


5-2018

# INVESTIGATING INVASION IN DUCTAL CARCINOMA IN SITU WITH TOPOGRAPHICAL SINGLE CELL GENOME SEQUENCING

Anna Casasent

Anna Casasent

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations](https://digitalcommons.library.tmc.edu/utgsbs_dissertations)

 Part of the [Cancer Biology Commons](#), [Genomics Commons](#), and the [Medicine and Health Sciences Commons](#)

---

## Recommended Citation

Casasent, Anna and Casasent, Anna, "INVESTIGATING INVASION IN DUCTAL CARCINOMA IN SITU WITH TOPOGRAPHICAL SINGLE CELL GENOME SEQUENCING" (2018). *UT GSBS Dissertations and Theses (Open Access)*. 842.  
[https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations/842](https://digitalcommons.library.tmc.edu/utgsbs_dissertations/842)

This Dissertation (PhD) is brought to you for free and open access by the Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in UT GSBS Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [laurel.sanders@library.tmc.edu](mailto:laurel.sanders@library.tmc.edu).

INVESTIGATING INVASION IN DUCTAL CARCINOMA IN SITU WITH  
TOPOGRAPHICAL SINGLE CELL GENOME SEQUENCING

by

*Anna Kristina Casasent, B.S.*

APPROVED:

---

Nicholas Navin, Ph.D.  
Advisory Professor

---

Keith Baggerly, Ph.D.

---

Mary Edgerton, M.D., Ph.D.

---

Vicki Huff, Ph.D.

---

Ralf Krahe, Ph.D.

APPROVED:

---

Dean, The University of Texas  
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

INVESTIGATING INVASION IN DUCTAL CARCINOMA IN SITU WITH  
TOPOGRAPHICAL SINGLE CELL GENOME SEQUENCING

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Anna Kristina Casasent, B.S.

Houston, Texas

May, 2018

## COPYRIGHT

© Copyright by Anna Kristina Casasent 2017-2018

All Rights Reserved



## DEDICATION

To my science father (you know who you are),  
and my amazing husband for your endless support and love.

## ACKNOWLEDGEMENTS

This dissertation is a culmination, not just of scientific endeavors for the last five to six years, but also of the support I have received both within and outside the lab. I will always be shaped by the years I spent under Dr. Nicholas Navin's mentorship. Dr. Navin gave me the opportunity to study in his lab and found time for numerous meetings about my project even as the lab expanded from a handful to over a dozen members. I am grateful to the Navin lab for help with all my ventures. It has been wonderful to be part of a group and watch us be transformed from individual students to a truly collaborative group, where biologists, statisticians, computer scientists, and others can come together and learn more about bioinformatics and cancer genetics.

I could not have completed this project as so quickly, without the help of Aislyn Schalck, Ruli Gao, Emi Sei, and my wonderful husband Tod Casasent. Aislyn deserves special mention for finding the tanglegram method to combine the spatial and genetic information from my project, and for spending countless hours helping me map cells, mark ducts, and make (and remake) tanglegrams based on new data. Dr. Ruli Gao helped with data analysis, especially the exome data. I will always appreciate that Dr. Gao took time to wade through my data, even when her schedule was also daunting. Dr. Emi Sei deserves credit for organizing the lab and helping out make the final set of libraries on a weekend, so that we could have then sequencing before we resubmitted the paper with revisions. My husband who drove me into work on weekends and stayed to help in the lab, who proofread, and who encouraged and believed in me, I love and appreciate.

I wish to thank my summer students that worked on this project Annie Long and Will Pangburn, teaching you reminded me that science is fun. Annie deserves great thanks for the two summers spent working on this project, helping amplify

hundreds of single cells and making hundreds more NGS libraries, and for the 9 months as a lab employee, helping out when she could. Will deserves special thanks for helping with the last final sprint to the finish line, for learning quickly and well, so that I could write, and helping polish image maps, even when his summer project was something else.

I have to thank Marco Leung for being the quiet center that held the lab together in the beginning. Alex Davis, I thank for always being ready to help with a statistics question and for the many "RedCap" lunches. Charissa Kim deserves credit for, no matter how busy she was, always having time for a few words. I thank Tessa Tsai who was always there with a hug and a pipette. Special thanks go to Jake Leighton and Darlan Conterno Minussi who keep me sane at work, a non-trivial task, with random jokes or silence, depending on what I needed. And I thank the rest of the lab, who was there during the rush with smiles and presentation feedback.

I wish to thank my advisory committee that has been with me throughout my graduate school process (Dr. Krahe, Dr. Huff, Dr. Edgerton, and Dr. Baggerly). I especially thank Dr. Edgerton who spent many hours explaining pathology to me and looking at numerous slides. Dr. Baggerly, who I worked with for over a year before I started graduate school, gets special thanks for calm, reasoned advice and details about data analysis and scientific rigor. It was time working with Dr. Baggerly that lead me to set course for a graduate degree in biomedical sciences. I must also thank Dr. Krahe, who always teaches me to look very carefully at my writing, and make sure I am truly saying precisely what I mean. I am grateful for Dr. Navin, Dr. Baggerly, and Dr. Edgerton taking the time to write numerous letters of recommendation for me for fellowships and postdoctoral applications.

I thank the people in my home life. I thank my sister and my (again) husband. I thank my sister, who is more than a sister but truly a combination sister, mother,

and best friend, I will be forever grateful that you are in my life. She took time to looked over my writing, even if she did not feel she knew the science. I thank my husband, who drove me in to the lab every weekday and some weekends, and then stayed at work with me on the weekend, looking over my scripts, and even trying his hand at sonication and running gels. My wonderful husband deserves so much for always being a support in my life, even when I felt too tired to help at home, for reminding me to eat and making me sleep, so that I could get up in the morning and work some more, and for supporting me in everything I want to do and believing in me so much that I learned to believe in myself.

Last, I thank my best friends, that I could ever imagine having that I found at graduate school, Beata Lerman and Lexy Plumer, who know how to make me laugh and make me think. Both of you amaze me. Beata helped build my confidence in so many ways, listened to me rant, and made me so proud to watch her take charge of her own life and make tough decisions look easy. Lexy (you fox) somehow snuck into being my best friend and the person I could tell anything. I love how bold she is and how she reminds me that being myself is the best.

I also thank my friendly editors Tod Casasent (who was always looking over my writing and helping to keep me on schedule), Lindsey Minter, and Julia Unruh.

# INVESTIGATING INVASION IN DUCTAL CARCINOMA IN SITU WITH TOPOGRAPHICAL SINGLE CELL GENOME SEQUENCING

Anna Kristina Casasent, B.S.

Advisory Professor: Nicholas Navin, Ph.D.

Synchronous Ductal Carcinoma *in situ* (DCIS-IDC) is an early stage breast cancer invasion in which it is possible to delineate genomic evolution during invasion because of the presence of both *in situ* and invasive regions within the same sample. While laser capture microdissection studies of DCIS-IDC examined the relationship between the paired *in situ* (DCIS) and invasive (IDC) regions, these studies were either confounded by bulk tissue or limited to a small set of genes or markers.

To overcome these challenges, we developed Topographic Single Cell Sequencing (TSCS), which combines laser-catapulting with single cell DNA sequencing to measure genomic copy number profiles from single tumor cells while preserving their spatial context. We applied TSCS to sequence 1,293 single cells from 10 synchronous DCIS patients. We also applied deep-exome sequencing to the *in situ*, invasive and normal tissues for the DCIS-IDC patients.

Previous bulk tissue studies had produced several conflicting models of tumor evolution. Our data support a multiclonal invasion model, in which genome evolution occurs within the ducts and gives rise to multiple subclones that escape the ducts into the adjacent tissues to establish the invasive carcinomas. In summary, we have developed a novel method for single cell DNA sequencing, which preserves spatial context, and applied this method to understand clonal evolution during the transition between carcinoma *in situ* to invasive ductal carcinoma.

## TABLE OF CONTENTS

Approval Sheet.....	i
Title Page.....	ii
Copyright.....	iii
Dedication .....	iv
Acknowledgements .....	v
Abstract.....	viii
Table of Contents .....	ix
List of Illustrations.....	xii
List of Tables.....	xiv
Abbreviations.....	xv
Text.....	1
1 Introduction.....	1
1.1 Breast Cancer.....	4
1.1.1 Model of Breast Cancer Progression.....	6
1.1.2 Pathology.....	8
1.1.3 Receptor Status .....	9
1.1.4 Nuclear Grades.....	10
1.1.5 Stages .....	11
1.1.6 DCIS Survival .....	12
1.2 IDC Genomics .....	13
1.2.1 IDC Aneuploidy.....	13
1.2.2 TCGA and Intertumor Heterogeneity.....	14
1.2.3 IDC Intratumor Heterogeneity .....	15
1.2.4 Synchronous DCIS-IDC .....	18
1.3 Single Cell Sequencing.....	18
1.3.1 SCS Beginnings.....	18
1.3.2 SCS Challenges .....	19
1.4 Laser Capture Microdissection Methods .....	21
1.4.1 LCM Sample Purity.....	21
1.4.2 LCM Type Selection.....	22
1.4.3 LCM Spatially Resolved Sequencing .....	22
1.5 Models of Genomic Lineage and Invasion .....	23
1.5.1 Independent Lineage Model.....	26

1.5.2	Dependent Genomic Lineage Model .....	27
1.6	Dissertation Summary .....	32
1.6.1	Spatially-Resolved Single Cell DNA Sequencing .....	32
1.6.2	Intratumor Heterogenetic during Invasion in Breast Cancer .....	33
2	Materials and Methods .....	34
2.1	Sample Selection .....	34
2.2	Human Samples Description .....	35
2.3	Single Cell Copy Number Protocol .....	40
2.3.1	Single Cell Isolation .....	40
2.3.2	Single Cell Processing .....	46
2.3.3	NGS Library Preparation .....	48
2.4	Single Cell Copy Number Data Analysis .....	49
2.4.1	Single Cell Copy Number Data Processing .....	49
2.4.2	Data Quality and Filtering .....	50
2.4.3	Clustering .....	50
2.4.4	Subclone Analysis .....	51
2.4.5	Topographical Analysis .....	53
2.5	Regional Exome Protocol .....	55
2.5.1	Exome Laser Capture Microdissection .....	56
2.5.2	Exome DNA Isolation .....	56
2.5.3	Exome Capture .....	57
2.6	Exome Regional Data Analysis .....	57
2.6.1	Exome Data Processing .....	57
2.6.2	Exome Regional Data Quality and Filtering .....	58
2.6.3	Exome Regional Mutation Calls .....	58
2.6.4	Exome Regional Amplicon Validation .....	58
3	Studying Synchronous DCIS using TSCS .....	60
3.1	Introduction .....	60
3.1.1	Rationale of Synchronous DCIS-IDC .....	63
3.1.2	Rationale of need for TSCS .....	64
3.2	Results .....	65
3.2.1	Copy Number Evolution During Invasion Polyclonal Tumors .....	65
3.2.2	Copy Number Evolution During Invasion Monoclonal Tumors .....	114
3.2.3	Copy Number Evolution Summary .....	133
3.2.4	Spatial Topography and Clonal Copy Number Genotypes .....	135

3.2.5	Regional Exome .....	136
3.3	Study Limitations.....	147
3.4	Conclusions .....	149
3.4.1	Using Topographical Single Cell Sequencing (TSCS).....	149
3.4.2	Single Cell of Origin .....	150
3.4.3	Multiclonal Invasion.....	151
3.4.4	Intraductal Punctuated Evolution.....	152
4	Discussion and Future Directions.....	158
4.1	Single Cells and Topographical Information .....	158
4.1.1	Technical Barrier: Scalability .....	159
4.1.2	Technical Barrier: DNA Mutations .....	161
4.1.3	Technical Barrier: TSCS and RNA .....	163
4.1.4	Technical Barrier: TSCS and Genome and Transcriptome Protocols.....	165
4.1.5	Technical Barrier: Spatial Genomics and Clinical Tools .....	166
4.2	Future Research Directions .....	168
4.2.1	ITH in Pure DCIS and Earlier Cancers .....	169
4.2.2	Examining Spatial and DNA Mutations.....	170
4.2.3	Profiling Geographic Heterogeneity .....	171
4.3	Closing Remarks.....	172
4.3.1	Synchronous DCIS-IDC .....	172
4.3.2	Technology .....	173
	Bibliography .....	174
	Vita.....	193



## LIST OF ILLUSTRATIONS

Figure 1 Models of Progression .....	3
Figure 2 Sequential Progression from DH to IDC .....	5
Figure 3 Examples of Pathology DCIS and DCIS-IDC .....	7
Figure 4 Models of Invasion.....	25
Figure 5 Estimates of Number of Samples Needs for Longitudinal Studies .....	37
Figure 6 Timeline of TSCS Protocol .....	41
Figure 7 3D Slide Stacking .....	43
Figure 8 UV Cutting Energy.....	45
Figure 9 TSCS Protocol.....	62
Figure 10 DC4 Copy Number Alteration Heatmap .....	68
Figure 11 DC4 Saturation Curve.....	69
Figure 12 DC4 MDS .....	70
Figure 13 DC4 TimeScape .....	71
Figure 14 DC4 Image Maps.....	72
Figure 15 DC4 Tanglegram .....	73
Figure 16 DC13 Copy Number Alteration Heatmap .....	76
Figure 17 DC13 Saturation Curve.....	77
Figure 18 DC13 MDS .....	78
Figure 19 DC13 TimeScape .....	79
Figure 20 DC13 Image Maps.....	80
Figure 21 DC13 Tanglegram .....	81
Figure 22 DC14 Copy Number Alteration Heatmap .....	83
Figure 23 DC14 Saturation Curve.....	84
Figure 24 DC14 MDS .....	85
Figure 25 DC14 TimeScape .....	86
Figure 26 DC14 Image Maps.....	87
Figure 27 DC14 Tanglegram .....	88
Figure 28 DC16 Copy Number Alteration Heatmap .....	92
Figure 29 DC16 Saturation Curve.....	93
Figure 30 DC16 MDS .....	94
Figure 31 DC16 TimeScape .....	95
Figure 32 DC16 Image Maps.....	96
Figure 33 DC16 Tanglegram .....	97
Figure 34 DC18 Copy Number Alteration Heatmap .....	100
Figure 35 DC18 Saturation Curve.....	101
Figure 36 DC18 MDS .....	102
Figure 37 DC18 TimeScape .....	103
Figure 38 DC18 Image Maps.....	104
Figure 39 DC18 Tanglegram .....	105
Figure 40 DC20 Copy Number Alteration Heatmap .....	108
Figure 41 DC20 Saturation Curve.....	109
Figure 42 DC20 MDS .....	110
Figure 43 DC20 TimeScape .....	111
Figure 44 DC20 Image Maps.....	112
Figure 45 DC20 Tanglegram .....	113
Figure 46 DC6 Copy Number Alteration Heatmap .....	115
Figure 47 DC6 Saturation Curve.....	116
Figure 48 DC6 MDS .....	117
Figure 49 DC6 Image Maps.....	118

Figure 50 DC12 Copy Number Alteration Heatmap .....	120
Figure 51 DC12 Saturation Curve.....	121
Figure 52 DC12 MDS .....	122
Figure 53 DC12 Image Maps.....	123
Figure 54 DC17 Copy Number Alteration Heatmap .....	125
Figure 55 DC17 MDS .....	126
Figure 56 DC17 Image Maps.....	127
Figure 57 DC19 Copy Number Alteration Heatmap .....	129
Figure 58 DC19 Saturation Curve.....	130
Figure 59 DC19 MDS .....	131
Figure 60 DC19 Image Maps.....	132
Figure 61 Copy Number Summary .....	134
Figure 62 Regional Microdissection .....	137
Figure 63 Regional Exome Oncomap .....	138
Figure 64 Regional Amplicon Validation .....	139
Figure 65 Regional Frequency Changes .....	144
Figure 66 Segmentation Data.....	155

## LIST OF TABLES

Table 1 aCGH and NGS DCIS Papers .....	17
Table 2 Clinical Information .....	38
Table 3 Exome Coverage .....	39
Table 4 Regional Invasive-Specific Mutations.....	140
Table 5 Regional Deep SNVs Genomics and Reads .....	141
Table 6 Regional DeepSNVs Result Details .....	142
Table 7 Regional Pre-Existing Mutations .....	145

## ABBREVIATIONS

aCGH – array copy genomic hybridization

ADH – atypical ductal hyperplasia

BaristaSeq - barcode *in situ* targeted sequencing

bp – Basepair

CBS – Circular Binary Segmentation

CIN – Chromosomal Instability

CN – Copy Number

CNA – Copy Number Aberration

DCIS – Ductal Carcinoma *in situ*

DH – Ductal Hyperplasia

DNA – Deoxyribonucleic Acid

DNA-SCS – Single Cell Genome Sequencing

DOP-PCR – Degenerate Oligonucleotide Primed Polymerase Chain Reaction

ER – Estrogen Receptor

FISH – Fluorescence *in situ* Hybridization

FISSEQ – Fluorescent *in situ* Sequencing

H&E – Hematoxylin and eosin stain

HER2 – Human Epidermal Growth Factor Receptor-2

IDC – Invasive Ductal Carcinoma

IHC – Immunohistochemistry

IR – Infrared

ITH – Intratumor Heterogeneity

LCM – Laser Capture Microdissection

LOH – Loss of Heterozygosity

MALBAC - Multiple Annealing and Looping Based Amplification Cycles

MDA-PCR - multiple strand displacement amplification polymerase chain reaction

MDS – Multidimensional Scaling

NGS – Next Generation Sequencing

PCR – Polymerase Chain Reaction

PR – Progesterone Receptor

RNA – Ribonucleic Acid

SCS – Single Cell Sequencing

SNV – Single Nucleotide Variant

STAR-FISH – Specific-to-allele PCR–FISH

TNBC – Triple Negative Breast Cancer

TSCS – Topographical Single Cell Sequencing

UV – Ultraviolet light

WGA – Whole Genome Amplification

## 1 Introduction

Portions of this introduction are adapted from the review paper "Genome evolution in ductal carcinoma *in situ*: invasion of the clones" published in the Journal of Pathology in 2017, by Casasent, Edgerton, and Navin<sup>2</sup>. Figures and text from this paper have been reused or modified under the journal's academic copyright license with permission from John Wiley & Sons, Ltd for the Pathological Society of Great Britain and Ireland. This is an expanded version of the review paper that focuses on spatially resolved single cell sequencing and the progression of breast cancer.

Understanding how tumors progress is vital to refining treatment in cancer. Since it is neither ethical nor feasible to sample patients longitudinally during tumor progression, it is imperative to have methods that can deduce as much as possible from a single time point and single tumor sample. The more data we can gather from a single sample the better chance we have at reconstructing tumor evolution to generate more extensive knowledge of tumor development and therefore hopefully improve treatments.

While we are limited by the lack of longitudinal data, it is still possible (1) to infer progression from a cohort of single cells<sup>4</sup> and (2) to gather regional information from surgically resected tumors<sup>5, 6</sup>. Reconstructing or inferring tumor progression from a tumor sample is possible because mutations accumulate as cells divide, leaving an imprint of tumor evolution<sup>4</sup>. While bulk regional information has been gathered during the dissection of tumors through tumor macrodissection and regional microdissection, the local spatial information is still lost<sup>6</sup>.

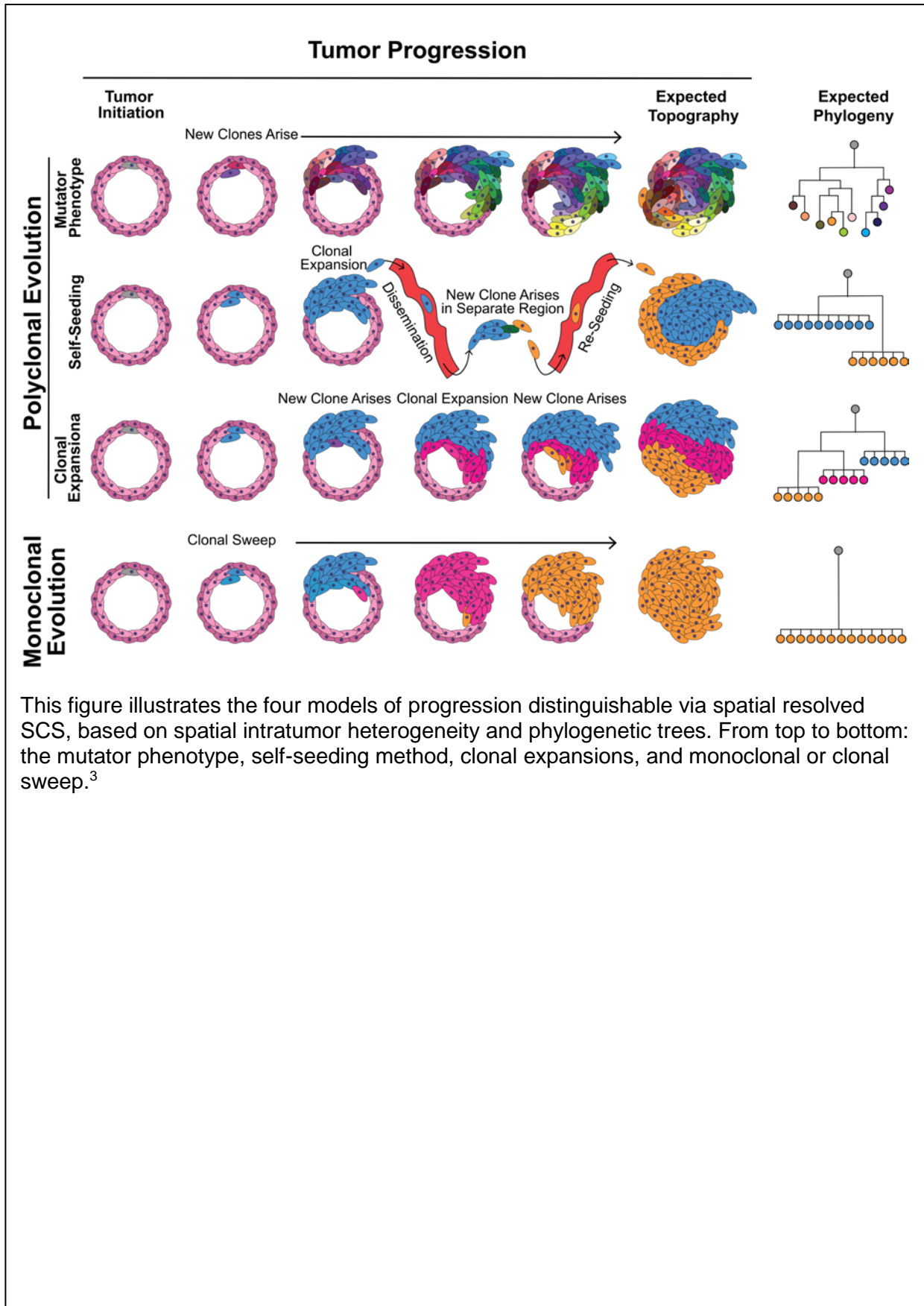
In the original single cell sequencing (SCS) studies, tumors were subdivided into 1mm cubic regions before being flow sorted by ploidy<sup>7, 4, 8, 9</sup>. The other common methods of single cell isolation are mouth pipetting, micromanipulators, and microfluidic<sup>10</sup>. These methods still require cell suspensions and lose spatial information. Previous methods that retain spatial information for deoxyribonucleic acid (DNA) alterations, such as fluorescence *in situ* hybridization (FISH)<sup>11</sup>, are limited to known targeted genetic alterations. Recent developments

in ribonucleic acid single cell sequencing (RNA-SCS) have provided two methods that retain spatial information while examining whole transcription alterations: (1) fluorescence *in situ* sequencing (FISSEQ)<sup>5</sup> and (2) a combination of laser capture microdissection (LCM) to isolate clumps of cells and of single-cell RNA-seq (Geo-Seq, ~10 cells)<sup>12, 13</sup> and LCM-Seq (1 to 50 cells)<sup>14</sup>. However, these methods are limited to ribonucleic acid (RNA) and have several technical challenges.

For DNA-SCS, the only previous study to attempt to retain spatial information divided the tumor into 100-micron thick sections and isolated clusters of morphologically distinct regions and flow sorted these clusters to examine single cells<sup>6</sup>. However, this method loses most of the spatial information within the tumor, and therefore is more of a purification method than a spatially resolved SCS method. In this thesis we will provide an alternative approach to retain spatial data called Topographical Single Cell Sequencing (TSCS)<sup>1</sup>.

Previous studies have shown that breast cancers have significant intratumor heterogeneity (ITH)<sup>15, 16, 8, 17</sup>, making them an ideal tumor to examine spatial ITH. We describe current topics for ITH in DCIS and IDC in 1.2.3 IDC Intratumor Heterogeneity. A recent review by Gulisa Turashvili and Edi Brogi describes the history of scientific understanding of ITH in breast cancer and how this has affected treatment<sup>18</sup>. While ITH can be a hindrance in treatment and studies that use bulk tissue, in our study, we use ITH as markers to help deconvolute tumor evolution based on the number of unique events that differentiate tumor cells from normal cells. By isolating single cells while retaining their spatial information, we can survey the spatial intermixture of tumor clones to help determine which model of tumor progression best fits our data (Figure 1 Models of Progression)<sup>3</sup>.

Figure 1 Models of Progression





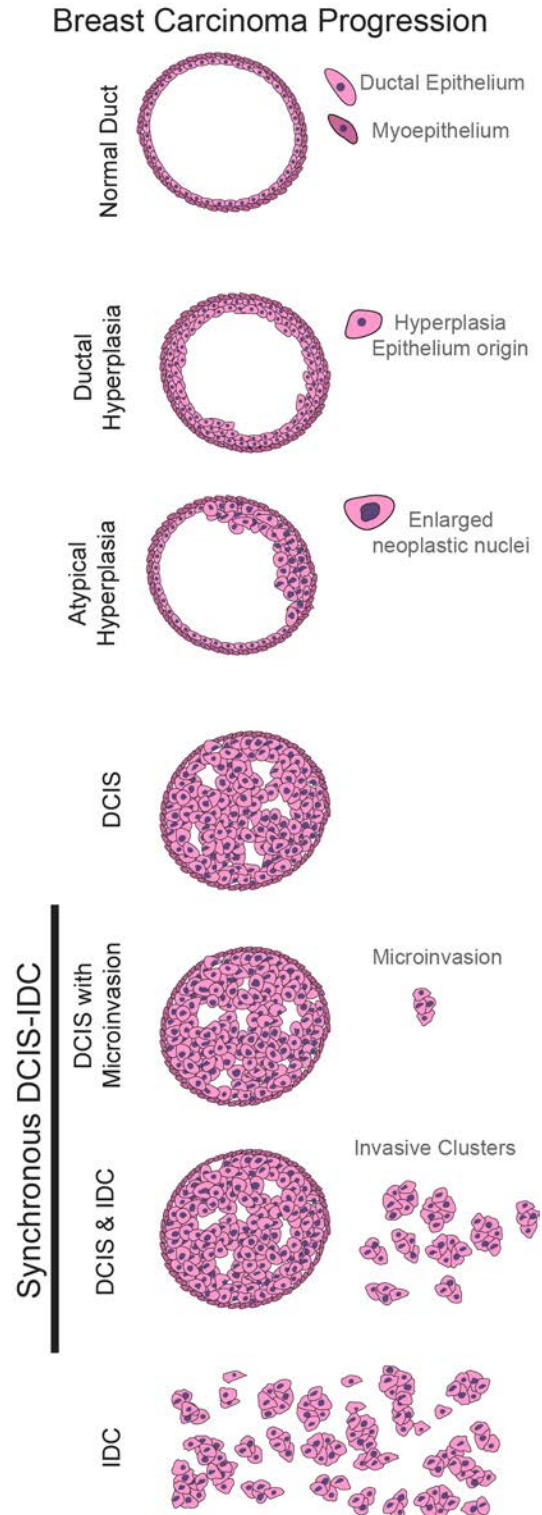
## 1.1 Breast Cancer

One of the most common female malignancies across the globe is breast carcinoma, with 1 in every 8 females experiencing breast cancer during their lifetime. In the United States alone, over a quarter-million women annually are diagnosed with invasive breast carcinoma and another 60,000 with early stage breast cancer as of 2018. While survival of this disease has increased steadily since 1989, about 50,000 women are expected to die from breast cancer in 2018<sup>19</sup>. Lung cancer is the only cancer with higher fatality rates than breast cancer<sup>20</sup>.

Breast cancer is a loss of genetic regulation of cell division and shares many of the same risk factors of other cancers<sup>21-23</sup>, such as family history; however, less than 15% of women with breast cancer report familial disease/incidence<sup>19, 24</sup>. Germline mutations in BRCA1 and BRCA2 increase lifetime risk of breast cancer from 12.5% in the general female population to 55-65% with a BRCA1 and 45% for those with BRCA2. However, only 5-10% of breast cancer cases have germline BRCA1 or BRCA2 mutations. Sadly, at this point the highest risk factors for breast cancer are being a woman and age<sup>19</sup>.

Breast cancer is defined by changes in ploidy, proliferation, and apoptosis of cells within the breast<sup>25</sup>. The more obvious "hallmarks of cancer" for breast cancer are sustaining proliferative signaling, evading growth suppressors, resisting cell death, genome instability (ploidy), and increased mutations<sup>21, 22</sup>. This thesis focuses on the genomic instability and intratumor heterogeneity (ITH) in synchronous ductal carcinoma, where both *in situ* and invasive regions are present in the same patient at the time of diagnosis and surgery.

Figure 2 Sequential Progression from DH to IDC



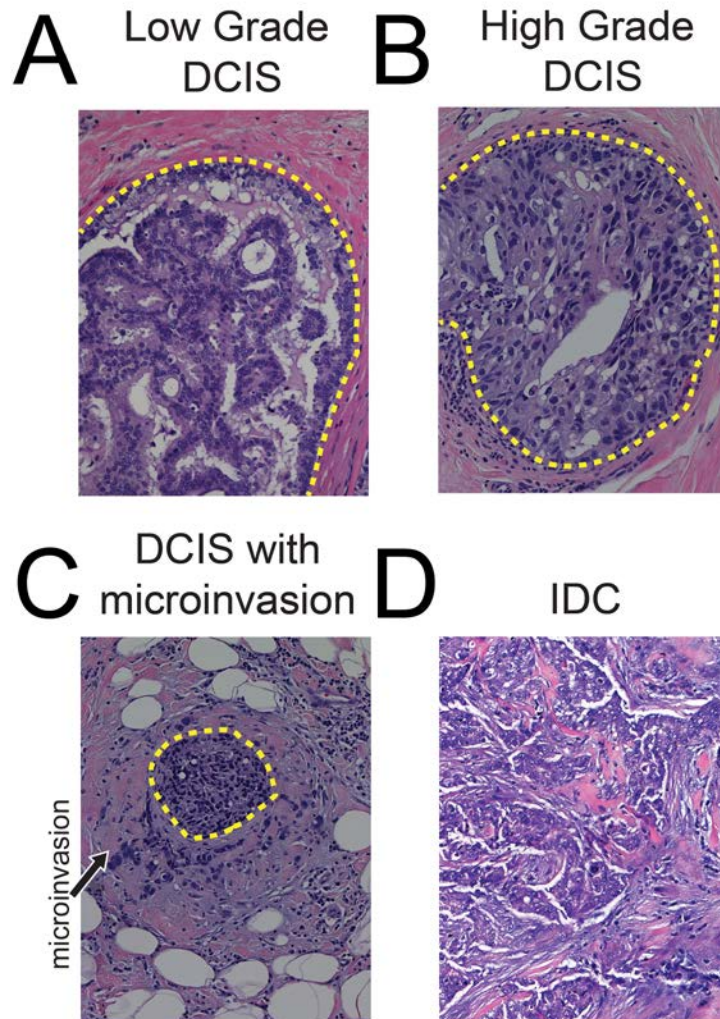
This figure is adapted from Casasent et al. 2016<sup>2</sup> and used by permission. From top to bottom this illustrates the histopathology of hypothesized progression of breast cancer.

### 1.1.1 Model of Breast Cancer Progression

Invasive ductal carcinoma (IDC) is the most common type of breast cancer and is thought to originate from ductal hyperplasia (DH) <sup>26</sup>, following a similar progression model as observed in colon cancer<sup>27</sup>. This model of breast cancer suggests a sequential progression from DH to IDC (see Figure 2 Sequential Progression from DH to IDC). Sequential models of progression are appealing, because it is simpler to assume a straightforward progression from step to step. Under this model, any increased risk for an individual with other breast proliferation dysregulations, such as atypical ductal hyperplasia (ADH) and ductal carcinoma *in situ* (DCIS), automatically increase risk of IDC. Epidemiological evidence however indicates that ADH and DCIS often never progress to IDC. The inconsistent progression of DCIS has caused DCIS to be termed a non-obligatory precursor to IDC. Some experts even question the use of the term 'carcinoma' to label DCIS; stating that DCIS is not actually cancer because by definition DCIS is confined to the ducts and therefore localized. Predicting progression is a major clinical challenge and affects the diagnosis and treatment of patients <sup>28-32</sup>.

Due to the inconsistent nature of DCIS progression to IDC, physicians must make the difficult decision to advise patients to either treat DCIS with aggressive therapy or utilize 'active surveillance'. Caution has led many oncologists to treat more aggressively, resulting in an epidemic of over-treatment in DCIS patients. While many studies have focused on identifying prognostic biomarkers by examining the progression of DCIS to IDC, few of these markers have been useful for treatment<sup>33-35</sup>.

Figure 3 Examples of Pathology DCIS and DCIS-IDC



This figure is adapted from Casasent et al. 2016<sup>2</sup> and used by permission.

(A-D) H&E-stained tissue sections of DCIS and IDC at 200× original magnification: (A) low-to intermediate-grade cribriform-type DCIS; (B) cross section of a duct involved by high-grade solid-type DCIS; (C) synchronous high-grade DCIS-IDC with microinvasion and (D) invasive ductal carcinoma (IDC).

### 1.1.2 Pathology

DCIS is considered a non-obligate precursor of IDC. This designation relies on the concept of sequential progression illustrated in Figure 2 Sequential Progression from DH to IDC. While DCIS is the most common diagnosed early breast cancer, IDC makes up about eighty percent of all diagnosed breast cancer and is the most commonly diagnosed invasive breast cancer.

Histopathology, using hematoxylin and eosin (H&E) staining, is used to determine, grade, stage and cell types of DCIS and IDC. Classification of DCIS and IDC are principally based on patterns observed in the overproliferated areas of the ducts (DCIS) and invasive cell clusters (IDC). The patterns are solid/comedo, cribriform, micropapillary, and mixed for DCIS and mucinous/colloid, medullary, tubular, and inflammatory for IDC.

Solid/comedo DCIS are where the ducts are filled with cells (solid) or filled except for a neurotic center (comedo). Comedo is one of the more invasive cell types, with an estimated about 40% to progress to invasive<sup>36</sup>. Cribriform DCIS as the name implies shows the pattern of spaces like stained glass window. The micropapillary or papillary DCIS also has an emptier center, with pultruding columns towards the center of the ducts. Mixed DCIS shows a mixture of these different cell types within the same patient.

Infiltrating or invasive ductal carcinomas is the most common type of breast cancer, representing about 78% of breast cancer diagnoses<sup>37</sup>. Mucinous IDC is defined by how the cells seem to drift in laden mucin areas, while medullary have a spongy appearance that resembles the brain tissue. Mucinous and medullary breast cancer have about 90% survival over 10 years<sup>38</sup>. Tubular IDC are defined by the tube-like structures that appear as the tumor expands. While tubular breast cancer is very rare, occurring in only 1-4% of breast cancers, it has the best survival rate, almost 100% over 15 years<sup>39</sup>. Inflammatory IDC is defined by the closely associated infatuation of immune-cells surrounding and even intermixed with the tumor cells. Inflammatory breast cancer is also rare (1-5%), but has the worst survival at 65% for 5 years and 35% for 10 years<sup>40</sup>. Inflammatory IDC is difficult to treat because discovery is usually

at stage 3 or 4 and it is often ER and PR negative and HER2 positive (and therefore resistant to many treatments). The recent anti-HER2 treatments appear to have improved the survival rate for patients with inflammatory breast cancers<sup>41</sup>.

