$$

where $p_o$ is the observed agreement, and $p_c$ is the chance agreement.

Matthew's Correlation Coefficient (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.5}$$

Positive Predictive Value (PPV):

$$PPV = \frac{TP}{TP + FP} \tag{3.6}$$

Negative Predictive Value (NPV):

$$NPV = \frac{TN}{TN + FN} \tag{3.7}$$

## 3.9 Statistical Analysis

Statistical hypothesis testing deals with the question of whether or not a specific null hypothesis can be explained by the available data. They result in a probability (P-value) for obtaining the observed or more extreme results, for example the quantitative difference between two groups, if the null hypothesis were true. For better or worse, a P-value of

P<0.05 is commonly interpreted as a difference being significant. Importantly, P-values only give the probability that an effect exists, but are not a measure of effect size. P-values are intricately linked with sample size in that large sample sizes can lead to small P-values even if the effect size is very small [487].

Based on their assumptions, statistical tests can be stratified into parametric and non-parametric tests. Parametric tests rely on the approximate normal distribution of the input data, while non-parametric tests do not. In studies I, III, and IV we relied on 1-sided and 2-sided Fisher's exact tests for all hypothesis tests. Additionally we performed survival analyses in studies III and IV, which will be described below. All statistical analyses in studies I, III, and IV were performed in R using diverse set of extension packages, most importantly the *survival* package.

### 3.9.1 Survival Analysis

Survival analysis refers to investigation of time to the occurrence of a specific event. In cancer the specific event depends on the selected endpoint, summarized in the guidelines of the Definition for the Assessment of Time-to-event Endpoints in CANcer trials (DATECAN) initiative [488]. For example the event may be patient death from any cause (overall survival, OS), or relapse of the disease (relapse-free survival, RFS).

The most common method for survival analysis is that by Kaplan and Meier (Kaplan-Meier, KM) [489]. It is a univariable method to estimate the survival function for a group of patients, for example stratified by the status of a biomarker, using patient status at last observation and the time to event. The KM method is used to calculate the fraction of patients still alive at a given time, for example after diagnosis of breast cancer. To test for significant differences in the survival curves between patient groups, the log-rank test is typically used.

To estimate the size of the effect of a variable such as biomarker status on the time to event, Cox proportional hazards models can be evaluated [490]. Cox models are often used in conjunction with KM-analysis to estimate the effect of the same variable visually analyzed using KM plots (univariable analysis). It can then be expanded to correct for additional variables, most importantly possible confounding variables (multivariable analysis). The effect is estimated as the hazard ratio (HR), which is a measure of relative risk and interpreted as follows. A HR of 1 for a variable means no risk difference between groups. A HR of 1.5 equals a 50% risk increase relative to the comparison group, whereas a HR of 0.5 means a reduction of risk by 50%. This interpretation depends on the adherence of the model to the proportional hazards assumption, so a change in HR of 0.1 approximately equals an increase/reduction of relative risk by 10%. Common methods to test that this assumption is upheld are QQ plots, Schoenfeld residuals [491], and Grambsch and Therneau's test for non-proportionality [492].

In studies III and IV we performed survival analysis using the KM method, log-rank tests, and Cox models, and OS as endpoint. In particular we evaluated the association of predicted biomarkers and somatic mutations in various constellations with patient survival.

# 4 | Results and Discussion

*Vimes felt that a comment was called for.*
*He said: 'Arrgh.'*

— Terry Pratchett, Guards! Guards!

## Study I
## The Sweden Cancerome Analysis Network–Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine

The SCAN-B initiative was launched in 2010 as a population-based study to sample biomaterial from breast cancer patients for molecular research. The purpose of study I was to describe the SCAN-B study and infrastructure, including the protocols and workflows, and to describe the clinical features of the patient population enrolled between the years 2010 and 2013. Many biomarkers and signatures for breast cancer classification have already been developed using transcriptome profiling techniques such as microarrays. We compared a sample of 49 SCAN-B tumors (six as technical replicates) processed by the SCAN-B RNA-seq pipeline with tumors profiled using Illumina HumanHT-12 v4 BeadChip microarrays, and evaluated whether array-developed techniques yield the same results when subjected to RNA-seq data. As examples for such array-based signatures we chose subtyping signatures based on the gene lists identified by Sørlie *et al* [137], Hu *et al* [140], and Parker *et al* (PAM50) [148]. We also highlighted the potential of mutation calling in RNA-seq samples.

In this study we showed that SCAN-B enrolled 85% of the eligible patients across the accruing sites in the early years of enrollment. By comparing the distribution of several clinical characteristics within all patients and enrolled patients we demonstrated the population-based nature of the study. In our hands, gene expression and subtyping between RNA-seq and microarrays was highly concordant. Results were highly reproducible between primary and replicate samples. In general, this study showed the feasibility of using RNA-seq as primary analytical tool within SCAN-B, its advantages over microarrays such as increased dynamic range, and the quality of the generated data. The routines and workflows described as part of this study paved the way for studies III and IV. In particular we demonstrated that mutation calling using the generated RNA-seq data is feasible, which lead us to explore this topic in depth in study IV.

# Study II
# TopHat-Recondition: A post-processor for TopHat unmapped reads

A principal part of the RNA-seq workflow is a computational pipeline that cleans the raw sequencing reads, aligns them to the human reference genome, and performs quality assessment. TopHat and TopHat2 were popular spliced-read mappers [366] for alignment with close to 20,000 combined citations. TopHat2 was used in studies I and III. All versions of TopHat/TopHat2 contain bugs that cause their output to diverge from the Binary Alignment/Map (BAM) format specification [380, 493]. Due to the design decision of the TopHat/TopHat2 authors to write aligned and unmapped reads to separate files, and the focus of most analyses on aligned reads, these problems remained undetected and uncorrected. This can make downstream analysis challenging and is relevant not only for ongoing sequencing, but also for the hundreds of sequencing datasets processed with TopHat/TopHat2 that have been deposited in archives such as the Gene Expression Omnibus (GEO) and the European Nucleotide Archive (ENA).

While most analyses focus on aligned reads, unmapped reads have a number of uses. First of all, they can be used for quality control purposes. A high number of unmapped reads may indicate quality issues such as low quality reads or cross-species contamination of input samples or reagents [494–496]. Recently, unmapped reads were used to improve the human reference genome by identifying sequences that are missing from the genome [249, 497], and to uncover missed indels [498]. Within structural variant calling, read-pairs with one unmapped read are being used to detect and refine breakpoints by re-aligning them to the genome sequence around the putative breakpoint to better localize the exact breakpoint coordinates [57, 499]. Lastly inspecting unmapped reads in detail can aid in improving alignment software itself.

We developed the software TopHat-Recondition as a post-processor for TopHat/TopHat2 files that can repair them so they conform to the specification, and thereby improve compatibility with important downstream software such as the Picard suite and GATK [500]. Through availability in the popular Bioconda software repository [501] and integration in the bcbio-nextgen [372] RNA-seq pipeline, the software is readily available for use.

Since the publication of this study, the SCAN-B pipeline has been updated to replace TopHat2 with its official successor, HISAT2. The BAM files written by this software do not have the same issue as those written by TopHat2, and consequently TopHat-Recondition has been retired from the pipeline. While even the original authors of TopHat/TopHat2 discourage the use of their software in favour of newer tools such as HISAT2, publications referencing TopHat2 and even TopHat are still being published, indicating they are still being used. As such, there remains a potential userbase for TopHat-Recondition beyond deposited data.

## Study III
## Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-based Multicenter Sweden Cancerome Analysis Network–Breast Initiative

The biomarkers estrogen receptor (ER), progesterone receptor (PgR), epidermal growth factor receptor 2 (HER2), Ki67, and Nottingham histologic grade (NHG) are established prognostic and predictive biomarkers in breast cancer care. Since current evaluation of these biomarkers by histopathology is imperfect, we thought to develop computational classifiers to predict these markers from tumor transcriptional profiles.

We performed a comprehensive histopathological evaluation of 405 primary breast tumors using technical replicates and readings by three pathologists to estimate inherent variability in clinical pathology and to generate reliable consensus scores for each biomarker. Using the consensus scores, we determined optimal expression cutoffs for the biomarkers with a single underlying gene (ER, PgR, HER2, and Ki67) resulting in single-gene classifiers (SGCs). We also trained multi-gene classifiers (MGCs) by fitting nearest shrunken centroid models [455], and performed cross-validation to determine optimal parameters. The performance of the SGC and MGC classification models was validated in an independent cohort of 3,273 tumors from the SCAN-B study by comparing classification results to the clinical pathology results and, importantly, to patient overall survival (52 months median follow-up time).

In this study we showed that histopathology for ER, PgR, and HER2 is highly concordant, but less concordant for Ki67 and NHG. Similarly, concordance between histopathology and the developed SGC and MGC models was high for ER, PgR, and HER2, and lower for Ki67 and NHG. Since the training labels for the models were based on histopathology, this result likely reflects the quality of the training data and the inherent variability within histopathology. Discordant results between classifiers and histopathology were associated with significant differences in patient overall survival in several biomarker and treatment groups. The MGC models have been integrated into the standard SCAN-B computational pipeline, and the classifications are part of preliminary RNA-seq-based clinical reports that can be automatically generated for each patient enrolled in SCAN-B [361, 362]. A pilot study for integrating SCAN-B reports into clinical practice was performed at Helsingborg Hospital in 2016 and included 113 patients.

## Study IV
## The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome

To expand the usefulness of RNA-seq beyond gene expression we strived to develop a bioinformatics approach to call somatic mutations in tumor-only RNA-seq datasets. Using 275 samples from 273 patients for which custom targeted capture tumor and normal DNA-seq data as well as RNA-seq data was available, we developed a computational pipeline for variant calling. By comparing DNA-seq and RNA-seq data, we optimized filters for removing germline calls and technical artifacts. Using the pipeline and filters we analyzed mutations in an independent population-based SCAN-B cohort of 3,217 tumors to describe the mutational landscape and relate mutations to patient overall survival (75 months median follow-up time).

Of the RNA-seq variants resulting from our 275 sample training cohort, 60.6% were identified as somatic in DNA, 17.0% as germline in DNA, and 22.4% as unique to RNA. The mutational landscape of the validation cohort was dominated by mutations in the genes *PIK3CA* and *TP53*. While mutation frequencies of oncogenes were comparable to previous DNA-based mutational profiling studies, we identified reduced mutation frequencies in tumor suppressor genes compared to DNA-based studies. Overall we identified mutations in genes with an existing drug targeting it in 86.6% of cases. Importantly we identified known treatment resistance mutations in the genes *ESR1* and *ERBB2* in early untreated breast cancer. Mutations were significantly associated with patient survival in several patient groups. To make our dataset useful for the wider research community we developed the web portal SCAN-B MutationExplorer, available at `https://oncogenomics.bmc.lu.se/MutationExplorer/`.

Building on the RNA-seq mutational profiling proof of concept work in study I, this study showed that RNA-seq mutational profiling is indeed feasible on a large scale. While tumor-only mutational profiling has several limitations, such as increased contamination by germline events, the overall mutational landscape was similar to previous studies on the DNA level. In particular the ability to detect known resistance mutations is clinically valuable and may be used to alter treatment regimens or increase surveillance for affected patients.

Similar to the work performed in study III, the computational pipeline defined during this study has been implemented in the standard SCAN-B workflow and mutations are now called for every patient enrolled in SCAN-B. We also extracted the pipeline from bcbio-nextgen into a stand-alone Snakemake [502] workflow that we have used in an unrelated project in renal cancer [503].

# 5 | Conclusions

*In a distant forest a wolf howled, felt embarrassed when no one joined in, and stopped.*

— Terry Pratchett, The Light Fantastic

The studies included in this thesis have helped advance the implementation of precision medicine within breast cancer care in Sweden as part of the SCAN-B infrastructure. With a focus on RNA-seq, SCAN-B has built a platform for large-scale transcriptome profiling, and we have evaluated current clinical biomarker assessment, developed and benchmarked expression-based tools for biomarker prediction, and described the mutational landscape of a large population-based primary cancer cohort. These findings provide the basis for clinical translation, and allow more advanced diagnostic tools to be developed in the future, for example through integration of gene expression and mutational data.

# 6 | Future Perspectives

*The phrase "Someone ought to do something" was not, by itself, a helpful one. People who used it never added the rider "and that someone is me".*

— Terry Pratchett, Hogfather

## Clinical Translation of Molecular Diagnostics

Molecular methods – particularly HTS – have taken the research world by storm and are increasingly being used in clinical decision-making. It is safe to assume that this adoption will continue as prices drop, while quality, read length, and sequencing speed increase. While the technical factors steadily improve, soft factors such as skilled personnel and training in how to interpret genomic information remain a limiting factor. The old issue of the "$1,000 genome but $100,000 analysis" [504] will continue for the time being, with analysis being difficult and demand for bioinformatics expertise being high. Leadership on all levels, not least on the clinical side, will be necessary to truly translate genomic methods such as expression-based subtyping into the clinical routine [505].

## SCAN-B

The SCAN-B project has come long way since its inception in 2009. Thousands of tumors have been profiled by RNA-seq, and many dozens of studies are underway that take advantage of this dataset and the SCAN-B biobank. Clinical implementation of the first genomic biomarkers is currently underway and will hopefully benefit patients in the near future. The use of true SSPs will be imperative for this purpose to achieve robust and reproducible classifications and predictions in a changing technological landscape.

The vision of precision medicine is to integrate data from as many layers as possible, for example patient characteristics, genome, methylome, transcriptome, proteome, and microbiome data, to make diagnoses/classifications/predictions as precise and accurate as possible. The completion of this vision is still ways ahead for both technological and economic reasons. Until then, RNA-seq may be a suitable proxy method that within a single analysis can profile the transcriptome and interrogate other layers such as DNA at least partially. The studies included in this thesis provide a first step in this direction. In the future we will hopefully see many more clinically meaningful tests, for example signatures for prediction of treatment response and resistance.

### Third Generation Sequencing

Third generation sequencing technologies such as the Pacific Biosciences SMRT and Oxford Nanopore platforms provide exciting research and diagnostic opportunities by enabling long read sequencing. Nanopore technology is particularly exciting in the context of transcriptome profiling, as it allows for direct sequencing of RNA [351–353] and direct detection of RNA modifications [354, 355]. These methods may help untangle the transcriptome and its involvement in oncogenesis.

### Bioinformatics

The field of bioinformatics is in an interesting situation. On the one hand it is indispensable for the life sciences and will be crucial for precision medicine to become a reality. On the other hand it suffers from lack of funding for maintenance of many crucial resources and software packages [298]. Further, lack of recognition and career options causes a drain of talent from academia to industry, or worse, other fields entirely. This problem is not necessarily unique to bioinformatics, but can be expanded to research software in general. Recently, private initiatives such as Essential Open Source Software for Science by the Chan Zuckerberg Initiative [506] have stepped in to fund several core research software projects, and research software engineering organizations have begun to form [507]. Taken together, these problems hint at failures on the side of universities and governments to provide adequate support for the field. This will need to be remedied for bioinformatics to advance and remain a reliable part of the life sciences.

From a technological point of view the adoption of graph genomes will be crucial to fully represent variation within populations and diseases such as cancer. These genome representations have the potential to alleviate reference allele bias and provide a more accurate way to represent complex structural events such as chromothripsis and non-trivial variation including the same allele between samples. For example one could envision a SCAN-B graph genome that represents the mutations in tumors of all enrolled patients.

### Liquid Biopsies

The development of liquid biopsy technologies, particularly using ctDNA, shows great promise for early cancer detection, detection of minimal residual disease, and monitoring of treatment response. Translation of these technologies into clinical practice as companion diagnostics will take well designed prospective clinical studies to validate their impact in improving patient treatment and survival.

# Acknowledgements

> *It would be nice, she thought wistfully, if*
> *someone could say 'thank you' occasionally.*
> — Terry Pratchett, Wyrd Sisters

A tremendous number of people have been involved – directly or indirectly – in the making of this thesis.

First of all, my thanks go out to the thousands of patients who enrolled in the SCAN-B and ABiM studies and chose to donate their biological material. Their donations laid the groundwork for this thesis and through the SCAN-B biobank and dozens of ongoing and future projects will help many patients worldwide.

*Lao Saal*, my main supervisor. He offered me to become a PhD student in his group, provided interesting research opportunities, supported me in my endeavours, and pushed me to push my projects further than I would have done on my own. He also did not mind my frequent ventures into the open source bioinformatics wilds.

*Håkan Axelson* and *Mattias Höglund*, my assistant supervisors, whose input was much appreciated.

*Åke Borg*, for initiating SCAN-B.

*Johan Vallon-Christersson* and *Jari Häkkinen*, for answering my many SCAN-B-related questions.

The *SCAN-B central laboratory* without whose hard work I would not have had data to analyze.

*Niklas Loman*, *Emma Niméus*, *Christian Ingvar*, *Håkan Olsson*, and *Dorthe Grabau* for letting me shadow them in the clinic, during surgery, and in pathology. Research – particularly the computational variety – can be very abstract and it is easy to forget that behind every datapoint is or was a human being. Those visits served as a reminder to always keep the patient in mind.

Current and former members of the extended Translational Oncogenomics group: *Yilun "Alan" Chen*, *Anthony George*, *Malin Dahlgren*, *Robert Rigo*, *Sergii Gladchuk*, *Heena Saini*, *Yoshiko Nanki*, *Christof Winter*, *Man-Hung Eric Tang*, *Eleonor Olsson*, *Barbara Lettiero*, *Stefano Misino*, *Willow Hight-Warburton*, *Michelle Lee*, *Madeline Dixon*, *Marina Villamor*, *Sofia Gruvberger-Saal*, and *Jill Howlin*.

*Daniel Filipazzi* and *Anders Kvist*, for keeping the servers running as long as possible (RIP *casa3*, *primolo{5–12}*, and *prime2*).

*Björn Frostner*, *Emelie Jalgén*, and *Susanne André* for administrative support.

All co-authors for their input, ideas, and work.

*Björn Canbäck*, for interesting discussions and teaching opportunities.

*John och Augusta Persson stiftelse*, *Maggie Stephens stiftelse*, the *Royal Physiographical Society in Lund*, and the *Lund University Medical Faculty PhD Student Travel Fund* for funding my conference travels.

The open source software community that makes the whole field of bioinformatics possible. In particular I would like to thank the hundreds of authors whose software contributed to this thesis, and the *Bioconda* and *conda-forge* communities for making this software available in a sane way.

The *FreeBSD*, *Biopython*, *Bioconda*, and *conda-forge* projects that allowed me to engage in productive procrastination.

*Merrilyn* and *David Hooley*, and the extended Hooley family for their support.

*Gabi* and *Markus*, my siblings, who bought a Commodore Plus/4 home computer in 1986 that planted the seed for my interest in computing.

*Hermann* und *Maria*, meine Eltern. Danke für alles.

*Emma*, for keeping me sane and/or caffeinated.

Lastly I would like to thank all Medicon Village staff who maintained and refilled the coffee machines, without whom this thesis would not have been possible.

*Christian Brüffer*
*Lund, December 2020*

# References

1. Murchison, E. P., Wedge, D. C., Alexandrov, L. B. *et al.* Transmissible dog cancer genome reveals the origin and history of an ancient cell lineage. *Science* **343,** 437–40 (2014) (cited on page 1).

2. Ekhtiari, S., Chiba, K., Popovic, S. *et al.* First case of osteosarcoma in a dinosaur: a multimodal diagnosis. *Lancet Oncology* **21,** 1021–1022 (2020) (cited on page 1).

3. Lee, C.-W., Efetova, M., Engelmann, J. C. *et al.* Agrobacterium tumefaciens Promotes Tumor Induction by Modulating Pathogen Defense in Arabidopsis thaliana. *Plant Cell* **21,** 2948–2962 (2009) (cited on page 1).

4. MacGregor, A. N. & Alexander, M. Formation of tumor-like structures on legume roots by Rhizobium. *Journal of Bacteriology* **105,** 728–732 (1971) (cited on page 1).

5. Monge, J., Kricun, M., Radovčić, J. *et al.* Fibrous Dysplasia in a 120,000+ Year Old Neandertal from Krapina, Croatia. *PLoS One* **8,** e64539 (2013) (cited on page 1).

6. Zink, A., Rohrbach, H., Szeimies, U. *et al.* Malignant tumors in an ancient Egyptian population. *Anticancer Research* **19,** 4273–4277 (1999) (cited on page 1).

7. Nerlich, A. G., Rohrbach, H., Bachmeier, B. *et al.* Malignant tumors in two ancient populations: An approach to historical tumor epidemiology. *Oncology Reports* **16,** 197–202 (2006) (cited on page 1).

8. Binder, M., Roberts, C., Spencer, N. *et al.* On the antiquity of cancer: Evidence for metastatic carcinoma in a young man from ancient Nubia (c. 1200bc). *PLoS One* **9,** e90924 (2014) (cited on page 1).

9. Domazet-Lošo, T., Klimovich, A., Anokhin, B. *et al.* Naturally occurring tumours in the basal metazoan Hydra. *Nature Communications* **5,** 4222 (2014) (cited on page 1).

10. Blackadar, C. B. Historical review of the causes of cancer. *World Journal of Clinical Oncology* **7,** 54 (2016) (cited on page 1).

11. Boveri, T. Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. *Verhandlungen der Physikalisch-Medizinischen Gesellschaft zu Würzburg* **35,** 67–90 (1902) (cited on page 1).

12. Hall, J. M., Lee, M. K., Newman, B. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250,** 1684–9 (1990) (cited on page 1).

13. Miki, Y., Swensen, J., Shattuck-Eidens, D. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994) (cited on page 1).

14. Wooster, R., Neuhausen, S. L., Mangion, J. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265,** 2088–90 (1994) (cited on page 1).

15. Wooster, R., Bignell, G., Lancaster, J. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–92 (1995) (cited on page 1).

16. Lander, E. S., Linton, L. M., Birren, B. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001) (cited on pages 1, 21).

17. Venter, J. C., Adams, M. D., Myers, E. W. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001) (cited on pages 1, 21).

18. World Health Organization. *International Classification of Diseases for Oncology (ICD-O) – 3rd Ed., 1st Rev.* `https://apps.who.int/iris/handle/10665/96612`, visited 2020-11-30 (cited on page 1).

19. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70 (2000) (cited on page 1).

20. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011) (cited on pages 1, 2).

21. Gerstung, M., Jolly, C., Leshchiner, I. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578,** 122–128 (2020) (cited on page 2).

22. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic aids: A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953) (cited on page 2).

23. Kimsey, I. J., Szymanski, E. S., Zahurancik, W. J. *et al.* Dynamic basis for dG•dT misincorporation via tautomerization and ionization. *Nature* **554,** 195–201 (2018) (cited on page 2).

24. Robinson, P. S., Cooren, T. H. H., Palles, C. *et al.* Elevated somatic mutation burdens in normal human cells due to defective DNA polymerases. *bioRxiv* (2020) (cited on page 2).

25. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009) (cited on page 2).

26. Akavia, U. D., Litvin, O., Kim, J. *et al.* An Integrated Approach to Uncover Drivers of Cancer. *Cell* **143,** 1005–17 (2010) (cited on page 2).

27. Gonzalez-Perez, A., Perez-llamas, C., Deu-Pons, J. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods* **10,** 1081–2 (2013) (cited on page 2).

28. Bailey, M. H., Tokheim, C., Porta-Pardo, E. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173,** 371–385 (2018) (cited on page 2).

29. Martínez-Jiménez, F., Muiños, F., Sentís, I. *et al.* A compendium of mutational cancer driver genes. *Nature Reviews Cancer* (2020) (cited on page 2).

30. Kumar, S., Warrel, J., Li, S. *et al.* Passenger mutations in 2500 cancer genomes: Overall molecular functional impact and consequences. *Cell* **180,** 915–927 (2020) (cited on page 2).

31. Dou, Y., Gold, H. D., Luquette, L. J. *et al.* Detecting Somatic Mutations in Normal Cells. *Trends in Genetics* **34,** 545–557 (2018) (cited on page 2).

32. García-Nieto, P. E., Morrison, A. J. & Fraser, H. B. The somatic mutation landscape of the human body. *Genome Biology* **20,** 298 (2019) (cited on page 2).

33. Martincorena, I., Roshan, A., Gerstung, M. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–6 (2015) (cited on pages 2, 3).

34. Tang, J., Fewings, E., Chang, D. *et al.* The genomic landscapes of individual melanocytes from human skin. *Nature* (2020) (cited on page 2).

35. Moore, L., Leongamornlert, D., Coorens, T. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580,** 640–646 (2020) (cited on page 2).

36. Martincorena, I., Fowler, J. C., Wabik, A. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362,** 911–917 (2018) (cited on pages 2, 3).

37. Lee-Six, H., Olafsson, S., Ellis, P. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574,** 532–537 (2019) (cited on pages 2, 3).

38. Lawson, A. R. J., Abascal, F., Coorens, T. H. H. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370,** 75–82 (2020) (cited on page 2).

39. Cereser, B., Tabassum, N., Del Bel Belluz, L. *et al.* Mutational landscapes of normal breast during age and pregnancy determine cancer risk. *bioRxiv* (2020) (cited on page 3).

40. Li, R., Du, Y., Chen, Z. *et al.* Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science* **370,** 82–89 (2020) (cited on page 3).

41. Salk, J. J., Loubet-Senear, K., Maritschnegg, E. *et al.* Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Reports* **28,** 132–144 (2019) (cited on page 3).

42. Risques, R. A. & Kennedy, S. R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genetics* **14,** e1007108 (2018) (cited on page 3).

43. Kennedy, S. R., Zhang, Y. & Risques, R. A. Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. *Trends in Cancer* **5,** 531–540 (2019) (cited on page 3).

44. Eilbeck, K., Lewis, S. E., Mungall, C. J. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology* **6,** R44 (2005) (cited on page 3).

45. Rheinbay, E., Nielsen, M. M., Abascal, F. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578,** 102–111 (2020) (cited on pages 3, 7).

46. Lebeuf-Taylor, E., McCloskey, N., Bailey, S. *et al.* The distribution of fitness effects among synonymous mutations in a gene under selection. *eLIFE* **8,** e45952 (2019) (cited on page 3).

47. Supek, F., Miñana, B., Valcárcel, J. *et al.* Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156,** 1324–1335 (2014) (cited on page 3).

48. Sharma, Y., Miladi, M., Dukare, S. *et al.* A pan-cancer analysis of synonymous mutations. *Nature Communications* **10,** 2569 (2019) (cited on page 3).

49. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12,** 363–376 (2011) (cited on page 4).

50. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics* **7,** 85–97 (2006) (cited on page 4).

51. Alkner, S., Tang, M.-H. E., Brueffer, C. *et al.* Contralateral breast cancer can represent a metastatic spread of the first primary tumor: determination of clonal relationship between contralateral breast cancers using next-generation whole genome sequencing. *Breast Cancer Research* **17,** 102 (2015) (cited on page 5).

52. Tang, M.-H. E., Dahlgren, M., Brueffer, C. *et al.* Remarkable similarities of chromosomal rearrangements between primary human breast cancers and matched distant metastases as revealed by whole-genome sequencing. *Oncotarget* **6,** 37169–37184 (2015) (cited on page 5).

53. Stephens, P. J., Greenman, C. D., Fu, B. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011) (cited on page 5).

54. Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–93 (2012) (cited on pages 5, 20).

55. Baca, S. C., Prandi, D., Lawrence, M. S. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153,** 666–677 (2013) (cited on page 5).

56. Hadi, K., Yao, X., Behr, J. M. *et al.* Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. *Cell* **183,** 197–210 (2020) (cited on page 5).

57. Olsson, E., Winter, C., George, A. *et al.* Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. *EMBO Molecular Medicine* **7,** 1034–1047 (2015) (cited on pages 5, 26, 60).

58. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: From mechanism to therapy. *Cell* **150,** 12–27 (2012) (cited on page 5).

59. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013) (cited on pages 5, 19, 20).

60. Burns, M. B., Lackey, L., Carpenter, M. a. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494,** 366–70 (2013) (cited on page 5).

61. Alexandrov, L. B., Ju, Y. S., Haase, K. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354,** 618–622 (2016) (cited on page 5).

62. Alexandrov, L. B., Kim, J., Haradhvala, N. J. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101 (2020) (cited on page 5).

63. Degasperi, A., Amarante, T. D., Czarnecki, J. *et al.* A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nature Cancer* **1,** 249–263 (2020) (cited on page 5).

64. Kucab, J. E., Zou, X., Morganella, S. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177,** 821–836 (2019) (cited on page 5).

65. Pich, O., Muiños, F., Lolkema, M. P. *et al.* The mutational footprints of cancer therapies. *Nature Genetics* **51,** 1732–1740 (2019) (cited on page 5).

66. Samstein, R. M., Lee, C.-h., Shoushtari, A. N. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics* **51,** 202–206 (2019) (cited on page 6).

67. Chalmers, Z. R., Connelly, C. F., Fabrizio, D. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine* **9,** 34 (2017) (cited on pages 6, 19).

68. Merino, D. M., McShane, L. M., Fabrizio, D. *et al.* Establishing guidelines to harmonize tumor mutational burden (TMB): In silico assessment of variation in TMB quantification across diagnostic platforms: Phase I of the Friends of Cancer Research TMB Harmonization Project. *Journal for ImmunoTherapy of Cancer* **8,** e000147 (2020) (cited on page 6).

69. Subbiah, V., Solit, D. B., Chan, T. A. *et al.* The FDA approval of pembrolizumab for adult and pediatric patients with tumor mutational burden (TMB) ≥10: a decision centered on empowering patients and their physicians. *Annals of Oncology* **31,** 1115–1118 (2020) (cited on page 6).

70. Prasad, V. & Addeo, A. The FDA approval of pembrolizumab for patients with TMB >10 mut/Mb: was it a wise decision? No. *Annals of Oncology* **31,** 1112–1114 (2020) (cited on page 6).

71. Gurjao, C., Tsukrov, D., Imakaev, M. *et al.* Limited evidence of tumour mutational burden as a biomarker of response to immunotherapy. *bioRxiv,* 260265 (2020) (cited on page 6).

72. Lindeboom, R. G., Vermeulen, M., Lehner, B. *et al.* The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nature Genetics* **51,** 1645–1651 (2019) (cited on pages 6, 7).

73. Litchfield, K., Reading, J. L., Lim, E. L. *et al.* Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. *Nature Communications* **11,** 3800 (2020) (cited on page 6).

74. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell* **43,** 853–866 (2011) (cited on page 6).

75. Xue, Z., Warren, R. L., Gibb, E. A. *et al.* Recurrent tumor-specific regulation of alternative polyadenylation of cancer-related genes. *BMC Genomics* **19,** 536 (2018) (cited on page 6).

76. Zhao, B. S. & He, C. Pseudouridine in a new era of RNA modifications. *Cell Research* **25,** 153–154 (2015) (cited on page 7).

77. Davalos, V., Blanco, S. & Esteller, M. SnapShot: Messenger RNA Modifications. *Cell* **174,** 498–498 (2018) (cited on page 7).

78. Boo, S. H. & Kim, Y. K. The emerging role of RNA modifications in the regulation of mRNA stability. *Experimental and Molecular Medicine* **52,** 400–408 (2020) (cited on page 7).

79. Wiener, D. & Schwartz, S. The epitranscriptome beyond m6A. *Nature Reviews Genetics* (2020) (cited on page 7).

80. Perou, C. M., Sørlie, T., Eisen, M. B. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000) (cited on pages 7, 13, 49).

81. Ross, D. T., Scherf, U., Eisen, M. B. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24,** 227–235 (2000) (cited on page 7).

82. Kahraman, A., Karakulak, T., Szklarczyk, D. *et al.* Pathogenic impact of transcript isoform switching in 1209 cancer samples covering 27 cancer types using an isoform-specific interaction network. *Scientific Reports* **10,** 14453 (2020) (cited on page 7).

83. Cherry, S. & Lynch, K. W. Alternative Splicing and Cancer: Insights, Opportunities and Challenges from an Expanding View of the Transcriptome. *Genes & Development* **34,** 1005–1016 (2020) (cited on page 7).

84. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative splicing in cancer — implications for care. *Nature Reviews Clinical Oncology* **17,** 457–474 (2020) (cited on page 7).

85. Monteuuis, G., Schmitz, U., Petrova, V. *et al.* Holding on to junk bonds: Intron retention in cancer and therapy. *Cancer Research* (2020) (cited on page 7).

86. Erson-Bensan, A. E. & Can, T. Alternative polyadenylation: Another foe in cancer. *Molecular Cancer Research* **14,** 507–517 (2016) (cited on page 7).

87. Slack, F. J. & Chinnaiyan, A. M. The Role of Non-coding RNAs in Oncology. *Cell* **179,** 1033–1055 (2019) (cited on page 7).

88. Liu, L., Wang, Q., Qiu, Z. *et al.* Noncoding RNAs: the shot callers in tumor immune escape. *Signal Transduction and Targeted Therapy* **5,** 102 (2020) (cited on page 7).

89. Eisenberg, E. & Levanon, E. Y. A-to-I RNA editing — immune protector and transcriptome diversifier. *Nature Reviews Genetics* **19,** 473–490 (2018) (cited on page 7).

90. Barbieri, I. & Kouzarides, T. Roles of RNA modifications in cancer. *Nature Reviews Cancer* **20,** 303–322 (2020) (cited on page 7).

91. Dong, Z. & Cui, H. The Emerging Roles of RNA Modifications in Glioblastoma. *Cancers* **12,** 736 (2020) (cited on page 7).

92. Popp, M. W. & Maquat, L. E. Nonsense-mediated mRNA Decay and Cancer. *Current Opinion in Genetics and Development* **48,** 44–50 (2018) (cited on page 7).

93. Harbeck, N., Cortes, J., Gnant, M. *et al.* Breast Cancer. *Nature Reviews Disease Primers* **5,** 66 (2019) (cited on page 7).

94. Ferzoco, R. M. & Ruddy, K. J. The Epidemiology of Male Breast Cancer. *Current Oncology Reports* **18,** 1 (2016) (cited on page 7).

95. Nilsson, C., Holmqvist, M., Bergkvist, L. *et al.* Similarities and differences in the characteristics and primary treatment of breast cancer in men and women - a population based study (Sweden). *Acta Oncologica* **50,** 1083–1088 (2011) (cited on page 7).

96. Johansson, I., Nilsson, C., Berglund, P. *et al.* High-resolution genomic profiling of male breast cancer reveals differences hidden behind the similarities with female breast cancer. *Breast Cancer Research and Treatment* **129,** 747–760 (2011) (cited on page 7).

97. Johansson, I., Ringnér, M. & Hedenfalk, I. The Landscape of Candidate Driver Genes Differs between Male and Female Breast Cancer. *PLoS One* **8,** e78299 (2013) (cited on page 7).

98. Hortobagyi, G. N. Breast Cancer: 45 Years of Research and Progress. *Journal of Clinical Oncology* **38,** 2454–2462 (2020) (cited on page 7).

99. Bray, F., Ferlay, J., Soerjomataram, I. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68,** 394–424 (2018) (cited on page 7).

100. Socialstyrelsen. *Cancer i siffror 2018* https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/statistik/2018-6-10.pdf, visited 2020-11-15 (cited on page 8).

101. Brenner, H. & Hakulinen, T. Very-Long-Term Survival Rates of Patients With Cancer. *Journal of Clinical Oncology* **20,** 4405–4409 (2002) (cited on pages 8, 25).

102. Pan, H., Gray, R., Braybrooke, J. *et al.* 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *New England Journal of Medicine* **377,** 1836–1846 (2017) (cited on page 8).

103. Socialstyrelsen. *Socialstyrelsen Cancer Statistics Database* https://sdb.socialstyrelsen.se/if_can/, visited 2020-11-16 (cited on page 9).

104. Hamajima, N., Hirose, K., Tajima, K. *et al.* Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology* **13,** 1141–1151 (2012) (cited on page 9).

105. Mørch, L. S., Skovlund, C. W., Hannaford, P. C. *et al.* Contemporary Hormonal Contraception and the Risk of Breast Cancer. *New England Journal of Medicine* **377,** 2228–2239 (2017) (cited on page 9).

106. Collaborative Group on Hormonal Factors in Breast Cancer. Type and timing of menopausal hormone therapy and breast cancer risk: individual participant meta-analysis of the worldwide epidemiological evidence. *The Lancet* **394,** 1159–1168 (2019) (cited on page 9).

107. Ewertz, M., Duffy, S. W., Adami, H. *et al.* Age at first birth, parity and risk of breast cancer: A meta-analysis of 8 studies from the Nordic countries. *International Journal of Cancer* **46,** 597–603 (1990) (cited on page 9).

108. Romieu, I., Scoccianti, C., Chajès, V. *et al.* Alcohol intake and breast cancer in the European prospective investigation into cancer and nutrition. *International Journal of Cancer* **137,** 1921–30 (2015) (cited on page 9).

109. Garaycoechea, J. I., Crossan, G. P., Langevin, F. *et al.* Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553,** 171–177 (2018) (cited on page 9).

110. Inoue-Choi, M., Sinha, R., Gierach, G. L. *et al.* Red and processed meat, nitrite, and heme iron intakes and postmenopausal breast cancer risk in the NIH-AARP Diet and Health Study. *International Journal of Cancer* **138,** 1609–1618 (2016) (cited on page 9).

111. Anderson, J., Darwis, N., Mackay, D. *et al.* Red and processed meat consumption and breast cancer: UK Biobank cohort study and meta-analysis. *European Journal of Cancer* **90,** 73–82 (2018) (cited on page 9).

112. Dossus, L., Boutron-Ruault, M. C., Kaaks, R. *et al.* Active and passive cigarette smoking and breast cancer risk: results from the EPIC cohort. *International Journal of Cancer* **134,** 1871–88 (2014) (cited on page 9).

113. Pearson-Stuttard, J., Zhou, B., Kontis, V. *et al.* Worldwide burden of cancer attributable to diabetes and high body-mass index: a comparative risk assessment. *The Lancet Diabetes & Endocrinology* **6,** e6–e15 (2018) (cited on page 9).

114. Iyengar, N. M., Arthur, R., Manson, J. E. *et al.* Association of Body Fat and Risk of Breast Cancer in Postmenopausal Women With Normal Body Mass Index. *JAMA Oncology* **5,** 155–163 (2019) (cited on page 9).

115. Bjørge, T., Häggström, C., Ghaderi, S. *et al.* BMI and weight changes and risk of obesity-related cancers: a pooled European cohort study. *International Journal of Epidemiology* **48**, 1872–1885 (2019) (cited on page 9).

116. Friedenreich, C. M., Neilson, H. K. & Lynch, B. M. State of the epidemiological evidence on physical activity and cancer prevention. *European Journal of Cancer* **46**, 2593–2604 (2010) (cited on page 9).

117. Wu, Y., Zhang, D. & Kang, S. Physical activity and risk of breast cancer: A meta-analysis of prospective studies. *Breast Cancer Research and Treatment* **137**, 869–882 (2013) (cited on page 9).

118. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820–823 (1971) (cited on page 9).

119. Nielsen, F. C., Van Overeem Hansen, T. & Sørensen, C. S. Hereditary breast and ovarian cancer: New genes in confined pathways. *Nature Reviews Cancer* **16**, 599–612 (2016) (cited on page 10).

120. Nilsson, M. P., Törngren, T., Henriksson, K. *et al.* BRCAsearch: written pre-test information and BRCA1/2 germline mutation testing in unselected patients with newly diagnosed breast cancer. *Breast Cancer Research and Treatment* **168**, 117–126 (2018) (cited on page 10).

121. Duffy, S. W., Tabár, L., Yen, A. M.-F. *et al.* Mammography Screening Reduces Rates of Advanced and Fatal Breast Cancers: Results in 549,091 Women. *Cancer* **126**, 2971–2979 (2020) (cited on page 10).

122. Løberg, M., Lousdal, M. L., Bretthauer, M. *et al.* Benefits and harms of mammography screening. *Breast Cancer Research* **17**, 63 (2015) (cited on page 10).

123. Johnson, K., Sarma, D. & Hwang, E. S. Lobular breast cancer series: Imaging. *Breast Cancer Research* **17**, 94 (2015) (cited on page 10).

124. Vourtsis, A. & Berg, W. A. Breast density implications and supplemental screening. *European Radiology* **29**, 1762–1777 (2019) (cited on page 10).

125. Kumar, V., Abbas, A., Fausto, N. *et al. Robbins and Cotran Pathologic Basis of Disease* 8th Ed (Elsevier Inc., 2009) (cited on page 11).

126. Makki, J. Diversity of breast carcinoma: Histological subtypes and clinical relevance. *Clinical Medicine Insights: Pathology* **8**, 23–31 (2015) (cited on page 11).

127. Elston, C. W. & Ellis, O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**, 403–410 (1991) (cited on page 11).

128. Rakha, E. A., El-Sayed, M. E., Lee, A. H. *et al.* Prognostic significance of nottingham histologic grade in invasive breast carcinoma. *Journal of Clinical Oncology* **26,** 3153–3158 (2008) (cited on page 11).

129. Rakha, E. A., Reis-Filho, J. S., Baehner, F. *et al.* Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research* **12,** 207 (2010) (cited on page 11).

130. Denoix, P. F. Enquête permanente dans les centres anticancéreux. *Bulletin. Institut National d'Hygiène* **1,** 70–75 (1946) (cited on page 11).

131. *AJCC Cancer Staging Manual* 8th Ed. (eds Amin, M. B., Edge, S., Greene, F. *et al.*) (2017) (cited on page 12).

132. Tarantino, P., Hamilton, E., Tolaney, S. M. *et al.* HER2-Low Breast Cancer: Pathological and Clinical Landscape. *Journal of Clinical Oncology* **38,** 1951–1962 (2020) (cited on page 13).

133. Bentzon, N., Düring, M., Rasmussen, B. B. *et al.* Prognostic effect of estrogen receptor status across age in primary breast cancer. *International Journal of Cancer* **122,** 1089–1094 (2008) (cited on page 13).

134. U.S. National Cancer Institute. *Surveillance, Epidemiology, and End Results (SEER) Program* `https://seer.cancer.gov/`, visited 2020-11-20 (cited on page 13).

135. Parise, C. A. & Caggiano, V. Breast Cancer Survival Defined by the ER/PR/HER2 Subtypes and a Surrogate Classification according to Tumor Grade and Immunohistochemical Biomarkers. *Journal of Cancer Epidemiology* **2014,** 469251 (2014) (cited on page 13).

136. Sørlie, T., Perou, C. M., Tibshirani, R. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 10869–10874 (2001) (cited on pages 13, 50).

137. Sørlie, T., Tibshirani, R., Parker, J. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 8418–8423 (2003) (cited on pages 13, 36, 49, 59).

138. Herschkowitz, J. I., Simin, K., Weigman, V. J. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology* **8,** R76 (2007) (cited on page 13).

139. Fougner, C., Bergholtz, H. & Norum, J. H. Re-definition of claudin-low as a breast cancer phenotype. *Nature Communications* **11,** 1787 (2020) (cited on page 13).

140. Hu, Z., Fan, C., Oh, D. S. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7,** 96 (2006) (cited on pages 13, 36, 49, 59).

141. Picornell, A. C., Echavarria, I., Alvarez, E. *et al.* Breast cancer PAM50 signature: Correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. *BMC Genomics* **20,** 452 (2019) (cited on page 13).

142. Weigelt, B., Hu, Z., He, X. *et al.* Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Research* **65,** 9155–9158 (2005) (cited on page 13).

143. Tobin, N. P., Harrell, J. C., Lövrot, J. *et al.* Molecular subtype and tumor characteristics of breast cancer metastases as assessed by gene expression significantly influence patient post-relapse survival. *Annals of Oncology* **26,** 81–88 (2015) (cited on page 13).

144. Prat, A., Cheang, M. C., Galván, P. *et al.* Prognostic Value of Intrinsic Subtypes in Hormone Receptor-Positive Metastatic Breast Cancer Treated With Letrozole With or Without Lapatinib. *JAMA oncology* **2,** 1287–1294 (2016) (cited on page 13).

145. Cejalvo, J. M., De Dueñas, E. M., Galván, P. *et al.* Intrinsic subtypes and gene expression profiles in primary and metastatic breast cancer. *Cancer Research* **77,** 2213–2221 (2017) (cited on pages 13, 20).

146. Holm, K., Hegardt, C., Staaf, J. *et al.* Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Research* **12,** R36 (2010) (cited on pages 13, 20).

147. Peppercorn, J., Perou, C. M. & Carey, L. A. Molecular subtypes in breast cancer evaluation and management: Divide and conquer. *Cancer Investigation* **26,** 1–10 (2008) (cited on page 14).

148. Parker, J. S., Mullins, M., Cheung, M. C. U. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27,** 1160–1167 (2009) (cited on pages 14, 15, 49, 50, 59).

149. Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology* **5,** 5–23 (2011) (cited on page 14).

150. Goldhirsch, A., Winer, E. P., Coates, A. S. *et al.* Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Annals of Oncology* **24,** 2206–2223 (2013) (cited on page 14).

151. Curigliano, G., Burstein, H. J., Winer, E. P. *et al.* De-escalating and escalating treatments for early-stage breast cancer: The St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. *Annals of Oncology* **28,** 1700–1712 (2017) (cited on page 14).

152. Ehinger, A., Malmström, P., Bendahl, P. O. *et al.* Histological grade provides significant prognostic information in addition to breast cancer subtypes defined according to St Gallen 2013. *Acta Oncologica* **56,** 68–74 (2017) (cited on page 14).

153. Guiu, S., Michiels, S., André, F. *et al.* Molecular subclasses of breast cancer: How do we define them? The IMPAKT 2012 working group statement. *Annals of Oncology* **23,** 2997–3006 (2012) (cited on page 14).

154. Prat, A., Pineda, E., Adamo, B. *et al.* Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast* **24,** S26–S35 (2015) (cited on page 14).

155. Lundgren, C., Bendahl, P.-O., Borg, Å. *et al.* Agreement between molecular subtyping and surrogate subtype classification: a contemporary population-based study of ER-positive/HER2-negative primary breast cancer. *Breast Cancer Research and Treatment* **178,** 459–467 (2019) (cited on pages 14, 29).

156. Rouzier, R., Perou, C. M., Symmans, W. F. *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical Cancer Research* **11,** 5678–5685 (2005) (cited on page 14).

157. Rody, A., Karn, T., Solbach, C. *et al.* The erbB2+ cluster of the intrinsic gene set predicts tumor response of breast cancer patients receiving neoadjuvant chemotherapy with docetaxel, doxorubicin and cyclophosphamide within the GEPARTRIO trial. *Breast* **16,** 235–240 (2007) (cited on page 14).

158. Prat, A., Fan, C., Fernández, A. *et al.* Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy. *BMC Medicine* **13,** 303 (2015) (cited on page 14).

159. Cejalvo, J. M., Pascual, T., Fernández-Martínez, A. *et al.* Clinical implications of the non-luminal intrinsic subtypes in hormone receptor-positive breast cancer. *Cancer Treatment Reviews* **67,** 63–70 (2018) (cited on page 14).

160. Fan, C., Oh, D. S., Wessels, L. *et al.* Concordance among Gene-Expression–Based Predictors for Breast Cancer. *New England Journal of Medicine* **355,** 560–569 (2006) (cited on page 14).

161. Wirapati, P., Sotiriou, C., Kunkel, S. *et al.* Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research* **10,** R65 (2008) (cited on page 14).

162. Huang, S., Murphy, L. & Xu, W. Genes and functions from breast cancer signatures. *BMC Cancer* **18**, 473 (2018) (cited on page 14).

163. Vallon-Christersson, J., Häkkinen, J., Hegardt, C. *et al.* Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. *Scientific Reports* **9**, 12184 (2019) (cited on pages 14, 29).

164. Sotiriou, C. & Pusztai, L. Gene-Expression Signatures in Breast Cancer. *New England Journal of Medicine* **360**, 790–800 (2009) (cited on page 14).

165. Prat, A., Ellis, M. J. & Perou, C. M. Practical implications of gene-expression-based assays for breast oncologists. *Nature Reviews Clinical Oncology* **9**, 48–57 (2012) (cited on page 14).

166. Matikas, A., Foukakis, T., Swain, S. *et al.* Avoiding over- and undertreatment in patients with resected node-positive breast cancer with the use of gene expression signatures: are we there yet? *Annals of Oncology* **30**, 1044–1050 (2019) (cited on page 14).

167. Fisher, B., Dignam, J., Wolmark, N. *et al.* Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer. *Journal of the National Cancer Institute* **89**, 1673–1682 (1997) (cited on page 14).

168. Paik, S., Shak, S., Tang, G. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* **351**, 2817–2826 (2004) (cited on page 14).

169. Sparano, J. A. & Paik, S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology* **26**, 721–728 (2008) (cited on page 14).

170. Van 't Veer, L. J., Dai, H., van de Vijver, M. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002) (cited on page 14).

171. Wallden, B., Storhoff, J., Nielsen, T. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Medical Genomics* **8**, 54 (2015) (cited on page 15).

172. Sparano, J. A. TAILORx: Trial Assigning Individualized Options for Treatment (Rx). *Clinical Breast Cancer* **7**, 347–350 (2006) (cited on page 15).

173. Cardoso, F., Piccart-Gebhart, M., Van't Veer, L. *et al.* The MINDACT trial: The first prospective clinical validation of a genomic tool. *Molecular Oncology* **1**, 246–251 (2007) (cited on page 15).

174. Wong, W. B., Ramsey, S. D., Barlow, W. E. *et al.* The value of comparative effectiveness research: Projected return on investment of the RxPONDER trial (SWOG S1007). *Contemporary Clinical Trials* **33,** 1117–1123 (2012) (cited on page 15).

175. Cardoso, F., van't Veer, L. J., Bogaerts, J. *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine* **375,** 717–729 (2016) (cited on page 15).

176. Stemmer, S. M., Steiner, M., Rizel, S. *et al.* Clinical outcomes in patients with node-negative breast cancer treated based on the recurrence score results: evidence from a large prospectively designed registry. *npj Breast Cancer* **3,** 33 (2017) (cited on page 15).

177. Sparano, J. A., Gray, R. J., Makower, D. F. *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *New England Journal of Medicine* **379,** 111–121 (2018) (cited on page 15).

178. Kun, Y., How, L. C., Hoon, T. P. *et al.* Classifying the estrogen receptor status of breast cancers by expression profiles reveals a poor prognosis subpopulation exhibiting high expression of the ERBB2 receptor. *Human Molecular Genetics* **12,** 3245–3258 (2003) (cited on pages 15, 51).

179. Gruvberger-Saal, S. K., Edén, P., Ringnér, M. *et al.* Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles. eng. *Molecular Cancer Therapeutics* **3,** 161–168 (Feb. 2004) (cited on pages 15, 51).

180. Roepman, P., Horlings, H. M., Krijgsman, O. *et al.* Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. *Clinical Cancer Research* **15,** 7003–7011 (2009) (cited on pages 15, 51).

181. Bastani, M., Vos, L., Asgarian, N. *et al.* A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. *PLoS One* **8,** e82144. (2013) (cited on pages 15, 51).

182. Wilson, T. R., Xiao, Y., Spoerke, J. M. *et al.* Development of a robust RNA-based classifier to accurately determine ER, PR, and HER2 status in breast cancer clinical samples. eng. *Breast Cancer Research and Treatment* **148,** 315–325 (Nov. 2014) (cited on pages 15, 51).

183. Viale, G., Slaets, L., Bogaerts, J. *et al.* High concordance of protein (by IHC), gene (by FISH; HER2 only), and microarray readout (by TargetPrint) of ER, PgR, and HER2: results from the EORTC 10041/BIG 03-04 MINDACT trial. *Annals of Oncology* **25,** 816–823 (2014) (cited on pages 15, 51).

184. Varga, Z., Lebeau, A., Bu, H. *et al.* An international reproducibility study validating quantitative determination of ERBB2, ESR1, PGR, and MKI67 mRNA in breast cancer using MammaTyper®. *Breast Cancer Research* **19,** 55 (2017) (cited on pages 15, 51).

185. Ivshina, A. V., George, J., Senko, O. *et al.* Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research* **66,** 10292–10301 (2006) (cited on page 15).

186. Sotiriou, C., Wirapati, P., Loi, S. *et al.* Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* **98,** 262–272 (2006) (cited on page 15).

187. Xiao-Jun, M., Salunga, R., Dahiya, S. *et al.* A five-gene molecular grade index and HOXB13.IL17BR are complementary prognostic factors in early stage breast cancer. *Clinical Cancer Research* **14,** 2601–2608 (2008) (cited on page 15).

188. Cardoso, F., Kyriakides, S., Ohno, S. *et al.* Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **30,** 1194–1220 (2019) (cited on pages 16–18).

189. Barker, A. D., Sigman, C. C., Kelloff, G. J. *et al.* I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology and Therapeutics* **86,** 97–100 (2009) (cited on page 17).

190. Gao, J. J., Cheng, J., Bloomquist, E. *et al.* CDK4/6 inhibitor treatment for patients with hormone receptor-positive, HER2-negative, advanced or metastatic breast cancer: a US Food and Drug Administration pooled analysis. *The Lancet Oncology* **21,** 250–260 (2020) (cited on page 17).

191. Juric, D., Rodon, J., Tabernero, J. *et al.* Phosphatidylinositol 3-Kinase $\alpha$–Selective Inhibition With Alpelisib (BYL719) in PIK3CA -Altered Solid Tumors: Results From the First-in-Human Study. *Journal of Clinical Oncology* **36,** 1291–1299 (2018) (cited on page 17).

192. André, F., Ciruelos, E., Rubovszky, G. *et al.* Alpelisib for PIK3CA-Mutated, Hormone Receptor–Positive Advanced Breast Cancer. *New England Journal of Medicine* **380,** 1929–1940 (2019) (cited on page 17).

193. Marra, A., Viale, G. & Curigliano, G. Recent advances in triple negative breast cancer: The immunotherapy era. *BMC Medicine* **17,** 90 (2019) (cited on page 19).

194. Zacharakis, N., Chinnasamy, H., Black, M. *et al.* Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nature Medicine* **24,** 724–730 (2018) (cited on page 19).

195. Chrétien, S., Zerdes, I., Bergh, J. *et al.* Beyond PD-1/PD-L1 inhibition: What the future holds for breast cancer immunotherapy. *Cancers* **11,** 628 (2019) (cited on page 19).

196. Postow, M. A., Sidlow, R. & Hellmann, M. D. Immune-related adverse events associated with immune checkpoint blockade. *New England Journal of Medicine* **378,** 158–168 (2018) (cited on page 19).

197. Wartewig, T., Kurgyis, Z., Keppler, S. *et al.* PD-1 is a haploinsufficient suppressor of T cell lymphomagenesis. *Nature* **552,** 121–125 (2017) (cited on page 19).

198. Ludin, A. & Zon, L. I. Cancer immunotherapy: The dark side of PD-1 receptor inhibition. *Nature* **552,** 41–42 (2017) (cited on page 19).

199. Hudson, T. J., Anderson, W., Aretz, A. *et al.* International network of cancer genome projects. *Nature* **464,** 993–8 (2010) (cited on page 19).

200. Sjöblom, T., Jones, S., Wood, L. D. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006) (cited on page 19).

201. Curtis, C., Shah, S. P., Chin, S.-F. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012) (cited on page 19).

202. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012) (cited on pages 19, 20).

203. Nik-Zainal, S., Davies, H., Staaf, J. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (2016) (cited on page 19).

204. Pereira, B., Chin, S.-F., Rueda, O. M. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications* **7,** 11479 (2016) (cited on page 19).

205. Staaf, J., Glodzik, D., Bosch, A. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nature Medicine* **25,** 1526–1533 (2019) (cited on pages 19, 29).

206. Campbell, P. J., Getz, G., Korbel, J. O. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93 (2020) (cited on page 19).

207. Nik-Zainal, S., Van Loo, P., Wedge, D. C. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012) (cited on page 19).

208. Shah, S. P., Roth, A., Goya, R. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–9 (2012) (cited on page 19).

209. Yates, L. R., Gerstung, M., Knappskog, S. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine* **21,** 751–759 (2015) (cited on page 19).

210. Kandoth, C., McLellan, M. D., Vandin, F. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–9 (2013) (cited on page 19).

211. Martínez-Saéz, O., Chic, N., Pascual, T. *et al.* Frequency and spectrum of PIK3CA somatic mutations in breast cancer. *Breast Cancer Research* **22,** 45 (2020) (cited on page 19).

212. Saal, L. H., Gruvberger-Saal, S. K., Persson, C. *et al.* Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair. *Nature Genetics* **40,** 102–107 (2008) (cited on page 19).

213. Jönsson, G., Staaf, J., Vallon-Christersson, J. *et al.* The retinoblastoma gene undergoes rearrangements in BRCA1-deficient basal-like breast cancer. *Cancer Research* **72,** 4028–36 (2012) (cited on page 19).

214. Stephens, P. J., McBride, D. J., Lin, M. L. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–1010 (2009) (cited on page 19).

215. Fimereli, D., Fumagalli, D., Brown, D. *et al.* Genomic hotspots but few recurrent fusion genes in breast cancer. *Genes, Chromosomes and Cancer* **57,** 331–338 (2018) (cited on pages 19, 20).

216. Edgren, H., Murumagi, A., Kangaspeska, S. *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology* **12,** R6 (2011) (cited on page 19).

217. Robinson, D. R., Kalyana-Sundaram, S., Wu, Y. M. *et al.* Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nature Medicine* **17,** 1646–1651 (2011) (cited on page 19).

218. Banerji, S., Cibulskis, K., Rangel-Escareno, C. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–9 (2012) (cited on pages 19, 20).

219. Matissek, K. J., Onozato, M. L., Sun, S. *et al.* Expressed Gene Fusions as Frequent Drivers of Poor Outcomes in Hormone Receptor Positive Breast Cancer. *Cancer Discovery* **8,** 336–353 (2018) (cited on page 19).

220. Persson, H., Søkilde, R., Häkkinen, J. *et al.* Frequent miRNA-convergent fusion gene events in breast cancer. *Nature Communications* **8,** 788 (2017) (cited on pages 19, 29).

221. Persson, H., Søkilde, R., Häkkinen, J. *et al.* Analysis of fusion transcripts indicates widespread deregulation of snoRNAs and their host genes in breast cancer. *International Journal of Cancer* **146,** 3343–3353 (2020) (cited on page 19).

222. Esteller, M., Silva, J. M., Dominguez, G. *et al.* Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *Journal of the National Cancer Institute* **92,** 564–569 (2000) (cited on page 20).

223. Glodzik, D., Bosch, A., Hartman, J. *et al.* Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. *Nature Communications* **11,** 3747 (2020) (cited on pages 20, 29).

224. Zhang, H. Y., Liang, F., Jia, Z. L. *et al.* PTEN mutation, methylation and expression in breast cancer patients. *Oncology Letters* **6,** 161–168 (2013) (cited on page 20).

225. Bardowell, S. A., Parker, J., Fan, C. *et al.* Differential methylation relative to breast cancer subtype and matched normal tissue reveals distinct patterns. *Breast Cancer Research and Treatment* **142,** 365–380 (2013) (cited on page 20).

226. Morganella, S., Alexandrov, L. B., Glodzik, D. *et al.* The topography of mutational processes in breast cancer genomes. *Nature Communications* **7,** 11383 (2016) (cited on page 20).

227. Glodzik, D., Morganella, S., Davies, H. *et al.* A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nature Genetics* **49,** 341–348 (2017) (cited on page 20).

228. Saal, L. H., Johansson, P., Holm, K. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 7564–7569 (2007) (cited on page 20).

229. Ciriello, G., Gatza, M. L., Beck, A. H. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163,** 506–519 (2015) (cited on page 20).

230. Toy, W., Shen, Y., Won, H. *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nature Genetics* **45,** 1439–1445 (2013) (cited on pages 20, 25).

231. Angus, L., Smid, M., Wilting, S. M. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nature Genetics* **51,** 1450–1458 (2019) (cited on page 20).