The difference in morphology of these types has been linked with risk and in some cases expression patterns. However, morphology alone is not enough to determine treatment without further staining for receptor status<sup>42, 43</sup>.

### **1.1.3 Receptor Status**

Due to the differential in treatment and prognosis between DCIS and IDC, macroscopic assessment of DCIS is more stringent than for IDC, especially at MD Anderson Cancer Center where they cut through most of the tumor to prevent missing possible misdiagnosis. Once a sample is known to be invasive, the treatment plan in the United States (US) is determined by the grade/stage and the status of three receptors for each tumor: the estrogen receptor (ER), the progesterone receptor (PR; ER and PR together are known as the hormone receptors), and the human epidermal growth factor 2 (HER2)<sup>44, 45</sup>. The hormone receptor status determines which treatments are most effective and patients are generally put into one of three groups: those which are positive for either ER or PR overexpression (ER/PR positive), those which are positive for HER2 overexpression and/or amplification (HER2 positive), and those which are negative for alterations in any of the three markers (triple negative breast cancers, or TNBC)<sup>46</sup>. Receptor status is very important for treatment decisions and determining genomic alterations.

The combination of grade and receptor status is strongly associated with divergent survival outcomes. ER and PR positive tumors usually expand very quickly but respond well to hormone therapy and chemotherapy. Therefore, ER/PR positive tumors have a better prognosis than TNBC tumors. Estrogen receptor (ER) positive patients make up the largest group of patients and have differential response to endocrine therapy<sup>47</sup>. Progesterone Receptor (PR) positive patients are rarely seen without also being ER positive.

HER2 tumors have a high expression of the oncogene HER2. HER2 is measured via fluorescent *in situ* hybridization (FISH) and is important because more copies of the HER2 growth gene is linked to normal cells becoming malignant<sup>48</sup>. HER2 tumors are particularly aggressive and account for about a third breast tumor cases<sup>49</sup>, and in general have a lower survival, but this has been significantly improved with trastuzumab treatment going from 75% without trastuzumab to 84% with trastuzumab for 10 year survival<sup>50-52</sup>. HER2 targeted therapies have vastly improved the clinical outcomes of patients with HER2 amplifications<sup>53, 54</sup>, which has been one of the major reasons for the recent push to examine CNA in breast cancer, in the hopes of finding more clinical targets.

The last group is triple negative breast cancer (TNBC) with a 5-year survival rate of 77%<sup>55</sup>. Negative status of receptors is defined as protein expression under 1% based on IHC or 5% prior to 2010<sup>56</sup>. TNBCs have no hormonal or targeted therapies, leaving only chemotherapy and resection options for treatment. TNBC patients are more likely to have BRCA1 germline<sup>57</sup>, <sup>58</sup> mutations, be of African ancestry with other cell cycle check point alterations<sup>59</sup>, and are more like to have extreme aneuploidy.

Previous reviews discussed differences between low and high-grade DCIS in detail <sup>60</sup>, <sup>61</sup>. Low-grade DCIS are more often ER+/PR+/HER2- with fewer CNA than high-grade patients<sup>62-64, 60</sup>. High-grade DCIS have more atypically shaped nuclei and is more often ER- and PR- <sup>65, 60</sup>.

#### **1.1.4 Nuclear Grades**

Another important prognostic factor for ductal carcinoma is nuclear grade. Nuclear grade uses the size and shape of overpopulated cells within the normal ductal structure. The lower the grade, the smaller the nuclei. Within grade 1 the size and shape of the cells are about the same as normal breast epithelial cells, except the cells are filling up the duct. At grade 2, cells are moderately larger (2 to 2.5 times normal breast epithelial cells) and the shape is more irregular. Grade 3 cells are 2.5+ times larger than normal breast epithelial cells and are highly

pleomorphic<sup>66</sup>. Nuclear grade is closely linked with risk of breast cancer progression, but not all high-grade DCIS tumors progress to IDC.<sup>67, 68</sup> Higher grades reflect more advanced cancer with higher risk, with low grades generally linked to higher survival<sup>69</sup>. In low grades (1 and 2), the breast cancer cells bear a closer resemblance to normal breast tissue, where cancer cells are "well-differentiated". High grade tumors (grade 3), cancer cells are "poorly differentiated" and bear little resemblance to normal cells. The speed of growth appears linked to grade, with grade 1 growing more slowly and grade 3 growing more rapidly<sup>70</sup>.

### **1.1.5 Stages**

Breast cancer stages are based on the extent of regions with tumor cells. Stages 0 to 1 are limited to a very localized area. Usually DCIS is referred to as stage 0 breast cancer, meaning the tumor cells are confined to the ducts. Stage 1 is separated into 1A and 1B, where 1A is less than 2 cm of tumor with no lymph nodes are involved, while 1B is also less than 2cm but with lymph-nodes involved. Beyond Stage 1, the stages are more complex, with the number of lymph-nodes involved and the size of the tumor being combined to provide different levels. Stages 2 and 3 are noted for having either many lymph-nodes involved or having larger than 2 cm tumors. Tumors between 2-5cm are Stage 2, unless there are many lymph-nodes involved. Stage 3 consists of 5cm or larger tumors that have not yet metastasized. Stage 4 is metastatic breast cancer. In breast cancer the most common sites for metastases are the brain, bones, lungs, and liver. The main cause of death in breast cancer is metastasis. As expected, tumor size is one of the driving factors of survival for patients with IDC. Patients with IDC tumors with 5cm or larger size had 50-60% survival over 20 years, while smaller tumors at 1cm or less resulted in about 90% survival over 20 years<sup>71</sup>. When one considers that larger margins are associated with less reoccurrence, suggesting that when margins are small or the tumor too large, breast cancer cells remain, thus resulting in better survival rates.

For DCIS, more factors are examined than for IDC, such as presence or absence of necrosis, size of the DCIS lesion, and distance to surgical margins. Recurrence is often of



significant concern during cancer treatment. Since the majority of DCIS recurrences are localized to the spot of resection, recurrence is thought to be caused by unclear margins and residual disease<sup>72</sup>. About half of these recurrences involve invasive disease, downgrading the prognosis for the patient, but the exact relationship between the IDC and DCIS is undetermined<sup>73</sup>. Due to this observation, there is no consensus about what margin should be used to consider a DCIS tumor "completely excised," although a margin size of less than 1mm has the highest recurrence and residual disease<sup>74</sup>.

Patients under the age of 40 and over the age of 80 have the worst prognosis<sup>75</sup>. For the under 40 age group, this is in large part linked to the increased likelihood of presenting with higher grade tumors, which are highly aggressive and usually negative for hormone receptors. Hence, these tumors do not respond to hormone therapies. For the over 80 age group, the treatments are enervating for elderly patients. The frailer a patient is at the start of therapy the less likely they are to recuperate from chemotherapy or surgeries necessary for treatment.

#### **1.1.6 DCIS Survival**

Long-term follow up studies of DCIS patients have shown a substantial difference in the progression of low-grade vs. high-grade DCIS with only 35% of low-grade DCIS patients progressing to IDC over 50 years, while 50% of high-grade DCIS progressed to IDC over 3 years<sup>76, 77</sup>. The grade of DCIS, as discussed in the previous section, is largely based on the size and shape of the nuclei, with larger, more polymorphic nuclei signaling higher grades. The grade is often associated with increased ploidy, or aneuploidy, meaning that there are increased copies of the genome, which follows the hallmarks of cancer (genome instability)<sup>21, 22</sup>. Unlike some colon cancers<sup>78</sup>, the mutational progression of DCIS is still not established. Previous autopsy studies have found moderate levels of DCIS (Average: 8.9%, range 0-14% of multiple studies examined) in undiagnosed women, suggesting that DCIS by itself does not affect quality of life<sup>79</sup>.

## 1.2 IDC Genomics

High-grade DCIS usually has many genome-wide copy number aberrations (CNAs), including frequent events in 1q+, 5p+, 8p-, 8q+, 11q-, 13q-, 14q-, and 17q+ and focal amplifications on 6q22, 8q22, 8q24, 11q13, 17q12, 17q22–24, and 20q13<sup>80, 62, 81, 82, 64, 83, 60</sup>. Mutational markers of IDC include mutations in *TP53*<sup>84</sup>, *PTEN*<sup>85</sup> and *PIK3CA*<sup>86, 84</sup> and amplifications of chromosome 17 and 11q<sup>87, 86, 84</sup>. While breast cancer has well-established genomic markers or even common mutations, most common alterations are specific by subtype because of the extreme amount of intratumor heterogeneity (ITH) in breast cancer. In the 2012 The Cancer Genome Atlas (TCGA) Study on breast cancer, several common CNAs were observed by subgroup and are discussed in more detail in the TCGA and Intertumor Heterogeneity section.

### 1.2.1 IDC Aneuploidy

Aneuploidy is a result of improper chromosomal segregation during proliferation<sup>25</sup>. One daughter cell with (2n+x) chromosomes and the other with (2n-x) chromosomes<sup>25</sup>. When DNA content is abnormal, meaning above or below diploid, DNA abnormality has been associated with adverse effects such as recurrence and low survival rates<sup>88</sup>. The definition of high-grade DCIS increases the likelihood of aneuploidy because grade is determined by size and shape of nuclei, and an abnormal ploidy effects the size and shape of the nucleus<sup>89</sup>. Previous literature suggests that CNAs are early events in tumorigenesis, specifically in breast cancer<sup>90-96</sup>.

Aneuploidy (the state of having an abnormal number of chromosomes) is commonly observed in breast cancer along with chromosomal instability (CIN) (the characteristic of being likely to change ploidy during cell division), and the most common CIN is the mis-segregation of 17, both for whole chromosome aneuploidy<sup>97</sup> and multiple aberrations<sup>98</sup>. Chromosome 17 is notable for harboring a number of genes, including the hormonal receptor *HER2*, as well as the well-known tumor suppressors genes *TP53* and *BRCA1*<sup>98</sup>.

*BRCA2* mutated breast cancers are more often tetraploid (4 copies of the genome) than sporadic breast cancers<sup>99</sup>. While *BRCA2* is associated with homologous recombination pathway for double-strand DNA repair, it also facilitates the formation of anaphase bridge mutations in *BRCA2*, which can cause mis-segregation of chromosomes<sup>100</sup>.

The TCGA study split the breast cancer samples into different subtypes, specifically Luminal A that was mostly diploid, while the other 3 subtypes Luminal B, Basal-like, and HER2-expressed were all highly aneuploid tumors. The most common alterations across all subtypes were gains in 1q and 8q, and loss in 8p<sup>101</sup>. One study found an early alteration commonly observed in DCIS, and in atypical ductal hyperplasia (ADH), is a loss of heterozygosity (LOH) in chromosome 11q13<sup>102</sup>. However, a later study found common LOHs in ADH were 16q and 17p, while 11q13 was infrequent but when present was clonal<sup>103</sup>. The small size and possible sample of original study probably accounts for the discordance between these two results. LOH (or deletion of one allele) of 6q13 and 6q26-27<sup>104-106</sup> is also common in breast cancers.

### **1.2.2 TCGA and Intertumor Heterogeneity**

Intertumor heterogeneity is the diversity and lack of common alterations across patients. The advent of personalized medicine is based on the idea that differences in genomic alterations between patients can determine specific treatments for individual patients. The TCGA study of breast cancer found few common mutations across the 825 patients analyzed<sup>101</sup>. There were 9 somatic mutations which occurred in over 10% of the patients. Some of these were mutations in well-known tumor genes like TP53 (36%), PIK3CA (34%), CDH1 (14.7%)<sup>101</sup>. Other mutations were found in lesser-known genes like TNN (24.7% for breast cancer), MUC4 (20% for breast cancer, but also seen in lung and cervical cancers), MUC16 (15% for breast cancer, but also seen in ovarian cancers), GATA3 (13.9% in breast cancer), MUC2 (11.6% for breast cancer, but also seen in ovarian and bladder cancer), and KMT2C (10% for breast cancer, but also seen in colon cancer)<sup>101</sup>.

While the TCGA study examined breast cancer samples regardless of their receptor status, receptor status correlates with specific mutations or alterations rather than histological subtype<sup>101</sup>. The TCGA paper shows high intertumor heterogeneity in DNA variants, mRNA expression, miRNA expression, DNA methylation status, as well as protein levels in invasive and metastatic breast cancer<sup>101</sup>.

### **1.2.3 IDC Intratumor Heterogeneity**

Intratumor heterogeneity (ITH) is frequently reported in IDC <sup>107, 108, 4, 109, 110, 17, 85</sup> and in DCIS studies profiling DNA, RNA, and protein levels <sup>15, 90, 111, 62, 112-118</sup>. ITH complicates diagnosis and treatment, but is beneficial for evolutionary studies since it provides a 'permanent record' of mutations during tumor growth<sup>9</sup>. Assuming mutational complexity increases over time and using phylogenetic inference, several studies showed clonal lineages and evolutionary histories can be inferred from a single time-point tumor sample <sup>16, 4, 17</sup>. This experimental approach is important for evolutionary studies of DCIS, where often only a single time point sample can be obtained <sup>114, 119</sup>.

Early studies of ITH used cytological and histopathological methods. These methods included FISH to measure DNA copy number of targeted genes or loci, and immunohistochemistry (IHC) to measure protein levels across tissue sections. Many FISH studies reported ITH in DNA copy number states of single tumor cells in the ducts of DCIS patients <sup>120-125, 116, 126</sup>. Multiple studies have reported ITH in receptors such as HER2 in DCIS <sup>121, 127, 125, 116</sup>. Heterogeneity in protein levels and targeted genes have also been reported using cytological and histological methods <sup>15, 128, 129, 4, 116</sup>. Allred et al. used IHC to stain specific proteins in DCIS and revealed spatial ITH in protein levels of TP53 and HER2 <sup>15</sup>. However, these methods were often qualitative and limited to single targeted genes or proteins.

Next Generation Sequencing (NGS) methods provided quantitative measurements of thousands of mutations and CNAs in parallel. Three different experimental NGS approaches have been developed to resolve ITH: 1) deep-sequencing, 2) multi-region sequencing, and 3)

single cell DNA sequencing. Deep-sequencing involves sequencing bulk tumor at high coverage depths to cluster mutation frequencies and identify clonal subpopulations. This approach has been applied to study ITH and clonal evolution in IDC patients<sup>130, 131, 110</sup>. Multi-region sequencing involves spatially sampling different macroscopic regions of tumor mass and sequencing each region independently to resolve geographic heterogeneity<sup>132, 85, 133</sup>. These methods enable the reconstruction of phylogenetic lineages to understand clonal evolution in breast cancer patients<sup>85</sup>. SCS methods can measure genome-wide copy number profiles<sup>7, 16</sup>, exome mutations<sup>134, 135</sup>, genomes<sup>17, 136</sup>, or targeted gene panels<sup>137</sup> in single cells. SCS methods can fully resolve ITH by reporting genomic information on individual tumor cells, but are more susceptible to sampling bias<sup>138</sup>. By sequencing and comparing multiple tumor cells, several studies have delineated the clonal substructure and evolutionary lineages of IDC<sup>16, 4, 17</sup>. However, these SCS approaches have not yet been applied to DCIS.

Table 1 aCGH and NGS DCIS Papers

Method	Samples Analyzed				First Author	Year	PMID
	DCIS	DCIS-IDC	IDC	Type			
SCS	0	10	0	Frozen	Casasent	2018	29307488
	0	3	1	FFPE	Martelotto	2017	28165479
NGS	6	5	0	Frozen	Kim	2015	25831047
	0	1	0	Frozen	Kroigard	2015	25730902
	0	1	50	FFPE	Yates	2015	26099045
	3	12	0	FFPE	Foschini	2013	23337025
	6	0	6	FFPE	Newburger	2013	23568837
aCGH	1	0	5	FFPE	Oikawa	2015	24402639
	0	13	0	Frozen	Hernandez	2012	22252965
	20	25	24	FFPE	Liao	2012	22887771
	52	0	0	FFPE	Hwang	2011	21496874
	0	21	0	FFPE	Johnson	2011	22052326
	31	42	36	Frozen	Muggerud	2010	20663721
	6	15		FFPE	Iakovlev	2008	18628458
	10	0	18	Frozen	Yao	2006	16618726

This table is adapted from Casasent et al. 2016<sup>2</sup> and used by permission.

This table contains a list of genomic studies from SCS, next-generation sequencing, and microarray CGH profiling of DCIS breast cancers. The columns are primary method used in the paper, number of samples analyzed, type of tissue, first author, year of publication, and PubMed ID (PMID).

The number of samples in each study is reported as DCIS-only, DCIS-IDC (synchronous, where both DCIS and IDC samples were assessed), and IDC-only samples.

FFPE=formalin-fixed, paraffin-embedded.

#### 1.2.4 Synchronous DCIS-IDC

In synchronous DCIS-IDC the patient has both *in situ* and invasive carcinoma. Genomic biomarker studies mainly used gene expression microarrays or array copy genomic hybridization (aCGH)<sup>87, 90, 120, 139, 140, 82, 126, 141</sup>. Many of these studies reported highly similar copy number profiles and gene expression signatures for synchronous DCIS-IDC regions, and analyzed both DCIS and IDC regions<sup>139, 142, 114, 84, 119</sup>. With the development of next-generation sequencing (NGS) technologies, studies have begun to apply higher resolution methods to study invasive-specific mutations and CNA in patients with synchronous DCIS-IDC<sup>84, 143, 119, 144, 85</sup>. Many of these studies have identified concordant and discordant mutations in synchronous DCIS-IDC patients<sup>114, 84, 119, 144, 85</sup>. However, these initial genomic studies faced several technical obstacles, including low tumor purity, the unavailability of fresh-frozen tissues, and ITH. Consequently, the genomic and molecular basis of invasion in DCIS breast cancers remains poorly understood. Table 1 aCGH and NGS DCIS Papers lists the papers that studied DCIS or Synchronous DCIS-IDC using aCGH or NGS methods.

### 1.3 Single Cell Sequencing

In the last decade, single cell sequencing (SCS) developed into a powerful genomics tool. In developmental biology, the ability to examine the transcriptome or genome of every cell within a specific organ can create a new understanding of organ-development and mosaicism. In cancer, SCS enables exploration of ITH previously inconceivable.

#### 1.3.1 SCS Beginnings

The first single cell full genome DNA sequencing experiment was completed in 2011 by Navin et al<sup>4</sup>. This project not only managed to create the first whole genome sequencing (WGS) of single cells but also attempted to infer the clonal evolution of a tumor from the single cells. The Tang et al 2009<sup>145</sup> whole single cell transcriptome paper is considered the beginnings of RNA-SCS while the 2011 Navin et al paper<sup>4</sup> is considered the beginning of DNA-SCS.

The SCS era allows examination of tumor heterogeneity, the mosaicism of tissues, full transcriptional heterogeneity of development, and the complexity of organs such as the brain. I will be focusing on DNA single cell sequencing (DNA-SCS), which we used in my project, and the many challenges for single cell DNA sequencing.

### **1.3.2 SCS Challenges**

The major issues that affect single cell genome sequencing (DNA-SCS) data are related to the limited amounts of genetic material and technical errors introduced by whole genome amplification (WGA) methods including: nonuniform coverage, allelic dropout, false-positive errors, false-negative errors, and cell type specific variations. To examine the effect of amplification methods on DNA-SCS, Biezuner et al<sup>146</sup> compared seven different commercial kits and Huang et al<sup>147</sup> compared five different kits. These experiments covered the top three DNA-SCS amplification methods: degenerate oligonucleotide primed polymerase chain reaction (DOP-PCR), multiple strand displacement amplification polymerase chain reaction (MDA-PCR), and Multiple Annealing and Looping Based Amplification Cycles (MALBAC)<sup>146, 148, 147</sup>. These studies showed that MDA and MALBAC had the lowest allelic dropout rates and appeared to be the best suited for DNA mutation (SNP and indel) analysis. With MDA methods producing the most mappable reads and the most coverage and the smallest number of false positive calls, but the least reproducibility between two cells especially in the case of copy number. After MDA, MALBAC protocols produced the decent coverage, with better reproducibility (according to Huang et al<sup>147</sup>), and with comparable allelic dropout rate to MDA, but a higher false positive rate and more unmapped regions. The last type of kit DOP-PCR is the type of kit we used. DOP-PCR results were highly reproducibility between cells but had did more poorly on the other statistics that were measured, however the reproducibility was the only statistic that measured how clean the copy number data was instead of the mutation data<sup>146, 147</sup>.



Another major issue for SCS is sampling. On the microscale, the increased information from SCS allows peeks into ITH. However, SCS of an entire tumor mass is cost prohibitive. Therefore, because the full population of the tumor is not sequenced, the result is an increase in sample bias, where small changes in samples can cause divergent results and therefore conclusions. Due to the expense of sequencing single cells, the first papers covered only a few dozen cells, but each year more and more cells are sequenced as part of SCS experiments. While it is now financially reasonable to sequence hundreds to thousands of cells per tumor mass, cost still prohibits sequencing the whole tumor mass. For the smaller tumors, most of this mass is used by pathologists for vital diagnostics. This is often true of pure DCIS tumors. Larger tumors are most often used for research and have vastly more cells than it is feasible to sequence.

Sampling issues led to techniques capable of determining the number of tumor cells necessary for good coverage of the mutation spectrum. Standard population metrics, often used in ecology, have also been redesigned for SCS. However, the major conclusion has been that sampling counts are unique to each tumor, resulting in several *ad hoc* calculations that have become standard at the end of individual studies. While this may answer the question of sample size on a case-by-case basis, the issue of sampling bias remains an issue for SCS experiments.

Even in the first DNA-SCS studies<sup>4</sup>, there were regional effects within a tumor measured on the macro level, in sections of about 1mm. Using microdissection, different sections of a tumor were flow sorted. The single cell profiles were mapped back to an approximate region. These methods demonstrated that tumors are not a homogenous mixture of cells and could have regional genomic differences<sup>4, 17</sup>. These results also suggested sampling bias could affect a study, which drove the search for a spatially aware SCS method. However, the majority of single cell isolation methods at the time used cell suspensions. Cell suspensions are created by separating or dissolving the extracellular matrix in order access the cells. For more details on the most common isolation techniques for SCS refer to the Wang and Navin 2015 Review<sup>10</sup>.

## **1.4 Laser Capture Microdissection Methods**

A major issue with current single cell techniques is loss of spatial information. Whether one is using flow sorting or nanowells, one of the first steps of traditional sample preparation is dissociation of the sample, disrupting the relationship between spatial coordinates and genetic or transcriptomic data. Laser capture microdissection (LCM) is one of the few methods that can maintain micron level spatial information. This section will discuss LCM and current sequencing techniques that use LCM.

As early as 1996, LCM became a major tool in the biological sciences for separating different cell types. The first techniques used crude methods of cell protection to prevent DNA damage by covering desired cells and exposing the remainder of the sample to UV to destroy the DNA of the undesired cell types<sup>149</sup>. Later techniques used touch-based methods to transfer the cells of interest by melting a film on the slide, allowing the region of interest to be peeled off with the film<sup>150</sup>.

### **1.4.1 LCM Sample Purity**

One of the major advantages to LCM is the increase in sample purity for the cell type of interest. Using bulk tissues containing normal tissue can cause inaccurate tumor profiles whether one is using next-generation sequencing or CGH<sup>89</sup>. The diversity of cell types within tumor samples has long been recognized. However, the ability to routinely separate cell populations was a substantial advancement for research, allowing scientists to focus on a cell subtype of interest<sup>150</sup>. While LCM techniques were predicted to provide an advancement in clinical applications such as increasing tumor purity during diagnostics, these have not yet emerged, partly due to the significant time and cost required for microdissection.

While LCM offers a much more efficient method compared to manual ink-stained sections (UV selective radiation ablation)<sup>149</sup> or manual microdissection, the current tools for LCM are expensive (between \$110K and \$250K for the instruments) and require large investments in training and regular maintenance. These limitations make LCM more time

intensive than clinical pathology reviews and LCM isolation is less systematic requiring intensive training and manual collection.

#### **1.4.2 LCM Type Selection**

The Zeiss PALM Microbeam system uses one type of laser to cut out around a region, cell, or nucleus and a second type of laser to catapult the cut-out into a capture cap. The Zeiss PALM Microbeam was selected as optimal for this project after comparison with other systems. The other systems in our limited initial testing produced limited, possibly contaminated, or no results.

#### **1.4.3 LCM Spatially Resolved Sequencing**

Preserving spatial information can help parse biologically interesting phenotypes and was first done with LCM-Seq<sup>14</sup> in 2016 and Geo-Seq (short for geographical sequencing) in 2016<sup>13</sup> and 2017<sup>12</sup>.

LCM-Seq was first used in a Nature Communications paper by Nichterwitz et al to isolate motor-neurons in mice<sup>14</sup>. The paper focused on the development of the LCM-Seq method, moving from a bulk purification method to a single cell method. The first step was to isolate clumps of cells ~120 cells and then to sequentially lower the number of input cells from 50, 30, 10, 5, 2, and finally 1 cell. While they were able to scale down to 1 cell, the total mapping and general data quality of the single cell RNA isolated was lower, just under 80% compared to 10 cells which had an equivalent mapping ratio to 120 cells. Since, cells were isolated from LCM slides, in theory topographical data could be collected. However, this paper did not utilize that aspect of LCM isolation, but instead used LCM to isolate rare cell types with known anatomic positions<sup>14</sup>.

Geo-Seq was first used in a 2016 Developmental Cell paper by Peng et al to examine spatial transcriptomes in mouse embryos<sup>13</sup>. In 2017, Nature Protocols published a paper by Chen et al covering a method designed to elucidate both cellular transcriptomic heterogeneity and spatial variance of the transcriptome<sup>12</sup>. The method takes about 5 days for collection and

processing of samples and 1-2 weeks for analysis. This method was used on mouse embryos and worked for small cellular input of 10 cells or less. They used a "zipcode mapping protocol" to allow the cell populations to be mapped back to an approximate location, and produce 4D data, with spatial (x,y,z) and genomic (a) components. However, all data shown was from cell clumps of 10 to 20 cells<sup>12, 13</sup>.

A method to examine single cells from microdissected tissue was published in 2017<sup>6</sup>. This method used standard microdissection instead of LCM and used H&E slides from an adjacent section to direct dissection of a DAPI stained tissue. Dissected cells were flow sorted so single cell genome data could be collected<sup>6</sup>. While this method allowed for some basic geographic data (to the nearest 1mm) to be collected, the biggest benefit was an increase in purity rather than precise spatially linked genomic data. However, this method is relatively high-throughput<sup>6</sup>.

Spatial information plays a major part in how cancer is treated. The difference between DCIS and IDC is determined solely by whether the abnormal cells or larger abnormally nuclei are inside the ductal structure (DCIS) or outside (IDC). Outside of breast cancer, other early cancers are often defined by the restriction of the cancer-like cells by a membrane. Outside of cancer, many questions relate to how adjacent cells diverge during development. For example, in neuroscience, the interaction between glia, astrocytes, and neurons is currently at the forefront of research. While many of these questions are better answered by single cell RNA sequencing, the development of spatial single cell DNA sequencing methods is the first step toward this.

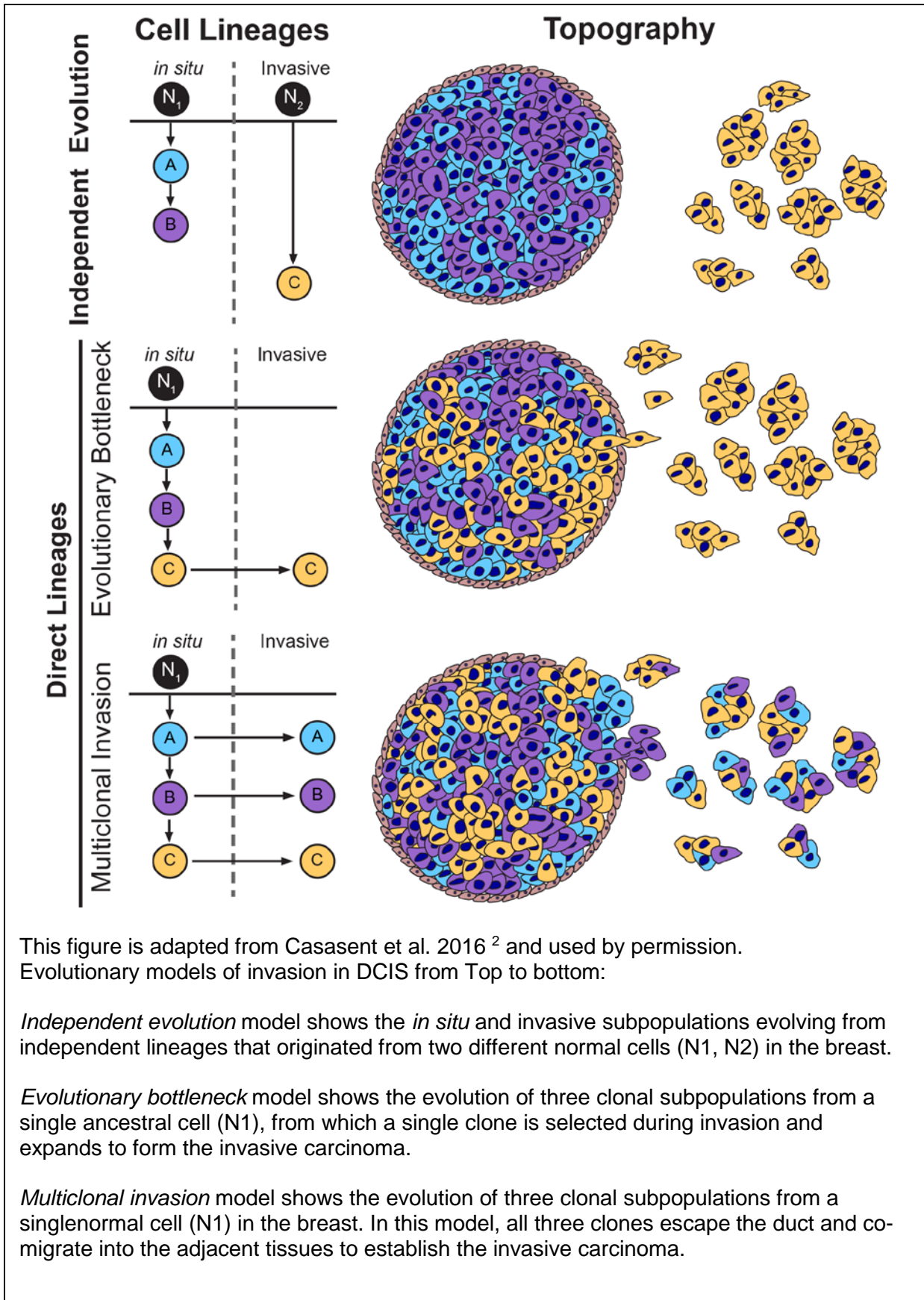
### 1.5 Models of Genomic Lineage and Invasion

Three proposed models of invasion during progression of DCIS to IDC: 1) *independent evolution*, 2) *evolutionary bottlenecks* and 3) *multiclonal invasion* (see **Error! Reference source not found.**). The *independent evolution* model postulates the presence of two different initiating cells ( $N_1$ ,  $N_2$ ) in normal breast tissue that separately evolve into DCIS and IDC

subpopulations. In contrast both direct lineage models (*evolutionary bottlenecks* and *multiclonal invasion*) postulate a single cell of origin. Specifically, direct lineage models presuppose a single normal breast cell ( $N_1$ ) that gives rise to both DCIS and IDC subpopulations.

The two direct lineage models are separated by how or if selection of subclones occurs during invasion. The *evolutionary bottleneck* model postulates that a subset of clones within the duct escape. Consequently, the founder's effect caused by the an invasive bottleneck would result in few individual cells and few subpopulations being present in the invasive regions. The results would be subclonal divergence and lower diversity in the invasive tumor than in the *in situ* region, also called the founder effect. The *multiclonal invasion* model postulates that all clones escape the duct. As a result, the invasive population would contain all subclones found in the original *in situ* population. Results that support the *multiclonal invasion* model would include the presence of all clones in both regions and similar clonal diversity between regions. While later diversity might occur, the *in situ* and invasive regions are seeded with the same set of clones, which would increase the likelihood of observing similar subclones and similar level of diversity in both regions. The escape described in the *multiclonal invasion* model could be a coordinated or stochastic process preceding the degradation of the ductal membrane.

Figure 4 Models of Invasion



### 1.5.1 Independent Lineage Model

Under the *independent lineage* model, DCIS and IDC independently evolve from two separate normal cells. Therefore, no somatic mutations or CNAs should occur in both cell lineages. The data supporting this model comes from histopathological sections, in which about 20% of these cases have DCIS and IDC in different regions of the breast<sup>151</sup>, and from a number of single marker studies showing discordance between synchronous *in situ* and invasive subpopulations<sup>15, 87</sup>. Topographically distant areas of synchronous DCIS-IDC cases can be classified as different grades and often display genetic and histopathological ITH which has been further stated as support for the *independent lineage* model<sup>15, 152, 113, 139, 153, 72, 85</sup>.

The emergence of multiple tumor lineages might be explained by cancer field effects giving rise to multiple tumor initiating cells. Cancer field effects have been reported in tumors with external mutagens, such as UV exposure in eye lid cancers<sup>154</sup> and cigarette smoke in lung adenomas<sup>155</sup>, as well as in tumors such as breast cancer, with no known external mutagens<sup>156</sup>. Cancers without external mutagens but with germline mutations (e.g. *BRCA1*, *BRCA2*, *TP53*) have generated multifocal tumors that often share few to no somatic mutations<sup>157-159</sup>.

Single targeted gene and protein studies often provide support for the *independent lineage* model, where the targeted gene or protein is discordant within a patient<sup>139, 160</sup>. Discordance of *PIK3CA* mutations with matched IDC and DCIS was found with only 30% concordance between regions<sup>160</sup>. Further support for the *independent lineage* model was generated by deep-sequencing of the mitochondrial D-loop in DCIS-IDC patients, where 61% of tumors were estimated to evolve from different clonal origins, supporting independent lineage<sup>139</sup>. The last set of support for the independent lineage model is from mathematical modeling of DCIS and IDC which also indicates *independent lineages*<sup>161</sup>. In summary, the experimental support for the independent lineages model is based mainly on single marker studies and mathematical models. These studies may not observe underlining genetic concordance due to profiling only selected mutations.

### 1.5.2 Dependent Genomic Lineage Model

Direct lineage models postulate that a single initiating cell ( $N_1$ ) in normal breast tissue gives rise to both DCIS and IDC subpopulations. Support for direct lineage models of synchronous DCIS-IDC comes from genomic studies using aCGH<sup>114, 162, 163, 86, 164-167</sup> and NGS<sup>84, 119, 144, 85</sup> which profile many markers across the whole genome. These studies report a high correlation in copy number profiles of *in situ* and invasive subpopulations in synchronous DCIS patients<sup>120, 168, 169, 86, 84, 164, 126, 167</sup>, and many concordant point mutations<sup>114, 84, 119, 170</sup>. A thirty-eight study meta-analysis found 67% of synchronous DCIS-IDC studies strongly supported direct lineage<sup>171</sup>, while the other 33% could possibly support independent lineages, especially in the 20% of cases where DCIS and IDC tumors were found in disparate regions.