232. Priedigkeit, N., Hartmaier, R. J., Chen, Y. *et al.* Intrinsic Subtype Switching and Acquired ERBB2/HER2 Amplifications and Mutations in Breast Cancer Brain Metastases. *JAMA Oncology* **3,** 666–671 (2017) (cited on page 20).

233. Klebe, M., Fremd, C., Kriegsmann, M. *et al.* Frequent Molecular Subtype Switching and Gene Expression Alterations in Lung and Pleural Metastasis From Luminal A–Type Breast Cancer. *JCO Precision Oncology* **4**, 848–859 (2020) (cited on page 20).

234. Gibbs, R. A. The Human Genome Project changed everything. *Nature Reviews Genetics* **21**, 575–576 (2020) (cited on page 20).

235. Rosenfeld, J. A., Mason, C. E. & Smith, T. M. Limitations of the human reference genome for personalized genomics. *PLoS One* **7**, e40294 (2012) (cited on pages 20, 21).

236. Chen, N.-C., Solomon, B., Mun, T. *et al.* Reducing reference bias using multiple population reference genomes. *bioRxiv* (2020) (cited on page 20).

237. Yang, X., Lee, W. P., Ye, K. *et al.* One reference genome is not enough. *Genome Biology* **20**, 19–21 (2019) (cited on pages 20–22).

238. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nature Reviews Genetics* **21**, 243–254 (2020) (cited on page 20).

239. Paten, B., Novak, A. M., Eizenga, J. M. *et al.* Genome graphs and the evolution of genome inference. *Genome Research* **27**, 665–676 (2017) (cited on pages 21, 22).

240. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004) (cited on page 21).

241. Osoegawa, K., Mammoser, A. G., Wu, C. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Research* **11**, 483–496 (2001) (cited on page 21).

242. Tuzun, E., Sharp, A. J., Bailey, J. A. *et al.* Fine-scale structural variation of the human genome. *Nature Genetics* **37**, 727–732 (2005) (cited on page 21).

243. Church, D. M., Schneider, V. A., Graves, T. *et al.* Modernizing reference genome assemblies. *PLoS Biology* **9**, e1001091 (2011) (cited on page 21).

244. Schneider, V. A., Graves-Lindsay, T., Howe, K. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**, 849–864 (2017) (cited on page 21).

245. Altshuler, D. M., Durbin, R. M., Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012) (cited on page 21).

246. Auton, A., Abecasis, G. R., Altshuler, D. M. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015) (cited on page 21).

247. Choudhury, A., Aron, S., Botigué, L. R. *et al.* High-depth African genomes inform human migration and health. *Nature* **586,** 741–748 (2020) (cited on page 21).

248. Huddleston, J., Chaisson, M. J., Steinberg, K. M. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research* **27,** 677–685 (2017) (cited on pages 21, 22).

249. Sherman, R. M., Forman, J., Antonescu, V. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics* **51,** 30–35 (2019) (cited on pages 21, 60).

250. Kehr, B., Helgadottir, A., Melsted, P. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nature Genetics* **49,** 588–593 (2017) (cited on page 21).

251. Ebbert, M. T., Jensen, T. D., Jansen-West, K. *et al.* Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology* **20,** 97 (2019) (cited on page 21).

252. Jain, M., Olsen, H. E., Turner, D. J. *et al.* Linear Assembly of a Human Y Centromere. *Nature Biotechnology* **36,** 321–323 (2018) (cited on page 22).

253. Logsdon, G. A., Vollger, M. R., Hsieh, P. *et al.* The structure, function, and evolution of a complete human chromosome 8. *bioRxiv* (2020) (cited on page 22).

254. Miga, K. H., Koren, S., Rhie, A. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* (2020) (cited on page 22).

255. Ameur, A., Che, H., Martin, M. *et al.* De novo assembly of two swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data. *Genes* **9,** 486 (2018) (cited on page 22).

256. Nordin, J., Ameur, A., Lindblad-Toh, K. *et al.* SweHLA: the high confidence HLA typing bio-resource drawn from 1000 Swedish genomes. *European Journal of Human Genetics* **28,** 627–635 (2019) (cited on page 22).

257. Svensson, D., Rentoft, M., Dahlin, A. *et al.* A whole-genome sequenced control population in northern Sweden reveals subregional genetic differences. *PLoS One* **15,** e0237721 (2020) (cited on page 22).

258. Maretty, L., Jensen, J. M., Petersen, B. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548,** 87–91 (2017) (cited on page 22).

259. Yamaguchi-Kabata, Y., Nariai, N., Kawai, Y. *et al.* iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing. *Human Genome Variation* **2,** 15050 (2015) (cited on page 22).

260. Nagasaki, M., Kuroki, Y., Shibata, T. F. *et al.* Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. *Human Genome Variation* **6,** 27 (2019) (cited on page 22).

261. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016) (cited on page 22).

262. Shendure, J., Balasubramanian, S., Church, G. M. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550,** 345–353 (2017) (cited on page 22).

263. Liu, Y., Siejka-Zielińska, P., Velikova, G. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature Biotechnology* **37,** 424–429 (2019) (cited on page 22).

264. UK Competition & Market Authority. *Looking forward to the future: investigating the proposed acquisition of PacBio by Illumina* `https://www.gov.uk/government/publications/analysis-of-the-investigation-of-the-proposed-acquisition-of-pacbio-by-illumina/looking-forward-to-the-future-investigating-the-proposed-acquisition-of-pacbio-by-illumina`, visited 2020-08-26 (cited on page 22).

265. Mahmoud, M., Gobet, N., Cruz-dávalos, D. I. *et al.* Structural variant calling: the long and the short of it. *Genome Biology* **20,** 246 (2019) (cited on page 22).

266. Olsen, N. D., Wagner, J., McDaniel, J. *et al.* precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map Regions. *bioRxiv,* 380741 (2020) (cited on page 22).

267. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (2009) (cited on pages 23, 24, 39).

268. Ståhl, P. L., Salmén, F., Vickovic, S. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016) (cited on page 23).

269. Gonorazky, H. D., Naumenko, S., Ramani, A. K. *et al.* Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *The American Journal of Human Genetics* **104,** 466–483 (2019) (cited on page 24).

270. Murdock, D. R., Dai, H., Burrage, L. C. *et al.* Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *Journal of Clinical Investigation* (2020) (cited on page 24).

271. Lee, H., Huang, A. Y., Wang, L. k. *et al.* Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine* **22,** 490–499 (2020) (cited on page 24).

272. Schischlik, F., Jäger, R., Rosebrock, F. *et al.* Mutational landscape of the transcriptome offers putative targets for immunotherapy of myeloproliferative neoplasms. *Blood* **134,** 199–210 (2019) (cited on page 24).

273. Wong, M., Mayoh, C., Lau, L. M. S. *et al.* Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. *Nature Medicine* (2020) (cited on page 24).

274. Roychowdhury, S., Iyer, M. K., Robinson, D. R. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science Translational Medicine* **3,** 111ra121 (2011) (cited on page 24).

275. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M. *et al.* Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nature Reviews Genetics* **17,** 257–271 (2016) (cited on page 24).

276. Cieślik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics* **19,** 93–109 (2018) (cited on page 24).

277. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20,** 631–656 (2019) (cited on page 24).

278. Marco-Puche, G., Lois, S., Benítez, J. *et al.* RNA-Seq Perspectives to Improve Clinical Diagnosis. *Frontiers in Genetics* **10,** 1152 (2019) (cited on page 24).

279. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *American Journal of Human Genetics* **93,** 641–651 (2013) (cited on pages 24, 39, 46).

280. Horvath, A., Pakala, S. B., Mudvari, P. *et al.* Novel insights into breast cancer genetic variance through RNA sequencing. *Scientific Reports* **3,** 2256 (2013) (cited on pages 24, 39, 46).

281. Radenbaugh, A. J., Ma, S., Ewing, A. *et al.* RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One* **9,** e111516 (2014) (cited on pages 24, 39, 46).

282. Wilkerson, M. D., Cabanski, C. R., Sun, W. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Research* **42,** e107 (2014) (cited on pages 24, 39, 46).

283. Sheng, Q., Zhao, S., Li, C. I. *et al.* Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics* **107,** 163–169 (2016) (cited on pages 24, 46).

284. Guo, Y., Zhao, S., Sheng, Q. *et al.* The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *BMC Genomics* **18,** 690 (2017) (cited on pages 24, 39, 46).

285. Siegel, M. B., He, X., Hoadley, K. A. *et al.* Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *Journal of Clinical Investigation* **128,** 1371–1383 (2018) (cited on pages 24, 39, 46).

286. Neums, L., Suenaga, S., Beyerlein, P. *et al.* VaDiR: an integrated approach to Variant Detection in RNA. *GigaScience* **7,** 1–13 (2018) (cited on pages 24, 39, 46).

287. Patel, A. P., Tirosh, I., Trombetta, J. J. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014) (cited on pages 24, 39).

288. Crowley, J. J., Zhabotynsky, V., Sun, W. *et al.* Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics* **47,** 353–360 (2015) (cited on pages 24, 39).

289. Talevich, E. & Shain, A. H. CNVkit-RNA: Copy number inference from RNA-Sequencing data. *bioRxiv* (2018) (cited on pages 24, 39).

290. Flensburg, C., Oshlack, A. & Majewski, I. J. Detecting copy number alterations in RNA-Seq using SuperFreq. *bioRxiv* (2020) (cited on pages 24, 39).

291. Ma, C., Shao, M. & Kingsford, C. SQUID: Transcriptomic Structural Variation Detection from RNA-seq. *Genome Biology* **19,** 52 (2018) (cited on pages 24, 39).

292. Cmero, M., Schmidt, B., Majewski, I. J. *et al.* MINTIE : identifying novel structural and splice variants in transcriptomes using RNA-seq data. *bioRxiv* (2020) (cited on pages 24, 39).

293. Eck, R. V. & Dayhoff, M. O. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152,** 363–366 (1966) (cited on page 24).

294. Hogeweg, P. The roots of bioinformatics in theoretical biology. *PLoS Computational Biology* **7,** e1002021 (2011) (cited on page 24).

295. Quackenbush, J. Open-source software accelerates bioinformatics. *Genome Biology* **4,** 336 (2003) (cited on page 24).

296. Douglas, C., Goulding, R., Farris, L. *et al.* Socio-Cultural characteristics of usability of bioinformatics databases and tools. *Interdisciplinary Science Reviews* **36,** 55–71 (2011) (cited on page 24).

297. Prlić, A. & Procter, J. B. Ten Simple Rules for the Open Development of Scientific Software. *PLoS Computational Biology* **8,** e1002802 (2012) (cited on page 24).

298. Siepel, A. Challenges in funding and developing genomic software: roots and remedies. *Genome Biology* **20,** 147 (2019) (cited on pages 25, 66).

299. Chang, J. *Core services: Reward bioinformaticians* `https://www.nature.com/news/core-services-reward-bioinformaticians-1.17251`, visited 2020-11-24 (cited on page 25).

300. Lewis, J., Bartlett, A. & Atkinson, P. Hidden in the Middle: Culture, Value and Reward in Bioinformatics. *Minerva* **54,** 471–490 (2016) (cited on page 25).

301. Dragon, J. A., Gates, C., Sui, S. H. *et al.* Bioinformatics core survey highlights the challenges facing data analysis facilities. *Journal of Biomolecular Techniques* **31,** 66–73 (2020) (cited on page 25).

302. *Holland-Frei Cancer Medicine* 9th Ed. (eds Bast Jr, R. C., Croce, C. M., Halt, W. N. *et al.*) (Wiley Blackwell, 2016) (cited on page 25).

303. Masood, S. Focusing on breast cancer overdiagnosis and overtreatment: The promise of molecular medicine. *The Breast Journal* **19,** 127–129 (2013) (cited on page 25).

304. Gluz, O., Kolberg-Liedtke, C., Prat, A. *et al.* Efficacy of deescalated chemotherapy according to PAM50 subtypes, immune and proliferation genes in triple-negative early breast cancer: Primary translational analysis of the WSG-ADAPT-TN trial. *International Journal of Cancer* **146,** 262–271 (2020) (cited on page 25).

305. Gerlinger, M. Targeted drugs ramp up cancer mutability. *Science* **366,** 1452–1453 (2019) (cited on page 25).

306. Bose, R., Kavuri, S. M., Searleman, A. C. *et al.* Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discovery* **3,** 224–237 (2013) (cited on page 25).

307. Nayar, U., Cohen, O., Kapstad, C. *et al.* Acquired HER2 mutations in ER+ metastatic breast cancer confer resistance to estrogen receptor–directed therapies. *Nature Genetics* **51,** 207–216 (2018) (cited on pages 25, 26).

308. Mao, P., Cohen, O., Kowalski, K. J. *et al.* Acquired FGFR and FGF alterations confer resistance to estrogen receptor (ER) targeted therapy in ER+ metastatic breast cancer. *Clinical Cancer Research* **26,** 5974–5989 (2020) (cited on page 25).

309. Razavi, P., Dickler, M. N., Shah, P. D. *et al.* Alterations in PTEN and ESR1 promote clinical resistance to alpelisib plus aromatase inhibitors. *Nature Cancer* **1,** 382–393 (2020) (cited on page 25).

310. Robinson, D. R., Wu, Y. M., Vats, P. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature Genetics* **45,** 1446–51 (2013) (cited on page 26).

311. Dahlgren, M., George, A. M., Brueffer, C. *et al.* Pre-existing somatic mutations of estrogen receptor alpha (ESR1) in early-stage primary breast cancer. *Manuscript* (cited on page 26).

312. Lehmann, B. D., Bauer, J. A., Chen, X. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation* **121,** 2750–2767 (2011) (cited on page 26).

313. Vagia, E., Mahalingam, D. & Cristofanilli, M. The landscape of targeted therapies in TNBC. *Cancers* **12,** 916 (2020) (cited on page 26).

314. Goncalves, A., Mezni, E. & Bertucci, F. Combining poly(ADP-ribose) polymerase inhibitors and immune checkpoint inhibitors in breast cancer: rationale and preliminary clinical results. *Current Opinion in Oncology* **32,** 585–593 (2020) (cited on page 26).

315. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nature Reviews Cancer* **16,** 110–120 (2016) (cited on page 26).

316. Davies, H., Glodzik, D., Morganella, S. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine* **23,** 517–525 (2017) (cited on page 26).

317. Garcia-Murillas, I., Schiavon, G., Weigelt, B. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine* **7,** 302ra133 (2015) (cited on page 26).

318. Hyman, D. M., Piha-Paul, S. A., Won, H. *et al.* HER kinase inhibition in patients with HER2- and HER3-mutant cancers. *Nature* **554,** 189–194 (2018) (cited on page 26).

319. Drilon, A., Laetsch, T. W., Kummar, S. *et al.* Efficacy of Larotrectinib in TRK Fusion–Positive Cancers in Adults and Children. *New England Journal of Medicine* **378,** 731–739 (2018) (cited on page 26).

320. Prasad, V., Kaestner, V. & Mailankody, S. Cancer Drugs Approved Based on Biomarkers and Not Tumor Type—FDA Approval of Pembrolizumab for Mismatch Repair-Deficient Solid Cancers. *JAMA Oncology* **4,** 157–158 (2018) (cited on page 26).

321. Prasad, V. The precision-oncology illusion. *Nature Outlook* **537,** S63 (2016) (cited on pages 26, 27).

322. Marquart, J., Chen, E. Y. & Prasad, V. Estimation of The Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology. *JAMA Oncology* **4,** 1093–1098 (2018) (cited on page 27).

323. Prasad, V., De Jesús, K. & Mailankody, S. The high price of anticancer drugs: Origins, implications, barriers, solutions. *Nature Reviews Clinical Oncology* **14,** 381–390 (2017) (cited on page 27).

324. Tay-Teo, K., Ilbawi, A. & Hill, S. R. Comparison of Sales Income and Research and Development Costs for FDA-Approved Cancer Drugs Sold by Originator Drug Companies. *JAMA Network Open* **2,** e186875 (2019) (cited on page 27).

325. Payne, K., Gavan, S. P., Wright, S. J. *et al.* Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nature Reviews Genetics* **19,** 235–246 (2018) (cited on page 27).

326. Marino, P., Touzani, R., Perrier, L. *et al.* Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study. *European Journal of Human Genetics* **26,** 314–323 (2018) (cited on page 27).

327. Flaherty, K. T., Gray, R. J., Chen, A. P. *et al.* Molecular Landscape and Actionable Alterations in a Genomically Guided Cancer Clinical Trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH). *Journal of Clinical Oncology* **38** (2020) (cited on page 27).

328. Rydén, L., Loman, N., Larsson, C. *et al.* Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. *British Journal of Surgery* **105,** e158–e168 (2018) (cited on page 27).

329. Xie, Y., Lynn, B. C. D., Moir, N. *et al.* Breast cancer gene expression datasets do not reflect the disease at the population level. *npj Breast Cancer* (2020) (cited on page 29).

330. Søkilde, R., Persson, H., Ehinger, A. *et al.* Refinement of breast cancer molecular classification by miRNA expression profiles. *BMC Genomics* **20,** 503 (2019) (cited on page 29).

331. Axelsson, U., Rydén, L., Johnsson, P. *et al.* A multicenter study investigating the molecular fingerprint of psychological resilience in breast cancer patients: Study protocol of the SCAN-B resilience study. *BMC Cancer* **18,** 789 (2018) (cited on page 29).

332. Dihge, L., Vallon-Christersson, J., Hegardt, C. *et al.* Prediction of lymph node metastasis in breast cancer by gene expression and clinicopathological models: Development and validation within a population based cohort. *Clinical Cancer Research* **25,** 6368–6381 (2019) (cited on page 29).

333. Förnvik, D., Aaltonen, K. E., Chen, Y. *et al.* Detection of circulating tumor cells and circulating tumor DNA before and after mammographic breast compression in a cohort of breast cancer patients scheduled for neoadjuvant treatment. *Breast Cancer Research and Treatment* **177,** 447–455 (2019) (cited on pages 29, 48).

334. Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research* **6,** 639–45 (1996) (cited on page 34).

335. Wu, C., Carta, R. & Zhang, L. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Research* **33,** e84 (2005) (cited on page 35).

336. Wei, T., Pearson, M. N., Armstrong, K. *et al.* Analysis of crucial factors resulting in microarray hybridization failure. *Molecular BioSystems* **8,** 1325–38 (2012) (cited on page 35).

337. Walsh, C., Hu, P., Batt, J. *et al.* Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays* **4,** 389–406 (2015) (cited on page 35).

338. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007) (cited on pages 35, 48).

339. Espín-Pérez, A., Portier, C., Chadeau-Hyam, M. *et al.* Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data. *PLoS One* **13,** e0202947 (2018) (cited on page 35).

340. Aittokallio, T. Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Briefings in Bioinformatics* **11,** 253–264 (2009) (cited on page 35).

341. Illumina Inc. *Illumina HumanHT-12 v4 BeadChip Product Information Sheet* `https://www.illumina.com/documents/products/product_information_sheets/product_info_humanht-12.pdf`, visited 2020-08-06 (cited on page 36).

342. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Molecular Cell* **58,** 586–597 (2015) (cited on page 36).

343. Chaitankar, V., Karakülah, G., Ratnapriya, R. *et al.* Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in Retinal and Eye Research* **55**, 1–31 (2016) (cited on page 37).

344. Ewing, B., Hillier, L. D., Wendl, M. C. *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**, 175–85 (1998) (cited on page 38).

345. Ewing, B., Hillier, L., Wendl, M. C. *et al.* Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186–194 (1998) (cited on page 38).

346. Illumina Inc. *Technical Note: Quality Scores for Next-Generation Sequencing* `https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf`, visited 2020-11-24 (cited on page 38).

347. Andrews, S. *Illumina 2 colour chemistry can overcall high confidence G bases* `https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases/`, visited 2020-11-25 (cited on page 38).

348. Vitting-Seerup, K. & Sandelin, A. The Landscape of Isoform Switches in Human Cancers. *Molecular Cancer Research* **15**, 1206–1220 (2017) (cited on page 39).

349. Ozsolak, F., Platt, A. R., Jones, D. R. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009) (cited on page 39).

350. Ozsolak, F. & Milos, P. M. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdisciplinary Reviews: RNA* **2**, 565–570 (2011) (cited on page 39).

351. Garalde, D. R., Snell, E. A., Jachimowicz, D. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods* **15**, 201–206 (2018) (cited on pages 39, 66).

352. Soneson, C., Yao, Y., Bratus-Neuenschwander, A. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nature Communications* **10**, 3359 (2019) (cited on pages 39, 66).

353. Hardwick, S. A., Joglekar, A., Flicek, P. *et al.* Getting the entire message: Progress in isoform sequencing. *Frontiers in Genetics* **10**, 709 (2019) (cited on pages 39, 66).

354. Leger, A., Amaral, P., Pandolfini, L. *et al.* RNA modifications detection by comparative Nanopore direct RNA sequencing. *bioRxiv* (2019) (cited on pages 39, 66).

355. Stephenson, W., Razaghi, R., Busan, S. *et al.* Direct detection of RNA modifications and structure using single molecule nanopore sequencing. *bioRxiv* (2020) (cited on pages 39, 66).

356. Parkhomchuk, D., Borodina, T., Amstislavskiy, V. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research* **37,** e123 (2009) (cited on page 39).

357. Levin, J. Z., Yassour, M., Adiconis, X. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7,** 709–715 (2010) (cited on page 39).

358. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* **40,** e72 (2012) (cited on pages 40, 47).

359. Ross, M. G., Russ, C., Costello, M. *et al.* Characterizing and measuring bias in sequence data. *Genome Biology* **14,** R51 (2013) (cited on pages 40, 47).

360. Saal, L. H., Troein, C., Vallon-Christersson, J. *et al.* BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biology* **3,** SOFTWARE0003 (2002) (cited on page 41).

361. Vallon-Christersson, J., Nordborg, N., Svensson, M. *et al.* BASE–2nd generation software for microarray data management and analysis. *BMC Bioinformatics* **10,** 330 (2009) (cited on pages 41, 61).

362. Häkkinen, J., Nordborg, N., Månsson, O. *et al.* Implementation of an Open Source Software solution for Laboratory Information Management and automated RNAseq data analysis in a large-scale Cancer Genomics initiative using BASE with extension package Reggie. *bioRxiv,* 038976 (2016) (cited on pages 41, 61).

363. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–20 (2014) (cited on page 41).

364. Smit, AFA and Hubley, R and Green, P. *RepeatMasker Open-4.0* `http://www.repeatmasker.org` (cited on page 42).

365. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012) (cited on page 42).

366. Kim, D., Pertea, G., Trapnell, C. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14,** R36 (2013) (cited on pages 42, 44, 60).

367. Trapnell, C., Williams, B. A., Pertea, G. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28,** 511–515 (2010) (cited on pages 42, 45).

368. Roberts, A., Trapnell, C., Donaghey, J. *et al.* Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology* **12** (2011) (cited on pages 42, 45).

369. Kim, D., Paggi, J. M., Park, C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37,** 907–915 (2019) (cited on pages 42, 44).

370. Pertea, M., Pertea, G. M., Antonescu, C. M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33,** 290–295 (2015) (cited on page 42).

371. Conesa, A., Madrigal, P., Tarazona, S. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17,** 13 (2016) (cited on pages 42, 45).

372. bcbio-nextgen. `https://github.com/bcbio/bcbio-nextgen` (cited on pages 43, 60).

373. Picard. `https://broadinstitute.github.io/picard` (cited on pages 43, 44, 48).

374. Del Fabbro, C., Scalabrin, S., Morgante, M. *et al.* An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One* **8,** e85024 (2013) (cited on page 43).

375. Liao, Y. & Shi, W. Read trimming is not required for mapping and quantification of RNA-seq reads. *NAR Genomics and Bioinformatics* **2,** lqaa068 (2020) (cited on page 43).

376. MacManes, M. D. On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics* **5,** 13 (2014) (cited on page 43).

377. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009) (cited on page 44).

378. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12,** 357–360 (2015) (cited on page 44).

379. Dobin, A., Davis, C. A., Schlesinger, F. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013) (cited on page 44).

380. Li, H., Handsaker, B., Wysoker, A. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009) (cited on pages 44, 60).

381. Burriesci, M. S., Lehnert, E. M. & Pringle, J. R. Fulcrum: Condensing redundant reads from high-throughput sequencing studies. *Bioinformatics* **28,** 1324–1327 (2012) (cited on page 44).

382. Tarasov, A., Vilella, A. J., Cuppen, E. *et al.* Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–2034 (2015) (cited on page 44).

383. Tischler, G. & Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* **9** (2014) (cited on page 44).

384. Faust, G. G. & Hall, I. M. SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics* **30,** 2503–2505 (2014) (cited on page 44).

385. Parekh, S., Ziegenhain, C., Vieth, B. *et al.* The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports* **6,** 25533 (2016) (cited on page 45).

386. Quinn, E. M., Cormican, P., Kenny, E. M. *et al.* Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS One* **8,** e58815 (2013) (cited on page 45).

387. Kivioja, T., Vähärautio, A., Karlsson, K. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9,** 72–74 (2012) (cited on page 45).

388. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research* **27,** 491–499 (2017) (cited on page 45).

389. umis. `https://github.com/vals/umis` (cited on page 45).

390. fgbio. `https://github.com/fulcrumgenomics/fgbio` (cited on page 45).

391. Islam, S., Zeisel, A., Joost, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11,** 163–166 (2014) (cited on page 45).

392. Bray, N. L., Pimentel, H., Melsted, P. *et al.* Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34,** 525–527 (2016) (cited on page 45).

393. Patro, R., Duggal, G., Love, M. I. *et al.* Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14,** 417–419 (2017) (cited on page 45).

394. Dillies, M.-A., Rau, A., Aubert, J. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* **14,** 671–83 (2013) (cited on page 45).

395. Li, X., Brock, G. N., Rouchka, E. C. *et al.* A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. *PLoS One* **12,** e0176185 (2017) (cited on page 45).

396. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics* **19**, 776–792 (2018) (cited on page 45).

397. Mortazavi, A., Williams, B. A., McCue, K. *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008) (cited on page 45).

398. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* **131**, 281–285 (2012) (cited on page 45).

399. Roberts, N. D., Kortschak, R. D., Parker, W. T. *et al.* A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* **29**, 2223–2230 (2013) (cited on page 46).

400. Xu, H., DiCarlo, J., Satya, R. V. *et al.* Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 244 (2014) (cited on page 46).

401. Alioto, T. S., Buchhalter, I., Derdak, S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6**, 10001 (2015) (cited on page 46).

402. Cai, L., Yuan, W., Zhang, Z. *et al.* In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports* **6**, 36540 (2016) (cited on page 46).

403. Krøigård, A. B., Thomassen, M., Lænkholm, A. V. *et al.* Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One* **11**, e0151664 (2016) (cited on page 46).

404. Sandmann, S., De Graaf, A. O., Karimi, M. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports* **7**, 43169 (2017) (cited on page 46).

405. Bian, X., Zhu, B., Wang, M. *et al.* Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **19**, 429 (2018) (cited on page 46).

406. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal* **16**, 15–24 (2018) (cited on page 46).

407. Chen, Z., Yuan, Y., Chen, X. *et al.* Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports* **10,** 3501 (2020) (cited on page 46).

408. Koboldt, D. C., Zhang, Q., Larson, D. E. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22,** 568–76 (2012) (cited on page 46).

409. Lai, Z., Markovets, A., Ahdesmaki, M. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research* **44,** e108 (2016) (cited on page 46).

410. Benjamin, D., Sato, T., Cibulskis, K. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* (2019) (cited on page 46).

411. Zook, J. M., Catoe, D., McDaniel, J. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* **3,** 160025 (2016) (cited on page 46).

412. Chapman, L. M., Spies, N., Pai, P. *et al.* A crowdsourced set of curated structural variants for the human genome. *PLoS Computational Biology* **16,** e1007933 (2020) (cited on page 46).

413. Quaglieri, A., Flensburg, C., Speed, T. P. *et al.* Finding a suitable library size to call variants in RNA-seq. *BMC Bioinformatics* **21,** 553 (2020) (cited on page 46).

414. Coudray, A., Battenhouse, A. M., Bucher, P. *et al.* Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ* **6,** e5362 (2018) (cited on page 46).

415. Yizhak, K., Aguet, F., Kim, J. *et al.* A comprehensive analysis of RNA sequences reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364,** eaaw0726 (2019) (cited on page 46).

416. Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Reviews Genetics* **19,** 269–285 (2018) (cited on page 47).

417. Brodin, J., Mild, M., Hedskog, C. *et al.* PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data. *PLoS One* **8,** e70388 (2013) (cited on page 47).

418. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research* **43,** e143 (2015) (cited on page 47).

419. Potapov, V. & Ong, J. L. Examining sources of error in PCR by single-molecule sequencing. *PLoS One* **12,** e0169774 (2017) (cited on page 47).

420. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology* **12,** R112 (2011) (cited on page 47).

421. Schirmer, M., D'Amore, R., Ijaz, U. Z. *et al.* Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17,** 125 (2016) (cited on page 47).

422. Sherry, S. T., Ward, M., Kholodov, M. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29,** 308–311 (2001) (cited on page 47).

423. Karczewski, K. J., Francioli, L. C., Tiao, G. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (2020) (cited on page 47).

424. Forbes, S. A., Beare, D., Gunasekaran, P. *et al.* COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43,** D805–D811 (2015) (cited on pages 47, 48).

425. Sondka, Z., Bamford, S., Cole, C. G. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18,** 696–705 (2018) (cited on page 47).

426. Griffith, M., Spies, N. C., Krysiak, K. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics* **49,** 170–174 (2017) (cited on pages 47, 48).

427. Adzhubei, I. A., Schmidt, S., Peshkin, L. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–9 (2010) (cited on page 47).

428. Cingolani, P., Platts, A., Coon, M. *et al.* 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) (cited on page 47).

429. McLaren, W., Gil, L., Hunt, S. E. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17,** 122 (2016) (cited on page 47).

430. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38,** e164 (Sept. 2010) (cited on page 47).

431. Pedersen, B. S., Layer, R. M., Quinlan, A. R. *et al.* Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology* **17,** 118 (2016) (cited on page 47).

432. Ameur, A., Dahlberg, J., Olason, P. *et al.* SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics* **25,** 1253–1260 (2017) (cited on page 47).

433. Barnell, E. K., Ronning, P., Campbell, K. M. *et al.* Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genetics in Medicine* **21,** 972–981 (2019) (cited on page 48).

434. Danos, A. M., Krysiak, K., Barnell, E. K. *et al.* Standard operating procedure for curation and clinical interpretation of variants in cancer. *Genome Medicine* **11,** 76 (2019) (cited on page 48).

435. Ståhlberg, A., Krzyzanowski, P. M., Jackson, J. B. *et al.* Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic Acids Research* **44,** e105 (2016) (cited on page 48).

436. Stasik, S., Schuster, C., Ortlepp, C. *et al.* An optimized targeted Next-Generation Sequencing approach for sensitive detection of single nucleotide variants. *Biomolecular Detection and Quantification* **15,** 6–12 (2018) (cited on page 48).

437. Arildsen, N. S., Martin de la Fuente, L., Måsbäck, A. *et al.* Detecting TP53 mutations in diagnostic and archival liquid-based Pap samples from ovarian cancer patients using an ultra-sensitive ddPCR method. *Scientific Reports* **9,** 15506 (2019) (cited on page 48).

438. Isaksson, S., George, A. M., Jönsson, M. *et al.* Pre-operative plasma cell-free circulating tumor DNA and serum protein tumor markers as predictors of lung adenocarcinoma recurrence. *Acta Oncologica* **58,** 1079–1086 (2019) (cited on page 48).

439. Pettersson, L., Chen, Y., George, A. M. *et al.* Subclonal patterns in follow-up of acute myeloid leukemia combining whole exome sequencing and ultrasensitive IBSAFE digital droplet analysis. *Leukemia and Lymphoma* **61,** 2168–2179 (2020) (cited on page 48).

440. Sheng, Q., Vickers, K., Zhao, S. *et al.* Multi-perspective quality control of Illumina RNA sequencing data analysis. *Briefings in Functional Genomics* **16,** 194–204 (2017) (cited on page 48).

441. FastQC. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/` (cited on page 48).

442. DeLuca, D. S., Levin, J. Z., Sivachenko, A. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28,** 1530–2 (2012) (cited on page 48).

443. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32,** 292–294 (2016) (cited on page 48).

444. Brueffer, C. *Quality Control and Analysis of RNA-seq Data from Breast Cancer Tumor Samples* MA thesis (Lund University, Department of Biology, July 2013) (cited on page 48).

445. Leek, J. T., Johnson, W. E., Parker, H. S. *et al.* The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012) (cited on page 48).

446. Lauss, M., Visne, I., Kriegner, A. *et al.* Monitoring of technical variation in quantitative high-throughput datasets. *Cancer Informatics* **12,** 193–201 (2013) (cited on page 48).

447. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research* **42,** e161 (2014) (cited on page 48).

448. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-Seq: batch effect adjustment for RNA-Seq count data. *NAR Genomics and Bioinformatics* **2,** 1 (2020) (cited on page 48).

449. Winter, C., Nilsson, M., Olsson, E. *et al.* Targeted sequencing of BRCA1 and BRCA2 across a large unselected breast cancer cohort suggests one third of mutations are somatic. *Annals of Oncology* **27,** 1532–1538 (2016) (cited on page 49).

450. Hofmann, A. L., Behr, J., Singer, J. *et al.* Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics* **18,** 8 (2017) (cited on page 49).

451. Ellrott, K., Bailey, M. H., Saksena, G. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems* **6,** 271–281 (2018) (cited on page 49).

452. Shi, W., Ng, C. K., Lim, R. S. *et al.* Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. *Cell Reports* **25,** 1446–1457 (2018) (cited on page 49).

453. Haibe-Kains, B., Desmedt, C., Loi, S. *et al.* A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute* **104,** 311–325 (2012) (cited on page 49).

454. Paquet, E. R. & Hallett, M. T. Absolute assignment of breast cancer intrinsic molecular subtype. *Journal of the National Cancer Institute* **107,** 357 (2015) (cited on pages 49, 50).

455. Tibshirani, R., Hastie, T., Narasimhan, B. *et al.* Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 6567–6572 (2002) (cited on pages 49, 53, 61).

456. Ali, H. R., Rueda, O. M., Chin, S.-F. *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology* **15,** 431 (2014) (cited on page 49).

457. Perou, C. M., Parker, J. S., Prat, A. *et al.* Clinical implementation of the intrinsic subtypes of breast cancer. *The Lancet Oncology* **11,** 718–719 (2010) (cited on page 49).

458. Staaf, J. & Ringnér, M. Making breast cancer molecular subtypes robust? *Journal of the National Cancer Institute* **107,** 21–22 (2015) (cited on page 49).

459. Weigelt, B., Mackay, A., A'hern, R. *et al.* Breast cancer molecular profiling with single sample predictors: A retrospective analysis. *The Lancet Oncology* **11,** 339–349 (2010) (cited on pages 49, 50).

460. MacKay, A., Weigelt, B., Grigoriadis, A. *et al.* Microarray-based class discovery for molecular classification of breast cancer: Analysis of interobserver agreement. *Journal of the National Cancer Institute* **103,** 662–673 (2011) (cited on pages 49, 50).

461. Sørlie, T., Borgan, E., Myhre, S. *et al.* The importance of gene-centring microarray data. *The Lancet Oncology* **11,** 719–720 (2010) (cited on page 50).

462. Weigelt, B., Mackay, A., Natrajan, R. *et al.* The importance of gene-centring microarray data - Authors' reply. *The Lancet Oncology* **11,** 720–721 (2010) (cited on page 50).

463. Lusa, L., McShane, L. M., Reid, J. F. *et al.* Challenges in projecting clustering results across gene expression-profiling datasets. *Journal of the National Cancer Institute* **99,** 1715–1723 (2007) (cited on page 50).

464. Franks, J. M., Cai, G. & Whitfield, M. L. Feature Specific Quantile Normalization Enables Cross-Platform Classification of Molecular Subtypes using Gene Expression Data. *Bioinformatics* **34,** 1868–1874 (2018) (cited on page 50).

465. Raj-Kumar, P. K., Liu, J., Hooke, J. A. *et al.* PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Scientific Reports* **9,** 7956 (2019) (cited on page 50).

466. Cascianelli, S., Molineris, I., Isella, C. *et al.* Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Scientific Reports* **10,** 14071 (2020) (cited on page 50).

467. Chen, R., Yang, L., Goodison, S. *et al.* Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics* **36,** 1476–1483 (2020) (cited on page 50).

468. Yu, Z., Wang, Z., Yu, X. *et al.* RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches. *Computational Intelligence and Neuroscience* **2020,** 4737969 (2020) (cited on page 50).

469. Gendoo, D. M. and Ratanasirigulchai, N. and Schroeder, M. S. and Pare, L. and Parker, Joel S. and Prat, A. and Haibe-Kains, B. genefu: Computation of Gene Expression-Based Signatures in Breast Cancer. `http://www.pmgenomics.ca/bhklab/software/genefu` (cited on page 50).

470. Meyerholz, D. K. & Beck, A. P. Principles and approaches for reproducible scoring of tissue stains in research. *Laboratory Investigation* **98,** 844–855 (2018) (cited on page 50).

471. Rakha, E. A., Pinder, S. E., Bartlett, J. M. *et al.* Updated UK recommendations for HER2 assessment in breast cancer. *Journal of Clinical Pathology* **68,** 93–99 (2015) (cited on page 51).

472. Leung, S. C., Nielsen, T. O., Zabaglo, L. A. *et al.* Analytical validation of a standardised scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicentre collaboration. *Histopathology* **75,** 225–235 (2019) (cited on page 51).

473. Rizzardi, A. E., Johnson, A. T., Vogel, R. I. *et al.* Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. *Diagnostic Pathology* **7,** 42 (2012) (cited on page 51).

474. Mungle, T., Tewary, S., Arun, I. *et al.* Automated characterization and counting of Ki-67 protein for breast cancer prognosis: A quantitative immunohistochemistry approach. *Computer Methods and Programs in Biomedicine* **139,** 149–161 (2017) (cited on page 51).

475. Van Eycke, Y. R., Allard, J., Salmon, I. *et al.* Image processing in digital pathology: An opportunity to solve inter-batch variability of immunohistochemical staining. *Scientific Reports* **7,** 42964 (2017) (cited on page 51).

476. Naik, N., Madani, A., Esteva, A. *et al.* Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nature Communications* **11,** 5727 (2020) (cited on page 51).

477. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* (2020) (cited on page 51).

478. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**, 321–332 (2015) (cited on page 51).

479. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface* **15**, 20170387 (2018) (cited on page 51).

480. Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* **8**, 1341–1390 (1996) (cited on page 51).

481. Lynam, A. L., Dennis, J. M., Owen, K. R. *et al.* Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagnostic and Prognostic Research* **4**, 6 (2020) (cited on page 51).

482. pamr. `https://cran.r-project.org/web/packages/pamr/` (cited on page 53).

483. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1–26 (2008) (cited on page 53).

484. Viera, A. J. & Garrett, J. M. Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* **37**, 360–363 (2005) (cited on pages 54, 55).

485. Cao, C., Chicco, D. & Hoffman, M. M. The MCC-F1 curve: a performance evaluation technique for binary classification. *arXiv* (2020) (cited on page 54).

486. Brodersen, K. H., Ong, C. S., Stephan, K. E. *et al. The balanced accuracy and its posterior distribution* in *Proceedings of the 20th International Conference on Pattern Recognition* (2010) (cited on page 54).

487. Lantz, B. The large sample size fallacy. *Scandinavian Journal of Caring Sciences* **27**, 487–492 (2013) (cited on page 56).

488. Gourgou-Bourgade, S., Cameron, D., Poortmans, P. *et al.* Guidelines for time-to-event end point definitions in breast cancer trials: Results of the DATECAN initiative (Definition for the Assessment of Time-to-event Endpoints in CANcer trials). *Annals of Oncology* **26**, 873–879 (2015) (cited on page 56).

489. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457–481 (1958) (cited on page 56).

490. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972) (cited on page 56).

491. Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241 (1982) (cited on page 56).

492. Grambsch, P. M. & Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994) (cited on page 56).

493. The SAM/BAM Format Specification Working Group. *Sequence Alignment/Map Format Specification* `https://samtools.github.io/hts-specs/SAMv1.pdf` (cited on page 60).

494. Strong, M. J., Xu, G., Morici, L. *et al.* Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens* **10** (2014) (cited on page 60).

495. Glassing, A., Dowd, S. E., Galandiuk, S. *et al.* Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens* **8** (2016) (cited on page 60).

496. Sangiovanni, M., Granata, L., Thind, A. S. *et al.* From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics* **20**, 168 (2019) (cited on page 60).

497. Eisfeldt, J., Ma, G., Ameur, A. *et al.* Discovery of Novel Sequences in 1,000 Swedish Genomes. *Molecular Biology and Evolution* **37**, 18–30 (2019) (cited on page 60).

498. Hasan, M. S., Wu, X. & Zhang, L. Uncovering missed indels by leveraging unmapped reads. *Scientific Reports* **9**, 11093 (2019) (cited on page 60).

499. Cameron, D. L., Baber, J., Shale, C. *et al.* GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. *bioRxiv* (2019) (cited on page 60).

500. McKenna, A., Hanna, M., Banks, E. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010) (cited on page 60).

501. Grüning, B., Dale, R., Sjödin, A. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* **15**, 475–476 (2018) (cited on page 60).

502. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012) (cited on page 62).

503. Nilsson, H., Lindgren, D., Axelson, H. *et al.* Features of increased malignancy in eosinophilic clear cell renal cell carcinoma. *The Journal of Pathology* **252**, 384–397 (2020) (cited on page 62).

504. Mardis, E. R. The $1,000 genome, the $100,000 analysis? *Genome Medicine* **2,** 84 (2010) (cited on page 65).

505. Best, S., Stark, Z., Brown, H. *et al.* The leadership behaviors needed to implement clinical genomics at scale: a qualitative study. *Genetics in Medicine* **22,** 1384–1390 (2020) (cited on page 65).

506. Chan Zuckerberg Initiative. *Essential Open Source Software for Science* `https://chanzuckerberg.com/eoss/`, visited 2020-11-15 (cited on page 66).

507. Research Software Engineers International. `https://researchsoftware.org/`, visited 2020-11-15 (cited on page 66).

# Part II

# Original Studies

# Study I

Saal *et al. Genome Medicine* (2015) 7:20
DOI 10.1186/s13073-015-0131-9

Genome Medicine

RESEARCH Open Access

# The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine

Lao H Saal[1,2,3], Johan Vallon-Christersson[1,2,3], Jari Häkkinen[1,2,3], Cecilia Hegardt[1,2,3], Dorthe Grabau[4], Christof Winter[1,2], Christian Brueffer[1,2], Man-Hung Eric Tang[1,2], Christel Reuterswärd[1,2,5], Ralph Schulz[1,2,5], Anna Karlsson[1,2,5], Anna Ehinger[1,2,6], Janne Malina[7], Jonas Manjer[8], Martin Malmberg[9], Christer Larsson[2,10], Lisa Rydén[2,11], Niklas Loman[1,2,9] and Åke Borg[1,2,3,5*]

## Abstract

**Background:** Breast cancer exhibits significant molecular, pathological, and clinical heterogeneity. Current clinicopathological evaluation is imperfect for predicting outcome, which results in overtreatment for many patients, and for others, leads to death from recurrent disease. Therefore, additional criteria are needed to better personalize care and maximize treatment effectiveness and survival.

**Methods:** To address these challenges, the Sweden Cancerome Analysis Network - Breast (SCAN-B) consortium was initiated in 2010 as a multicenter prospective study with longsighted aims to analyze breast cancers with next-generation genomic technologies for translational research in a population-based manner and integrated with healthcare; decipher fundamental tumor biology from these analyses; utilize genomic data to develop and validate new clinically-actionable biomarker assays; and establish real-time clinical implementation of molecular diagnostic, prognostic, and predictive tests. In the first phase, we focus on molecular profiling by next-generation RNA-sequencing on the Illumina platform.

**Results:** In the first 3 years from 30 August 2010 through 31 August 2013, we have consented and enrolled 3,979 patients with primary breast cancer at the seven hospital sites in South Sweden, representing approximately 85% of eligible patients in the catchment area. Preoperative blood samples have been collected for 3,942 (99%) patients and primary tumor specimens collected for 2,929 (74%) patients. Herein we describe the study infrastructure and protocols and present initial proof of concept results from prospective RNA sequencing including tumor molecular subtyping and detection of driver gene mutations. Prospective patient enrollment is ongoing.

**Conclusions:** We demonstrate that large-scale population-based collection and RNA-sequencing analysis of breast cancer is feasible. The SCAN-B Initiative should significantly reduce the time to discovery, validation, and clinical implementation of novel molecular diagnostic and predictive tests. We welcome the participation of additional comprehensive cancer treatment centers.

**Trial registration:** ClinicalTrials.gov identifier NCT02306096.

* Correspondence: ake.borg@med.lu.se
[1]Department of Clinical Sciences, Division of Oncology and Pathology, Lund University, Medicon Village 404-A2, SE-22381 Lund, Sweden
[2]Lund University Cancer Center, SE-22381 Lund, Sweden
Full list of author information is available at the end of the article

## Background

Breast carcinoma is one of the most common cancers worldwide and a leading cause of cancer-related death in women. Approximately one in nine women will be diagnosed with breast cancer during their lifetime, and in Sweden it accounted for 7,087 new diagnoses and 1,401 deaths in 2011 alone [1]. Contemporary treatment, consisting of surgery, radiotherapy, endocrine therapy, chemotherapy, as well as targeted agents, is driven by standardized clinicopathological criteria and has led to a modest decrease in mortality the last two decades. For example, in the Nordic countries the 5-year survival rate is over 85% [2]. Despite this encouraging statistic, the complete portrait is less than ideal. Unfortunately, approximately 25% of women who survive 5 years will, within the subsequent 15 years, die from recurrent disease [3]. This is in stark contrast to many other cancer types where a 5-year survival is essentially a cure (for example, uterine cancer). On the other hand, it is also recognized that a significant proportion of breast cancer patients are being overtreated: many patients are likely cured by locoregional therapy alone, but are enduring the side effects of unnecessary additional systemic therapies [4]. Our inability to reliably identify such patients has a significant impact on patient quality of life, and adds significantly to the direct economic costs of treating breast cancer as well as the indirect effects on societal productivity [5,6]. Furthermore, we also have limited tools to predict which patients will fail on an indicated therapy due to inherent resistance, or to predict which therapy among statistically equivalent options will be the most effective for an individual patient. Thus, there is still a pressing need for improved biomarkers in breast cancer.

Like all malignancy, breast carcinoma is caused by aberrations in the genome of formerly healthy cells. These aberrations include changes in the normal DNA genetic sequence (for example gene mutations or gains or losses of genetic material) as well as changes in the accessibility and regulation of DNA (such as hypermethylation and chromatin marks). These genomic aberrations affect gene function, and, in concert, also manifest themselves by markedly changing the expression levels of thousands of genes in the tumor from what is the normal pattern in the healthy tissue. Moreover, many of these gene-, genomic-, and gene expression alterations (termed collectively here as biomarkers) are believed to relate to the patient's prognosis and response to therapy. In breast cancer, the study of gene expression alterations and their relation to clinical outcomes is the most mature, whereas DNA copy number aberrations and clinical course has not advanced as far (with one notable exception, HER2), and much less is understood about somatic mutations and therapy response and survival. Despite much study, there are only a handful of examples of breast cancer biomarkers in clinical use today (for example, the estrogen receptor and HER2).

Recent technological advances have opened exciting new possibilities for studying carcinogenesis at an unprecedented molecular detail, and for developing new clinical tools to improve cancer diagnosis, prognosis, and treatment decision-making. One of the most significant of these new technologies is massively-parallel sequencing, also called next-generation sequencing or deep sequencing [7]. Deep sequencing allows one to 'read' the sequence of nucleotide bases of DNA or RNA molecules and identify abnormal sequence variations such as gene mutations and chromosomal rearrangements. Moreover, deep sequencing is also quantitative: the number of sequencing reads that map to a given sequence is proportional to the number of nucleic acid molecules (DNA or RNA) with that sequence in the original sample. Therefore, by sequencing a tumor's DNA one can measure the DNA copy number of each segment of the genome, and by sequencing a tumor's messenger RNA (mRNA), one can quantitate the expression level of each gene transcript. In contrast to microarray methods, where expression level or copy number can only be reported for the pre-determined probe sequences that are present on the microarray, an added advantage of deep sequencing is that it operates at the whole-genome scale where a complete representation of the population of DNA or RNA molecules in a sample can be queried simultaneously. Most next-generation technologies are also several orders of magnitude more efficient and less costly than prior sequencing approaches. The cost of gene expression profiling by RNA sequencing (RNA-seq) is similar to the cost of a microarray gene expression experiment. On the other hand, whole-genome sequencing or targeted exome sequencing remains significantly more costly per sample than RNA-seq; however sequencing costs continue to fall. Therefore, routine clinical tests based on tumor deep sequencing can be economically viable, especially considering that many different test results could be reported from a single sequencing analysis.

A major challenge to translating a new cancer biomarker to the clinic is the study sample size. In the development phase, the number of patients studied and its representation of the natural biological and clinical diversity has been inadequate in many studies, usually numbering in only a few hundred samples in the largest studies, and often suffering from various types of selection biases. Validation phase studies often suffer from similar issues. This leads to several consequences, for example the failure to discover potential biomarkers in the development phase, overfitting of data and non-generalizability of biomarkers, and biomarker failure at the validation phase. As a result, thus far there are few multigene assays for breast cancer in limited clinical use: the two most commonly used are a microarray-based test, MammaPrint (Agendia BV), and the OncotypeDX qRT-PCR assay (Genomic Health, Inc). However, these assays are expensive (approximately €3,000 per test), and due to the fact that they were developed based on relatively small study populations (MammaPrint: 78 patients) [8] or only one subgroup of the disease (OncotypeDX: patients with estrogen-receptor-positive tumors with no involved lymph nodes and treated with tamoxifen) [9], the overall clinical utility of these tests beyond the selection of patients with limited benefit from adjuvant chemotherapy, and whether better tests could be developed, has been debated [10,11].

To address these clinical and practical challenges and to continue our efforts to develop improved clinical biomarker tests for breast cancer [12-16], we initiated the multicenter and multidisciplinary consortium, Sweden Cancerome Analysis Network - Breast (SCAN-B) [17] (ClinicalTrials.gov identifier NCT02306096). Launched in the autumn of 2010, the study is fully integrated in the clinical routine and has enrolled more than 6,000 patients to date and collected tumors and blood specimens at a rate of 25 to 30 per week, representing approximately 85% of all breast cancer diagnoses in southern Sweden. Based on our prior experiences, in the first phase we are performing whole-transcriptome RNA-seq and have sequenced over 3,000 breast tumors to date. The primary objectives are to develop, validate, and implement clinically beneficial molecular tumor analyses into the routine healthcare setting for patients with breast cancer in order to improve their care, quality of life, and outcome. Herein we present an overview of the SCAN-B Initiative, our optimized protocols, the status of patient accrual and sample processing for the first 3 years, and the results of initial proof of concept RNA-seq analyses for 49 consecutive patient tumors analyzed in parallel on gene expression microarrays including tumor subtyping and analysis of mutations in cancer-associated genes.

## Methods

### Ethics statement

The study was conducted in accordance with the Declaration of Helsinki and has been approved by the Regional Ethical Review Board of Lund (diary numbers 2007/155, 2009/658, 2009/659, 2014/8), the county governmental biobank center, and the Swedish Data Inspection group (diary number 364-2010). Written information is given by trained health professionals and all patients provided written informed consent.

### Infrastructure

SCAN-B involves researchers, clinicians, and healthcare professionals at Lund University Hospital, Division of Oncology and Pathology, the South Sweden Breast Cancer Group [18], the Regional Cancer Center South, and all seven hospital centers treating breast cancer patients in the Southern Healthcare Region (Malmö, Lund, Helsingborg, Kristianstad, Halmstad, Växjö and Karlskrona), and operates under the auspices of the South Sweden Breast Cancer Group and Regional Cancer Center South. An overview of the study infrastructure is presented in Figure 1.

### Patients and samples

Patient enrollment is integrated and performed as part of the clinical routine (Figure 1). From 30 August 2010, breast cancer patients across the south of Sweden have been offered inclusion in SCAN-B. The eligibility criterion was a preoperative diagnosis of primary invasive breast cancer, and since the autumn of 2012, patients with a preoperative suspicion for breast cancer are also eligible as well as patients receiving neoadjuvant therapy. Patients who participate in SCAN-B receive the same standard of care as patients who do not participate, and at the present time, results from this prospective study are not used to alter any clinical decisions. The study affects the clinical routine minimally. At time of routine preoperative/pre-biopsy blood work, three additional study blood tubes are collected and biobanked as whole blood, buffy coat, plasma, and serum. Clinical routines at surgery, radiology, pathology, and oncology proceed normally. After the routine assessment of the surgical specimen by the pathologist, remainder tumor-cell enriched fresh specimen(s) is placed in a study sample tube(s) containing RNAlater reagent (Ambion) and the time to preservation is recorded. Very small tumors do not always yield excess material for the study, and due to clinical considerations, at present it is difficult to sample cases that appear to be purely or primarily carcinoma *in situ*. For included patients undergoing preoperative biopsy, additional study biopsies are taken and placed in RNAlater. Sample tubes, identified by barcodes, are shipped twice per week at 4°C via inter-hospital transport to the central research laboratory of the Canceromics Branch, Division of Oncology and Pathology, Lund University Cancer Center [19]. Clinical and pathological information tied to the patient and diagnosis as well as follow-up data is retrieved from the national quality registry for cancer patients (INCA) (Figure 1). Postoperative blood samples are also collected as above at 6 months, 12 months, and 36 months after primary surgery. In accordance with ethics and privacy guidelines and laws, clinical and sample information are coded and strictly confidential.

### Tumor sample processing

Tumor specimens sent in RNAlater are processed continuously in our central laboratory (see Additional file 1

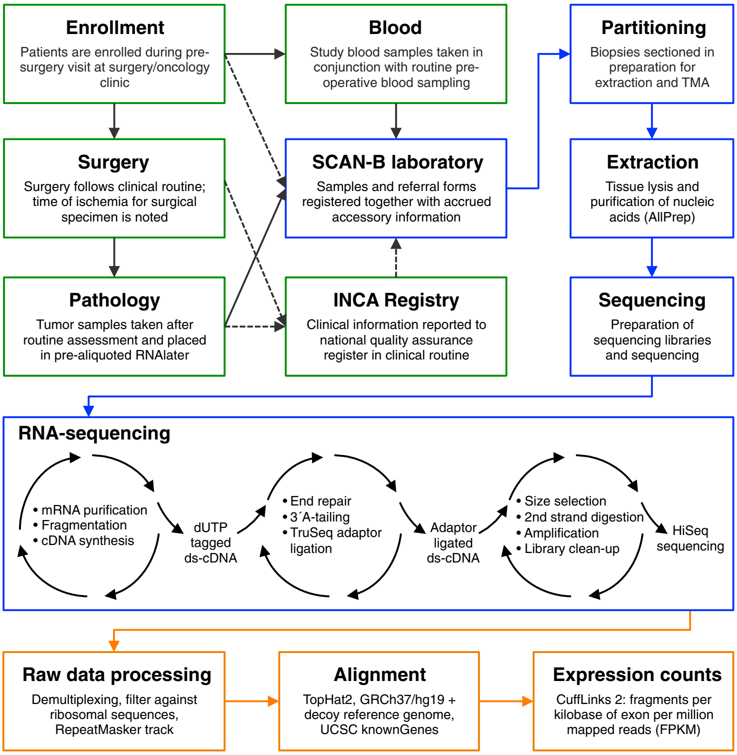

Enrollment
Patients are enrolled during pre-surgery visit at surgery/oncology clinic
Blood
Study blood samples taken in conjunction with routine pre-operative blood sampling
Partitioning
Biopsies sectioned in preparation for extraction and TMA
Surgery
Surgery follows clinical routine; time of ischemia for surgical specimen is noted
SCAN-B laboratory
Samples and referral forms registered together with accrued accessory information
Extraction
Tissue lysis and purification of nucleic acids (AllPrep)
Pathology
Tumor samples taken after routine assessment and placed in pre-aliquoted RNAlater
INCA Registry
Clinical information reported to national quality assurance register in clinical routine
Sequencing
Preparation of sequencing libraries and sequencing
RNA-sequencing
• mRNA purification
• Fragmentation
• cDNA synthesis
dUTP tagged ds-cDNA
• End repair
• 3´A-tailing
• TruSeq adaptor ligation
Adaptor ligated ds-cDNA
• Size selection
• 2nd strand digestion
• Amplification
• Library clean-up
HiSeq sequencing
Raw data processing
Demultiplexing, filter against ribosomal sequences, RepeatMasker track
Alignment
TopHat2, GRCh37/hg19 + decoy reference genome, UCSC knownGenes
Expression counts
CuffLinks 2: fragments per kilobase of exon per million mapped reads (FPKM)

**Figure 1 Overview of the SCAN-B infrastructure.** Shown are the SCAN-B clinical (green boxes), laboratory (blue), and computational and analytical (orange) components. Solid black arrows indicate flow of material, and dashed black lines indicate flow of information. Enrollment and sampling of patients at time of preoperative (neoadjuvant) biopsy is not shown. ds, double-stranded; INCA, Swedish national breast cancer registry; TMA, tissue microarray.

for detailed protocols) with handling standards that meet or exceed recommendations of the Breast International Group (BIG). Each tumor specimen is weighed, and when possible, partitioned into three parts: one 30 mg (approximately) piece for simultaneous isolation of DNA, RNA, and protein; one adjacent 10 mg (approximately) piece used for manufacture of a formalin-fixed paraffin-embedded low-density tissue microarray (TMA); and any remainder is stored frozen for future use. The TMA is used for estimation of tumor cellularity and as a research resource. Nucleic acids and protein fraction are isolated from tumor specimen using the AllPrep method and automated using QIAcube machines (Qiagen). RNA and DNA quality control is performed by NanoDrop spectrophotometry and BioAnalyzer (Agilent) or Caliper LabChip XT (PerkinElmer) capillary gel analysis. The extracted RNA, DNA, and flow-through portion that contains proteins and short nucleic acids, are stored frozen for future use. All study information, sampling information, and analysis information are recorded in a secure relational data management and analysis system, BASE [20-22], and user-friendly sample and protocol workflows are interactively generated by the system to ensure standard laboratory operating procedures and efficiency.

### Library preparation for RNA-sequencing

Customized protocols for RNA-seq using 1 μg of starting total RNA were developed and automated for a high-throughput workflow (Figure 1). The complete methods and protocols are described in the Additional file 1. In brief, poly(A) mRNA is isolated from the total RNA in up to 96-well microtiter plate format by two rounds of purification with Dynabeads Oligo $(dT)_{25}$ (Invitrogen) using a KingFisher Flex magnetic particle processor (ThermoScientific). Zinc-mediated fragmentation (Ambion) is performed and the fragmented mRNA retrieved using column purification (Zymo-Spin I-96 plates; Zymo). The sequencing library generation protocol is a modification of the dUTP method, which importantly retains the directionality (stranded-ness) of the sequenced RNA molecules [23,24]. First strand cDNA synthesis is performed using random hexamers and standard dNTP mix followed by cleanup using Sephadex gel filtration (Illustra AutoScreen-96A plates; GE Healthcare), and second strand cDNA synthesis is performed using dUTP in place of dTTP in the dNTP-mix and cleanup using Zymo-Spin I-96 plates. The cDNA is end-repaired and A-tailed, and diluted TruSeq adapters with barcodes are ligated using a modified protocol (Illumina) [23]. Adapter-ligated cDNA is then size-selected to remove short oligonucleotides using carboxylic acid (CA) paramagnetic beads (Invitrogen) and polyethylene glycol (PEG), similar to the previously described methods [25], and automated on the KingFisher Flex. The second cDNA strand is digested using uracil-DNA glycosylase and the product is enriched by 12 PCR cycles (Illumina). The PCR product undergoes two cycles of size selection using CA-beads and varying concentrations of PEG, first to exclude DNA fragments >700 bp and then to exclude fragments <200 bp. Quality control is performed on control libraries using Qubit fluorometric measurement (Life Technologies) and Caliper LabChip XT microcapillary gel electrophoresis. Typically, 10 to 24 barcoded libraries are included in a pool and each pool is sequenced in at least one lane across dual flowcells. Paired-end sequencing of 50 bp read-length is performed on an Illumina HiSeq 2000 instrument.

### RNA-seq gene expression measurements

Raw sequencing read data are demultiplexed using an in-house software and collated by library barcode into sample data sets (Figure 1). Each data set is filtered to remove reads that align (using Bowtie 2 [26] with default parameters except *-k 1 –phred33 –local*) to ribosomal RNA/DNA (GenBank loci NR_023363.1, NR_003285.2, NR_003286.2, NR_003287.2, X12811.1, U13369.1), phiX174 Illumina control (NC_001422.1), and sequences contained in the UCSC hg19 RepeatMasker track (downloaded 14 March 2011). The remaining reads are aligned using TopHat2 [27] to the human genome reference GRCh37/hg19 (with b37 masked chromosome Y and hs37d5 decoy sequences) together with 80,884 transcript annotations from the UCSC knownGenes table (downloaded 10 September 2012). Default TopHat2 parameters are used except for *–mate-inner-dist* (average size with adapters 355, range 268 to 465, measured for each sample individually) *–mate-std-dev 100 –library-type fr-firststrand –keep-fasta-order –no-coverage-search*. Cufflinks v2.1.1 [28] is used to calculate expression levels, fragments per kilobase of exon per million mapped reads (FPKM), using default settings except *–frag-bias-correct –multi-read-correct –library-type fr-firststrand –compatible-hits-norm*. Unmapped reads are processed to be usable by downstream analysis tools using custom software [29]. Read duplication statistics and routine quality assessment were performed using the Bioconductor *Rsamtools* v1.12.4 package [30]. Herein we present analysis for 55 sample libraries generated from 49 tumor specimens, six run with technical replicates using separate aliquots of total RNA. RNA-seq read statistics are presented in Additional file 2: Table S1. Gene expression data were pre-processed by collapsing on 27,979 unique gene symbols (sum of FPKM values of each matching transcript), adding to each gene's expression measurement 0.1 FPKM, performing a $\log_2$ transformation, and centering the gene expression values by subtracting the row-wise (gene) median (calculated across the 49 primary data sets) from the values in each row of data.

### Microarray gene expression measurements

To compare to RNA-seq, the same 55 RNA samples as above (49 tumors, six as technical replicates) were analyzed on Human HT12 v4 BeadChip microarrays following the manufacturer's standard protocol (Illumina). Data from each microarray were pre-processed in BASE [20-22]: background correction was performed, and a constant of 11 was added to each intensity measurement. Genes with missing values in >10% of samples were excluded; otherwise missing values were imputed using k-nearest neighbors implemented in the *impute* R package. The data were quantile normalized using the *preprocessCore* R package, log2 transformed, and each gene was median centered across samples as for the RNA-seq data.

### Molecular subtyping

Intrinsic molecular subtyping was performed by nearest centroid method and Spearman correlation within the *genefu* R package, using three published gene lists (Sørlie, Hu, and PAM50) [31-33]. Mapping of genes between data sets was performed using the *probemapper* 1.0.0 R package [34]. For PAM50, all 50 genes were used for subtyping on both RNA-seq and HT12 platforms; for the Sørlie classification, 432 genes were matched for RNA-seq and 434 genes for HT12; and for Hu classification, 225 genes and 229 genes, respectively. To facilitate unbiased between-platform comparison, each tumor was assigned to the class with the highest correlation. For visualization, hierarchical clustering was performed using the *ConsensusClusterPlus* R package with 1,000 sub-samplings of 80% of samples (or genes; run independently), Pearson distance metric, Ward linkage, and the RNA-seq PAM50 expression values as input. Clusters stabilized at five sample and seven gene clusters.

### RNA-seq mutation analysis

Sequence variants were investigated in known and likely breast cancer driver genes. A list of candidate driver genes of interest was compiled based on the union of genes identified in several large studies: the TCGA breast cancer study (supplementary table 2 in [35]); the Sanger 100 breast cancer exome study (supplementary table 4 in [36]); the Cancer Gene Census of breast cancer drivers [37]; and additional genes with evidence for hereditary breast cancer predisposition. The union resulted in 90 genes (see Additional file 2: Table S2). Using the TopHat-aligned BAM files, pileup files restricted to the exonic regions (plus padding of 10 bases) for these 90 genes were created for each sample using *samtools* v0.1.18 and read metrics were calculated using *bam-readcount* [38]. VarScan v2.3.5 [39] was used to call single nucleotide variants (SNVs) and indels using the following settings: *–min-coverage 2 –min-reads2 2 –min-avg-qual 10 –min-var-freq 0.05 –p-value 1*. The first six bases of each read were ignored for mutation analysis in subsequent steps as mismatches can be introduced by random hexamer priming during library preparation. The local reference sequence around each variant was retrieved using BEDTools [40]. Variant calls were annotated using ANNOVAR [41] with the databases *refGene*, *snp137NonFlagged*, and *cosmic65* from the ANNOVAR website; additional databases of SNVs and indels were created, *tcgaBreast* using data from the Level 2.5.1.0 MAF file from the TCGA Data Portal [42], and *stephens2012* using data from supplementary table 4 in reference [36]. To reduce false positive mutation calls, variants were excluded if they matched any of the following criteria: present in *dbSNP137NonFlagged*, not present in *cosmic65* or *tcgaBreast* or *stephens2012*, located in 5′ or 3′ UTRs, synonymous variants, variants with adjacent homopolymer stretches of ≥5 bases, SNVs with an average base quality of the variant allele <20, and variants with an average distance to the 3′ end of the read <5% of the total read length (after clipping). Thus, only previously identified somatic mutations remained. For plotting amino acid variants, Pfam protein domains were obtained using the *biomaRt* R package by first mapping RefSeq transcript identifiers to UniProt entries within the Ensembl *hsapiens_gene_ensembl* data set, and then querying the InterPro protein data set with these UniProt entries.

### Statistics

Enrollment statistics are based on study records in our relational data management system BASE. The counts for blood samples and tumor specimens are based on the number of patients with at least one sample or specimen collected. To compare the distribution of patient and clinicopathological annotations between sets of patients, Fisher's exact test was used. A *P* value less than 0.05 was considered significant.

### Bioinformatics implementation

Customized Bash shell scripts, Python, and R code, as well as relevant software packages as described above, were used to perform all bioinformatics analyses.

### Data availability

The RNA-seq and microarray gene expression data herein are available from the NCBI Gene Expression Omnibus [43] under accession GSE60789.

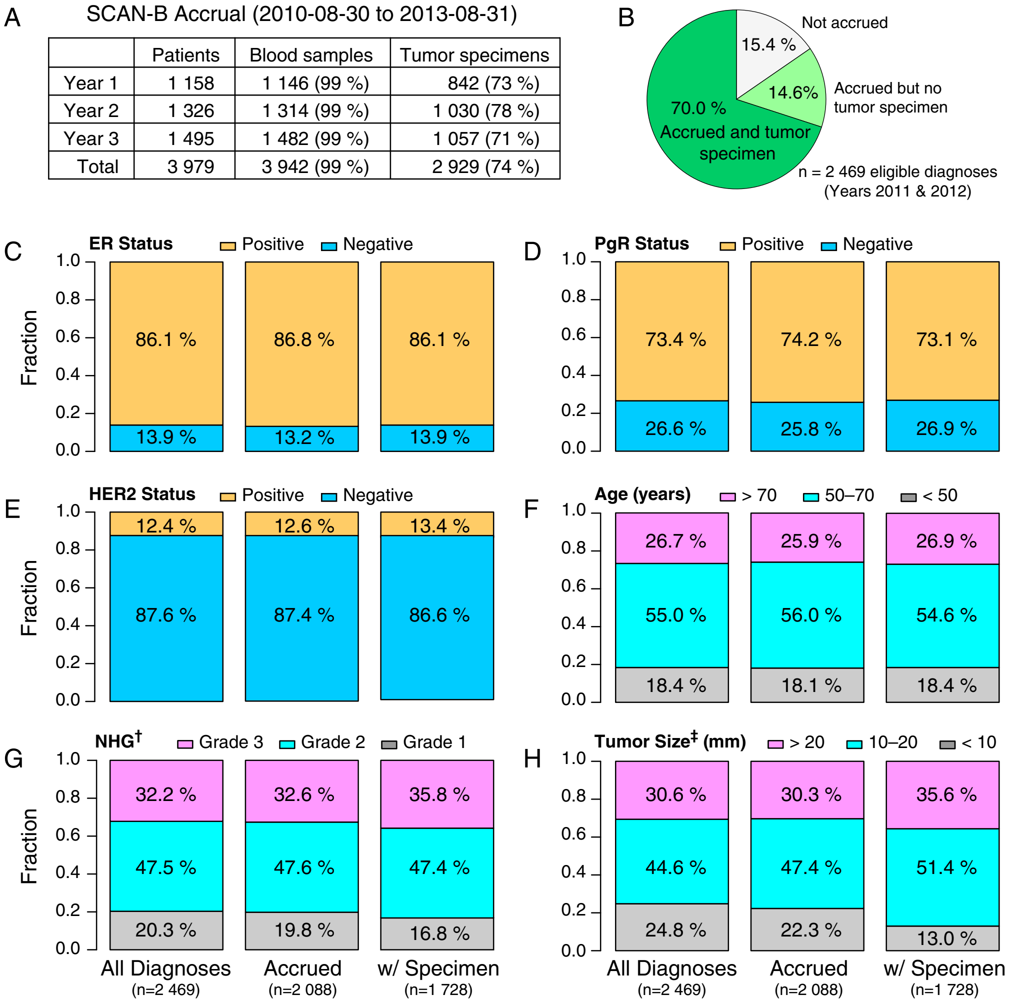
## Results

### Population-based enrollment

We summarize here the results for the first three years of patient accrual, from 30 August 2010 to 31 August 2013. During this period, 3,961 women and 18 men enrolled in SCAN-B (Figure 2A). For the 2011 and 2012 calendar years, where it was possible to match complete annual records to the Swedish national breast cancer registry (INCA), this represents an estimated 85% of the eligible patient population (with a preoperative diagnosis) within the catchment region (Figure 2B). Approximately 3% of eligible patients decline participation in the study, and 12% are lost to enrollment. There is no bias in terms of clinical variables (estrogen receptor status, progesterone receptor status, HER2 status, patient age, Nottingham grade, or tumor size) between the included patients and the population of all eligible breast cancer diagnoses (Figure 2C-H).

Patients diagnosed with breast cancer are prospectively enrolled at the rate of 25 to 30 per week. For 99% of included patients, preoperative blood samples are biobanked (Figure 2A). For 2,929 patients (74%; Figure 2A), at least one tumor specimen has been submitted to the central laboratory, usually within 1 to 3 days after biopsy/surgery. Most commonly, the reasons for not submitting a tumor specimen include it being judged by the clinical pathologist to be too small for sampling (73%) or the tumor appearing to be carcinoma *in situ* only (7%). The subgroup of patients for which a tumor specimen was collected does not differ significantly from the population of enrolled patients with respect to all clinical variables with the exception of tumor size and Nottingham grade (Figure 2C-H). The median tumor specimen ischemia time, from excision from the patient to placement in preservative solution, is 46 min (interquartile range (IQR), 32 to 65 min), and the median specimen weight is 63 mg (IQR, 34 to 108 mg).

Processing of tumor specimens is performed in near real-time in our central laboratory. As of 31 August 2013, 2,890 of 2,929 tumor specimens (99%) had been partitioned for AllPrep, TMA, and reserve piece, processed,



SCAN-B Accrual (2010-08-30 to 2013-08-31)

| | Patients | Blood samples | Tumor specimens |
|---|---|---|---|
| Year 1 | 1 158 | 1 146 (99 %) | 842 (73 %) |
| Year 2 | 1 326 | 1 314 (99 %) | 1 030 (78 %) |
| Year 3 | 1 495 | 1 482 (99 %) | 1 057 (71 %) |
| Total | 3 979 | 3 942 (99 %) | 2 929 (74 %) |

**Figure 2 Study demographics and clinical variables. (A)** For the period 30 August 2010 to 31 August 2013, yearly (non-calendar) summary of the number of enrolled patients, the number of patients with preoperative blood sample collected, and number of patients with tumor specimen collected. **(B-H)** For the two complete calendar years 2011 and 2012 that could be matched to the INCA Swedish national breast cancer registry, **(B)** chart of all cases with a preoperative diagnosis of primary breast cancer within the catchment region divided into those that were accrued or not accrued. Comparison of baseline clinical variables between all eligible breast cancer patients, patients accrued, and patients accrued with tumor sample, for **(C)** estrogen receptor (ER) status, **(D)** progesterone receptor (PgR) status, **(E)** HER2 status, **(F)** age at diagnosis, **(G)** Nottingham histological grade (NHG), and **(H)** tumor size. †,‡ Significant differences were identified between all diagnoses and accrued with tumor specimen for NHG ($P = 0.005$) and tumor size ($P < 0.001$), and between patients accrued and accrued with biopsy for NHG ($P = 0.025$) and tumor size ($P < 0.001$).

and the nucleic acids isolated (see Methods and Additional file 1). In the first round of processing (with half of the specimen lysate stored frozen for future use), a median of 8.5 μg total RNA (IQR, 3.7 to 16.5 μg) and 15.4 μg DNA (IQR, 7.6 to 25.5 μg) has been isolated per tumor specimen. The isolated nucleic acids are of high purity, with a median 260/280 ratio of 2.05 (IQR, 2.03 to 2.07) and 260/230 ratio 1.93 (IQR, 1.61 to 2.08) for the RNA, and median 260/280 ratio of 1.87 (IQR, 1.86 to 1.88) and 260/230 ratio 1.78 (IQR, 1.38 to 1.99) for the DNA. All unused tissue, lysates, and extracted nucleic acids are stored frozen for future use. The focus of our molecular analyses is initially on whole-transcriptome RNA sequencing using the Illumina HiSeq 2000 platform. For this purpose, greater than 1 μg of total RNA was isolated from 95% of patient samples in the first round of processing, and the median RNA quality score (RQS) is 8.4 (IQR, 7.8 to 8.7).

## RNA sequencing of breast cancer transcriptomes

We have developed a customized high-throughput RNA-seq library generation protocol (Additional file 1). Thus far, it has been used to sequence the transcriptomes of over 3,000 breast tumors. Here we present initial proof of concept results for a representative series of 49 population-based breast cancer patients whose primary surgery occurred during the fall of 2011 and that we analyzed in parallel by RNA-sequencing and gene expression microarrays. From these 49 cases, six were sequenced in technical replicates making for a total of 55 libraries. For each library, a median of 47.6 million passed-filter (PF) paired-reads of 50 bp length were analyzed (IQR, 43.4 to 54.2 million) (Additional file 2: Table S1). An average of 83.0% (range, 73.0% to 88.6%) of the paired-reads remain after initial filtering against a database of non-mRNA targets (passing contamination filter, PCF). Of the remaining reads, a median of 68.0% (IQR, 65.3% to 75.3%) can be mapped to the reference transcriptome map. The average base quality Q-score per read cycle was never below 29, and the duplication rate was low, with a median 63.3% read-pairs being unique (IQR, 55.1 to 69.1%).

Quantitative gene expression levels, in the form of fragments per kilobase of exon per million mapped reads (FPKM), were derived from the aligned RNA-seq data. As anticipated, the molecular subtypes of breast cancer were readily apparent when classifying tumors using several of the published molecular signatures, such as PAM50 or the intrinsic gene lists of Sørlie *et al.* and Hu *et al.* (Figure 3A and Additional file 2: Figure S2) [31-33]. To compare our RNA-seq method to a prior standard for gene expression profiling, in parallel we performed microarray analysis using Illumina HT12 BeadChips with the same RNA from these 49 tumors, including performing the same six cases in replicate. Concordance of molecular subtypes between RNA-seq and microarray platforms was high using the PAM50, Sørlie, or Hu signatures (90%, 92%, and 96%, respectively) (see also Additional file 2: Figure S2). Gene expression

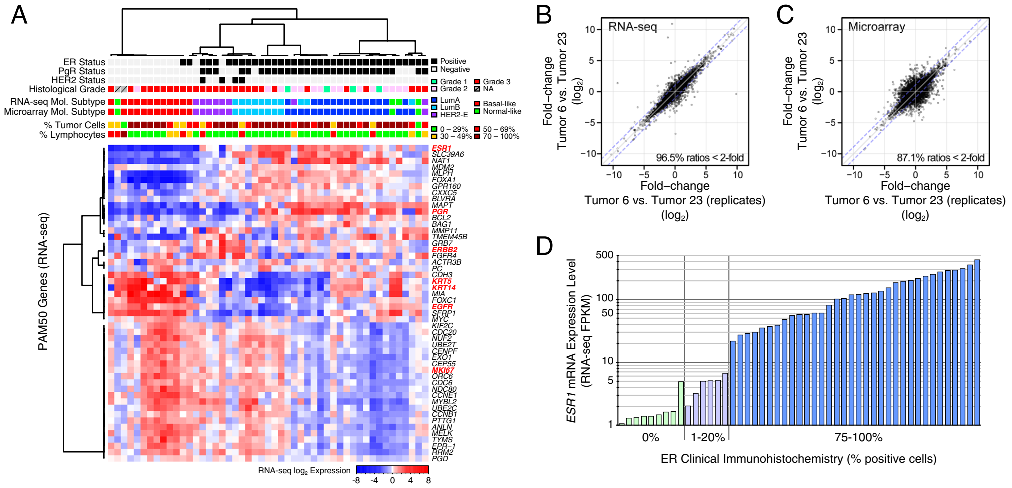


**Figure 3 RNA sequencing and microarray analysis for population-based breast tumors. (A)** Hierarchical clustering of 49 primary breast tumors (clustered columns) using the RNA-seq gene expression measurements and the PAM50 intrinsic gene signature (clustered rows). Clinical annotations for estrogen receptor (ER), progesterone receptor (PgR), and HER2 are indicated below the sample dendrogram, and PAM50 intrinsic subtyping is shown for classification using RNA-seq data as well as using microarray data generated from the same input RNA (90% concordant; results for Sørlie (92%) and Hu (96%) signatures are presented in Additional file 2: Figure S2). Genes of interest are highlighted in red, and relative expression level is indicated by the box color (see color key below the heatmap). For six tumor samples, technical replicates from the same RNA sources were performed for both RNA-seq and microarrays; plotted in **(B)** and **(C)** are representative examples comparing the fold-change for all RefSeq genes between two tumors (Y axis), and the fold-change between the replicated experiments for the same two tumors (X axis). Consistently, RNA-seq demonstrated values closer to the ideal line of identity and for a broader dynamic range. The +/- 2 fold-change ($|\log_2| = 1$) thresholds are indicated by blue dashed lines. **(D)** RNA-seq-derived expression level of *ESR1*, which encodes the ER alpha protein, is shown compared to the clinical ER IHC score for each of the 49 tumors. See Additional file 2: Figure S3 for corresponding plots for progesterone receptor and *ERBB2* (HER2).

levels derived from RNA-seq compared favorably to microarray-derived gene expression levels. Replicate experiments for six tumors, performed on both platforms, show that the measurement range and reproducibility are higher for RNA-seq, with less apparent noise, as compared to microarrays (Figure 3B and C). The gene expression levels for *ESR1*, which encodes the estrogen receptor alpha (ER) receptor, was compared to the clinical ER immunohistochemistry scores and illustrates the wide dynamic range of RNA-seq for an important breast cancer biomarker (Figure 3D). Corresponding plots are shown for *PGR* (progesterone receptor (PgR)) and *ERBB2* (HER2) in Additional file 2: Figure S3.

## Mutation screening by RNA sequencing

In addition to these technical attributes, RNA-seq data can be used to detect gene mutations, splice variants, and fusion transcripts, opening up new avenues of study. As proof of principle, we utilized the 49 tumor RNA-seq data to screen for sequence alterations in 90 genes known to be mutated in human breast cancers (Figure 4A; Additional file 2: Table S2; GSE60789). Typical mutations were detected in these breast cancers with the expected frequencies and association to clinicopathological characteristics: for example, 17/49 (35%) cases were determined to harbor mutations in the oncogene *PIK3CA* and these occurred most frequently in luminal A (8/14), HER2-enriched (3/8), and normal-like (2/4) tumors [35,44]. Similarly, *TP53* was found to be mutated in 17/49 (35%), almost exclusively in grade 3 tumors (16/17), most frequently within the basal-like (9/12), HER2-enriched (4/8) and luminal B (3/11) subtypes, and least frequently in normal-like (0/4) and luminal A tumors (1/14). The detected spectrum of mutations was in-line with expectations: for example, *PIK3CA* mutations affecting residue H1047 in the kinase domain of p110-alpha protein were the most frequently observed, whereas none of the *TP53* mutations were observed more than once in this

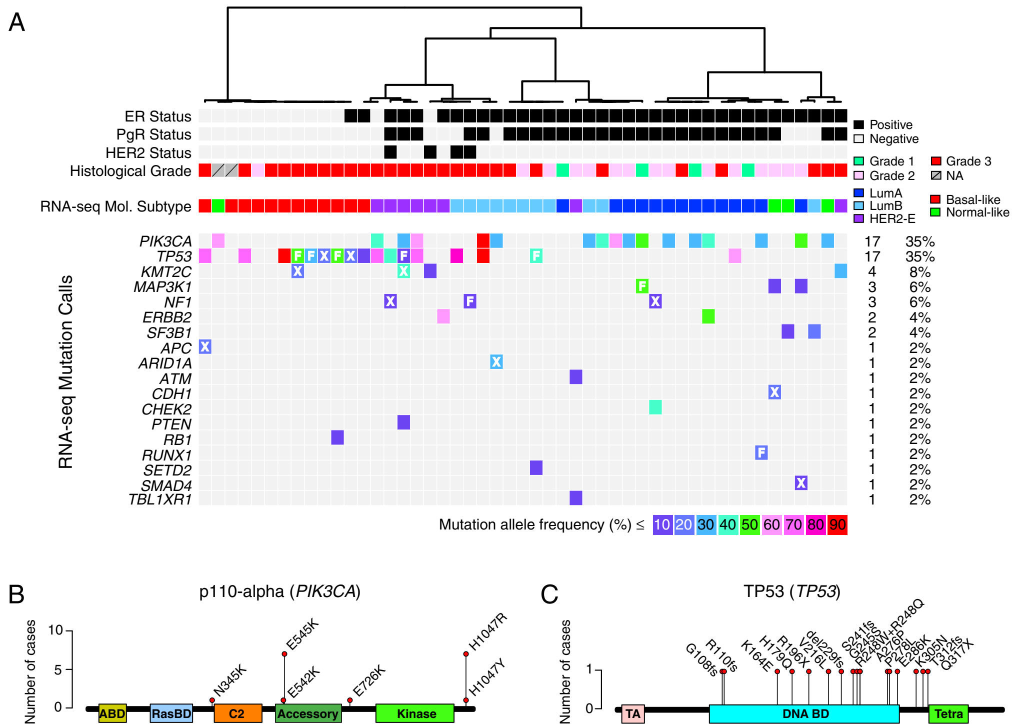


**Figure 4 Detection of mutations by RNA-seq. (A)** Eighteen genes with at least one mutation (out of 90 genes screened) across the 49 population primary breast tumors are shown, in order of frequency (see totals and percentages to the right of each gene row). Mutant allele frequency is indicated by the box color (see key below matrix). All mutations are non-synonymous missense mutations except those indicated by F (frameshift) and X (nonsense). Tumor sample dendrogram is as in Figure 3A. Predicted mutant amino acids are shown for **(B)** *PIK3CA* which encodes the p110-alpha catalytic subunit of the phosphatidylinositol-4,5-bisphosphate 3-kinase oncogene, and **(C)** *TP53* which encodes the tumor suppressor TP53.

series of population-based cases (Figure 4B and C). Mutations in 18 out of 90 genes investigated could also be reliably detected in the transcriptome, such as *KMT2C* (*MLL3*), *MAP3K1*, *ERBB2*, *ARID1A*, *PTEN*, and *RB1*, and 39/49 (80%) of tumors had at least one of these 18 genes mutated (Figure 4A).

Tissue microarrays, constructed from a piece of each tumor adjacent to that used for nucleic acid extractions, were evaluated for cellularity composition by hematoxylin and eosin staining and scored for invasive tumor, *in situ* tumor, normal epithelium, lymphocyte, stroma, and adipocyte content. Generally, approximately 75% of cases contain >50% tumor cells and 15% of cases contain less than 30% tumor cells. Few cases appear to be overtly affected by tumor cell content with respect to supervised analyses, for example in the intrinsic molecular subtyping or mutation analysis (Figures 3A and 4A).

## Discussion

We have developed a mature infrastructure for prospective, multicenter, population-based, enrollment of breast cancer patients, coupled to an optimized genomics platform for gene expression profiling and mutation analysis by RNA sequencing. Powerful biomarker discovery projects will be possible after we have studied many hundreds to many thousands of breast tumors and related these data to patient characteristics, treatment response, and outcomes. The established infrastructure will enable SCAN-B-derived biomarker tests to be validated using independent series of population-based cases from the ongoing prospective SCAN-B study. In a similar way, biomarker tests (such as gene expression signatures) from the literature can be tested and validated within our patient material. For validated biomarker tests that are proven to be clinically relevant, the goal is to perform the analysis as a diagnostic test and communicate the result back to the treating physicians within a clinically-actionable time-frame (within weeks after surgery or biopsy). Thus, within the framework of an initiative such as SCAN-B, the cycle time from biomarker discovery, to independent validation, to clinical implementation can be made more rapid and efficient.

With the current participating sites, the SCAN-B Initiative has and will continue to assemble a very large series of breast cancer cases over many years, prospectively analyzed with the same methods and platforms. The first phase of SCAN-B prioritizes the sequencing of expressed mRNAs because of our prior experience and interest, the maturity of the field, experimental cost, as well as the fact that expression level as well as isoform and variant status can be ascertained simultaneously. The wealth of small and long non-coding RNAs, DNA-level aberrations, and epigenetic changes are not yet investigated. Future analyses will investigate global mutational portraits and differential expression of gene isoforms, and ample study material is stored for future genomic, transcriptomic, and proteomic analyses such as whole-genome and targeted exome sequencing, sequencing of non-coding RNAs, and studies of active proteins. The SCAN-B Initiative will enable numerous types of investigations that are population-based and appropriately powered. For example, we aim to identify and validate RNA and DNA biomarkers predicting exceptionally favorable prognosis without need for adjuvant therapy, biomarkers for resistance to specific therapies, such as trastuzumab resistance or resistance to endocrine therapy, and biomarkers to refine the intermediate prognosis cases, such as tumors of histological grade 2. Within the coming years, we will have amassed many cohorts of hundreds to thousands of patients receiving any particular standard treatment, linked to >5 years follow-up history, and with corresponding RNA-sequencing data. Gene expression patterns and mutational patterns, and other biologically relevant information discernible from the SCAN-B data such as expression of alternatively spliced transcripts, fusion genes, and allele-specific expression, will be analyzed in the context of the clinicopathological information, therapy, and patient outcome in order to develop and validate new biomarker tests for eventual clinical use. We also aim to use the SCAN-B infrastructure to identify patients who may benefit from participation in specific clinical trials, for example to select patients whose gene expression or gene mutation status suggests sensitivity to an emergent therapeutic.

Tumor tissues, grossly dissected at pathology, are the most practical samples to analyze in a large-scale setting as compared to microdissection; moreover, the non-tumoral gene expression signals, such as from immune cells and stroma, may be highly biologically and clinically relevant. For example, it has been shown that immune response signatures can be predictive of outcome across a wide range of cancer forms [45-48]. Therefore, depending on the purpose, we foresee the importance of interpreting genomic biomarker results in the context of the estimated compartmental cellularity of each analyzed specimen.

Comparison of mutation analysis at the level of mRNA- versus DNA-sequence warrants further investigation and is currently underway. Based on our early experiences as well as the work of others, we posit that mutations of oncogenes should be efficiently detectable by RNA-seq, but that some mutations in tumor suppressor genes may be more difficult to detect due to lowered expression levels, loss of heterozygosity, and/or nonsense mediated decay [49-51]. We hope to add DNA-level profiling to the SCAN-B routine in the future. For example, the BRCAsearch subproject is investigating, after additional informed consent, the consecutive testing of germline BRCA mutations. Another project currently in progress is comparing the classification of the five conventional clinical biomarkers (clinical determinations for ER, PgR, HER2, Ki67, and histological grade) to paired classifications based on RNA-seq gene expression signatures. The influence of intra-tumoral heterogeneity, subclonality, other cell types such as non-tumoral epithelial and stromal cells, and the limitations of sampling require further study. For patients receiving primary medical treatment, we are currently implementing an extensive sampling program including sequential blood sampling and an additional tumor biopsy after two cycles of preoperative chemotherapy. We also plan to soon begin systematic collection and analysis of metastatic breast cancer samples upon disease relapse, which will provide a rich and informative platform for studying tumor progression and tumor evolution. We believe these results and future results from SCAN-B will complement the existing clinical and pathological evaluation and can become another part of our armamentarium in diagnosing, evaluating, and treating breast cancer.

We present the feasibility of large-scale multicenter collection of population-based breast cancer patient material and analysis with next-generation genomic analytical methods. In our hands, 85% of new diagnoses are enrolled across a wide geography of Sweden, and the specimen collection reflects well the clinicopathological characteristics of breast cancer in the catchment region. Due to primacy of the clinical diagnostic evaluation, very small/low-grade tumors are slightly under-sampled. We anticipate improvements in patient enrollment, and increases in the fraction where a tumor specimen can be collected, as the procedures become further integrated into the healthcare routine and the importance of tissue and blood sampling for genomic analyses becomes further evident through forthcoming studies from us and others. We are expanding the infrastructure to include patients diagnosed with metastatic breast cancer, and also drawing blood samples at routine intervals during the clinical course for liquid biopsy studies.

Lastly, we extend an open invitation to other hospital systems in Sweden and the Nordic countries to join the SCAN-B network. Most recently, Uppsala County joined the network in October 2013. SCAN-B may also serve as a model for similar translational projects in other types of cancer and diseases.

## Conclusions

In summary, we present the successful implementation of a multicenter infrastructure for genomic biomarker development in breast cancer across a wide geography of Sweden, and the optimization of RNA-seq protocols for high-throughput analyses. To our knowledge, this is the largest endeavor of its kind and is distinctive in its population-based approach.

## Additional files

**Additional file 1: Figure S1.** Materials and Methods.

**Additional file 2: Figure S2.** Figure S3, Table S1, and Table S2.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

LHS, JVC, JH, CH, DG, JaM, JoM, MM, CL, LR, NL, and ÅB conceived of the study. LHS, JVC, JH, CH, CW, CB, and MHET analyzed data. DG, AE, and JaM evaluated histopathology. LHS, JVC, JH, CR, RS, and AK participated in protocol development. CR and AK performed sequencing experiments. LHS, JVC, JH, CH, DG, AE, JaM, JoM, MM, CL, LR, NL, and ÅB participated in establishing the clinical infrastructure. LHS, CH, DG, JoM, MM, CL, LR, NL, and ÅB are members of the SCAN-B Steering Committee. LHS drafted the manuscript with assistance from JVC, JH, and CH. All authors critically revised and approved the final manuscript.

### Acknowledgements

We thank the patients who have elected to be a part of this study. We also thank the personnel of the SCAN-B laboratory and the Division of Oncology and Pathology, Lund University, for handling samples, genomic analyses, and administrative support; SCIBLU Genomics for performing microarray experiments; and all of the SCAN-B Collaborators within surgery, oncology, pathology, and clinical chemistry at Hallands Hospital Halmstad, Helsingborg Hospital, Blekinge County Hospital, Central Hospital Kristianstad, Skåne University Hospital Lund/Malmö, and Central Hospital Växjö, for inclusion of patients and sampling of tissue and blood; Unilabs for preoperative biopsies; the Region Skåne Biobank for storage of blood samples; and the Regional Cancer Center South for INCA data. This study was generously supported by the Mrs. Berta Kamprad Foundation, and was also funded in part by the Swedish Foundation for Strategic Research, Swedish Research Council, Swedish Cancer Society, Knut and Alice Wallenberg Foundation, VINNOVA, Governmental Funding of Clinical Research within National Health Service, Crafoord Foundation, Lund University Medical Faculty, Gunnar Nilsson Cancer Foundation, Skåne University Hospital Foundation, BioCARE Research Program, King Gustav Vth Jubilee Foundation, and the Krapperup Foundation.

### Author details

[1]Department of Clinical Sciences, Division of Oncology and Pathology, Lund University, Medicon Village 404-A2, SE-22381 Lund, Sweden. [2]Lund University Cancer Center, SE-22381 Lund, Sweden. [3]CREATE Health Strategic Centre for Translational Cancer Research, Lund University, SE-22381 Lund, Sweden. [4]Department of Pathology, Skåne University Hospital, SE-22185 Lund, Sweden. [5]Department of Clinical Sciences, SCIBLU Genomics, Lund University, SE-22381 Lund, Sweden. [6]Department of Pathology and Cytology, Blekinge County Hospital, SE-37185 Karlskrona, Sweden. [7]Department of Pathology, Skåne University Hospital, SE-20502 Malmö, Sweden. [8]Department of Surgery, Lund University and Skåne University Hospital, SE-20502 Malmö, Sweden. [9]Department of Oncology, Skåne University Hospital, SE-22185 Lund, Sweden. [10]Department of Laboratory Medicine, Division of Molecular Pathology, Lund University, SE-22185 Lund, Sweden. [11]Department of Surgery, Lund University and Skåne University Hospital, SE-22185 Lund, Sweden.

Received: 18 September 2014 Accepted: 15 January 2015
Published online: 02 February 2015

### References

1. Engholm G, Ferlay J, Christensen N, Bray F, Gjerstorff ML, Klint A, et al. NORDCAN–a Nordic tool for cancer information, planning, quality control and research. Acta Oncol. 2010;49:725–36.
2. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. Lancet. 2011;377:127–38.

3. Brenner H, Hakulinen T. Very-long-term survival rates of patients with cancer. J Clin Oncol. 2002;20:4405–9.
4. Dodwell D, Thorpe H, Coleman R. Refining systemic therapy for early breast cancer: difficulties with subtraction. Lancet Oncol. 2009;10:738–9.
5. Gordon L, Scuffham P, Hayes S, Newman B. Exploring the economic impact of breast cancers during the 18 months following diagnosis. Psychooncology. 2007;16:1130–9.
6. Armstrong K. Can genomics bend the cost curve? JAMA. 2012;307:1031–2.
7. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010;11:685–96.
8. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
9. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.
10. Ahmed AA, Brenton JD. Microarrays and breast cancer clinical studies: forgetting what we have not yet learnt. Breast Cancer Res. 2005;7:96–9.
11. Reis-Filho JS, Westbury C, Pierga JY. The impact of expression profiling on prognostic and predictive testing in breast cancer. J Clin Pathol. 2006;59:225–31.
12. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene-expression profiles in hereditary breast cancer. N Engl J Med. 2001;344:539–48.
13. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. 2001;61:5979–84.
14. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proc Natl Acad Sci U S A. 2007;104:7564–9.
15. Staaf J, Ringner M, Vallon-Christersson J, Jonsson G, Bendahl PO, Holm K, et al. Identification of subtypes in human epidermal growth factor receptor 2–positive breast cancer reveals a gene signature prognostic of outcome. J Clin Oncol. 2010;28:1813–20.
16. Jonsson G, Staaf J, Vallon-Christersson J, Ringner M, Gruvberger-Saal SK, Saal LH, et al. The retinoblastoma gene undergoes rearrangements in BRCA1-deficient basal-like breast cancer. Cancer Res. 2012;72:4028–36.
17. Sweden Cancerome Analysis Network - Breast. Available at: http://scan.bmc.lu.se/.
18. Alkner S, Bendahl PO, Ferno M, Manjer J, Ryden L. Prediction of outcome after diagnosis of metachronous contralateral breast cancer. BMC Cancer. 2011;11:114.
19. Lund University. Faculty of Medicine - Oncology and Pathology. Available at: http://www.med.lu.se/canceromics.
20. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol. 2002;3: SOFTWARE0003.
21. Troein C, Vallon-Christersson J, Saal LH. An introduction to BioArray Software Environment. Methods Enzymol. 2006;411:99–119.
22. Vallon-Christersson J, Nordborg N, Svensson M, Hakkinen J. BASE–2nd generation software for microarray data management and analysis. BMC Bioinformatics. 2009;10:330.
23. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37:e123.
24. Nalpas NC, Park SD, Magee DA, Taraktsoglou M, Browne JA, Conlon KM, et al. Whole-transcriptome, high-throughput RNA sequence analysis of the bovine macrophage response to Mycobacterium bovis infection in vitro. BMC Genomics. 2013;14:230.
25. Borgstrom E, Lundin S, Lundeberg J. Large scale library generation for high throughput sequencing. PLoS One. 2011;6:e19119.
26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
27. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
28. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.
29. Brueffer C. TopHat Recondition. Python script. Available at: https://github.com/cbrueffer/tophat-recondition.
30. Morgan M, Pages H. Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import. R package version 1.12.4. Available at: http://www.bioconductor.org/packages/release/bioc/html/Rsamtools.html.
31. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003;100:8418–23.
32. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006;7:96.
33. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
34. Allen JD, Wang S, Chen M, Girard L, Minna JD, Xie Y, et al. Probe mapping across multiple microarray platforms. Brief Bioinform. 2012;13:547–54.
35. Atlas TCG. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
36. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486:400–4.
37. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.
38. bam-readcount. Available at: https://github.com/genome/bam-readcount/.
39. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.
40. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
41. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
42. The Cancer Genome Atlas. Available at: https://tcga-data.nci.nih.gov/.
43. Gene Expression Omnibus. Available at: http://www.ncbi.nlm.nih.gov/geo/.
44. Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, et al. PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. Cancer Res. 2005;65:2554–9.
45. Budhu A, Forgues M, Ye QH, Jia HL, He P, Zanetti KA, et al. Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. Cancer Cell. 2006;10:99–111.
46. Jais JP, Haioun C, Molina TJ, Rickman DS, de Reynies A, Berger F, et al. The expression of 16 genes related to the cell of origin and immune response predicts survival in elderly patients with diffuse large B-cell lymphoma treated with CHOP and rituximab. Leukemia. 2008;22:1917–24.
47. Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, et al. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clin Cancer Res. 2009;15:284–90.
48. Criscitiello C, Azim Jr HA, Schouten PC, Linn SC, Sotiriou C. Understanding the biology of triple-negative breast cancer. Ann Oncol. 2012;23:vi13–18.
49. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486:395–9.
50. Tang X, Baheti S, Shameer K, Thompson KJ, Wills Q, Niu N, et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. Nucleic Acids Res. 2014;42:e172.
51. Wilkerson MD, Cabanski CR, Sun W, Hoadley KA, Walter V, Mose LE, et al. Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res. 2014;42:e107.

# The Sweden Canceromics Analysis Network – Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine

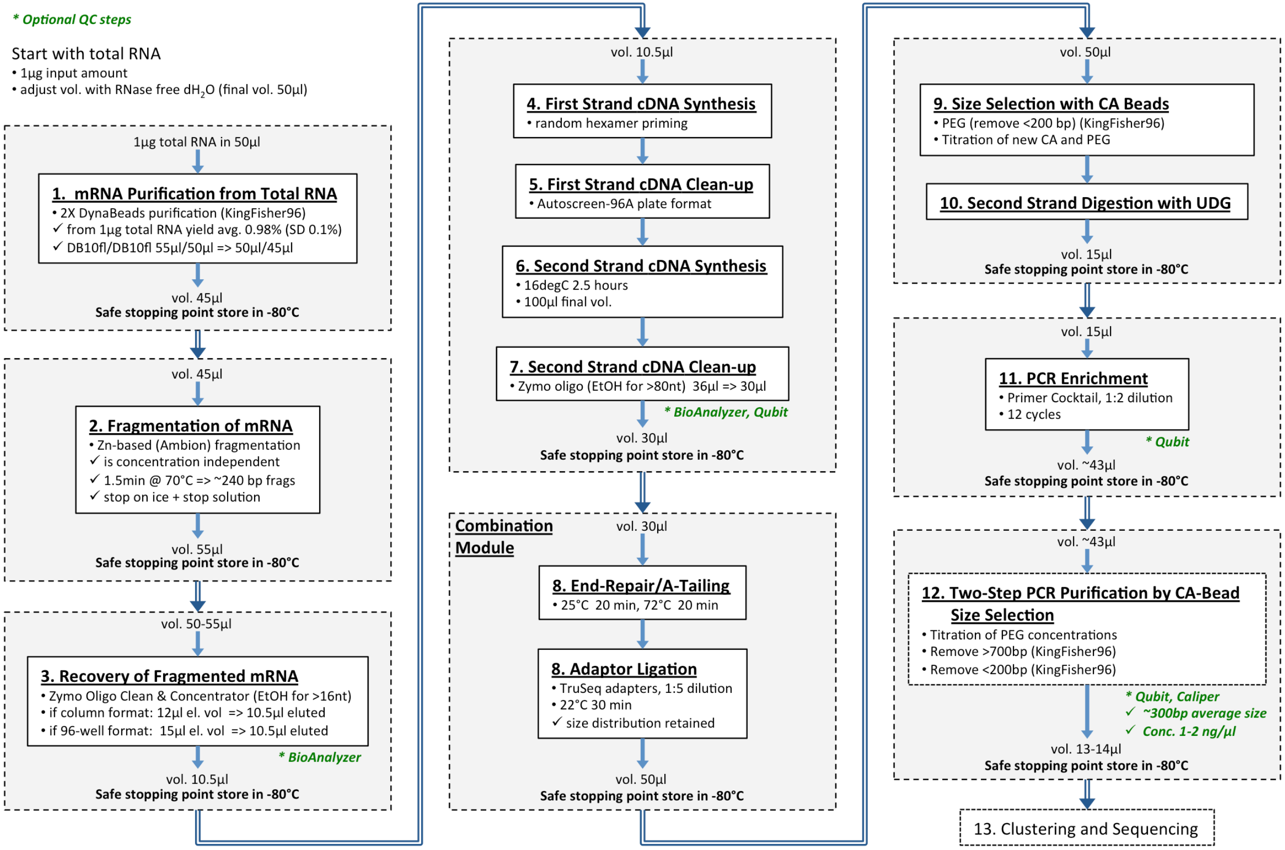
## ADDITIONAL FILE 1

Lao H. Saal, Johan Vallon-Christersson, Jari Häkkinen, Cecilia Hegardt, Dorthe Grabau, Christof Winter, Christian Brueffer, Man-Hung Eric Tang, Christel Reuterswärd, Ralph Schulz, Anna Karlsson, Anna Ehinger, Janne Malina, Jonas Manjer, Martin Malmberg, Christer Larsson, Lisa Rydén, Niklas Loman, and Åke Borg

This appendix contains Figure S1 and two SCAN-B protocols, Tumor Sample Processing, and RNA-Sequencing Library Preparation.

(next page)
Figure S1 – Schema for RNA-sequencing library preparation.

# SCAN-B directional RNA-seq Library Preparation



** Optional QC steps*

Start with total RNA
- 1µg input amount
- adjust vol. with RNase free dH$_2$O (final vol. 50µl)

1µg total RNA in 50µl

**1. mRNA Purification from Total RNA**
- 2X DynaBeads purification (KingFisher96)
- ✓ from 1µg total RNA yield avg. 0.98% (SD 0.1%)
- ✓ DB10fl/DB10fl 55µl/50µl => 50µl/45µl

vol. 45µl
**Safe stopping point store in -80°C**

vol. 45µl

**2. Fragmentation of mRNA**
- Zn-based (Ambion) fragmentation
- ✓ is concentration independent
- ✓ 1.5min @ 70°C => ~240 bp frags
- ✓ stop on ice + stop solution

vol. 55µl
**Safe stopping point store in -80°C**

vol. 50-55µl

**3. Recovery of Fragmented mRNA**
- Zymo Oligo Clean & Concentrator (EtOH for >16nt)
- if column format: 12µl el. vol => 10.5µl eluted
- if 96-well format: 15µl el. vol => 10.5µl eluted

** BioAnalyzer*

vol. 10.5µl
**Safe stopping point store in -80°C**

vol. 10.5µl

**4. First Strand cDNA Synthesis**
- random hexamer priming

**5. First Strand cDNA Clean-up**
- Autoscreen-96A plate format

**6. Second Strand cDNA Synthesis**
- 16degC 2.5 hours
- 100µl final vol.

**7. Second Strand cDNA Clean-up**
- Zymo oligo (EtOH for >80nt) 36µl => 30µl

** BioAnalyzer, Qubit*

vol. 30µl
**Safe stopping point store in -80°C**

**Combination Module**

vol. 30µl

**8. End-Repair/A-Tailing**
- 25°C 20 min, 72°C 20 min

**8. Adaptor Ligation**
- TruSeq adapters, 1:5 dilution
- 22°C 30 min
- ✓ size distribution retained

vol. 50µl
**Safe stopping point store in -80°C**

vol. 50µl

**9. Size Selection with CA Beads**
- PEG (remove <200 bp) (KingFisher96)
- Titration of new CA and PEG

**10. Second Strand Digestion with UDG**

vol. 15µl
**Safe stopping point store in -80°C**

vol. 15µl

**11. PCR Enrichment**
- Primer Cocktail, 1:2 dilution
- 12 cycles

** Qubit*

vol. ~43µl
**Safe stopping point store in -80°C**

vol. ~43µl

**12. Two-Step PCR Purification by CA-Bead Size Selection**
- Titration of PEG concentrations
- Remove >700bp (KingFisher96)
- Remove <200bp (KingFisher96)

** Qubit, Caliper*
- *✓ ~300bp average size*
- *✓ Conc. 1-2 ng/µl*

vol. 13-14µl
**Safe stopping point store in -80°C**

13. Clustering and Sequencing

# Table of Contents

# Introduction

This is a custom protocol for the tissue sample treatment of tumor samples for genomic analysis of breast cancer within the SCAN-B Initiative. The procedure is divided into preservation, partitioning, histology, and isolation using the AllPrep method of total RNA for fragments > 200bp, DNA isolation, and possibility for protein isolation and small RNA isolation from the retained flow-through fraction.

Samples: Preserved tumor samples (or fresh/snap frozen specimens),
optimal weight: 5-30 mg (min 1 mg, max 50 mg).

# 1. Tissue Preservation

## 1.1. Clinical Pathology

1.1.1. Tissues should be transported and manipulated on ice.

1.1.2. Tissue specimens of approximately 3x3x3 mm are collected but a single smaller piece may suffice. For efficient preservation, the widest dimension of a collected sample should not exceed 5 mm.

1.1.3. Place specimen(s) in pre-aliquoted SCAN-B tubes (Corning 2.0 ml cryogenic vials, cat# 430659) containing 1 ml RNAlater (at least 5 parts RNAlater to 1 part specimen).

1.1.4. Use separate tubes for multiple specimens if they are considered separate samples, e.g., a multifocal tumor or a lymph node. Annotate details on barcode stickers and referral form.

1.1.5. Apply barcode sticker and fill in SCAN-B pathology form.

1.1.6. Place tube in 4°C and record date/time. RNAlater should be allowed to penetrate specimen for at least 16 hours prior to processing or freezing.

1.1.7. Ship to Lund Pathology following schedule and using inter-hospital transport.

1.1.8. At Lund Pathology, all SCAN-B specimen tubes are placed in Lund SCAN-B 4°C refrigerator to await pick-up by SCAN-B personnel.

1.1.9. Tested RNA stability after the 16h @ 4°C incubation:

| Temperature | Stability |
|---|---|
| 4°C | = up to 4 weeks (we aim to store at 4°C for < 5 days) |
| RT | = up to 5 days |
| -80°C | = long term |

## 2. Collection and LIMS Registration

Specimens and referral forms are delivered in cold-pack insulated carrier bags.

### 2.1. Collection and Registration in LIMS

2.1.1. Match collected specimen tubes with SCAN-B pathology forms. If a pathology form or specimen tube is missing, notify the clinical liaison.

2.1.2. Note time of surgery and check which specimens can be partitioned the same day. Specimens should be stored at 4°C for at least 16 hours before partitioning or freezing.

2.1.3. Log in to: *BASE → Extensions → Reggie → Sample processing wizards → Specimen tube registration.*

2.1.4. *Enter Case Name (step 1 of 3)*. Use the barcode reader to enter Case name (“studie löp nr”) → *Next* to proceed.

2.1.5. *Enter Case information (step 2 of 3)*. Fill in: Number of tubes, Arrival date, Sampling date and time, RNA Later date and time, Laterality, Specimen type, Biopsy type, Other path note; → *Next* to proceed.

2.1.6. *Enter tube information (step 3 of 3)*. Assigned box number in freezer will come up. Fill in delivery comment → *Create* to save.

2.1.7. When the referrals have been registered in BASE, leave them in the "UT" tray. Unregistered referrals may be left in the "IN" tray. *Note: The referrals must always be kept in the secure location.*

2.1.8. Go to the Brady computer and log in to *BASE* → *Extensions* → *Reggie* → *Sample processing wizards* → *Partition registration wizard*. Select specimen tubes that you want to process → *next* → *download label files*. Save downloaded file on desktop as CSV with the date as filename (e.g., 20120221.csv).

2.1.9. Open the software Multiple_Brady_Printer → *Data sources* → *Database* → *Create*. Under *Select a datasource* choose csv and select the file that was downloaded and saved on the desktop → *ok*.

2.1.10. In software Multiple_Brady_Printer, Select *Print direct* under Print to check if the number is in the correct position on the printed labels, preferably as far down as possible. If needed change position on screen or move the label in the printer. When it is ok select → *Print* → *More* → *all records* → *ok*.

2.1.11. Collect printed labels for lab tubes: one for AllPrep, Histology, and the Remainder.

# 3. Specimen Partitioning

Optimally we process a ~30 mg piece of each specimen for extraction of nucleic acids using the AllPrep method. The tumor pieces are delivered in tubes containing 1 ml of RNAlater. Store them at 4°C for at least 16 hours from operation time to allow full penetration of tissue with the preservative. All received samples must be properly registered in BASE prior to sample processing.

## 3.1. Before you start

3.1.1. Spray lab benches with Ambion RNAse Away (cat# 10328-011) and wipe them dry.

3.1.2. Put a sterile blanket on the bench; change when needed (if possible every day).

3.1.3. Prepare 2.0 ml Eppendorf Safe-Lock tubes with pre-printed barcoded labels for AllPrep, Histology, and Remainder. Add 500 µl RNAlater to tubes for Histology and Remainder (no RNAlater for AllPrep tubes). *Write last 3 digits of barcode on top of tubes.*

3.1.4. Prepare for tissue dissection by having petri dishes (35 x 10 mm) and single use scalpels (blade no. 10) for each tissue sample to be dissected.

### 3.2. Partitioning

3.2.1. Log in to the laptop next to the scale (partitioning lab bench), "wake" the scale by pressing any button. Press the green icon (Metler Toledo Balance link software) select → *Configuration → interface → Port, COM3 →ok.*

3.2.2. Log in to *BASE → Extensions → Reggie → Sample processing wizards → Partition registration wizard.* Select specimens to be partitioned and press → *Next*.

3.2.3. Put fresh petri dish on the bench and get the first tissue piece out of its tube and onto the dish. Make note of the appearance of the tumor and enter the *Number of pieces* (NofPieces) in the BASE form.

3.2.4. Cut off desired sections with a scalpel. In priority order: one representative piece for AllPrep (~30 mg), one section immediately adjacent to be used for Histology (5-10 mg), and one for Remainder (if there is any left). The AllPrep piece should be subpartitioned into smaller pieces to enhance tissue lysis. Use single-use scalpels and forceps to handle each sample. *Note: check the box 'Mult' if sampling multiple pieces is required to get material for AllPrep.*

3.2.5. If the delivered specimen is <15 mg, then only take a piece for AllPrep. If the specimen is exactly 15 mg, then take 10 mg to Allprep and 5 mg to Histology.

3.2.6. Dispose of used single-use utilities and discard the scalpels in the yellow sharps box.

3.2.7. Repeat for each tissue sample.

3.2.8. As prompted by BASE, weigh each tube containing the 500 µl RNAlater, tare, and place the appropriate cut section(s) into the labeled tube, and record in BASE by pressing print on the scale.

3.2.9. If Histology piece cannot be sampled, write 0 in the HisWeight field.

3.2.10. Place the tubes with the pieces for AllPrep, Histology, and Remainder in their assigned storage locations in the -80°C freezer (AllPrep and Remainder) or 4°C fridge (Histology).

## 4. Tissue Disruption and Lysis

Tissue samples are lysed and homogenized with the TissueLyzer (TL) and the RNA/DNA is extracted using the AllPrep method, automated on the QIAcube (all from Qiagen). Flow-though fraction is saved for eventual isolation of small RNAs and proteins.

### 4.1. Before you start

4.1.1. Log in to *BASE* → *Extensions* → *Reggie* → *Sample processing wizards* → *DNA/RNA extraction wizards* → *Lab Tracking Protocol for Allprep isolation*.

4.1.2. Select the 12 samples that are next in line under *Select unprocessed lysate items* → *Finish* → *Print*.

4.1.3. Make sure that the samples selected are in the AllPrep storage box in the -80°C freezer and check the "ApWeight" (printed in remark). If any sample is less than 10 mg or more than 50 mg this sample should be treated differently (described later in this protocol).

4.1.4. Chill the TissueLyser (TL) adapter at -20°C for at least 1 hour.

4.1.5. Take out samples from freezer to thaw at room temperature if they are in RNAlater. If fresh frozen tissue is being used, it should at all times be on dry ice.

4.1.6. Retrieve the remaining pre-printed labels for all samples to be processed.

### 4.2. Prepare lysis buffer

4.2.1. Always use Eppendorf 2.0 ml Safe-Lock tubes for disruption/lysis. Mark two tubes/sample:

- One with the label ending with ".l" (lysate) and the number of the sample on the lid (together with position in Qiacube).
- The other with the last three numbers of the sample on the side and the QIAcube number on the lid (for AllPrep isolation)

4.2.2. From this step work on the ventilated bench. Prepare the "Lysis buffer mix" (800 µl/sample). Mix 790 µl RLT Plus lysis buffer + 8 µl 2-Mercaptoethanol (BME). Make a lysis mastermix for 13 samples if you are processing 12 samples (10,270 µl RLT + 104 µl BME in a 15 ml tube). Be aware that, if any of the samples are over 50 mg, additional buffer will be needed; therefore check this before preparing the mastermix.

4.2.3. Rinse the steel beads (5 mm) to be used (2 per sample) in 1 ml fresh RLT Plus lysis buffer using a 2 ml tube and a clean pair of forceps (maximum 12 beads in one tube).

### 4.3. Tissue disruption

4.3.1. Quickspin the thawed sample tubes and remove the RNAlater solution (~600 µl). Be sure that all the salt precipitate is dissolved.

4.3.2. Add 2ul Reagent DX-Antifoaming reagent to each sample.

4.3.3. Place 2 steel beads in each sample tube and store the samples in the pre-cooled TL adapter (12 samples/run).

4.3.4. Add 400 µl lyses buffer mix and disrupt the samples in the TissueLyser, disrupt the samples at 50 Hz for 2 x 4 minutes (pause in-between).

4.3.5. After disruption centrifuge the samples briefly and keep at room temperature

4.3.6. For samples 10-50 mg, add an additional 400 µl lysis buffer and mix well

4.3.7. For samples <10 mg, do not add additional lysis buffer.

4.3.8. For samples >50 mg, add 600 µl lysis buffer.

4.3.9. Add up to 800 µl lysed sample to the pre-labeled QIAshredder columns.

4.3.10. Centrifuge at 16000 g for 5 minutes at room temperature. Spin longer if precipitate appears loose in order to pellet it.

4.3.11. Flow-through contains RNA/DNA and protein. Continue with 350 µl of flow-through lysate for RNA/DNA isolation in a new tube, and store the remaining ~390 µl in a new labeled tube for storage (suffix ".l"), taking care to leave behind any precipitate. If the specimen was <10 mg, then there will be no lysate for storage. If the specimen was >50 mg, then repeat loading the same QIAshredder column and split the flow-through such that 350 µl is used immediately for RNA/DNA isolation and the remainder lysate is stored for future use.

4.3.12. Store the homogenized samples for at least 30 minutes at -80°C or on dry ice before continuing with RNA/DNA isolation.

4.3.13. The homogenized samples for storage should be put in the "lys" box in the position indicated by BASE.

4.3.14. Dispose of BME waste appropriately.

# 5. AllPrep RNA/DNA/flow-through isolation using the QIAcube

This is the protocol for isolation of RNA and DNA from 350 μl homogenized tissue sample in RLT Plus buffer using the AllPrep method. Customized QIAcube protocols were developed for SCAN-B which modified slightly the standard protocols provided by Qiagen. Additional instructions are found in the QIAcube Protocol Sheet and in the Customized Protocol General Information ID 1608 v2. Protocols are available under:

RNA → AllPrep DNA/RNA Mini Kit → Animal tissues and cells →
Part A (DNA Purification): → Custom part A
Part B (protein/smallRNA FT):→ AllPrep Mod 1 part B
Part C (RNA Purification): → AllPrep Mod 1 part C

## 5.1. Preparation

5.1.1. Thaw the samples processed above and bring them to RT. Continue using the "Lab Tracking Protocol for Allprep isolation" printed out from BASE.

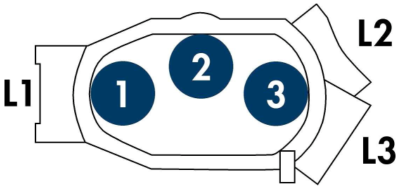
5.1.2. Retrieve the labels previously printed out.

5.1.3. The label ending with ".d" (DNA) and label ending with ".r" (RNA) should go on 1.5 ml safe-lock tubes (cat# , Eppendorf), and mark the lids with the sample ID.

5.1.4. The last label, ending with ".ft" (flow through) should be put on a 1.5 ml protein low-binding safe-lock tube (cat# , Eppendorf). Mark the lid with the number.

5.1.5. Mark (using pen) two Rotor adapter and two 1.5 ml collection tubes (from the AllPrep kit) for each sample, with QIAcube position number. Put them on two separate rotor adapter holders.

5.1.6. Put marked collection tube in position 3 on respective Rotor adapter and fasten lid in position L3 (see picture below or "Protocol sheet").



5.1.7. Fill up with Filter-tips (1000 μl) in the QIAcube.

5.1.8. Prepare one 2 ml safe-lock tube (cat# Eppendorf) with RNAse free water (for 12 samples 1336 μl) for the RNA elution step and put the tube in slot A in the QIAcube.

For fewer samples than 12 see volume "Protocol Sheet". It is important to have the exact volume that is described in the protocol sheet.

5.1.9. Check the "QIAcube-AllPrep-Buffer Exchange List" if any reagents needs to be exchanged. Put the bottles in the Reagent Bottle Rack in correct position (see Protocol Sheet). Fill up with solutions from kit if needed (it should be at least 2/3 in each bottle). Note: mark the bottle with date if a new bottle is opened.

5.1.10. Remove the caps from the reagent bottles and put them in the assigned box. Put the reagent bottle rack in the QIAcube.