Large genomic synchronous DCIS-IDC studies also report high numbers of discordant mutations or region-specific CNAs<sup>172, 114, 162, 163, 86, 84, 119, 164, 165, 144, 166, 167, 85</sup>. However, these discordances could arise through later divergent evolution. These data could therefore be explained by both direct lineage models, *evolutionary bottleneck* through selection of minor clones with invasive phenotypes and *multiclonal invasion* through migration of multiple dominant clones. The limited ability of bulk genomic studies to resolve ITH and trace clonal lineages makes distinguishing between *evolutionary bottleneck* and *multiclonal invasion* models challenging.

#### 1.5.2.1 Evolutionary Bottleneck

Population bottlenecks (*evolutionary bottlenecks*) are illustrated in natural species often via allopatric speciation in which a small population (the founding population) moves to a distant region and adapts to the new environment selecting for new traits<sup>173</sup>. Population dynamics often refers to the original change as the founder effect. *Evolutionary bottlenecks* within tumors have been reported in studies in metastatic dissemination<sup>174, 119, 85</sup> and as a result of cancer therapy<sup>175-178</sup>. The selection of minor clones may occur during invasion, resulting in a founder effect that suggests higher discordances or longer evolutionary distances



between regions. A few studies of synchronous DCIS-IDC have failed to detect mutations and CNAs prevalent in IDC that are absent in the DCIS region <sup>84, 119, 144, 85</sup>.

Concordant mutations are often referred to as ‘truncal’ mutations, a term taken from evolutionary biology, because ‘truncal’ mutations can be traced to the last common ancestor in the tumor. Truncal events, which are traceable throughout the lineage, were reported by Sidow and colleagues based on whole-genome-sequencing (WGS) of six breast cancer patients with matched longitudinal samples of atypical ductal hyperplasia (ADH), DCIS, and IDC that produced concordance early in the lineage, with late lineage discordant events consistent with the *evolutionary bottleneck* model <sup>144, 179</sup>. A multi-region (DCIS and IDC) sequencing study performed lineage-tracing experiments in ER+/PR/HER2- synchronous DCIS patients and found two distinct *PTEN* mutations in IDC regions that were absent in DCIS regions, which was used to suggest that a minor clone without a PTEN mutation was selected during invasion <sup>85</sup>.

In another study, higher concordance of CNAs than point mutations were reported in a study of 6 patients with synchronous DCIS-IDC that strongly supports a direct lineage model <sup>84</sup>. About 40% of the mutations were concordant with *TP53* having the highest concordance, and *FANCE*, *ATM*, *BCOR*, *PDGFRA*, and *PMS1* being the least concordant mutations<sup>84</sup>. Reis-Filho and colleagues used aCGH to profile 13 synchronous patients and found 77% of patients had highly similar genome-wide copy number profiles between regions <sup>114</sup>. The remaining patients mostly showed additional amplifications in invasive subpopulations (1q41, 2q24.2, 6q22.31, 7q11.21, 8q21.2 and 9p13.3), which is consistent with an evolutionary bottleneck <sup>180</sup>.

Further support for the *evolutionary bottleneck* version of the direct lineage model comes from reported concordance of mutations and CNAs in patients with synchronous DCIS-IDC, in studies without evolutionary analysis. Microdissected DCIS and IDC regions processed with aCGH presented overall concordant CNAs with a few invasive-specific amplifications of oncogenes and deletions of tumor suppressors <sup>114, 162, 163, 86, 164, 165, 181, 166, 167</sup>. Similarly, sequencing and genotyping analysis of synchronous DCIS-IDC regions for a majority of shared

events with a few invasive-specific events which is still consistent with an *evolutionary bottleneck*<sup>87, 120, 182, 113, 114, 169, 163, 86, 84, 144, 85</sup>.

Collectively, these data are consistent with an evolutionary bottleneck model, in which a clone is selected during invasion, leading to the expansion of a minor genotype in the invasive carcinoma. An alternative explanation for discordant data is invasive clones continuing to evolve new mutations and CNAs after tumor cells escape from the ducts. To distinguish between these possibilities, higher resolution genomic methods are required to resolve ITH and perform lineage reconstruction, and to determine if the invasive genotype was pre-existing in the ducts in a minor subclone.

### 1.5.2.2 Multiclonal Invasion

*Multiclonal invasion* is a direct lineage model in which multiple cells from different subclones escape the duct to establish the invasive carcinoma. In the *multiclonal invasion* model the migration of clones is often thought to be preceded by the breakdown of the basement membrane. In *multiclonal invasion*, or parallel invasion<sup>6</sup>, all clones escaped the ducts and are therefore observed in both the ducts and invasive regions.

Two scenarios comply with the *multiclonal invasion*. The first is the cooperative scenario, in which multiple DCIS clones coordinating through non-cell autonomous paracrine or juxtacrine interactions cooperatively escape the basement membrane (the ducts) to become invasive. In the cooperative scenario, multiple clones escape the duct and the cooperation could be between clones or with the tumor microenvironment. Support for the cooperative scenario was found in functional experiments where IDC showed non-cell autonomous interactions of clones which secreted growth factors and cytokines promoting tumor growth<sup>183</sup> or in mouse models, where *WNT* signaling drove tumor progression<sup>184</sup>.

The second is the leader scenario. In the leader scenario, a "leader clone" breaks down the basement membrane. Then the leader clone and "follower clones", which would be incapable of escaping the membrane by themselves, together establish the invasive legion.

The leader clone scenario does not require direct cooperative interactions between clones. Evidence for this process is supported by histopathological images showing a complete breakdown of basement membrane and myoepithelial layers in some DCIS cases. Both scenarios contain similar proportions of clones and result in DCIS and IDC regions with similar mutations and variant allele frequencies (VAFs).

Highly correlated copy number profiles between DCIS and IDC regions in aCGH studies provided genomic evidence for the *multiclonal invasion* model<sup>114, 162, 163, 86, 164, 165, 181, 166, 167</sup>.

Oikawa et al used aCGH and reported a similarly high 97% concordance<sup>166</sup>. Hernandez et al found 10 of 13 patient's aCGH copy number profiles to be highly correlated. An average of 83% concordance was found by Johnson et al via aCGH for 23 patients with synchronous DCIS-IDC, with many having extremely similar copy number profiles<sup>86</sup>. *Multiclonal invasion* model is also supported by high concordance of DCIS and IDC mutations and subclonal mutation frequencies found in NGS studies<sup>84, 119, 144</sup>. While the data is inferential and provides only indirect evidence, the data is consistent with a *multiclonal invasion* model.

### 1.5.2.3 Clinical Implications of Invasion Models

Proper diagnosis and therapy for DCIS patients hinges on the proper model of invasion. Under the *independent lineage* model DCIS and IDC are genetically unrelated. Treatments targeting DCIS genetics, which is unlikely to become invasive, are not particularly useful. However, both dependent lineage models (*evolutionary bottleneck* and *multiclonal evolution*) suggest genetics in common between DCIS and IDC. Truncal mutations from early in the tumor lineage and carried throughout the lineage can be targeted and used to eliminate both DCIS and IDC cells<sup>185, 186</sup>. The TRACERx clinical trial is currently investigating just such an approach in lung cancer<sup>187</sup>.

Under the *evolutionary bottleneck* version of the dependent lineage model, invasive clones could have specific mutations allowing invasive clones to be treated. For example, in a multi-region sequencing study of a synchronous DCIS-IDC patient, the authors identified loss-

of-function mutations in the *PTEN* tumor suppressor in invasive subpopulations not present in ducts<sup>85</sup>. This invasive-specific *PTEN* mutation could potentially be targeted with *PIK3CA*, *AKT* or *mTOR* inhibitors to treat the cancer<sup>188</sup>.

The *multiclonal invasion* version of the dependent lineage model suggests a unique method of invasion. If the cooperative scenario holds true, it could be possible to prevent further invasion by inhibiting cooperative interactions to therapeutically hinder invasion. This could be achieved by interfering with cooperative clonal interactions via drugs or antibodies targeting secreted factors or receptors that cells use in paracrine or juxtacrine interactions. Conversely, if the leader clone scenario is true, then identifying the leader clone will be required to target therapeutic intervention. However, such an approach requires a mechanistic knowledge of underlying cell interactions and signaling pathways used for cooperation, requiring detailed studies using *in vitro* or *in vivo* systems, such as xenografts.

Direct lineage models also have important prognostic implications for measuring ITH using diversity indexes based on genome type<sup>107, 16</sup>. These models suggest DCIS patients with high diversity indexes<sup>189, 190</sup> (such as, Shannon<sup>191, 192</sup> or Simpson's Index<sup>193</sup>) would be more likely to progress to IDC due to the increased chance of an invasive clone evolving (specifically in the leader clone scenario)<sup>116</sup>. Similarly, high diversity indexes were correlated with potential to metastasize or present with poor response to therapy<sup>107, 194, 127</sup>. Conversely, in a direct lineage model, a low genomic diversity would expect to predict a lower risk of invasion in DCIS patients. Future studies would be required to determine if high genomic diversity predicts progression.

## 1.6 Dissertation Summary

In return for large input of hundreds or thousands of cells, NGS methods provided quantitative measurements of thousands of mutations and CNAs in parallel. However, for cases with a heterogeneous population, NGS methods created a new problem, how to deconvolute a set of unknown rare mutations. SCS provides some experimental resolution to this problem.

This thesis focuses on two matters (1) creating a SCS method to retain spatial context and (2) using this method to examine intratumor evolution and heterogeneity in synchronous DCIS-IDC breast cancer. We hope that further studies will build on these methods to identify prognostic biomarkers by comparing, *in situ* only, synchronous, invasive only, and metastatic breast cancers.

### 1.6.1 Spatially-Resolved Single Cell DNA Sequencing

Since a major issue with the current single cell techniques is the loss of spatial information, I developed a true single cell method to examine the spatial, morphologic, and genomic data from single cells. While I was not the first to study this, my method provides more precise spatial information than the 20 cell clumps of the 2016 Geo-Seq method<sup>12, 13</sup> or than regional microdissection of thousands of cells paired with flow sort by Martelotto et al in 2017<sup>6</sup>. While these methods allowed for basic geographic information to be retained, they lose single cell morphology and precise location information to providing higher-throughput.

The method I have developed is able to isolate single tumor cells from frozen tissue sections and preserve their precise spatial positions and morphology. The development of Topographical Single Cell Sequencing (TSCS) adds precise spatial mapping to morphological and genomic analysis<sup>1</sup>.

TSCS combines laser-capture-microdissection, laser-catapulting, whole-genome-amplification (WGA), and single cell DNA sequencing to generate spatially resolved single cell genomic data. Using whole-tissue and cutting slides, TSCS can provide an estimated Z-axis,

while collecting the X and Y locations and morphology via images and mapping the genome of each single cell isolated from a tumor. The full description of the method and data analysis for TSCS are provided in the Materials and Methods section.

### **1.6.2 Intratumor Heterogenetic during Invasion in Breast Cancer**

The second issue my thesis covers is using TSCS to examine intratumor evolution and heterogeneity in synchronous DCIS-IDC breast cancer. Since breast cancer, like other cancers, is profiled by the loss of genetic control, specific changes in ploidy, proliferation, and apoptosis of cells within the breast have been documented. NGS studies have examined the genetic aspects of breast cancer across patients (intertumor heterogeneity) and within a tumor (ITH).

ITH has been frequently reported in invasive and even *in situ* breast cancers, making breast cancer a good candidate to study cellular evolution. We used TSCS to investigate the three models of invasion discussed earlier: independent evolution and the evolutionary bottleneck and multiclonal invasion versions of direct lineage. Based on our data, we concluded the direct lineage model with a single initiating cell (N1) is the most probable. We were further able to delineate between the population bottleneck and multiclonal evolution versions of the direct lineage model. Due to the limited number of changes in the frequency of clones in each population, we concluded that the multiclonal invasion version of the direct lineage model is most probable.

In *multiclonal invasion*, multiple clonal populations are migrating from the ducts into invasive regions. While my data is consistent with a *multiclonal invasion* version of direct lineage, much is being inferred based on the lack of significant change in most subpopulations. Further sampling is needed to completely distinguish between the population bottleneck and multiclonal invasion version of direct lineage model. Given the heterogeneity found in cancer, the multiclonal model could apply to some patients and the population bottleneck model to others.

## 2 Materials and Methods

This section is based on the research paper "Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing" published in the Cell in 2018, by Casasent et al<sup>1</sup>. Figures from this paper have been reused or modified under the journal's academic copyright license for student thesis usage. This section is expanded from paper and included details and tips about the TSCS protocol.

In this section I cover the materials and methods required for my project to examine the intratumor heterogeneity (ITH) and evolution of subclones in synchronous ductal carcinoma. The study was approved by the IRB at the University of Texas MD Anderson Cancer Center.

### 2.1 Sample Selection

We used synchronous breast cancer samples, since they provide several advantages over 'pure DCIS' and recurrent IDC samples for our purposes. First, synchronous samples are from the same time point, while a 'pure DCIS' sample with an accompanying recurrent IDC sample are collected years apart, often after confounding treatment.

In addition, the cohort size for a well powered longitudinal (pure or recurrent DCIS) study requires a larger number of samples than synchronous DCIS-IDC. Using the estimate of likelihood for low grade DCIS to progress to IDC (15% over 10 years)<sup>76, 195</sup> and high grade (50% over 3 years)<sup>76, 195, 196</sup>, we estimate that for a low-grade study we would need 103 enrolled patients and for high grade 28 enrolled patients to have a 95% confidence of getting at least 10 samples for the study. At MD Anderson, the recurrence rate of DCIS is reported to be 6%<sup>197</sup>, in this case we would need 260 enrolled patients to have a 95% confidence of getting at least 10 samples for the study. We calculated these numbers using the negative binomial function in R selecting of selecting at least 10 samples within a defined probability (Figure 5 Estimates of Number of Samples Needs for Longitudinal Studies). Even with 260 patients enrolled, this would require an impractical 100% compliance of these patients over 10 years.

In addition to the difficulty of tumor collection, the results from longitudinal studies may be confounded by (1) intervening therapies, (2) continued evolution, and (3) changes in sampling area. The confounding factors could lead to coincidental mutations improperly associated with invasion. The value of using synchronous samples (temporally and spatially matched samples) to study invasion in breast cancer have been highlighted in several papers<sup>114, 86, 119, 6</sup>.

Therefore, we used treatment-naïve synchronous (DCIS-IDC) tissue samples to infer tumor evolution and progression during invasion.

## **2.2 Human Samples Description**

We examined 10 treatment-naïve synchronous (DCIS-IDC) tissue samples with paired normal samples from adjacent breast tissue, obtained from the University of Texas MD Anderson Cancer Center Tissue Bank. Frozen tissue selection was based on the presence of both *in situ* and invasive regions, validated by a pathologist before processing of samples. We also required samples to have paired normal adjacent breast tissue, for normal control in exome regional sequencing.

For all tumors, we scored ER, PR, and HER2 status scored separately for the *in situ* and invasive regions. Only one tumor DC17 had any difference observed in the staining of these receptors. Negative ER and PR status of <1% was determined by IHC following the 2010 American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations<sup>198</sup>. Negative HER2 amplification status was defined through FISH analysis using a CEP-17 centromere control probe (ratio of Her2/CEP17 < 2.2). Five of the ten samples were classified as TNBC based on negative staining for ER, PR, and HER2. Receptor status and clinical parameters such as age, stage, grade, and number of cells collected per region are in Table 2 Clinical Information.

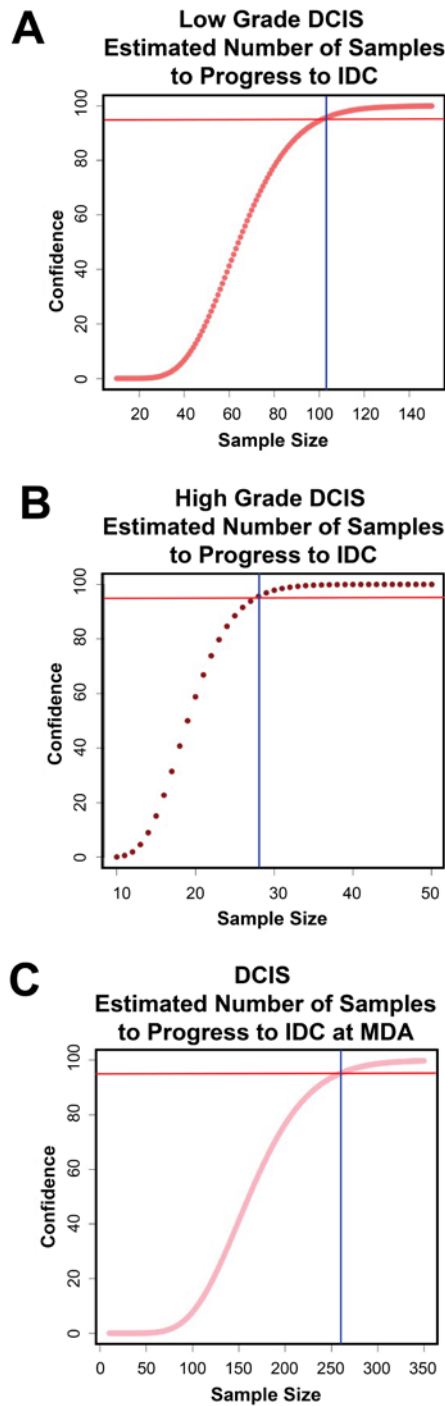
Our patient cohort consisted of 5 TNBC (ER-, PR-, HER2-) and 5 estrogen-receptor positive (ER+) breast cancer patients. Most of the patients had high grade tumors, apart from



DC12 and DC13, which were grade 1 and 2, respectively. On average, about 129 single cells per patient passed our filtration criteria and had genome-wide copy number clonality analyzed.

In addition to single cell analysis of CNA, we collected laser-capture-microdissected clumps of thousands of tumor cells from the *in situ* and invasive regions. We then isolated DNA for exome analysis from the laser captured bulk regions to compare *in situ*, invasive, and matched normal mutations. The bulk region exome DNA was sequenced at high coverage depth (mean=162.8X, SEM=18.9) to detect somatic point mutations (SNPs). Matched normal tissues were sequenced at a slightly lower coverage (mean=144.1X, SEM=20.3) to identify and filter germline variants (Table 3 Exome Coverage).

Figure 5 Estimates of Number of Samples Needs for Longitudinal Studies



This figure depicts the negative binomial to represent the probability of selecting at least 10 samples for different expected probabilities of progression needed for a longitudinal study of different expected rates of progression. The x-axis represents the number of samples. The y-axis shows the cumulative distribution function (cdf) of the negative binomial distribution with the parameters specified, with values rescaled by 100 to be interpretable as percentages. The horizontal red line represents 95% confidence. The vertical blue line shows the selected number of samples closest to the 95% confidence. (A) Low grade (15%). (B) High-grade (50%). (C) Reported DCIS progression at University of Texas MD Anderson Cancer Center (6%).

Table 2 Clinical Information

Synchronous DCIS-IDC Patient Cohort								
Patient	Age	TNBC	ER	PR	HER2	grade	Stage	Cells
DC4	57	Y	-/-	-/-	-/-	3/3	IIB	57
DC6	36	N	+/+	+/+	+/+	3/3	IIB	114
DC12	64	N	+/+	+/+	-/-	1/1	IV	102
DC13	66	N	+/+	+/+	-/-	2/2	IIIC	104
DC14	47	N	+/+	-/-	-/-	3/3	IIA	148
DC16	77	Y	-/-	-/-	-/-	3/3	IIA	204
DC17	66	N	-/+	-/-	-/-	3/3	IIIC	112
DC18	62	Y	-/-	-/-	-/-	3/3	IIA	235
DC19	49	Y	-/-	-/-	-/-	3/3	IIA	96
DC20	48	Y	-/-	-/-	-/-	3/3	IIA	122

This table is adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

This table contains clinical information on the 10 patients with synchronous DCIS-IDC tumors that were analyzed by single cell and exome sequencing in this Thesis.

Clinical parameters listed include patient age, triple-negative breast cancer status, estrogen, progesterone and HER2 receptor status, tumor grade and tumor stage.

The receptor status and grade were scored independently for the DCIS and IDC regions and are displayed on the left (DCIS) and right-hand (IDC) side in these columns. The total number of single cells analyzed by TSCS is also indicated in the last column. DC17 is the only sample with a change in receptor status between *in situ* and invasive regions. This change in the ER receptor status was from less than 1% to close to 10% and, previous to 2010, both would have been marked as negative.

Table 3 Exome Coverage

Exome Sequencing Metrics						
Sample Number	Normal		In situ		Invasive	
	Depth	Breadth	Depth	Breadth	Depth	Breadth
DC4	107	0.9599	116	0.9628	58	0.9626
DC6	76	0.9601	142	0.9553	124	0.9536
DC12	100	0.9587	187	0.9665	116	0.9597
DC13	153	0.9543	136	0.9538	140	0.9524
DC14	104	0.9571	46	0.9458	125	0.9526
DC16	87	0.9558	180	0.9607	89	0.9539
DC17	177	0.9552	211	0.9561	136	0.9501
DC18	144	0.9584	315	0.965	105	0.9577
DC19	280	0.9638	335	0.9635	298	0.9621
DC20	213	0.9608	287	0.9628	110	0.9528
Mean	144.1	0.95841	195.5	0.95923	130.1	0.95575

This table is adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

This table shows the exome sequencing metrics for the 10 DCIS-IDC patients, in which laser-capture-microdissection was used to isolate *in situ* and invasive regions from frozen tissue sections. Matched normal breast tissue was sequenced in parallel.

Coverage depth was calculated for the *in situ* regions (mean=195.5X, SEM=29.4), invasive regions (mean=130.1X, SEM=20.1) and normal tissues (mean=144X, SEM=20.3).

Coverage breadth, or physical coverage, was also calculated from the targeted exon regions for each patient and tissue region.

Coverage breadth is defined as the percentage of the targeted capture regions in which at least 1X coverage depth was achieved after sequencing.

## **2.3 Single Cell Copy Number Protocol**

The single cell copy number workflow takes about 5 days to generate data for about 48 cells. One of the most time-consuming steps is single cell collection which usually takes 4-5 hours for ~28 cells. A maximum of 48 single cells were collected by LCM in one sitting. When increasing the number of cells collected per sitting the percent of cells that passed quality control appeared to decrease, while processing 24-32 cells at one time provided consistent results. After collection, LCM collected cells were amplified using single cell WGA (6-8 hours). After WGA, quality control was used to filter out low quality samples. These steps were repeated until we had 48 to 96 cells. Then Illumina NGS libraries were prepared. The purified WGA single cell DNA was sonicated (~2.5 minutes per single cell, which is 2.5-3 hours for 48 cells or 4.5-6 hours for 96 cells) followed by NEB Illumina NGS library preparation protocol, which takes about 5 hours. Figure 6 Timeline of TSCS Protocol provides the time line of this workflow.

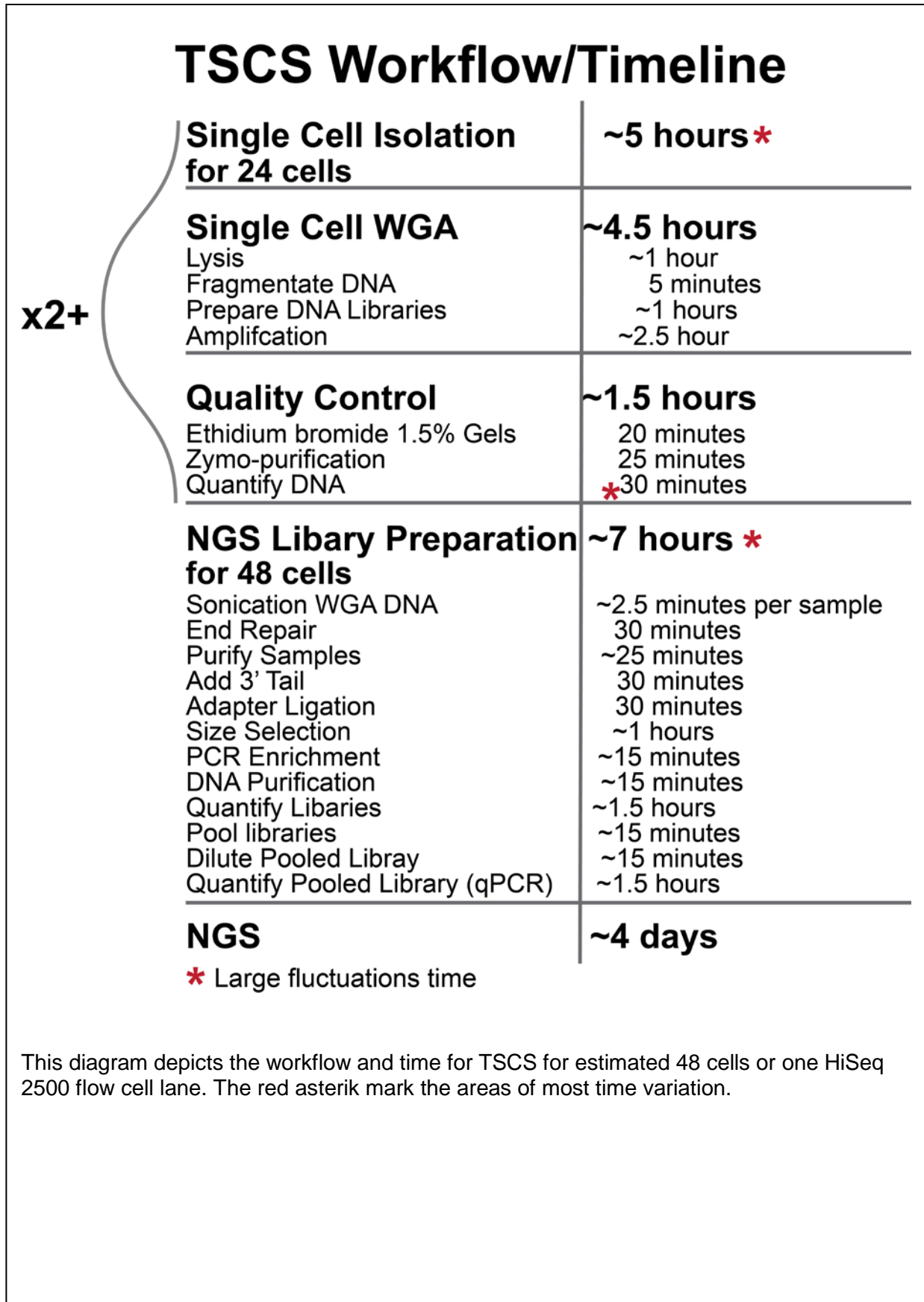
### **2.3.1 Single Cell Isolation**

The first step of any single cell protocol is the isolation of single cells. In this case, we started with fresh frozen tumor tissue. In this section I will detail the protocols for Slide Preparation, Tissue Staining, and Single Cell Isolation with Zeiss PALM Robo System.

#### **2.3.1.1 Slide Preparation and Staining**

Since we desired single cells with spatial information and morphology, individual vials were divided into sectors (<1mm cubes). The tissue sector was mounted on OCT Compound (Tissue-Tek, Cat# 25608-930) and allowed to equilibrate to the Thermo Scientific CryoStar (NX70) or Leica Cryostat (cm3050S) temperature. To reduce smearing of the fat tissue in breast samples, we cut tissue at -23°C to -27°C with blade temperature at least 2°C below the ambient temperature. Each sector was then divided into sections (slices of tissue).

Figure 6 Timeline of TSCS Protocol



During the 15 minutes needed for tissue to equilibrate to the cryostat temperature, we treated the LCM PEN-membrane slides (Carl Zeiss Microscopy, Cat# 415190-9041-001) for 15 minutes with ultraviolet light (UV). UV treatment was recommended by the manufacture<sup>199</sup>, to prevent contamination and help tissue adhere to the slide membrane. Slides were cut to generate

- (1) a set of 1-2 slides with two or three 6-micron thick sections per slide for H&E visualization staining,
- (2) 1-3 LCM slides with 4-6 sections per slide at 12-14 microns thick,
- (3) an additional visualization slide was prepared after every 3 LCM slides or every 16-24 tissue sections, and at the end of every set of sections we cut another visualization slide.

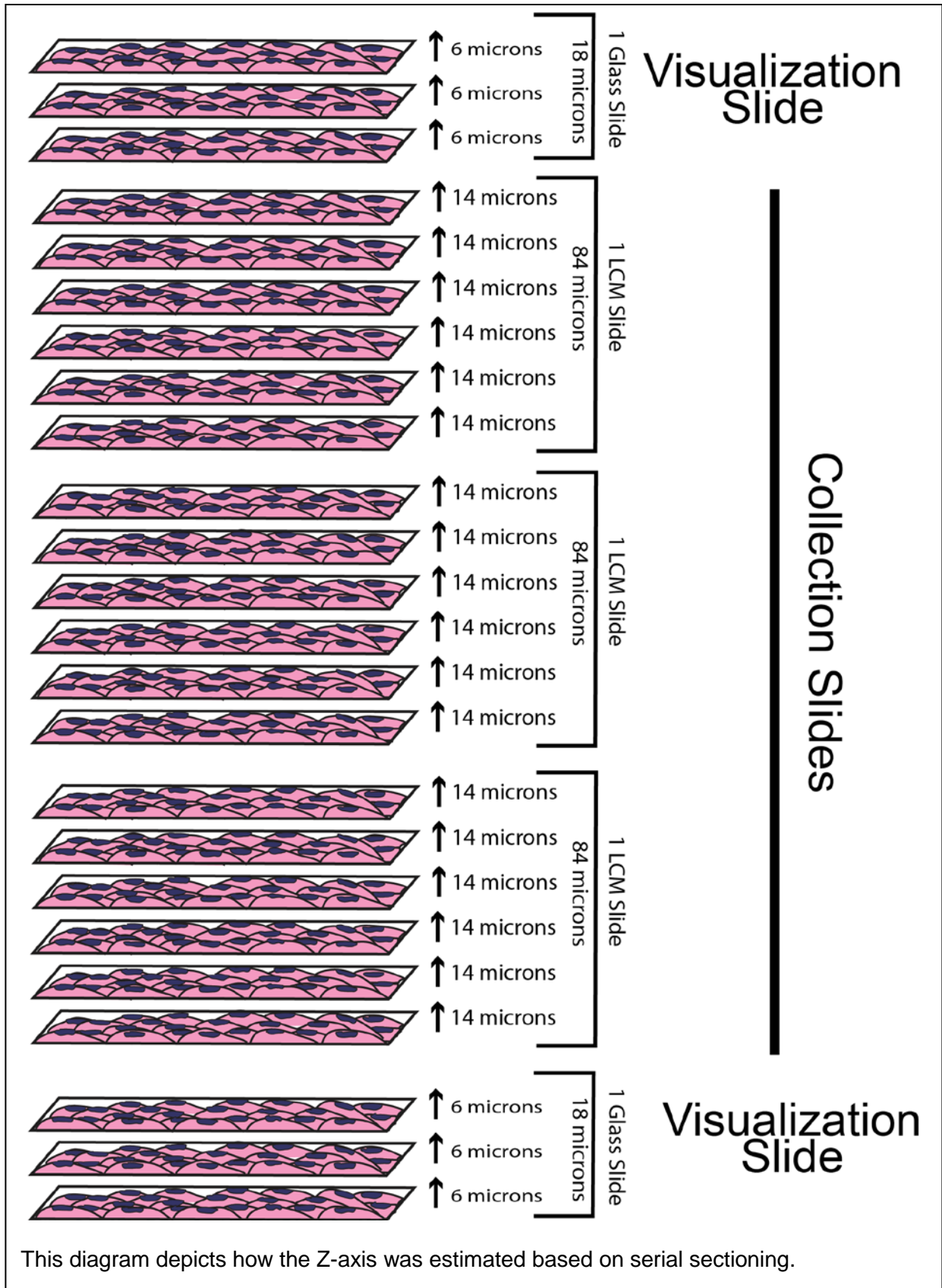
Figure 7 3D Slide Stacking demonstrates the cutting order for 2 rounds of LCM slides.

After cutting sections, slides were air dried at room temperature for about 30 seconds to increase adherence to the slide, and then slides were placed in the cryostat to keep temperature below -20°C until fixed. All PEN-membrane slides were fixed in 70-75% ethanol, instead of the 95% ethanol used on the glass visualization slides. Zeiss PALM DNA Handling Protocol<sup>199</sup> suggests using a ~70% ethanol concentration to prevent ethanol from damaging the membrane. All slides were stained using Harris' Alum Hematoxylin (VWR Cat#638A-71) and Eosin Y (VWR Cat#588X-75).

To save glass visualization slides for future use, slides were placed in the -80°C freezer in a sealed container prior to fixing. Glass and PEN-slides can be fixed and stored in the -20°C freezer until use (usually 1 week). If fixed and stained, slides were placed in the 4°C (PEN-slides) or maintained at room-temperature (glass slides).

Note: Do not place the PEN membrane slides into the -80°C freezer, since this causes bubbles in the membrane that interfere with LCM.

Figure 7 3D Slide Stacking





### 2.3.1.2 Single Cell Isolation via Laser Capture Microdissection

Some tips for LCM single cell capture efforts are:

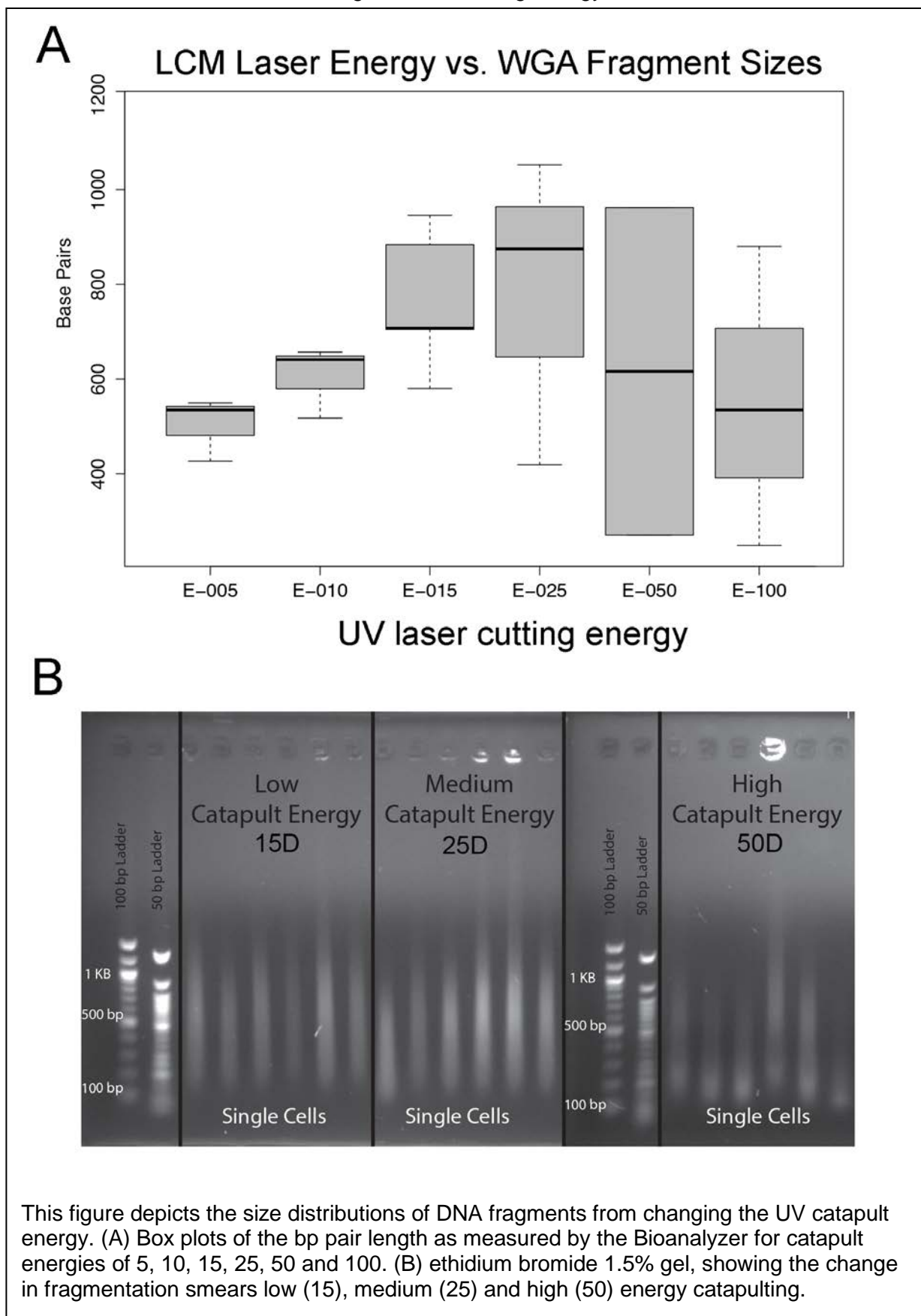
- (1) the LCM system must be in a room with stable temperature and humidity;
- (2) the LCM system must be turned on for at least 1 hour before laser use;
- (3) the LCM system has a minor slant in the robo-mover; therefore, when utilizing the 12x8 cap strip collector, only the 1<sup>st</sup> three cap strips should be used;
- (4) finally, the LCM collector should be centered and the orientation of X, Y, and Z calibrated before each use.

All tissues were scanned at 10X, which was used to identify single cells as *in situ* or invasive and facilitate selection via histology (size, shape, and location to nearest duct). *In situ* regions were marked with a blue flag and invasive regions were marked with a red flag at 10X magnification. To examine single cells, we used the 63X setting. For selection as grade 3, all tumor cells had to be 2x to 2.5x the size of a normal lymphocytes. Nuclei were selected to be oddly shaped, overly large, and least 1 micron (preferably >2 microns) from surrounding nuclei. All nuclei of interest were marked and labeled as *in situ* or invasive and recorded in the elements file and a separate Excel spreadsheet. Brightfield images at 63X were collected before and after capture of each single cell.

PALM Robo wizard (Carl Zeiss) was used to optimize the UV cutting parameters. The optimal energy for laser-catapulting single cells was set between 20-25 delta to reduce DNA fragmentation and increase collection efficiency. Delta settings below 15 resulted in frequent cell transfer failures by laser catapulting (Figure 8 UV Cutting Energy).

Details for each collection were recorded, such as if the lysis buffer was centered or to the side of the cap, the number of catapult attempts, the results of a visual inspection of the collections, and any other notes. A post collection image was collected to record results of the collection (missed or fragmented nucleus) or if other cells were damaged during capture.

Figure 8 UV Cutting Energy



After collection, the strip of caps for the 8-well Polymerase Chain Reaction (PCR) tubes were sealed and placed on ice with cap side down. If cell collection took longer than 3 hours, the first subset of sealed tubes were started on the thermocycler, and collection continued with a second subset. When collection was complete and all cells were on the thermocycler, a full image section and additional zonal element images were recorded for mapping cell locations.

The objective was then changed from 63x to 10x and a full image section was taken by using the scan feature on the PALM system. Next, the show element option was selected and the visual area was moved to contain as many elements (collected cell markers) as possible, to make a zonal elemental image. Each of these zonal elemental images must be captured manually. The 10x full section scans and zonal element images were used in mapping to the image space across tissue sections. In addition, the element files were saved as PALM and text format. Scans were saved with standardized file names for record keeping.

### **2.3.2 Single Cell Processing**

As described above, single cells were laser-catapulted into 8 well strips of 0.2 ml PCR tube caps containing 10 $\mu$ l of lysis solution from Sigma-Aldrich GenomePlex<sup>®</sup> WGA4 kit (cat# WGA4-50RXN) using CapCollector 12x8 attachment (Zeiss Ref# 415101-2000-911) for robotic automation with Zeiss PALM Microbeam RoboMover. After capture, the cells and lysis buffer were spun down at 12,000 rpm for 30 seconds. Single cell DNA was amplified using Degenerative-Oligonucleotide-Primer PCR (DOP-PCR from Sigma-Aldrich GenomePlex<sup>®</sup> WGA4 kit (cat# WGA4-50RXN)). The details of the Single Nucleus Sequencing (SNS) protocol were previously described in Nature Methods by Baslan et al<sup>7, 4</sup>.

We made six alterations to the SNS protocol:

- (1) We collected from LCM slides.
- (2) We collected less than 96 cells at any one given time. We usually collected 24-32 cells in one sitting.
- (3) We collected into the caps of 8 well strips of 0.2 ml PCR tube, not directly into plates.  
Because we captured into caps, we had to make sure to seal the caps to the tubes before spinning the lysis buffer and cells down.
- (4) We eliminated wells considered to have malfunctions during sample collection, either from LCM cut or catapulting based on visual examination during the procedure. This visual examination worked to minimize incomplete captures and over captures, such as doublets or extra-cellular contamination from neighboring cells. In addition, cells that required too many catapults or splattered during catapulting were eliminated.
- (5) In step 11-12 of the Nature Methods paper<sup>7</sup>, we did not mix a master mix for the library preparation buffer and enzyme. Instead we add 2uL of the buffer, spin down, flick the strip to mix, pulse spin down a second time, add 1uL of enzyme, pulse spin down, flick to mix, and pulse spin before placing on the thermocycler. This process tended to produce better results than premixing the enzyme but did lead to a little more library enzyme loss. Since the library enzyme was not the limiting reagent in the kits, the minor loss was not an issue.
- (6) Since we eliminated wells, did not use a 96 well plate and expected only around 70% of captures to succeed, we did not follow the 96 well purification protocol in the Nature Methods paper<sup>7</sup> (items 19-30). Instead, we began with a quality control step for WGA DNA in which size distributions were determined through electrophoresis and only samples with fragment sizes >300bp were selected and purified (Genesee Cat # 11-303). Purified WGA DNA was measured on a Qubit 2.0 Fluorometer (Fisher Cat#Q32854) and samples containing > 200ng of DNA were selected for library construction and next-generation sequencing.

### 2.3.3 NGS Library Preparation

Using the single cell amplified DOP-PCR products that passed quality control in the previous step we measured the concentration of the samples and used 200ng-500ng of DNA in 87uL of water. DNA was sonicated to 250bp using the S220 acoustic sonicator (Covaris). After sonication we followed methods like those used by Goa et al 2016<sup>200</sup> in which the DNA fragments were treated for end repair (New England BioLabs, E6050L). After repair, DNA was purified to remove the end-repair enzyme using DNA Clean and Concentrator-5 kit (Genesee, 11-303 or 11-306). NGS libraries were made using the NEBNext DNA library prep enzymes (New England BioLabs), 3' adenylation (E6053L), ligation (E6056L/M0202L) followed by PCR amplification (M0541L). These protocols were based on the manufacturer's instructions.

Our three alterations to the protocol from the version published in Gao et al 2016<sup>200</sup> were:

- (1) we increased the ligation time (20°C for 30 minutes) during library preparation,
- (2) we based the PCR amplification cycles on input DNA (8 cycles for 1ug, 9 cycles for 500ng, and 10 cycles for 200ng), and
- (3) we used P7 adaptors for each single-cell library used unique 6-bp and a common P5 adaptor to allow sequencing.

After ligation and before PCR amplification, size selection was used to remove over-ligated DNA strands and primer-dimers using AMPure XP beads (Beckman Coulter, A63881), 0.7X (removes large fragments) and 0.15X (removes small fragments). After PCR amplification, DNA was purified using AMPure XP beads (Beckman Coulter, A63881) at 1X. DNA concentrations were measure using the Qubit 2.0 fluorometer and concentration was used to pool 48 libraries to equimolar amounts. The pooled libraries were measured using KAPA Library Quantification kit (KAPA Biosystems, KK4835) and ABI PRISM real-time PCR machine (Applied Biosystems 7900HT).

Multiplexed libraries were sequenced for 76 cycles using single-end or paired-end flow cell lanes on the HiSeq2000 or HiSeq4000 systems (Illumina, Inc.).

## **2.4 Single Cell Copy Number Data Analysis**

In this section I discuss the data processing steps and analysis performed on single cell copy number data. There will be some similarities between this section and that for exome sequencing. However, aside from initial data processing steps, the analysis of regional exome sequencing is divergent enough to warrant a separate section.

### **2.4.1 Single Cell Copy Number Data Processing**

This section covers data processing for single cell copy number data.

#### **2.4.1.1 Genome Alignment**

Multiplexed single-cell FASTQ files corresponding to the single cell samples were deconvoluted using 1 mismatch of the 6pb barcodes. The deconvoluted FASTQ files were aligned to hg19 (NCBS build 36) using Bowtie 2 (2.1.0) alignment software<sup>201</sup>. The aligned reads were converted from SAM files to BAM files, then sorted using SAMtools (0.1.16). PCR duplicates were marked and removed using SAMtools<sup>202</sup>. The sequencing data was processed following the ‘variable binning’ pipeline<sup>7, 203</sup>. After aligning reads, genomic regions were separated into ~220kb variable bins and the number of reads per bin was counted. The script used for ‘variable binning’ is directly from the Wang et al paper<sup>17</sup>.

#### **2.4.1.2 Circular Binary Segmentation (CBS)**

Unique normalized read counts were segmented using the circular binary segmentation (CBS) method from R Bioconductor ‘DNAcopy’ package<sup>204</sup>. The CBS algorithm was developed by Olshen et al in 2004<sup>205</sup>. CBS takes continuous or binary data and recursively splits each segment (in our case each chromosome) into either 2 or 3 sub-segments based on the maximum t-statistic, and then each sub-segment is compared between a permuted reference distribution and the sub-segment’s actual distribution to determine if it should be split. Whether or not to eliminate a split is decided by whether the two segment means are distinct enough. An alpha( $\alpha$ ) of 0.0001 was selected for probability of a Type I Error, which produces very

hypersensitive segmentation. We used "undo.prune" of 0.05 to reduce the sensitivity of segmentation for splits where the proportion of sum of squares between splits increases by less than 0.05. This was followed by use of MergeLevels to join adjacent segments with non-significant differences in segmented ratios to further reduce over segmentation. Default parameters were used for MergeLevels, which also removed erroneous chromosome breakpoints. This script was also used in the Gao et al paper<sup>16</sup>.

## **2.4.2 Data Quality and Filtering**

Data was filtered to remove data with more than 100 break points or identified as noise. Density-based spatial clustering of applications (DBSCAN) uses the density of the data points in a user set space, to group together the most tightly packed points with their nearest neighbors. If points are outside of the resulting groups, they are marked as outliers. The R package for DBSCAN<sup>206-208</sup> used 'dbscan' (v1.1-1)<sup>209</sup> for the noise portion. We examined the k-nearest neighbors plots (with k set from 2 to 15) to find the elbow and recorded this value for selecting the distance allowed from a point to the edge of the nearest cluster or "eps". Once the eps was selected, data was filtered to exclude all "noise points". Using the eps number, dbscan determined which single cell copy number segmented samples exhibited too much technical noise and filtered approximately 20% of the total datasets for each patient.

## **2.4.3 Clustering**

This section covers our clustering performed on single cell copy number data.

### **2.4.3.1 Determining K**

K-means clustering requires an optimal k to be set by the user to partition the data. In order to find the optimal 'k' (number of clusters between 1-15), we used the R-package 'cluster' and started with the clusGap function in that package (K.max=15, B = 100, d.power = 2, FUNcluster = kmeans). We ran the results from clusGap through maxSE (method="firstSEmax") to select the best number of clusters for k-means clustering. Our

subclones were defined using K-means clustering with multiple start sites with the number of clusters ( $k$ ) selected by the smallest  $k$  which the  $f(k)$  is no more than 1 standard error away from the first local maximum, this forces the  $k$  to be lower and prevents over clustering<sup>210, 211</sup>.

#### **2.4.3.2 K-means Clustering**

After we determined the optimal number of clusters ( $k$ ), we calculated K-means with  $k+1$  using the previously selected  $k$  to provide relationships within clusters. K-means clustering was done by splitting the data into the given number of clusters ( $k$ ) with each data point being assigned to the nearest cluster using centroid distances. However, the original start sites for the clustering will cause the partitioning to be different even for the same dataset. Therefore, we calculated a k-means matrix using 500 original start sites for  $k+1$ .

Next, we used ward.D2 clustering to generate the genetic trees based on the k-means matrix. The tree was cut into  $k$  clusters to define "subclones". The internal Pearson and Spearman correlation of the samples within each "subclone" was calculated. Most cells with technical noise were removed in the previous filtering steps; however, in a few cases, we identified additional cells with an internal correlation of Pearson and Spearman of less than 0.2, which were excluded from further analysis as "noisy profiles". There are a few possibilities why these profiles could be occurring: (1) technical noise from cutting too close to another nucleus, (2) technical noise from UV cutting or catapulting damaging the DNA during transfer, or (3) a biological rationale for this noise such as a mutator phenotype or biological dead-ends.

#### **2.4.4 Subclone Analysis**

A major part of this project was examination of subclones. Major questions concerned the differences between subclones genetically and spatially relative to other subclones. Subclones were defined using k-means as described above. Before further genetic analysis, we determined if we had collected enough single cells per tumor to identify rare clonal types.



#### 2.4.4.1 Sample / Power Calculation

Since the required number of cells depends on how many clones were discovered, we first sequenced 30-50 cells per region (DCIS and IDC). We analyzed these results and estimated the number of subclones per tumor using the k-means clustering described previously. To determine if we sequenced enough cells to discover the "major" subpopulations in both *in situ* and invasive regions, we performed a *post hoc* saturation analysis as described by Gao et al in 2016<sup>16</sup>.

Based on the results for subclones, we defined the total number of subclones and fractions of each region for each patient. We used these values to calculate a cumulative multinomial distribution (an expansion of the generalization of a binomial distribution from 2 variables to many) probability of observing at least 2 tumor cells in each subpopulation, given the numbers of cells sequenced in our experiments. Note that if this is increased to a higher number, 3 cells for example, more cells would be required.

The multinomial distribution method requires at least 2 subclones or subpopulations. Even in our "monoclonal" samples, we usually collected a few normal cells which were usable in this calculation as a second subclone. However, in one case, DC17, only tumor cells were collected, so DC17 was excluded from this analysis. To provide the best estimate of the number of tumor cells required, we calculated the cumulative distribution for both the *in situ* and invasive regions, and then also pooled the regions with a weighted average to obtain a total number of cells needed per patient with our current number of subclones. Dr. Ruli Gao provided significant assistance with this portion of the analysis.

Figures providing summaries of the *ad hoc* saturation analysis for all tumor samples used in this project are provided in the individual tumor sections, except for the excluded DC17.

#### **2.4.4.2 Calculating Diversity**

Within each tumor, the amount of subclonal diversity was defined to represent ITH, the number of subclones with the normal subclone cluster removed. The normal subclone cluster was defined by having a high internal Pearson correlation, but low Spearman correlation. These normal profiles were removed from subclone analysis, while the remaining  $k$  or  $k-1$  clones were considered tumor subclones.

Then, we calculated the subclonal diversity index for each tumor by first determining the proportion ( $p$ ) of cells that belong to each distinct subclone within a given tumor. We used the Shannon Index<sup>191, 192</sup> ( $Dc = -\sum_i (p_{ix} \ln p_i)$ ) to calculate diversity within the tumor. The Shannon diversity index uses larger values to signal higher subclonal diversity. The Shannon diversity index represents both numbers of clones and equality of clones<sup>191, 192</sup>. Therefore, the highest diversity measurements would be tumors that have both more clones and most equal proportions of each clone.

#### **2.4.4.3 Cancer Genes**

Cancer genes were annotated using the 413 genes compiled from multiple databases including the Cancer Gene Census<sup>212</sup>, The Cancer Gene Atlas Project (TCGA), and the NCI cancer gene index (Sophic Systems Alliance Inc., Biomax Informatics A.G) used in previous publications<sup>16, 17</sup>.

#### **2.4.5 Topographical Analysis**

We collected topographical data during the collection of single cells. The spatial XY coordinates of each cell were defined by a projection of the stacked tissue section layers to project the XY coordinates into the same space. We mapped zonal elemental images to the 10x section scans images for each respective section. When multiple tissue sections were used for collection, we projected X and Y coordinates to the same space. Projection was accomplished by mapping the tissue section scans onto an H&E tissue scan facsimile.

The H&E facsimile was selected from the section scans either (1) from LCM H&E tissue sections from which cells were collected, or (2) from the H&E visualization slide which was used to verify regional pathology. The H&E facsimile was the section with the largest tissue area to make sure all cells were mappable to the scan. The Z-axis was estimated based on the number of sequential sections cut. The Z-axis can also be estimated based on changes in known ducts between the H&E facsimile and the LCM section from where the cells were collected. In cases where tissue sections had different orientations (because the sections are often slightly rotated between each cut), we rotated the spatial coordinates and transposed the coordinate values to project them into the correct space. We call these projections "Image Maps".

#### **2.4.5.1 Image Maps**

Image Maps are the projection of the XY coordinates of cells based on subclones for all stacked tissue section from a tissue sector onto the H&E section facsimile. The appropriate XY coordinates were color coded according to clonal genotypes using the H&E section facsimile as the coordinate space. We facilitated tracking of the three-dimensional duct network by enumerating the duct and using false-color outlines. Image map figures are provided within each tumor section below.

#### **2.4.5.2 Tanglegrams**

Tanglegrams provide another method to visualize the relationship between location and genetic information. Tanglegrams were designed to visualize co-evolution between samples<sup>213</sup>. We used tanglegrams to compare the spatial coordinates to the subclones. We created spatial trees using Euclidean distance between cell coordinates and clustered with the R `hclust` function using "ward.D2" linkage. We occasionally had multiple vials for the same tumor. Cells of the same tumor from different sample vials were given an artificial distance to buffer samples in which the distance between the regions of the tissue sectors was unknown. The second tree

we created was the subclone tree, based on k-means clustering of copy number profiles with 500 different start sites.

Finally, the spatial and subclone trees were compared to examine the relationship between subclones and location using Tanglegram version 1.5.2 from the dendextend package in R<sup>213</sup>. Untangling the tanglegrams makes the relationship between space and subclone more discrete. We performed untangling by flipping nodes to minimize branch crossing. The minimization method first tried 100 random shuffles, selecting the one with the lowest crossings. From the initial shuffle, a local stepwise method was used to reduce crosses at each node. Aislyn Schalck contributed significantly to this analysis. Tanglegrams are provided within each tumor section below.

#### **2.4.5.3 Morphological Analysis Options**

As described earlier, in addition to capturing 10x tissue section scans, we also collected brightfield images at 63X magnification before and after laser-catapulting. The brightfield images assisted in confirming collection of single cells were complete and without adjacent material from neighboring cells. Additionally, we used these images to validate collection location from *in situ*, invasive, or stroma regions.

### **2.5 Regional Exome Protocol**

The current Topographical Single Cell Sequencing (TSCS) method does not examine single cell mutations (single nucleotide variants (SNV) or Indels). Therefore, we used a more standard microdissection method to compare regional *in situ*, invasive, and normal DNA for exome mutations.

### 2.5.1 Exome Laser Capture Microdissection

Regional exome capture was done on fixed H&E stained slides after single cell collection was completed. Using the same slides ensures the exome samples are as close to matched as possible with the single cell sample from these same regions.

Before LCM, all tissue slides were scanned, with each region of interest marked as *in situ*, invasive, or stroma. Each region was selected using the UV cutting laser and captured using the UV catapult components of the PALM System (Carl Zeiss). Validation of collection from proper regions was done via review of the 6µm visualization H&E slide with a pathologist (Dr. Mary Edgerton). validating the *in situ* and invasive regions prior to LCM regional collection. Thousands of cells from both *in situ* and invasive regions were catapulted into 2 mL adhesive PCR tube caps (Item #: Zeiss 415190-9181-000 Wor 415190-9191-000).

Since relatively large sections were being collected, the catapult energy could be higher and the focus spread across the collection region. We used 50-100 delta instead of the lower 15-25 delta energy used for single cells. In addition, we increased the UV laser cut energy and spread to 72-81.

### 2.5.2 Exome DNA Isolation

Selected regions were captured into PCR caps and the DNA was isolated. We used QIAamp DNA Micro Kit (QIAGEN Cat# 56304). DNA was incubated overnight at 56°C to increase DNA yield.

Since all the LCM collected tissue was from *in situ* or invasive regions, additional tissue was needed as a matched normal. A small 0.25mm<sup>3</sup> to 0.5mm<sup>3</sup> of fresh frozen adjacent normal tissue was macerated in preparation for DNA isolation. DNA was isolated using the DNeasy Blood & Tissue Kit (QIAGEN Cat# 69506).

DNA concentrations were quantified by Qubit 2.0, while normal DNA was isolated using the QIAGEN DNeasy protocol (Cat # 69506).

### 2.5.3 Exome Capture

Exome libraries were constructed for the *in situ*, invasive, and normal DNA. The DNA was sonicated into 200 bp fragments using the Covaris Sonicator. We manufactured Next-Generation Sequencing (NGS) libraries from this sonicated DNA. The sonicated DNA was treated with NEBNext end repair (NEB, E6050L), which was removed using Zymo DNA Clean & Concentrator Column Kit (Genesee Cat # 11-303). Libraries were constructed by adding dA-tailing module (NEB, E6053L) and quick ligation module (NEB, E6056L) with barcodes to multiplex libraries. Barcoded next-generation sequencing library amplification was via NEBNext HiFi 2x PCRmix (NEB, M0541L). Libraries were measured using Qubit 2.0 Fluorometer and measured by quantitative PCR using the KAPA Library Quantification Kit (KAPA Biosystems, KK4835) before pooling. The 3-8 barcoded samples were pooled in equimolar concentrations for exome capture.

Exome capture was via Nimblegen's SeqCap EZ Exome V2 kit (Roche, 05860482001), which according the company website was designed using GRCh37 (hg19) to capture more than 20K genes. The total capture covers about 44.1 Megabases.

The Exome Captured pooled samples were measured by quantitative PCR using the KAPA Library Quantification Kit (KAPA Biosystems, KK4835) and diluted to 10nM before processing by the University of Texas MD Anderson Cancer Center Sequencing Core. Sample sequencing was performed on a 100 paired-end flowcell on the Illumina HiSeq4000 system.

## 2.6 Exome Regional Data Analysis

The regional exome data was provided to us as sequence reads in FASTQ files from the University of Texas MD Anderson Cancer Center Sequencing Core. Below are the details of the processing of the regional exome data from FASTQ files to adjusted variant calls.

### 2.6.1 Exome Data Processing

The FASTQ files for *in situ*, invasive, and normal samples were then aligned to the hg19 using the Bowtie 2 alignment software<sup>201</sup>. To increase processing speed, we converted the

aligned data using Samtools (0.1.16)<sup>202</sup> from SAM files to compressed BAM files. BAM files were sorted by coordinate from the hg19 alignment. This conversion allowed duplicates to be marked and removed by Picard<sup>214</sup>.

### **2.6.2 Exome Regional Data Quality and Filtering**

One of the major issues for bioinformatics analysis is determination of sample or read quality. We required each sample to have 20 reads per SNV with at least 5 reads for a variant, which is fairly strict and intended to prevent false positives. We also filtered germline SNPs identified in the matched normal tissue samples and the tumor samples. Reads were detected using GATK. GATK was run with default parameters for depth (maximum read coverage = 250x). We generated a multi-sample VCF file. Next, the variant quality score recalibration was performed using training sets for SNVs or indels. The output from GATK improves variant quality scores.

### **2.6.3 Exome Regional Mutation Calls**

Resulting mutation calls required 5 variant reads present, with at least 20 read depth at each site of interest in all regions. The mutation calls were annotated using ANNOVAR<sup>215</sup>. All mutations of interest were examined visually in IGV to check for mapping errors in the mutation sites.

### **2.6.4 Exome Regional Amplicon Validation**

After defining the variants, we used targeted deep-amplicon sequencing to validate the regional specific mutations. We focused mostly on the invasive specific mutations, because invasion specific mutations might provide information on why a clone or clones could escape and survive in outside the ducts.

#### **2.6.4.1 Exome Regional Design Primers**

Primer design used Primer 3<sup>216</sup>, with five base pairs upstream and downstream from the SNP location used as a target. The amplicon size range was limited to 125-250bp.

#### 2.6.4.2 Exome Regional Testing Primers

Primers pair testing for the best primer set used four combinations of primers ( $F_1+R_1$ ,  $F_1+R_2$ ,  $F_2+R_1$ ,  $F_2+R_2$ ), the best temperature for the different pairs (a gradient from 62C to 72C), and the best yield with the start DNA of 1ug and a DNA smear within the 150-300bps. We confirmed the minimum start DNA (a dilution series 0.25ug-10ug of DNA) with the best primer set. For all tests, the DNA amplification used the different primer pairs with TaqMan for 35 cycles. The products were run out on a gel and clearest band of around 200bps was chosen.

#### 2.6.4.3 Exome Regional Amplicon NSG Prep

The amplicons from different regions (*in situ* and invasive) were pooled in equimolar amounts and sequencing libraries were constructed using NEBNext® DNA library Prep enzymes (NEB, #E6050L, E6053L, E6056L/M0202L, and M0541 for end-repair, 3' adenylation, ligation and PCR amplification). Following ligation, DNA underwent a negative and positive selection with Ampure XP beads (Beckman Coulter, #A63881), 0.7x and 0.15x respectively, prior to PCR amplification. Final library concentrations were measured using the Qubit 2.0 Fluorometer. Samples were diluted to 10nM and sequenced on the MiSeq system (Illumina, 150 paired-end) to obtain a target coverage depth of >100,000X.

#### 2.6.4.4 Exome Regional Deep SNV

Statistical significance of observed variants was calculated using deepSNV version 1.16.0, which detects variants assuming a beta-binomial model<sup>217</sup>. To estimate the over dispersion parameter of the model, data from the targeted sites plus flanking regions of 20bp on either side were used. DeepSNV was used to calculate p-values for the null hypothesis (that the targeted variant was equally frequent in primary tumor and paired normal) using separate one-tailed likelihood ratio tests for each strand orientation, and combining the p-values using Fisher's method, comparing the variant read against the background caused by amplification errors and other technical noise and then against the reference variant.



### 3 Studying Synchronous DCIS using TSCS

This section is based on the research paper "Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing" published in the Cell in 2018, by Casasent et al<sup>1</sup>. Figures from this paper have been reused or modified under the journal's academic copyright license for student thesis usage. This section is expanded from the paper to go into more details about each tumor samples and presents much of the data which was only covered in the supplements of the paper.

#### 3.1 Introduction

The genomic and evolutionary basis of invasion and progression from DCIS to IDC remains uncertain. Several technical challenges caused by using bulk tissue have made reconstructing genomic tumor progression difficult, including the extensive intratumor heterogeneity (ITH), the limited number of tumor cells in early cancers, and the large number of stromal cells.

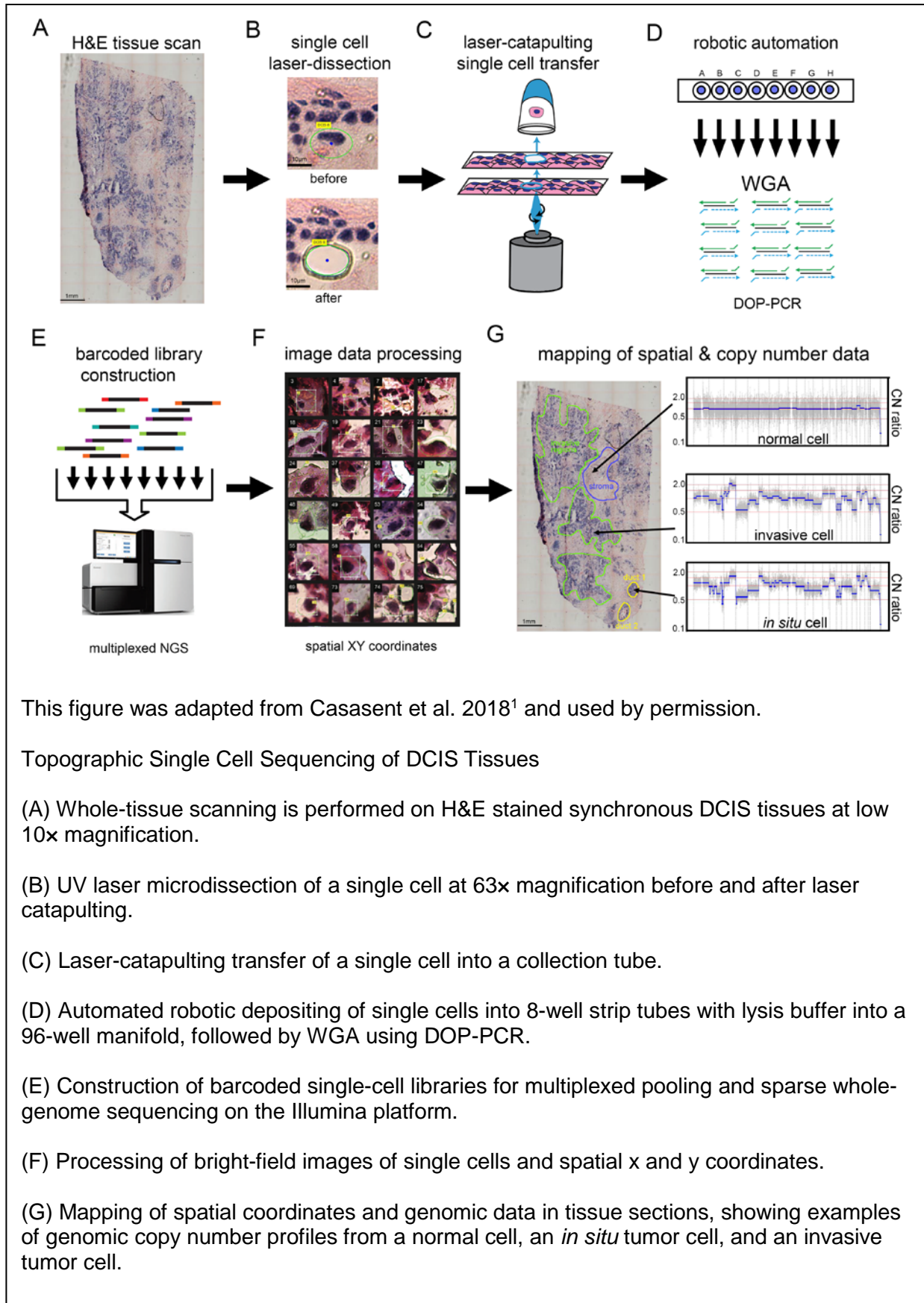
Two major evolutionary hypotheses have been proposed for cancer progression: the independent lineage hypothesis and the direct lineage hypothesis. The independent lineage model postulated that different initiating cells give rise to the *in situ* and invasive subpopulations separately. The direct lineage model postulates a single cell *in situ* gives rise to a cell or cells that invade the surrounding tissue. The classic direct lineage hypothesis of invasion suggests that an evolutionary bottleneck gives rise to the invasive tumor<sup>180</sup>. The evolutionary bottleneck model states that an *in situ* cell or a select few *in situ* cells, invade into the adjacent tissues.

While single-cell DNA sequencing methods have emerged as powerful tools for resolving ITH<sup>4, 135, 136</sup>, delineating stromal cell types<sup>218, 219</sup>, and detecting rare subpopulations<sup>220, 221</sup>, these methods are limited by single-cell isolation methods which require cell suspensions<sup>10</sup>. These procedures inherently lose all spatial information, which is critical for studies of early stage cancers. To address this limitation, we developed Topographical Single Cell Sequencing (TSCS)<sup>1</sup> an approach that combines laser catapulting<sup>222</sup> and single-cell DNA sequencing to

measure genomic copy number profiles of single tumor cells while preserving their spatial information in tissue sections, as illustrated in Figure 9 TSCS Protocol.

We hypothesized that invasive cells share a direct genomic lineage with one (or more) single cells in the ducts. To investigate this question, we applied TSCS, along with deep-exome sequencing, to trace clonal evolution during invasion in 10 high-grade frozen tumor samples from synchronous DCIS-IDC patients. Our results support a direct genomic lineage between the *in situ* and invasive tumor cell subpopulations and further show that most mutations and CNAs evolved within the ducts prior to invasion. These data suggest that multiple clones escaped from the ducts and migrated into the adjacent tissues to establish invasive carcinomas, leading us to postulate the multiclonal model of invasion. The model of multiclonal invasion, postulates that all clones invade the surrounding tissue to form the tumor mass.

Figure 9 TSCS Protocol



### 3.1.1 Rationale of Synchronous DCIS-IDC

Patient samples with synchronous DCIS-IDC provide a golden opportunity to study the genomic and molecular basis of invasion without confounding effects of inter-patient heterogeneity. Synchronous DCIS-IDC samples are, by definition, breast cancer samples with both *in situ* and invasive regions. While using longitudinal pure DCIS samples that later progress to IDC might appear to have advantages over synchronous DCIS-IDC samples for studying clonal evolution during invasion, there are too many issues with confounding effects which would prevent clean analysis of genomic evolution during invasion to make longitudinal recurrent breast cancer samples practical. Longitudinal samples undergo selection from the following confounding effects (1) time, (2) space, and (3) therapy. Additionally, longitudinal samples are (4) less practical to collect.

The first and most fundamental of these advantages is that synchronous DCIS-IDC samples are matched in time, while longitudinal recurrent breast cancer (DCIS to IDC) samples can be separated by a decade or even more, during which many random or passenger mutations may have accumulated.

Second, longitudinal recurrent breast cancer samples are not collected from the same geographical regions. While they might be from the same breast, the distance from the original DCIS and the recurrent IDC tumor can only be measured in large approximations. This spatial distance could result in increased mutations simply because of geographical separation of clones rather than invasion. In synchronous DCIS-IDC, the tumor cells are directly adjacent in geographical space, which minimizes spatial effects.

Third, synchronous DCIS-IDC samples will not be affected by the confounding effects of intervening therapy which can cause selection and result in therapy selection or resistance being confused with invasion. Cancer therapy including radiation, hormonal or chemotherapy, have been noted to cause mutations or selection in tumors.

Fourth, longitudinal recurrent breast cancer samples are logistically very difficult to collect. At the University of Texas MD Anderson Cancer Center the recurrence rate of DCIS

cancer is only 6% and most patients are seen at different hospitals when they have recurrent disease. In addition, most of the pure DCIS is usually used by pathology, and the samples are usually fixed and not collected as fresh-frozen material, which is necessary for SCS studies. Since patient treatment is the first and most important concern with samples, pathologists at the University of Texas MD Anderson Cancer Center tend to section through the entirety of a sample considered DCIS-only to make sure that the tumor samples have no invasive regions present. This leaves no residual samples for research purposes<sup>197</sup>.

Lastly, many previous studies have used synchronous IDC-DCIS to study invasion, and this has been widely accepted in the field to study invasion and overcome limitations associated with the analysis of longitudinal samples<sup>223, 139, 114, 86, 119, 6, 224</sup>. These justifications provide strong rationale for the biological and technical advantages of using synchronous DCIS-IDC samples over longitudinal samples to study clonal evolution during invasion.

### **3.1.2 Rationale of need for TSCS**

Even in the first single cell studies regional ITH was observed on a macrolevel (1mm cube sectors). Current single-cell DNA sequencing methods require cell suspensions making microlevel investigation impossible because the inherent loss of microlevel spatial information. However, by pairing SCS with LCM and laser catapulting, single cells can be isolated passed on morphology and location. Topographical Single Cell Sequencing (TSCS) was developed to preserve location while measuring genomic copy number profiles of single tumor cells allows us to examine the genomes of *in situ* and invasive tumor cells.

## 3.2 Results

In this section I will discuss the single cell copy number results for each tumor and the regional exome results. We examined synchronous DCIS-IDC samples, the receptor status and other clinical information was presented earlier in Table 2 Clinical Information. Here, we split the data into two groups, polyclonal tumors and monoclonal tumors, before examining these results in more detail. The number of clones in each sample was determined earlier using k-means clustering, with the k being selected by the first standard error max as describe in the methods. Next, we will show the results for the regional exome data.