5.1.11. Cut the lids off the AllPrep DNA spin columns.

5.1.12. Work on the ventilated bench after running each QIAcube program.

### 5.2. Part A (DNA isolation)

5.2.1. Place the labeled elution tubes in position 3 and the lidless AllPrep DNA spin columns into position 2 of each rotor adapters. (Position 1 is empty).

5.2.2. Place the rotor adapters into the QIAcube centrifuge.

5.2.3. Quickspin the thawed samples lysates and place them in their correct positions in the QIAcube shaker.

*5.2.4.* Close the QIAcube and start the program → Custom part A. Run time is approximately 30 minutes for 12 samples. *Note: This is a modified program; importantly an additional 1 minute incubation following addition of wash buffer was added.*

5.2.5. At the end of the program take out the rotor adapters and place into the rotor adapter holder.

5.2.6. Remove the DNA spin column and save the labeled DNA elution tube on ice.

5.2.7. Transfer the eluted DNA to a new labeled 1.5 ml safe-lock DNA tube.

5.2.8. Transfer the flow-through (containing total RNA and protein) from the *used rotor adapter Position 2* (approximately 310 μl) into *Position 2 of a fresh rotor adapter*.

### 5.3. Part B (collection of flow-through for protein/smallRNA)

5.3.1. Prepare a second set of Rotor adapters.

5.3.2. Place clean RNeasy spin columns in position 1 with the lid fastened in position L1 of the new adapters.

5.3.3. Place RNA elution tubes into position 3 of the new adapters.

5.3.4. Transfer the flow-through (containing total RNA and protein) from the *used rotor adapter Position 2* (~310 µl) from above into *Position 2 of the new rotor adapter*.

5.3.5. Place the rotor adapters into the correct positions in the QIAcube.

5.3.6. Close the QIAcube and start the program → AllPrep Mod 1 part B. Run time is approximately 13 minutes for 12 samples. *Note: This is a modified program to allow for saving of the flow-through which contains the protein and small RNA fractions.*

5.3.7. At the end of the program take out the rotor adapters and place into the rotor adapter holder.

5.3.8. Save the protein/small RNA flow-through (550 µl) from the *space between the adapter positions* and transfer to a new pre-labeled 1.5 ml protein low-binding safe-lock tube and store at -80°C according to position indicated by BASE.

5.3.9. Place the rotor adapters back into the QIAcube maintaining the same centrifuge adapter holder positions as before.

### 5.4. Part C (complete RNA isolation)

5.4.1. Close the Cube and start the program → AllPrep Mod 1 part C. Run time is approximately 23 minutes for 12 samples.

5.4.2. At the end of the program take out the rotor adapters and place into the rotor adapter holder.

5.4.3. Remove the RNeasy spin columns and save the labeled RNA elution tubes on ice.

5.4.4. Transfer the eluted RNA into a new pre-labeled 1.5 ml safe-lock tube and store on ice.

5.4.5. Take out reagent bottles and put the lids on to prevent ethanol evaporation.

5.4.6. Clean the QIACube with 70% ethanol if needed and then clean out used tips from drawer and clean the inside of the drawer with 70% EtOH and put it back again. Dispose of waste appropriately.

5.4.7. Store all samples at -80°C in labeled boxes and in positions as directed by BASE.

# 6. RNA and DNA QC

## 6.1. Nanodrop

6.1.1. Directly after isolation of RNA and DNA, log in to the 8-channel Nanodrop (ND-8000) computer and log in to *BASE → Extensions → Reggie → Sample processing wizards → DNA/RNA extraction wizards → DNA/RNA/Flowthrough registration.*

6.1.2. Select the 12 samples that have been extracted under Select unprocessed lysate items → *Next*; fill in information on Lysis and Qiacube run under Common information from Lysate and Qiacube step → *Next*.

6.1.3. Under *RNA/DNA/Flowthrough details → NanoDrop Sample ID → click Download* → save file on desktop (remove file when finished).

6.1.4. Open the ND-8000 program and upload the file when prompted (Load Sample ID file). The DNA samples are now listed on row 1 and 2 and the RNA samples are listed on 3 and 4. *If any of the samples get an error message or look strange do the re-measurement in the same position. Measurements that are done in other positions will not be uploaded into BASE.*

6.1.5. After measurement of samples save the report as a "txt" file on the server *sky1:\scanb\SCAN-B\Nanodrop_resultat* in folder with appropriate "month-year". Name the file with "date of extraction + _DNA_RNA". Print the file.

6.1.6. Go back to *BASE Extensions → Reggie → RNA/DNA/FlowThrough details → NanoDrop values → click 'Browse' → Select the results file from ND measurement → open.* The concentrations should now appear for each sample.

6.1.7. Check that the values are the same as on the ND results file. If any of the samples have been re-measured BASE will automatically take the measurement with the highest value, *if this is not correct change the value in BASE.*

6.1.8. Check that the lysate volumes are correct. If samples had a weight < 10 mg the Lysate *Total* is only 350 µl and if a sample had a weight > 50 mg the Lysate *Total* is 900 µl. For normal samples the lysate is 700 µl.

6.1.9. If anything needs to be changed or a comment is to be added *press → Edit… → change what's needed → ok → Finish.*

### 6.2. Preparation of RNA aliquots for Caliper run

6.2.1. In *BASE* → *Reggie* → *Sample processing wizards* → *RNA quality control wizards* → *Create aliquots on Bioanalyzer/Caliper plates* → Select the RNA extracts to be aliquoted → *Next.*

6.2.2. Positions of samples on Caliper plate and concentrations appear and if the RNA sample concentration is <35 ng/µl the HS (High Sensitivity assay on Caliper) box is ticked.

6.2.3. Write down the positions of the samples on the Caliper plate in the comment box on the "Lab Tracking Protocol for Allprep isolation" for current run. Also note if any of the samples are to be run with the HS assay; press → *Finish*.

6.2.4. Prepare 12-tube strips for RNA quality check on Caliper, write the 3 last numbers of the sample on each tube on the strip. For HS, aliquot 2 µl per sample to the strip. For Std take 2 µl up to 200 ng/µl, and if at a higher concentration then dilute to <200 ng/µl and aliquot 2 µl to the Caliper strip.

6.2.5. Put strip caps on and label them with plate position and the ends with column number and put them in the correct position in the specific Caliper plate in the -80°C freezer.

6.2.6. When the plate is complete it will be run on Caliper. The results are stored directly in BASE. For Caliper instructions see protocol "Caliper LabChip GX HT RNA Assay user instructions".

6.2.7. Store all samples at -80°C in correct positions in labeled boxes as indicated by BASE.

6.2.8. Put the completed "Lab Tracking Protocol for Allprep isolation" with the print out of the NanoDrop results in folder named "SCAN-B QC".

## 7. Equipment, Consumables, and Reagents

**Equipment:**

- QIAcube (cat# 9001293, Qiagen)
- TissueLyser LT (cat# 85600, Qiagen)
- Caliper LabChip GX (cat# 122000, PerkinElmer)

**Consumables and Reagents:**

- Cyogenic vial 2 ml (cat# 430659, Corning)
- Brady label stickers (cat# BPT-628-461, Brady)
- Petri dish 35 x 10 mm (cat# 353001, Falcon)
- Single-use no. 10 scalpels (cat# REF0501, Swann-Morton)
- AllPrep DNA/RNA Mini Kit (cat# 80204, Qiagen)
- Stainless Steel Beads 5 mm (cat# 69989, Qiagen)
- TissueLyser 2x24 adapter set (cat# 69982, Qiagen)
- Reagent DX (cat# 19088, Qiagen)
- Buffer RLT Plus (cat# 1053393, Qiagen)
- RNAlater 500 ml (cat# AM7021, Ambion)
- Safe-lock Tubes 2.0 ml (cat# 0030 123.344, Eppendorf)
- Safe-lock Tubes 1.5 ml (cat# 0030 123.328, Eppendorf)
- Protein low-bind Safe-lock Tubes 1.5 ml (cat# 0030 097.221, Eppendorf)
- 2-Mercaptoethanol >98% (cat# M3148-100ml, Sigma)
- QIAshredder columns (250 pcs) (cat# 79656, Qiagen)
- QIAcube Filter Tips 1000 µl (cat# 990352, Qiagen)
- QIAcube Rotor Adapters (240 pcs) (cat# 990394, Qiagen)
- QIAcube Reagent Bottle (30 ml) (cat# 990393, Qiagen)
- 12-tube strips (200 µl) (cat# 732-0553, VWR)

# Table of Contents

# Introduction

This is the SCAN-B RNA-seq protocol for preparation of Illumina-compatible sequencing libraries from total RNA in a high-throughput format with an integrated workflow together with the BioArray Software Environment (BASE) laboratory information management and analysis web-based platform. The library preparation method is an adaptation of the dUTP strand-specific method as described by Parkhomchuk et al. (*Nucleic Acids Res* 2009) and modified by Lohan and colleagues (Nalpas et al., *BMC Genomics* 2013). Size selection using polyethylene glycol and carboxylic acid (CA)-beads is adapted from described methods by Borgstrom et al. (*PLOS One* 2011).

# 1. mRNA Purification from Total RNA

This is a modification of the Invitrogen protocol "Dynabeads mRNA DIRECT Kit (rev008)" (cat# 61012) for automatization on the ThermoScientific KingFisher Flex Magnetic Particle Processor (KF-Flex). Input material is 1.1 µg (1-2 µg can readily be used) total RNA diluted in 50 µl nuclease-free water. Two rounds of mRNA purification are performed to reduce ribosomal RNA. For breast cancer samples, the expected yield of mRNA after the 2nd round is approximately 1-2% of total RNA. Relative humidity in the KF-Flex should be stabilized by placing open plates with water inside prior to starting.

## 1.1. Diluting total RNA samples

1.1.1. Use BASE to create the worksheet for the next batch of total RNA samples to be purified: *BASE → Extensions → Reggie → Library preparation wizards → Lab protocols for mRNA and cDNA preparation*. Select mRNA bioplate and input concentration of the Stratagene Universal Human Reference RNA (cat# 740000, Agilent) being used. Print out List layout and Plate layout.

1.1.2. Dilute all samples in a 96-well 4titude PCR plate, 1.1 µg total RNA in 50 µl nuclease free water.

*Note: this is an optional stopping-point: plate can be sealed and stored at -80℃.*

1.1.3. Vortex plate and spin it down before incubation.

1.1.4. Incubate the samples at 65°C for 5 min in a thermocycler (e.g. Eppendorf vapo.protect).

1.1.5. Place denatured diluted total RNA samples on ice.

## 1.2. Preparation of Binding Plates

1.2.1. In advance, contents of Ambion mRNA Purification Kit should be brought to room temperature (allow at least 20 min).

1.2.2. Vortex the Dynabeads and transfer 100 µl Dynabeads per well into the Binding Plate1 (KingFisher 96 plate 200 µl, cat# 97002540).

*Note: 100 µl/well Dynabeads is prepared to cover both round #1 and round #2.*

1.2.3. Remove suspension buffer by applying plate to 96-position magnetic stand and wait until the solution is completely clear (1-2 min); then discard the supernatant by pipetting.

1.2.4. Wash beads: add 100 µl Binding Buffer and mix by pipetting.

1.2.5. Remove Binding Buffer by applying plate to 96-position magnetic stand, wait until clear (1-2 min), and discard the supernatant by pipetting.

1.2.6. Add 100 µl Binding Buffer and mix by pipetting.

1.2.7. Per well, transfer 50 µl of the resuspended Dynabeads to Binding Plate2 for later use in round #2 (see section below).

1.2.8. Add 50 µl of the diluted total RNA samples to each well in Binding Plate1 and mix by pipetting (total volume 100 µl).

## 1.3. Preparing KF-Flex for round #1 of mRNA purification

1.3.1. Label and fill the KF 96 plates as follows:

| | | |
|---|---|---|
| 1 | - Tip Plate: | Put 1 new KingFisher 96 tip comb on to the TIP Plate |
| 2 | - Elution Plate: | 55 µl nuclease free water |
| 3 | - Wash Plate 2: | 60 µl Washing Buffer B |
| 4 | - Wash Plate 1: | 60 µl Washing Buffer B |
| 5 | - Binding Plate: | 50 µl prepared Dynabeads + 50 µl prepared Samples (as described above) |

## 1.4. Run KF-Flex, purification round #1

1.4.1. Make sure that the A1 positions on the plates are in the same corner as the A1 positions of the KF-Flex disc.

1.4.2. Start program mRNA_DB10fl, follow on-screen instructions.

1.4.3. After completed run, remove elution plate immediately and place on 96-position magnetic stand and wait until the solution is completely clear (1-2 min).

1.4.4. Transfer 50 µl of the samples (all) into a new PCR plate, seal the plate, and store it on ice until round #2.

*Note: when proceeding directly to round #2 samples are already denatured from the elution step, otherwise a separate denaturation should be added by incubation at 65℃ for 5 min in a thermocycler.*

## 1.5. Preparing KF-Flex for round #2

*Note: Dynabeads for Binding Plate2 is prepared together with Binding Plate for round #1 (above). Prepare for round #2 while round #1 is running in the KF-Flex.*

1.5.1. Label and fill the KF 96 plates as follows (note lower volume in Elution Plate compared to round #1):

| | | |
|---|---|---|
| 1 | - Tip Plate: | Put 1 new KingFisher 96 tip comb on the TIP Plate |
| 2 | - Elution Plate: | 50 µl nuclease free water |
| 3 | - Wash Plate 2: | 60 µl Washing Buffer B |
| 4 | - Wash Plate 1: | 60 µl Washing Buffer B |
| 5 | - Binding Plate2: | 50 µl prepared Dynabeads + 50 µl Sample from round #1 |

## 1.6. Run KF-Flex 2nd Round

1.6.1. Make sure that the A1 positions on the plates are in the same corner as the A1 positions of the KF-Flex disc.

1.6.2. Start program mRNA_DB10fl, follow on-screen instructions.

1.6.3. After completed run, remove elution plate immediately and place on 96-position magnetic stand and wait until the solution is completely clear (1-2 min).

1.6.4. Transfer 45 µl of the samples (all) into a new PCR plate, seal the plate, and store it on ice.

**Safe stopping point – may store at -80°C**

## 1.7. Comments

After two rounds of mRNA purification using Dynabeads, mRNA yield will be between 1-5% of the total RNA.

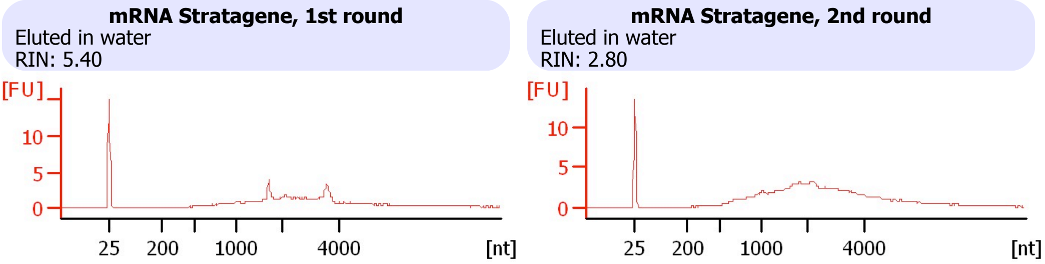


**Figure 1.** Example BioAnalyzer analysis of mRNA purification after 1 round and 2 rounds.

## 2. Fragmentation of mRNA

Purified mRNA is fragmented to ~240 bp fragments using the Ambion buffered zinc fragmentation reagents (cat# AM8740). In our hands, incubation for 1.5 minutes yields optimally-sized fragments. Fragmentation is robust to variation in input-RNA concentration and to delay in addition of Stop Buffer as long as samples are put on ice immediately after incubation.

### 2.1. Fragmentation of mRNA

2.1.1. Assemble the following reaction in PCR tube/strip/plate format:

| | |
|---|---|
| • mRNA | 45 µl |
| • 10X Fragmentation Reagent | 5 µl |

2.1.2. Incubate at 70°C for 1.5 minutes in a thermocycler.

2.1.3. Place the tube/plate on ice.

2.1.4. Add 5 µl of Stop Buffer.

**Safe stopping point – may store at -80°C**

## 3. Recovery of Fragmented mRNA

The Zymo Oligo Clean & Concentrator with high EtOH (~70%) is used to clean and concentrate fragmented mRNA larger than 16 nucleotides. This can be performed in either column (cat# D4061, Zymo Research) or 96-well plate (cat# D4063) format.

### 3.1. Recovery of fragmented mRNA

3.1.1. Add 100 µl Oligo Binding Buffer to a new deep well collection plate (from Zymo kit).

3.1.2. Transfer fragmented mRNA (~55 µl) to the collection plate and mix by pipetting.

3.1.3. Add 400 µl **absolute ethanol** to the collection plate with Oligo Binding Buffer and fragmented mRNA, mix briefly by pipetting and transfer mixture to the corresponding well of a Zymo-Spin I-96 Plate.

3.1.4. Centrifuge at 5000 g for 5 minutes. Discard the flow-through.

3.1.5. Add 800 µl Wash Buffer to all the wells on the plate.

3.1.6. Centrifuge the plate at 5000 g for 5 minutes and discard the flow-through.

3.1.7. To completely dry the plate, centrifuge at 5000 g for 7 minutes and discard any flow-through.

3.1.8. Transfer the plate onto an Elution Plate (in this case we use a PCR plate). Add 15 µl water directly to the column matrix (*make sure that the entire volume is on the matrix*) and let the plate stand for 1 minute. Centrifuge at 5 000 g for 7 minutes to elute the mRNA.

3.1.9. Transfer 10.5 µl of the eluate to a new PCR plate.
*Note: transfer to a new PCR plate is not required if an appropriate stand/rack is used to support (prevent damage) the PCR plate during centrifugation; however, verify that volumes are even across wells.*

### 3.2. Quality Control (optional)

Quality control of fragmented mRNA can be performed by RNA Pico Chip on the Bioanalyzer. Since an amount of the sample is lost when running quality control, samples must be pre-selected from the start and compensated by inputting an extra amount of total-RNA in the mRNA purification (add an extra 11% of total-RNA).

3.2.1. Aliquot 1.2 µl for selected samples to perform quality control on a RNA Pico Chip on the Bioanalyzer. Add back 1.2 µl water to these samples to compensate the total volume.

3.2.2. Use BASE to register mRNA plate and upload Bioanalyzer PDF file: *BASE → Extensions → Reggie → Library preparation wizards → mRNA registration and quality control results.*

**Safe stopping point – may store at -80°C**

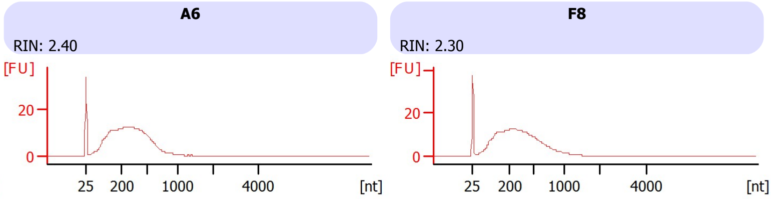


**Figure 2.** Example of typical QC results for fragmented mRNA from well A6 (left) and F8 (right).

# 4. First Strand cDNA Synthesis

First strand cDNA is synthesized using random hexamer priming, dNTP mix, and SuperScript II Reverse Transcriptase reagents (cat# 18064-014, Life Technologies).

## 4.1. Priming

4.1.1. Assemble the following reaction in a PCR tube/PCR plate

| | |
|---|---|
| • Fragmented mRNA | 10.5 µl |
| • Random hexamers (3 µg/µl) | 1 µl |

4.1.2. Mix well with a pipette.

4.1.3. Incubate 65°C for 5 minutes in a thermocycler.

4.1.4. Place tube/plate on ice.

## 4.2. cDNA synthesis

4.2.1. Mix the following 1st Strand Synthesis Mix, per sample (use SCAN-B calculator):

| | |
|---|---|
| • 5X First Strand buffer | 4 µl |
| • 0.1 M DTT | 2 µl |
| • 10 mM dNTP mix | 1 µl |
| • RNaseOUT (40 U/µl) | 0.5 µl |

4.2.2. Add 7.5 µl of 1st Strand Synthesis Mix to the mRNA+hexamers sample, mix well (total volume 19 µl).

4.2.3. Place in thermocycler, run the following program:

- 25°C 2 min
- **PAUSE**, add 1 µl SuperScript II, take care to mix well (mix w larger volume)
- Resume program, 25°C 10 min
- 42°C 50 min
- 70°C 15 min
- 4°C Hold

4.2.4. After completed program store samples on ice and proceed directly to clean-up.

## 5. First Strand cDNA Clean-up

First strand clean-up is performed using the Illustra AutoScreen-96A plate (cat# 25-9005-98, GE Healthcare) to remove unincorporated dNTPs. **Note:** Set out AutoScreen-96A plates ahead of time (~2 hr) as they must be used at room temperature or else performance will be erratic.

### 5.1. Prepare AutoScreen-96A Well Plate

5.1.1. Remove the AutoScreen-96A from the foil storage pouch.

5.1.2. Remove both the top and bottom adhesive seals and place it directly on a collection plate (U-bottom plate).

5.1.3. Centrifuge for 5 min at 910 g.

5.1.4. Add, drop wise, 150 µl of distilled water.

5.1.5. Centrifuge for 5 min at 910 g.

5.1.6. Discard collection plate and replace it with a fresh PCR plate (4titude PCR plate).

5.1.7. Slowly apply 20 µl of first strand cDNA to the center of the column resin bed in the AutoScreen-96A plate.

5.1.8. Centrifuge the samples for 5 min at 910 g.

5.1.9. Take note of eluted volume for each well. If needed add water to final volume of 16ul.

5.1.10. Store first strand collection plate on ice and proceed immediately to 2nd strand synthesis.

## 6. Second Strand cDNA Synthesis

Second strand synthesis incorporates dUTP instead of dTTP. To make a 400 µl mix of 10 nM of each nucleotide substituting dUTP for dTTP, take 40 µl of each 100 nM stock of dATP + dGTP + dCTP + dUTP, plus 240 µl $H_2O$. Prepare nucleotide mix in advance.

*Note: for a full 96-well plate a total of 306 µl is typically needed.*

### 6.1. Second strand synthesis

*Note: the recipe for 2nd Strand Synthesis Mix includes water for a final reaction volume of 100 µl. However, if elution volumes from first cDNA clean-up are variable, the volume of water in the 2nd Strand Synthesis Mix may be reduced and the final reaction volume adjusted with water after the 1$^{st}$ strand cDNA reaction sample has been added.*

6.1.1. Mix the following 2nd Strand Synthesis Mix, per sample (use SCAN-B calculator):

| | |
|---|---|
| 5X First Strand buffer | 1.3 µl |
| 5X Second Strand buffer | 20 µl |
| 10 mM dUTP+dATP+dGTP+dCTP mix | 3 µl |
| 0.1 M DTT | 1 µl |
| DNA Pol I (10 U/µl) | 5 µl |
| RNaseH (10 U/µl) | 0.2 µl |
| $H_2O$ | 53.5 µl |

6.1.2. Chill 2nd Strand Master Mix on ice for at least 5 minutes before aliquoting.

6.1.3. Add 84 µl of 2nd Strand Master Mix to each 1st strand cDNA reaction and mix (final total volume 100 µl).

6.1.4. Incubate at 16°C in a thermocycler for 2.5 hours (no heated lid).

6.1.5. Proceed to second strand cDNA clean-up.

# 7. Second Strand cDNA Clean-up

Second strand cDNA clean-up is performed using the Zymo Research Oligo Clean & Concentrator and low EtOH (~57%) to concentrate and clean oligonucleotides >80 bp. This can be performed in either column (cat# D4061, Zymo Research) or 96-well plate (cat# D4063) format.

## 7.1. Second strand cDNA clean-up

7.1.1. Add 200 µl Oligo Binding Buffer to a new collection plate.

7.1.2. Transfer 100 µl sample to the collection plate containing Oligo Binding Buffer.

7.1.3. Add 400 µl **absolute EtOH** to the collection plate with sample and Oligo Binding Buffer, mix briefly by pipetting and transfer mixture to the corresponding well of a Zymo-Spin I-96 Plate.

7.1.4. Centrifuge at 5000 g for 5 minutes. Discard the flow-through.

7.1.5. Add 800 µl Wash Buffer to all the wells on the plate.

7.1.6. Centrifuge the plate at 5000 g for 5 minutes and discard the flow-through.

7.1.7. To completely dry the plate, centrifuge at 5000 g for 7 minutes and discard any flow-through.

7.1.8. Transfer the plate onto an Elution Plate (in this case we use a PCR plate). Add 34 µl water directly to the column matrix (*make sure that the entire volume is on the matrix*) and let the plate stand for 1 minute. Centrifuge at 5 000 g for 7 minutes to elute the cDNA.

7.1.9. Verify that volumes are even across wells, or transfer 30 µl to a new PCR plate.

7.1.10. Use BASE to register cDNA plate: *BASE → Extensions → Reggie → Library preparation wizards → cDNA registration.*

### 7.2. Quality Control (optional)

7.2.1. For selected standard samples (e.g. Stratagene Reference RNA) measure concentration using Qubit (will only consume 1 µl of sample, replace used volume with water). Concentrations can be tracked over time across multiple plates for quality assurance.

7.2.2. For selected standard samples (e.g. Stratagene Reference RNA) run BioAnalyzer (will only consume 1 µl of sample). Results can be tracked over time across multiple plates for quality assurance.

**Safe stopping point – may store at -80°C**

# 8. End-Repair/A-Tailing and Adaptor Ligation

As a combination module, end-repair, A-tailing, and adapter ligation are performed in an additive reaction. Illumina TruSeq barcoded adapters kits A and B (FC-121-2001 and FC-121-2002) are used, with the adapters first diluted 1:5. Use BASE to assign barcoded adapters samples: *BASE → Extensions → Reggie → Library preparation wizards → Assign barcodes to cDNA plate,* select cDNA bioplate and appropriate preconfigured barcode layout. Use BASE to create worksheets for barcoded adapters: *BASE → Extensions → Reggie → Library preparation wizards → Lab protocols and files for library preparation,* select cDNA bioplate and Print out List layout and Plate layout.

## 8.1. End-Repair/A-Tailing

8.1.1. Make an end-repair/A-tailing master mix, per sample (use SCAN-B calculator):

| | |
|---|---|
| 10X T4 Ligase Buffer | 4 µl |
| 10 mM dNTP mix | 2 µl |
| ATP (10 mM) | 1 µl |
| T4 DNA pol (5U/ul) | 1 µl |
| T4 PNK (10U/ul) | 1 µl |
| Taq DNA pol | 1 µl |

8.1.2. Take 10 µl of mastermix to 30 µl cDNA, mix well (final volume 40 µl).

8.1.3. Place in thermocycler, run the following program:

- 25°C 20 min
- 72°C 20 min
- 12°C Hold

## 8.2. Adaptor Ligation

8.2.1. The adaptors are first diluted 1:5 with water and we have 24 different adaptors available from Illumina TruSeq DNA LT Sample Prep Kit A and B.

8.2.2. Make an adapter ligation master mix, per sample (use SCAN-B calculator):

| | |
|---|---|
| T4 DNA Ligase (5 U/µl) | 3 µl |
| 10X T4 DNA Ligase buffer | 1 µl |
| $H_2O$ | 5 µl |

8.2.3. Add 1 µl of diluted adaptors to each 40 µl sample

8.2.4. Add 9 µl of the adaptor ligation mix, mix well, final volume of 50 µl.

8.2.5. Place in thermocycler, run the following program:

- 22°C 30 min
- 4°C Hold

8.2.6. Proceed to size selection.

**Safe stopping point – may store at -80°C**

## 9. Size Selection with CA Beads

Size selection is performed using polyethylene glycol 8000 (PEG) at appropriate (titrated) final concentration and carboxylic acid (CA)-beads to remove fragments <200 bp. We prepare a PEG stock solution (40%) that is sterile-filtered (0.22 µm filter) and aliquoted in 15 ml tubes: we aliquot 13 ml in each tube to cover use for both size-selection and 2-step PCR purification. Aliquot tubes are labeled with preparation date, PEG lot# and stored at 4°C protected from light. Before first use, each new PEG stock solution must be titrated to evaluate size-selection properties to determine appropriate final working concentration, as this can vary between stock solutions.

### 9.1. Prepare Samples

9.1.1. Add appropriate volume of $H_2O$ to the wells (1-2 µl) to bring up to 50 µl.

9.1.2. Store the samples on ice until the KF96-Binding plate has been prepared.

### 9.2. Prepare PEG mastermix and 80% EtOH

9.2.1. Vortex a new PEG stock-solution tube (15 ml tube) vigorously. Let sit for several minutes (in dark). The PEG stock-solution is going to be used for both size selection and for 2-step PCR purification step so calculate required volume and make sure that enough volume in available.

9.2.2. Make fresh 40 ml 80% EtOH, by taking 32 ml absolute EtOH + 8 ml $H_2O$.

### 9.3. Prepare PEG mastermix

*9.3.1.* Prepare the mastermix, **MM**, using JVC-1-step-CA-MM-calculator-v2.0.xlsx to calculate the master mix.

*Note: the mastermix is prepared using PEG stock-solution, 5M NaCl stock-solution, and water to permit preparation of a binding reaction (final volume 100 µl) by combining sample (50 µl) and mastermix (50 µl) to achieve an appropriate final PEG concentration at 0.9M NaCl. The final PEG concentration in the binding reaction will vary with every new batch of PEG stock-solution and is typically in the range of 7-10% PEG.*

### 9.4. Preparing KF-Flex for size selection run

9.4.1. Allow sufficient time to allow the beads to equilibrate to RT. Shake the magnetic beads bottle for at least 30 min at RT (Heidolph Vibramax 100 Shaker, 300 rpm).

9.4.2. Label and fill the KF 96 plates as follows:

| | | |
|---|---|---|
| 1 | - Tip Plate: | Put 1 new King Fisher 96 tip comb on to the TIP Plate |
| 2 | - Elution Plate: | 15 µl EB |
| 3 | - Wash plate: | 180 µl fresh 80% EtOH |
| 4 | - Binding Plate: | 50 µl **MM** + 50 µl sample, mix by pipetting (thorough mixing is crucial) |
| 5 | - CA bead plate | 40 µl CA beads + 160 µl EB |

### 9.5. Run KF-Flex

9.5.1. Make sure that the A1 positions on the plates are in the same corner as the A1 positions of the KF-Flex disc.

9.5.2. Start program CA_HP6fl (it will take around 23 minutes), follow on-screen instructions.

9.5.3. After completed run, remove elution plate immediately and place on 96-position magnetic stand and wait until the solution is completely clear (1-2 min).

9.5.4. Transfer 13 µl of sample from the elution plate into a UDG plate (prepared in advance) containing UDG master mix, mix well, and store it on ice.

9.5.5. The Binding Plate can be sealed and stored at 4°C and used for further analyses.

**Prepare Second Strand Digestion with UDG while the KF-Flex is running.**

## 10. Second Strand Digestion with UDG

The second strand with incorporated dUTP is specifically digested using uracil-DNA glycosylase (New England Bio Labs, cat# M02805).

### 10.1. Digestion with Uracil-DNA Glycosylase

10.1.1. Make an UDG master mix, per sample (use SCAN-B calculator):

| | |
|---|---|
| 10X UDG Buffer | 1.5 µl |
| UDG (5 U/µl) | 0.1 µl |
| $H_2O$ | 0.4 µl |

10.1.2. Take 2 µl of the master mix to a new PCR plate and transfer 13 µl of DNA from CA purification (final volume 15 µl) and store on ice. Mix well.

10.1.3. Place in thermocycler, run the following PCR program:

- 37°C 15 min
- 94°C 10 min
- 4°C Hold

**Safe stopping point – may store at -80°C**

## 11. PCR Enrichment

Single-stranded cDNA is amplified by PCR.

### 11.1. PCR enrichment mastermix

11.1.1. Make a PCR enrichment mastermix, per sample (use SCAN-B calculator):

| | |
|---|---|
| Illumina Primer Cocktail (1:2 dilution) | 2.625 µl |
| 10 mM dNTP mix | 0.9 µl |
| Phusion Mix (NEB) | 22.5 µl |
| $H_2O$ | 4.875 µl |

11.1.2. Add 30.9 µl of PCR mastermix to 15 µl UDG digested cDNA, final volume of 45.9 µl.

11.1.3. Place in thermocycler, run the following PCR program:

- 98°C 3 min
- 12 cycles of: 98°C 30 sec, 60°C 30 sec, 72°C 30 sec
- 72°C 10 min
- 4°C Hold

11.1.4. After PCR place sample tube/plate on ice.

11.1.5. Proceed to two-step purification with CA size selection.

### 11.2. Quality Control (optional)

11.2.1. For selected standard samples (e.g. Stratagene Reference RNA) measure concentration using Qubit (will only consume 1 µl of sample). Concentrations can be tracked over time across multiple plates for quality assurance.

**Safe stopping point – may store at -20°C**

## 12. Two-Step PCR Purification by CA-Bead Size Selection

The PCR product undergoes two cycles of size selection using CA-beads and varying concentrations of PEG, first to exclude DNA fragments >700 bp and then to exclude fragments <200 bp. Due to variations in ambient humidity and temperature and between batches of PEG preparations, this two-step PCR purification should always be done first with QC libraries (e.g. Stratagene Reference RNA, cat# 740000, Agilent) with different PEG concentrations to determine the optimal solution to proceed with for the entire plate of sample libraries on the same date.

### 12.1. Sample preparation

12.1.1. Make sure that the sample volume (PCR product) is at least 43 µl (expected volume ~45 µl). Top off wells with water if necessary.

### 12.2. PEG Mastermix-1 (MM-1) and Mastermix-2 (MM-2)

12.2.1. Note: The volumes will vary with every new batch of PEG solution. Prepare the mastermixes, **MM-1** and **MM-2** using JVC-2-step-CA-MM-calculator-v5.0.xlsx. *Note: MM1 is prepared to permit set-up of the 1st binding reaction (final volume 80 µl) by combining sample (43 µl) and MM1 (37 µl) to achieve an appropriate final PEG concentration at 0.9M NaCl. Conversely, MM2 is prepared to permit set-up of the 2nd binding reaction (final volume 160 µl) by combining sample (70 µl from 1st binding reaction) and MM2 (90 µl) to bring final PEG concentration to an appropriate level at 0.9M NaCl. Note: final PEG concentrations in the binding reactions will vary with every new batch of PEG stock-solution and is typically in the* *range of 7-9% for the 1st binding reaction and in the range of 8-12% in the 2nd binding reaction.*

### 12.3. Preparation of KF-Flex 2-step run, step 1

12.3.1. Place the beads at RT for at least 2 hours. Shake the bottle with the magnetic beads for at least 30 min at RT (Heidolph Vibramax 100 Shaker, 300 rpm)

12.3.2. Label and fill the KF 96 plates as follows:

| | | |
|---|---|---|
| 1 | - Tip Plate: | Put 1 new King Fisher 96 tip comb on to the TIP Plate |
| 2 | - Binding Plate: | 37 µl **MM-1** + 43 µl sample, mix by pipetting (thorough mixing is crucial) |
| 3 | - CA Bead Plate: | 40 µl CA beads + 160 µl EB, mix by pipetting |

### 12.4. Run KF-Flex 2-step, first step

12.4.1. Make sure that the A1 positions on the plates are in the same corner as the A1 positions of the KF-Flex disc.

12.4.2. Start program CA2ST_C_fl, follow on-screen instructions.

12.4.3. After completed run, remove plates immediately and prepare for KF-Flex step 2 run.

12.4.4. Store the **Binding plate** that will be used in the second step run.

### 12.5. Preparation of KF-Flex 2-step run, step 2

12.5.1. Label and fill the KF 96 plates as follows:

| | | |
|---|---|---|
| 1 | - Tip Plate: | Put 1 new King Fisher 96 tip comb on to the TIP Plate |
| 2 | - Elution Plate: | 15 µl EB |
| 3 | - Wash Plate: | 180 µl fresh 80% EtOH |
| 4 | - Binding Plate: | Binding plate from step 1 run + 90 µl **MM-2**, mix by pipetting (thorough mixing is crucial) |
| 5 | - CA Bead Plate: | 60 µl CA beads + 140 µl EB, mix by pipetting |

### 12.6. Run KF-Flex 2-step, second step

12.6.1. Make sure that the A1 positions on the plates are in the same corner as the A1 positions of the KF-Flex disc.

12.6.2. Start program CA2ST_B_fl, follow on-screen instructions.

12.6.3. After completed run, remove elution plate immediately and place on 96-position magnetic stand and wait until the solution is completely clear (1-2 min).

12.6.4. Transfer 13-14 µl (all) of the purified library samples in a new low-binding deep-well Eppendorf PCR plate (cat# 0030503104), seal, and store on ice.

12.6.5. The Binding Plate can be sealed and stored at 4°C and used for further analyzes.

12.6.6. Run the Bioanalyzer DNA High Sensitivity Chip (cat# 5067-4626) on tester Stratagene Reference RNA libraries to verify which PEG concentration should be used for the entire plate.

## 12.7. Quality Control (optional)

12.7.1. Qubit: 1 µl to Qubit 2.0 Fluorometer (cat# Q32866, Life Technologies). Measure with Qubit prior to aliquoting to Caliper. If the samples have a concentration <500 ng/ml, SpeedVac (Savant) them down to 8 µl, measure once again on Qubit, then aliquot to Caliper.

12.7.2. Caliper: Prefill a blue frame 4titude PCR plate (cat# 4TI-0960/B) with 20 µl nuclease free water; add 2 µl of the samples. Run Caliper according to protocol: Caliper HT DNA High Sensitivity Labchip GX Assay (cat# CLS760672, PerkinElmer) user instruction.

**Safe stopping point – may store at -20°C**

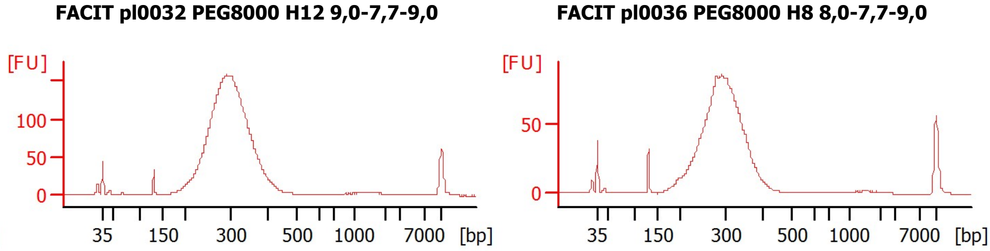


**Figure 3**. Example of typical QC results for Stratagene Reference RNA libraries from well H12 (left) and H8 (right); size-selection at 9% and 8% (left and right, respectively) and 2-step purification at 7.7% followed by 9%.

## 13. Clustering and Sequencing

We use BASE to track libraries and pooled libraries and to create protocols for pooling. Concentration and fragment size for libraries are registered in BASE and used to dynamically create pooling recipes. Libraries are diluted to 2nM and pooled according to pre-configured layouts; typically 21 libraries are pooled together.

- Use BASE to register quality control for Stratagene Reference RNA test samples: *BASE → Extensions → Reggie → Library preparation wizards → Register quality control results*.
- Use BASE to register results from Qubit and Caliper results for the whole plate: *BASE → Extensions → Reggie → Library preparation wizards → Library registration*.
- Use BASE to create pooling schema for libraries according to pre-configured layouts: *BASE → Extensions → Reggie → Pooling wizards → Create pooled libraries*.
- Use BASE to generate and download pooling protocols: *BASE → Extensions → Reggie → Pooling wizards → Lab protocols for pooling*.
- Register library pools: *BASE → Extensions → Reggie → Pooling wizards → Register pooled libraries*.

We use *Clustering* and *Sequencing* wizards in BASE to track pools when clustering flow-cells and to track flow-cells through sequencing. Typically each pool is clustered on a total of 4 lanes across 2 separate flow-cells (2 lanes per flow-cell). Pools are diluted to 12 pM according to Illumina's standard procedure and spiked with 0.5% PhiX control. We typically achieve a cluster density between 800-900 K/mm$^2$ for a total number of PF clusters per lane between 180-200 M (HiSeq 2000).

## 14. Equipment, Consumables, and Reagents

**Equipment:**

- KingFisher Flex Magnetic Particle Processor (KF-Flex; ThermoScientific)
- PCR themocycler (Eppendorf vapo.protect)
- 96-positition magnetic stand, Dynal MPC-96S

- Sigma centrifuge 4K1S
- BioAnalyzer
- NanoDrop
- Qubit
- Caliper LabChip GX

## Consumables:

| | | |
|---|---|---|
| 96-well 4titude PCR plate | Saveen Werner | 4ti-0740 |
| KingFisher 96 plate 200 µl (Elution-, Wash-, and Binding Plate) ??? | VWR | FINN97002540 |
| KingFisher 96 tip comb | VWR | 733-3015 |
| Low-binding deep-well Eppendorf PCR plate | VWR | 0030503.104 |

## Reagents:

| | | |
|---|---|---|
| **mRNA purification** | | |
| Dynabeads mRNA Purification Kit | Life Technologies | cat# 61006 |
| Column-based Oligo Clean & Concentrator | Zymo Research | cat# D4061 |
| RNA Fragmentation Reagents | Ambion | cat# AM8740 |
| Oligo Clean & Concentrator (single columns) | Zymo Research | cat# D4061 |
| ZR-96 Oligo Clean & Concentrator | Zymo Research | cat# D4063 |
| **First strand cDNA Synthesis and Clean-up** | | |
| Random hexamer | Invitrogen | Custom Oligo NNNNNN, 10U |
| SuperScript II Reverse Transcriptase | Life Technologies | cat# 18064-014 |
| Superscript II RT 1000U | Invitrogen | cat# 18064014 |
| 5X First strand buffer and 0.1 M DDT provided with SuperScript II Reverse Transcriptase | | |
| RNase OUT 40U/ul | Invitrogen | cat# P/N 100000840 |
| 10 mM dNTP | Thermo Scientific | cat# R0192 |
| Illustra AutoScreen-96A plate | GE Healthcare | cat# 25-9005-98 |
| **Second Strand cDNA Synthesis** | | |
| 5X First strand buffer and 0.1 M DDT provided with SuperScript II Reverse Transcriptase | | |
| 5X Second-Strand Buffer | Life Technologies (Invitrogen) | cat# 10812-014 |
| dATP 10mM | Thermo Scientific | cat# R0141 |
| dUTP 10mM | Thermo Scientific | cat# R0133 |
| dGTP 10mM | Thermo Scientific | cat# R0161 |
| dCTP 10mM | Thermo Scientific | cat# R0151 |
| DNA polymerase I 10U/ul | New England Bio Labs | cat# M0209L |
| RNaseH 10U/ul | Ambion | cat# AM2293 |
| | | |
| **End-Repair/A-Tailing and Adaptor Ligation** | | |
| T4 DNA Ligase 5 Weiss U/µl +10x Buffer T4 DNA ligase | Thermo Scientific | cat# EL0011 |
| T4 DNA polymerase 5U/µl | Thermo Scientific | cat# EP0062 |
| T4 PNK 10U/ul | New England Bio Labs | cat# M0201L |
| Taq DNA polymerase 5U/ul | Thermo Scientific | cat# EP0402 |

| | | |
|---|---|---|
| 10 mM dNTP | Thermo Scientific | cat# R0192 |
| ATP 100mM | Thermo Scientific | cat# R0441 |
| Adaptors. Included in Illumina TruSeq DNA LT Sample Prep Kit A and B | Illumina | FC-121-2001 and FC-121-2002 |
| **Carboxylic acid (CA)-bead purification** | | |
| Polyethylene glycol 8000 (PEG) | Sigma | cat# P1458-50ml |
| Dynabeads MyOne Carboxylic acid | Life Technologies (Invitrogen) | cat# 65012 |
| Sodium chloride solution | Sigma | cat# 71386-1L |
| **Second strand Digestion with UDG** | | |
| 10x UDG Reaction Buffer | New England Bio Labs | cat# B0280S |
| UDG 5U/ul | New England Bio Labs | cat# M02805 |
| **PCR Enrichment** | | |
| PCR Primer Coctail. Included in Illumina TruSeq DNA LT Sample Prep Kit A and B | Illumina | cat# FC-121-2001 and FC-121-2002 |
| 2x Phusion Master Mix w. HF buffer | Thermo Scientific | cat# F-531L |
| **Other** | | |
| Stratagene Universal Human Reference RNA | Agilent | cat# 740000 |
| EB Buffer | Qiagen | Cat# 19086 |

# The Sweden Canceromics Analysis Network – Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine

## ADDITIONAL FILE 2

Lao H. Saal, Johan Vallon-Christersson, Jari Häkkinen, Cecilia Hegardt, Dorthe Grabau, Christof Winter, Christian Brueffer, Man-Hung Eric Tang, Christel Reuterswärd, Ralph Schulz, Anna Karlsson, Anna Ehinger, Janne Malina, Jonas Manjer, Martin Malmberg, Christer Larsson, Lisa Rydén, Niklas Loman, and Åke Borg

This appendix contains Figure S2, Figure S3, Table S1, and Table S2. The RNA-seq and microarray gene expression data with clinical and mutational annotations are available from the NCBI Gene Expression Omnibus under accession GSE60789.

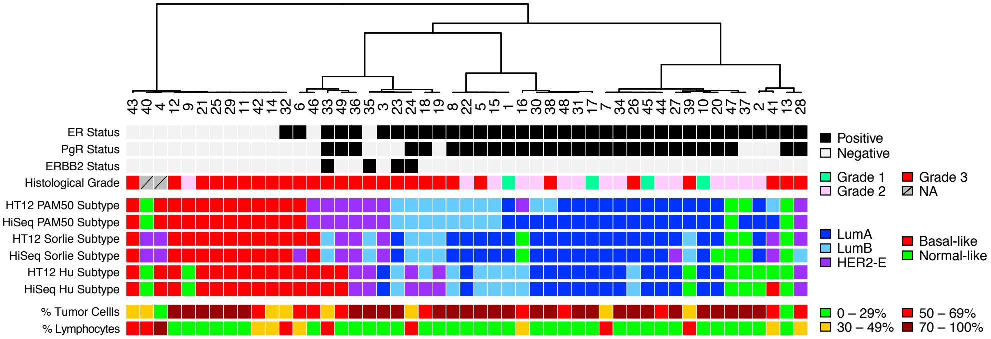


Figure S2 – Hierarchical clustering of 49 primary breast tumors using the RNA-seq gene expression measurements and the PAM50 intrinsic gene signature as in Figure 3.
Here, each tumor's molecular subtype is shown for three different signatures (PAM50, Sørlie, and Hu) using data from either RNA-seq (HiSeq) or microarray (HT12) platforms.  See Materials and Methods section Molecular subtyping.

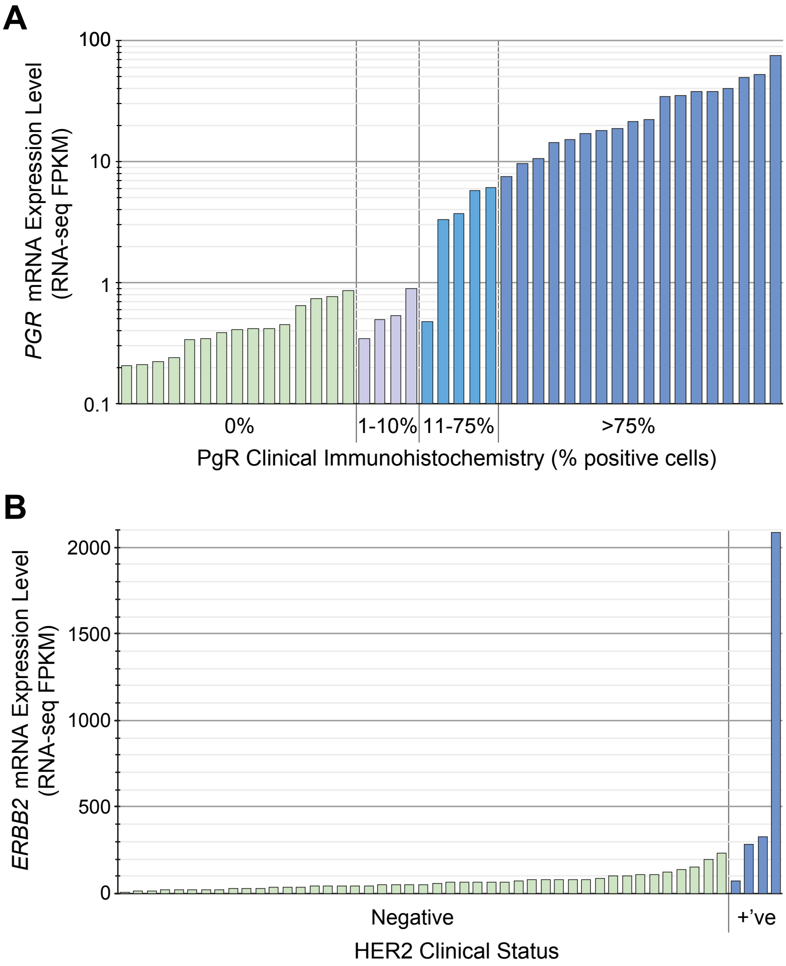


Figure S3 – RNA-seq-derived expression level of (A) *PGR*, which encodes the progesterone receptor (PgR), is shown compared to the clinical PgR IHC score for each tumor. Cases with missing percentage positive cells are not shown. In (B) the expression level of *ERBB2*, encoding the human epidermal growth factor receptor 2 (HER2), is shown compared to the clinical HER2 status.

## Table S1. RNA-Seq Data

| Tumor Number | Raw Sequencing Reads (million) | PF Rate | PF Reads (million) | PCF Rate | PCF Reads (million) | TopHat Alignment Rate | TopHat Aligned Reads (million) | Fraction PF Reads Mappable | Total Reads Aligned (million) | Fraction Unique Read-Pairs (non-duplicates) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 74.9 | 82.5% | 61.8 | 83.3% | 51.5 | 68.9% | 35.5 | 74.1% | 45.8 | 70.2% |
| 2 | 58.9 | 82.3% | 48.5 | 81.9% | 39.7 | 65.3% | 26.0 | 71.6% | 34.7 | 66.1% |
| 3 | 54.3 | 89.3% | 48.5 | 84.0% | 40.8 | 72.3% | 29.5 | 76.8% | 37.2 | 69.8% |
| 4 | 80.5 | 79.4% | 63.9 | 87.2% | 55.8 | 77.2% | 43.0 | 80.1% | 51.2 | 51.2% |
| 5 | 53.9 | 81.5% | 43.9 | 78.2% | 34.3 | 70.6% | 24.2 | 77.0% | 33.8 | 74.6% |
| 6 | 68.1 | 84.2% | 57.3 | 76.8% | 44.1 | 68.9% | 30.4 | 76.1% | 43.6 | 70.2% |
| 7 | 21.0 | 89.6% | 18.8 | 84.0% | 15.8 | 61.9% | 9.8 | 68.0% | 12.8 | 24.9% |
| 8 | 38.2 | 89.8% | 34.3 | 80.3% | 27.6 | 63.9% | 17.6 | 71.0% | 24.4 | 66.7% |
| 9 | 48.4 | 89.5% | 43.3 | 84.4% | 36.6 | 67.4% | 24.7 | 72.4% | 31.4 | 58.3% |
| 10 | 47.5 | 88.6% | 42.1 | 84.1% | 35.4 | 64.9% | 22.9 | 70.5% | 29.6 | 56.1% |
| 11 | 53.2 | 84.9% | 45.2 | 86.4% | 39.0 | 76.0% | 29.7 | 79.3% | 35.8 | 54.9% |
| 12 | 55.6 | 82.1% | 45.7 | 87.8% | 40.1 | 73.9% | 29.7 | 77.1% | 35.2 | 57.7% |
| 13 | 68.2 | 82.1% | 55.9 | 88.6% | 49.6 | 65.3% | 32.4 | 69.2% | 38.7 | 66.2% |
| 14 | 59.4 | 84.1% | 49.9 | 84.2% | 42.1 | 68.0% | 28.6 | 73.0% | 36.5 | 72.4% |
| 15 | 47.5 | 84.3% | 40.0 | 82.3% | 33.0 | 61.8% | 20.4 | 68.5% | 27.4 | 40.4% |
| 16 | 58.6 | 81.1% | 47.5 | 85.6% | 40.6 | 67.8% | 27.5 | 72.4% | 34.4 | 74.1% |
| 17 | 71.3 | 83.4% | 59.5 | 79.4% | 47.2 | 71.0% | 33.5 | 76.9% | 45.8 | 68.3% |
| 18 | 53.0 | 83.1% | 44.0 | 82.2% | 36.2 | 63.1% | 22.8 | 69.6% | 30.7 | 55.3% |
| 19 | 61.2 | 83.6% | 51.2 | 87.1% | 44.6 | 76.9% | 34.2 | 79.8% | 40.9 | 51.1% |
| 20 | 66.6 | 84.1% | 56.0 | 86.6% | 48.5 | 76.7% | 37.2 | 79.8% | 44.7 | 62.9% |
| 21 | 56.6 | 83.4% | 47.2 | 82.1% | 38.7 | 78.1% | 30.3 | 82.0% | 38.7 | 53.9% |
| 22 | 45.5 | 83.4% | 38.0 | 87.3% | 33.1 | 75.2% | 24.9 | 78.3% | 29.7 | 52.4% |
| 23 | 48.8 | 83.5% | 40.7 | 81.8% | 33.3 | 61.6% | 20.5 | 68.6% | 27.9 | 47.9% |
| 24 | 47.0 | 79.3% | 37.2 | 79.5% | 29.6 | 56.1% | 16.6 | 65.1% | 24.3 | 80.8% |
| 25 | 59.8 | 89.7% | 53.7 | 79.6% | 42.7 | 65.4% | 27.9 | 72.5% | 38.9 | 55.4% |
| 26 | 67.3 | 80.6% | 54.3 | 84.7% | 46.0 | 72.1% | 33.1 | 76.4% | 41.5 | 77.2% |
| 27 | 69.2 | 83.8% | 58.0 | 78.4% | 45.5 | 75.5% | 34.3 | 80.8% | 46.8 | 65.8% |
| 28 | 55.2 | 83.2% | 45.9 | 82.2% | 37.7 | 71.1% | 26.8 | 76.3% | 35.0 | 75.8% |
| 29 | 60.5 | 82.6% | 50.0 | 84.9% | 42.4 | 64.1% | 27.2 | 69.5% | 34.7 | 43.6% |
| 30 | 62.6 | 81.4% | 50.9 | 86.5% | 44.1 | 67.4% | 29.7 | 71.8% | 36.6 | 65.1% |
| 31 | 49.5 | 84.4% | 41.8 | 81.1% | 33.9 | 64.6% | 21.9 | 71.3% | 29.8 | 44.2% |
| 32 | 63.4 | 85.2% | 54.1 | 76.1% | 41.2 | 77.9% | 32.1 | 83.2% | 45.0 | 66.7% |
| 33 | 55.9 | 84.6% | 47.2 | 79.6% | 37.6 | 76.6% | 28.8 | 81.4% | 38.5 | 63.3% |
| 34 | 69.0 | 84.5% | 58.3 | 85.7% | 50.0 | 77.1% | 38.5 | 80.4% | 46.9 | 59.4% |
| 35 | 75.8 | 82.0% | 62.2 | 85.8% | 53.4 | 65.4% | 34.9 | 70.4% | 43.8 | 65.8% |
| 36 | 56.9 | 83.7% | 47.6 | 73.0% | 34.8 | 65.8% | 22.9 | 75.1% | 35.7 | 67.8% |
| 37 | 55.8 | 81.9% | 45.7 | 83.2% | 38.0 | 67.5% | 25.7 | 73.0% | 33.3 | 70.9% |
| 38 | 52.4 | 82.9% | 43.4 | 84.9% | 36.8 | 67.3% | 24.8 | 72.3% | 31.4 | 73.6% |
| 39 | 67.4 | 84.5% | 57.0 | 83.2% | 47.4 | 65.3% | 31.0 | 71.1% | 40.5 | 61.8% |
| 40 | 69.2 | 84.7% | 58.6 | 86.4% | 50.7 | 81.0% | 41.0 | 83.6% | 49.0 | 63.5% |
| 41 | 42.1 | 86.0% | 36.3 | 80.1% | 29.1 | 76.3% | 22.2 | 81.0% | 29.4 | 44.6% |
| 42 | 51.2 | 84.2% | 43.1 | 79.9% | 34.5 | 77.5% | 26.7 | 82.0% | 35.4 | 66.0% |
| 43 | 65.7 | 80.3% | 52.8 | 78.1% | 41.2 | 73.6% | 30.3 | 79.4% | 41.9 | 59.3% |
| 44 | 70.8 | 85.0% | 60.1 | 84.7% | 50.9 | 79.4% | 40.4 | 82.6% | 49.7 | 60.7% |
| 45 | 58.2 | 83.3% | 48.5 | 87.0% | 42.2 | 67.7% | 28.6 | 71.9% | 34.9 | 72.8% |
| 46 | 57.0 | 82.6% | 47.1 | 80.9% | 38.1 | 69.8% | 26.6 | 75.6% | 35.6 | 79.3% |
| 47 | 51.6 | 82.9% | 42.8 | 82.7% | 35.4 | 61.6% | 21.8 | 68.2% | 29.2 | 55.5% |
| 48 | 50.3 | 80.8% | 40.6 | 83.8% | 34.0 | 65.9% | 22.4 | 71.4% | 29.0 | 65.3% |
| 49 | 56.9 | 83.5% | 47.5 | 78.2% | 37.2 | 68.0% | 25.3 | 75.0% | 35.6 | 53.0% |
| 3-replicate | 47.6 | 89.8% | 42.7 | 82.6% | 35.3 | 72.9% | 25.7 | 77.6% | 33.1 | 55.5% |
| 10-replicate | 50.1 | 89.3% | 44.7 | 87.0% | 38.9 | 63.0% | 24.5 | 67.8% | 30.3 | 47.7% |
| 18-replicate | 59.1 | 83.5% | 49.3 | 81.6% | 40.2 | 67.2% | 27.0 | 73.2% | 36.1 | 68.0% |
| 22-replicate | 70.8 | 82.7% | 58.6 | 87.9% | 51.5 | 77.1% | 39.8 | 79.9% | 46.8 | 59.2% |
| 38-replicate | 60.4 | 84.6% | 51.1 | 84.4% | 43.1 | 66.7% | 28.8 | 71.9% | 36.7 | 76.0% |
| 45-replicate | 64.0 | 84.2% | 53.9 | 87.9% | 47.4 | 62.9% | 29.8 | 67.4% | 36.3 | 50.1% |
| **Minimum** | 21.0 | 79.3% | 18.8 | 73.0% | 15.8 | 56.1% | 9.8 | 65.1% | 12.8 | 24.9% |
| **Maximum** | 80.5 | 89.8% | 63.9 | 88.6% | 55.8 | 81.0% | 43.0 | 83.6% | 51.2 | 80.8% |
| **Mean** | 57.9 | 84.0% | 48.5 | 83.0% | 40.3 | 69.6% | 28.2 | 74.7% | 36.4 | 61.4% |
| **STDEV** | 10.3 | 2.6% | 8.2 | 3.4% | 7.3 | 5.7% | 6.3 | 4.8% | 7.2 | 10.9% |
| **Median** | 57.0 | 83.5% | 47.6 | 83.3% | 40.1 | 68.0% | 27.9 | 74.1% | 35.7 | 63.3% |
| **75th Percentile** | 66.1 | 84.6% | 54.2 | 85.7% | 45.0 | 75.3% | 31.5 | 79.3% | 41.2 | 69.1% |
| **25th Percentile** | 51.4 | 82.5% | 43.4 | 80.6% | 35.4 | 65.3% | 24.6 | 71.2% | 31.4 | 55.1% |