This data was generated by TSCS. On all these data we provided the following: single cell heatmap with the subclones and regions marked, followed by the consensus heatmap and line plots, the saturation curve for the number of single cells per region, 2-dimensional cluster of the data using multidimensional scaling (MDS) and Image Maps to show the histology ducts and locations of each clones. For polyclonal tumors I show two extra plots, the change a frequency plots by TimeScape, and the spatial and clonal relationships by tanglegram.

### 3.2.1 Copy Number Evolution During Invasion Polyclonal Tumors

In this section I discuss each polyclonal tumor in detail.

#### 3.2.1.1 DC4

We investigated copy number evolution during invasion in patient DC4, a TNBC grade 3 sample. We collected the fewest number of cells from this tumor, because of lack of tumor tissue. This was also the first tumor we examined using TSCS. After filtering and analysis, we examined 57 total cells, 19 from *in situ* and 38 invasive from 2 tumor sectors (R1, R2). While this number appears very small, the saturation analysis showed only 50 cells were sufficient to detect all the subclones (see Figure 11 DC4 Saturation Curve). In DC4, we did observe some regional effects, since DC4 had only 3 aneuploid cells, all of which were clone B, while the other sector R2 had clones from both A, B, and normal cells (see Figure 14 DC4 Image Maps and Figure 15 DC4 Tanglegram).

Consensus copy number profiles (Figure 10 DC4 Copy Number Alteration Heatmap) showed that both clones shared a common amplification of chromosome 1q and 5p, as well as common deletions of 13 and 18q, suggesting a common ancestor between clones A and B. In clone A, we identified many unique focal amplifications in chromosome 8 (MTDH, MYC, and PTPRD) and 17p (ERBB2 or HER2) and larger deletions of 3p, 6q, 8p, and 17. However, clone A does have some variations. Within the clones there are changes in the focal deletion on chromosome 4 and 6q, and focal amplifications of 11q and 12q.

Also, while amplification of 5p was a common alteration between clones A and B, it was not observed in all cells in clone A. In clone B, we identified many unique focal amplifications, such as chromosome 1 near the centromere (MCL1, SHC1), 8q, 9p, and 18p, as well as larger amplifications of 10p (GATA3), 12p (CDKN1B, KRAS), and 19q, in addition to the many deletions of chromosome 4, 5q, and 20. However clone B also had some variations, suggesting that we might be under clustering or under sampling. The variations in clone B were the deletions of chromosome 4 and 6q, and focal deletion 3q, and focal amplification on 12q.

Based on our current clustering, this data showed that genomic copy number evolution occurred within the ducts and gave rise to 2 major tumor subpopulations. During invasion, the frequency of clone A increased from 40% to 55%, while the frequency of clone B decreased from 60% to 45% in the invasive tissues (See Figure 13 DC4 TimeScape).

MDS (Figure 12 DC4 MDS) identified 3 distinct clusters that corresponded to the normal cells (N) and the 2 tumor clones (A, B). The MDS plot showed that each clonal genotype was composed of both *in situ* and invasive tumor cells, with no specific genotype associated with either region. Two or three of the B clones were closer to the normal population, suggesting a potential misclassification. Next, we mapped the clonal genotypes to their spatial coordinates on an image map for DC4 sectors (R1, R2), which showed that both clones were located within the ductal and invasive regions in R2. If we focus on sector R2, we can see that within the ducts clones A and B are very close together and even intermixed. While we see some

intermixture between clones A and B in the invasive regions, clone B is found further away from ducts (see Figure 14 DC4 Image Maps and Figure 15 DC4 Tanglegram).

One point of interest about DC4 is that the sample is marked as TNBC. However, clone A has ERBB2 or HER2 amplification, which is not present in clone B, demonstrating heterogeneity in receptor status (Table 2 Clinical Information).



Figure 10 DC4 Copy Number Alteration Heatmap

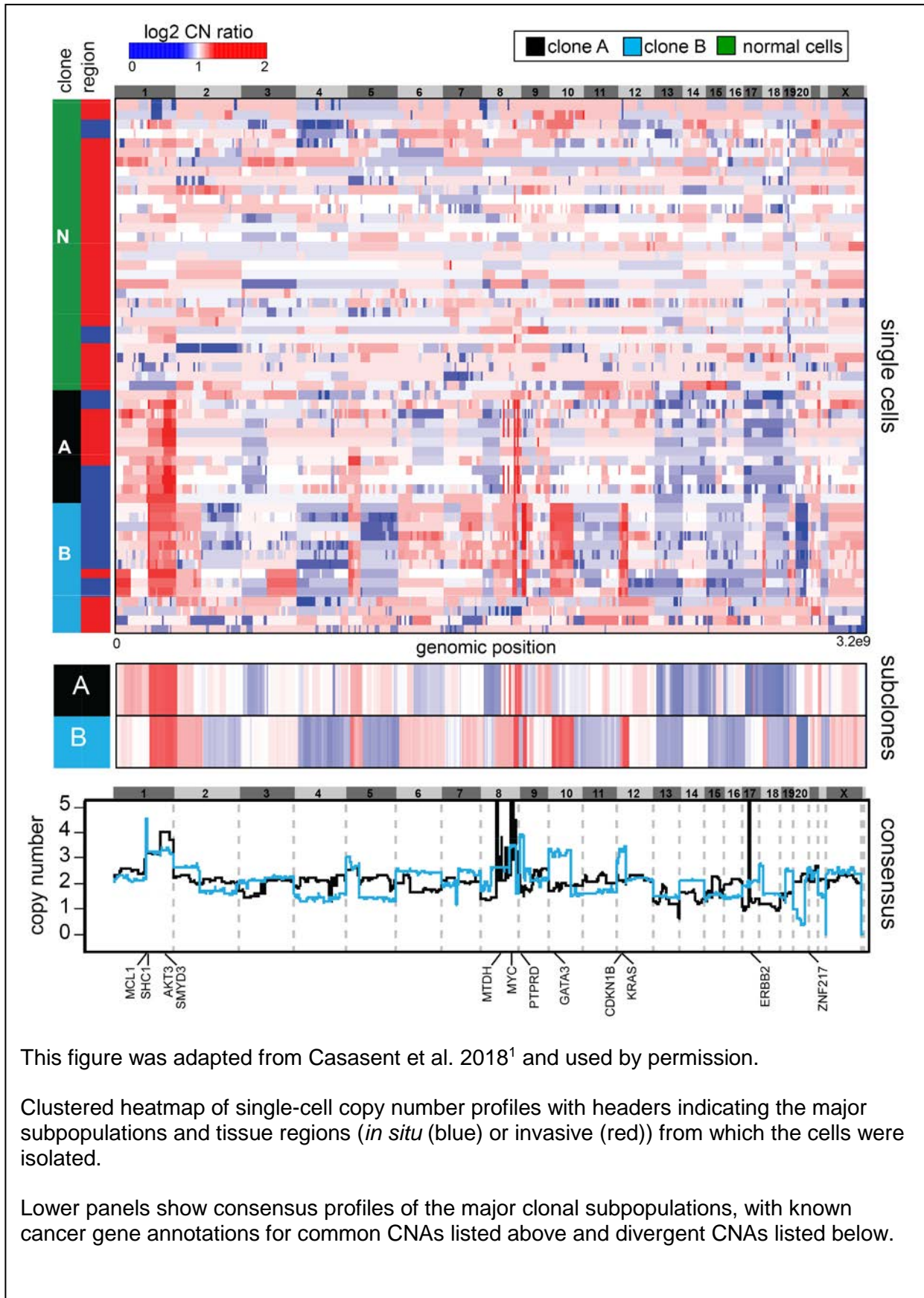
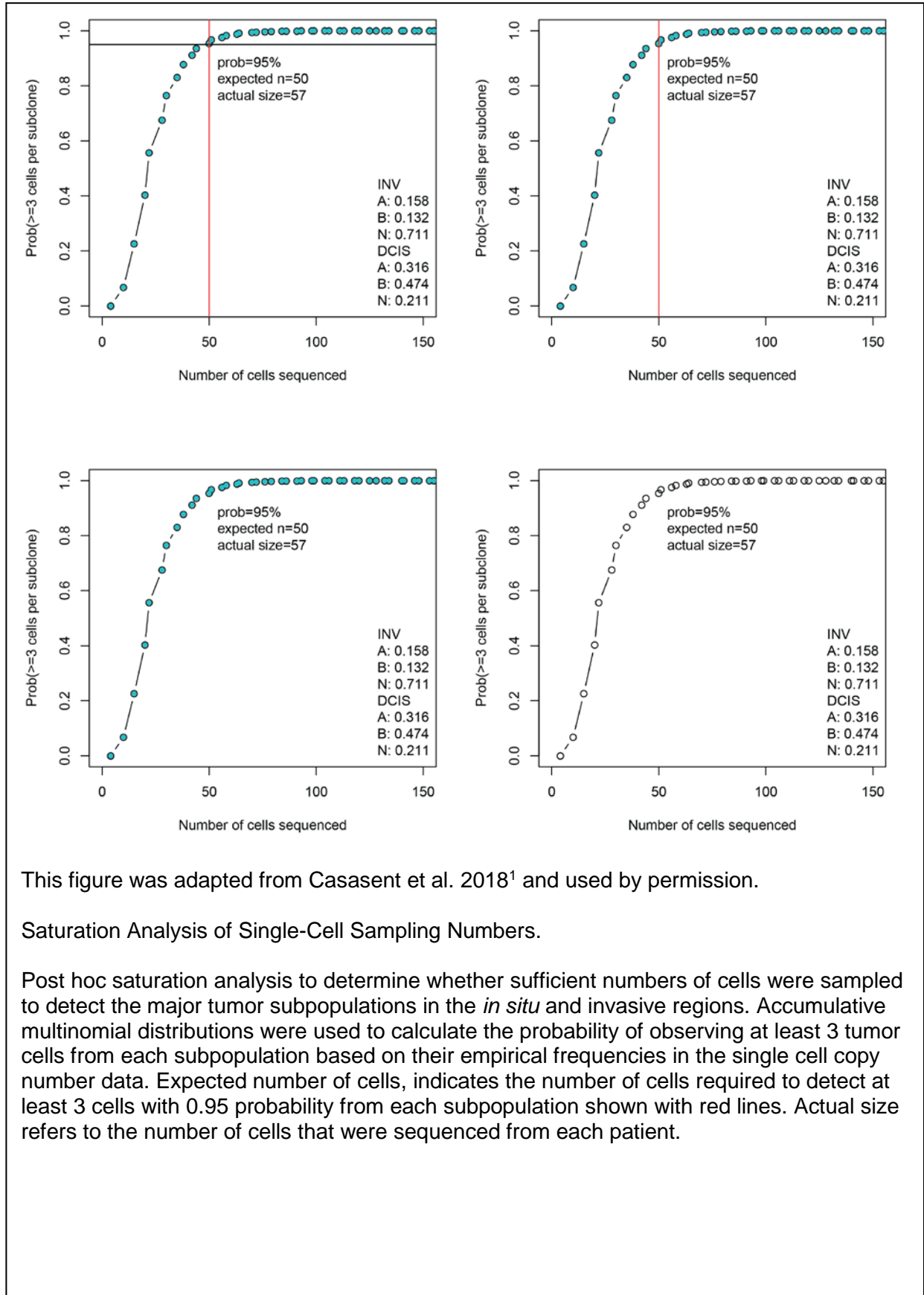


Figure 11 DC4 Saturation Curve



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Saturation Analysis of Single-Cell Sampling Numbers.

Post hoc saturation analysis to determine whether sufficient numbers of cells were sampled to detect the major tumor subpopulations in the *in situ* and invasive regions. Accumulative multinomial distributions were used to calculate the probability of observing at least 3 tumor cells from each subpopulation based on their empirical frequencies in the single cell copy number data. Expected number of cells, indicates the number of cells required to detect at least 3 cells with 0.95 probability from each subpopulation shown with red lines. Actual size refers to the number of cells that were sequenced from each patient.

Figure 12 DC4 MDS

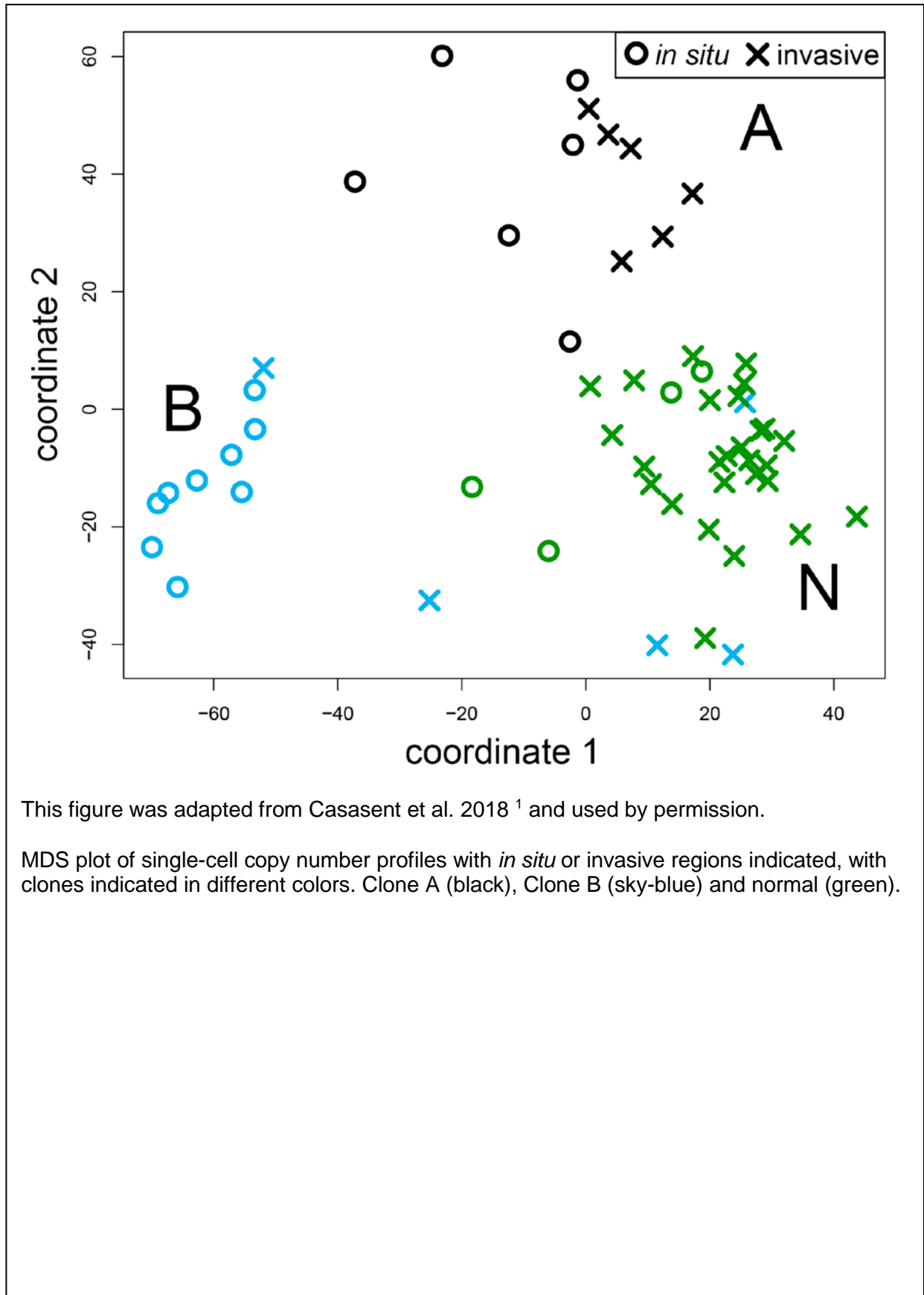


Figure 13 DC4 TimeScape

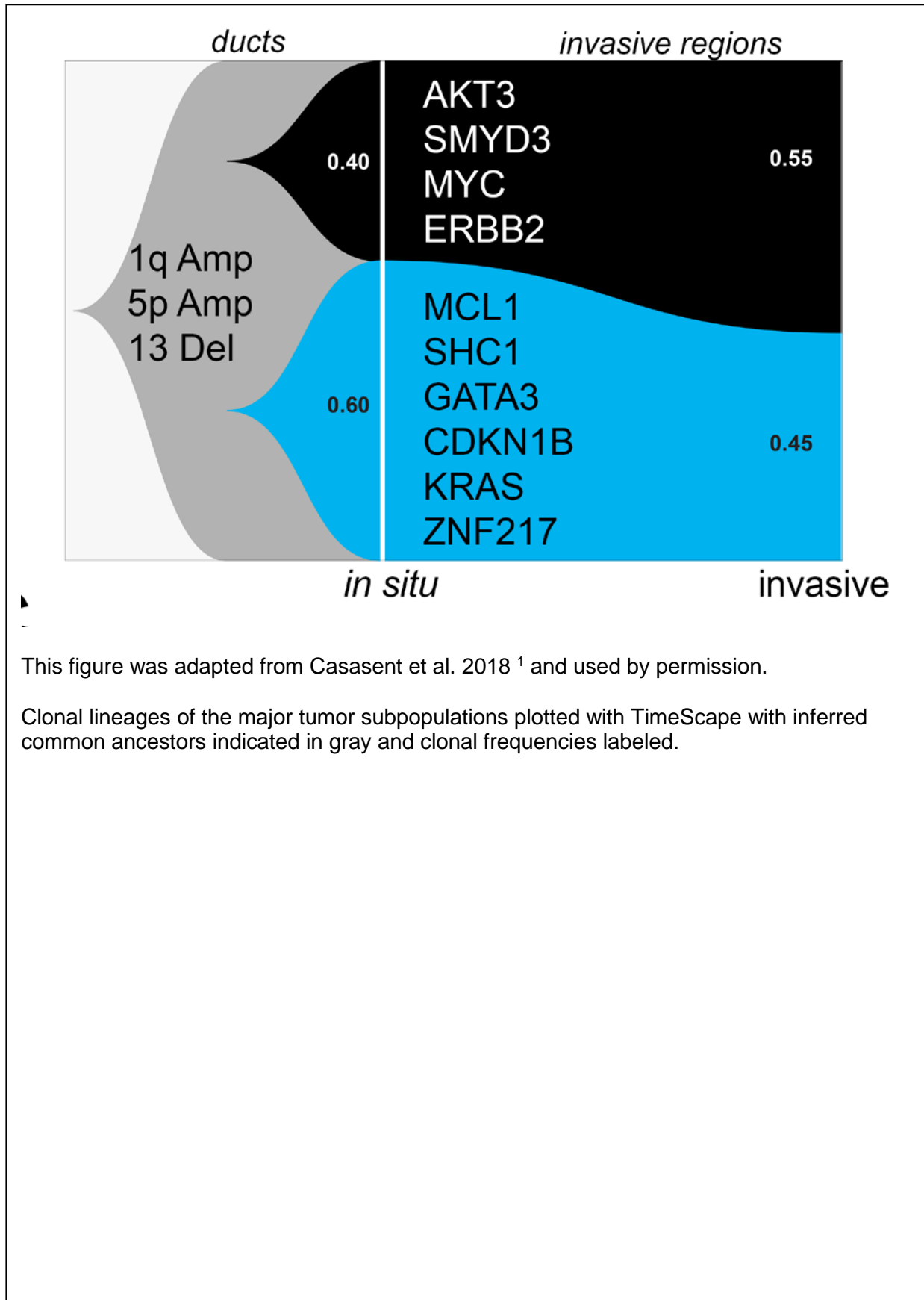


Figure 14 DC4 Image Maps

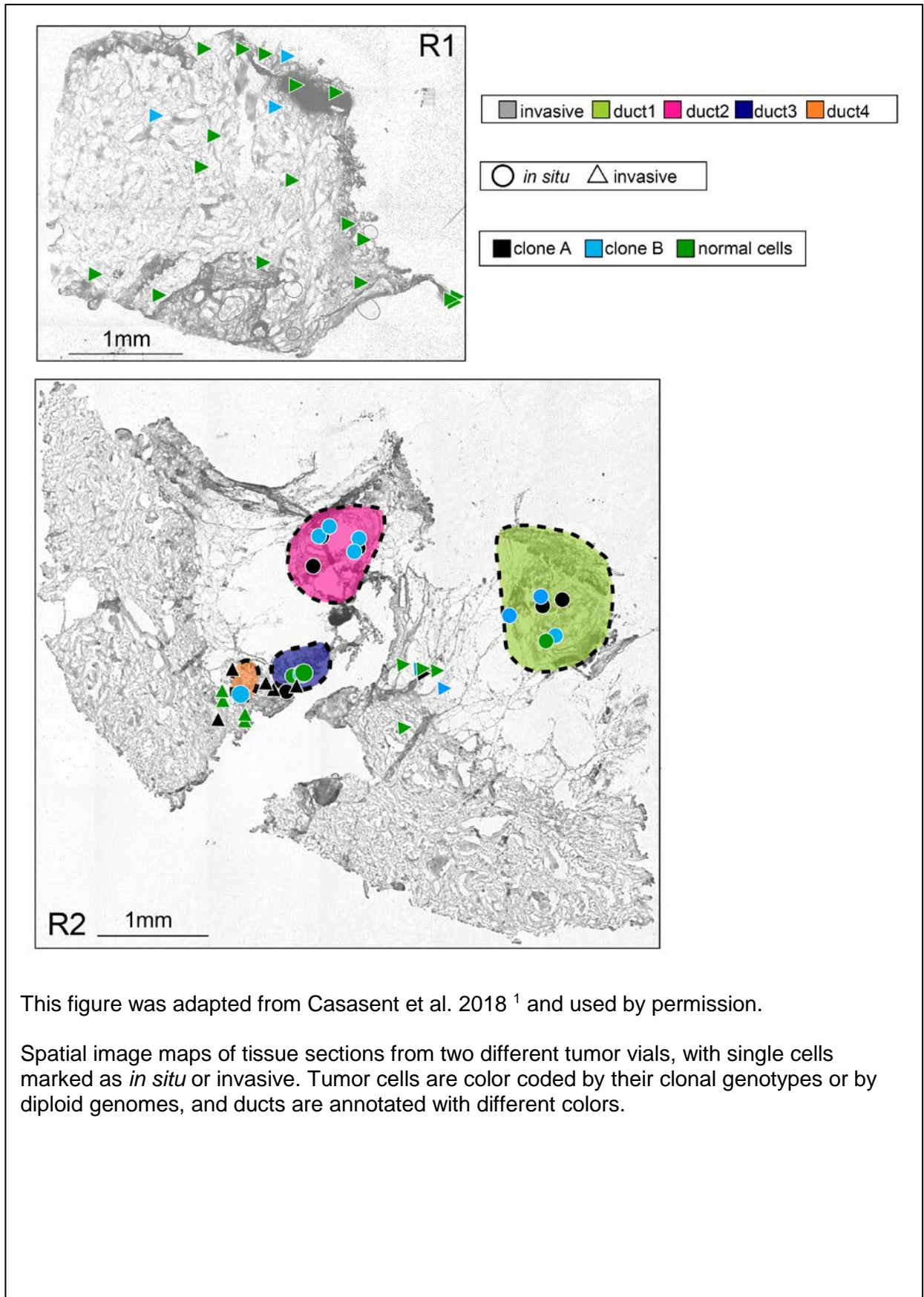
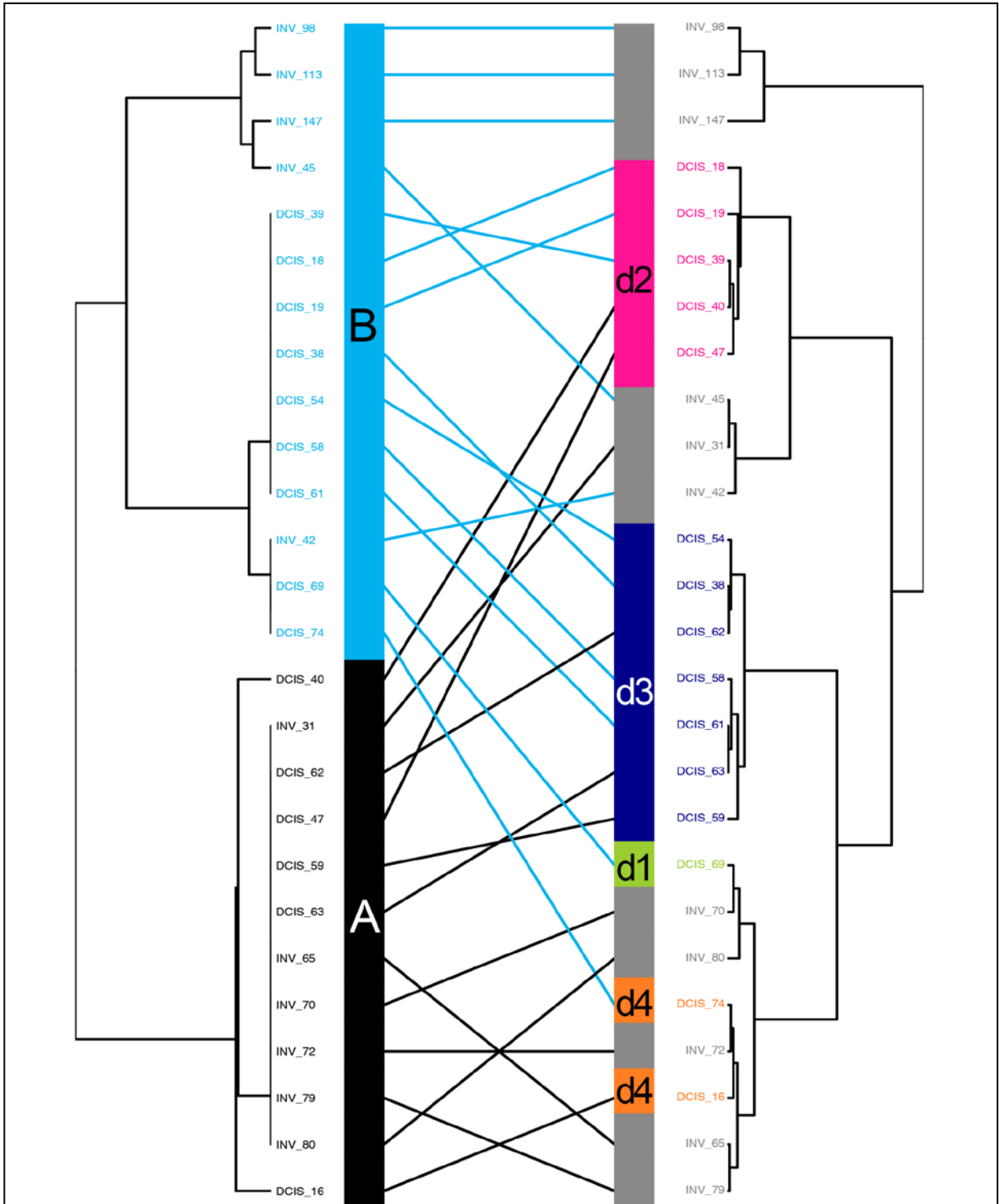


Figure 15 DC4 Tanglegram



This figure was adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

Genotype trees are located on the left side for each patient, with clonal subpopulations indicated by color. Spatial trees are located on the right side with different ducts indicated by colors and the invasive regions colored in gray. Mapping of cells coordinates and genotypes were performed by minimizing overlapping connections.



### 3.2.1.2 DC13

We investigated copy number evolution during invasion in patient DC13, a grade 2 ER/PR positive, HER2 negative sample, using TSCS to sequence 46 *in situ* cells and 58 invasive cells from two tumor regions (R1 and R2) (See Figure 17 DC13 Saturation Curve). Hierarchical clustering of single cell copy number profiles identified one subpopulation of diploid cells (N) and two aneuploid tumor subpopulations (A, B) (Figure 16 DC13 Copy Number Alteration Heatmap). Within each subpopulation, the single cell copy number profiles showed high correlations (A= 0.89, B=0.60, Pearson correlations) representing stable clonal expansions. Consensus copy number profiles showed that both clones shared a common amplification of chromosome 1p (*MDM4*, *ABL2*), in addition to many subpopulation-specific CNAs. In clone A we identified many focal amplifications, including chromosome 3q (*EVI1*), 4p (*CPEB2*), 11q (*CASP12*), and 13q (*PCDH17*), as well as an amplification of chromosome 12q (*CDK2*, *MDM2*). In contrast, clone B harbored many large hemizygous chromosomal deletions including 3p (*SETD2*, *FHIT*), 4 (*FGFR3*, *NEK1*), 5q (*PIK3R1*, *APC*), 14q (*AKT1*), 15q (*NTRK3*), 16q (*CDH1*), 17p (*TP53*, *MAP2K4*), 18 (*SMAD4*), and 22 (*NF2*).

Clonal lineages, inferred from the major subpopulations, identified a common ancestor with an amplification of chromosome 1q that gave rise to the two tumor subpopulations in the ducts: one that had many focal amplifications of cancer genes including *MDM2* and *CDK2* (clone A), and another that had many large hemizygous deletions, including *CDH1*, *TP53*, *FHIT*, and *SMAD4* (clone B). This showed that genomic copy number evolution occurred within the ducts and gave rise to two major tumor subpopulations. During invasion, the frequency of clone B increased from 16% to 67%, while the frequency of clone A decreased from 84% to 33% in the invasive tissues (Figure 19 DC13 TimeScape).

MDS identified three distinct clusters that corresponded to the normal cells (N) and the two tumor clones (A, B). The MDS plot (Figure 18 DC13 MDS) showed that each clonal genotype was composed of both *in situ* and invasive tumor cells, with no specific genotype associated with either region. Next, we mapped the clonal genotypes to their spatial

coordinates in the two tissue sections (R1, R2), which showed that both clones were located in the ductal and invasive regions. This map also showed that in region 1 most of the normal diploid cells were localized to the invasive regions, which may reflect the difficulty in distinguishing stromal from tumor cells in these regions by histopathology (Figure 20 DC13 Image Maps). Furthermore, these data showed that clone A was highly localized to the four ducts (d1 – d4) in region 2, while clone B was more prevalent in the invasive regions (Figure 21 DC13 Tanglegram). Consistent with the invasive spatial localization, we found that clone B had deletions in several cancer genes involved in cell migration, including *AKT1*, *APC*, *FGFR3*, *CDH1*, and *SMAD4*.



Figure 16 DC13 Copy Number Alteration Heatmap

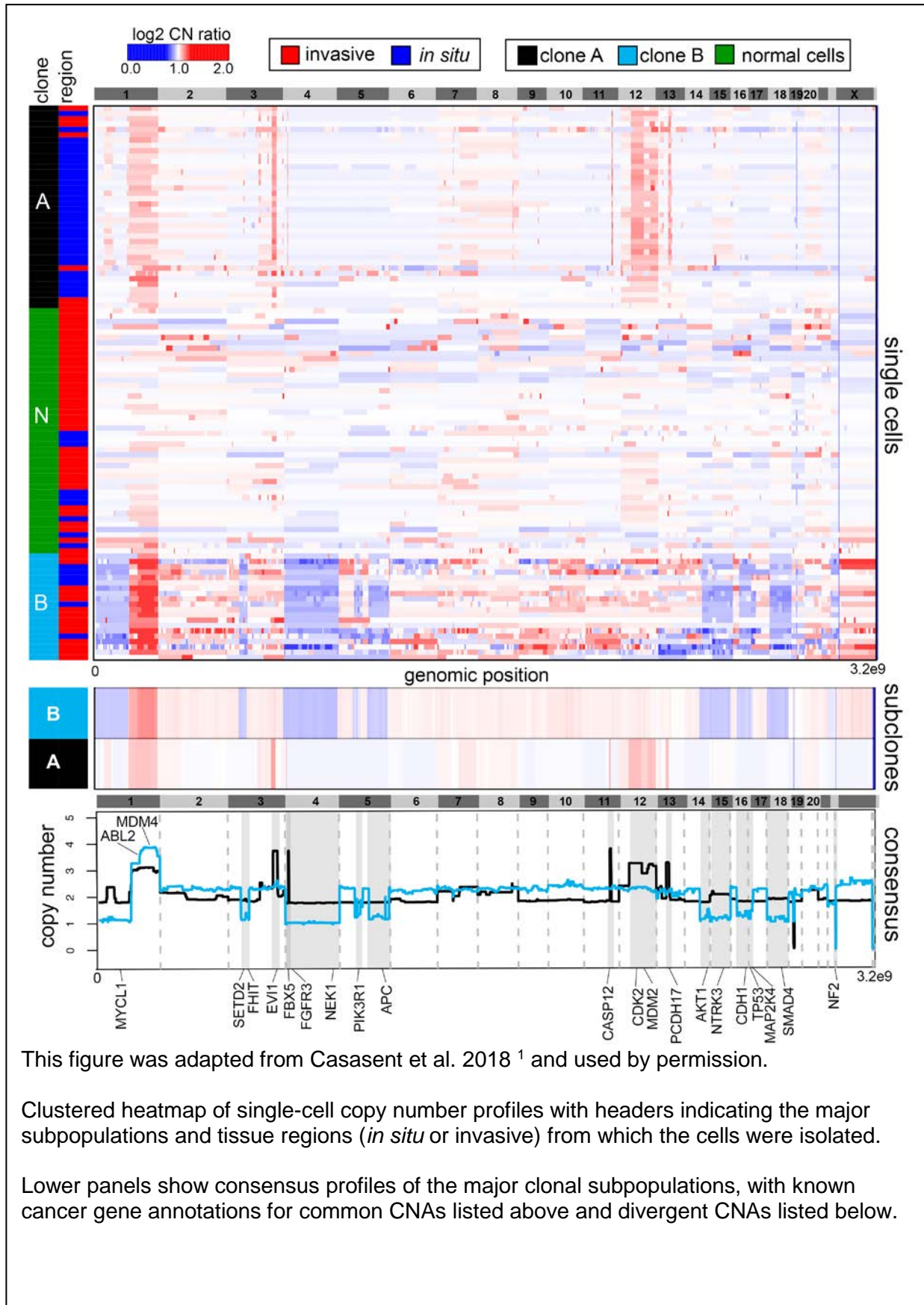
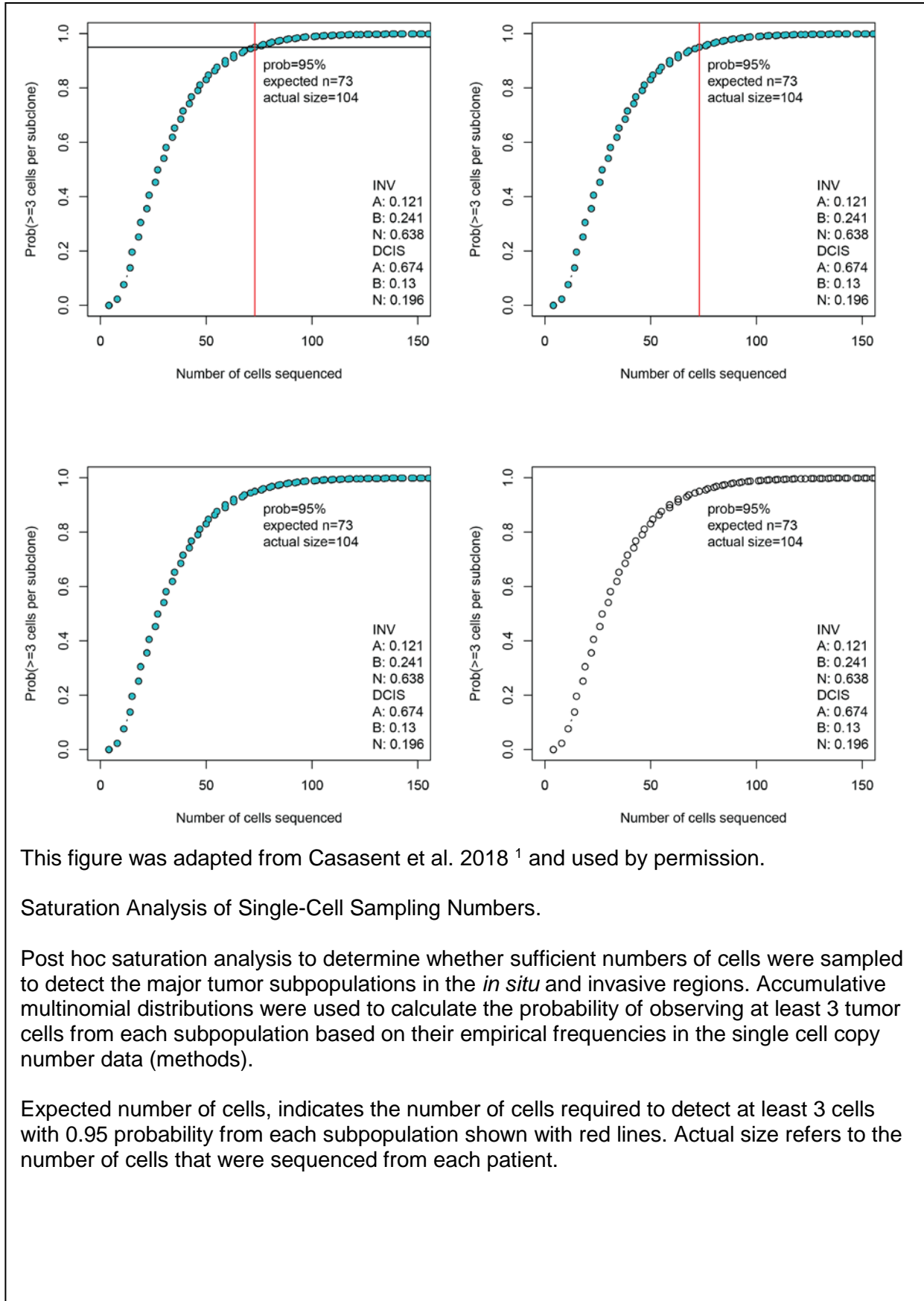


Figure 17 DC13 Saturation Curve



This figure was adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

#### Saturation Analysis of Single-Cell Sampling Numbers.

Post hoc saturation analysis to determine whether sufficient numbers of cells were sampled to detect the major tumor subpopulations in the *in situ* and invasive regions. Accumulative multinomial distributions were used to calculate the probability of observing at least 3 tumor cells from each subpopulation based on their empirical frequencies in the single cell copy number data (methods).

Expected number of cells, indicates the number of cells required to detect at least 3 cells with 0.95 probability from each subpopulation shown with red lines. Actual size refers to the number of cells that were sequenced from each patient.

Figure 18 DC13 MDS

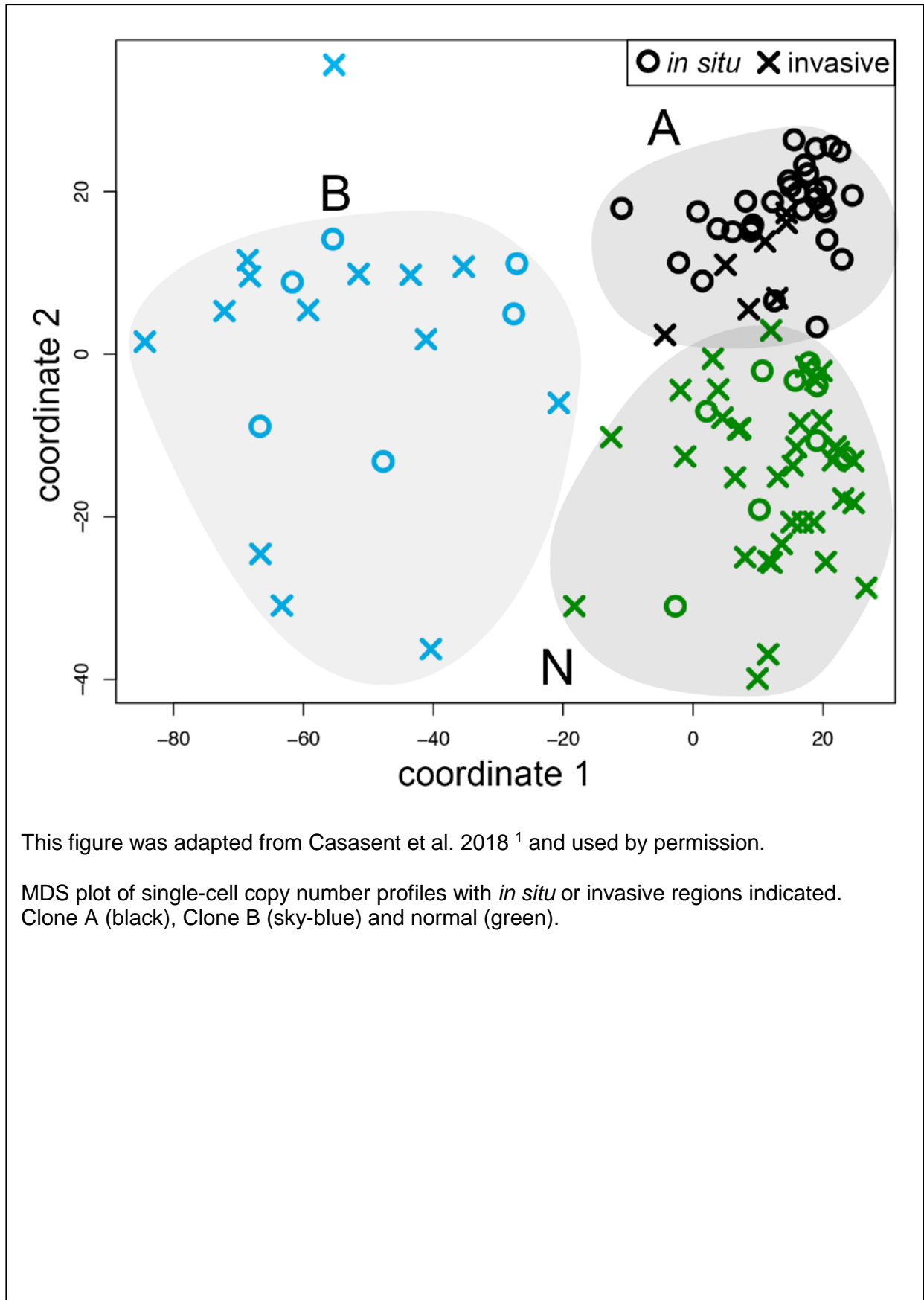
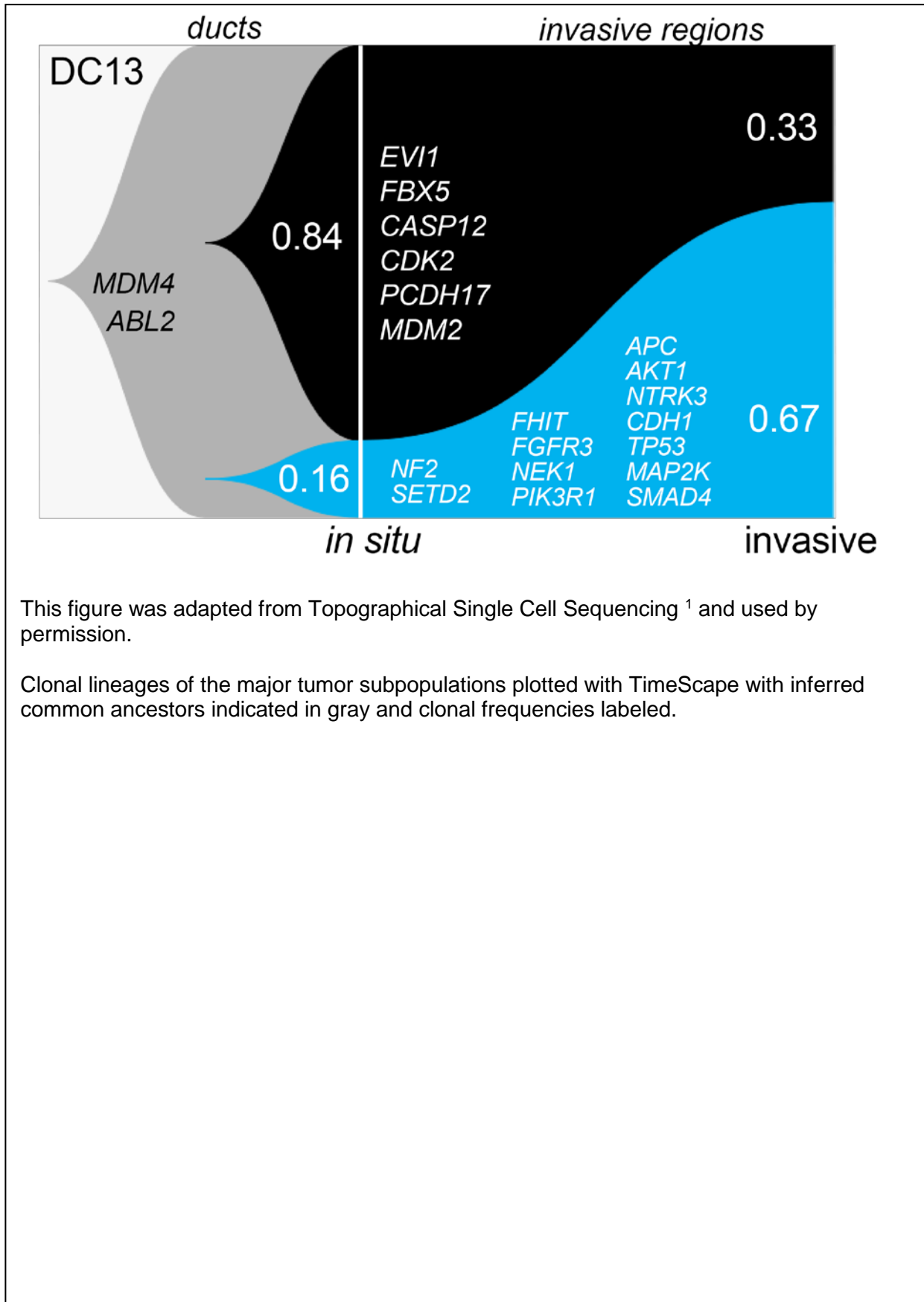


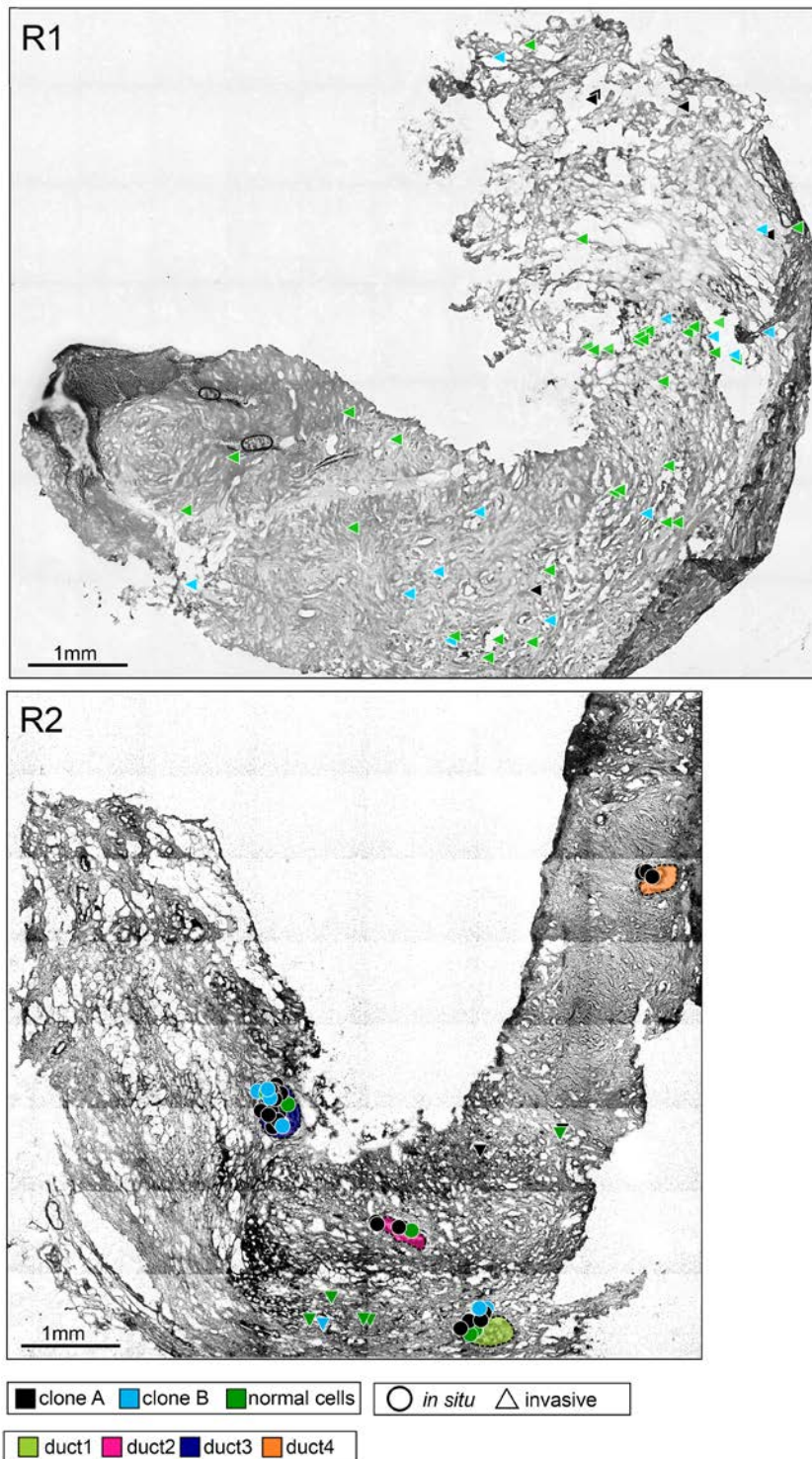
Figure 19 DC13 TimeScape



This figure was adapted from Topographical Single Cell Sequencing <sup>1</sup> and used by permission.

Clonal lineages of the major tumor subpopulations plotted with TimeScape with inferred common ancestors indicated in gray and clonal frequencies labeled.

Figure 20 DC13 Image Maps

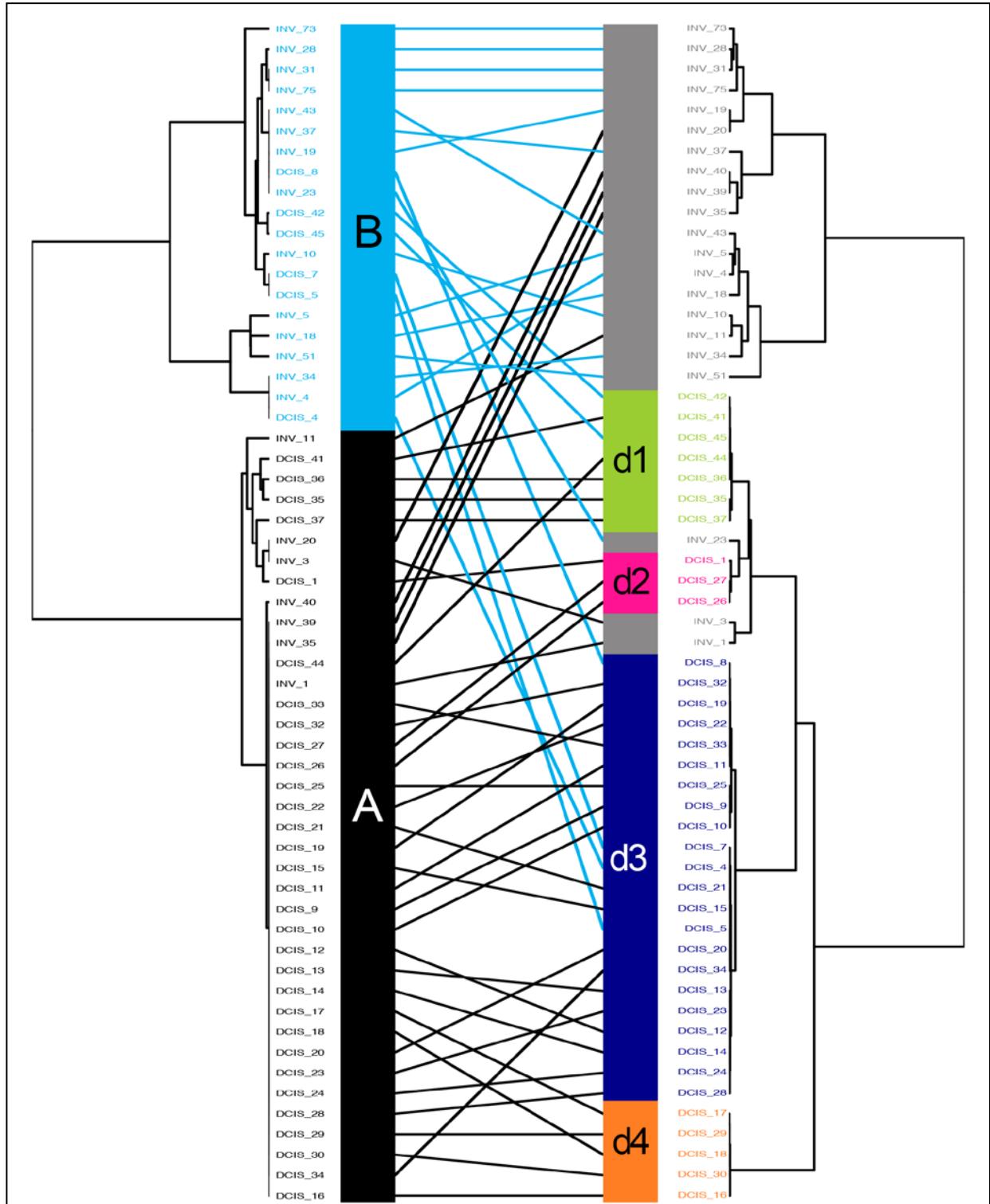


This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Spatial maps of tissue sections from two different tumor vials, with single cells marked as *in situ* or invasive. Tumor cells are color coded by their clonal genotypes or by diploid genomes, and ducts are annotated with different colors.



Figure 21 DC13 Tanglegram



This figure was adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

Genotype trees are located on the left side for each patient, with clonal subpopulations indicated by color. Spatial trees are located on the right side with different ducts indicated by colors and the invasive regions colored in gray. Mapping of cells coordinates and genotypes were performed by minimizing overlapping connections.

### 3.2.1.3 DC14

While DC14, grade 3 with ER positive, and PR and HER2 negative, was found to be a polyclonal tumor, the two clonal populations found (A, B) were highly correlated, suggesting they were from the same single clone of origin. In patient DC14, we examined, after filtering, 148 total cells, 70 from *in situ* and 78 invasive from 3 tumor sectors (R1, R2, R3) (See Figure 23 DC14 Saturation Curve). In clone A we identified a few unique amplifications or deletions. Clone A was very stable and suggested a very strong clonal expansion. Clone B while highly related to clone A, but had more variations within the single cell profiles. Specifically, in clone B we identified only identified two alterations that distinguished it from clone A: (1) less complete deletion of chromes 21 and a stronger deletion of chromosome 13 (Figure 22 DC14 Copy Number Alteration Heatmap).

Our clustering showed that genomic copy number evolution occurred within the ducts and gave rise to 2 major tumor subpopulations. During invasion, the frequency of clone A and B stayed very stable, with clone A increasing slightly from 76% to 80%, and clone B decreasing slightly from 24% to 20% in the invasive tissues. These changes are very small, suggesting both clones can survive in *in situ* and invasive regions equally (Figure 25 DC14 TimeScape).

MDS (Figure 24 DC14 MDS) identified 3 distinct clusters that corresponded to the normal cells (N) and the 2 tumor clones (A, B). The MDS plot showed that each clonal genotype was composed of both *in situ* and invasive tumor cells, with no specific genotype associated with either region. However, clone A closely localized in the MDS plot, suggesting high clonality, while the normal and clone B cells had more spread, suggesting more diversity or noise within these profiles.

Next, we mapped the clonal genotypes to their spatial coordinates on an image map (Figure 26 DC14 Image Maps) for DC14 sectors (R1, R2, R3), with R1 and R2 showing *in situ* populations and R3 showing invasive populations. We see some intermixture of clones A and B in R2. However, R1 and R3 appear to be mostly made up of clone A (Figure 27 DC14 Tanglegram).

Figure 22 DC14 Copy Number Alteration Heatmap

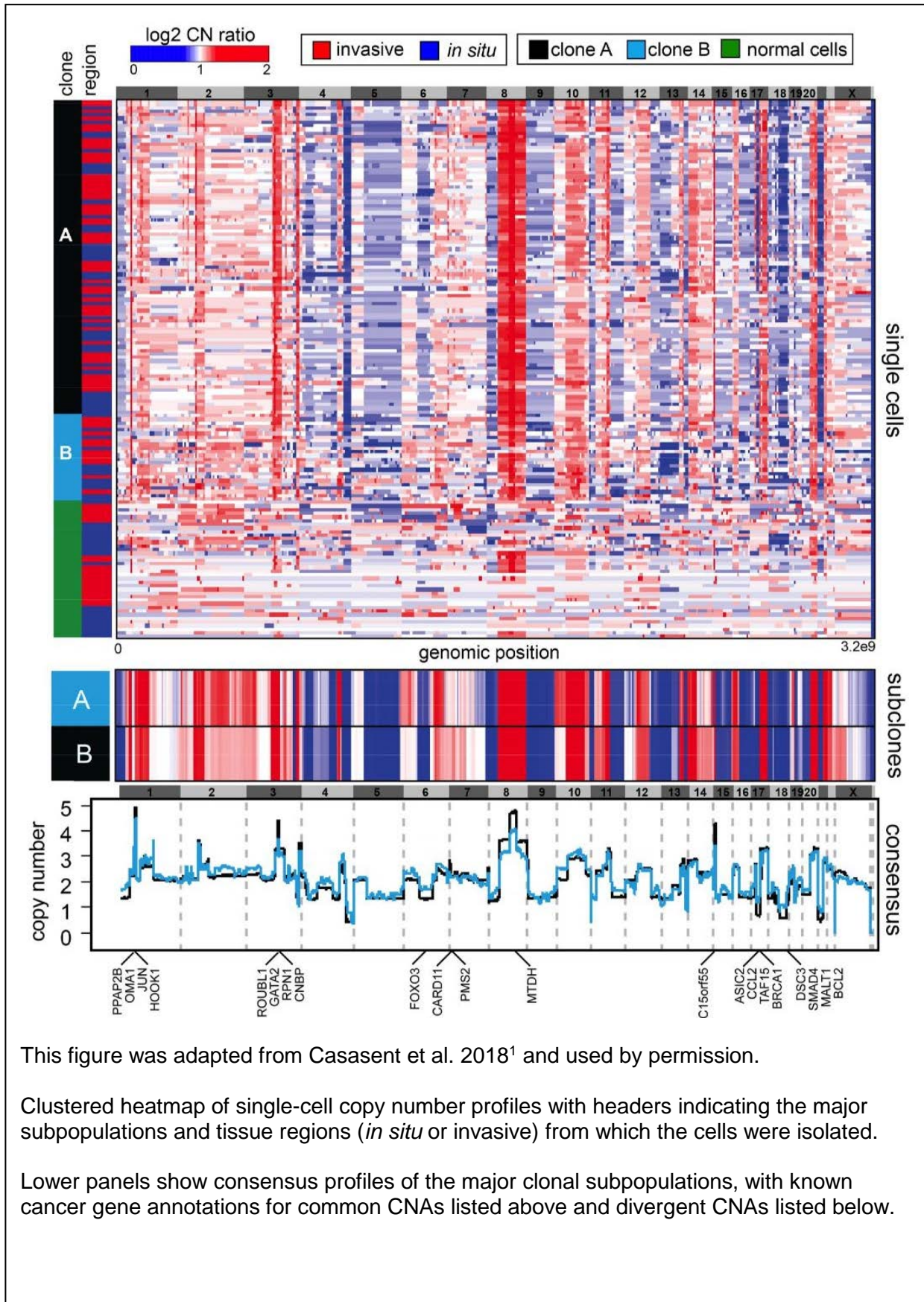




Figure 23 DC14 Saturation Curve

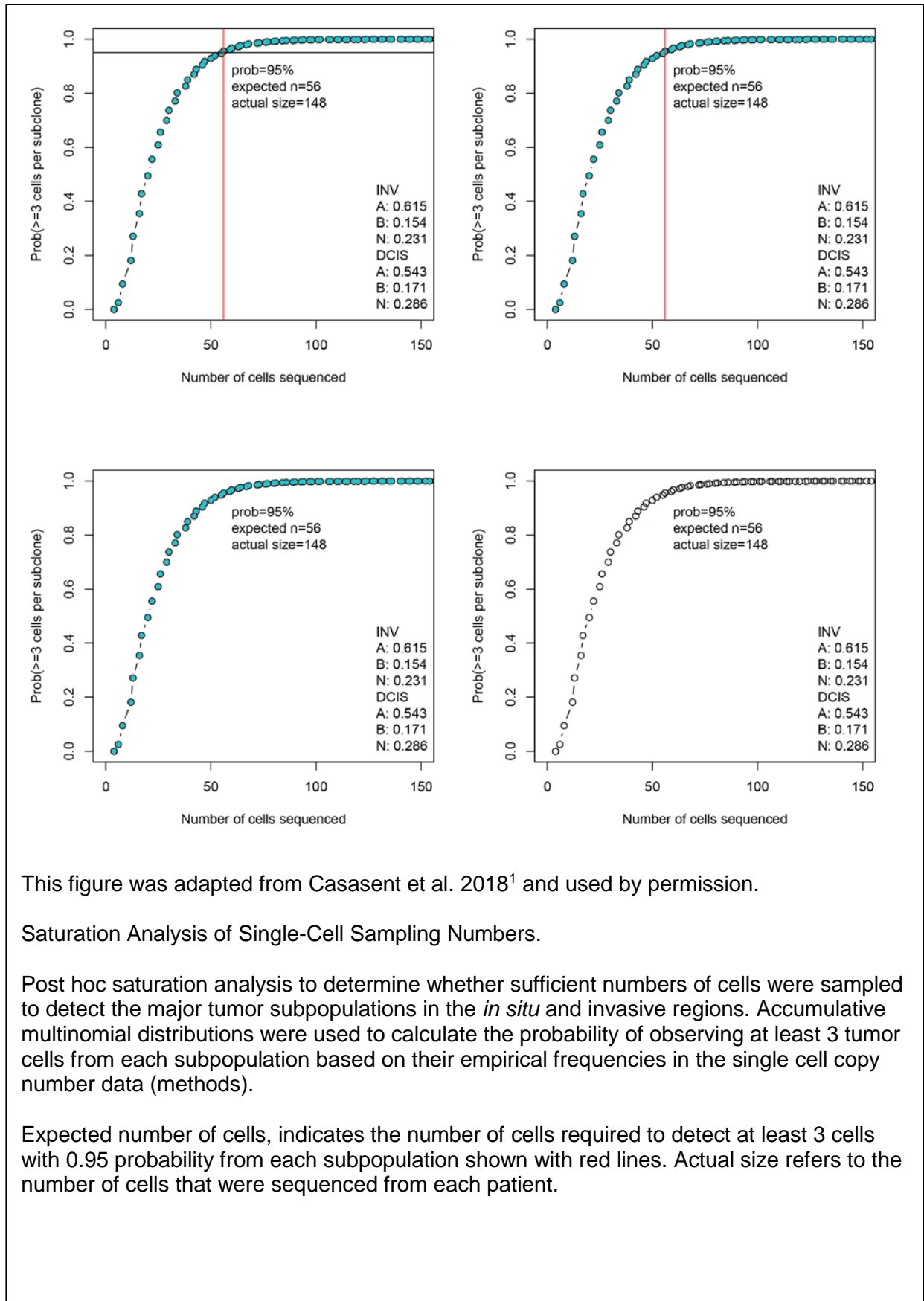
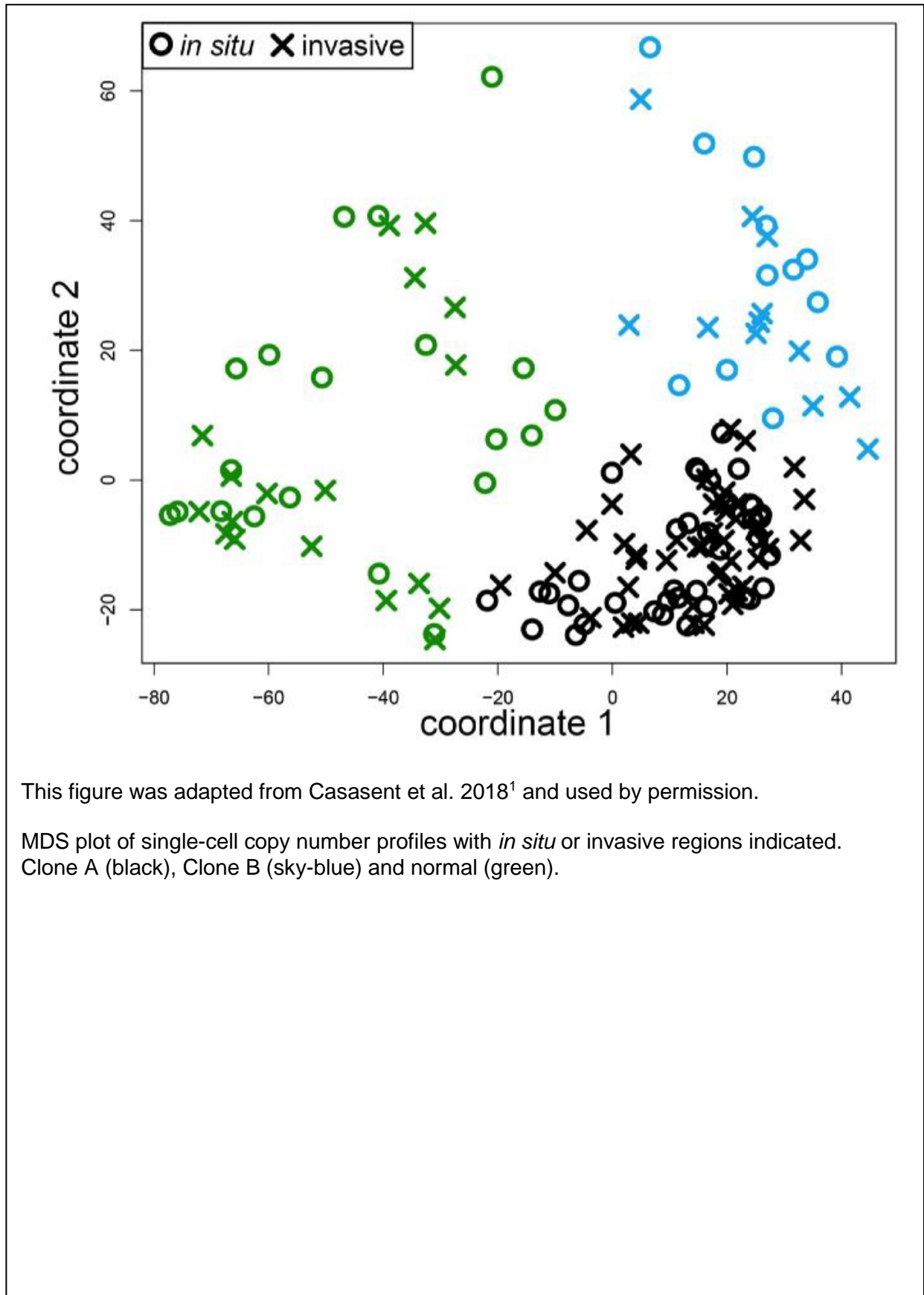


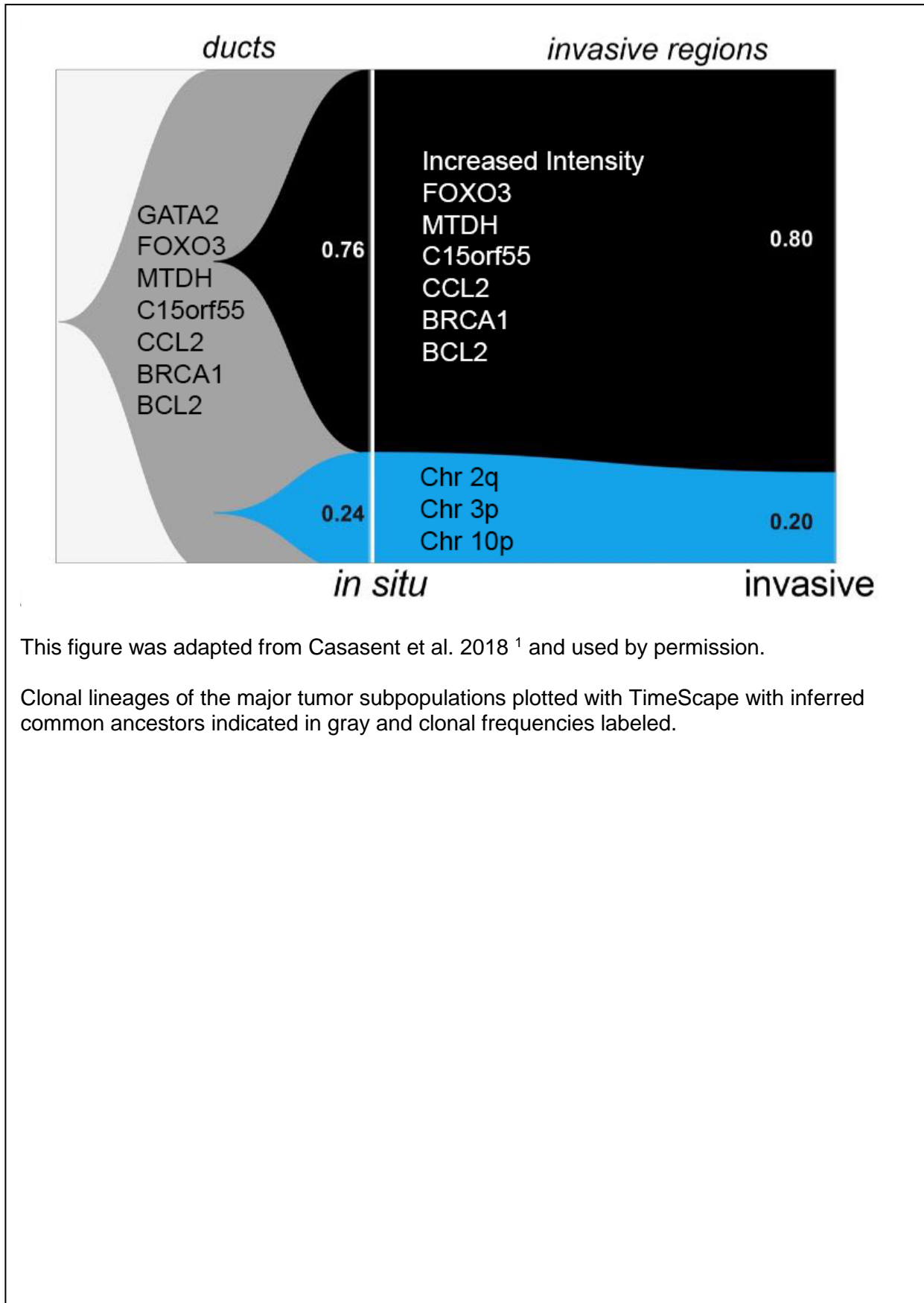
Figure 24 DC14 MDS



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

MDS plot of single-cell copy number profiles with *in situ* or invasive regions indicated. Clone A (black), Clone B (sky-blue) and normal (green).

Figure 25 DC14 TimeScape



This figure was adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

Clonal lineages of the major tumor subpopulations plotted with TimeScape with inferred common ancestors indicated in gray and clonal frequencies labeled.