## Table S2. 90 Genes Screened for Mutations

| Gene symbol | Gene name | Location | HGNC ID |
|---|---|---|---|
| AFF2 | AF4/FMR2 family, member 2 | Xq28 | HGNC:3776 |
| AKAP3 | A kinase (PRKA) anchor protein 3 | 12p13.3 | HGNC:373 |
| AKT1 | v-akt murine thymoma viral oncogene homolog 1 | 14q32.32-q32.33 | HGNC:391 |
| AKT2 | v-akt murine thymoma viral oncogene homolog 2 | 19q13.1-q13.2 | HGNC:392 |
| APC | adenomatous polyposis coli | 5q21-q22 | HGNC:583 |
| ARID1A | AT rich interactive domain 1A (SWI-like) | 1p36.1-p35 | HGNC:11110 |
| ARID1B | AT rich interactive domain 1B (SWI1-like) | 6q25.3 | HGNC:18040 |
| ARID2 | AT rich interactive domain 2 (ARID, RFX-like) | 12q13.11 | HGNC:18037 |
| ASXL1 | additional sex combs like 1 (Drosophila) | 20q11 | HGNC:18318 |
| ATM | ataxia telangiectasia mutated | 11q22-q23 | HGNC:795 |
| ATN1 | atrophin 1 | 12p | HGNC:3033 |
| ATP2B2 | ATPase, Ca++ transporting, plasma membrane 2 | 3p25.3 | HGNC:815 |
| BAP1 | BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase) | 3p21.31-p21.2 | HGNC:950 |
| BARD1 | BRCA1 associated RING domain 1 | 2q34-q35 | HGNC:952 |
| BRCA1 | breast cancer 1, early onset | 17q21.31 | HGNC:1100 |
| BRCA2 | breast cancer 2, early onset | 13q12-q13 | HGNC:1101 |
| BRIP1 | BRCA1 interacting protein C-terminal helicase 1 | 17q22.2 | HGNC:20473 |
| CASP8 | caspase 8, apoptosis-related cysteine peptidase | 2q33-q34 | HGNC:1509 |
| CBFB | core-binding factor, beta subunit | 16q22.1 | HGNC:1539 |
| CCND1 | cyclin D1 | 11q13 | HGNC:1582 |
| CCND3 | cyclin D3 | 6p21 | HGNC:1585 |
| CDH1 | cadherin 1, type 1, E-cadherin (epithelial) | 16q22.1 | HGNC:1748 |
| CDKN1B | cyclin-dependent kinase inhibitor 1B (p27, Kip1) | 12p13.1-p12 | HGNC:1785 |
| CDKN2A | cyclin-dependent kinase inhibitor 2A | 9p21 | HGNC:1787 |
| CHEK2 | checkpoint kinase 2 | 22q12.1 | HGNC:16627 |
| CLEC19A | C-type lectin domain family 19, member A | 16p12.3 | HGNC:34522 |
| CTCF | CCCTC-binding factor (zinc finger protein) | 16q21-q22.3 | HGNC:13723 |
| DCAF4L2 | DDB1 and CUL4 associated factor 4-like 2 | 8q21.3 | HGNC:26657 |
| DGKG | diacylglycerol kinase, gamma 90kDa | 3q27-q28 | HGNC:2853 |
| EP300 | E1A binding protein p300 | 22q13.2 | HGNC:3373 |
| ERBB2 | v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 | 17q11.2-q12 | HGNC:3430 |
| ETV6 | ets variant 6 | 12p13 | HGNC:3495 |
| FAM157B | family with sequence similarity 157, member B | 9q34 | HGNC:34080 |
| FAM47C | family with sequence similarity 47, member C | Xp21.1 | HGNC:25301 |
| FOXA1 | forkhead box A1 | 14q12-q13 | HGNC:5021 |
| GATA3 | GATA binding protein 3 | 10p15 | HGNC:4172 |
| GPR32 | G protein-coupled receptor 32 | 19q13.33 | HGNC:4487 |
| GPS2 | G protein pathway suppressor 2 | 17p13.1 | HGNC:4550 |
| HIST1H1C | histone cluster 1, H1c | 6p21.3 | HGNC:4716 |
| HIST1H2BC | histone cluster 1, H2bc | 6p22.1 | HGNC:4757 |
| KCNB2 | potassium voltage-gated channel, Shab-related subfamily, member 2 | 8q13.2 | HGNC:6232 |
| KRAS | Kirsten rat sarcoma viral oncogene homolog | 12p12.1 | HGNC:6407 |
| MAP2K4 | mitogen-activated protein kinase kinase 4 | 17p12 | HGNC:6844 |
| MAP3K1 | mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase | 5q11.2 | HGNC:6848 |
| MAP3K13 | mitogen-activated protein kinase kinase kinase 13 | 3q27 | HGNC:6852 |
| MED23 | mediator complex subunit 23 | 6q22.33-q24.1 | HGNC:2372 |
| MICA | MHC class I polypeptide-related sequence A | 6p21.3 | HGNC:7090 |
| KMT2D | lysine (K)-specific methyltransferase 2D | 12q13.12 | HGNC:7133 |
| KMT2C | lysine (K)-specific methyltransferase 2C | 7q36 | HGNC:13726 |
| MRE11A | MRE11 meiotic recombination 11 homolog A (S. cerevisiae) | 11q21 | HGNC:7230 |
| MYB | v-myb avian myeloblastosis viral oncogene homolog | 6q22-q23 | HGNC:7545 |
| NCOR1 | nuclear receptor corepressor 1 | 17p11.2 | HGNC:7672 |
| NF1 | neurofibromin 1 | 17q11.2 | HGNC:7765 |
| NTRK3 | neurotrophic tyrosine kinase, receptor, type 3 | 15q24-q25 | HGNC:8033 |
| OR2G3 | olfactory receptor, family 2, subfamily G, member 3 | 1q44 | HGNC:15008 |
| OR2L2 | olfactory receptor, family 2, subfamily L, member 2 | 1q44 | HGNC:8266 |
| OR6A2 | olfactory receptor, family 6, subfamily A, member 2 | 11p15.4 | HGNC:15301 |
| PALB2 | partner and localizer of BRCA2 | 16p12.1 | HGNC:26144 |
| PBRM1 | polybromo 1 | 3p21 | HGNC:30064 |
| PIK3CA | phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha | 3q26.3 | HGNC:8975 |
| PIK3R1 | phosphoinositide-3-kinase, regulatory subunit 1 (alpha) | 5q13.1 | HGNC:8979 |
| PIWIL1 | piwi-like RNA-mediated gene silencing 1 | 12q24.33 | HGNC:9007 |
| PNPLA3 | patatin-like phospholipase domain containing 3 | 22q13.31 | HGNC:18590 |
| PTEN | phosphatase and tensin homolog | 10q23 | HGNC:9588 |
| PTPN22 | protein tyrosine phosphatase, non-receptor type 22 (lymphoid) | 1p13.2 | HGNC:9652 |
| PTPRD | protein tyrosine phosphatase, receptor type, D | 9p24.1-p23 | HGNC:9668 |
| RAD50 | RAD50 homolog (S. cerevisiae) | 5q23-q31 | HGNC:9816 |
| RAD51C | RAD51 paralog C | 17q25.1 | HGNC:9820 |
| RAD51D | RAD51 paralog D | 17q11 | HGNC:9823 |
| RB1 | retinoblastoma 1 | 13q14.2 | HGNC:9884 |
| RPGR | retinitis pigmentosa GTPase regulator | Xp11.4 | HGNC:10295 |
| RUNX1 | runt-related transcription factor 1 | 21q22.3 | HGNC:10471 |
| RYR2 | ryanodine receptor 2 (cardiac) | 1q43 | HGNC:10484 |
| SEPT7P2 | septin 7 pseudogene 2 | 7p12.3 | HGNC:32339 |
| SETD2 | SET domain containing 2 | 3p21.31 | HGNC:18420 |
| SF3B1 | splicing factor 3b, subunit 1, 155kDa | 2q33.1 | HGNC:10768 |
| SMAD4 | SMAD family member 4 | 18q21.1 | HGNC:6770 |
| SMARCD1 | SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1 | 12q13-q14 | HGNC:11106 |
| SRPR | signal recognition particle receptor (docking protein) | 11q24-q25 | HGNC:11307 |
| STK11 | serine/threonine kinase 11 | 19p13.3 | HGNC:11389 |
| TBL1XR1 | transducin (beta)-like 1 X-linked receptor 1 | 3q26.33 | HGNC:29529 |
| TBX3 | T-box 3 | 12q24.21 | HGNC:11602 |
| TLR4 | toll-like receptor 4 | 9q33.1 | HGNC:11850 |
| TP53 | tumor protein p53 | 17p13.1 | HGNC:11998 |
| TPRX1 | tetra-peptide repeat homeobox 1 | 19q13.33 | HGNC:32174 |
| TRIM53AP | tripartite motif containing 53A, pseudogene | 11q14.3 | HGNC:19025 |
| TRIM6-TRIM34 | TRIM6-TRIM34 readthrough | 11p15.4 | HGNC:33440 |
| USH2A | Usher syndrome 2A (autosomal recessive, mild) | 1q41 | HGNC:12601 |
| WNT7A | wingless-type MMTV integration site family, member 7A | 3p25 | HGNC:12786 |
| ZFP36L1 | ZFP36 ring finger protein-like 1 | 14q22-q24 | HGNC:1107 |

# Study II

Brueffer and Saal *BMC Bioinformatics* (2016) 17:199
DOI 10.1186/s12859-016-1058-x

BMC Bioinformatics

SOFTWARE Open Access

# TopHat-Recondition: a post-processor for TopHat unmapped reads

Christian Brueffer and Lao H. Saal*

## Abstract

**Background:** TopHat is a popular spliced junction mapper for RNA sequencing data, and writes files in the BAM format – the binary version of the Sequence Alignment/Map (SAM) format. BAM is the standard exchange format for aligned sequencing reads, thus correct format implementation is paramount for software interoperability and correct analysis. However, TopHat writes its unmapped reads in a way that is not compatible with other software that implements the SAM/BAM format.

**Results:** We have developed TopHat-Recondition, a post-processor for TopHat unmapped reads that restores read information in the proper format. TopHat-Recondition thus enables downstream software to process the plethora of BAM files written by TopHat.

**Conclusions:** TopHat-Recondition can repair unmapped read files written by TopHat and is freely available under a 2-clause BSD license on GitHub: https://github.com/cbrueffer/tophat-recondition.

**Keywords:** RNA-seq, Deep sequencing, Sequence alignment, Sequence analysis

## Background

RNA sequencing (RNA-seq) has become as a cornerstone of genomics research. TopHat and TopHat2 [1, 2] (jointly referred to as TopHat from here on) is a highly-cited spliced read mapper for RNA-seq data that is used in many large-scale studies around the world, for example in breast cancer [3]. A search for the term "TopHat" in the NCBI Gene Expression Omnibus (GEO) and the European Nucleotide Archive (ENA) yields 288 and 197 datasets using TopHat, respectively, with the true number being likely much higher.

TopHat writes read data in the BAM format – the binary version of the Sequence Alignment/Map (SAM) format [4], but unlike other read mappers, it writes separate files for reads it could map to the reference genome (`accepted_hits.bam`) and reads it could not map (`unmapped.bam`). Although many analyses focus on mapped reads alone, often it is necessary to consider unmapped reads, for example to perform quality assurance, to deposit the data in online archives, or to analyze the unmapped reads themselves.

However, all released versions of TopHat to date (version $\leq$ 2.1.1) generate `unmapped.bam` files that are incompatible with common downstream software, e.g., the Picard suite (http://broadinstitute.github.io/picard), SAMtools [4], or the Genome Analysis Toolkit (GATK) [5]. Even if the problems leading to the incompatibility are corrected in future versions of TopHat, an immense amount of data has already been aligned with affected versions and would need to be realigned, and potentially reanalyzed. TopHat-Recondition is a post-processor for TopHat unmapped reads that corrects the compatibility problems, and restores the ability to process BAM files containing unmapped reads.

## Implementation

TopHat-Recondition is implemented in Python using the Pysam library (https://github.com/pysam-developers/pysam) and requires Python 2.6 or higher. The simplified workflow of the software is shown in Fig. 1. First, the `unmapped.bam` file is loaded into memory, both for performance reasons and to enable random access to the unmapped reads. In the first pass over the unmapped reads the `/1` and `/2` suffixes are removed from read

*Correspondence: lao.saal@med.lu.se
Division of Oncology and Pathology, Department of Clinical Sciences, Lund University Cancer Center, Lund University, Medicon Village Building 404-B2, 223 81 Lund, Sweden

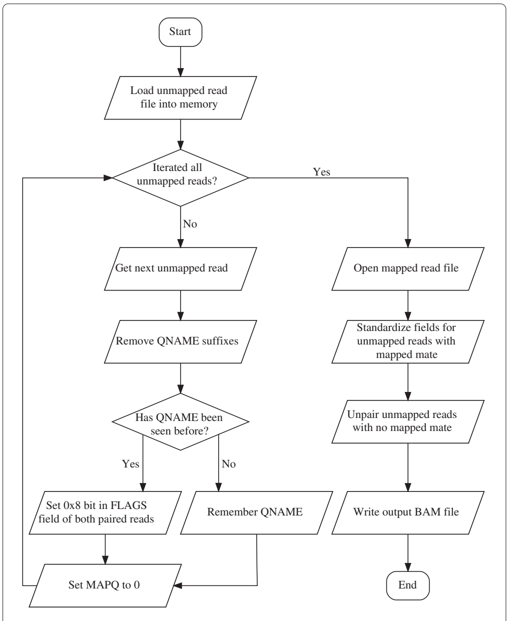

Start
Load unmapped read file into memory
Iterated all unmapped reads?
Yes
No
Get next unmapped read
Remove QNAME suffixes
Has QNAME been seen before?
Yes
No
Set 0x8 bit in FLAGS field of both paired reads
Remember QNAME
Set MAPQ to 0
Open mapped read file
Standardize fields for unmapped reads with mapped mate
Unpair unmapped reads with no mapped mate
Write output BAM file
End

**Fig. 1** Simplified workflow of TopHat-Recondition

names (only TopHat prior to version 2.0.7), MAPQ is set to 0, missing 0x8 flags are added to unmapped read-pairs, and the reads are indexed by their read names (QNAME). In the second pass all unmapped reads with mapped mate are recorded to enable detection of missing mapped mates. The accepted_hits.bam file is read sequentially to obtain information to correct unmapped reads with mapped mate; the previously built index is used to quickly access the unmapped mate of the current mapped read. The mate-related bits (0x1, 0x2, 0x8, 0x20, 0x40, 0x80) in the FLAGS field of unmapped reads for which the mapped paired read could not be found are unset, effectively making them unpaired. Additionally, the RNAME, RNEXT, PNEXT and POS fields are modified as described above. The corrected unmapped reads are written as unmapped_fixup.bam in the specified directory (by default the input BAM file directory), along with a log file detailing the performed modifications. TopHat-Recondition can process a library with 50 million reads in ten minutes on a standard PC, with the disk read performance being the limiting factor.

## Results and discussion

TopHat's unmapped.bam incompatibility with other tools has three origins: software bugs resulting in violations of the SAM/BAM specification (https://samtools.github.io/hts-specs/SAMv1.pdf), divergences from the specification's recommended practices, and different interpretation of acceptable values for some of the file format's fields between software.

Two TopHat issues impair compatibility: First, all unmapped read-pairs lack the 0x8 bit (next segment in the template unmapped) in their FLAGS field. This leads to downstream software incorrectly assuming the reads to be mapped. Second, for unmapped reads where the FLAGS field declares the paired read to be mapped, this mapped paired read may be missing from the sequence files. This makes the unmapped read's fields invalid and can lead to software searching for, and failing to find the paired read.

The SAM/BAM specification contains a section on recommended practices for implementing the format. For read-pairs with one mapped and one unmapped read, TopHat does not follow the recommendations that RNAME and POS of the unmapped read should have the same field values as the mapped read. Additionally we found that setting RNEXT to the mapped read's RNEXT value, and PNEXT to 0 improves compatibility.

Lastly, there are differing interpretations of which field values are acceptable in certain conditions between software packages. For example, the valid range of values for the BAM mapping quality (MAPQ) is 0-255. For unmapped reads, TopHat always sets the MAPQ value of unmapped reads to 255, and BWA [6] sets the value to greater than 0 in certain conditions, while the Picard suite asserts that this value be 0 and returns an error when encountering such a read, which can confuse users.

Some BAM-processing software, e.g., Picard and GATK can be configured to accept reads that do not conform to its expectations by ignoring errors, thus allowing processing to succeed. However, the resulting BAM files remain non-compliant to the specification which can lead to issues in later analysis steps that are difficult to debug.

The occurrence of these problems is dependent on both the sequencing depth and the percentage of unmapped reads in the dataset; a higher value in either category can result in a higher rate of errors.

TopHat-Recondition either repairs or works around these problems, which allows processing to complete with all SAM/BAM-compliant software without relying on reducing strictness requirements.

Usage information and a walk-through example can be found in Additional file 1.

## Conclusions

TopHat-Recondition enables easy and fast post-processing for TopHat unmapped reads. The tool can be used to process TopHat-written unmapped reads to make them compatible with downstream tools such as samtools, the Picard suite and GATK, which is currently not possible with the stock unmapped reads. This will increase the utility of the immense amount of RNA-seq data that has been analyzed by TopHat.

## Availability and requirements

**Project name:** TopHat-Recondition
**Project home page:** https://github.com/cbrueffer/tophat-recondition
**Operating system(s):** Platform independent
**Programming language:** Python
**Other requirements:** Pysam
**License:** 2-clause BSD
**Any restrictions to use by non-academics:** none

## Additional file

**Additional file 1:** Usage information and walk-through example. (PDF 166 kb)

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

CB wrote the TopHat-Recondition software. CB and LHS wrote the manuscript. Both authors read and approved the final manuscript.

### Acknowledgements

The authors would like to thank the TopHat developers for their software and permission to use the name TopHat as part of TopHat-Recondition, and Christof Winter for fruitful discussions about the SAM/BAM file format.

### Funding

This work was supported by the Swedish Cancer Society, Swedish Research Council, Governmental Funding of Clinical Research within National Health Service, and Mrs. Berta Kamprad Foundation.

Received: 13 January 2016 Accepted: 20 April 2016
Published online: 04 May 2016

### References

1. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
2. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):36.
3. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, Brueffer C, Tang M-HE, Reuterswärd C, Schulz R, Karlsson A, Ehinger A, Malina J, Manjer J, Malmberg M, Larsson C, Rydén L, Loman N, Borg Å. The Sweden Cancerome Analysis Network - Breast (SCAN-B) initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med. 2015;7(1): 1–12.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, the 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

# Supplementary Materials for TopHat-Recondition 1.0

Christian Brueffer and Lao H Saal

## Usage

TopHat-Recondition can be obtained from GitHub (https://github.com/cbrueffer/tophat-recondition/). Here we assume it is available as ~/tophat-recondition/tophat-recondition.py.
The only required argument for the software is a directory containing the TopHat output files accepted_hits.bam and unmapped.bam, such as the default TopHat tophat_out output directory. A full list of options can be obtained by running tophat-recondition.py without arguments.

```
$ ~/tophat-recondition/tophat-recondition.py
Usage:

tophat-recondition.py [-hqv] [-l logfile] tophat_output_dir [result_dir]

-h                  print this usage text and exit (optional)
-l                  log file (optional, default: result_dir/tophat-recondition.log)
-q                  quiet mode, no console output (optional)
-v                  print the script name and version, and exit (optional)
tophat_output_dir: directory containing accepted_hits.bam and unmapped.bam
result_dir:         directory to write unmapped_fixup.bam to (optional, default: tophat_output_dir)
```

By default, TopHat-Recondition will write the corrected unmapped read file unmapped_fixup.bam to the directory containing the input BAM files.

## Example Run

To show the usage and operation of TopHat-Recondition, we use the workflow and data outlined in the TopHat tutorial:
Tutorial: http://ccb.jhu.edu/software/tophat/tutorial.shtml
Data: http://ccb.jhu.edu/software/tophat/downloads/test_data.tar.gz

### Running TopHat

We extract the data and run TopHat 2.1.0 as instructed in the tutorial.

```
$ tar zxvf test_data.tar.gz
$ cd test_data
$ tophat -r 20 test_ref reads_1.fq reads_2.fq

[2015-10-30 12:58:40] Beginning TopHat run (v2.1.0)
-----------------------------------------------
[2015-10-30 12:58:40] Checking for Bowtie
      Bowtie version:  2.2.5.0
[2015-10-30 12:58:40] Checking for Bowtie index files (genome)..
  Found both Bowtie1 and Bowtie2 indexes.
[2015-10-30 12:58:40] Checking for reference FASTA file
[2015-10-30 12:58:40] Generating SAM header for test_ref
[2015-10-30 12:58:40] Preparing reads
   left reads: min. length=75, max. length=75, 100 kept reads (0 discarded)
  right reads: min. length=75, max. length=75, 100 kept reads (0 discarded)
[2015-10-30 12:58:40] Mapping left_kept_reads to genome test_ref with Bowtie2
[2015-10-30 12:58:41] Mapping left_kept_reads_seg1 to genome test_ref with Bowtie2 (1/3)
[2015-10-30 12:58:41] Mapping left_kept_reads_seg2 to genome test_ref with Bowtie2 (2/3)
[2015-10-30 12:58:41] Mapping left_kept_reads_seg3 to genome test_ref with Bowtie2 (3/3)
[2015-10-30 12:58:41] Mapping right_kept_reads to genome test_ref with Bowtie2
[2015-10-30 12:58:41] Mapping right_kept_reads_seg1 to genome test_ref with Bowtie2 (1/3)
[2015-10-30 12:58:41] Mapping right_kept_reads_seg2 to genome test_ref with Bowtie2 (2/3)
```

```
[2015-10-30 12:58:41] Mapping right_kept_reads_seg3 to genome test_ref with Bowtie2 (3/3)
[2015-10-30 12:58:41] Searching for junctions via segment mapping
[2015-10-30 12:58:41] Retrieving sequences for splices
[2015-10-30 12:58:42] Indexing splices
Building a SMALL index
[2015-10-30 12:58:42] Mapping left_kept_reads_seg1 to genome segment_juncs with Bowtie2 (1/3)
[2015-10-30 12:58:42] Mapping left_kept_reads_seg2 to genome segment_juncs with Bowtie2 (2/3)
[2015-10-30 12:58:42] Mapping left_kept_reads_seg3 to genome segment_juncs with Bowtie2 (3/3)
[2015-10-30 12:58:42] Joining segment hits
[2015-10-30 12:58:42] Mapping right_kept_reads_seg1 to genome segment_juncs with Bowtie2 (1/3)
[2015-10-30 12:58:43] Mapping right_kept_reads_seg2 to genome segment_juncs with Bowtie2 (2/3)
[2015-10-30 12:58:43] Mapping right_kept_reads_seg3 to genome segment_juncs with Bowtie2 (3/3)
[2015-10-30 12:58:43] Joining segment hits
[2015-10-30 12:58:43] Reporting output tracks
-----------------------------------------------
[2015-10-30 12:58:43] A summary of the alignment counts can be found in ./tophat_out/align_summary.txt
[2015-10-30 12:58:43] Run complete: 00:00:02 elapsed
```

## Running TopHat-Recondition

TopHat writes its output files — accepted_hits.bam and unmapped.bam — to the directory tophat_out. We run TopHat-Recondition with this directory as argument. By not specifying a separate output directory, the corrected unmapped read file — unmapped_fixup.bam — will be written to the input directory tophat_out.

```
$ tophat-recondition.py tophat_out
2015-10-30 12:59:45 - Starting run of tophat-recondition 1.0
2015-10-30 12:59:45 - Command: tophat-recondition.py tophat_out
2015-10-30 12:59:45 - Current working directory: /home/chris/test_data
2015-10-30 12:59:45 - Writing logfile: tophat_out/tophat-recondition.log
2015-10-30 12:59:45 - Opening unmapped BAM file: tophat_out/unmapped.bam
2015-10-30 12:59:45 - Loading unmapped BAM file into memory: tophat_out/unmapped.bam
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_150_290_0
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_96_238_3
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_75_235_21
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_48_207_39
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_94_291_40
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_33_189_4a
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_172_294_4f
2015-10-30 12:59:45 - Setting missing 0x8 flag for unmapped read-pair: test_mRNA_4_191_5d
2015-10-30 12:59:45 - Opening mapped BAM file: tophat_out/accepted_hits.bam
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_5_197_46
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_11_190_1a
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_21_208_24
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_23_186_42
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_28_188_11
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_28_206_1f
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_30_231_3c
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_33_223_4e
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_44_225_1e
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_44_193_3f
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_46_195_17
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_51_194_49
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_57_231_8
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_58_234_7
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_58_220_3d
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_65_238_2e
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_69_229_23
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_81_228_3a
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_82_255_2
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_89_230_b
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_89_245_15
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_92_266_43
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_92_250_44
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_97_275_26
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_114_277_5b
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_16_194_10
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_131_260_33
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_39_219_5c
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_50_224_2d
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_51_248_14
```

```
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_128_252_36
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_52_261_1b
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_110_267_22
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_111_268_d
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_104_274_1c
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_85_275_38
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_75_277_3b
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_125_280_48
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_151_286_e
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_125_293_60
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_111_297_61
2015-10-30 12:59:45 - Standardizing flags of unmapped read: test_mRNA_145_300_37
2015-10-30 12:59:45 - Writing corrected BAM file: tophat_out/unmapped_fixup.bam
2015-10-30 12:59:45 - Program finished successfully.
```

### Verifying the Result

The log details the modifications performed on the reads. To verify them, we can compare the original `unmapped.bam` file and the corrected `unmapped_fixup.bam` file.

In the original `unmapped.bam` file, unmapped read pairs cannot be identified by the bits set in their `FLAGS` fields (both reads having the "mate is unmapped" bit set), even though it clearly contains eight such pairs.

```
$ cd tophat_out
$ samtools view -f 0x8 unmapped.bam
$
$ samtools view unmapped.bam | cut -f 1 | sort | uniq --repeated
test_mRNA_150_290_0
test_mRNA_172_294_4f
test_mRNA_33_189_4a
test_mRNA_4_191_5d
test_mRNA_48_207_39
test_mRNA_75_235_21
test_mRNA_94_291_40
test_mRNA_96_238_3
```

The corrected `unmapped_fixup.bam` file shows the unmapped read pairs correctly.

```
$ samtools view -f 0x8 unmapped_fixup.bam
test_mRNA_150_290_0 77  * 0 0 * * 0 0
    TCCTAAAAAGTCCGCCTCGGTCTCAGTCTCAAGTAGAAAAAGTCCCGTTGGCGATCCGTCTACGTCCGAGTAAGA
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_150_290_0 141 * 0 0 * * 0 0
    TACGTATTTGTCGCGCGGCCCTACGGCTGAGCGTCGAGCTTGCGATCCGCCACTATTACTTTATTATCTTACTCG
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_96_238_3  141 * 0 0 * * 0 0
    GATGCAGCGACTGGACTATTTAGGACGATCGGACGGAGGAGGGCAGTAGGACGCTACGTATTTGGCGCGCGGACC
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_96_238_3  77  * 0 0 * * 0 0
    GATCCGTCTACGTCCGCGTAAGATAATAAAGTACTAGTAGCGTATCGCAAGCTCGACGCTCAGCCGTAGGGCCGC
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_75_235_21 77  * 0 0 * * 0 0
    ACGGACGGACTTAGAGCGTCAGATGCAGCGACTGGACTATTTAGCACGATCGGACTGAGGAGGGCAGTAGAACGT
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_75_235_21 141 * 0 0 * * 0 0
    CCGTCTACGTCCGAGTAAGATAATAAAGTAATAGTGGCGTATCGCAAGCTCAACGCTCAGCCGTAGGGCCGTGCG
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_48_207_39 77  * 0 0 * * 0 0
    GCCCCTACGGGGATGACGACTAGGACTACGGACGGATTTAGACCGTCAGATGCAGCGACTGGACTATTTAGGACG
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_48_207_39 141 * 0 0 * * 0 0
    TAAGAGTGGCGTATCGCAAGATCGACGCTCAGCCGTAGGGCCGCGCGCCAAATACGTAGCGTCCTACTTCCCTCC
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_94_291_40 141 * 0 0 * * 0 0
    GTCCCAAAAAGTCCGCCTCGATCCCAGTCTCAAGTAGAAAATGTCGCGTTGCCGATCCGTCTACGTCCCAGGAAG
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_94_291_40 77  * 0 0 * * 0 0
    CAGATGCAGCGACTGTACTATTTAGGACGACCTGACTGAGGAGGGTAGTAGGACGCTACGTATTTGGCGCGCGGC
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_33_189_4a 77  * 0 0 * * 0 0
    AGCCCGACGCTCAGCCGTAGGGCCGCGCGCCAAATAGGTAGCGTCCTACTGCCCTCCTCAGTCCGATCGTCCTAA
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

```
test_mRNA_172_294_4f  77  * 0 0 * * 0 0
    ACGGATGAGCGTCGAGCTTGCGATACGCCACTATTACTTTATTATCTTCCTCGGACGTAGACGGATCGCCAACGG
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_33_189_4a 141 * 0 0 * * 0 0
    ACTGAGCTAGGACGTGCCACTACGGGGATTACCACTAGGGCTACGGACGGACTTAGAGCGTCAGATGCAGCGACT
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_172_294_4f  141 * 0 0 * * 0 0
    CCCGTCCTAAAACGTCCGCCTCGATCCCAGTCTCAAGTAGAAAAAGTCCCGCTGCCGACCCGTCTACGTCCGAGT
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_4_191_5d  77  * 0 0 * * 0 0
    CAAGCTCGACGCTCAGCCGTAGGGCCGCGCGCCAAATACGTAGTGTCCTACTGCCCTACTCAGTCCGATCGTCCT
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
test_mRNA_4_191_5d  141 * 0 0 * * 0 0
    ACTATCTGACGAGACTGGAGGCGCTTGCGACTGAGCTAGGACGTACCATTACGCGGATGACGACTAGGACTACGG
    IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

## Example Use Case: Picard AddOrReplaceReadGroups

We can try to add a basic read group header to a merged file `merged.bam`, generated by merging the `accepted_hits.bam` with either the original `unmapped.bam` or the corrected `unmapped_fixup.bam` file.
With the original `unmapped.bam`:

```
$ samtools merge merged.bam accepted_hits.bam unmapped_fixup.bam
$ samtools sort merged.bam merged_refsort
$
$ java -jar ~/software/picard-tools-1.115/AddOrReplaceReadGroups.jar INPUT=merged_refsort.bam OUTPUT=
    merged_refsort_rg.bam RGLB=1 RGPL=illumina RGPU=NA RGSM=LU
[Thu Nov 12 10:40:56 CET 2015] picard.sam.AddOrReplaceReadGroups INPUT=merged_refsort.bam OUTPUT=
    merged_refsort_rg.bam RGLB=1 RGPL=illumina RGPU=NA RGSM=LU    RGID=1 VERBOSITY=INFO QUIET=false
    VALIDATION_STRINGENCY=STRICT COMPRESSION_LEVEL=5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=false
    CREATE_MD5_FILE=false
[Thu Nov 12 10:40:56 CET 2015] Executing as chris@host on Linux 2.6.32-358.23.2.el6.x86_64 amd64;
    OpenJDK 64-Bit Server VM 1.7.0_45-mockbuild_2013_10_23_08_18-b00; Picard version: 1.115(30
    b1e546cc4dd80c918e151dbfe46b061e63f315_1402927010) JdkDeflater
INFO  2015-11-12 10:40:56 AddOrReplaceReadGroups  Created read group ID=1 PL=illumina LB=1 SM=LU

[Thu Nov 12 10:40:56 CET 2015] picard.sam.AddOrReplaceReadGroups done. Elapsed time: 0.00 minutes.
Runtime.totalMemory()=376963072
To get help, see http://picard.sourceforge.net/index.shtml#GettingHelp
Exception in thread "main" htsjdk.samtools.SAMFormatException: SAM validation error: ERROR: Record 143,
     Read name test_mRNA_150_290_0, Mapped mate should have mate reference name
  at htsjdk.samtools.SAMUtils.processValidationErrors(SAMUtils.java:452)
  at htsjdk.samtools.BAMFileReader$BAMFileIterator.advance(BAMFileReader.java:643)
  at htsjdk.samtools.BAMFileReader$BAMFileIterator.next(BAMFileReader.java:628)
  at htsjdk.samtools.BAMFileReader$BAMFileIterator.next(BAMFileReader.java:598)
  at htsjdk.samtools.SamReader$AssertingIterator.next(SamReader.java:514)
  at htsjdk.samtools.SamReader$AssertingIterator.next(SamReader.java:488)
  at picard.sam.AddOrReplaceReadGroups.doWork(AddOrReplaceReadGroups.java:107)
  at picard.cmdline.CommandLineProgram.instanceMain(CommandLineProgram.java:183)
  at picard.cmdline.CommandLineProgram.instanceMainWithExit(CommandLineProgram.java:124)
  at picard.sam.AddOrReplaceReadGroups.main(AddOrReplaceReadGroups.java:74)
```

As the error indicates, Picard AddOrReplaceReadGroups cannot process the merged BAM file containing the original `unmapped.bam` file. Running AddOrReplaceReadGroups with the `VALIDATION_STRINGENCY=LENIENT` option would work by simply ignoring the errors, but the result would be a BAM file with the same issues as the input files.
On the other hand, with the corrected `unmapped_fixup.bam` file, the command succeeds:

```
$ samtools merge merged.bam accepted_hits.bam unmapped_fixup.bam
$ samtools sort merged_fixup.bam merged_fixup_refsort
$
$ java -jar ~/software/picard-tools-1.115/AddOrReplaceReadGroups.jar INPUT=merged_fixup_sort.bam OUTPUT
    =merged_fixup_refsort_rg.bam RGLB=1 RGPL=illumina RGPU=NA RGSM=LU
[Wed Nov 11 17:43:33 CET 2015] picard.sam.AddOrReplaceReadGroups INPUT=merged_fixup_refsort.bam OUTPUT=
    merged_fixup_refsort_rg.bam RGLB=1 RGPL=illumina RGPU=NA RGSM=LU    RGID=1 VERBOSITY=INFO QUIET=
    false VALIDATION_STRINGENCY=STRICT COMPRESSION_LEVEL=5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=false
     CREATE_MD5_FILE=false
[Wed Nov 11 17:43:33 CET 2015] Executing as chris@host on Linux 2.6.32-358.23.2.el6.x86_64 amd64;
    OpenJDK 64-Bit Server VM 1.7.0_45-mockbuild_2013_10_23_08_18-b00; Picard version: 1.115(30
    b1e546cc4dd80c918e151dbfe46b061e63f315_1402927010) JdkDeflater
```

```
INFO  2015-11-11 17:43:33 AddOrReplaceReadGroups  Created read group ID=1 PL=illumina LB=1 SM=LU

[Wed Nov 11 17:43:33 CET 2015] picard.sam.AddOrReplaceReadGroups done. Elapsed time: 0.00 minutes.
Runtime.totalMemory()=376963072
```

In conclusion, the `unmapped_fixup.bam` or `merged_fixup.bam` files containing the corrected unmapped reads can be used as input for further BAM processing and analysis software, e.g., Picard, GATK, or quality assessment software like RNA-SeQC (`https://www.broadinstitute.org/cancer/cga/rna-seqc`). This can be done without the need for reduced strictness requirements that could mask other problems in the data file, or discarding non-conforming reads from the file, both of which would lead to ignoring potentially useful data. The corrected files can also be deposited in a sequencing archive like NCBI Gene Expression Omnibus (GEO) or the European Nucleotide Archive (ENA), without the need for others to deal with the problems described in this paper.

# Study III

original reports

# Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative

Christian Brueffer

Johan Vallon-Christersson

Dorthe Grabau†

Anna Ehinger

Jari Häkkinen

Cecilia Hegardt

Janne Malina

Yilun Chen

Pär-Ola Bendahl

Jonas Manjer

Martin Malmberg

Christer Larsson

Niklas Loman

Lisa Rydén

Åke Borg

Lao H. Saal

Author affiliations and support information (if applicable) appear at the end of this article.

†Deceased.

C.B. and J.V.-C. contributed equally to this work.

(continued)

abstract

**Purpose** In early breast cancer (BC), five conventional biomarkers—estrogen receptor (ER), progesterone receptor (PgR), human epidermal growth factor receptor 2 (HER2), Ki67, and Nottingham histologic grade (NHG)—are used to determine prognosis and treatment. We aimed to develop classifiers for these biomarkers that were based on tumor mRNA sequencing (RNA-seq), compare classification performance, and test whether such predictors could add value for risk stratification.

**Methods** In total, 3,678 patients with BC were studied. For 405 tumors, a comprehensive multi-rater histopathologic evaluation was performed. Using RNA-seq data, single-gene classifiers and multigene classifiers (MGCs) were trained on consensus histopathology labels. Trained classifiers were tested on a prospective population-based series of 3,273 BCs that included a median follow-up of 52 months (Sweden Cancerome Analysis Network—Breast [SCAN-B], ClinicalTrials.gov identifier: NCT02306096), and results were evaluated by agreement statistics and Kaplan-Meier and Cox survival analyses.

**Results** Pathologist concordance was high for ER, PgR, and HER2 (average κ, 0.920, 0.891, and 0.899, respectively) but moderate for Ki67 and NHG (average κ, 0.734 and 0.581). Concordance between RNA-seq classifiers and histopathology for the independent cohort of 3,273 was similar to interpathologist concordance. Patients with discordant classifications, predicted as hormone responsive by histopathology but non–hormone responsive by MGC, had significantly inferior overall survival compared with patients who had concordant results. This extended to patients who received no adjuvant therapy (hazard ratio [HR], 3.19; 95% CI, 1.19 to 8.57), or endocrine therapy alone (HR, 2.64; 95% CI, 1.55 to 4.51). For cases identified as hormone responsive by histopathology and who received endocrine therapy alone, the MGC hormone-responsive classifier remained significant after multivariable adjustment (HR, 2.45; 95% CI, 1.39 to 4.34).

**Conclusion** Classification error rates for RNA-seq–based classifiers for the five key BC biomarkers generally were equivalent to conventional histopathology. However, RNA-seq classifiers provided added clinical value in particular for tumors determined by histopathology to be hormone responsive but by RNA-seq to be hormone insensitive.

**JCO Precis Oncol. © 2018 by American Society of Clinical Oncology** Licensed under the Creative Commons Attribution 4.0 License

**Corresponding author:** Lao H. Saal, MD, PhD, Department of Clinical Sciences Lund, Division of Oncology and Pathology, Lund University Cancer Center, Medicon Village 404-B2, SE-22381 Lund, Sweden; e-mail: lao.saal@med.lu.se; Twitter: @LaoSaal.

## INTRODUCTION

Histopathologic analysis of breast cancers (BCs) for estrogen receptor (ER) and progesterone receptor (PgR) content, human epidermal growth factor receptor 2 (HER2) gene amplification, and Nottingham histologic grade (NHG) are the mainstays of current clinical practice.[1] Increasingly, assessment of the proliferation antigen Ki67 is clinically recommended.[2] These five biomarkers carry prognostic and predictive information and are used in combination with other clinicopathological factors for risk stratification and therapy selection.[1]

Current evaluation of these BC biomarkers is imperfect. Immunohistochemistry (IHC) is the principal method for ER, PgR, HER2, and Ki67 measurement, and in situ hybridization (ISH) methods are used to refine HER2 IHC. Among laboratories, significant differences exist in, for example, fixation, antigen retrieval, antibodies, chemistries, scoring systems, and interpretation. Accuracy and reproducibility are concerns, with up to 20% false-positive or false-negative ER/PgR IHC determinations.[3] Varying discordance has been reported for HER2 IHC and fluorescent ISH (FISH).[4-7] Accordingly, consensus guidelines emphasize standardization and validation of analytic performance.[1,2,8] Lack of standardization has slowed the entrance of Ki67 into clinical routines.[9] For example, Ki67 status was only moderately concordant in an interlaboratory reproducibility analysis.[10] Thresholds for Ki67 positivity are evolving; cutoffs between 20% and 29% were recommended by the 2015 St Gallen/Vienna panel for laboratories with a quality assurance program.[11] Swedish quality assurance program guidelines recommend that each laboratory calibrate a cutoff yearly such that one third of 100 consecutive occurrences are Ki67-high. The NHG system was developed to establish better standards and improve reproducibility, and it is the recommended method for BC grading today. NHG reproducibility studies[12] have reported modest agreements (pairwise κ, 0.43 to 0.83), which correspond to 15% to 30% discordance.

Microarray and reverse transcriptase polymerase chain reaction–based gene expression analyses of BCs have yielded many signatures for tumor subtyping, prognosis, and survival, as well as for individual biomarkers, such as ER, PgR, HER2, and PTEN.[13-16] Massively parallel sequencing of mRNA (RNA-seq) has advantages compared with earlier methods, including greater dynamic range and reproducibility and the ability to discover and quantify transcripts without a priori sequence knowledge. In 2010, toward implementation of molecular profiling in the clinical routine, we launched the Sweden Cancerome Analysis Network Breast Initiative (SCAN-B; ClinicalTrials.gov identifier: NCT02306096), an ongoing population-based multicenter observational study covering a wide geography of Sweden that prospectively invites all patients with BC to participate.[17] To date, approximately 85% of the eligible catchment population are included, more than 11,000 patients have enrolled, and blood and fresh tumor tissues are sampled for molecular research. In the first phase, all tumors are analyzed by RNA-seq generally within 1 week after surgery. Thus, for each BC, it will be possible to report a multitude of biomarker tests simultaneously on the basis of its RNA-sequencing data and within a clinically actionable time frame.

Herein, we aimed to validate the SCAN-B multicenter infrastructure and provide molecular analyses of clinical value by developing RNA-seq–derived classifiers for the conventional histopathologic BC biomarkers ER, PgR, HER2, Ki67, and NHG. For this purpose, both single-gene classifiers (SGCs) and multigene classifiers (MGCs) were developed by using a training cohort, the prediction accuracy was compared against current clinical practice across a large independent prospective cohort, and the classifier predictions and their discrepancies to histopathology were evaluated with respect to patient survival.

## METHODS

### Patients

The study (Fig 1) was approved by the Regional Ethical Review Board of Lund at Lund University and the Swedish Data Inspection group. Health professionals provided patient information, and patients gave written informed consent. Clinical data were retrieved from the Swedish National Breast Cancer Registry. Diagnostic pathology slides, snap-frozen surgical tumor specimens, and formalin-fixed paraffin-embedded tissue blocks were retrieved for 405 patient cases, selected for classifier training with an over-representation of HER2-positive

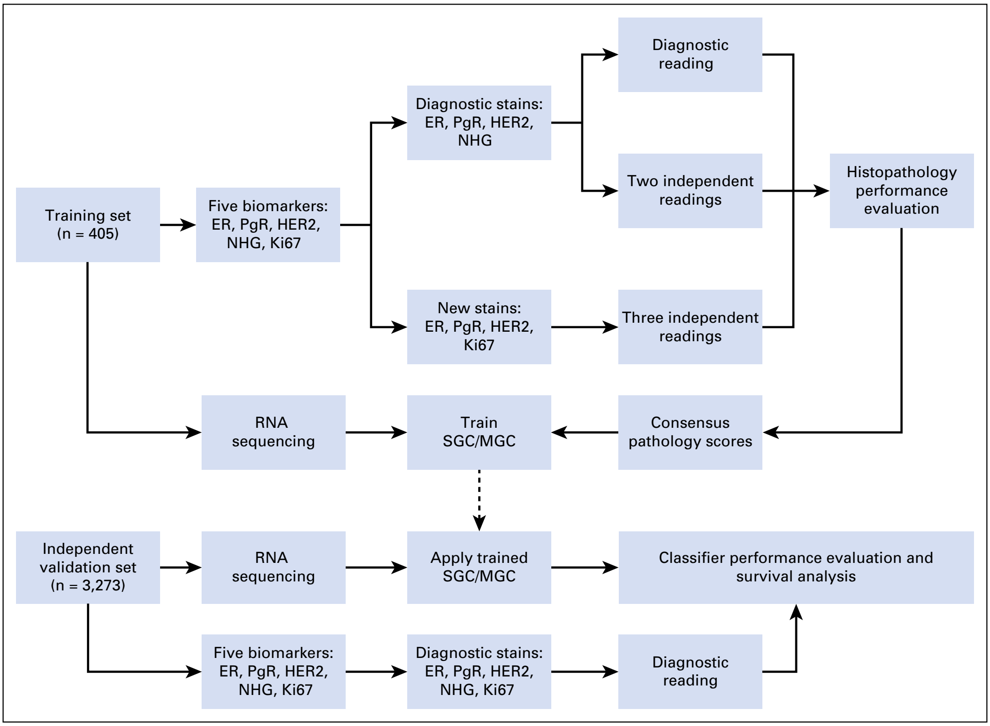


**Fig 1.** Study design flow diagram. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; Ki67, proliferation antigen Ki67; MGC, multigene classifier; NHG, Nottingham histologic grade; PgR, progesterone receptor; SGC, single-gene classifier.

and ER-negative tumors (training cohort; Data Supplement). For classifier testing, an independent, prospective, and population-based modern cohort of 3,273 patients with early BC was assembled from the ongoing SCAN-B study[17] (validation cohort; Appendix Fig A1; Data Supplement).

## Histopathology

For the training cohort, all biomarkers with the exception of Ki67 were evaluated at time of diagnosis. In addition, new formalin-fixed paraffin-embedded slides were analyzed for ER, PgR, and Ki67 IHC and for HER2 silver ISH, all performed at a central laboratory (Helsingborg Hospital). The diagnostic slides and newly stained slides were each scored in total by three pathologists independently by using 1% or greater tumor cell staining threshold for hormone receptor positivity, standard HER2 HercepTest (Agilent/Dako, Santa Clara, CA) and ISH criteria (Roche/Ventana, Tucson, AZ), greater than 20% positive nuclei for Ki67-high status, and the NHG scoring system (Data Supplement). On the basis of all evaluations, a consensus score for each biomarker was determined with the majority scores.

## Tumor Processing and RNA Sequencing

Snap-frozen (training cohort) or RNAlater-preserved (validation cohort) tumor specimens were processed and sequenced, and the raw data (Data Supplement) was processed as described previously.[17,18] All data are available from the NCBI Gene Expression Omnibus (Accession Nos. GSE81538 and GSE96058).

## Classifiers

Within the 405-patient training set, SGCs were built for the ER, PgR, HER2, and Ki67 biomarkers by determining the optimal expression thresholds for the genes *ESR1*, *PGR*, *ERBB2*, and *MKI67* that maximized concordance to the respective histopathology consensus score (Data Supplement). MGCs for ER, PgR, HER2, Ki67, and NHG were built by training nearest shrunken centroid (NSC)[19] models with the

5,000 most varying genes across the training cohort (Data Supplement) and the histopathology consensus scores as training labels. Within the training set, optimal model parameters were determined by using cross-validation and then were used to train prediction models with all training samples. The resulting four SGCs and five MGCs were used to predict the biomarker status of 3,273 independent validation BC samples. The biologic functional annotation clusters of each MGC signature were evaluated with the DAVID Bioinformatics Resource.[20]

### Statistical Analysis

Histopathology evaluations and single-gene and multigene predictions were compared with agreement statistics[21] (defined in the Data Supplement) and balanced statistics—Cohen's κ and Matthews correlation coefficient (MCC)—and were interpreted according to Viera and Garrett.[22] The κ and MCC values were comparable (Data Supplement), so we focused on κ. Kaplan-Meier and Cox regression survival analyses were performed with overall survival as the end point. Multivariable Cox models included the variables age at diagnosis, lymph node status, tumor size, ER, PgR, HER2, and NHG as covariates, as relevant (Data Supplement). All calculations were performed with R 3.2.3. *P* values of ≤ .05 were considered significant.

## RESULTS

### Clinical Histopathology

To estimate the inherent variability within clinical histopathology and to determine a consensus score for each BC biomarker for classifier training, a comprehensive histopathologic analysis was performed for 405 patient breast tumors with three readings of up to two independent stains for the five conventional biomarkers: ER, PgR, HER2, Ki67, and NHG (Fig 1). With the diagnostic evaluation as the reference, agreement statistics were calculated (Table 1; Data Supplement). Concordance for histopathologic evaluation of ER, PgR, and HER2 into positive and negative groups was high; the average pairwise agreements were 97.3% (average κ [Aκ], 0.920), 95.5% (Aκ, 0.891), and 96.6% (Aκ, 0.899), respectively, whereas agreements were lower for Ki67 (86.8%; Aκ, 0.734) and NHG (74.8%; Aκ, 0.581). As expected with minimization of technical and heterogeneity factors, within-slide concordances were slightly better than between-slide concordances (Data Supplement).

### Classifier Training

Whole-transcriptome expression profiles were generated for the 405 training samples using RNA-seq. For the SGCs, optimal thresholds were determined for *ESR1* (which encodes the ER protein), *PGR* (PgR), *ERBB2* (HER2), and *MKI67* (Ki67) (Data Supplement). Next, MGCs were trained, and the training-cohort cross-validation accuracy was determined (balanced accuracy or accuracy ± standard deviation; Data Supplement) as follows: ER, 95.3% ± 2.4%; PgR, 90.4% ± 2.9%; HER2, 88.5% ± 3.8%; Ki67, 84.9% ± 3.4%; and NHG, 73.8% ± 3.9%. For MGCs, the NSC method has the property of eliminating noninformative genes (zero weight for the classification). The ER classifier had 459 weighted genes; PgR, 184; HER2, 312; Ki67, 273; and NHG, 206 (Data Supplement). In total, 869 genes had nonzero weights in at least one MGC classifier. The constituent biologic themes for each MGC classifier were investigated with functional annotation clustering (Data Supplement).

### Performance on Independent Data

To evaluate the classifiers, we tested them on RNA-seq data generated for 3,273 independent tumors from the prospective population-based multicenter SCAN-B study (n = 136 tumors were analyzed in technical replicates). Concordance between the diagnostic histopathologic results and the SGC predictions was substantial for ER (overall agreement, [OA], 96.1%; κ, 0.730) and HER2 (OA, 94.92%; κ, 0.749) and moderate for PgR (OA, 89.6%; κ, 0.588) and Ki67 (OA, 76.7%; κ, 0.516; Fig 2; Appendix Figs A2 and A3; Data Supplement). Similarly, for the MGCs, concordance was substantial for ER (OA, 91.9%; κ, 0.606) and HER2 (OA, 92.4%; κ, 0.667), moderate for PgR (OA, 88.7%; κ, 0.568) and NHG (OA, 67.7%; κ, 0.418), and fair for Ki67 (OA, 66.3%; κ, 0.370). For RNA-seq replicates, 534 (98.2%) of 544 SGC classifications and 675 (99.3%) of 680 MGC classifications were concordant (Data Supplement). Similar results were obtained when an ER/PgR IHC cutoff of 10% or greater positive cells (current Swedish standard) was used.

Table 1. Concordance Among Three Pathologist Evaluations for Five Biomarkers and Multiple Stains Within the Training Cohort

| Biomarker Staining Pathology | Overall Agreement | | | Concordance | | |
|---|---|---|---|---|---|---|
| | % | 95% CI | | κ | 95% CI | |
| ER (diagnostic IHC) | | | | | | |
| Routine (reference) | — | — | — | — | — | — |
| Versus 2 | 98.8 | 97.1 | 99.6 | 0.965 | 0.931 | 0.993 |
| Versus 3 | 98.8 | 97.1 | 99.6 | 0.965 | 0.931 | 0.993 |
| ER (new IHC) | | | | | | |
| Versus 1 | 95.8 | 93.4 | 97.5 | 0.873 | 0.810 | 0.929 |
| Versus 2 | 96.5 | 94.3 | 98.1 | 0.898 | 0.842 | 0.947 |
| Versus 3 | 96.5 | 94.3 | 98.1 | 0.898 | 0.842 | 0.947 |
| ER summarized | | | | | | |
| Average (*v* reference) | 97.3 | 95.2 | 98.6 | 0.920 | 0.871 | 0.962 |
| Complete concordance | 94.1 (381 of 405) | | | | | |
| PgR (diagnostic IHC) | | | | | | |
| Routine (reference) | — | — | — | — | — | — |
| Versus 2 | 96.0 | 93.7 | 97.7 | 0.904 | 0.855 | 0.947 |
| Versus 3 | 96.0 | 93.7 | 97.7 | 0.902 | 0.851 | 0.945 |
| PgR (new IHC) | | | | | | |
| Versus 1 | 96.0 | 93.7 | 97.7 | 0.905 | 0.857 | 0.949 |
| Versus 2 | 93.8 | 91.0 | 96.0 | 0.853 | 0.793 | 0.906 |
| Versus 3 | 95.3 | 92.8 | 97.2 | 0.889 | 0.836 | 0.934 |
| PgR summarized | | | | | | |
| Average (*v* reference) | 95.5 | 93.0 | 97.3 | 0.891 | 0.838 | 0.936 |
| Complete concordance | 91.1 (369 of 405) | | | | | |
| HER2 (diagnostic IHC) | | | | | | |
| Routine (reference) | — | — | — | — | — | — |
| Versus 2 | 72.8 | 68.2 | 77.1 | 0.628 | 0.568 | 0.686 |
| Versus 3 | 75.3 | 70.8 | 79.4 | 0.661 | 0.602 | 0.717 |
| HER2 (new SISH) | | | | | | |
| Clinical status (reference) | — | — | — | — | — | — |
| Versus 1 | 96.6 | 94.3 | 98.2 | 0.902 | 0.844 | 0.95 |
| Versus 2 | 96.4 | 94.1 | 98.0 | 0.895 | 0.837 | 0.945 |
| Versus 3 | 96.6 | 94.3 | 98.2 | 0.901 | 0.844 | 0.95 |
| HER2 SISH summarized | | | | | | |
| Average (*v* reference) | 96.6 | 94.2 | 98.1 | 0.899 | 0.842 | 0.948 |
| Complete concordance | 96.3 (360 of 374) | | | | | |
| Ki67 (new IHC) | | | | | | |
| Reader 1 (reference) | — | — | — | — | — | — |
| Versus 2 | 85.9 | 82.2 | 89.2 | 0.717 | 0.648 | 0.783 |
| Versus 3 | 87.7 | 84.0 | 90.7 | 0.751 | 0.684 | 0.811 |

(Continued on following page)

**Table 1.** Concordance Among Three Pathologist Evaluations for Five Biomarkers and Multiple Stains Within the Training Cohort (Continued)

| Biomarker Staining Pathology | Overall Agreement | | | Concordance | | |
|---|---|---|---|---|---|---|
| | % | 95% CI | | κ | 95% CI | |
| Ki67 summarized | | | | | | |
| Average (*v* reference) | 86.8 | 83.1 | 89.9 | 0.734 | 0.666 | 0.797 |
| Complete concordance | 80.0<br>(324 of 405) | | | | | |
| NHG (diagnostic H&E) | | | | | | |
| Routine (reference) | — | — | — | — | — | — |
| Versus 2 | 75.3 | 70.8 | 79.4 | 0.589 | 0.520 | 0.655 |
| Versus 3 | 74.3 | 69.8 | 78.5 | 0.573 | 0.504 | 0.642 |
| NHG summarized | | | | | | |
| Average (*v* reference) | 74.8 | 70.3 | 79.0 | 0.581 | 0.512 | 0.649 |
| Complete concordance | 62.0<br>(251 of 405) | | | | | |

NOTE. Within a biomarker staining group (left-most column headings), all comparisons presented are the reference evaluation (the diagnostic reading made in the clinical routine, or reader 1 in the case of Ki67) versus each specified reader number. Overall agreement was defined as the number of concordant determinations (assigned to the same class) divided by the total sample size. Complete concordance was defined as the number of occurrences for which all readings were concordant across all stains divided by the total sample size (Data Supplement).

Abbreviations: ER, estrogen receptor; H&E, hematoxylin and eosin; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; NHG, Nottingham histologic grade; PgR, progesterone receptor; SISH, silver in situ hybridization.

### Survival Analysis

To evaluate the possible clinical utility of our classifiers, we analyzed our classifier predictions within the validation cohort with respect to overall survival. Kaplan-Meier analysis revealed comparable patient stratification for both diagnostic histopathology and SGCs for the five biomarkers across the entire validation cohort, whereas the MGCs had a noticeably richer stratification, particularly for the hormone receptors and the hormone-responsive group, defined by ER positivity and PgR positivity (Appendix Figs A4 and A5). Therefore, and to reduce the number of comparisons, we focused on the MGCs for each biomarker and within the major treatment groups. Patients with tumors discrepant for hormone responsiveness (hormone responsive by pathology but not responsive by MGC) had significantly worse outcomes across the entire cohort (hazard ratio [HR], 1.64; 95% CI, 1.17 to 2.28; log-rank $P$ = .0034) as well as within subgroups defined by adjuvant treatment: no systemic therapy (HR, 3.19; 95% CI, 1.19 to 8.57; $P$ = .015) and only endocrine therapy (HR, 2.64; 95% CI, 1.55 to 4.51; $P$ < .001; Fig. 3A). Furthermore, MGC predictions added value to predictions of HER2, Ki67, and NHG (Figs 3B to 3D). After adjusting for important covariates in multivariable Cox analyses, the MGC prediction for hormone nonresponsiveness was a significant stratifier among patients with histopathologic hormone-responsive disease who were treated with endocrine therapy, as were the MGC predictions discordant for HER2-negative or Ki67-high status in patients who received chemotherapy with or without trastuzumab and/or endocrine therapy. Conversely, the NHG MGC became nonsignificant (Fig 3).

## DISCUSSION

Despite efforts to develop better standards for clinical histopathologic evaluation of breast tumors, intra/interlaboratory and -reader variation remain problematic. Previously, several gene expression–based approaches for determination of known treatment-predictive biomarkers have been developed[16,23-26]; however, they are not widely used clinically in most countries. Supplementation of histopathologic biomarkers with biomarkers determined from RNA-seq profiling is becoming feasible today: costs are less than $300 per transcriptome, and projects, such as SCAN-B and others, that use RNA-seq in the clinic are emerging.[17,27,28] In this study, we

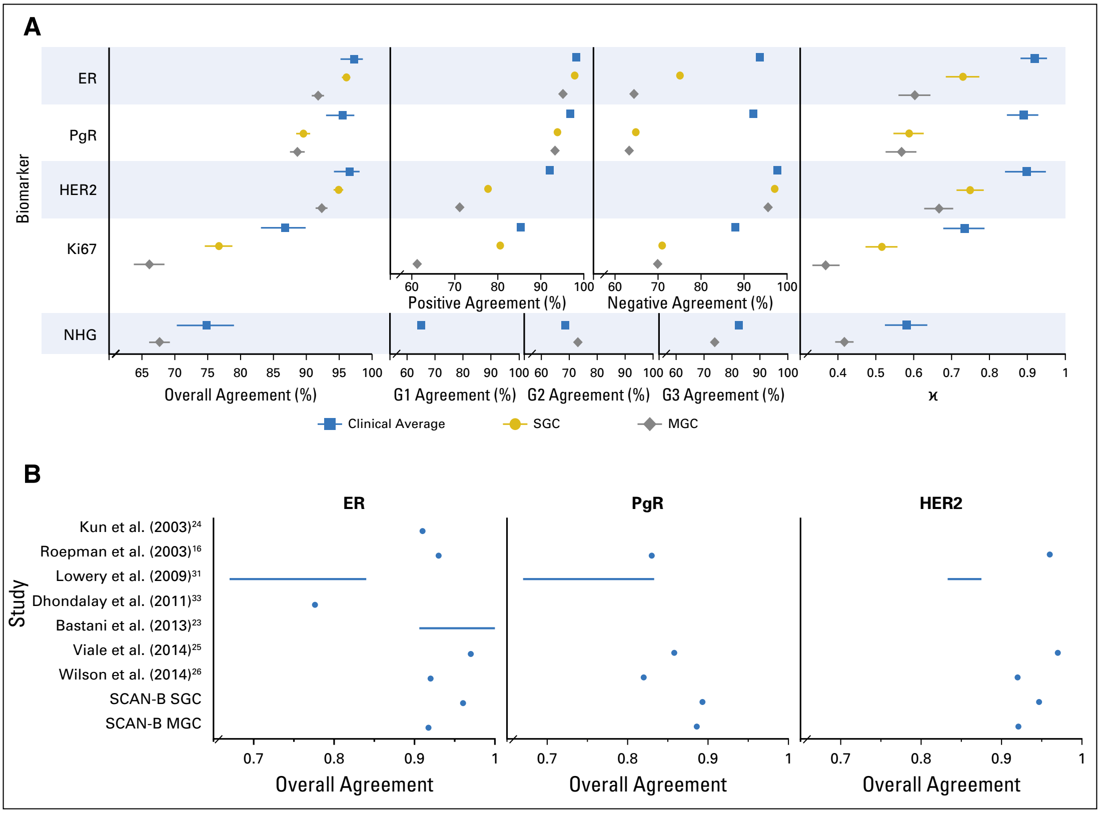


**Fig 2.** Performance of trained classifiers in the 3,273-tumor independent validation cohort. (A) Forest plots of concordance statistics for histopathologic evaluation in the training set (blue square markers), and single-gene classifiers (SGCs; gold circles) and multigene classifiers (MGCs; gray diamonds) in the validation cohort, which plots overall agreement with 95% CIs, specific agreements (positive and negative agreements for estrogen receptor [ER], progesterone receptor [PgR], human epidermal growth factor receptor 2 [HER2], and Ki67) and Nottingham histologic grade (NHG) category agreements (grade [G] 1, G2, and G3), and κ values with 95% CIs. Overall agreement is defined as the number of concordant determinations (assigned to the same class) divided by the total sample size. Positive, negative, and G1/G2/G3 agreements are the proportions of agreement specific to the given category (Data Supplement). (B) Overall agreement of classifiers from the literature compared with our SGCs and MGCs. SCAN-B, Sweden Cancerome Analysis Network—Breast.

demonstrated that accurate classifiers for ER, PgR, HER2, Ki67 and NHG can be built with RNA-seq data, can provide a valuable complement to traditional histopathology, and represent the first of many potential clinical reports that can be delivered from a single RNA-seq measurement. In the future, we foresee the development, validation, and clinical implementation of a multitude of signatures, classifiers, and mutational profiles within the SCAN-B population-based infrastructure and RNA-seq platform.[17,18] We also aim to use RNA-seq analyses in the performance of interventional clinical trials.[29]

The quality of machine-learned classifiers is crucially dependent on the quality of the labels on which they have been trained. To ensure highly accurate pathology labels, we sought to reduce variance by generating consensus scores for each biomarker. Matched against routine histopathologic evaluation, repeated ER, PgR, and HER2 readings showed good concordance, whereas Ki67 and NHG had notably lower concordance between pathologists (Table 1). Reproducibility of tumor grading systems has long been debated,[30] and Ki67 has been shown to have high intralaboratory but low interlaboratory

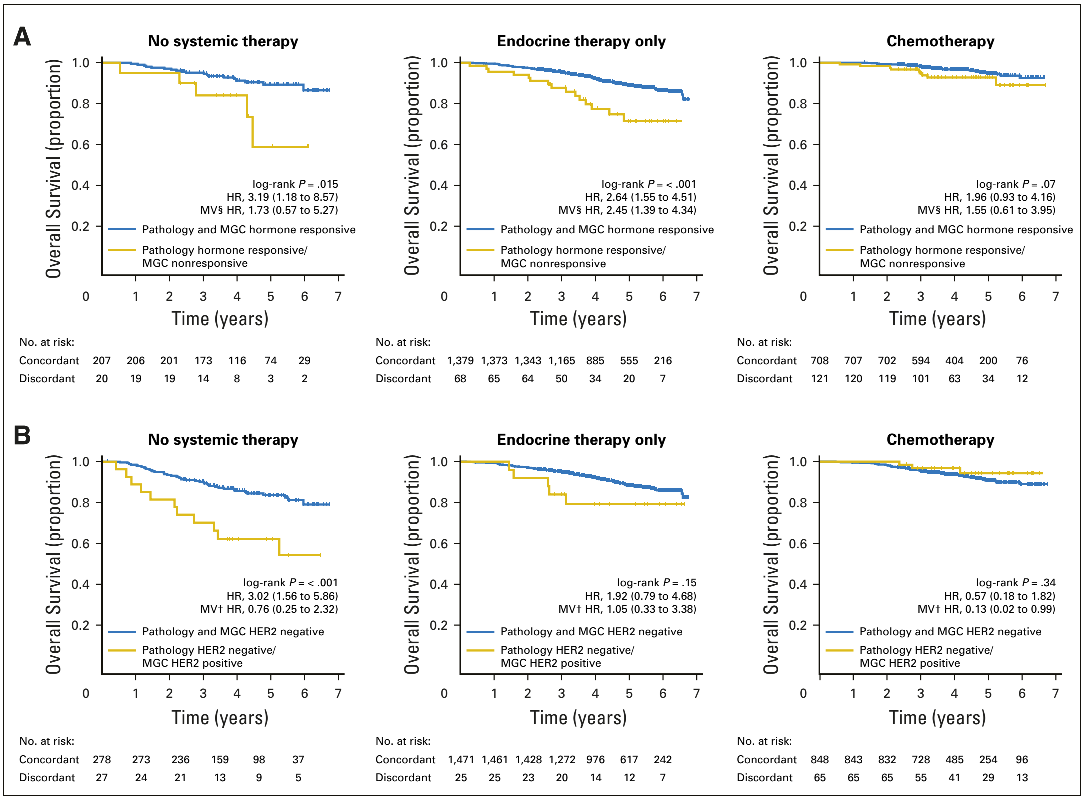


**Fig 3.** Kaplan-Meier overall survival estimates and Cox regression survival analysis for multigene classifiers (MGCs) within the independent validation cohort. (A) Histopathologically hormone responsive (defined as estrogen receptor [ER] positive and progesterone receptor [PgR] positive) group stratified by MGC hormone responsive classification (concordant [blue curve] or discordant [gold curve] to histopathology) within the subgroup of patients who received (left) no adjuvant systemic therapy, (middle) endocrine therapy alone, or (right) chemotherapy with or without trastuzumab or endocrine therapy. (B) Human epidermal growth factor receptor 2 [HER2]–negative histopathology group stratified by HER2 MGC for the same three treatment subgroups as in A.

reproducibility.[10] Here, the histopathologic variability was highest for Ki67 and NHG, which added uncertainty even to our consensus scores. It is unlikely that a classifier would perform better than the quality of training labels; therefore, it is not surprising that our classifiers had the worst performance for Ki67 and NHG. Moreover, because we benchmarked our biomarker predictions in the validation cohort to the clinical diagnostic histopathology results that contained this inherent variability, we could not expect our classifiers to have higher accuracy than what is achievable within histopathology.

Generally, SGCs performed comparably to clinical diagnostic pathology. The SGC ER and HER2 classifiers had substantial κ agreement compared with the clinical average, and PgR and Ki67 had moderate agreement. Likewise, our MGCs had comparable performance. The MGC ER and HER2 classifiers had substantial agreement in line with the clinical average, whereas PgR and NHG classifiers had moderate agreement, and the Ki67 classifier had fair agreement. Earlier work on mRNA-based classifiers for ER, PgR, and HER2 has been performed with microarrays, quantitative reverse-transcriptase polymerase chain reaction, and, recently, with RNA-seq and mainly has been restricted to signatures of either one[16,31] or few[23,24,26,32,33] genes. The performance of our classifiers generally were in line with the results of these previous studies, which indicates the suitability of our RNA-seq approach (Fig 2B).

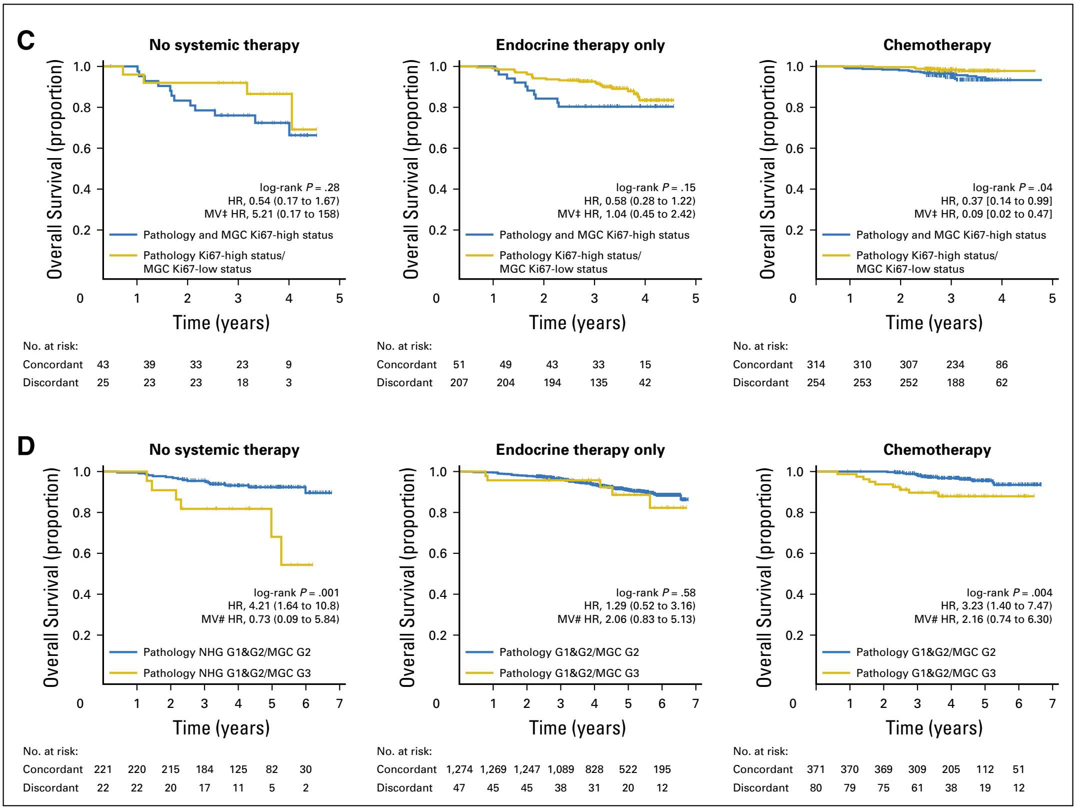


**Fig 3.** (Continued). (C) Ki67-high histopathology group stratified by Ki67 MGC for the same three treatment subgroups as in A. (D) Nottingham histologic grade (NHG) combined grade [G] 1 and G2 histopathology group stratified by NHG MGC for the same three treatment subgroups as in A. In each Kaplan-Meier plot, the histopathology to MGC concordant tumor cases are plotted in blue, the discordant tumor cases are plotted in gold, the log-rank *P* value is given, and the hazard ratio (HR) for discordant-versus-concordant result is given with a 95% CI and after multivariable (MV) Cox regression adjustment. Covariables included in the MV analysis were age at diagnosis, lymph node status, tumor size, and the variables denoted by the following symbols: †, ER, PgR, and NHG; ‡, ER, PgR, HER2, and NHG; §, HER2 and NHG; #, ER, PgR, and HER2.

Discrepancies between RNA-seq–based classifications and histopathology may be a result of staining and reader variations, as discussed in this paper. Discrepancies may also develop from tissue sampling and heterogeneity, in which the specimen used for sequencing may not be representative of the piece selected for histopathology. Another consideration is the biologic layer at which biomarker status is assessed: mRNA versus protein abundance or DNA copy number. The consequence is that a mismatch between mRNA biomarker prediction and histopathology may be influenced by various mechanisms active between these layers, for example RNA silencing/interference/translation, protein stability and epitope availability, or tumor heterogeneity.

Despite these possible explanations for discrepancies, when benchmarked against patient outcome, our classifiers exemplified enhanced stratification of patients with significant differences in overall survival (Fig 3; Appendix Figs A4 and A5). The fact that MGCs performed best overall suggests that a multigene signature captures the biologic signaling up- and downstream of the biomarker in question in a more consequential way than the expression of the single gene or protein alone. This conclusion is supported by each signature's underlying biologic themes and pathways (Data Supplement), and by our observation for technical replicates, in which MGCs had near-perfect reproducibility and an error rate that was approximately half that of SGCs (0.7% *v* 1.8%). Ultimately, these results can be used to identify patients who may benefit from additional treatment. Another approach is to use clinical outcome as the training labels to develop new prognostic/predictive signatures.[13,34] The SCAN-B material is excellently suited to evaluate previously published signatures; as we accrue longer follow-up, we aim to develop RNA-seq signatures trained on clinical outcomes.

Ki67 has been introduced relatively recently in international guidelines.[11] To our knowledge, this study is the first to develop a validated predictor for Ki67 status. The lower concordance between our Ki67 predictions compared with the clinical reference is related to the relatively larger Ki67 interrater disagreement seen within our consensus pathology evaluation, which is likely a consequence of the continuous nature of Ki67 expression and of the spectrum of proliferation activity and pathways in BC.

NHG is distinct from the other biomarkers. It has no single underlying gene but rather is a compound biomarker that consists of three morphologic properties: tubular differentiation, nuclear pleomorphism, and mitotic count. Moreover, NHG prediction is a three-class problem. Even for pathologists, NHG can be difficult to determine, as evidenced by the moderate κ and OA results within clinical pathology, in line with the literature.[12] Most misclassified tumor cases in this study were histologically grade 1 (G1) or grade 3 that were misclassified as grade 2 (G2) by our predictor. Large interrater disagreement, especially for G1 and G2, could explain the results of our classifier with only moderate OA to histopathology (67.7%). All histologic G1 occurrences were misclassified, which may have been a result of the imbalanced composition of the training set for NHG (48 of 405 samples consensus-scored G1), or may have occurred because G1 is not a discrete entity but rather the lower end of an underlying continuous scale. Indeed, Kaplan-Meier analysis showed that the curves G1 and G2 largely overlapped in the validation cohort (Appendix Fig A4). Another approach, instead of recapitulation of the pathology grading scheme, could be to reduce the problem to a binary classification of either low or high grade. This approach has been suggested by others as a viable gene-expression–based alternative to NHG for translation into a clinical setting[35,36] and essentially is what our NHG predictor has become.

An important question when building classifiers is how many genes to use. We compared single-gene and multigene classifiers. When compared with clinical pathology, SGCs have slightly better concordance than MGCs for ER and HER2, whereas the SGC and MGC performances were comparable for PgR and Ki67. This difference may have developed because these biomarkers are faithfully represented by their associated single genes. Another consideration for classifiers is robustness toward missing values. MGCs may be more robust than SGCs, because they are able to classify tumors correctly even when the main gene that underlies a biomarker is poorly measured in a particular analysis. When clinical outcome was considered, the survival analyses indicated that our MGCs generally contained greater potential clinical utility than SGCs to complement histopathology.

In summary, we have performed a systematic pathologic evaluation of 405 BC tumors, which resulted in consensus scores for the five conventional BC biomarkers and estimated a well-controlled best-case scenario for the inherent uncertainty within clinical histopathology. With tumor RNA-seq data and the consensus scores, we trained SGCs and MGCs and evaluated the classifiers on an independent set of 3,273 tumors. The accuracy of our classifiers was comparable to the inherent accuracy of clinical pathology and was highly reproducible. Classifiers based on the expression of single genes performed slightly better than MGCs for concordance to histopathology, but MGCs performed significantly better for stratification of patients into groups with clinically meaningful differences in survival, in particular for histopathologic hormone-responsive BCs. In conclusion, RNA-seq–based classifiers may be suitable complementary diagnostics for BC, in particular for difficult diagnoses in which the classifier can add an additional vote toward the therapeutic choice. For future implementation of our MGCs in the clinical routine, additional health economics analyses and external validation are needed.

DOI: https://doi.org/10.1200/PO.17.00135
Published online on ascopubs.org/journal/po on March 9, 2018.

## AUTHOR CONTRIBUTIONS

**Conception and design:** Johan Vallon-Christersson, Dorthe Grabau, Anna Ehinger, Pär-Ola Bendahl, Niklas Loman, Lisa Rydén, Åke Borg, Lao H. Saal

**Collection and assembly of data:** Christian Brueffer, Johan Vallon-Christersson, Dorthe Grabau, Anna Ehinger, Jari Häkkinen, Cecilia Hegardt, Janne Malina, Jonas Manjer, Christer Larsson, Niklas Loman, Lisa Rydén, Åke Borg, Lao H. Saal

**Provision of study material or patients:** Dorthe Grabau, Jonas Manjer, Christer Larsson, Niklas Loman, Lisa Rydén, Åke Borg, Lao H. Saal

**Data analysis and interpretation:** Christian Brueffer, Johan Vallon-Christersson, Dorthe Grabau, Anna Ehinger, Jari Häkkinen, Yilun Chen, Pär-Ola Bendahl, Jonas Manjer, Martin Malmberg, Lisa Rydén, Åke Borg, Lao H. Saal

**Financial support:** Åke Borg, Lao H. Saal

**Administrative support:** Lao H. Saal

**Manuscript writing:** All authors

**Final approval of manuscript:** All authors

**Agree to be accountable for all aspects of the work:** All authors

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.

**Christian Brueffer**
**Employment:** SAGA Diagnostics AB

**Johan Vallon-Christersson**
No relationship to disclose

**Dorthe Grabau**
No relationship to disclose

**Anna Ehinger**
No relationship to disclose

**Jari Häkkinen**
No relationship to disclose

**Cecilia Hegardt**
No relationship to disclose

**Janne Malina**
**Employment:** Unilabs

**Honoraria:** AstraZeneca

**Yilun Chen**
No relationship to disclose

**Pär-Ola Bendahl**
No relationship to disclose

**Jonas Manjer**
No relationship to disclose

**Martin Malmberg**
No relationship to disclose

**Christer Larsson**
**Honoraria:** Lilly (I)

**Research Funding:** Diamyd Medical AB (I)

**Travel, Accommodations, Expenses:** Lilly (I)

**Niklas Loman**
**Honoraria:** AstraZeneca

**Consulting or Advisory Role:** Amgen

**Lisa Rydén**
**Research Funding:** Roche

**Åke Borg**
**Honoraria:** Roche, AstraZeneca

**Travel, Accommodations, Expenses:** Roche, AstraZeneca

**Lao H. Saal**
**Employment:** SAGA Diagnostics AB

**Leadership:** SAGA Diagnostics AB

**Stock and Other Ownership Interests:** SAGA Diagnostics AB

**Patents, Royalties, Other Intellectual Property:** Patent filed for methods related to ultrasensitive quantification of nucleotide sequence variants.

## ACKNOWLEDGMENT

We thank the patients who were part of this study and the SCAN-B study. We also thank Karin Annersten, Minerva Li, Inger Remse, Ralph Schulz, Jeanette Valcich, Cecilia Wahlström, Olle Månsson, Nicklas Nordborg, Anna Karlsson, Christel Reuterswärd, Frida Rosengren, and Ingrid Wilson of the SCAN-B laboratory and the Division of Oncology and Pathology, Lund University, for handling samples, genomic analyses, and database and administrative support, as well as Cristina Ciornei-Karlsson for retrieving pathology slides and patient reports, Daniel Filipazzi and Anders Kvist for help with computing infrastructure, and Johan Staaf for help with classifier development. We acknowledge Roche Diagnostics Scandinavia for providing *HER2* silver in situ hybridization reagents. We thank the South Sweden Breast Cancer Group and all SCAN-B collaborators at Hallands Hospital Halmstad, Helsingborg Hospital, Blekinge County Hospital, Central Hospital Kristianstad, Skåne University Hospital Lund/Malmö, Central Hospital Växjö, for inclusion of patients and sampling of tissue for this study, and we thank the Swedish National Breast Cancer Registry and Regional Cancer Center South for clinical data.

This work is dedicated to the memory of Dorthe Grabau, who sadly passed away during the writing of this paper.

## Affiliations

**Christian Brueffer**, **Johan Vallon-Christersson**, **Anna Ehinger**, **Jari Häkkinen**, **Cecilia Hegardt**, **Yilun Chen**, **Pär-Ola Bendahl**, **Jonas Manjer**, **Christer Larsson**, **Niklas Loman**, **Lisa Rydén**, **Åke Borg**, and **Lao H. Saal**, Lund University, Lund; **Dorthe Grabau**, **Anna Ehinger**, **Martin Malmberg**, **Niklas Loman**, and **Lisa Rydén**, Skåne University Hospital Lund, Lund; **Anna Ehinger**, Blekinge County Hospital, Karlskrona; and **Janne Malina** and **Jonas Manjer**, Skåne University Hospital Malmö, Malmö, Sweden.

**Support**

Supported by the Mrs. Berta Kamprad Foundation and funded in part by the Swedish Foundation for Strategic Research, Swedish Research Council, Swedish Cancer Society, Knut and Alice Wallenberg Foundation, VINNOVA, Governmental Funding of Clinical Research within National Health Service, Scientific Committee of Blekinge County Council, Crafoord Foundation, Lund University Medical Faculty, Gunnar Nilsson Cancer Foundation, Skåne University Hospital Foundation, BioCARE Research Program, King Gustav Vth Jubilee Foundation, Maggie Stephens Foundation, Royal Physiographic Society in Lund, and the Krapperup Foundation.

## REFERENCES

1. Gradishar WJ, Anderson BO, Blair SL, et al: Breast cancer, version 3.2014. J Natl Compr Canc Netw 12:542-590, 2014
2. Senkus E, Kyriakides S, Ohno S, et al: Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 26:v8-v30, 2015
3. Hammond MEH, Hayes DF, Dowsett M, et al: American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol 28:2784-2795, 2010
4. Press MF, Sauter G, Bernstein L, et al: Diagnostic evaluation of HER-2 as a molecular target: An assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. Clin Cancer Res 11:6598-6607, 2005
5. Perez EA, Suman VJ, Davidson NE, et al: HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. J Clin Oncol 24:3032-3038, 2006
6. Rydén L, Haglund M, Bendahl P-O, et al: Reproducibility of human epidermal growth factor receptor 2 analysis in primary breast cancer: A national survey performed at pathology departments in Sweden. Acta Oncol 48:860-866, 2009
7. Ekholm M, Grabau D, Bendahl P-O, et al: Highly reproducible results of breast cancer biomarkers when analysed in accordance with national guidelines: A Swedish survey with central re-assessment. Acta Oncol 54:1040-1048, 2015
8. Wolff AC, Hammond MEH, Hicks DG, et al: Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol 31:3997-4013, 2013
9. Dowsett M, Nielsen TO, A'Hern R, et al: Assessment of Ki67 in breast cancer: Recommendations from the International Ki67 in Breast Cancer working group. J Natl Cancer Inst 103:1656-1664, 2011
10. Polley MYC, Leung SCY, McShane LM, et al: An international Ki67 reproducibility study. J Natl Cancer Inst 105:1897-1906, 2013
11. Gnant M, Thomssen C, Harbeck N: St Gallen/Vienna 2015: A brief summary of the consensus discussion. Breast Care (Basel) 10:124-130, 2015
12. Rakha EA, Reis-Filho JS, Baehner F, et al: Breast cancer prognostic classification in the molecular era: The role of histological grade. Breast Cancer Res 12:207, 2010
13. van 't Veer LJ, Dai H, van de Vijver MJ, et al: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536, 2002
14. Saal LH, Johansson P, Holm K, et al: Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proc Natl Acad Sci USA 104:7564-7569, 2007
15. Parker JS, Mullins M, Cheang MC, et al: Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27:1160-1167, 2009
16. Roepman P, Horlings HM, Krijgsman O, et al: Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. Clin Cancer Res 15:7003-7011, 2009

17. Saal LH, Vallon-Christersson J, Häkkinen J, et al: The Sweden Cancerome Analysis Network Breast (SCAN-B) initiative: A large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med 7:20, 2015

18. Häkkinen J, Nordborg N, Månsson O, et al: Implementation of an open source software solution for laboratory information management and automated RNAseq data analysis in a large-scale cancer genomics initiative using BASE with extension package Reggie. bioRxiv doi: 10.1101/038976 [epub on February 6, 2016]

19. Tibshirani R, Hastie T, Narasimhan B, et al: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99:6567-6572, 2002

20. Huang W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57, 2009

21. Cicchetti DV, Feinstein AR: High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43:551-558, 1990

22. Viera AJ, Garrett JM: Understanding interobserver agreement: The kappa statistic. Fam Med 37:360-363, 2005

23. Bastani M, Vos L, Asgarian N, et al: A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status. PLoS One 8:e82144, 2013

24. Kun Y, How LC, Hoon TP, et al: Classifying the estrogen receptor status of breast cancers by expression profiles reveals a poor prognosis subpopulation exhibiting high expression of the ERBB2 receptor. Hum Mol Genet 12:3245-3258, 2003

25. Viale G, Slaets L, Bogaerts J, et al: High concordance of protein (by IHC), gene (by FISH; HER2 only), and microarray readout (by TargetPrint) of ER, PgR, and HER2: Results from the EORTC 10041/BIG 03-04 MINDACT trial. Ann Oncol 25:816-823, 2014

26. Wilson TR, Xiao Y, Spoerke JM, et al: Development of a robust RNA-based classifier to accurately determine ER, PR, and HER2 status in breast cancer clinical samples. Breast Cancer Res Treat 148:315-325, 2014

27. Cieslik M, Chugh R, Wu YM, et al: The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. Genome Res 25:1372-1381, 2015

28. Roychowdhury S, Iyer MK, Robinson DR, et al: Personalized oncology through integrative high-throughput sequencing: A pilot study. Sci Transl Med 3:111ra121, 2011

29. Cardoso F, van't Veer LJ, Bogaerts J, et al: 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med 375:717-729, 2016

30. Boiesen P, Bendahl PO, Anagnostaki L, et al: Histologic grading in breast cancer: Reproducibility between seven pathologic departments. Acta Oncol 39:41-45, 2000

31. Lowery AJ, Miller N, Devaney A, et al: MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. Breast Cancer Res 11:R27, 2009

32. Rantalainen M, Klevebring D, Lindberg J, et al: Sequencing-based breast cancer diagnostics as an alternative to routine biomarkers. Sci Rep 6:38037, 2016

33. Dhondalay GK, Tong DL, Ball GR: Estrogen receptor status prediction for breast cancer using artificial neural network. Proc Int Conf Mach Learn Cybern 2:727-731, 2011

34. Paik S, Shak S, Tang G, et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817-2826, 2004

35. Ivshina AV, George J, Senko O, et al: Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 66:10292-10301, 2006

36. Sotiriou C, Wirapati P, Loi S, et al: Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262-272, 2006

## Appendix

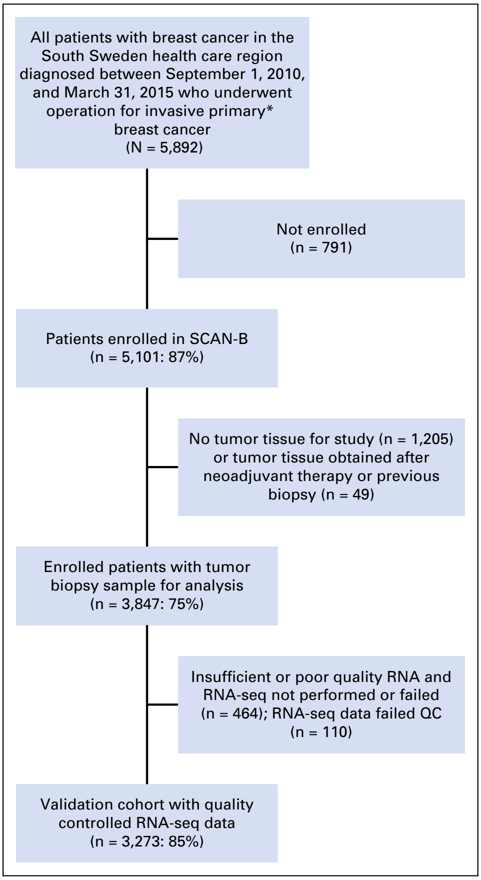

All patients with breast cancer in the South Sweden health care region diagnosed between September 1, 2010, and March 31, 2015 who underwent operation for invasive primary* breast cancer (N = 5,892)
Not enrolled (n = 791)
Patients enrolled in SCAN-B (n = 5,101: 87%)
No tumor tissue for study (n = 1,205) or tumor tissue obtained after neoadjuvant therapy or previous biopsy (n = 49)
Enrolled patients with tumor biopsy sample for analysis (n = 3,847: 75%)
Insufficient or poor quality RNA and RNA-seq not performed or failed (n = 464); RNA-seq data failed QC (n = 110)
Validation cohort with quality controlled RNA-seq data (n = 3,273: 85%)

**Fig A1.** Flow diagram for Sweden Cancerome Analysis Network—Breast (SCAN-B) population-based 3,273-tumor independent validation cohort. (*) Nonmetastatic primary unilateral breast cancer, which excluded patient cases that had a diagnosis of synchronous (< 3 months) contralateral invasive breast cancer. QC, quality control.

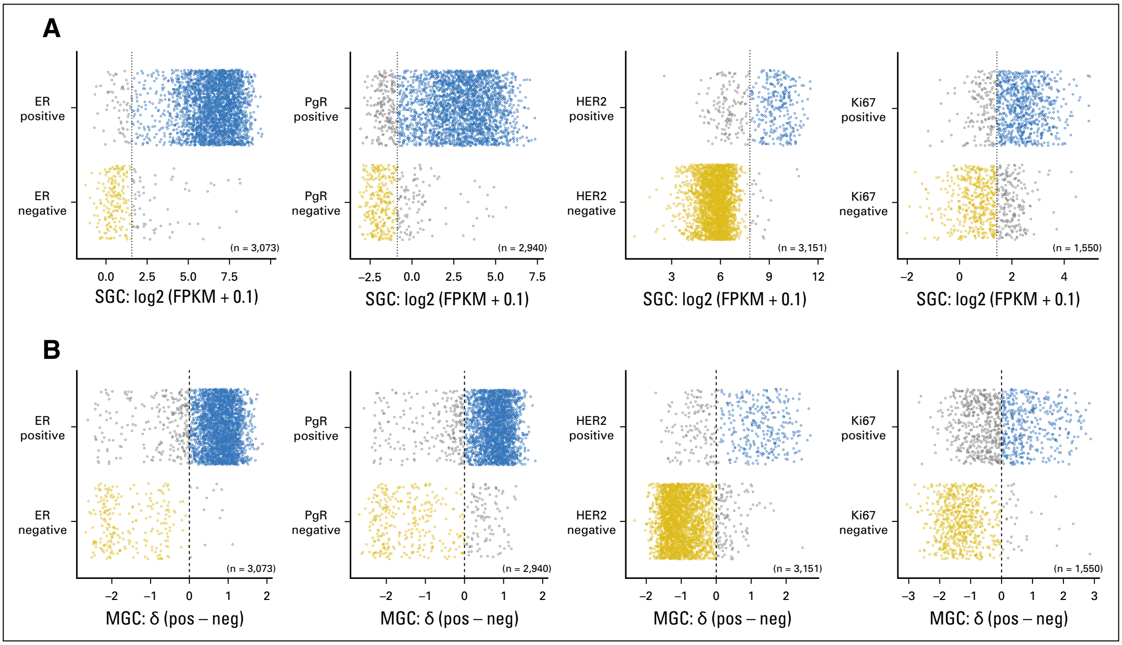
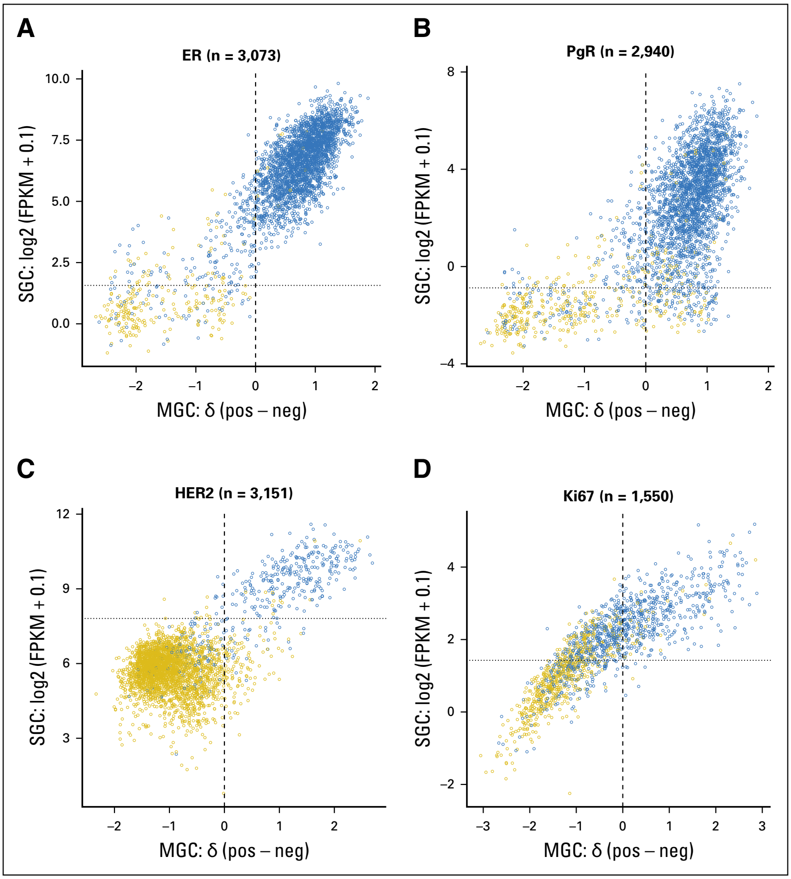


**Fig A2.** Prediction of biomarker status in the 3,273-case independent validation cohort. For estrogen receptor (ER), progesterone receptor (PgR), human epidermal growth factor receptor 2 (HER2), and Ki67 clinical histopathology diagnostic results (y-axis), the single-gene classifier (SGC) gene expression (x-axis) (A) or the transformed multigene classifier (MGC) score (x-axis) (B) is plotted for the validation cohort (circles). Within a biomarker prediction, gold circles were concordantly biomarker negative, blue circles were concordantly positive, and gray circles were discordant by the classifier or histopathology. Vertical dotted (SGC) and dashed (MGC) lines represent the classifier threshold that distinguished the classes. FPKM, fragments per kilobase of transcript per million mapped reads.

**Fig A3.** Transformed multigene classifier (MGC) score (x-axis) versus single-gene classifier (SGC) gene expression (y-axis) in the 3,273 samples of the independent validation cohort (circles) for (A) estrogen receptor (ER), (B) progesterone receptor (PgR), (C) human epidermal growth factor receptor 2 (HER2), and (D) Ki67. Gold circles are negative or low by histopathology, and blue circles are positive or high by histopathology. Vertical dashed lines are drawn at the MGC score threshold of 0 to distinguish the classes, and horizontal dotted lines are drawn at the SGC gene expression thresholds determined from the training cohort. FPKM, fragments per kilobase of transcript per million mapped reads.



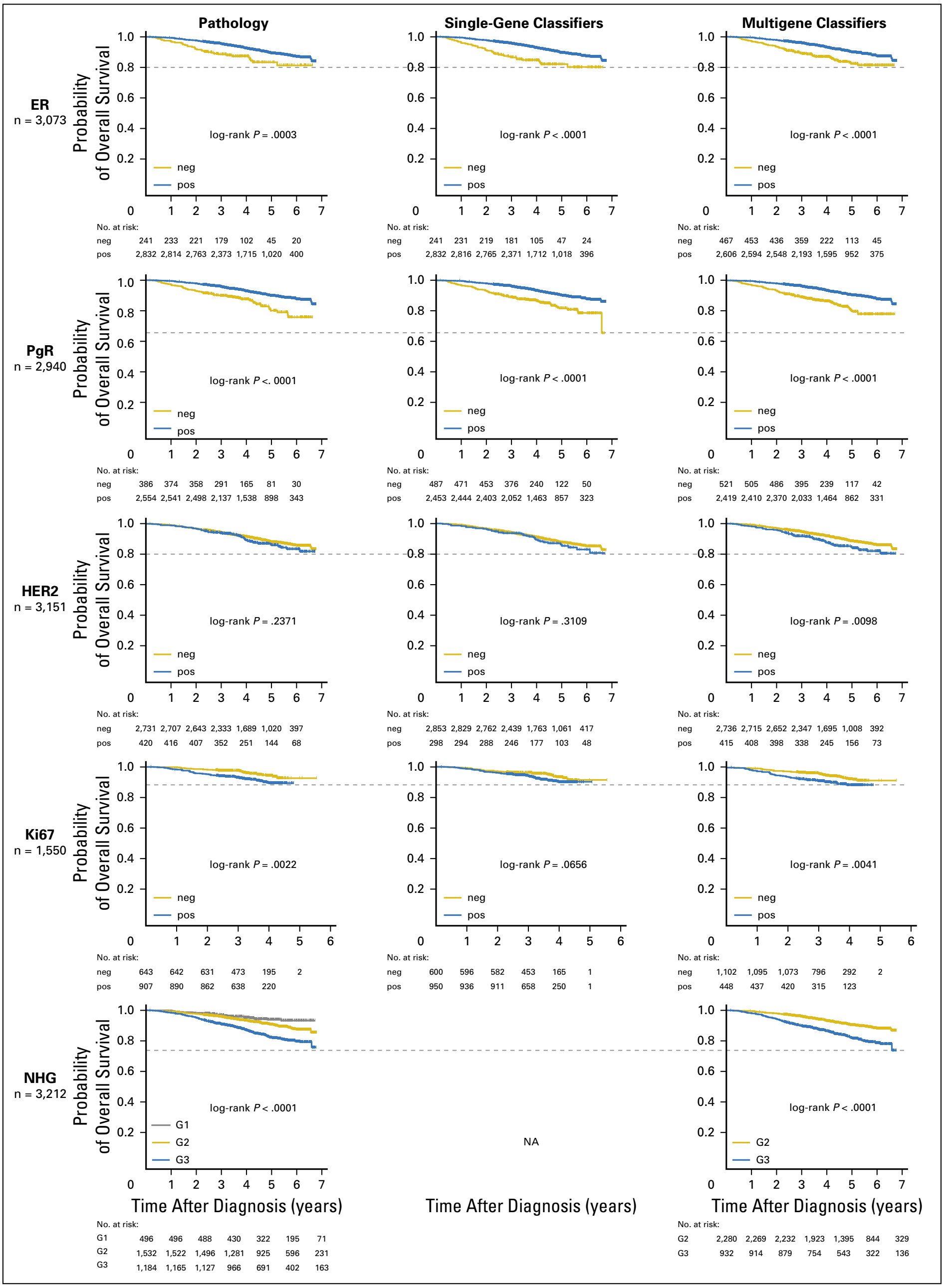
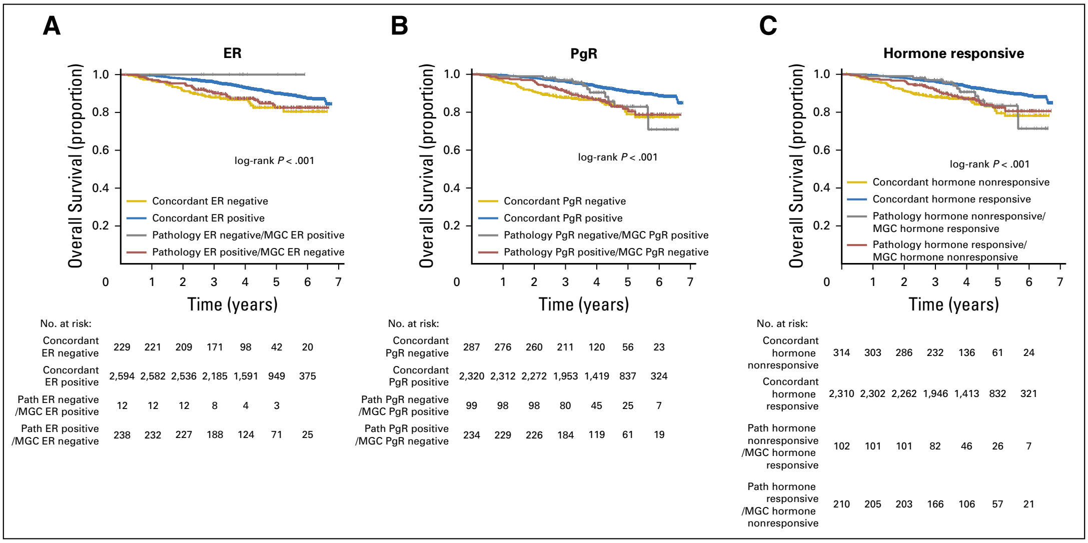


**Fig A4.** Kaplan-Meier overall survival estimates for histopathology, single-gene classifiers (SGCs), and multigene classifiers (MGCs) within the validation cohort (neg, classified as negative; pos, classified as positive; grade [G]1, G2, or G3). The biomarker is indicated at the far left, and the number of tumor cases with complete data across pathology, SGC, and MGC for a given biomarker is shown below each biomarker name. In columns are plotted the Kaplan-Meier survival curves for each classification: (left) pathology, (middle) SGC, and (right column) MGC. The log-rank *P* value is displayed, and horizontal dashed lines are drawn to aid identification of Kaplan-Meier estimates with the poorest outcome classification group within each row. Generally, histopathology and SGCs had similar curves, whereas the MGCs had noticeably improved stratification, for the hormone receptors, in particular.



No. at risk:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Concordant ER negative | 229 | 221 | 209 | 171 | 98 | 42 | 20 |
| Concordant ER positive | 2,594 | 2,582 | 2,536 | 2,185 | 1,591 | 949 | 375 |
| Path ER negative /MGC ER positive | 12 | 12 | 12 | 8 | 4 | 3 | |
| Path ER positive /MGC ER negative | 238 | 232 | 227 | 188 | 124 | 71 | 25 |

No. at risk:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Concordant PgR negative | 287 | 276 | 260 | 211 | 120 | 56 | 23 |
| Concordant PgR positive | 2,320 | 2,312 | 2,272 | 1,953 | 1,419 | 837 | 324 |
| Path PgR negative /MGC PgR positive | 99 | 98 | 98 | 80 | 45 | 25 | 7 |
| Path PgR positive /MGC PgR negative | 234 | 229 | 226 | 184 | 119 | 61 | 19 |

No. at risk:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Concordant hormone nonresponsive | 314 | 303 | 286 | 232 | 136 | 61 | 24 |
| Concordant hormone responsive | 2,310 | 2,302 | 2,262 | 1,946 | 1,413 | 832 | 321 |
| Path hormone nonresponsive /MGC hormone responsive | 102 | 101 | 101 | 82 | 46 | 26 | 7 |
| Path hormone responsive /MGC hormone nonresponsive | 210 | 205 | 203 | 166 | 106 | 57 | 21 |

**Fig A5.** Kaplan-Meier overall survival estimates for groups defined by pathology (path) versus multigene classifiers (MGCs) within the validation cohort; the log-rank *P* value is given. (A) The entire validation cohort stratified by concordance or discordance between estrogen receptor (ER) histopathology and the ER MGC. (B) Progesterone receptor (PgR) status stratified by histopathology and PgR MGC. (C) Hormone responsiveness status stratified by histopathology and MGC; responsive is defined as ER and PgR positive; nonresponsive, as ER negative or PgR negative.

# Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative

Christian Brueffer, MSc, Johan Vallon-Christersson, PhD, Dorthe Grabau, MD PhD, Anna Ehinger, MD, Jari Häkkinen, PhD, Cecilia Hegardt, PhD, Janne Malina, MD, Yilun Chen, MSc, Pär-Ola Bendahl, PhD, Jonas Manjer, MD PhD, Martin Malmberg, MD PhD, Christer Larsson, PhD, Niklas Loman, MD PhD, Lisa Rydén, MD PhD, Åke Borg, PhD, and Lao H. Saal, MD PhD

## SUPPLEMENTARY METHODS

## SUPPLEMENTARY METHODS

### Patients

The study schema is presented in Figure 1. The study was approved by the Regional Ethical Review Board of Lund at Lund University (diary numbers 2007/155, 2009/658, 2009/659, 2010/383, 2012/58, 2013/459) and the Swedish Data Inspection group (364-2010). Trained health professionals provided patient information and patients gave written informed consent. Clinical records were retrieved from the Swedish National Cancer Registry (NKBC). Diagnostic pathology slides, snap-frozen surgical tumor specimens, and formalin-fixed paraffin-embedded (FFPE) tissue blocks were retrieved for 405 patients (training cohort; Supplementary Table 1) diagnosed between 2006 and 2010 and treated at the Skåne University Hospital in Malmö and Lund. The 405 cohort was assembled for classifier training purposes and not for survival analysis, and thus an overrepresentation of HER2+ and ER– cases was selected for.

### Independent validation cohort

For testing of the classifiers and for survival analyses, an independent, prospective, and population-based cohort of 3273 primary breast tumors, diagnosed between September 2010 and March 2015, was assembled from the ongoing SCAN-B study[1] (validation cohort; Supplementary Table 1 and Supplementary Figure 1).

### Histopathology

For the 405 training cases, all biomarkers with the exception of Ki67 had been evaluated at time of diagnosis. The original clinical diagnostic pathology slides and scores were retrieved. For ER and PgR, the diagnostic IHC results were classified into the categories 0%, 1-10%, 11-50%, and >50% positive cells, and the international threshold of ≥1% was used to define positive status. Routine HER2 IHC was evaluated according to the HercepTest criteria using standard local practices, with follow-up HER2 fluorescent *in situ* hybridization (FISH) as needed. Tumor grade was scored according to the Nottingham histological grade (NHG) system, which involves semiquantitative evaluation of three morphological features, tubule formation, nuclear pleomorphism, and mitotic count, using a 3 grade scoring scheme for each feature.[2] For this study, new 4-micron slides were cut from the archival pathology blocks for ER, PgR, HER2, and Ki67 IHC, and for HER2 silver *in situ* hybridization (SISH). For NHG only the diagnostic hematoxylin and eosin slides were used. The new immunostainings were performed by a central laboratory (Helsingborg Hospital) using Ventana instrumentation: ER IHC used antibody SP1 (Ventana); PgR IHC used antibody clone 16 (Leica); HER2 SISH was performed using the INFORM Her2 Dual ISH assay (Ventana); Ki67 IHC used antibody MIB-1 (Dako). Each set of slides for each biomarker, whether the original diagnostic slides or the newly stained slides, were scored in total by three pathologists. The diagnostic slides were scored in the clinical routine, counting as the first reading, and then re-evaluated independently by two pathologists for this study (D.G. and A.E. or J.M.); the new stains were evaluated independently by all three pathologists. ER, PgR, and HER2 were evaluated as described above. Ki67 was evaluated by estimating the percentage of positive nuclei within hotspot regions, with semi-quantitative percentage scores recorded as whole numbers from 0% to 10%, then by bins of 5 from 15% to 100%. The cutoff for Ki67 was determined to be >20% high, ≤20% low, based on the internal Quality Assurance Program cutoff following the procedure recommended by the Swedish guidelines wherein one-third of cases should be high and two-thirds of cases low (see Introduction). A ‘consensus score’ for each biomarker was determined using majority voting from all evaluations.

For validation cohort patients, the diagnostic histopathological records for ER, PgR, HER2, Ki67, and NHG were retrieved from NKBC. For ER and PgR a cutoff for positivity of ≥1% positive cells was applied. Clinical HER2 status, positive or negative, was based on IHC and/or ISH analysis following standard guidelines. Ki67 status was based on percent positive nuclei of tumor cells as recorded in the clinical routine, with Ki67 scores thresholded at >20% being high and ≤20% as low.

## Therapies

All patients were treated uniformly according to common regional guidelines that in turn were based on national and international guidelines during the years 2006 through 2016. For the period 2010 onwards, HER2-positive patients herein generally received HER2-directed treatment with trastuzumab concomitantly with chemotherapy, with a total treatment period of 12 months for trastuzumab. ER-positive cases generally were prescribed endocrine therapy (premenopausal: 5-years tamoxifen; postmenopausal: aromatase inhibitor alone or followed by tamoxifen for a total of 5-years; extended endocrine treatment was introduced for node positive patients during the period). Most patients with ER-negative disease received chemotherapy (taxane/anthracycline), and chemotherapy for ER-positive cases was based on risk for recurrence as estimated by tumor size, nodal status, and NHG.

### Tumor sample processing and RNA-sequencing

Tumor specimens were macrodissected at the pathology departments and processed in our central laboratory with handling standards that meet or exceed the recommendations of the Breast International Group, as described previously[1], with the exception that samples in the training cohort were snap-fresh-frozen instead of being preserved in RNAlater. In brief, nucleic acids were isolated using the AllPrep method and automated using QIAcube machines (Qiagen). Quality control was performed by NanoDrop spectrophotometry and BioAnalyzer (Agilent) analysis; all RNA was highly intact with RNA Integrity Number (RIN) ≥6. Starting from 1 $\mu$g total RNA, sequencing libraries for RNA-seq were generated using customized strand-specific protocols, automated for a high-throughput workflow, which have been previously described in detail.[1,3] Sequencing clusters were generated using the Illumina cBot instrument, and paired-end data were generated using an Illumina HiSeq 2000 or NextSeq 500 instrument. Sequencing statistics are presented in Supplementary Table 2.

### RNA-seq gene expression measurements

Raw sequencing read data was analyzed as previously described.[1,3] To be more consistent with ongoing prospective RNA-seq data being generated within the SCAN-B initiative, for this study we truncated long sequencing reads to 2x50 bp. In brief, raw sequencing data was demultiplexed and filtered using Bowtie 2 against ribosomal, phiX174, and UCSC RepeatMasker sequences. The remaining reads were aligned using TopHat2 2.0.5 (training cohort) or 2.0.12 and 2.0.13 (validation cohort) to the GRCh37/hg19 (with b37 masked chromosome Y and hs37d5 decoy sequences; training cohort) or the GRCh38 (validation cohort) genome together with 80,883 transcript annotations from the UCSC knownGenes table (downloaded September 10, 2012, training cohort) or 104,133 transcript annotations from the UCSC knownGenes table (downloaded September 22, 2014, validation cohort). Cufflinks v2.1.1 (training cohort) or v2.2.1 (validation cohort) was used to calculate expression levels in the form of fragments per kilobase of exon per million mapped reads (FPKM). Isoform-level gene expression data were collapsed on 27,979 (training cohort) or 30,865 (validation cohort) unique gene symbols (sum of FPKM values of each matching transcript).

### Classifiers

Using only the 18,802 genes contained in the NCBI RefSeq NM category (mRNA), the gene expression FPKM values for training samples were transformed by adding a constant 0.1 to each expression value and then applying log2. Single-gene classifiers were built for the four biomarkers that have a single corresponding underlying gene by determining the optimal expression threshold for the genes *ESR1*, *PGR*, *ERBB2*, and *MKI67* that maximizes concordance with the respective histopathological consensus score within the 405 training cohort. Multi-gene classifiers for ER, PgR, HER2, Ki67, and NHG were built by training nearest shrunken centroid[4] models – using *pamr* 1.55 driven by the *caret* 6.0-47 R package – on the gene expression data of the 5000 most varying genes across all 405 samples (Supplementary Table 3), using the histopathological consensus scores for the respective biomarker as labels. In training, support vector machines (SVM) and random forests (RF) were evaluated but provided no improvement (data not shown). Model parameters were determined by performing 10 rounds of 4-fold cross-validation. During cross-validation, the 405 samples were randomly divided into four sub-cohorts, where classifiers trained on the union of three of the sub-cohorts were used to predict the biomarker status in the remaining cohort. This procedure was repeated so that every sub-cohort was predicted by the remaining sub-cohorts exactly once, constituting one round; the result was a mean summary metric (balanced accuracy for biomarkers with dichotomous classes: ER, PgR, HER2 and Ki67; accuracy for NHG) and standard deviation (SD) for each round. The threshold yielding the round with the best mean summary metric within each biomarker (Supplementary Table 6) was used to train a prediction model using all 405 samples. The resulting nine classifiers were used to predict the IHC biomarker status of 3273 independent samples using their gene expression data normalized as described above.

The biological and functional themes of each MGC signature were evaluated using Database for Annotation, Visualization, and Integrated Discovery v6.8 (DAVID).[5] Functional annotation clustering was performed using the Entrez identifiers for each MGC signature (all genes with non-zero weight) and the default DAVID settings and annotation categories, with the following two changes to reduce the number of identified clusters: 'Classification stringency' was increased to High, and the 'Enrichment Thresholds EASE' score was decreased to 0.1.

For representational purposes, MGC classifier scores (the output of the *pamr* discriminant function) were scaled by first calculating the delta score (positive class score – negative class score) for each sample, and scaling the within-class delta scores to have a mean of 1 and -1 for the positive class and the negative class, respectively. The delta score distribution was then shifted so that a delta score < 0 represents a negative classification.

### Statistical analysis

All calculations were performed using R 3.2.3. Matthews correlation coefficients (MCC) were calculated using the generalized method by Gorodkin.[6] Kappa statistics were determined using the *irr* 0.84 and *psy* 1.1 packages. Confidence intervals for kappa and MCC were calculated by bootstrapping using the *boot* 1.3 package and

10,000 bootstrap iterations. Pathology evaluations, multi-gene and single-gene predictions were compared using agreement statistics, MCC and Cohen's kappa, which were interpreted according to Viera and Garrett.[7] *P*-values ≤0.05 were considered significant.

Overall survival (OS) was used as end point for survival analysis and calculated from the date of diagnosis. Kaplan-Meier (KM) analysis and Cox proportional hazards regression were performed using the *survival* 2.38-3 package. Survival times were compared among classes using the logrank test. Multivariate Cox models included the variables age at diagnosis (continuous), lymph node status (positive vs negative), and tumor size (continuous) as covariates, as well as ER, PgR and HER2 status (all positive vs negative), and NHG (G1-G3), as relevant depending on the analyzed model (e.g., the model for the histopathologically HER2-negative group excluded HER2 status). Cases with missing data in any of the included variables were excluded from KM and Cox analysis. All models were checked for proportional hazards using Grambsch and Therneau’s test for non-proportionality and Schoenfeld residuals.[8]

For calculation of concordance statistics, the following definitions are used:

**Balanced Accuracy** $\frac{sensitivity+specificity}{2}$

**Overall Agreement** $\frac{\sum_1^{NClasses}\sum SamplesTruePositiveforClassX}{\sum Samples}$

**Specific Agreement for Class X** (e.g. “positive” or “negative”; “G1”, “G2”, or “G3”)

$$\frac{2 * \sum TruePositivesforClassX}{\sum PositiveReadingsforClassXbyReader1 + \sum PositiveReadingsforClassXbyReader2}$$

**Expected Agreement**

$$\frac{\sum_1^{NClasses}(\sum ReferenceSamplesofClassX * \sum PredictedSamplesofClassX)}{(\sum Samples)^2}$$

**Kappa** $\frac{OverallAgreement+ExpectedAgreement}{1-ExpectedAgreement}$

REFERENCES

**1.** Saal LH, Vallon-Christersson J, Häkkinen J, et al: The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Medicine 7:1-12, 2015

**2.** Elston CW, Ellis IO: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19:403-10, 1991

**3.** Häkkinen J, Nordborg N, Månsson O, et al: Implementation of an Open Source Software solution for Laboratory Information Management and automated RNAseq data analysis in a large-scale Cancer Genomics initiative using BASE with extension package Reggie. bioRxiv:1-23, 2016

**4.** Tibshirani R, Hastie T, Narasimhan B, et al: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America 99:6567-6572, 2002

**5.** Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57, 2009

**6.** Gorodkin J: Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem 28:367-74, 2004

**7.** Viera AJ, Garrett JM: Understanding interobserver agreement: the kappa statistic. Fam Med 37:360-3, 2005

**8.** Grambsch PM, Therneau TM: Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika 81:515-526, 1994

# Study IV

Report

TRANSPARENT PROCESS

OPEN ACCESS

EMBO Molecular Medicine

# The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome

Christian Brueffer[1,2], Sergii Gladchuk[1,2], Christof Winter[1,2,†], Johan Vallon-Christersson[1,2,3], Cecilia Hegardt[1,2,3], Jari Häkkinen[1,2], Anthony M George[1,2], Yilun Chen[1,2], Anna Ehinger[1,2,4], Christer Larsson[2,5], Niklas Loman[1,2,6], Martin Malmberg[6], Lisa Rydén[1,2,7], Åke Borg[1,2,3] & Lao H Saal[1,2,3,*]

## Abstract

**Breast cancer is a disease of genomic alterations, of which the panorama of somatic mutations and how these relate to subtypes and therapy response is incompletely understood. Within SCAN-B (ClinicalTrials.gov: NCT02306096), a prospective study elucidating the transcriptomic profiles for thousands of breast cancers, we developed a RNA-seq pipeline for detection of SNVs/indels and profiled a real-world cohort of 3,217 breast tumors. We describe the mutational landscape of primary breast cancer viewed through the transcriptome of a large population-based cohort and relate it to patient survival. We demonstrate that RNA-seq can be used to call mutations in genes such as *PIK3CA*, *TP53*, and *ERBB2*, as well as the status of molecular pathways and mutational burden, and identify potentially druggable mutations in 86.8% of tumors. To make this rich dataset available for the research community, we developed an open source web application, the SCAN-B MutationExplorer (http://oncogenomics.bmc.lu.se/MutationExplorer). These results add another dimension to the use of RNA-seq as a clinical tool, where both gene expression- and mutation-based biomarkers can be interrogated in real-time within 1 week of tumor sampling.**

**Keywords** breast cancer; mutation; RNA-seq; survival; transcriptome
**Subject Categories** Cancer; Chromatin, Transcription & Genomics
**DOI** 10.15252/emmm.202012118 | Received 2 February 2020 | Revised 8 August 2020 | Accepted 13 August 2020 | Published online 14 September 2020
**EMBO Mol Med (2020) 12: e12118**

## Introduction

Mutations in the cancer genome, including single nucleotide variants (SNVs) and small insertions and deletions (indels), can shed light on cancer biology, tumor evolution and susceptibility or resistance to therapeutic agents (The Cancer Genome Atlas, 2012; Bose *et al*, 2013; Robinson *et al*, 2013). Mutations can now even be used to track circulating tumor DNA in the blood of patients (Garcia-Murillas *et al*, 2015; Förnvik *et al*, 2019). In recent years, the characterization of the mutational landscape of breast cancer has been performed primarily on the DNA level (The Cancer Genome Atlas, 2012; Cheng *et al*, 2015; Ciriello *et al*, 2015). Adoption of massively parallel RNA sequencing (RNA-seq) as a clinical tool has been slower, despite several complementary advantages over DNA-seq. In addition to gene and isoform expression profiling and detection of *de novo* transcripts such as fusion genes, RNA-seq can approximate classical DNA-seq capabilities in the detection of SNVs, indels, as well as structural variants (Ma *et al*, 2018) and coarse copy number (preprint: Talevich & Shain, 2018). This makes RNA-seq an excellent tool for biomarker development (Brueffer *et al*, 2018) and potential clinical deployment (Byron *et al*, 2016; Cieślik & Chinnaiyan, 2018).

For these reasons, among others, in 2010, the Sweden Cancerome Analysis Network–Breast (SCAN-B) initiative (Clinica lTrials.gov ID NCT02306096) selected RNA-seq as the primary analytic tool (Saal *et al*, 2015; Rydén *et al*, 2018). SCAN-B is a prospective real-world and population-based multicenter study with the aim of developing, validating, and clinically implementing novel biomarkers. To this end, SCAN-B collects tumor tissue and blood samples from enrolled patients with a diagnosis of primary breast cancer (BC). To date, over 15,000 patients have been enrolled, and messenger RNA (mRNA) sequencing is performed on patient tumors within 1 week of surgery. All patients are treated uniformly according to the Swedish national standard of care regimen.

Expression profiling is an excellent tool to develop gene signatures for established and novel biomarkers (Sotiriou *et al*, 2006; Roepman *et al*, 2009; Brueffer *et al*, 2018), and many such signatures can be applied to a single RNA-seq dataset. However, for the detection of SNVs and indels from RNA-seq data, there are several challenges.