Figure 26 DC14 Image Maps

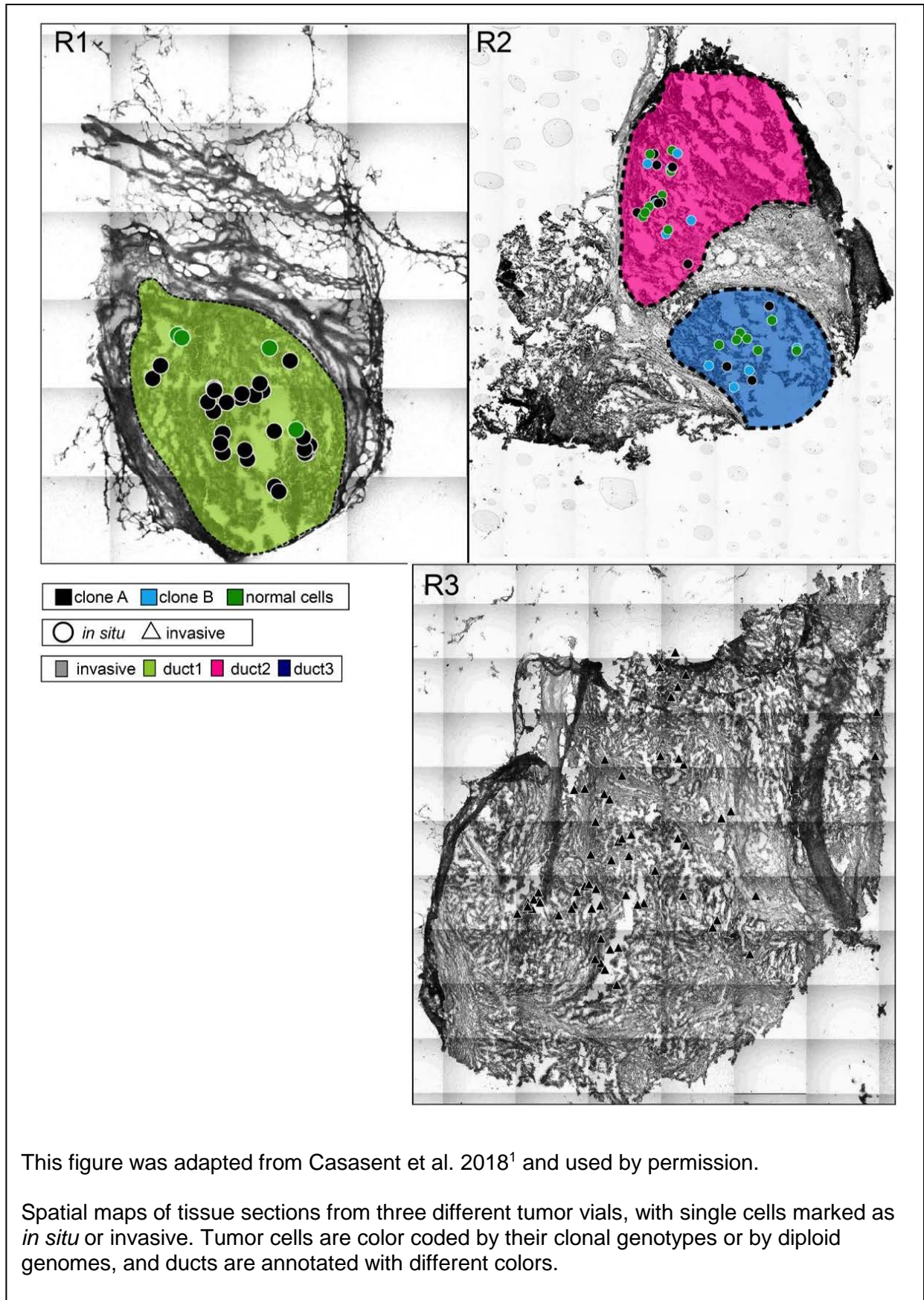
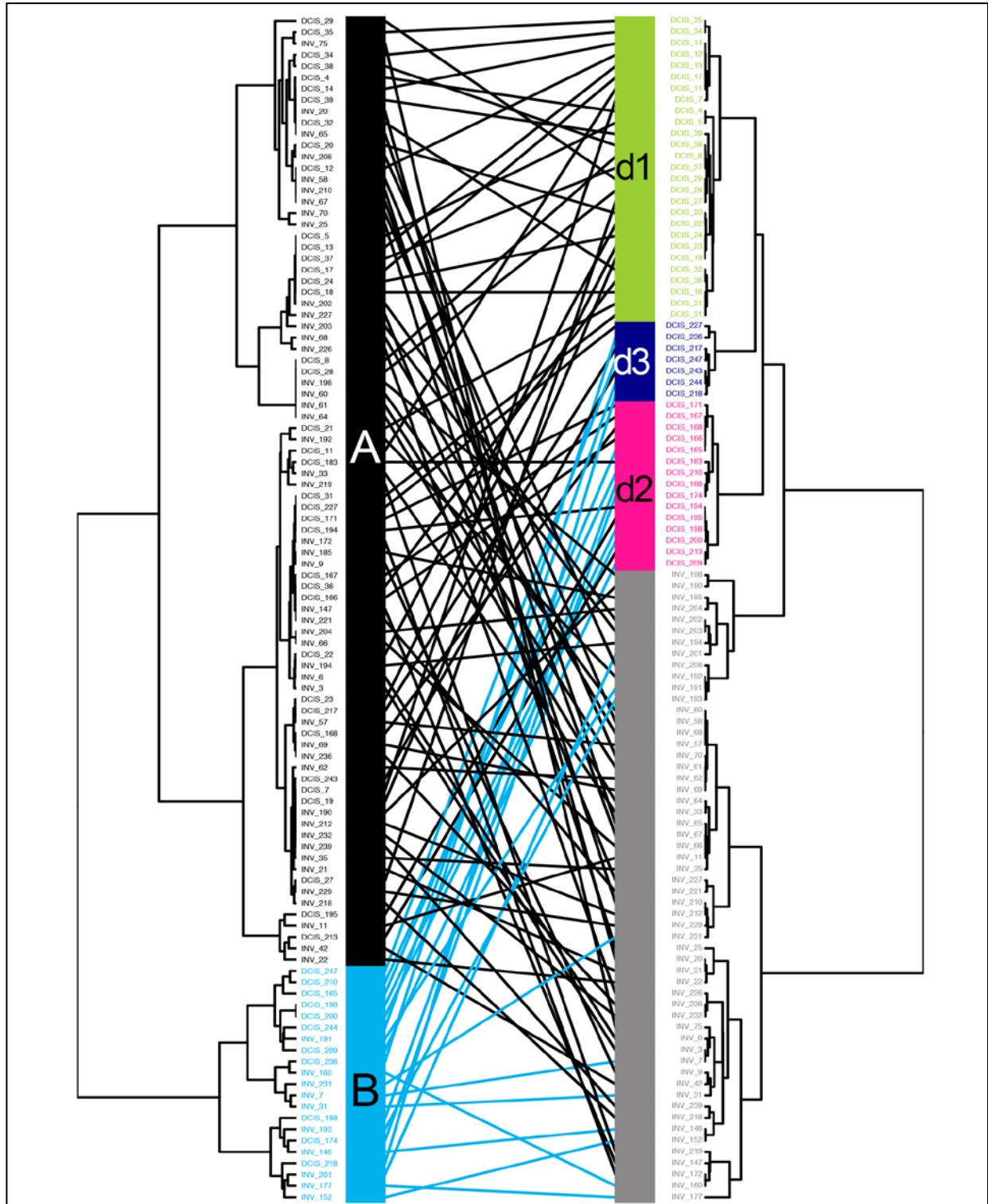




Figure 27 DC14 Tanglegram



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Genotype trees are located on the left side for each patient, with clonal subpopulations indicated by color. Spatial trees are located on the right side with different ducts indicated by colors and the invasive regions colored in gray. Mapping of cells coordinates and genotypes were performed by minimizing overlapping connections.

#### 3.2.1.4 DC16

We see some pseudo-diploid cells in DC16, something that was also observed by Goa et al<sup>200</sup>. DC16 is a grade 3 TNBC tumor and had the high number of clones observed. While the clones shared many similar events, these events further shared the common zig-zag chromosome amplification/deletion often found in TNBC samples (Figure 28 DC16 Copy Number Alteration Heatmap). Due to the high number of copy number clones observed, when we first calculated our saturation indices, we had too few cells and collect about 100 more cells with about another 50 per region to prevent under-sampling. With the new cells and k-means re-clustered, we found 5 tumor populations and once again a suggestion for 280 more cells (Figure 29 DC16 Saturation Curve). However, we stopped collection of this tumor at 204 cells total, which was enough if we only require 2 cells for each subclone instead of 3. While this is less than ideal, the number of cells needed for this tumor was much higher than other tumors, suggesting much higher diversity. Therefore, all frequency changes within this tumor are very general, despite the large number of cells collected per region. As stated before, in tumor 204 cells passed filtering, with 82 *in situ* cells and 122 invasive cells from 3 tumor sectors (R1, R2, R3).

All 5 subclones (A, B, C, D, E) in D16 are highly related and shared several similar copy number events (Figure 28 DC16 Copy Number Alteration Heatmap). In addition, because of the number of populations that were observed, we were able to observe more of a branching structure in DC16. Clone B in DC16 had fewer strong events for the first common ancestor, with the clonal events of loss of chromosome 4, amplification of 6p and 13, several focal deletions on 15, deletion of 18q, amplification of 19, 21, and 22, and a deletion of the X Chromosome. Only deletion of 18q was shared by all subpopulations. This marked sequential changes in the cells suggested this tumor went through several clonal branching. The next major population was clone C, which appears to have had several branching off points into clones A, D, and E.

Within the normal population, we would like to suggest that there was an additional clonal structure, specifically the pseudo-diploid population marked by the loss of chromosome X, a loss that is stronger in this population. Loss of chromosome X was also found in the tumor subclones, suggesting that it might be an initiating step of CIN in this tumor. In the MDS plots, we saw two normal clones separated from the group of tumor cells, which strongly supported our suggestion of separating the normal clones into normal and pseudo-normal (Chromosome X loss). The tumor cells were not as easily separated in the MDS plots, suggesting either that we are seeing sequential accumulations of alterations or intermediates, or noise which is causing the number of clusters. MDS identified 3 distinct clusters that corresponded to the normal cells (2 normals  $N_N$  and  $N_p$ ) and one cluster for the tumor clones (A, B, C, D, and E). While this lack of separation of the clonal population was marked, it suggests that these subclones are highly related, which is further supported by the correlation across subclones.

However, the MDS plot (Figure 30 DC16 MDS) did show that each clonal genotype was composed of both *in situ* and invasive tumor cells and that these were intermixed suggesting *in situ* and invasive cells arose from the same cell and were all able to escape into the surrounding tissue. While most clones stayed relatively stable, we observed a strong change in subclone E frequency which was high in *in situ* and very low in invasive regions changing from 37% to 3% (Figure 31 DC16 TimeScale). This dichotomy suggested that perhaps this clone either (1) arose after most other clones escaped or (2) was not able to escape or survive outside of the ducts. When we mapped the clones to the different sectors (R1, R2, R3), we observed that some of the sectors had more of one clone, suggesting regional effects. R2 was predominantly clone D. While we do observe intermixture of the clones in R3, many of the ducts are mostly made up of one clone type, suggesting a localized clonal expansion, with the exception of clone duct 3 in R3, which is very intermixed (See Figure 32 DC16 Image Maps and Figure 33 DC16 Tanglegram).

We observed shared amplifications of chromosomes 1, 6p, 14, 19, and 21, as well as deletions of chromosomes 4, 15, and X between all clones. However, for each subclone, we

observed unique amplifications or deletions. For clone A, we found amplifications of 3p (EVI1 and PIK3CA, shared with clones C, D, and E) and 13 (BRCA2, RB1 and ERCC5, shared with B, C, and D) with a deletion of a more complete deletion of all of chromosome 2 unique to A. We saw the most noise of any of the clones in B but saw a more complete deletion of chromosome 4 and strong deletion of 18 (MATI1 and BCL2, shared with C and E). Clone C had the deletion in the first part of 2q (ERCC3, PNS1, and CREB1, shared with E) and a specific focal deletion on 6q (MLT4). For clone D, we saw an amplification of 13q (ERCC5, shared with E). Lastly, clone E had the deletion of chromosome 5 (Figure 28 DC16 Copy Number Alteration Heatmap). To delineate clonal evolution during invasion, we inferred genomic lineages from these consensus plots and constructed a possible phylogeny where first clone B arose followed by clone C, from which clones A, B and E arose (Figure 31 DC16 TimeScape).

Most notably each clone consisted of single cells from both the *in situ* and invasive cells. While there was a decrease in the frequency of clone E, we observed that all clones were present in all regions and most likely arose within the ducts prior to invasion and the clones shared so much similarity that it was difficult to separate them in the MDS plot.



Figure 28 DC16 Copy Number Alteration Heatmap

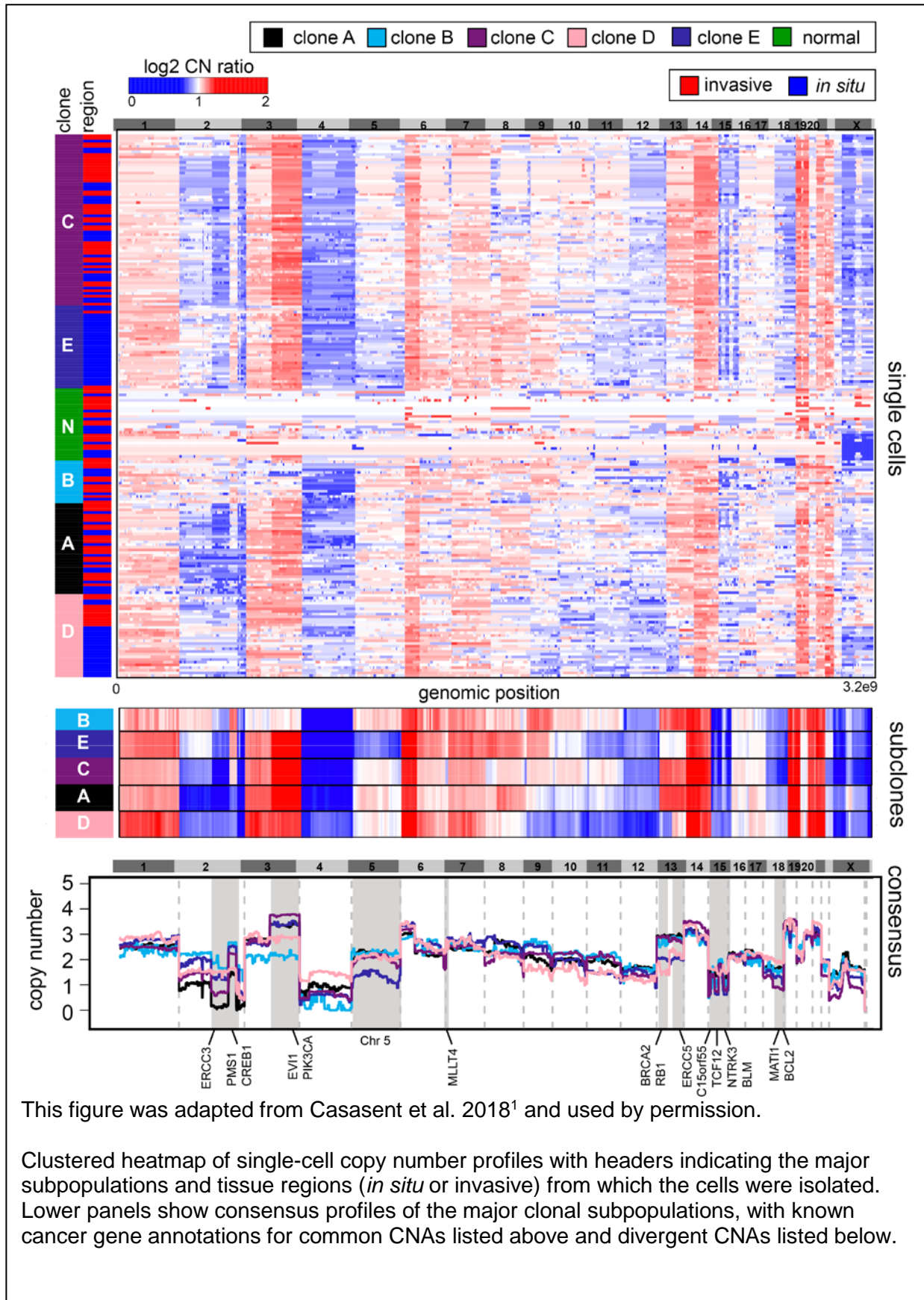
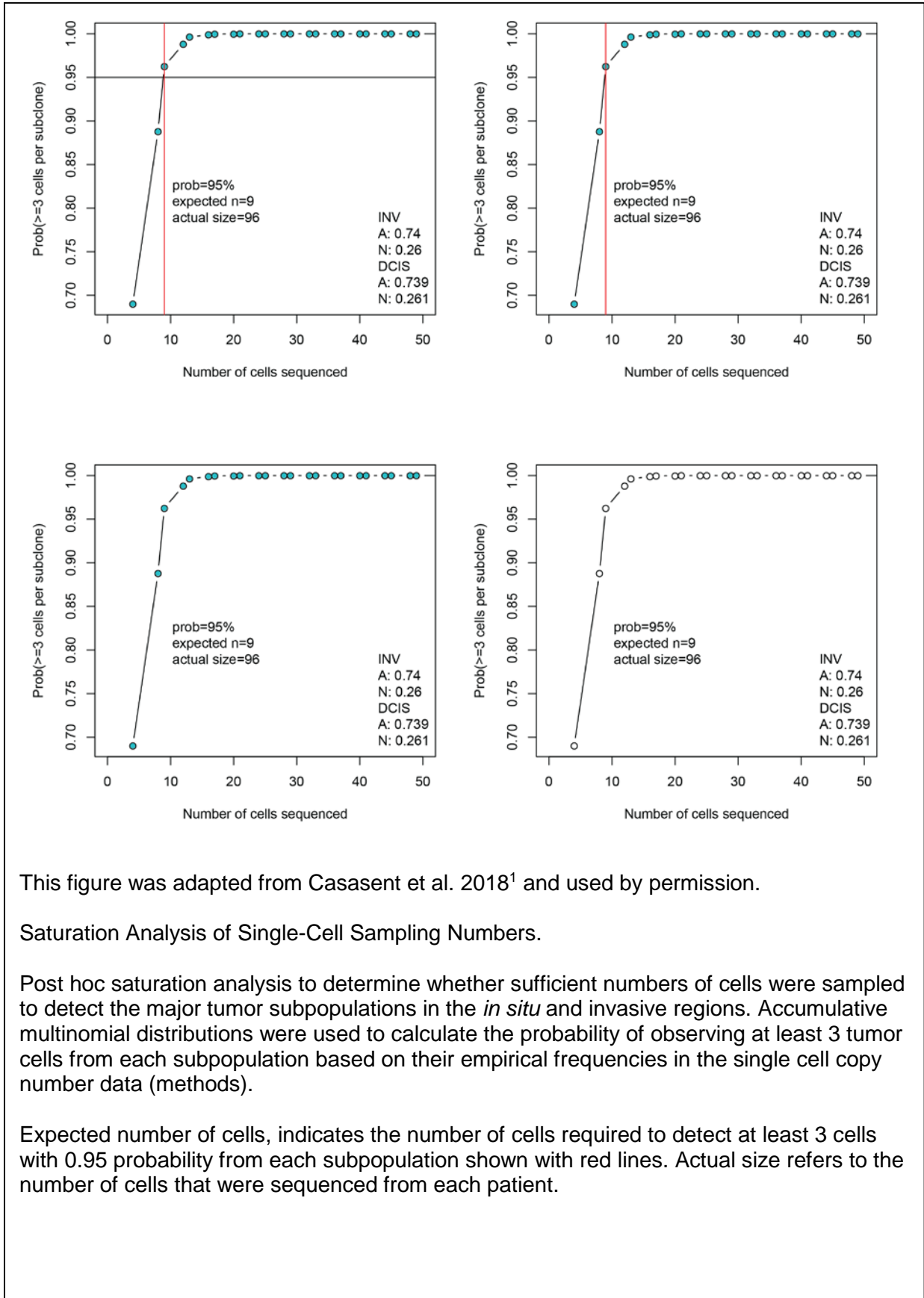


Figure 29 DC16 Saturation Curve



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Saturation Analysis of Single-Cell Sampling Numbers.

Post hoc saturation analysis to determine whether sufficient numbers of cells were sampled to detect the major tumor subpopulations in the *in situ* and invasive regions. Accumulative multinomial distributions were used to calculate the probability of observing at least 3 tumor cells from each subpopulation based on their empirical frequencies in the single cell copy number data (methods).

Expected number of cells, indicates the number of cells required to detect at least 3 cells with 0.95 probability from each subpopulation shown with red lines. Actual size refers to the number of cells that were sequenced from each patient.

Figure 30 DC16 MDS

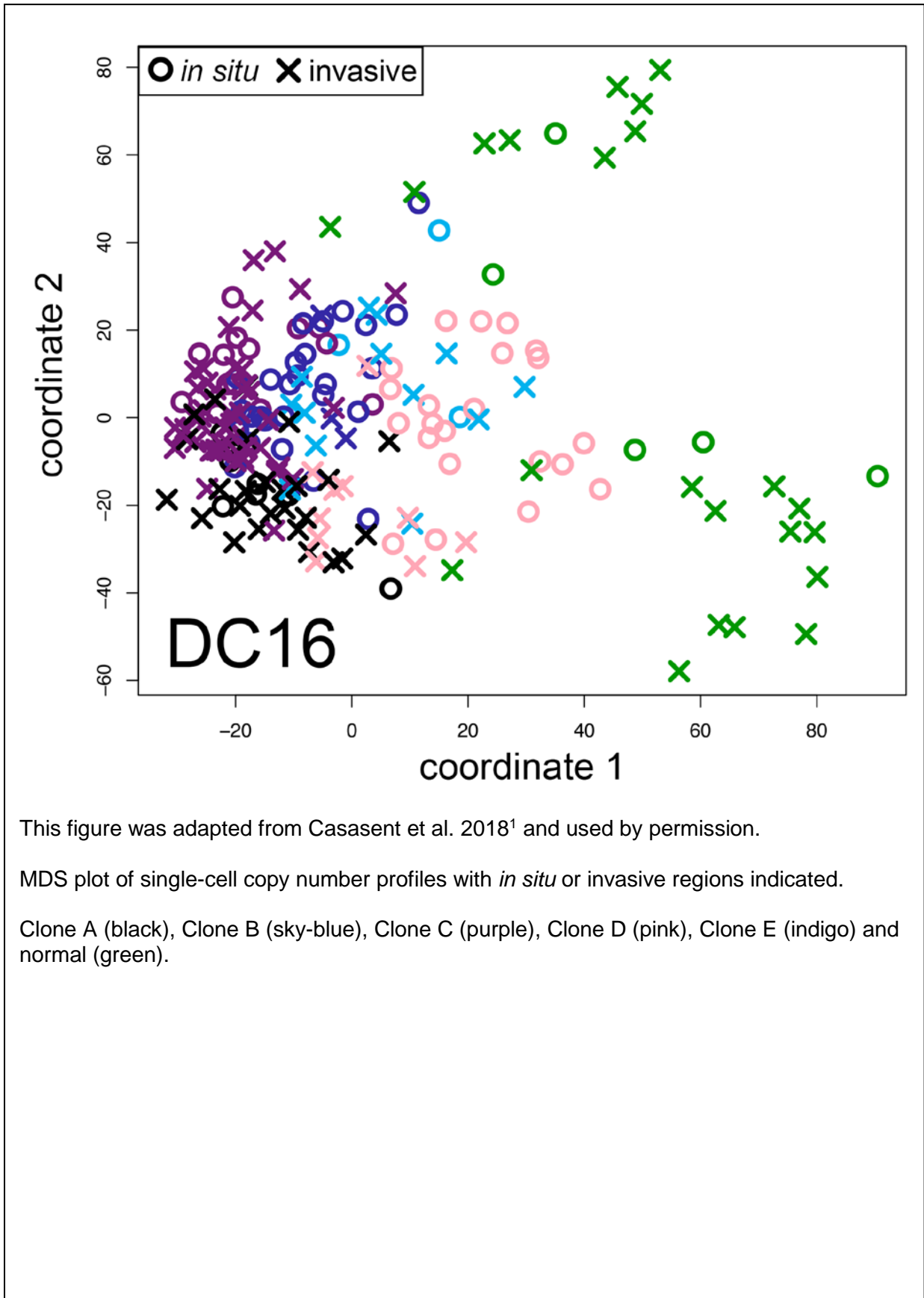
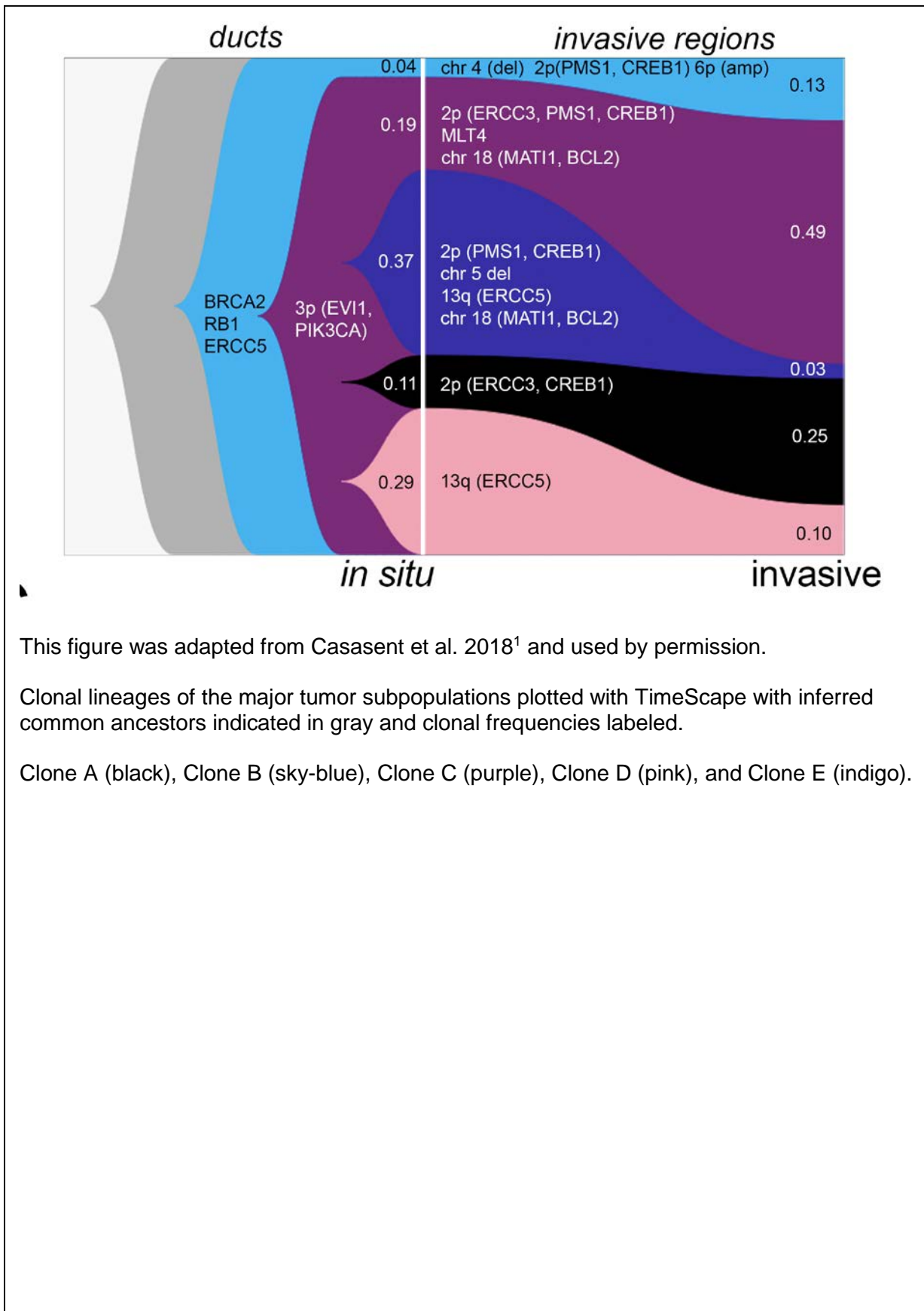


Figure 31 DC16 TimeScape



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Clonal lineages of the major tumor subpopulations plotted with TimeScape with inferred common ancestors indicated in gray and clonal frequencies labeled.

Clone A (black), Clone B (sky-blue), Clone C (purple), Clone D (pink), and Clone E (indigo).

Figure 32 DC16 Image Maps

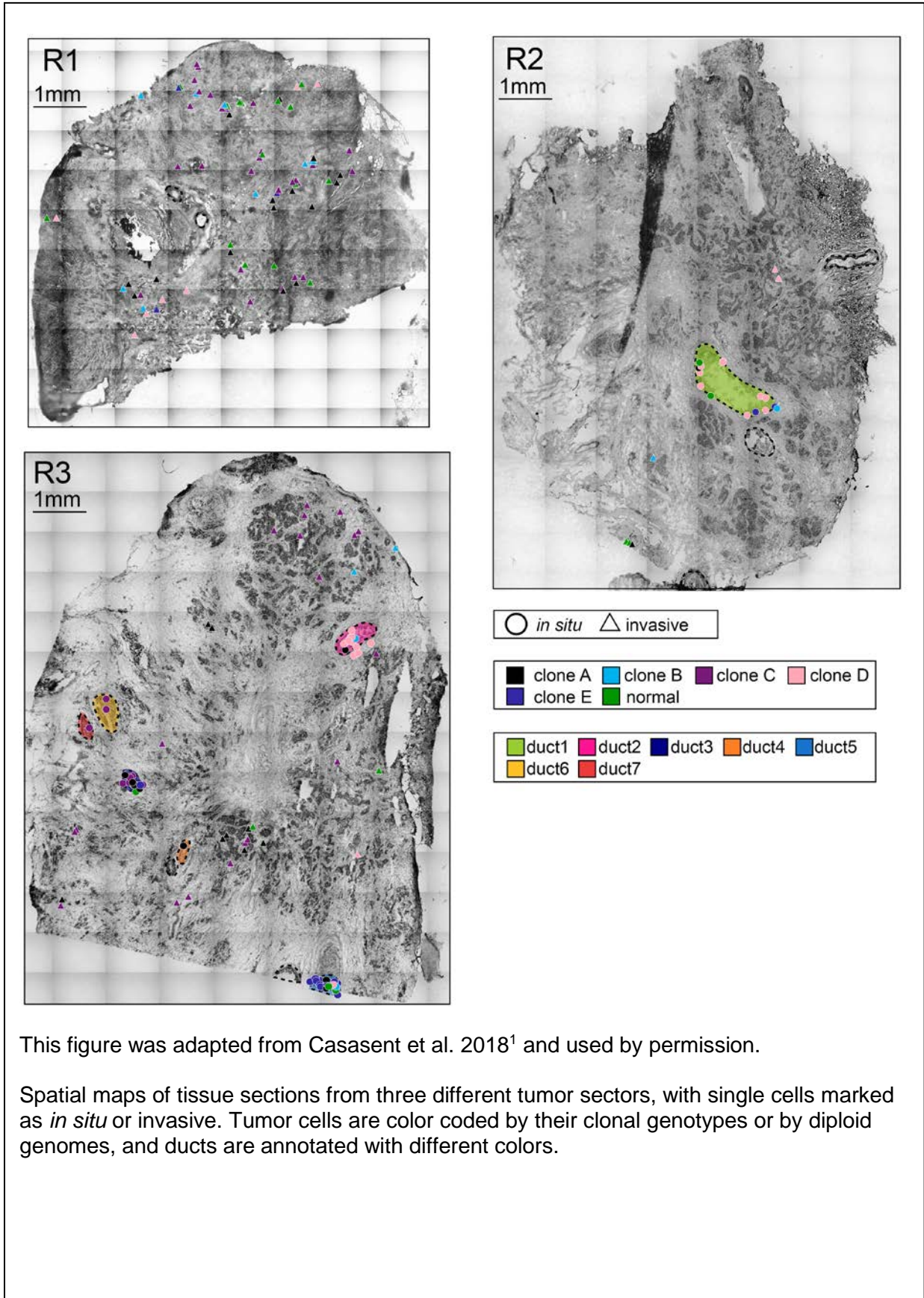
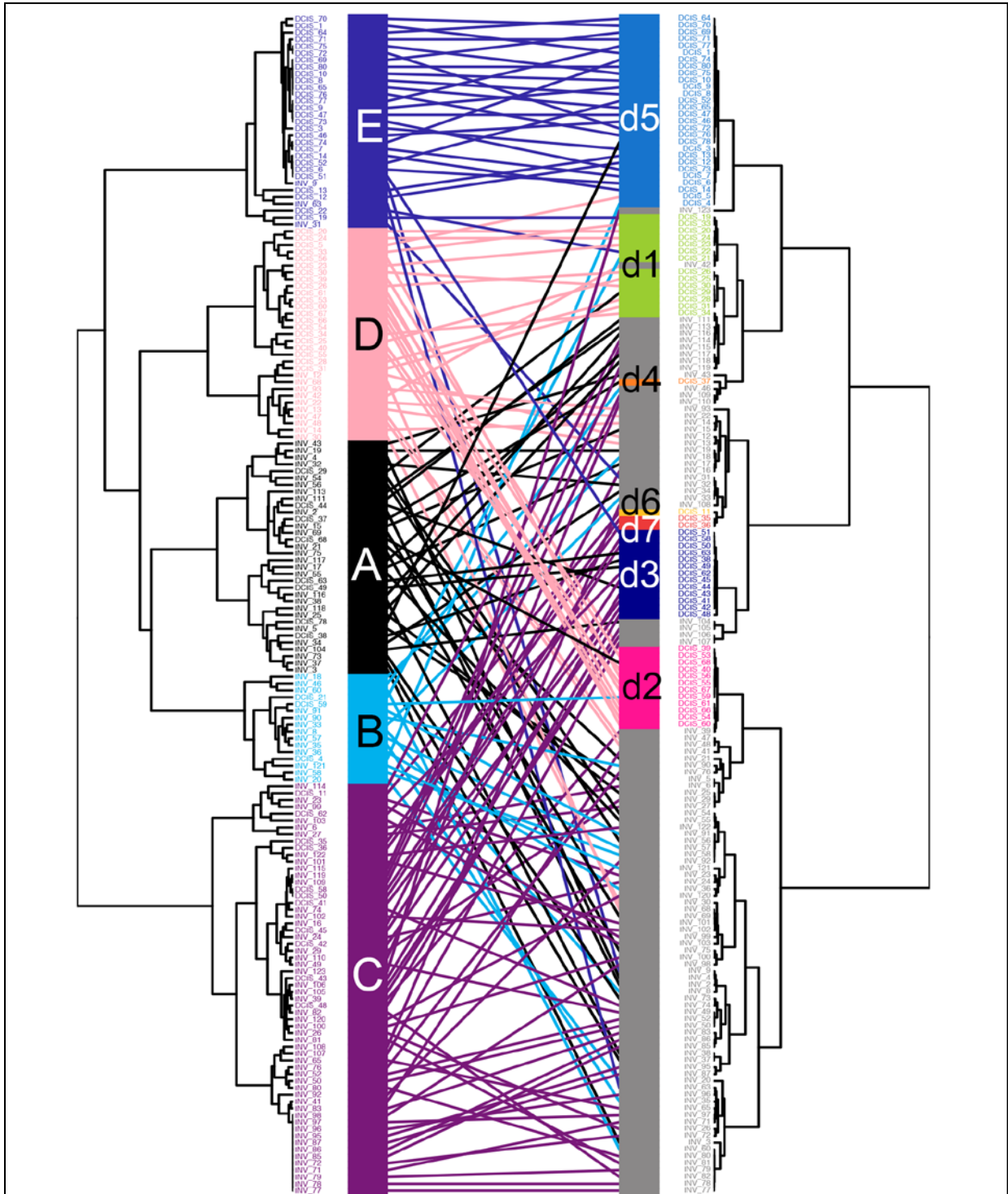




Figure 33 DC16 Tanglegram



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Genotype trees are located on the left side for each patient, with clonal subpopulations indicated by color. Spatial trees are located on the right side with different ducts indicated by colors and the invasive regions colored in gray. Mapping of cells coordinates and genotypes were performed by minimizing overlapping connections. Clone A (black), Clone B (sky-blue), Clone C (purple), Clone D (pink), and Clone E (indigo).