1 Division of Oncology, Department of Clinical Sciences, Lund University, Lund, Sweden
2 Lund University Cancer Center, Lund, Sweden
3 CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden
4 Department of Pathology, Skåne University Hospital, Lund, Sweden
5 Division of Molecular Pathology, Department of Laboratory Medicine, Lund University, Lund, Sweden
6 Department of Oncology, Skåne University Hospital, Lund, Sweden
7 Department of Surgery, Skåne University Hospital, Lund, Sweden
*Corresponding author. Tel: +46 46 2220365: E-mail: lao.saal@med.lu.se
†Present address: Institut für Klinische Chemie und Pathobiochemie, Klinikum rechts der Isar, Technische Universität München, München, Germany

Unlike DNA-seq, where whole-genome or targeted sequencing reads are distributed approximately uniformly and in proportion to DNA copy number, the abundance of reads in RNA-seq is proportional to the expression of each gene or locus. Consequently, only variants in expressed transcripts of sufficient level can be detected. In cancer, this means that variants in oncogenes can likely be detected, whereas those in tumor suppressor genes, e.g., *TP53*, *BRCA1*, or *BRCA2*, are more likely to be missed. For example, mutations inducing premature stop codons can lead to nonsense-mediated decay, causing loss of expression and subsequently false-negative calls. The transcriptome is also more complex and challenging than the genome. RNA structures, such as alternative splicing, add computational challenges to alignment, and RNA editing can contribute to false-positive variant calls. Finally, there is the lack of benchmark datasets for RNA-seq, as are available for DNA from the Genome in a Bottle consortium and others (Zook *et al*, 2016; Li *et al*, 2018).

The aim of this study was to optimize RNA-seq somatic mutation calling through comparison to matched targeted DNA-seq, discern the mutational landscape of the early breast cancer transcriptome across a large cohort of 3,217 treatment-naïve SCAN-B cases with sufficient follow-up time, and to make the resulting vast dataset available for exploration by the wider research community. To demonstrate the power of the methodology and dataset, we assessed the impact of mutations in important breast cancer driver genes and pathways, as well as tumor mutational burden (TMB) on patient overall survival (OS).

## Results

An outline of the study design, which comprised DNA sequencing and RNA sequencing of 275 samples from the ABiM cohort, and RNA sequencing of 3,217 samples from the SCAN-B cohort, is shown in Fig 1.

### Variant filter performance

Mutation calling in the 275 sample ABiM cohort resulted in 3,478 somatic post-filter mutations from the matched tumor/normal targeted capture DNA, and 1,459 variants from tumor RNA-seq in the DNA capture regions (Table 1 and Fig EV1A). Comparing these DNA and RNA variants resulted in 1,132 mutations that were present both in DNA and RNA in the capture regions and whose frequencies were generally in line with previous studies such as The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas, 2012) (Fig EV1B). Of the 1,459 RNA-seq variants, 884 (60.6%) were identified as somatic in DNA, 248 (17.0%) as germline in DNA, and 327 (22.4%) as unique to RNA. These RNA-unique variants are a mix of somatic mutations missed in DNA-seq, e.g., due to regional higher sequencing coverage in RNA-seq or tumor heterogeneity, unfiltered RNA editing sites, or artifacts caused by PCR, sequencing, or alignment and variant calling.

### Landscape of somatic mutations in the breast cancer transcriptome

We applied the filters derived from the 275 sample set to the entire RNA-seq SCAN-B 3,217 sample set, resulting in 144,593 total variants comprised of 141,095 SNVs, 1,112 insertions, and 2,386 deletions (Table 1). The number of mutations per sample in the SCAN-B set was lower compared to the ABiM set, likely due to the ABiM set being sequenced to a higher depth (Table EV1). The SNVs comprised 50,270 missense, 2,311 nonsense, 1,042 splicing, 68,819 affecting 3′/5′ untranslated regions (UTRs), 17,057 synonymous mutations, as well as 1,596 mutations predicted otherwise. The majority of indels were predicted to cause frameshifts or affect 3′/5′ UTRs (Table EV2). After removing synonymous mutations, the number of mutations was reduced to 127,536 variants in the SCAN-B set, i.e., an average of 40 mutations per tumor.

We analyzed the contribution of the six nucleotide substitution types (C>A, C>G, C>T, T>A, T>C, and T>G) to SNVs in the ABiM and SCAN-B sets (Fig 2A). Compared to DNA, RNA-seq-based variant calls showed a relative under-representation of C>T substitutions and an over-representation of T>C substitutions.

In accordance with published studies of primary BC, the most frequently mutated genes were the known BC drivers *PIK3CA* (34% of samples), *TP53* (23%), *MAP3K1* (7%), *CDH1* (7%), *GATA3* (7%), and *AKT1* (5%) (Fig 3). As reported before (Ciriello *et al*, 2015), disruptive alterations in *CDH1* were a hallmark of lobular carcinomas (135/386 [35.0%] of samples), while alterations in *TP53*, *MAP3K1*, and *GATA3* were more common in the ductal type. 86.8% of SCAN-B samples had at least one mutation in a gene targeted by an approved or experimental drug, based on the Database of Gene-drug Interactions (DGI).

### Somatic mutations in important BC genes

We examined known driver BC genes more closely and found our RNA-seq-based mutation calls to recapitulate known mutation rates and hot spots, summarized in Table 2, Table EV2, and Fig 2C–F. Associations of mutated genes and clinical and molecular biomarkers are summarized in Table EV3, and several examples are highlighted below.

*PIK3CA* was the most frequently mutated gene, with 1,163 nonsynonymous mutations in 1,095 patient samples (34% of patients). As expected, and in line with previous studies (Saal *et al*, 2005; The Cancer Genome Atlas, 2012; Pereira *et al*, 2016), the majority of alterations were the known hot spot mutations H1047R/L, E545K, and E542K (Table 2, Fig 2D), which lead to constitutive signaling (Bader *et al*, 2006). All hot spot mutations and the vast majority of other *PIK3CA* alterations were missense mutations. Mutations were associated with lobular, ER$^+$, PgR$^+$, HER2$^-$, and Luminal A (LumA) BC (Table EV3).

*TP53* is frequently disrupted by somatic SNVs; however, a few hot spot mutations exist (Giacomelli *et al*, 2018). The mutation frequency in BC is estimated to be 35.4-37% (The Cancer Genome Atlas, 2012; Pereira *et al*, 2016), which we could confirm in our DNA-seq ABiM filter-definition cohort (37%). Likely due to nonsense-mediated decay (NMD), loss of heterozygosity, and/or decreased mRNA transcription, in the 3,217 cases, the frequency of *TP53* mutations was lower at 23% (782 mutations in 733 samples). Despite underdetection by RNA-seq, the identified hot spot residues were the same as reported in the IARC TP53 database (release R20) (Bouaoun *et al*, 2016). The most often mutated amino acids we observed were R273, R248, R175 (50, 49, and 24 mutations respectively, total 123/782 [15.7%]), followed by positions Y220 (21/782

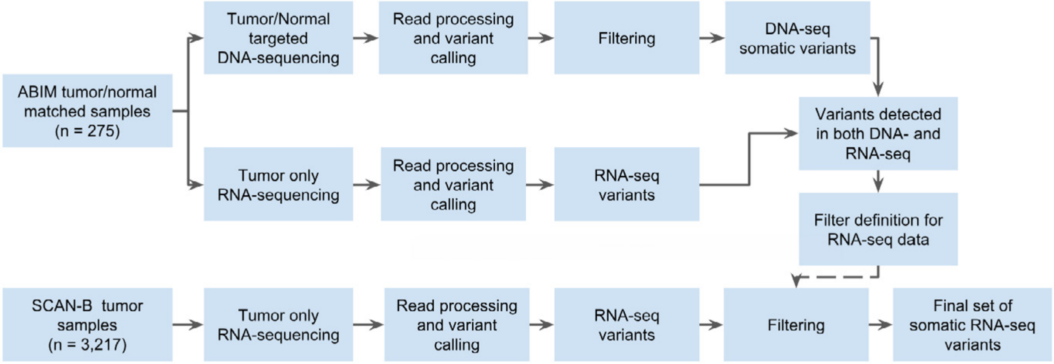

ABIM tumor/normal matched samples (n = 275)
Tumor/Normal targeted DNA-sequencing
Read processing and variant calling
Filtering
DNA-seq somatic variants
Tumor only RNA-sequencing
Read processing and variant calling
RNA-seq variants
Variants detected in both DNA- and RNA-seq
Filter definition for RNA-seq data
SCAN-B tumor samples (n = 3,217)
Tumor only RNA-sequencing
Read processing and variant calling
RNA-seq variants
Filtering
Final set of somatic RNA-seq variants

**Figure 1. Study design.**
Study design flow diagram for DNA-seq-informed optimization of RNA-seq variant calling.

**Table 1. Number of mutations in the ABiM (DNA-seq and RNA-seq) and SCAN-B (RNA-seq) cohorts.**

| Cohort | Source | Coverage | Total mutations | SNVs | Insertions | Deletions | Samples with mutations | Mutations per sample |
|---|---|---|---|---|---|---|---|---|
| ABiM | DNA | Capture regions | 3,478 | 3,173 | 50 | 173 | 274 | 12.7 |
| ABiM | RNA | Capture regions | 1,459 | 1,304 | 57 | 98 | 265 | 5.5 |
| ABiM | RNA | Whole mRNA | 16,683 | 15,764 | 235 | 684 | 275 | 60.7 |
| SCAN-B | RNA | Whole mRNA | 144,593 | 141,095 | 1,112 | 2,386 | 3,217 | 44.9 |

Sample numbers differ from total cohort sizes due to filtering resulting in samples with no remaining post-filter mutations.

[2.7%]), R280 (19/782 [2.4%]), and R342 (17/782 [2.2%]) (Table 2, Fig 2C). Most detected mutations are in the DNA binding domain, and 77.6% of overall mutations are missense mutations, likely leading to protein loss of function (LoF). As anticipated, *TP53* mutations were associated with ductal, ER$^-$, PgR$^-$, HER2$^+$, hormone receptor positive (HoR$^+$)/HER2$^+$ (HoR$^+$ defined as ER$^+$ and PgR$^+$, HoR$^-$ otherwise), HoR$^-$/HER2$^+$, triple-negative BC (TNBC), and the basal-like and HER2-enriched PAM50 subtypes (Table EV3), as reported before (The Cancer Genome Atlas, 2012).

*PTEN* is a crucial tumor suppressor gene and regulator of PI3K activity, and PTEN protein expression is associated with poor outcome (Saal *et al*, 2007). In our dataset, we found 124 non-synonymous mutations in 116/3,217 (3.6%) samples, including hot spot mutations in H303 and H266 of unknown significance (Fig 2E). Mutations were significantly associated with HER2$^-$ disease (Table EV3).

*ERBB2* (HER2) mutations have emerged as a novel biomarker and occur by the majority in patients without *ERBB2* amplification (Bose *et al*, 2013), but also in *ERBB2*-amplified cases (Cocco *et al*, 2018). Evidence is mounting that recurrent *ERBB2* mutations lead to increased activation of the HER2 receptor in tumors classified as HER2 normal (Bose *et al*, 2013; Wen *et al*, 2015; Pahuja *et al*, 2018). Activating *ERBB2* mutations have been shown to confer therapy resistance against standard of care drugs such as trastuzumab and lapatinib (Cocco *et al*, 2018), but can be overcome using pan-HER tyrosine kinase inhibitors (TKIs) such as neratinib (Bose *et al*, 2013; Ben-Baruch *et al*, 2015; Ma *et al*, 2017; Cocco *et al*, 2018). *ERBB2* mutations have also been shown to confer resistance to endocrine therapy in the metastatic setting (Nayar *et al*, 2018), where HER2-directed drugs are effective (Murray *et al*, 2018). We identified 117 non-synonymous *ERBB2* mutations in 103 patients (3.2%), higher than the previously reported incidence rates of 1.6%-2.4% (Bose *et al*, 2013; Wen *et al*, 2015; Ross *et al*, 2016), but lower than in metastatic BC where rates as high as ~7% have been reported (Cocco *et al*, 2018). Two hot spots, L755S (28/117) and V777L (24/117) that cause constitutive HER2 signaling (Fig 2F) (Bose *et al*, 2013; Wen *et al*, 2015), accounted for 44.4% of total *ERBB2* mutations. Co-occurrence of *ERBB2* mutation and amplification has been reported before, however mainly in the metastatic setting (Cocco *et al*, 2018). In our untreated, early BC cohort, we observed *ERBB2* mutation and amplification in 12 tumors, demonstrating that co-incident *ERBB2* mutation and amplification is rare but can occur in early, treatment-naïve BC. Mutation and amplification were not mutually exclusive ($P$ = 0.88), and interestingly *ERBB2* mutations occurred predominantly in tumors classified as PAM50 HER2-enriched subtype ($P$ = 0.0001). Moreover, *ERBB2* mutation was significantly associated with PgR$^-$ and lobular BC (Table EV3).

Loss of E-cadherin (CDH1) protein expression is a hallmark of the lobular BC phenotype (Ciriello *et al*, 2015). With 12% of our cohort being of lobular type, we observed 137 of total 233 *CDH1* mutations in lobular BCs (58.8%, $P$ = 1.6E-72). The mutations were mostly comprised of nonsense mutations (37.2%) and frameshift indels (35.4%), suggesting they contribute to CDH1

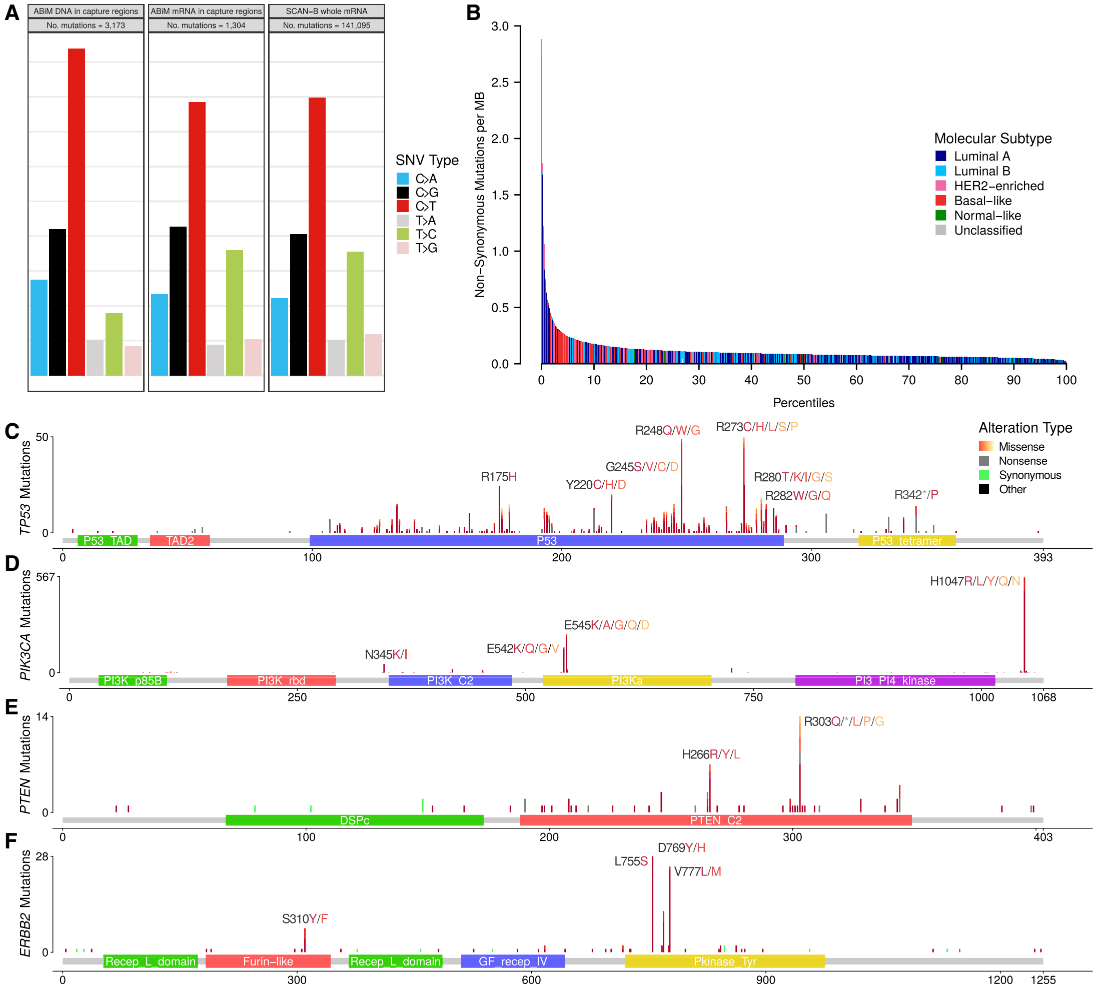


**Figure 2. Overview of non-synonymous mutations in terms of base substitution signatures, molecular subtype, and protein impact.**

A Contribution of base change types to the overall SNV composition in the ABiM cohort for captured DNA regions and mRNA in the captured DNA regions, as well as SCAN-B whole mRNA.

B Number of non-synonymous mutations per sample. Bars are colored by PAM50 subtypes Luminal A (dark blue), Luminal B (light blue), HER2-enriched (pink), basal-like (red), Normal-like (green) and Unclassified (gray).

C–F Lollipop plots showing the location, abundance, and impact of SNVs in (C) *TP53*, (D) *PIK3CA*, (E) *PTEN*, and (F) *ERBB2* on the respective encoded protein. Protein change labels are shown for the most mutated amino acid positions, with residues ordered left to right by mutation frequency within each label.

expression loss and drive the lobular phenotype. We observed one nonsense mutation hot spot (Q23*, $n = 18$), and this residue was also hit by a rare missense mutation (Q23K, $n = 1$). In addition to lobular BC, *CDH1* mutations were associated with ER$^+$, HER2$^-$, and HoR$^+$/ HER2$^-$ status, and the LumA subtype (Table EV3).

Other notable mutated genes in our set were *MAP3K1*, *AKT1*, *ESR1*, *GATA3*, *FOXA1*, *SF3B1*, and *CBFB*. *MAP3K1* is a regulator of signaling pathways and regularly implicated in various cancer types. Loss of *MAP3K1* expression activates the PI3K/AKT/mTOR pathway and desensitizes the tumor to PI3K inhibition (Avivar-Valderas *et al*, 2018), thus mutation status of this gene may affect efficacy of PI3K-targeting drugs. We observed a high rate of frameshift indels, and missense mutations mostly clustered in the kinase domain. Co-mutation of *MAP3K1* and *PIK3CA* occurred in 108 tumors (3.4%), and inactivating (frameshift/nonsense) *MAP3K1* alterations occurred in 77 of 1,095 (7%) of *PIK3CA*-mutant tumors. *AKT1* is a

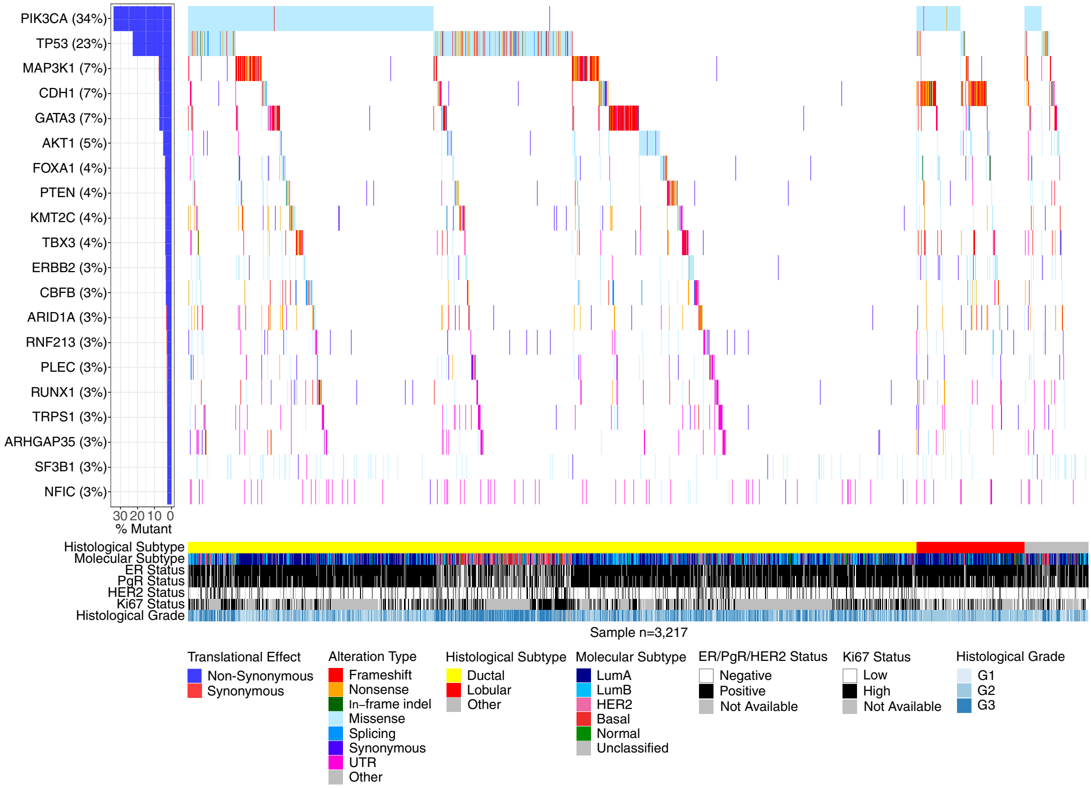


**Figure 3. Overview of frequently mutated genes across 3,217 SCAN-B samples.**

Waterfall plot of the 20 most frequently mutated genes (rows) across 3,217 SCAN-B samples (columns). Genes are ranked from top to bottom by mutation frequency. Samples are sorted by histological subtype and alteration occurrence. Mutations are colored by predicted functional impact.

common oncogene with 156 (4.8%) mutated samples and featured the fourth most mutated hot spot (E17K, 121 mutations) in the SCAN-B cohort. These mutations are predictive of sensitivity to AKT inhibitors (Hyman *et al*, 2017). *ESR1* encodes the estrogen receptor (ER) alpha, perhaps the most important clinical BC biomarker. Seventy-seven tumors harbored 81 *ESR1* variants, including known endocrine treatment resistance mutations, that are discussed elsewhere in detail (M. Dahlgren, AM. George, C. Brueffer, S. Gladchuk, Y. Chen, J. Vallon-Christersson, C. Hegardt, J. Häkkinen, L. Rydén, M. Malmberg, C. Larsson, SK. Gruvberger-Saal, A. Ehinger, N. Loman, Å. Borg, LH. Saal, submitted). Relatedly, *GATA3* and *FOXA1* are frequently mutated transcription factors that are directly involved in modulating ER signaling, and their expression is independently associated with beneficial survival in ER$^+$ tumors (Hisamatsu *et al*, 2012). We identified 246 *GATA3* mutations, including known recurrent frameshift mutations (P409fs, $n$ = 30 and D336fs, $n$ = 10) and the M294K/R missense mutation ($n$ = 15), as well as 10 splice site variants. In *FOXA1*, we detected 146 total mutations, including known recurrent S250F ($n$ = 23) and F266L/C ($n$ = 12) missense mutations. Most mutations occurred in the forkhead DNA binding domain. While the role of mutations in these genes has not been thoroughly characterized, Takaku *et al* (2018) suggest that *GATA3* can function as either oncogene or tumor suppressor depending on the mutations the gene accumulated, and which part of the protein product is impacted. According to their classification, the most frequent mutation in our cohort, the P409fs frameshift mutation, results in an elongated protein product compared to *GATA3*-wt that has favorable survival compared to mutations of the second Zinc finger domain. In line with their involvement in ER signaling, mutations in *GATA3*, *FOXA1*, *MAP3K1*, and *ESR1* were associated with ER$^+$ and PgR$^+$ disease. Further, *GATA3, MAP3K1,* and *ESR1* were associated with HoR$^+$/ HER2$^-$, and *GATA3* and *MAP3K1* with ductal BC, while *ESR1* and *FOXA1* were more common in lobular BC. All these genes were associated with the LumA subtype, with the exception of *GATA3* which was associated to Luminal B (LumB) (Table EV3).

*SF3B1* encodes a subunit of the spliceosome and mutations in this gene have been identified as potentially interesting treatment

Table 2. The most occurring non-synonymous mutations in the genes *PIK3CA*, *AKT1*, *SF3B1*, *GATA3*, *ERBB2*, *TP53*, *FOXA1*, and *CDH1* in 3,217 SCAN-B samples.

| Gene | AA change | Number of mutations | Mut. samples (%) | Mut. in gene (%) |
|---|---|---|---|---|
| *PIK3CA* | H1047R | 483 | 15 | 41.5 |
| | E545K | 212 | 6.6 | 18.2 |
| | E542K | 142 | 4.4 | 12.2 |
| | H1047L | 77 | 2.4 | 6.6 |
| | N345K | 49 | 1.5 | 4.2 |
| | E726K | 26 | 0.8 | 2.2 |
| | C420R | 20 | 0.6 | 1.7 |
| | E453K | 13 | 0.4 | 1.1 |
| | G1049R | 11 | 0.3 | 0.9 |
| | E545A | 10 | 0.3 | 0.9 |
| | Q546K | 10 | 0.3 | 0.9 |
| | M1043I | 8 | 0.2 | 0.7 |
| | Other | 102 | 3.2 | 8.8 |
| *AKT1* | E17K | 121 | 3.8 | 76.1 |
| | Other | 38 | 1.2 | 23.9 |
| *SF3B1* | K700E | 60 | 1.9 | 74.1 |
| | Other | 21 | 0.7 | 25.9 |
| *GATA3* | P409fs | 30 | 0.9 | 12.2 |
| | M294K | 14 | 0.4 | 5.7 |
| | D336fs | 10 | 0.3 | 4.1 |
| | D332fs | 10 | 0.3 | 4.1 |
| | Other | 182 | 5.7 | 74 |
| *ERBB2* | L755S | 28 | 0.9 | 23.9 |
| | V777L | 24 | 0.7 | 20.5 |
| | D769Y | 9 | 0.3 | 7.7 |
| | Other | 56 | 1.7 | 47.9 |
| *TP53* | R273C | 25 | 0.8 | 3.2 |
| | R248Q | 25 | 0.8 | 3.2 |
| | R175H | 24 | 0.7 | 3.5 |
| | R248W | 22 | 0.7 | 3.1 |
| | R273H | 19 | 0.6 | 2.4 |
| | Y220C | 17 | 0.5 | 2.2 |
| | F134L | 14 | 0.4 | 1.8 |
| | E285K | 13 | 0.4 | 1.7 |
| | R213* | 12 | 0.4 | 1.5 |
| | R282W | 12 | 0.4 | 1.5 |
| | R306* | 10 | 0.3 | 1.3 |
| | Y163C | 10 | 0.3 | 1.3 |
| | L194R | 9 | 0.3 | 1.2 |
| | R342* | 9 | 0.3 | 1.2 |
| | E286K | 8 | 0.2 | 1 |
| | G245S | 8 | 0.2 | 1 |
| | H179R | 8 | 0.2 | 1 |
| | Q331* | 8 | 0.2 | 1 |
| | Other | 529 | 16.4 | 65.1 |
| *FOXA1* | S250F | 23 | 0.7 | 15.8 |
| | F266L | 11 | 0.3 | 7.5 |
| | Other | 112 | 3.5 | 76.7 |
| *CDH1* | Q23* | 18 | 0.6 | 7.7 |
| | I650fs | 8 | 0.2 | 3.4 |
| | P127fs | 8 | 0.2 | 3.4 |
| | Other | 199 | 6.2 | 85.4 |

Shown are the total number of mutations, the frequency of the mutations in the SCAN-B cohort (Mut. samples), and the frequency of a particular mutation within all mutations in the gene (Mut. in gene).

targets after having been observed in myelodysplastic syndromes and chronic lymphocytic leukemia. We identified 81 *SF3B1* mutations in 79 tumors, 60 of which were K700E hot spot mutations that deregulate splicing and result in differential splicing patterns in BC (Maguire *et al*, 2015). Alterations in this gene are associated with $ER^+$ disease (Maguire *et al*, 2015) and affect alternative splicing patterns (Alsafadi *et al*, 2016). The cohort frequency of 1.9% K700E mutations matches up with previously reported 1.8% in an unselected breast cancer cohort (Maguire *et al*, 2015). We could not confirm the reported prevalence of *SF3B1* mutations in $ER^+$ tumors in the total $ER^+$ group ($P = 0.052$), but in the $ER^+$/ $HER2^-$ subgroup (68/79 mutated tumors $ER^+$/ $HER2^-$, $P = 0.021$), as well as the association with non-ductal, and non-lobular subtypes ($P = 0.0033$). Additionally, *SF3B1* mutations were associated with LumB tumors ($P = 0.0006$) (Table EV3).

CBFB is a transcriptional co-activator of RUNX2, an expression regulator of several genes involved in metastatic processes such as cell migration. Increased CBFB expression has been identified as essential for cell invasion in BC (Mendoza-Villanueva *et al*, 2010). Recurrent *CBFB* mutations have recently been reported in $ER^+$/ $HER2^-$ disease; however, the significance of these mutations is unknown (Griffith *et al*, 2018). We could confirm this finding showing 107 mutations (3.3% cohort frequency), 95 of which were in $ER^+$/ $HER2^-$ samples (4% of $ER^+$/ $HER2^-$ samples, $P = 0.0005$). We also found them to be associated with the LumA subtype (Table EV3); however, we did not observe the splice site mutation described by Griffith *et al* (2018), perhaps due to degradation of the spliced mRNA by NMD.

### Mutations in molecular pathways

We were interested whether the mutational data, when considered from the perspective of mutated pathways, could reveal new biological correlates. To test this, we mapped mutation status to important BC pathways as defined in the Reactome database (Fabregat *et al*, 2018; Jassal *et al*, 2020). We called a pathway mutated when at least one of the member genes had a non-synonymous mutation and clustered samples by pathway mutation status using Euclidean distance and Ward linkage. Notable clusters that emerged were co-mutated hedgehog signaling, p53-independent DNA repair, and hypoxia response pathways, as well as a cluster of NOTCH1/2/3 signaling mutated tumors, both in mostly basal-like and HER2-enriched tumors. Both clusters are linked in their relation to cancer stem cell development (Habib & O'Shaughnessy, 2016; Locatelli & Curigliano, 2017), which, in addition to the NOTCH and Hedgehog pathways themselves, has emerged as a novel treatment target, particularly in TNBC. Another co-mutation cluster was made up of PI3K/AKT, MET, RET, EGFR, ERBB2, and ERBB4 signaling pathways that occurred in a subset of Luminal A and B tumors (Fig 4; see Table EV4 for Reactome pathway IDs). Activation of these pathways is involved in the development of ER$^+$ BC through proliferation-inducing signaling, or endocrine therapy resistance, e.g., via activating *ERBB2* mutations (Nayar *et al*, 2018).

### Tumor mutational burden

Tumor mutational burden is increasingly of interest due to its association to neoantigen burden and response to immunotherapies. We used the median number of non-synonymous mutations per transcriptome megabase (rnaMB), 0.082 mutations/rnaMB, to stratify all SCAN-B samples into TMB-high and TMB-low groups. Samples with HER2-enriched and basal-like PAM50 subtypes were enriched in the top 10% of samples with the highest TMB compared to the lowest 90% ($P = 2.2\text{E-}16$, Fig 2B), supporting previous results and indicating that immunotherapy may have higher activity in these two PAM50 subtypes (The Cancer Genome Atlas, 2012).

### Mutational landscape and patient outcomes

Next, we were interested in the association between mutations in important BC genes and patient outcome under various treatments. Below we show the results for *TP53*, *PIK3CA*, *ERBB2*, and *PTEN* with OS of SCAN-B patients in clinical biomarker and treatment groups (Figs 5 and EV2), as well as selected pathways (Figs 6 and EV3). Specific treatments stratified by clinical biomarker and treatment groups are detailed in Table EV4. The web tool SCAN-B MutationExplorer may be used to query any gene(s) and pathway(s) of interest.

In line with expectations, *TP53* mutation predicted poor survival in untreated patients (hazard ratio [HR] 2.39, 95% CI [1.5–3.79], $P = 0.00014$), patients treated with endocrine- and chemotherapy (HR: 1.83 [1.09–3.05], $P = 0.02$), as well as the HoR$^+$/HER2$^-$ biomarker subgroup (HR: 1.43 [1.06–1.94], $P = 0.019$). After adjusting for important covariates in multivariable (MV) Cox analyses, *TP53* mutations remained a significant stratifier among patients receiving endocrine- and chemotherapy.

In early-stage breast cancer, *PIK3CA* mutations have been associated with slightly better 5-year OS than *PIK3CA*-wt tumors in univariable analysis, but not when correcting for clinicopathological and treatment variables (Zardavas *et al*, 2018). In our hands, we saw a similar univariable effect in patients who did not receive systemic treatment (HR: 0.54 [0.32–0.91], $P = 0.018$), but not when adjusting for covariates. Additionally, *PIK3CA* mutations in HER2 ± any treated patients became significant in multivariable analysis.

*ERBB2* mutations were indicators of poor prognosis in endocrine therapy only (HR: 1.85 [1.08–3.18], $P = 0.023$) and endocrine- and chemotherapy-treated (HR: 3.49 [1.4–8.72], $P = 0.0042$) patients, as well as in the HoR$^+$/HER2$^-$ subgroup (HR: 1.96 [1.14–3.35], $P = 0.013$). After multivariable adjustment, they remained a significant predictor in the endocrine-only-treated patient subgroup.

*PTEN* mutations alone were associated with poor survival in the patient group not receiving systemic treatment (HR: 2.56 [1.03–6.33], $P = 0.036$), but not in any of the other treatment or clinical biomarker groups (Fig 5 and EV2). While loss of PTEN protein expression or non-functional PTEN protein can be caused by SNVs and indels, it can also be caused by other mechanisms such as large structural variants (Saal *et al*, 2008) and promoter methylation (Zhang *et al*, 2013) that have not been investigated in this study. To account for this, we defined a new subgroup *PTEN*-MutExp, where a status of "low" identifies cases with either *PTEN* mutation or gene expression in the lower quartile within the cohort, and "normal" otherwise. The *PTEN*-MutExp low group, incorporating gene expression, showed improved stratification in the no systemic treatment group (HR: 1.88 [1.2–2.95], $P = 0.0053$), and significantly lower OS in patients receiving only endocrine treatment (HR: 1.63 [1.26–2.12], $P = 0.00021$), as well as HoR$^+$/HER2$^-$ patients (HR: 1.54 [1.2–1.99], $P = 0.00076$). Most of the prognostic value is provided by the gene expression, however mutation data improved stratification (Fig EV4). After multivariable adjustment, *PTEN* mutations in the no systemic-treated subgroup, as well as the *PTEN*-MutExp "low" group in HoR$^+$/HER2$^-$ and HoR$^+$/HER2$^+$ patients, remained significant.

Abstracting from mutations in individual genes, we investigated the effect of mutated pathways on OS in patient subgroups stratified by treatment (Fig 6) and clinical subgroup (Fig EV3). Mutated WNT (Fig 6A, HR: 2.14 [1.18–3.89], $P = 0.01$), Hedgehog (Fig 6B, HR: 1.68 [1.06–2.68], $P = 0.026$), and NOTCH2 (Fig 6C, HR: 2.31 [1.27–4.2], $P = 0.0047$) pathways, as well as the p53-independent DNA damage repair pathway (Fig 6D, HR: 2.03 [1.3–3.17], $P = 0.0015$) were associated with worse survival in patients not receiving systemic treatment. Additionally, NOTCH2 signaling (Fig 6C, HR: 1.65 [1.19–2.3], $P = 0.0026$) was associated with worse OS in patients receiving only endocrine treatment, and TGF$\beta$ signaling (Fig 6E, HR: 1.79 [1.08–2.96], $P = 0.021$) with worse OS in patients treated with endocrine- and chemotherapy. Further, WNT signaling was associated with worse OS in HoR$^+$/HER2$^+$ (HR: 2.57 [1.04–6.33], $P = 0.034$) and TNBC patients (HR: 2.5 [1.27–4.91], $P = 0.0061$; Fig EV3). In multivariable analysis, WNT pathway mutations in HoR$^+$/HER2$^+$ and TNBC patients, NOTCH2 pathway mutations in endocrine-only-treated patients, and TGF$\beta$ pathway mutations in endocrine + chemo ± any treated patients remained significant stratifiers.

Given its importance as an emerging biomarker for response to immune checkpoint therapy (Goodman *et al*, 2017), we investigated whether TMB could also provide response information with respect to conventional treatment regimens (Fig 7). When stratified into TMB-high and TMB-low by the SCAN-B cohort median TMB per rnaMB, low TMB was favorable to OS independent of treatment across the cohort (HR, 1.54 [1.28–1.86], $P = 0.0000033$), as well as in patients not systemically treated (HR: 2.53, [1.58–4.05], $P = 0.000066$), treated with endocrine therapy only (HR: 1.55 [1.22–

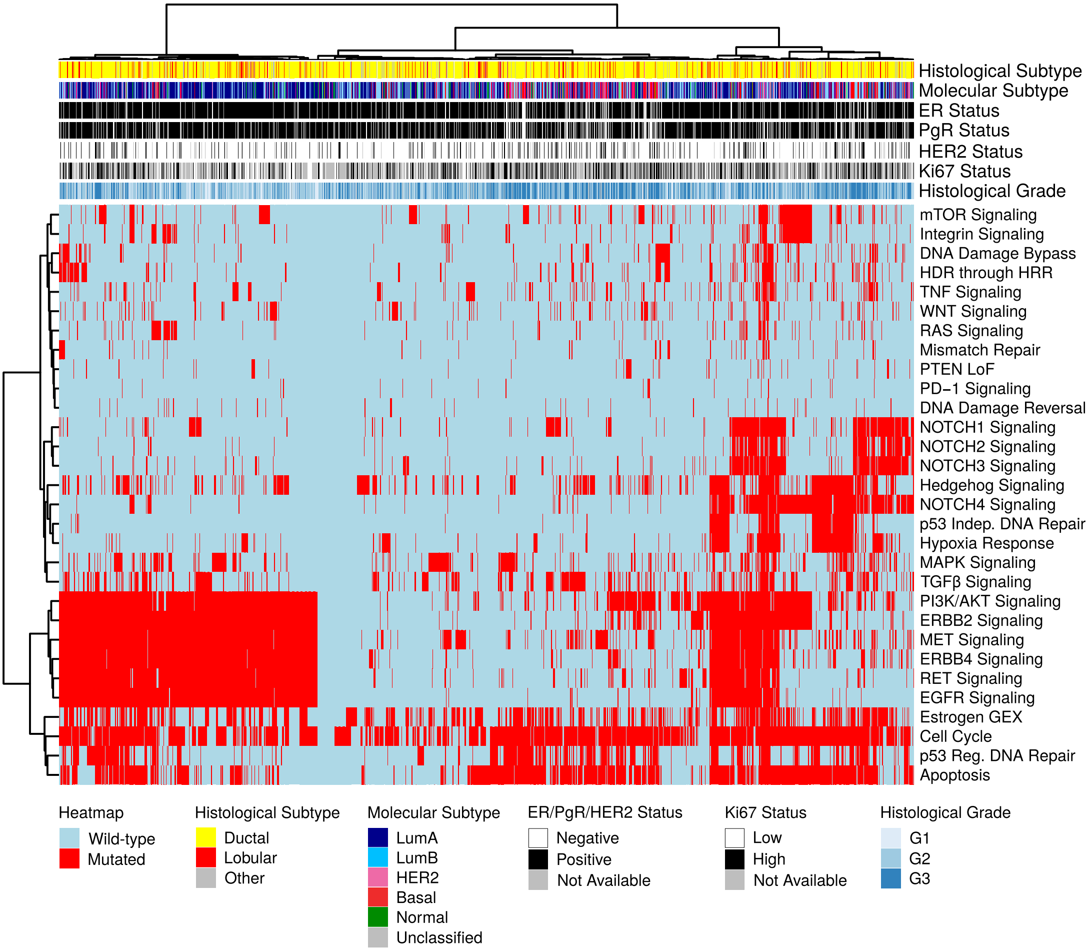


**Figure 4. Binary heatmap of mutation status of important breast cancer pathways in 3,217 samples.**

Binary heatmap of mutation status of important BC pathways in 3,217 samples. Samples with wild-type (wt) pathway status (defined as all member genes being wt) are colored blue, those with mutated pathways (at least one member gene mutated) are colored red. Samples and pathways were clustered using Euclidean distance and Ward linkage. Reactome IDs for the pathways can be found in Table EV4.

1.98], $P = 0.00036$), endocrine ± any therapy (HR: 1.4 [1.12–1.74], $P = 0.0028$), and chemotherapy ± any therapy (HR: 1.66 [1.12–2.47], $P = 0.011$). High TMB is typically associated with improved survival in TNBC, possibly due to increased neoantigen load enabling a stronger immune response. However, we observed no such effect in TNBC patients within the SCAN-B cohort ($P = 0.34$, Fig EV5). Mutational load was a significant survival stratifier across the Nottingham Histological Grade (NHG) grading scheme (G1, HR: 0.38 [0.16–0.9], $P = 0.022$; G2, HR: 1.46 [1.1–1.94], $P = 0.0078$; G3, HR: 1.53 [1.13–2.07], $P = 0.0055$), and within the ER$^+$ (HR: 1.41 [1.15–1.74], $P = 0.00097$), PgR$^+$ (HR: 1.28 [1.02–1.59], $P = 0.031$), HER2$^-$ (HR: 1.53 [1.25–1.86], $P = 0.000024$), and Ki67-high (HR: 1.76 [1.17–2.65], $P = 0.0064$) patient subgroups (Fig EV5). Interestingly, LumB patients with high TMB showed worse survival (HR: 1.58 [1.13–2.21], $P = 0.0064$), whereas TMB was not a significant stratifier for any other molecular subtype (Fig EV5). LumB tumors were also the only subgroup where TMB remained a significant stratifier in multivariable analysis.

### SCAN-B MutationExplorer

To enable public exploration and re-use of our rich mutational dataset, we developed the web-based application SCAN-B MutationExplorer (available at http://oncogenomics.bmc.lu.se/MutationExplorer; Fig 8).

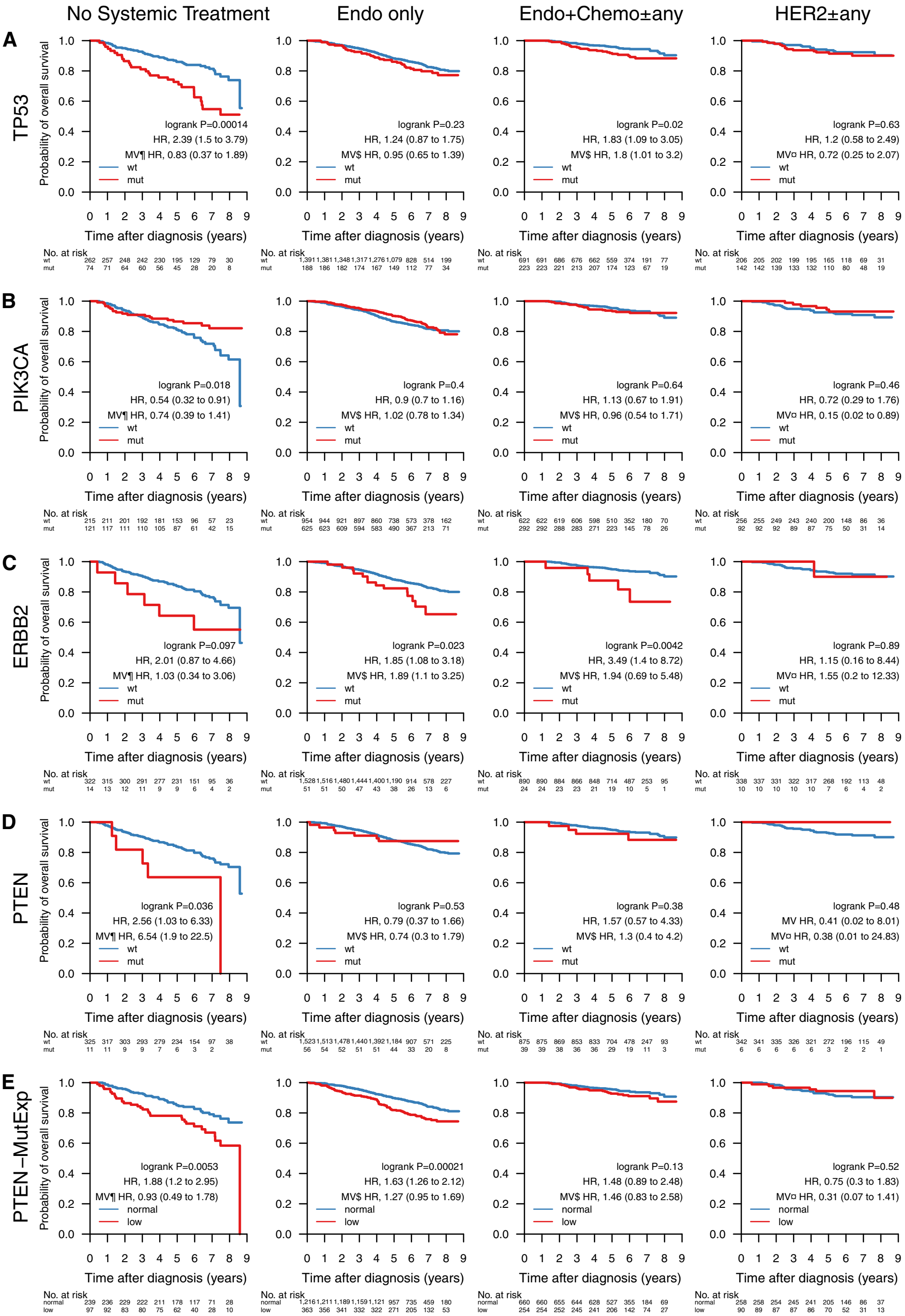

No Systemic Treatment
Endo only
Endo+Chemo±any
HER2±any
A
TP53
B
PIK3CA
C
ERBB2
D
PTEN
E
PTEN-MutExp
Probability of overall survival
Time after diagnosis (years)
No. at risk
logrank P=0.00014
HR, 2.39 (1.5 to 3.79)
MV¶ HR, 0.83 (0.37 to 1.89)
logrank P=0.23
HR, 1.24 (0.87 to 1.75)
MV$ HR, 0.95 (0.65 to 1.39)
logrank P=0.02
HR, 1.83 (1.09 to 3.05)
MV$ HR, 1.8 (1.01 to 3.2)
logrank P=0.63
HR, 1.2 (0.58 to 2.49)
MV¤ HR, 0.72 (0.25 to 2.07)
logrank P=0.018
HR, 0.54 (0.32 to 0.91)
MV¶ HR, 0.74 (0.39 to 1.41)
logrank P=0.4
HR, 0.9 (0.7 to 1.16)
MV$ HR, 1.02 (0.78 to 1.34)
logrank P=0.64
HR, 1.13 (0.67 to 1.91)
MV$ HR, 0.96 (0.54 to 1.71)
logrank P=0.46
HR, 0.72 (0.29 to 1.76)
MV¤ HR, 0.15 (0.02 to 0.89)
logrank P=0.097
HR, 2.01 (0.87 to 4.66)
MV¶ HR, 1.03 (0.34 to 3.06)
logrank P=0.023
HR, 1.85 (1.08 to 3.18)
MV$ HR, 1.89 (1.1 to 3.25)
logrank P=0.0042
HR, 3.49 (1.4 to 8.72)
MV$ HR, 1.94 (0.69 to 5.48)
logrank P=0.89
HR, 1.15 (0.16 to 8.44)
MV¤ HR, 1.55 (0.2 to 12.33)
logrank P=0.036
HR, 2.56 (1.03 to 6.33)
MV¶ HR, 6.54 (1.9 to 22.5)
logrank P=0.53
HR, 0.79 (0.37 to 1.66)
MV$ HR, 0.74 (0.3 to 1.79)
logrank P=0.38
HR, 1.57 (0.57 to 4.33)
MV$ HR, 1.3 (0.4 to 4.2)
logrank P=0.48
MV HR, 0.41 (0.02 to 8.01)
MV¤ HR, 0.38 (0.01 to 24.83)
logrank P=0.0053
HR, 1.88 (1.2 to 2.95)
MV¶ HR, 0.93 (0.49 to 1.78)
logrank P=0.00021
HR, 1.63 (1.26 to 2.12)
MV$ HR, 1.27 (0.95 to 1.69)
logrank P=0.13
HR, 1.48 (0.89 to 2.48)
MV$ HR, 1.46 (0.83 to 2.58)
logrank P=0.52
HR, 0.75 (0.3 to 1.83)
MV¤ HR, 0.31 (0.07 to 1.41)
wt
mut
normal
low
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6 7 8 9
wt 262 257 248 242 230 195 129 79 30
mut 74 71 64 60 56 45 28 20 8
wt 1,391 1,381 1,348 1,317 1,276 1,079 828 514 199
mut 188 186 182 174 167 149 112 77 34
wt 691 691 686 676 662 559 374 191 77
mut 223 223 221 213 207 174 123 67 19
wt 206 205 202 199 195 165 118 69 31
mut 142 142 139 133 132 110 80 48 19
wt 215 211 201 192 181 153 96 57 23
mut 121 117 111 110 105 87 61 42 15
wt 954 944 921 897 860 738 573 378 162
mut 625 623 609 594 583 490 367 213 71
wt 622 622 619 606 598 510 352 180 70
mut 292 292 288 283 271 223 145 78 26
wt 256 255 249 243 240 200 148 86 36
mut 92 92 92 89 87 75 50 31 14
wt 322 315 300 291 277 231 151 95 36
mut 14 13 12 11 9 9 6 4 2
wt 1,528 1,516 1,480 1,444 1,400 1,190 914 578 227
mut 51 51 50 47 43 38 26 13 6
wt 890 890 884 866 848 714 487 253 95
mut 24 24 23 23 21 19 10 5 1
wt 338 337 331 322 317 268 192 113 48
mut 10 10 10 10 10 7 6 4 2
wt 325 317 303 293 279 234 154 97 38
mut 11 11 9 9 7 6 3 2
wt 1,523 1,513 1,478 1,440 1,392 1,184 907 571 225
mut 56 54 52 51 51 44 33 20 8
wt 875 875 869 853 833 704 478 247 93
mut 39 39 38 36 36 29 19 11 3
wt 342 341 335 326 321 272 196 115 49
mut 6 6 6 6 6 3 2 2 1
normal 239 236 229 222 211 178 117 71 28
low 97 92 83 80 75 62 40 28 10
normal 1,216 1,211 1,189 1,159 1,121 957 735 459 180
low 363 356 341 332 322 271 205 132 53
normal 660 660 655 644 628 527 355 184 69
low 254 254 252 245 241 206 142 74 27
normal 258 258 254 245 241 205 146 86 37
low 90 89 87 87 86 70 52 31 13

**Figure 5.**

**Figure 5. Impact of gene mutations on overall survival across treatment groups.**

A–E Overall survival (OS) of patients with tumors containing mutations in the genes (A) *TP53*, (B) *PIK3CA*, (C) *ERBB2*, and (D) *PTEN*. (E) OS by *PTEN*-MutExp genotype ("low" defined as *PTEN* mutation or *PTEN* expression in the lower quartile across the cohort, "normal" otherwise) stratified by groups receiving no systemic treatment (*n* = 336), endocrine therapy only (Endo only; *n* = 1,579), endocrine- and chemotherapy (Endo + Chemo ± any; *n* = 914), as well as HER2 treatment with any other treatment or none (HER2 ± any; *n* = 348). Specific treatments in these groups are detailed in Table EV5. In each Kaplan–Meier plot, wild-type (wt) and normal cases are plotted in blue, mutated (mut) and low cases are plotted in red, the log-rank *P* value is given, and the hazard ratio (HR) for mutation/low is given with a 95% CI and after univariable and multivariable (MV) Cox regression adjustment. Covariables included in the MV analysis were age at diagnosis, lymph node status, tumor size, and the variables denoted by the following symbols: ¶, ER, PgR, HER2, and NHG; ¤, ER, PgR, and NHG; $, HER2 and NHG. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; NHG, Nottingham histological grade; PgR, progesterone receptor.

With this interactive application, a user can filter the 3,217 SCAN-B samples based on combinations of clinicopathological and molecular markers (histological type, ER, PgR, HER2, Ki67, NHG, and PAM50 subtype), treatments (endocrine, chemotherapy, HER2 treatment), and mutations based on mutation type (e.g., nonsense or missense) and COSMIC occurrence. From the filtered data, the user can create mutational landscape waterfall plots and conduct survival analysis using KM analysis and log-rank tests based on mutations in single genes, pathways as defined in the Reactome database or custom, as well as TMB, either using the absolute number of mutations, or mutations per expressed MB of genome, using a user-defined threshold. Mutations can also be plotted from a protein point of view using user-defined occurrence cutoffs for showing and annotating mutations. Plots in PDF format as well as the mutation set underlying the currently active plot in tab-separated values (TSV) format can be downloaded for further analysis. The application is based on R Shiny and the source code is available under the BSD 2-clause open source license at http://github.com/cbrueffer/MutationExplorer.

## Discussion

Tumor somatic mutation status is a crucial piece of information for the future of precision medicine to guide treatment selection and give insight into tumor evolution. Analysis of DNA is the gold standard for detecting SNVs, indels, and larger structural variants. However, many interesting tumor properties are only accessible on the transcriptome level and cannot be interrogated using DNA; most prominently gene expression at the isoform and gene level, as well as *de novo* transcripts originating from gene fusions. The SCAN-B initiative (Saal *et al*, 2015) decided early on to perform RNA-seq on the tumors of all enrolled patients. Based on this, we have developed, refined, and benchmarked gene expression signatures (Brueffer *et al*, 2018; Dihge *et al*, 2019; Lundgren *et al*, 2019; Søkilde *et al*, 2019; Vallon-Christersson *et al*, 2019), and detected recurring fusions affecting miRNAs (Persson *et al*, 2017). Herein, we described the development of a pipeline for detection of somatic SNVs and indels based on RNA-seq, adding another layer to information that can now be obtained from a single sequencing analysis within 1 week of surgery (Saal *et al*, 2015).

To date, several approaches for RNA-seq mutation calling, mostly in combination with matched DNA, have been developed (Horvath *et al*, 2013; Piskol *et al*, 2013; Radenbaugh *et al*, 2014; Wilkerson *et al*, 2014; Guo *et al*, 2017; Siegel *et al*, 2018); however, calling from RNA-seq alone, particularly from tumor-only samples, is still a challenge. With the advance of targeted and whole exome sequencing into the clinics, and efforts such as TCGA, MSK-Impact, and others, variant calling from DNA-seq has improved in recent years, although discordance between detection pipelines still exists (Hofmann *et al*, 2017; Ellrott *et al*, 2018; Shi *et al*, 2018). Part of this improvement is the availability of validation resources such as the Genome in a Bottle datasets (Zook *et al*, 2016). With clinical interest in RNA-seq only recently picking up, e.g., as shown by two recent review articles (Byron *et al*, 2016; Cieślik & Chinnaiyan, 2018), comparably well-characterized RNA-seq datasets for validation do not yet exist to our knowledge.

The strategy for mutation calling herein was to perform initial variant calling with low requirements on coverage and base quality to increase sensitivity while allowing false positives. To increase specificity, we then applied stringent *post hoc* filtering that can be easily amended as further annotation data become available, or as existing sources receive updates. The advantage of this two-step strategy is the possibility to accommodate different research and clinical questions in the future that may have different filtering needs.

Two major contributors of false-positive mutation calls are germline SNPs/indels and RNA editing. Common approaches for dealing with germline events are calling mutations from matched tumor/normal samples, or filtering SNPs present in databases such as dbSNP. The latter is problematic, since some dbSNP entries with a low variant allele frequency (VAF) may be legitimate somatic mutations. On the other hand, filtering on the dbSNP "common" flag (at least 1% VAF in any of the 1,000 genomes populations) can lead to many low-VAF germline SNPs remaining. We tried to address this issue by combining the dbSNP and COSMIC databases, and only filtering variants present in dbSNP if they were not present in COSMIC. We filtered out known RNA editing sites using publicly available databases; however, there is still an overabundance of T>C substitutions in our RNA-based calls compared to DNA-based calls, suggesting many unknown editing sites and insufficient filtering (Fig 2B). Approaches have been developed to identify RNA editing sites using DNA/RNA-trained machine learning models (Sun *et al*, 2016) or RNA-seq data alone (Ramaswami *et al*, 2013), which may provide ways to improve filtering in the future by creating a SCAN-B RNA editing database.

The overall landscape of somatic mutations in our study looked similar to that reported previously from DNA (The Cancer Genome Atlas, 2012; Pereira *et al*, 2016), with the two most frequently mutated genes *PIK3CA* (34% of samples) and *TP53* (23%), followed by other known drivers *MAP3K1* (7%), *CDH1* (7%), *GATA3* (7%), and *AKT1* (5%) (Fig 2). While mutation frequencies in oncogenes such as *PIK3CA* are generally in line with previous reports, frequencies in tumor suppressor genes were generally lower in RNA-seq than would be expected from our study population. For example,

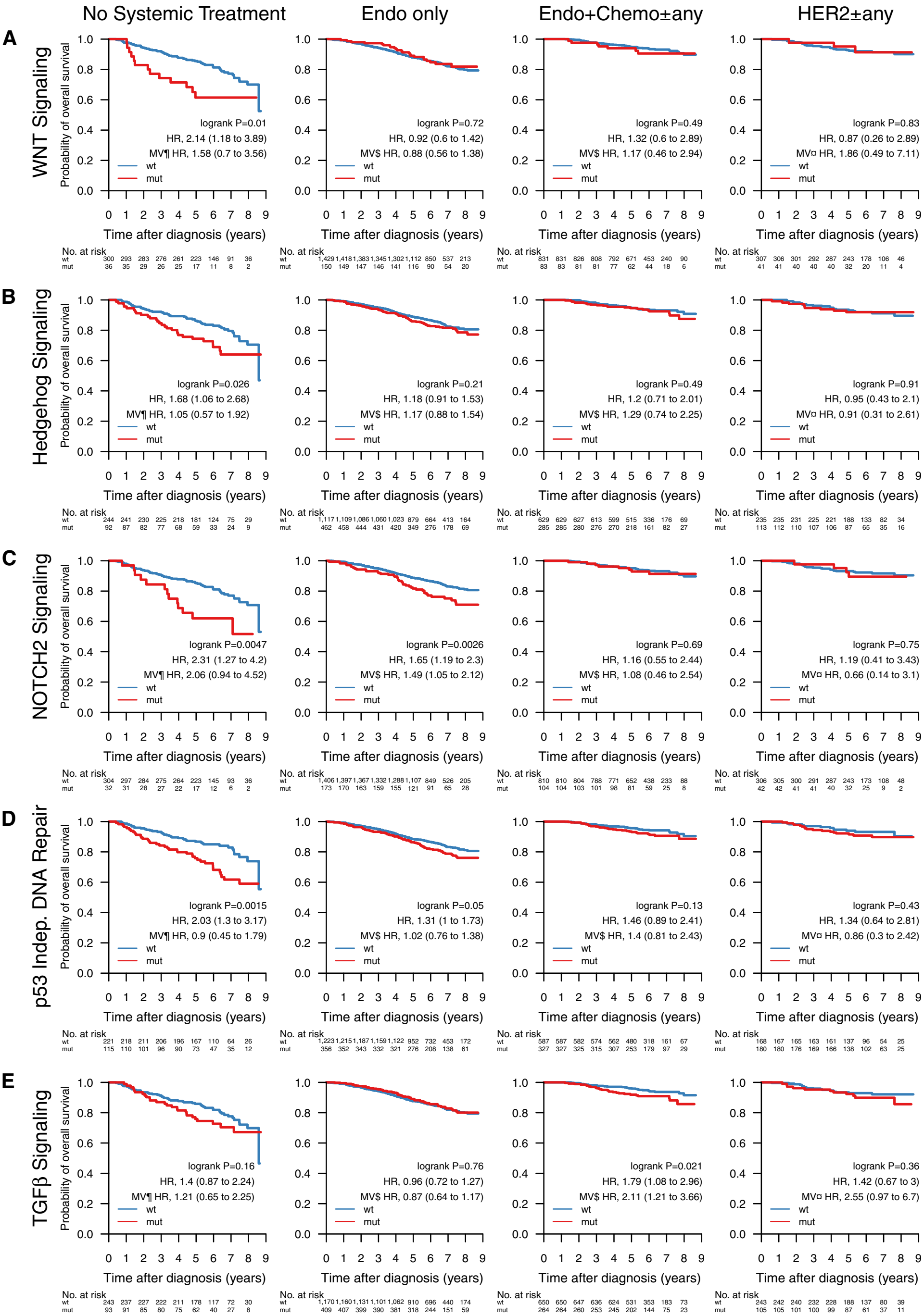

No Systemic Treatment
Endo only
Endo+Chemo±any
HER2±any
A
WNT Signaling
Probability of overall survival
1.0
0.8
0.6
0.4
0.2
0.0
0 1 2 3 4 5 6 7 8 9
Time after diagnosis (years)
wt
mut
logrank P=0.01
HR, 2.14 (1.18 to 3.89)
MV¶ HR, 1.58 (0.7 to 3.56)
No. at risk
wt 300 293 283 276 261 223 146 91 36
mut 36 35 29 26 25 17 11 8 2
logrank P=0.72
HR, 0.92 (0.6 to 1.42)
MV$ HR, 0.88 (0.56 to 1.38)
No. at risk
wt 1,429 1,418 1,383 1,345 1,302 1,112 850 537 213
mut 150 149 147 146 141 116 90 54 20
logrank P=0.49
HR, 1.32 (0.6 to 2.89)
MV$ HR, 1.17 (0.46 to 2.94)
No. at risk
wt 831 831 826 808 792 671 453 240 90
mut 83 83 81 81 77 62 44 18 6
logrank P=0.83
HR, 0.87 (0.26 to 2.89)
MV¤ HR, 1.86 (0.49 to 7.11)
No. at risk
wt 307 306 301 292 287 243 178 106 46
mut 41 41 40 40 40 32 20 11 4
B
Hedgehog Signaling
logrank P=0.026
HR, 1.68 (1.06 to 2.68)
MV¶ HR, 1.05 (0.57 to 1.92)
No. at risk
wt 244 241 230 225 218 181 124 75 29
mut 92 87 82 77 68 59 33 24 9
logrank P=0.21
HR, 1.18 (0.91 to 1.53)
MV$ HR, 1.17 (0.88 to 1.54)
No. at risk
wt 1,117 1,109 1,086 1,060 1,023 879 664 413 164
mut 462 458 444 431 420 349 276 178 69
logrank P=0.49
HR, 1.2 (0.71 to 2.01)
MV$ HR, 1.29 (0.74 to 2.25)
No. at risk
wt 629 629 627 613 599 515 336 176 69
mut 285 285 280 276 270 218 161 82 27
logrank P=0.91
HR, 0.95 (0.43 to 2.1)
MV¤ HR, 0.91 (0.31 to 2.61)
No. at risk
wt 235 235 231 225 221 188 133 82 34
mut 113 112 110 107 106 87 65 35 16
C
NOTCH2 Signaling
logrank P=0.0047
HR, 2.31 (1.27 to 4.2)
MV¶ HR, 2.06 (0.94 to 4.52)
No. at risk
wt 304 297 284 275 264 223 145 93 36
mut 32 31 28 27 22 17 12 6 2
logrank P=0.0026
HR, 1.65 (1.19 to 2.3)
MV$ HR, 1.49 (1.05 to 2.12)
No. at risk
wt 1,406 1,397 1,367 1,332 1,288 1,107 849 526 205
mut 173 170 163 159 155 121 91 65 28
logrank P=0.69
HR, 1.16 (0.55 to 2.44)
MV$ HR, 1.08 (0.46 to 2.54)
No. at risk
wt 810 810 804 788 771 652 438 233 88
mut 104 104 103 101 98 81 59 25 8
logrank P=0.75
HR, 1.19 (0.41 to 3.43)
MV¤ HR, 0.66 (0.14 to 3.1)
No. at risk
wt 306 305 300 291 287 243 173 108 48
mut 42 42 41 41 40 32 25 9 2
D
p53 Indep. DNA Repair
logrank P=0.0015
HR, 2.03 (1.3 to 3.17)
MV¶ HR, 0.9 (0.45 to 1.79)
No. at risk
wt 221 218 211 206 196 167 110 64 26
mut 115 110 101 96 90 73 47 35 12
logrank P=0.05
HR, 1.31 (1 to 1.73)
MV$ HR, 1.02 (0.76 to 1.38)
No. at risk
wt 1,223 1,215 1,187 1,159 1,122 952 732 453 172
mut 356 352 343 332 321 276 208 138 61
logrank P=0.13
HR, 1.46 (0.89 to 2.41)
MV$ HR, 1.4 (0.81 to 2.43)
No. at risk
wt 587 587 582 574 562 480 318 161 67
mut 327 327 325 315 307 253 179 97 29
logrank P=0.43
HR, 1.34 (0.64 to 2.81)
MV¤ HR, 0.86 (0.3 to 2.42)
No. at risk
wt 168 167 165 163 161 137 96 54 25
mut 180 180 176 169 166 138 102 63 25
E
TGFβ Signaling
logrank P=0.16
HR, 1.4 (0.87 to 2.24)
MV¶ HR, 1.21 (0.65 to 2.25)
No. at risk
wt 243 237 227 222 211 178 117 72 30
mut 93 91 85 80 75 62 40 27 8
logrank P=0.76
HR, 0.96 (0.72 to 1.27)
MV$ HR, 0.87 (0.64 to 1.17)
No. at risk
wt 1,170 1,160 1,131 1,101 1,062 910 696 440 174
mut 409 407 399 390 381 318 244 151 59
logrank P=0.021
HR, 1.79 (1.08 to 2.96)
MV$ HR, 2.11 (1.21 to 3.66)
No. at risk
wt 650 650 647 636 624 531 353 183 73
mut 264 264 260 253 245 202 144 75 23
logrank P=0.36
HR, 1.42 (0.67 to 3)
MV¤ HR, 2.55 (0.97 to 6.7)
No. at risk
wt 243 242 240 232 228 188 137 80 39
mut 105 105 101 100 99 87 61 37 11

**Figure 6.**

**Figure 6. Impact of pathway mutations on overall survival across treatment groups.**

A–E Overall survival of patients with tumors containing mutations in pathways (A) WNT signaling, (B) Hedgehog signaling, (C) NOTCH2 signaling, (D) p53 independent DNA damage repair, (E) TGFβ signaling, stratified by groups receiving no systemic treatment (*n* = 336), endocrine therapy only (Endo only; *n* = 1,579), endocrine- and chemotherapy (Endo + Chemo ± any; *n* = 914), as well as HER2 treatment with any other treatment or none (HER2 ± any; *n* = 348). Specific treatments in these groups are detailed in Table EV4. In each Kaplan–Meier plot, wild-type (wt) cases are plotted in blue, mutated (mut) cases are plotted in red, the log-rank *P* value is given, and the hazard ratio (HR) for mutation is given with a 95% CI and after univariable and multivariable (MV) Cox regression adjustment. Covariables included in the MV analysis were age at diagnosis, lymph node status, tumor size, and the variables denoted by the following symbols: ¶, ER, PgR, HER2, and NHG; ¤, ER, PgR, and NHG; $, HER2 and NHG. See Table EV3 for Reactome pathway IDs. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; NHG, Nottingham histological grade; PgR, progesterone receptor.

our *TP53* RNA-seq somatic mutation frequency of 23% (reference: 36%, cBioPortal.org) suggests we may be missing a significant fraction of *TP53* mutations present in DNA. Similar trends can be seen in *PTEN* (observed: 3.6%, reference: 4.6%), *BRCA1* (observed: 0.2%, reference: 1.6%), and *BRCA2* (observed: 0.03%, reference: 2.2%). This is not surprising since only mutations in sufficiently highly expressed genomic regions can be detected by RNA-seq and loss of expression of tumor suppressor genes is a hallmark of oncogenesis. Furthermore, truncated mRNAs caused by nonsense mutations are typically removed by nonsense-mediated decay before they can be captured for sequencing. Thus, our findings do not reflect the true mutational spectrum of tumor suppressor genes. Despite these limitations, we could identify a putative mutation in at least one gene targeted by an existing drug in the majority of patient tumors (86.8%), demonstrating that it should be feasible to match most patients to targeted treatments using RNA-seq analyses.

One of the major oncogenic pathways in breast cancer is PI3K/AKT/mTOR, which is frequently upregulated by activating mutations in *PIK3CA*, *MAP3K1*, and *AKT1*, or inactivating mutations in *PTEN*, leading to increased growth signaling. This pathway is being targeted by multiple drugs, such as alpelisib (Novartis) (Juric *et al*, 2018) in HoR$^+$/HER2$^-$ *PIK3CA* mutant tumors in combination with fulvestrant (André *et al*, 2019), and the AKT1 inhibitor AZD5363 (AstraZeneca) (Hyman *et al*, 2017). The strength of RNA-seq in mutation profiling lies within oncogenes, and we demonstrate that alterations in drug targets such as *PIK3CA* and *AKT*, as well as genes potentially modulating drug efficacy, such as *MAP3K1*, can be detected. Eventually, RNA-seq may be used as companion diagnostic for oncogene-targeting drugs such as these. While we also detected mutations in *PTEN*, these only showed significant prognostic power when combined with low gene expression in the *PTEN*-MutExp low group, suggesting either SNVs and indels are a minor mechanism of PTEN loss in early BC compared to structural rearrangements (Saal *et al*, 2008), and other means of *PTEN* expression loss. Taken together, we detected mutations in multiple PI3K/AKT/mTOR signaling nodes that lead to increased pathway activation and have emerging clinical utility in luminal BC, e.g., through combination with EGFR inhibition as demonstrated in basal-like BC (She *et al*, 2016).

Loss of p53 activity, either through LoF mutations, dominant-negative mutations, or low expression, is a major contributor to tumorigenesis. While RNA-seq generally underdetects *TP53* mutations, the identified hot spot residues remain the same as reported in the IARC TP53 database. Clinically these mutations could already be actionable, as *TP53* mutations are a sign of DNA damage repair deficiency and may be prognostic for sensitivity to PARP inhibition (Holstege *et al*, 2010; Severson *et al*, 2015). Patients with *TP53*-mutant tumors had significantly worse OS in the patient subgroups treated only with endocrine therapy, or no systemic treatment at all (Fig 5), and HoR$^+$/HER2$^-$ patients (Fig EV2), suggesting that *TP53* mutations identify a subgroup of patients that are spared chemotherapy or systemic therapy overall by appearing low risk, but are in fact high-risk patients that should be treated accordingly.

Endocrine treatment is the most important first-line treatment in BC. Resistance to these treatments leads to disease progression and recurrence and has been studied extensively. Drivers for endocrine resistance include activating mutations in *ESR1* and *ERBB2* which have been studied mostly in the metastatic setting. We show that mutations in these genes already occur in early, untreated BC, with 177 (5.5%) of patients in our population-based cohort having a mutation in either gene. We further demonstrate that patients with these mutations that received only endocrine treatment have inferior OS, suggesting drug resistance. Detecting these patients early could open up additional treatment options that have shown efficacy in the metastatic setting, such as selective estrogen receptor degraders (SERDs) in *ESR1*-mutated tumors, or TKIs such as neratinib in *ERBB2*-mutated BC.

The role of alternative splicing in tumorigenesis has recently garnered increased attention, and the extend of isoform switching in several cancer types, including BC, has been characterized (Vitting-Seerup & Sandelin, 2017). Mutations such as the *SF3B1* K700E hot spot mutation deregulate splicing and result in differential splicing patterns in BC (Maguire *et al*, 2015). The clinical effect of these mutations is unclear, and we did not detect significant survival stratification in important biomarker or treatment groups. However, the fact that mutations in splicing-related genes can be detected from RNA-seq make this method attractive for research and possible clinical use, as they can be correlated with expression originating from the same sequencing experiment.

Individual mutations, particularly in infrequently mutated genes, affect a smaller number of molecular pathways to achieve the classical hallmarks of cancer such as sustained proliferative signaling. Mutation status of several individual pathways was associated with reduced OS in different treatment subgroups. In patients not systemically treated or only treated with endocrine therapy WNT, NOTCH2, p53-independent DNA repair pathway mutation status, and Hedgehog signaling mutation status may identify patients diagnosed as low risk who may benefit from more adjuvant treatment (Fig 6). While these stratification profiles were visible in treatment subgroups, they mostly did not yield significant results in clinical biomarker subgroups (Fig EV3). This may indicate that current risk stratification in histopathological biomarker subgroups is inadequate and should take molecular information into account—something we and others have also shown on the level of gene expression (Brueffer *et al*, 2018). Identifying the mutation status of pathways and pathway clusters may aid in future clinical trials and treatment, e.g., by aiding selection of treatments that exploit synthetic lethality (Weidle *et al*, 2011).

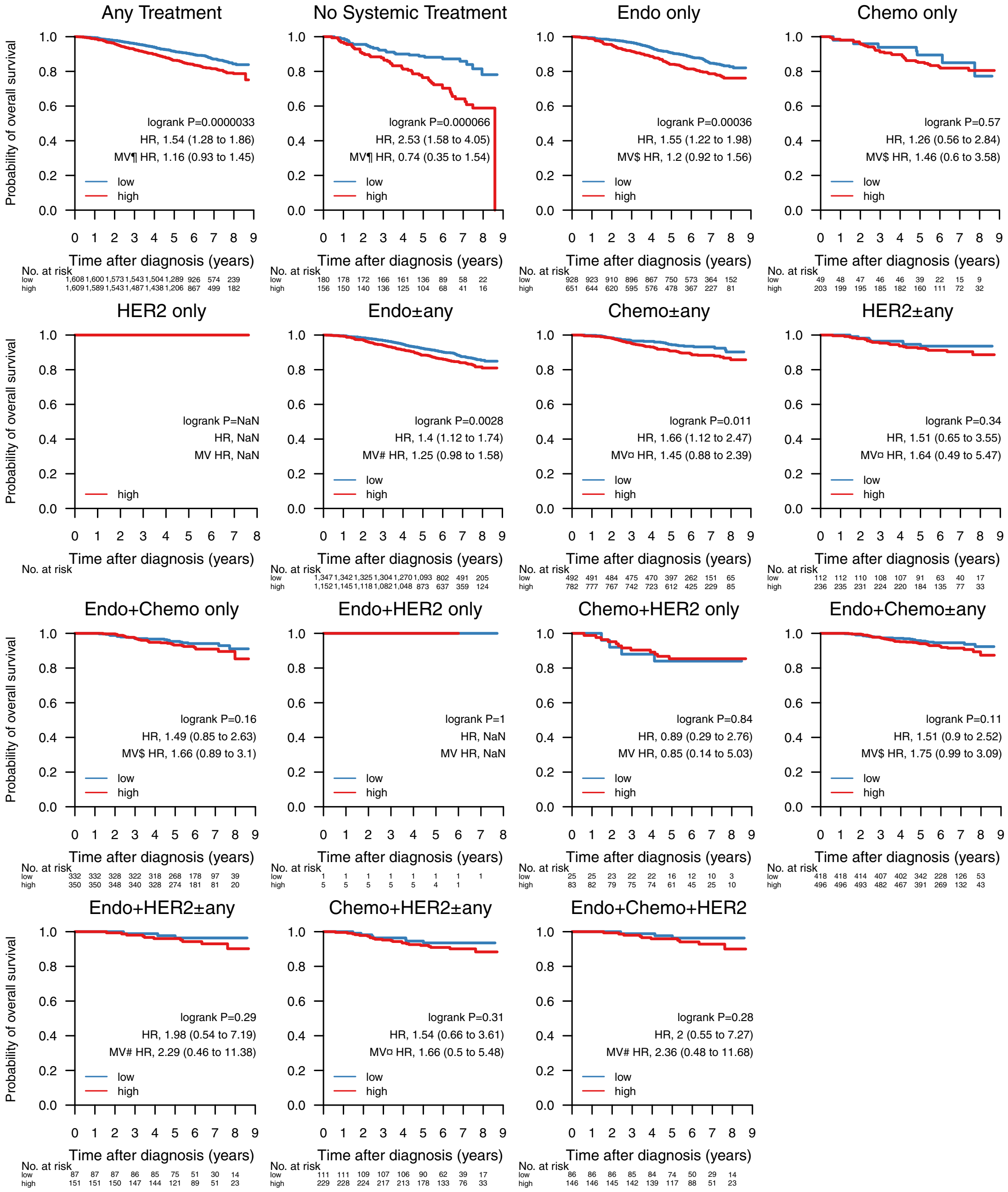
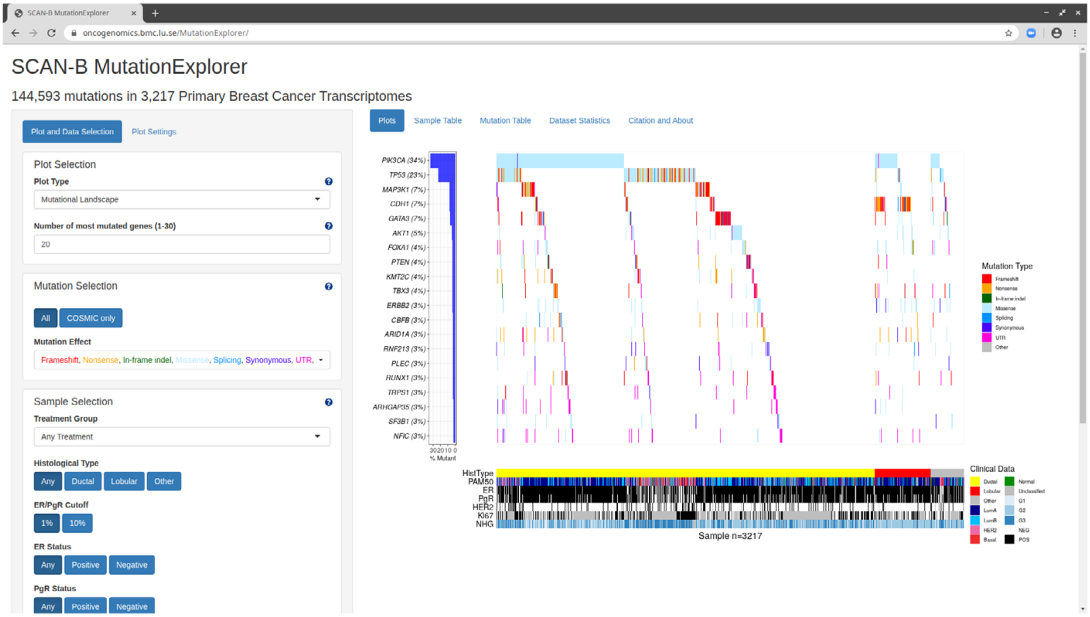


**Figure 7.**

**Figure 7. Impact of tumor mutational burden on overall survival across treatment groups.**

Overall survival stratified by tumor mutational burden (TMB) across treatment groups in 3,217 patients. Samples were classified as TMB-high if the amount of non-synonymous mutations per expressed MB (rnaMB) was ≥ the median number of non-synonymous mutations per rnaMB across the whole SCAN-B cohort (0.082 mutations per rnaMB) and TMB-low otherwise. In each Kaplan–Meier plot, TMB-low cases are plotted in blue, TMB-high cases are plotted in red, the log-rank *P* value is given, and the hazard ratio (HR) for TMB high is given with a 95% CI and after univariable and multivariable (MV) Cox regression adjustment. Covariables included in the MV analysis were age at diagnosis, lymph node status, tumor size, and the variables denoted by the following symbols: ¶, ER, PgR, HER2, and NHG; ¤, ER, PgR, and NHG; $, HER2 and NHG; #, NHG. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; HoR, hormone receptor; NHG, Nottingham histological grade; PgR, progesterone receptor; TMB, tumor mutational burden; TNBC, triple-negative breast cancer.



SCAN-B MutationExplorer

oncogenomics.bmc.lu.se/MutationExplorer/

SCAN-B MutationExplorer

144,593 mutations in 3,217 Primary Breast Cancer Transcriptomes

Plot and Data Selection | Plot Settings

Plot Selection

Plot Type: Mutational Landscape

Number of most mutated genes (1-30): 20

Mutation Selection

[✓ All] COSMIC only

Mutation Effect: Frameshift, Nonsense, In-frame indel, Missense, Splicing, Synonymous, UTR,

Sample Selection

Treatment Group: Any Treatment

Histological Type: [✓ Any] Ductal Lobular Other

ER/PgR Cutoff: [✓ 1%] 10%

ER Status: [✓ Any] Positive Negative

PgR Status: [✓ Any] Positive Negative

Plots | Sample Table | Mutation Table | Dataset Statistics | Citation and About

**Figure 8. The SCAN-B MutationExplorer.**

The SCAN-B MutationExplorer web-based application for interactive exploration of mutations, and their association with clinicopathological subgroups and overall survival. As an example, generation of the image used in Fig 2 is shown.

High TMB has been identified as a predictive biomarker for response to immune checkpoint therapy in diverse solid tumors (Goodman *et al*, 2017; Lauss *et al*, 2017; Hellmann *et al*, 2018; Thomas *et al*, 2018; Zacharakis *et al*, 2018). Using RNA-seq to assess mutational burden may be a useful capability for clinical trials and eventual clinical implementation in BC (Schmid *et al*, 2018). Questions remain however, as TMB is influenced by many biological and technical factors such as ploidy, tumor heterogeneity and clonality (Conroy *et al*, 2019), sample tumor cell content, sequencing depth, and variant filtering. Which cutoff to use for stratifying patients into TMB groups is also still emerging (Panda *et al*, 2017; Schmid *et al*, 2018), and specifically has not been addressed to our knowledge in RNA-seq data. Due to this, and to account for different expression profiles per tumor, we decided to use the median number of non-synonymous mutations per MB of transcriptome across the cohort to stratify patients into TMB-high and TMB-low groups and use it to study OS in different conventional treatment and biomarker subgroups. In several of these groups, high TMB was significantly associated with worse survival, confirming previous reports (Xu *et al*, 2018), however interestingly not in TNBC. These tumors typically show higher TMB than other clinical BC subtypes, likely because many of them have impaired DNA damage repair mechanisms. Shah and colleagues (Shah *et al*, 2012) showed that only ~ 36% of mutations in TNBCs are expressed; we speculate that due to this, we may underestimate TMB in several of our TMB-low patients. Additionally, RNA-seq underdetects truncating mutations such as frameshift indels that are a major source of neoantigens. Immune checkpoint therapy is a particularly attractive treatment approach in patients with TNBC and basal-like tumors for which currently no targeted therapy exists. For these patients, determination of TMB using DNA-seq may be a better option than relying on RNA-seq.

Large-scale projects such as TCGA and SCAN-B generate vast amounts of data, but bioinformatics skills are required to make efficient use of them. Web portals such as cBioPortal (Cerami *et al*, 2012) have emerged to make these huge datasets explorable without specialized skills. In this spirit, we developed the open source web application SCAN-B MutationExplorer to make our mutation dataset easily accessible for other researchers. We hope that SCAN-B MutationExplorer will aid knowledge generation and the development of better BC biomarkers in the future. The open source nature of the portal allows developers to adopt the code for their own purposes, and we welcome contributions of any kind.

### Limitations

The mutation calling we have performed herein tries to achieve sensitive variant calling by using lenient parameters, and heavy filtering of the resulting variants based on stringent quality factors, annotations, and curated databases. This approach has several limitations. While our 275 patient cohort for filter development had matched tumor and normal DNA sequencing data, the SCAN-B cohort only consisted of tumor RNA-seq data. This made accounting for PCR and sequencing artifacts more challenging. Further while many germline events can be filtered by comparing to general databases such as dbSNP, and population-specific ones such as SweGen, these databases are incomplete, and it is thus not possible to remove all germline events this way. As these databases improve, our filters can be upgraded to increase performance. Herein, we also applied filters developed in a matched DNA/RNA set of targeted capture sequencing of 1,697 genes and 1,047 miRNAs (275 sample ABiM cohort) to whole mRNA-seq (3,217 sample SCAN-B cohort). This assumes the transcriptional characteristics of the captured regions are representative for the whole mRNA.

### Conclusion

In summary, we present a tumor-only RNA-seq variant calling strategy and resulting mutation dataset from a large population-based early breast cancer cohort. Although variant calling from RNA-seq data is limited to expressed regions of the genome, mutations in important BC genes such as *PIK3CA*, *TP53*, and *ERBB2*, as well as pathways can be reliably detected, which may be used to inform clinical trials and eventual reporting to the clinic. Mutations in *TP53*, *PIK3CA*, *ERBB2*, and *PTEN* provided prognostic information in several treatment and biomarker patient subgroups, demonstrating the utility of the dataset for research. We make this dataset available for analysis and download via the open source web application SCAN-B MutationExplorer, accessible at http://oncogenomics.bmc.lu.se/MutationExplorer.

## Materials and Methods

### Patients

The study was approved by the Regional Ethics Review Board of Lund at Lund University (diary numbers 2007/155, 2009/658, 2009/659, 2010/383, 2012/58, 2013/459). We analyzed data from two previously described cohorts. For 273 patients, including two patients with bilateral disease (thus 275 tumors), enrolled in the All Breast Cancer in Malmö (ABiM) study from 2007 to 2009, matched snap-frozen primary breast tumor tissue and blood samples were collected as previously described (Winter *et al*, 2016). A cohort of 3,273 SCAN-B primary breast tumors described previously (Brueffer *et al*, 2018) was reduced to 3,217 samples following additional quality controls. All patients provided informed consent, and the study conforms to the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report. Tissue collection, preservation in RNAlater, sequencing, expression estimation, and molecular subtyping using the PAM50 gene list were performed as previously reported (Saal *et al*, 2015; Brueffer *et al*, 2018). Clinical records were retrieved from the Swedish National Cancer Registry (NKBC). Estrogen receptor (ER) and progesterone receptor (PgR) status was categorized using an immunohistochemical staining cutoff of 1%. Patients in the SCAN-B cohort had median 74.5 months follow-up, and patient demographics for both cohorts are detailed in Table 3.

### Library preparation and sequencing

For the 275 sample ABiM cohort, tumor and normal DNA was sequenced using a custom targeted capture panel of 1,697 genes and 1,047 miRNAs as described (Winter *et al*, 2016). For the same tumors, RNA-seq was performed as described (Brueffer *et al*, 2018) (a subset of the 405 sample cohort therein). In short, strand-specific dUTP libraries were prepared and sequenced on an Illumina HiSeq 2000 sequencer to an average of 50 million 101 bp reads per sample (Parkhomchuk *et al*, 2009; Saal *et al*, 2015).

For the 3,217 sample SCAN-B cohort, RNA-seq data were generated as previously described (Brueffer *et al*, 2018). In short, strand-specific dUTP mRNA-seq libraries were prepared (Parkhomchuk *et al*, 2009; Saal *et al*, 2015), and an average 38 million 75 bp reads were sequenced on an Illumina HiSeq 2000 or NextSeq 500 instrument (Table EV1).

### Sequence data processing

For tumor and normal DNA, reads were aligned to the GRCh37 reference genome using Novoalign 2.07.18 (Novocraft Technologies, Malaysia). Using a modified version of the variant workflow of the bcbio-nextgen NGS framework (https://github.com/bcbio/bcbio-nextgen, modified version https://github.com/cbrueffer/bcbio-nextgen/tree/v1.0.2-scanb-calling) utilizing Bioconda for software management (Grüning *et al*, 2018), duplicate reads were marked using biobambam v2.0.62 (Tischler & Leonard, 2014) and variants were called from paired tumor/normal samples using VarDict-Java 1.5.0 (Lai *et al*, 2016) (with default options except -f 0.02 -N ${SAMPLE} -b ${BAM_FILE} -c 1 -S 2 -E 3 -g 4 -Q 10 -r 2 -q 20), which internally performs local realignment around indels. Variant coordinates were converted to the GRCh38 reference genome using CrossMap 2.5 (Zhao *et al*, 2014). Raw RNA-seq reads were trimmed and filtered as described previously (Brueffer *et al*, 2018) and then processed using the modified bcbio-nextgen 1.0.2 variant workflow. Reads were aligned to a version of the GRCh38.p8 reference genome that included alternative sequences and decoys and was patched with dbSNP Build 147 common SNPs, and the GENCODE 25 transcriptome model using HISAT2 2.0.5 (Kim *et al*, 2015) (with default

**Table 3. Patient demographics and clinicopathological variables in the ABiM and SCAN-B cohorts.**

| | ABiM cohort (275 Samples) | | SCAN-B cohort (3,217 Samples) | |
|---|---|---|---|---|
| | Patient count | Percent (%) | Patient count | Percent (%) |
| Age (years) | | | | |
| <50 | 64 | 23.3 | 597 | 18.6 |
| ≥50 | 211 | 76.7 | 2,620 | 81.4 |
| Tumor size (mm) | | | | |
| ≤20 | 145 | 52.7 | 2,080 | 64.7 |
| 21–50 | 120 | 43.6 | 1,018 | 31.6 |
| >50 | 6 | 2.2 | 77 | 2.4 |
| Missing | 3 | 1.1 | 42 | 1.3 |
| Positive lymph nodes (number) | | | | |
| 0 | 151 | 54.9 | 1,974 | 61.4 |
| 1–3 | 60 | 21.8 | 851 | 26.5 |
| ≥4 | 44 | 16.0 | 290 | 9.0 |
| Missing | 20 | 7.3 | 102 | 3.2 |
| Histological type | | | | |
| Ductal | 215 | 78.2 | 2,602 | 80.9 |
| Lobular | 23 | 8.4 | 386 | 12.0 |
| Other | 28 | 10.2 | 229 | 7.1 |
| Missing | 9 | 3.3 | 0 | 0.0 |
| ER status (1% cutoff) | | | | |
| Positive | 223 | 81.1 | 2,786 | 86.6 |
| Negative | 48 | 17.5 | 233 | 7.2 |
| Missing | 4 | 1.5 | 198 | 6.2 |
| PgR status (1% cutoff) | | | | |
| Positive | 204 | 74.2 | 2,509 | 78.0 |
| Negative | 64 | 23.3 | 379 | 11.8 |
| Missing | 7 | 2.5 | 329 | 10.2 |
| HER2 status | | | | |
| Positive | 44 | 16.0 | 414 | 12.9 |
| Negative | 197 | 71.6 | 2,651 | 82.4 |
| Missing | 34 | 12.4 | 152 | 4.7 |
| Nottingham histological grade | | | | |
| Grade 1 | 31 | 11.3 | 483 | 15.0 |
| Grade 2 | 97 | 35.3 | 1,509 | 46.9 |
| Grade 3 | 146 | 53.1 | 1,161 | 36.1 |
| Missing | 1 | 0.4 | 64 | 2.0 |
| Ki67 status | | | | |
| High | 109 | 39.6 | 887 | 27.6 |
| Low | 153 | 55.6 | 627 | 19.5 |
| Missing | 13 | 4.7 | 1,703 | 52.9 |
| Molecular subtype | | | | |
| Luminal A | 109 | 39.6 | 1,545 | 48.0 |
| Luminal B | 83 | 30.2 | 899 | 27.9 |
| HER2-enriched | 30 | 10.9 | 279 | 8.7 |
| Basal-like | 35 | 12.7 | 318 | 9.9 |

**Table 3** (continued)

| | ABiM cohort (275 Samples) | | SCAN-B cohort (3,217 Samples) | |
|---|---|---|---|---|
| | Patient count | Percent (%) | Patient count | Percent (%) |
| Normal-like | 13 | 4.7 | 112 | 3.5 |
| Unclassified | 5 | 1.8 | 64 | 2.0 |

options except --rna-strandness RF --rg-id ${ID_NAME} --rg PL:illumina --rg PU:${UNIT} --rg SM:${SAMPLE}). BAM index files were generated using Sambamba 0.6.6 (Faust & Hall, 2014), and duplicate reads were marked using SAMBLASTER 0.1.24 (Tarasov *et al*, 2015). Variants were called using VarDict-Java 1.5.0 with default options except -f 0.02 -N ${SAMPLE} -b ${BAM_FILE} -c 1 -S 2 -E 3 -g 4 -Q 10 -r 2 -q 20 callable_bed, where callable_bed was a sample-specific BED file containing all regions of depth ≥ 4.

All variants were annotated using a Snakemake (Köster & Rahmann, 2012) workflow around vcfanno 0.3.1 (Pedersen *et al*, 2016) and the data sources dbSNP v151 (Sherry *et al*, 2001), Genome Aggregation Database (gnomAD) (Karczewski *et al*, 2020), Catalogue of Somatic Mutations in Cancer (COSMIC) v87 (Forbes *et al*, 2015; Sondka *et al*, 2018), CIViC (Griffith *et al*, 2017), MyCancerGenome (release March 2016, http://www.mycancergenome.org), SweGen version 20171025 (Ameur *et al*, 2017), the Danish Genome Project population reference (Maretty *et al*, 2017), RNA editing databases (Kiran & Baranov, 2010; Ramaswami & Li, 2014; Sun *et al*, 2016; Picardi *et al*, 2017), UCSC low complexity regions, IntOGen breast cancer driver gene status (Gonzalez-Perez *et al*, 2013) (accessed 2018-08-02), and the drug gene interaction database (DGIdb) v3.0.2 (Cotto *et al*, 2017). We used SnpEff v4.3.1t (with default parameters except hg38 -t -canon) (Cingolani *et al*, 2012b) to predict functional variant impact on canonical transcripts as defined by SnpEff.

To filter out recurrent artifacts introduced during library preparation or sequencing, we constructed a panel of "normal" tissues consisting of all variants enumerated from RNA-seq analysis of adjacent non-tumoral breast tissues sampled from 10 SCAN-B patients.

Gene expression data in fragments per kilobase of transcript per million mapped reads (FPKM) for the ABiM and SCAN-B cohorts were generated as previously reported and is available from the NCBI Gene Expression Omnibus, accession GSE81540 (Brueffer *et al*, 2018).

### Variant filtering

The strategy we applied for developing DNA-seq-informed filters is outlined in Fig 1. Due to the lenient settings used for sensitive initial variant calling, we developed and applied rigid filters to reduce false-positive calls resulting from either sequencing or PCR artifacts, RNA editing, or germline variants. To this end, variants called from 275 matched tumor/normal targeted capture DNA datasets were filtered, among other parameters, for low complexity regions, SNP status (dbSNP "common", SweGen and COSMIC SNPs, high gnomAD allele frequency), allele frequency ≥ 0.05, depth ≥ 8, homopolymer environments, and RNA editing sites. Using the resulting DNA variants as reference, we developed filters for the 275 sample RNA-seq variants by permuting values of the sequencing, variant calling, and annotation variables, and for each permutation calculating the concordance to the DNA mutations. Following these "negative" filters, we applied a range of "positive" filters to rescue filtered variants, e.g., to retain a variant if it is present in the curated MyCancerGenome database of clinically important mutations. Finally, we selected the combination of "negative" and "positive" filter settings with the best balance of sensitivity and specificity. Using SnpSift (Cingolani *et al*, 2012a), we applied the filters to RNA-seq mutation calls from the 3,217 patient cohort. A complete list of final filter variables and values for both the tumor/normal DNA variant calls, as well as the RNA-seq variant calls can be found in Table EV6.

### Data analysis

All analyses were performed using R 3.5.1. Waterfall, heatmap, and lollipop plots were made using the *GenVisR* 1.14.2 (Skidmore *et al*, 2016), *pheatmap* 1.0.12, and *RTrackLayer* 1.42.1 packages. Substitution signatures were analyzed using the *MutationalPatterns* 1.8.0 package (Blokzijl *et al*, 2018). Survival analysis was conducted using OS as endpoint. Overall survival was analyzed using the Kaplan–Meier (KM) method, two-sided log-rank tests, and Cox models, all implemented in the *survival* 2.44-1.1 package. Multivariable Cox models included the variables age at diagnosis, lymph node status, and tumor size as covariables, as well as ER, PgR, HER2, and NHG as relevant. All models were checked for proportional hazards using Grambsch and Therneau's test for non-proportionality and Schoenfeld residuals (Grambsch and Therneau, 1994). Associations were tested using one-tailed and two-tailed Fisher's exact test. *P*-values < 0.05 were considered significant. The web application SCAN-B MutationExplorer was written in R using the *Shiny*, *GenVisR,* and *SurvMiner* packages.

## Data availability

The datasets produced and used in this study are available in the following databases:

- Clinical data and mutation calls: http://oncogenomics.bmc.lu.se/MutationExplorer
- Gene expression data: NCBI Gene Expression Omnibus GSE81540 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81540; Brueffer *et al*, 2018).

Raw patient sequencing data cannot be provided due to Swedish data protection laws.

**Expanded View** for this article is available online.

**The paper explained**

**Problem**
Breast cancer is a disease of genomic alterations, of which the complete panorama of somatic mutations and how these relate to molecular subtypes, therapy response, and clinical outcomes is incompletely understood. RNA sequencing is a powerful technique for profiling tumor transcriptomes; however, using it for reliable detection of single nucleotide variants and small insertions and deletions is challenging.

**Results**
Within the Sweden Cancerome Analysis Network-Breast project (SCAN-B; ClinicalTrials.gov NCT02306096), we developed an optimized bioinformatics pipeline for detection of single nucleotide variants and small insertions and deletions from RNA-seq data. From this, we describe the mutational landscape of 3,217 primary breast cancer transcriptomes and relate it to patient overall survival in a real-world setting (median follow-up 75 months, range 2–105 months). We demonstrate that RNA-seq can be used to call mutations in important breast cancer genes such as *PIK3CA*, *TP53*, *ESR1*, and *ERBB2*, as well as mutation status of key molecular pathways and tumor mutational burden. We identify mutations in one or more potentially druggable genes in 86.8% of cases and reveal significant relationships to patient outcome within specific treatment groups, such as occurrence of mutations inducing resistance to standard of care drugs in untreated patients. To make this rich and growing mutational portraiture of breast cancer available for the wider research community, we developed an open source interactive web application, SCAN-B MutationExplorer, publicly accessible at http://oncogenomics.bmc.lu.se/MutationExplorer.

**Impact**
These results add another dimension to the use of RNA-seq as a potential clinical tool, where both gene expression-based signatures and gene mutation-based biomarkers can be interrogated simultaneously and in real-time within 1 week of tumor sampling. Treatment resistance mutations can be detected in early disease and could inform clinical decision-making.

## Acknowledgements

We thank the patients who were part of this study and the SCAN-B study, the employees of the SCAN-B laboratory, the South Sweden Breast Cancer Group, and all SCAN-B collaborators at Hallands Hospital Halmstad, Helsingborg Hospital, Blekinge County Hospital, Central Hospital Kristianstad, Skåne University Hospital Lund/Malmö, Central Hospital Växjö, for inclusion of patients and sampling of tissue for this study. We also thank the Swedish National Breast Cancer Registry and Regional Cancer Center South for clinical data. The SweGen allele frequency data were generated by Science for Life Laboratory. This work was supported by the Mrs. Berta Kamprad Foundation (to ÅB, LHS) and funded in part by the Swedish Research Council (ÅB, LHS), Swedish Cancer Society (ÅB, LHS), Swedish Foundation for Strategic Research (ÅB), Knut and Alice Wallenberg Foundation (ÅB), VINNOVA (ÅB), Governmental Funding of Clinical Research within National Health Service (ALF) (ÅB, LHS), Scientific Committee of Blekinge County Council (AE), Crafoord Foundation (LHS), Lund University Medical Faculty (LHS), Gunnar Nilsson Cancer Foundation (LHS), Skåne University Hospital Foundation (LHS), BioCARE Research Program (LHS), King Gustav Vth Jubilee Foundation (LHS), and the Krapperup Foundation (LHS).

## Author contributions

CB, CW, and LHS conceived the study. CB, SG, CW, JV-C, JH, AMG, YC, NL, and LHS analyzed data. JV-C, CH, JH, AE, CL, NL, MM, LR, ÅB, and LHS established the SCAN-B initiative. CB, CW, and LHS established the DNA-seq analyses. CB, SG, and LHS established the RNA-seq mutation calling pipeline and filters. JV-C, CH, JH, AE, CL, NL, MM, LR, ÅB, and LHS provided clinical information. LHS supervised the project, and CB and LHS wrote the report with assistance from all authors. All authors discussed, critically revised, and approved the final version of the report for publication.

## Conflict of interest

CB, SG, AMG, YC, and LHS are shareholders and/or employees of SAGA Diagnostics AB. LHS has received honorarium from Novartis and Boehringer-Ingelheim. All remaining authors have declared no conflicts of interest.

# References

Alsafadi S, Houy A, Battistella A, Popova T, Wassef M, Henry E, Tirode F, Constantinou A, Piperno-Neumann S, Roman-Roman S *et al* (2016) Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun* 7: 10615

Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kähäri AK, Lundin P, Che H *et al* (2017) SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet* 25: 1253 – 1260

André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, Iwata H, Conte P, Mayer IA, Kaufman B *et al* (2019) Alpelisib for PIK3CA-mutated, hormone receptor-positive advanced breast cancer. *N Engl J Med* 380: 1929 – 1940

Avivar-Valderas A, McEwen R, Taheri-Ghahfarokhi A, Carnevalli LS, Hardaker EL, Maresca M, Hudson K, Harrington EA, Cruzalegui F (2018) Functional significance of co-occurring mutations in PIK3CA and MAP3K1 in breast cancer. *Oncotarget* 9: 21444 – 21458

Bader AG, Kang S, Vogt PK (2006) Cancer-specific mutations in PIK3CA are oncogenic *in vivo*. *Proc Natl Acad Sci USA* 103: 1475 – 1479

Ben-Baruch E, Bose R, Kavuri SM, Ma CX, Ellis MJ (2015) HER2-mutated breast cancer responds to treatment with single-agent neratinib, a second-generation HER2/EGFR tyrosine kinase inhibitor. *J Natl Compr Canc Netw* 13: 1061 – 1064

Blokzijl F, Janssen R, Boxtel RV, Cuppen E (2018) MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* 10: 33

Bose R, Kavuri SM, Searleman AC, Shen W, Shen D, Koboldt DC, Monsey J, Goel N, Aronson AB, Li S *et al* (2013) Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discov* 3: 224 – 237

Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, Olivier M (2016) TP53 variations in human cancers: new lessons from the IARC TP53 database and genomics data. *Hum Mutat* 37: 865 – 876

Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, Malina J, Chen Y, Bendahl P-O, Manjer J *et al* (2018) Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter sweden cancerome analysis network—breast initiative. *JCO Precis Oncol* 2: 1 – 18

Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW (2016) Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 17: 257 – 271

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E *et al* (2012) The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2: 401 – 404

Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN *et al* (2015) Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molcular oncology. *J Mol Diagn* 17: 251 – 264

Cieślik M, Chinnaiyan AM (2018) Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet* 19: 93 – 109

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X (2012a) Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* 3: 35

Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012b) 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

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C *et al* (2015) Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* 163: 506 – 519

Cocco E, Carmona FJ, Razavi P, Won HH, Cai Y, Rossi V, Chan C, Cownie J, Soong J, Toska E *et al* (2018) Neratinib is effective in breast tumors bearing both amplification and mutation of ERBB2 (HER2). *Sci Signal* 11: eaat9773

Conroy JM, Pabla S, Glenn ST, Nesline M, Burgher B, Lenzo FL, Papanicolau-Sengos A, Gardner M, Morrison C (2019) Tumor mutational burden (TMB): assessment of inter and intra-tumor heterogeneity. *J Clin Oncol*, 37(suppl 8; abstr 27): 27

Cotto KC, Griffith OL, Wollam A, Wagner AH, Griffith M, Feng Y-Y, Spies G, Coffman AC, Spies NC, Kiwala S (2017) DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res* 46: D1068 – D1073

Dihge L, Vallon-Christersson J, Hegardt C, Saal LH, Häkkinen J, Larsson C, Ehinger A, Loman N, Malmberg M, Bendahl P-O *et al* (2019) Prediction of lymph node metastasis in breast cancer by gene expression and clinicopathological models: development and validation within a population based cohort. *Clin Cancer Res* 25: 6368 – 6381

Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M *et al* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst* 6: 271 – 281

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B *et al* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res* 46: D649 – D655

Faust GG, Hall IM (2014) SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 30: 2503 – 2505

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S *et al* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43: D805 – D811

Förnvik D, Aaltonen KE, Chen Y, George AM, Brueffer C, Rigo R, Loman N, Saal LH, Rydén L (2019) Detection of circulating tumor cells and circulating tumor DNA before and after mammographic breast compression in a cohort of breast cancer patients scheduled for neoadjuvant treatment. *Breast Cancer Res Treat* 177: 447 – 455

Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, Cheang M, Osin P, Nerurkar A, Kozarewa I *et al* (2015) Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med* 7: 302ra133

Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, Howard TP, Takeda DY, Ly SH, Kim E *et al* (2018) Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet* 50: 1381 – 1387

Gonzalez-Perez A, Perez-llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* 10: 1081 – 1082

Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, Stephens PJ, Daniels GA, Kurzrock R (2017) Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol Cancer Ther* 16: 2598 – 2608

Grambsch PM, Therneau TM (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81: 515 – 526

Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L *et al* (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 49: 170 – 174

Griffith OL, Spies NC, Anurag M, Griffith M, Luo J, Tu D, Yeo B, Kunisaki J, Miller CA, Krysiak K *et al* (2018) The prognostic effects of somatic mutations in ER-positive breast cancer. *Nat Commun* 9: 3476

Grüning B, Dale R, Sjödin A, Rowe J, Chapman BA, Tomkins-Tinch CH, Valieris R, Team TB, Köster J (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 15: 475 – 476

Guo Y, Zhao S, Sheng Q, Samuels DC, Shyr Y (2017) The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *BMC Genom* 18: 690

Habib JG, O'Shaughnessy JA (2016) The hedgehog pathway in triple-negative breast cancer. *Cancer Med* 5: 2989 – 3006

Hellmann M. D., Ciuleanu T.-E., Pluzanski A., Lee J. S., Otterson G. A., Audigier-Valette C., Minenza E., Linardou H., Burgers S., Salman P. *et al* (2018) Nivolumab plus Ipilimumab in lung cancer with a high tumor mutational burden. *N Engl J Med* 378: 2093 – 2104

Hisamatsu Y, Tokunaga E, Yamashita N, Akiyoshi S, Okada S, Nakashima Y, Aishima S, Morita M, Kakeji Y, Maehara Y (2012) Impact of FOXA1 expression on the prognosis of patients with hormone receptor-positive breast cancer. *Ann Surg Oncol* 19: 1145 – 1152

Hofmann AL, Behr J, Singer J, Kuipers J, Beisel C, Schraml P, Moch H, Beerenwinkel N (2017) Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers. *BMC Bioinformatics* 18: 1 – 15

Holstege H, Horlings HM, Velds A, Langerød A, Børresen-Dale AL, van de Vijver MJ, Nederlof PM, Jonkers J (2010) BRCA1-mutated and basal-like breast cancers have similar aCGH profiles and a high incidence of protein truncating TP53 mutations. *BMC Cancer* 10: 654

Horvath A, Pakala SB, Mudvari P, Reddy SDN, Ohshiro K, Casimiro S, Pires R, Fuqua SA, Toi M, Costa L *et al* (2013) Novel insights into breast cancer genetic variance through RNA sequencing. *Sci Rep* 3: 2256

Hyman DM, Smyth LM, Donoghue MTA, Westin SN, Bedard PL, Emma J, Bando H, El-khoueiry AB, Mita A, Schellens JHM *et al* (2017) AKT inhibition in solid tumors with AKT1 mutations. *J Clin Oncol* 35: 2251 – 2259

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R *et al* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res* 48: D498 – D503

Juric D, Rodon J, Tabernero J, Janku F, Burris HA, Schellens JH, Middleton MR, Berlin J, Schuler M, Gil-Martin M *et al* (2018) Phosphatidylinositol 3-kinase $\alpha$–selective inhibition with alpelisib (BYL719) in PIK3CA -altered solid tumors: results from the first-in-human study. *J Clin Oncol* 36: 1291 – 1299

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP *et al* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581: 434 – 443

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12: 357 – 360

Kiran A, Baranov PV (2010) DARNED: a database of RNA editing in humans. *Bioinformatics* 26: 1772 – 1776

Köster J, Rahmann S (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* 28: 2520 – 2522

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

Lauss M, Donia M, Harbst K, Andersen R, Mitra S, Rosengren F, Salim M, VallonChristersson J, Törngren T, Kvist A *et al* (2017) Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. *Nat Commun* 8: 1738

Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier LD, Neale B, MacArthur D (2018) A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* 15: 595 – 597

Locatelli M, Curigliano G (2017) Notch inhibitors and their role in the treatment of triple negative breast cancer: promises and failures. *Curr Opin Oncol* 29: 411 – 427

Lundgren C, Bendahl P-O, Borg Å, Ehinger A, Hegardt C, Larsson C, Loman N, Malmberg M, Olofsson H, Saal LH *et al* (2019) Agreement between molecular subtyping and surrogate subtype classification: a contemporary population-based study of ER-positive/HER2-negative primary breast cancer. *Breast Cancer Res Treat* 178: 459 – 467

Ma CX, Bose R, Gao F, Freedman RA, Telli ML, Kimmick G, Winer E, Naughton M, Goetz MP, Russell C *et al* (2017) Neratinib efficacy and circulating tumor DNA detection of HER2 mutations in HER2 nonamplified metastatic breast cancer. *Clin Cancer Res* 23: 5687 – 5695

Ma C, Shao M, Kingsford C (2018) SQUID: transcriptomic structural variation detection from RNA-seq. *Genome Biol* 19: 52

Maguire SL, Leonidou A, Wai P, Marchiò C, Ng CK, Sapino A, Salomon A-V, ReisFilho JS, Weigelt B, Natrajan RC (2015) SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J Pathol* 235: 571 – 580

Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, Skov L, Belling K, Theil Have C, Izarzugaza JM *et al* (2017) Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. *Nature* 548: 87 – 91

Mendoza-Villanueva D, Deng W, Lopez-Camacho C, Shore P (2010) The Runx transcriptional co-activator, CBFbeta, is essential for invasion of breast cancer cells. *Mol Cancer* 9: 171

Murray EM, Cherian MA, Ma CX, Bose R (2018) HER2 activating mutations in estrogen receptor positive breast cancer. *Curr Breast Cancer Rep* 10: 41 – 47

Nayar U, Cohen O, Kapstad C, Cuoco MS, Waks AG, Wander SA, Painter C, Freeman S, Persky NS, Marini L *et al* (2018) Acquired HER2 mutations in ER+ metastatic breast cancer confer resistance to estrogen receptor–directed therapies. *Nat Genet* 51: 207 – 216

Pahuja KB, Nguyen TT, Jaiswal BS, Prabhash K, Thaker TM, Senger K, Chaudhuri S, Kljavin NM, Antony A, Phalke S *et al* (2018) Actionable activating oncogenic ERRB2/HER2 transmembrane and juxtamembrane domain mutations. *Cancer Cell* 34: 792 – 806

Panda A, Betigeri A, Subramanian K, Ross JS, Pavlick DC, Ali S, Markowski P, Silk A, Kaufman HL, Lattime E *et al* (2017) Identifying a clinically applicable mutational burden threshold as a potential biomarker of response to immune checkpoint therapy in solid tumors. *JCO Precis Oncol* 1: 1 – 13

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37: e123

Pedersen BS, Layer RM, Quinlan AR, Li H, Wang K, Li M, Hakonarson H, McLaren W, Pritchard B, Rios D *et al* (2016) Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol* 17: 118

Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut S-J *et al* (2016) The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun* 7: 11479

Persson H, Søkilde R, Häkkinen J, Pirona AC, Vallon-Christersson J, Kvist A, Mertens F, Borg Å, Mitelman F, Höglund M *et al* (2017) Frequent miRNA-convergent fusion gene events in breast cancer. *Nat Commun* 8: 788

Picardi E, Erchia AMD, Giudice CL, Pesole G (2017) REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res* 45: D750 – D757

Piskol R, Ramaswami G, Li JB (2013) Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet* 93: 641 – 651

Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, Haussler D (2014) RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS ONE* 9: e111516

Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB (2013) Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods* 10: 128 – 132

Ramaswami G, Li JB (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* 42: D109 – D113

Robinson DR, Wu YM, Vats P, Su F, Lonigro RJ, Cao X, Kalyana-Sundaram S, Wang R, Ning Y, Hodges L *et al* (2013) Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat Genet* 45: 1446 – 1451

Roepman P, Horlings HM, Krijgsman O, Kok M, Bueno-de-Mesquita JM, Bender R, Linn SC, Glas AM, van de Vijver MJ (2009) Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. *Clin Cancer Res* 15: 7003 – 7011

Ross JS, Gay LM, Wang K, Ali SM, Chumsri S, Elvin JA, Bose R, Vergilio JA, Suh J, Yelensky R *et al* (2016) Nonamplification ERBB2 genomic alterations in 5605 cases of recurrent and metastatic breast cancer: an emerging opportunity for anti-HER2 targeted therapies. *Cancer* 122: 2654 – 2662

Rydén L, Loman N, Larsson C, Hegardt C, Vallon-Christersson J, Malmberg M, Lindman H, Ehinger A, Saal LH, Borg Å (2018) Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. *Br J Surg* 105: e158 – e168

Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, Yu JS, Malmström PO, Mansukhani M, Enoksson J *et al* (2005) PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Res* 65: 2554 – 2559

Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, Koujak S, Ferrando AA, Malmstrom P, Memeo L *et al* (2007) Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc Natl Acad Sci USA* 104: 7564 – 7569

Saal LH, Gruvberger-Saal SK, Persson C, Lövgren K, Jumppanen M, Staaf J, Jönsson G, Pires MM, Maurer M, Holm K *et al* (2008) Recurrent gross mutations of the PTEN tumor suppressor gene in breast cancers with deficient DSB repair. *Nat Genet* 40: 102 – 107

Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, Brueffer C, Tang M-HE, Reuterswärd C, Schulz R *et al* (2015) The

Sweden Cancerome Analysis Network Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med* 7: 20

Schmid P, Adams S, Rugo HS, Schneeweiss A, Barrios C, Iwata H, Diéras V, Hegg R, Im S, Wright GS *et al* (2018) Atezolizumab and Nab-paclitaxel in advanced triple-negative breast cancer. *N Engl J Med* 379: 2108 – 2121

Severson TM, Peeters J, Majewski I, Michaut M, Bosma A, Schouten PC, Chin SF, Pereira B, Goldgraben MA, Bismeijer T *et al* (2015) BRCA1-like signature in triple negative breast cancer: molecular and clinical characterization reveals subgroups with therapeutic potential. *Mol Oncol* 9: 1528 – 1538

Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G *et al* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 486: 395 – 399

She QB, Gruvberger-Saal SK, Maurer M, Chen Y, Jumppanen M, Su T, Dendy M, Lau YKI, Memeo L, Horlings HM *et al* (2016) Integrated molecular pathway analysis informs a synergistic combination therapy targeting PTEN/PI3K and EGFR pathways for basal-like breast cancer. *BMC Cancer* 16: 587

Sherry ST, Ward M, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308 – 311

Shi W, Ng CK, Lim RS, Jiang T, Kumar S, Li X, Wali VB, Piscuoglio S, Gerstein MB, Chagpar AB *et al* (2018) Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep* 25: 1446 – 1457

Siegel MB, He X, Hoadley KA, Hoyle A, Pearce JB, Garrett AL, Kumar S, Moylan VJ, Brady CM, Swearingen AEDV *et al* (2018) Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *J Clin Invest* 128: 1 – 13

Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, Griffith M (2016) GenVisR: genomic visualizations in R. *Bioinformatics* 32: 3012 – 3014

Søkilde R, Persson H, Ehinger A, Pirona AC, Fernö M, Hegardt C, Larsson C, Loman N, Malmberg M, Rydén L *et al* (2019) Refinement of breast cancer molecular classification by miRNA expression profiles. *BMC Genom* 20: 1 – 12

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* 18: 696 – 705

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B *et al* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 98: 262 – 272

Sun J, De Marinis Y, Osmark P, Singh P, Bagge A, Valtat B, Vikman P, Spégel P, Mulder H (2016) Discriminative prediction of A-To-I RNA editing events from DNA sequence. *PLoS ONE* 11: 1 – 18

Takaku M, Grimm SA, Roberts JD, Chrysovergis K, Bennett BD, Myers P, Perera L, Tucker CJ, Perou CM, Wade PA (2018) GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. *Nat Commun* 9: 1 – 14

Talevich E, Shain AH (2018) CNVkit-RNA: copy number inference from RNA-sequencing data. *bioRxiv* https://doi.org/10.1101/408534 [PREPRINT]

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31: 2032 – 2034

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

Thomas A, Routh ED, Pullikuth A, Jin G, Su J, Chou JW, Hoadley KA, Print C, Knowlton N, Black MA *et al* (2018) Tumor mutational burden is a determinant of immunemediated survival in breast cancer. *OncoImmunology* 7: e1490854

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

Vallon-Christersson J, Häkkinen J, Hegardt C, Saal LH, Larsson C, Ehinger A, Lindman H, Olofsson H, Sjöblom T, Wärnberg F *et al* (2019) Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. *Sci Rep* 9: 12184

Vitting-Seerup K, Sandelin A (2017) The landscape of isoform switches in human cancers. *Mol Cancer Res* 15: 1206 – 1220

Weidle UH, Maisel D, Eick D (2011) Synthetic lethality-based targets for discovery of new cancer therapeutics. *Cancer Genomics Proteomics* 8: 159 – 171

Wen W, Chen WS, Xiao N, Bender R, Ghazalpour A, Tan Z, Swensen J, Millis SZ, Basu G, Gatalica Z *et al* (2015) Mutations in the kinase domain of the HER2/ERBB2 gene identified in a wide variety of human cancers. *J Mol Diagn* 17: 487 – 495

Wilkerson MD, Cabanski CR, Sun W, Hoadley KA, Walter V, Mose LE, Troester MA, Hammerman PS, Parker JS, Perou CM *et al* (2014) Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res* 42: e107

Winter C, Nilsson M, Olsson E, George A, Chen Y, Kvist A, Törngren T, Vallon-Christersson J, Hegardt C, Häkkinen J *et al* (2016) Targeted sequencing of BRCA1 and BRCA2 across a large unselected breast cancer cohort suggests one-third of mutations are somatic. *Ann Oncol* 27: 1532 – 1538

Xu J, Guo X, Jing M, Sun T (2018) Prediction of tumor mutation burden in breast cancer based on the expression of ER, PR, HER-2, and Ki-67. *Onco Targets Ther* 11: 2269 – 2275

Zacharakis N, Chinnasamy H, Black M, Xu H, Lu Y-C, Zheng Z, Pasetto A, Langhan M, Shelton T, Prickett T *et al* (2018) Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat Med* 24: 724 – 730

Zardavas D, te Marvelde L, Milne RL, Fumagalli D, Fountzilas G, Kotoula V, Razis E, Papaxoinis G, Joensuu H, Moynahan ME *et al* (2018) Tumor PIK3CA genotype and prognosis in early-stage breast cancer: a pooled analysis of individual patient data. *J Clin Oncol* 36: 981 – 990

Zhang HY, Liang F, Jia ZL, Song ST, Jiang ZF (2013) PTEN mutation, methylation and expression in breast cancer patients. *Oncol Lett* 6: 161 – 168

Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30: 1006 – 1007

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N *et al* (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 3: 160025