### 3.2.1.5 DC18

We investigated copy number evolution during invasion in patient DC18, a grade 3 TNBC, using TSCS to profile 85 *in situ* cells and 150 invasive cells from tissue sections from four different tumor regions (R1-R4) (Figure 35 DC18 Saturation Curve). We performed 1-dimensional clustering, which revealed 1 major population of diploid cells (N) and 3 clonal aneuploid tumor subpopulations (A, B, C). Within each subpopulation (A, B, C), the copy number profiles were highly correlated (A=0.64, B=0.71, C=0.80, Pearson correlations), representing stable clonal expansions. Consensus profiles were calculated and compared from each tumor subpopulation, which identified shared amplifications on chromosome 2p (*ALK*), 8q (*MYC*), 14q (*FOXA1*), and 21q (*RUNX1*), in addition to many subpopulation-specific CNAs. Clone A had focal deletions in chromosome 4p (*RHOH*), 9p (*CDKN2A*), and Xq (*COL4A5*), as well as focal amplifications on chromosome 17p (*MAP2K3*, *NF1*, *BCAS3*), 12p (*ALG10B* and *ERBB3*), and chromosome Xq (*AR*). Clone B had deletions on chromosome 3p (*FHIT*), 13 (*RB1*), and 8p (*DBC2*), as well as amplifications on chromosomes 2q (*GALNT13*), 11p (*WT1*), and Xp (*PDK3*). Clone C shared many CNA events with clone B, including an amplification on 7p (*EGFR*) (See Figure 34 DC18 Copy Number Alteration Heatmap).

To delineate clonal evolution during invasion, we inferred genomic lineages and plotted the data using TimeScape (Figure 37 DC18 TimeScape). This analysis identified a common ancestor that evolved in the ducts with amplifications of *ALK*, *MYC*, *FOXA1*, and *RUNX1* that subsequently diverged to form clones A and C. Clone B was a common ancestor of clone C, but diverged and evolved additional CNAs in *RB1*, *FHIT*, and *DBC2*. This data showed that all 3 subclones evolved in the ducts from a common ancestor prior to invasion, and subsequently migrated into the surrounding tissues where they underwent stable clonal expansions. These data did not detect any new CNAs acquired in the clones during invasion, but did reveal a decreased frequency of subclone A (40% to 5%) in the invasive regions.

To understand the relationship between the clonal genotypes and their spatial positions, we performed multi-dimensional scaling (MDS), which identified 4 discrete clusters

corresponding to different subpopulations (1 normal cells and 3 tumor subpopulations; See Figure 36 DC18 MDS). Each subpopulation consisted of single cells isolated from both the *in situ* and invasive cells, with no clonal genotype specifically associated with the *in situ* or invasive regions. MDS showed that subpopulations C and B were adjacent in high-dimensional space, while subpopulation A was the most distant.

The clonal genotypes were mapped to their spatial coordinates in the four tissue sections (R1-R4) to delineate their topography, which showed that all three tumor clones were localized to both the ductal and the invasive regions, with no single genotype mapping exclusively to one region (Figure 38 DC18 Image Maps and Figure 39 DC18 Tanglegram). However, clone A was more restricted to the ductal regions (R3), while clones B and C were more frequent in the invasive regions. Consistent with the spatial distributions, we found that clones B and C had an amplification of *EGFR* previously shown to be associated with cell migration<sup>225</sup>, while clone C had an additional deletion of *FHIT* known to suppress EMT and cell migration<sup>226</sup>.



Figure 34 DC18 Copy Number Alteration Heatmap

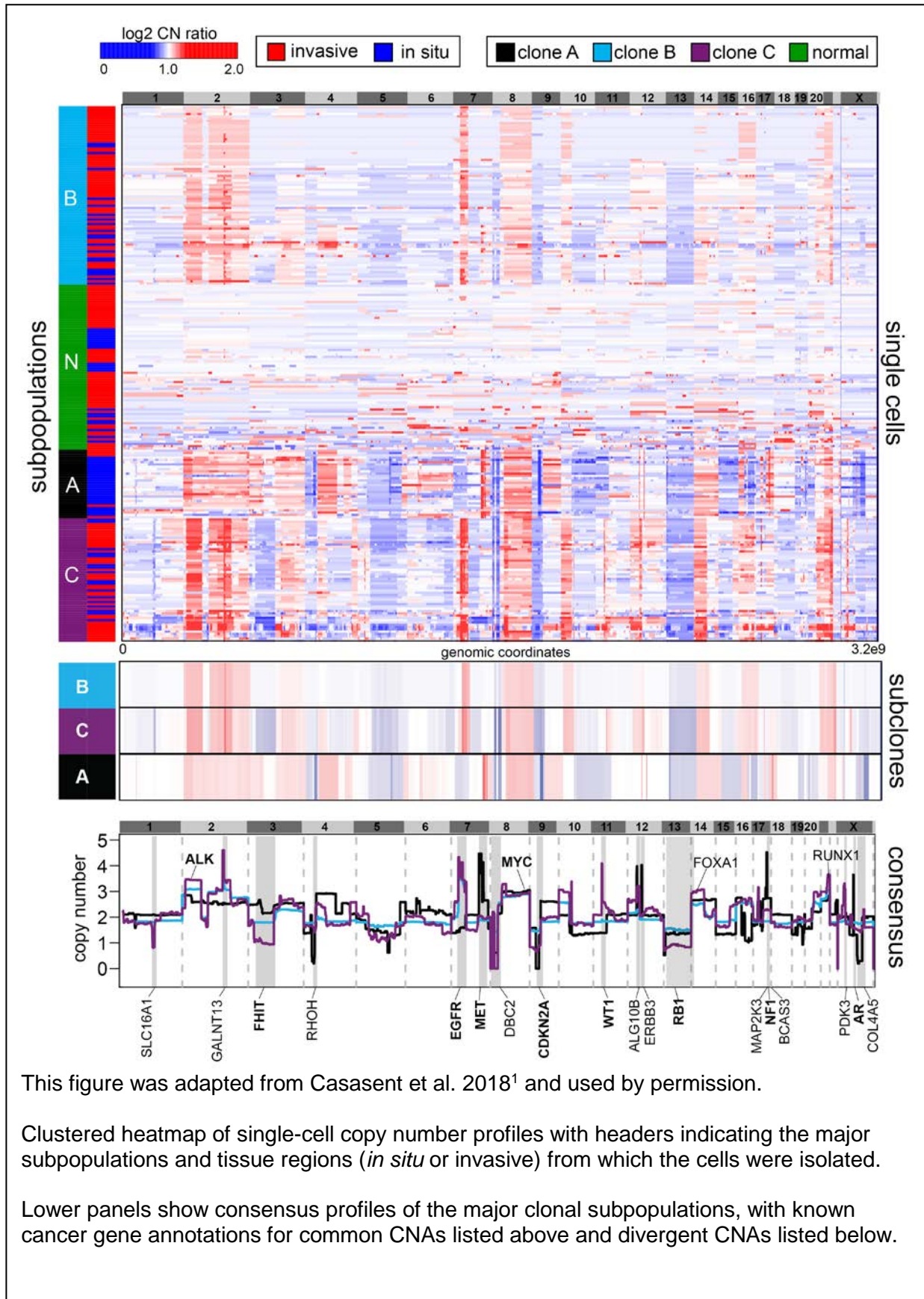
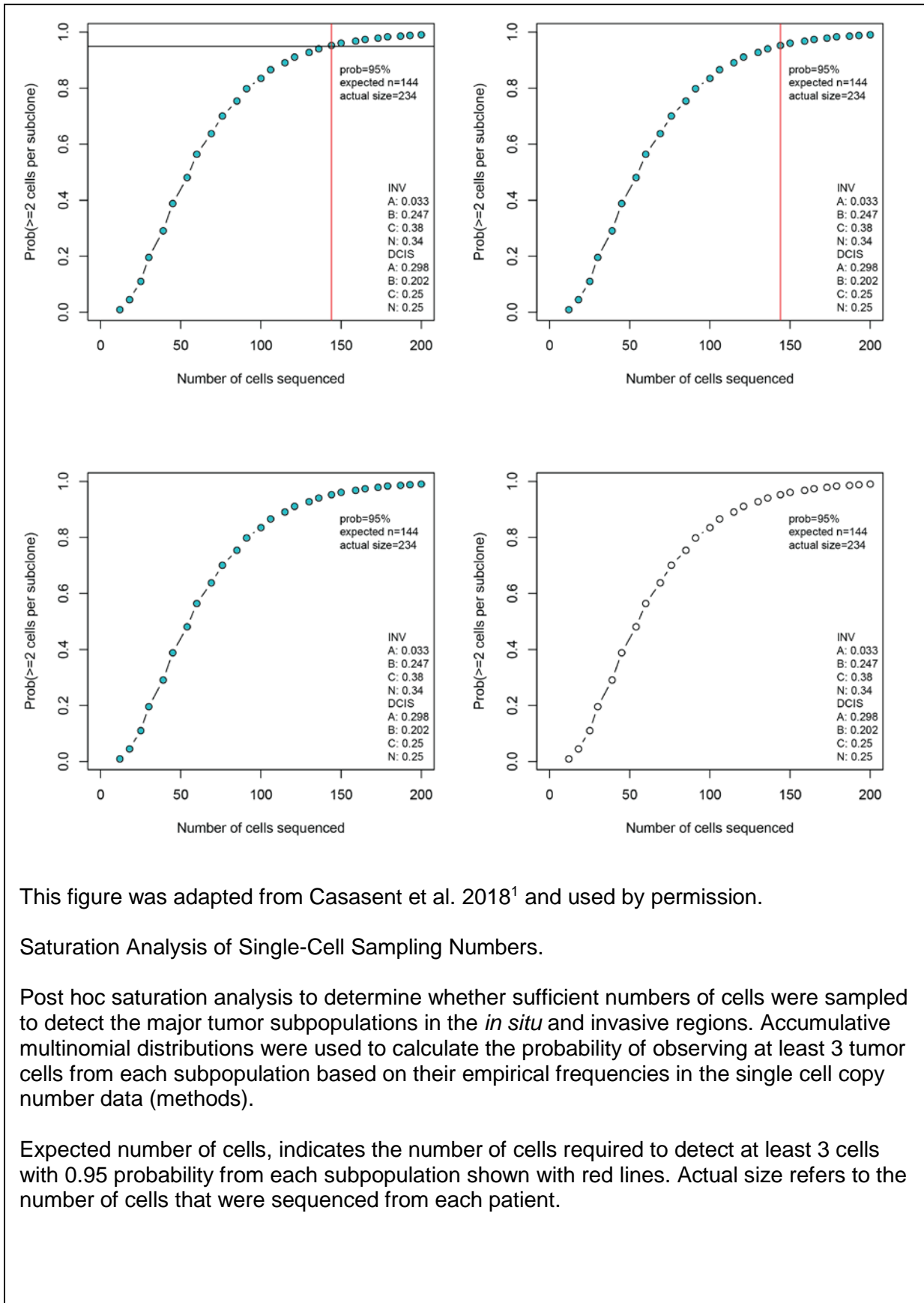


Figure 35 DC18 Saturation Curve



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Saturation Analysis of Single-Cell Sampling Numbers.

Post hoc saturation analysis to determine whether sufficient numbers of cells were sampled to detect the major tumor subpopulations in the *in situ* and invasive regions. Accumulative multinomial distributions were used to calculate the probability of observing at least 3 tumor cells from each subpopulation based on their empirical frequencies in the single cell copy number data (methods).

Expected number of cells, indicates the number of cells required to detect at least 3 cells with 0.95 probability from each subpopulation shown with red lines. Actual size refers to the number of cells that were sequenced from each patient.

Figure 36 DC18 MDS

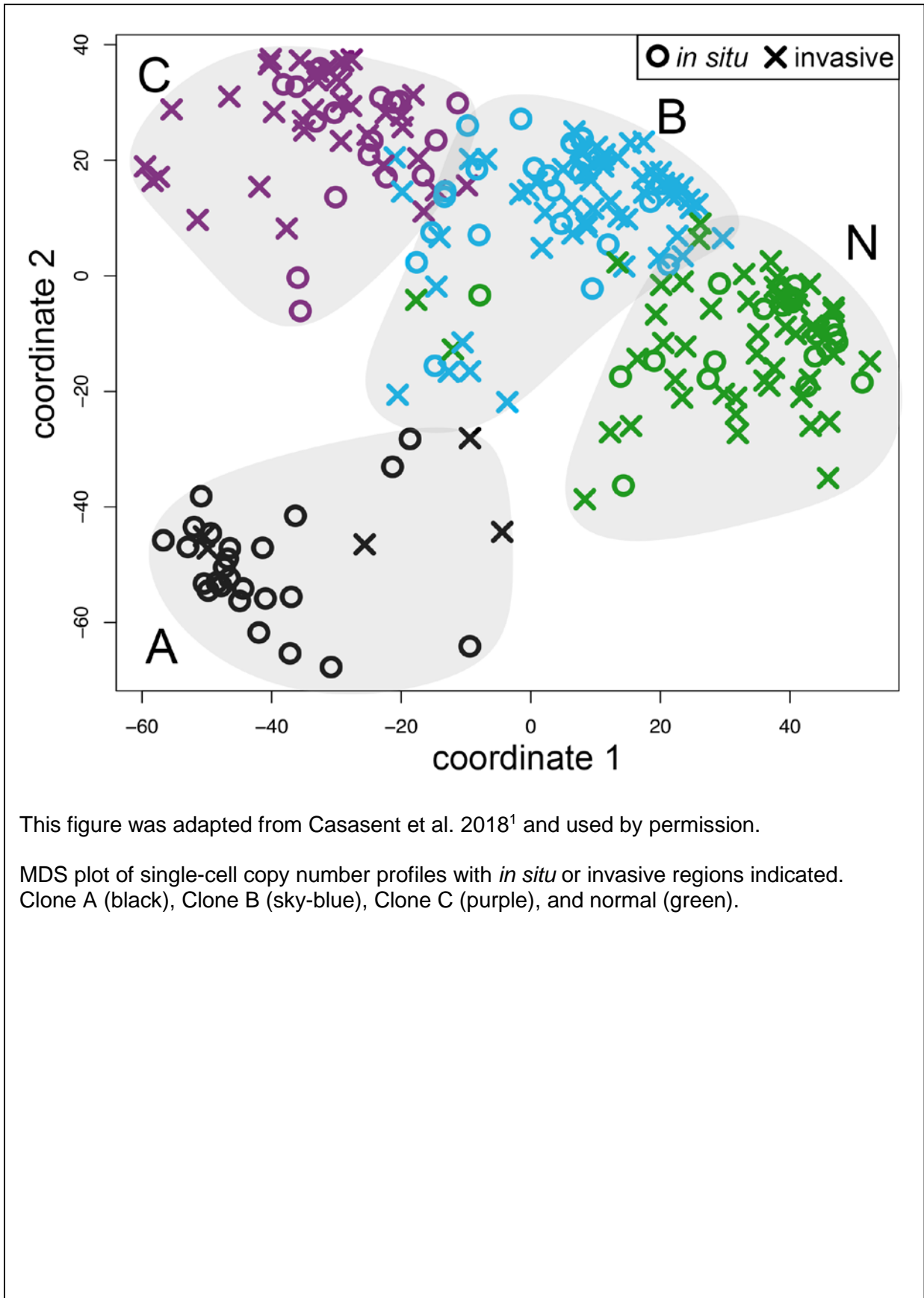
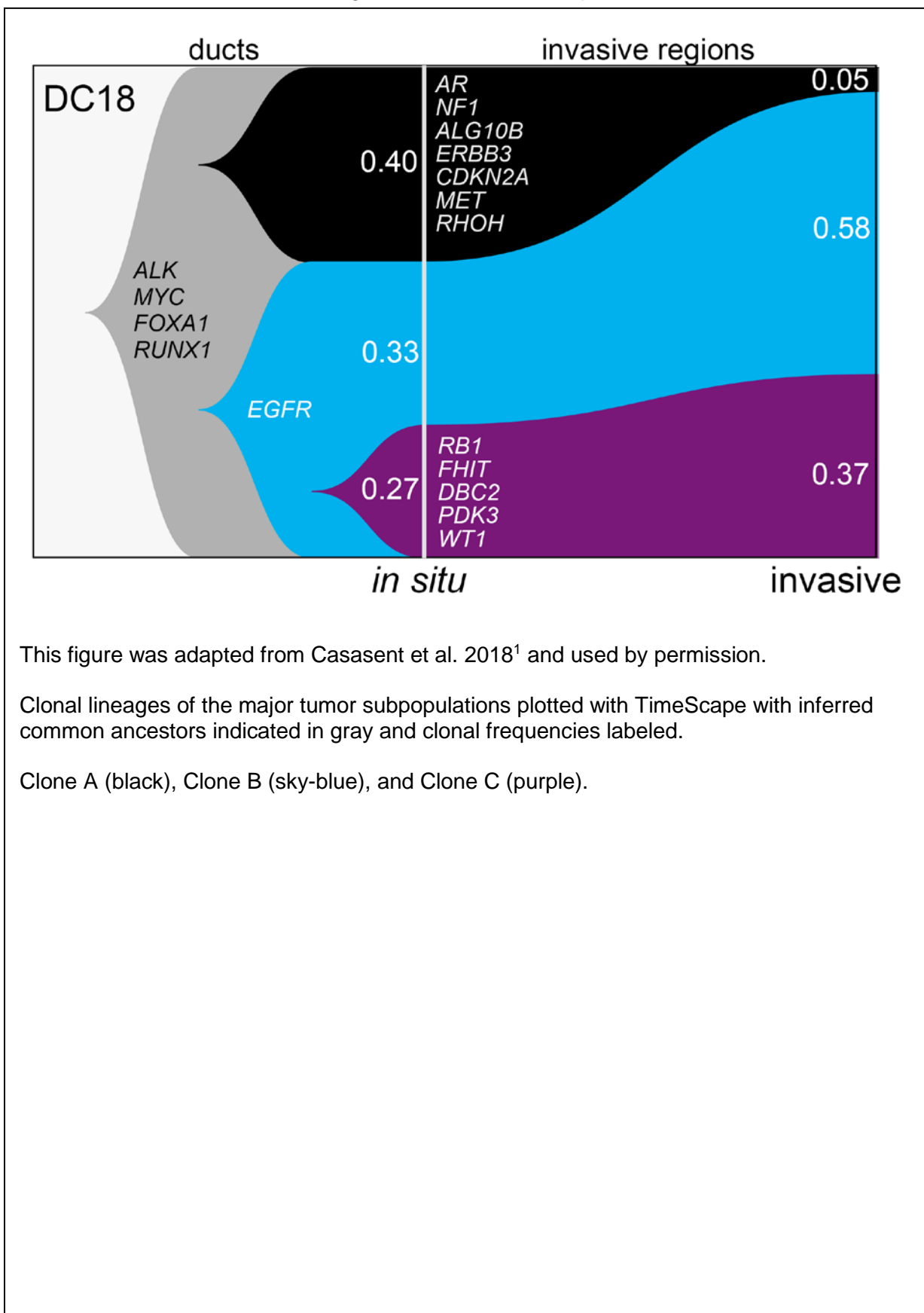


Figure 37 DC18 TimeScape



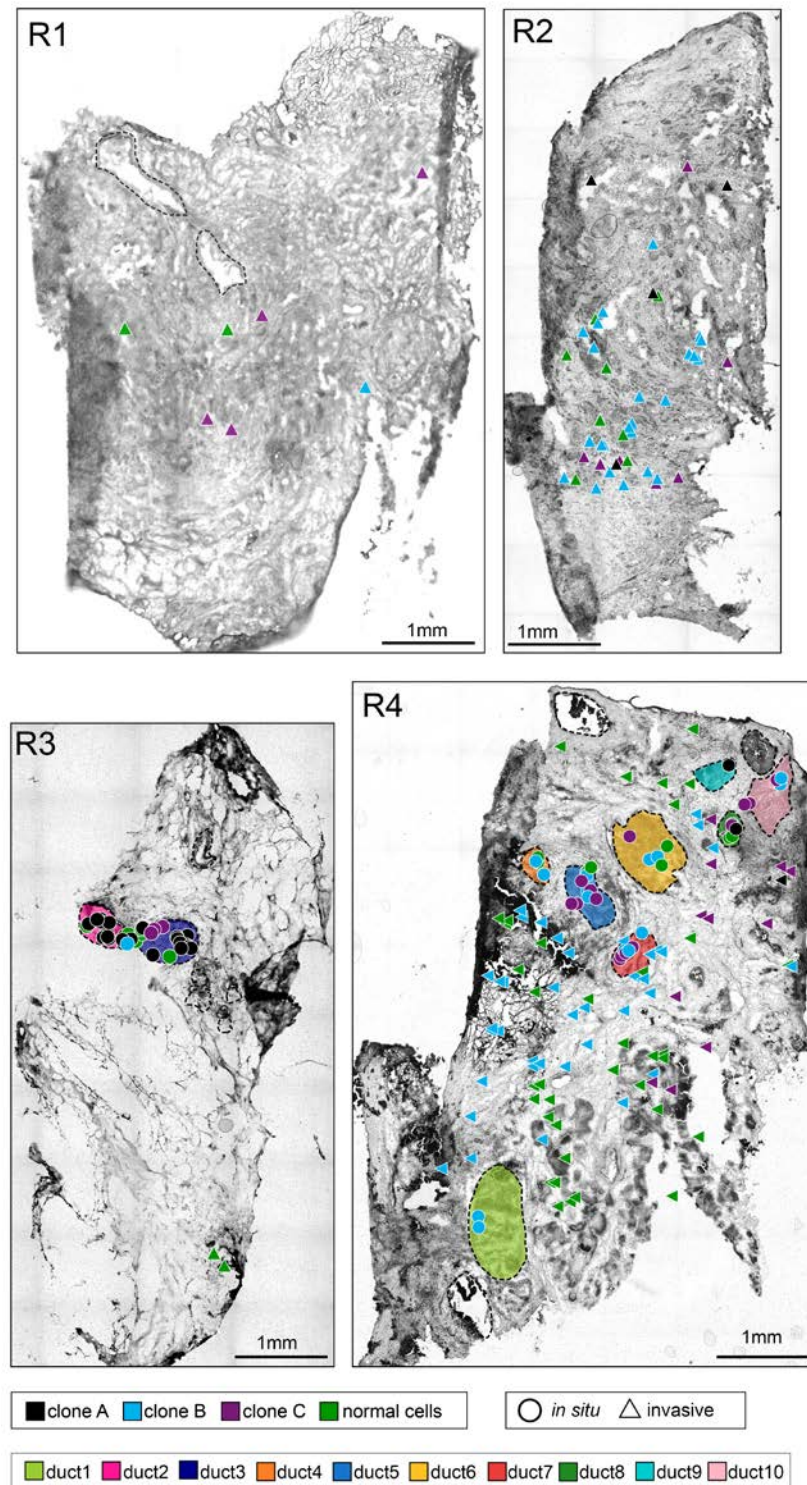
This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Clonal lineages of the major tumor subpopulations plotted with TimeScape with inferred common ancestors indicated in gray and clonal frequencies labeled.

Clone A (black), Clone B (sky-blue), and Clone C (purple).



Figure 38 DC18 Image Maps



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Spatial maps of tissue sections from four different tumor regions, with single cells marked as *in situ* or invasive. Tumor cells are color coded by their clonal genotypes or by diploid genomes, and ducts are annotated with different colors.



### 3.2.1.6 DC20

We found DC20, grade 3 TNBC, to be a polyclonal tumor. The three clonal populations (A, B, C) shared several events between all three clones, strongly suggesting the same single clone of origin (Figure 40 DC20 Copy Number Alteration Heatmap). Most notably, all clones shared a zig-zag chromosome amplification, a deletion pattern commonly noted in TNBC samples. More specifically, multiple CNAs occurred in almost every chromosome, with shared breakpoints at larger events on chromosomes 1, 7, 10, 12, 17, 18, and 21, as well as focal share alterations on 2, 3, 6, 7, 11, and 13.

The close relationship of these 3 populations is demonstrated in the MDS plot, which clearly shows the 2 "normal" cell populations and 1-2 tumor cell clones slightly overlapped, suggesting a common cell of origin (Figure 42 DC20 MDS). The normal cells in the case of DC20 appear to be either (1) very noisy, (2) intermediates, or (3) alternative cells of origin. The normal cells do not cluster well together and N<sub>1</sub> shows strong amplifications of chromosome 1, 7, 19, and X, while N<sub>2</sub> shows deletions of chromosome 4, 10, 20, and X. Since these break points are not shared with the other clones, it suggests these chromosome alterations might have occurred but not expanded or that the profiles are noisy.

We examined DC20 in three different sectors (R1, R2, and R3), and we saw strong intermixture of the different clones, suggesting these clones stayed intermixed throughout progression (Figure 44 DC20 Image Maps). The consensus copy number profiles between clone A, B, and C were highly correlated, and on the MDS plot it is difficult to separate clones A and C, while clone B appears to be more distinct. When examining the single cell copy number heatmap in detail, we can see that clones A and C appear more clonal, while B has more variation. However, even with clone A, we still distinct changes between deletions on chromosome 5 (APC), which is shared with clone C but not B, but is highly variant in clone A and B. This variance suggests that the deletion might in fact represent a different set of subclones. Clone A could also be broken up into 3 clones based on deletions in chromosome 7. These deletions are not shared with clone C, which has an amplification at the same

position, or clone B, which is not changed in this position (Figure 40 DC20 Copy Number Alteration Heatmap). These results suggest that we might need to look at more sensitive clustering metrics to truly delineate clonal substructure. Lastly, when we examined subclone B, we saw several unique CNA not shared with any profile, suggesting either that these profiles are noisy or that these profiles represent several intermediates or dead-ends.

If there are in fact more subclones than we observed, our current requirements for number of cells based on the saturation index is obsolete, and we would require many more cells. We found some clonal change between *in situ* and invasive frequency, the largest being the decreases of clone A from 69% to 31% and the reciprocal increases of clone B from 8% to 26% and clone C 23% to 43% (Figure 43 DC20 TimeScape). This might suggest that clones B and C are more invasive or merely that clone A was not able to survival as well outside of the ducts.

Next, we mapped the clonal genotypes to their spatial coordinates on an image map for DC20 sectors (R1, R2, R3), with R2 showing *in situ* populations and R1 and R3 showing both *in situ* and invasive populations. We see some intermixture of all clones both in the *in situ* and invasive sectors, but clone A is more highly localized to a region (R2) which had very few invasive cells collected, perhaps suggesting instead a regional bias caused by non-random sampling (Figure 44 DC20 Image Maps and Figure 45 DC20 Tanglegram). In R2, very few invasive cells were collected due to two issues: (1) most of the cells surrounding the ducts appeared to be smaller normal cells and (2) the tumor cells of notice were tightly clustered (less than 1 micron distance between nuclei), making isolating a single cell not practical.



Figure 40 DC20 Copy Number Alteration Heatmap

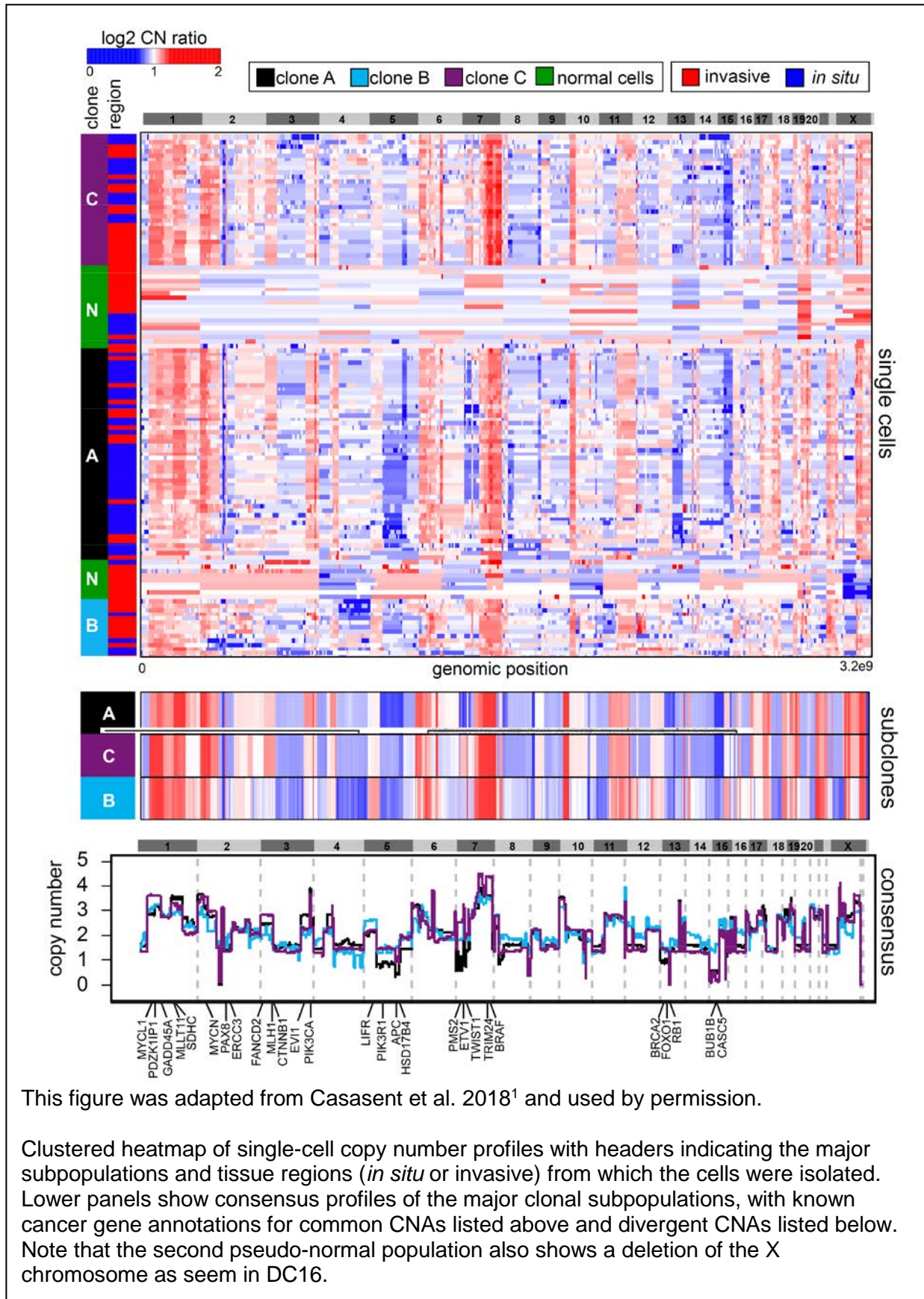
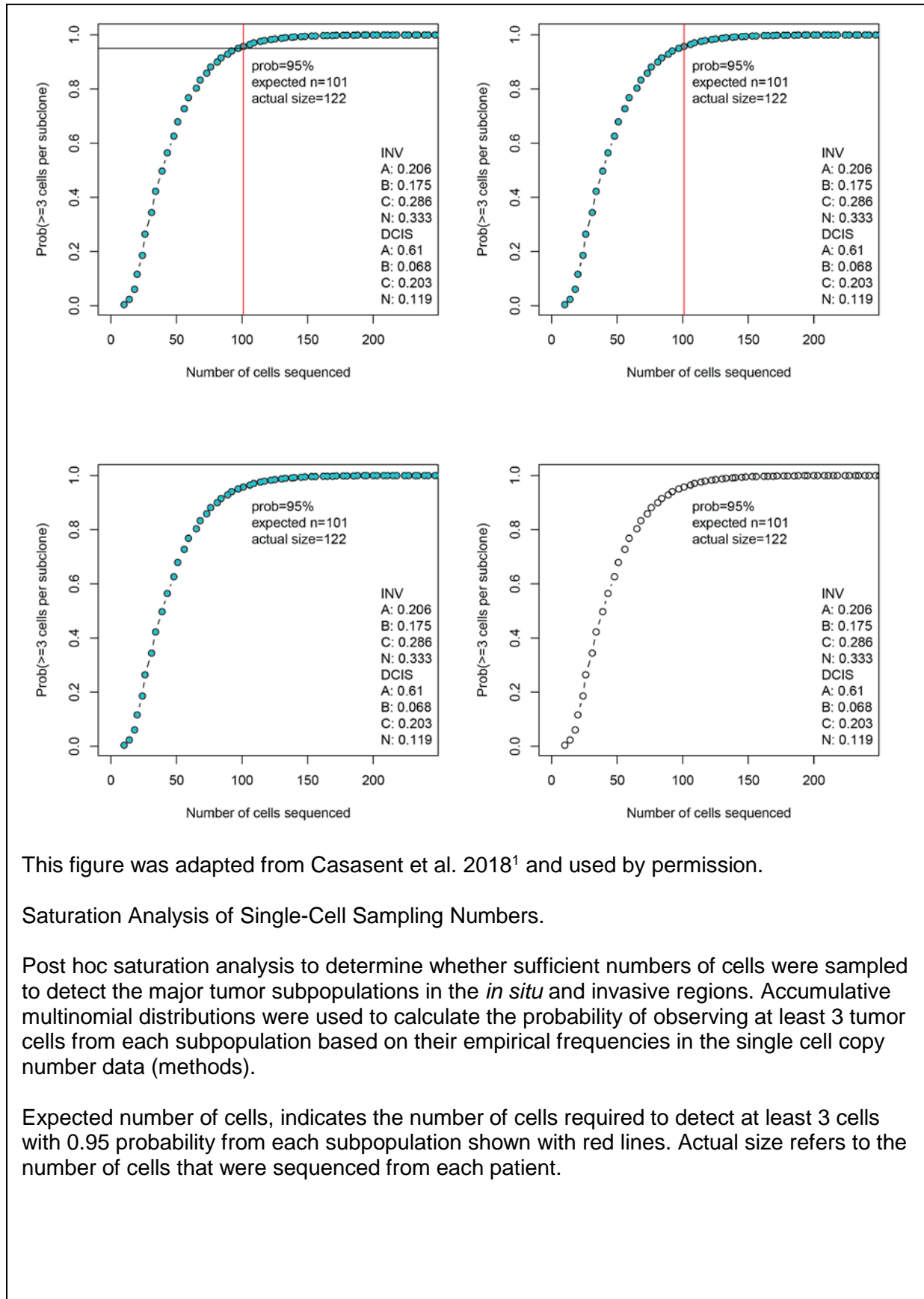


Figure 41 DC20 Saturation Curve



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Saturation Analysis of Single-Cell Sampling Numbers.

Post hoc saturation analysis to determine whether sufficient numbers of cells were sampled to detect the major tumor subpopulations in the *in situ* and invasive regions. Accumulative multinomial distributions were used to calculate the probability of observing at least 3 tumor cells from each subpopulation based on their empirical frequencies in the single cell copy number data (methods).

Expected number of cells, indicates the number of cells required to detect at least 3 cells with 0.95 probability from each subpopulation shown with red lines. Actual size refers to the number of cells that were sequenced from each patient.

Figure 42 DC20 MDS

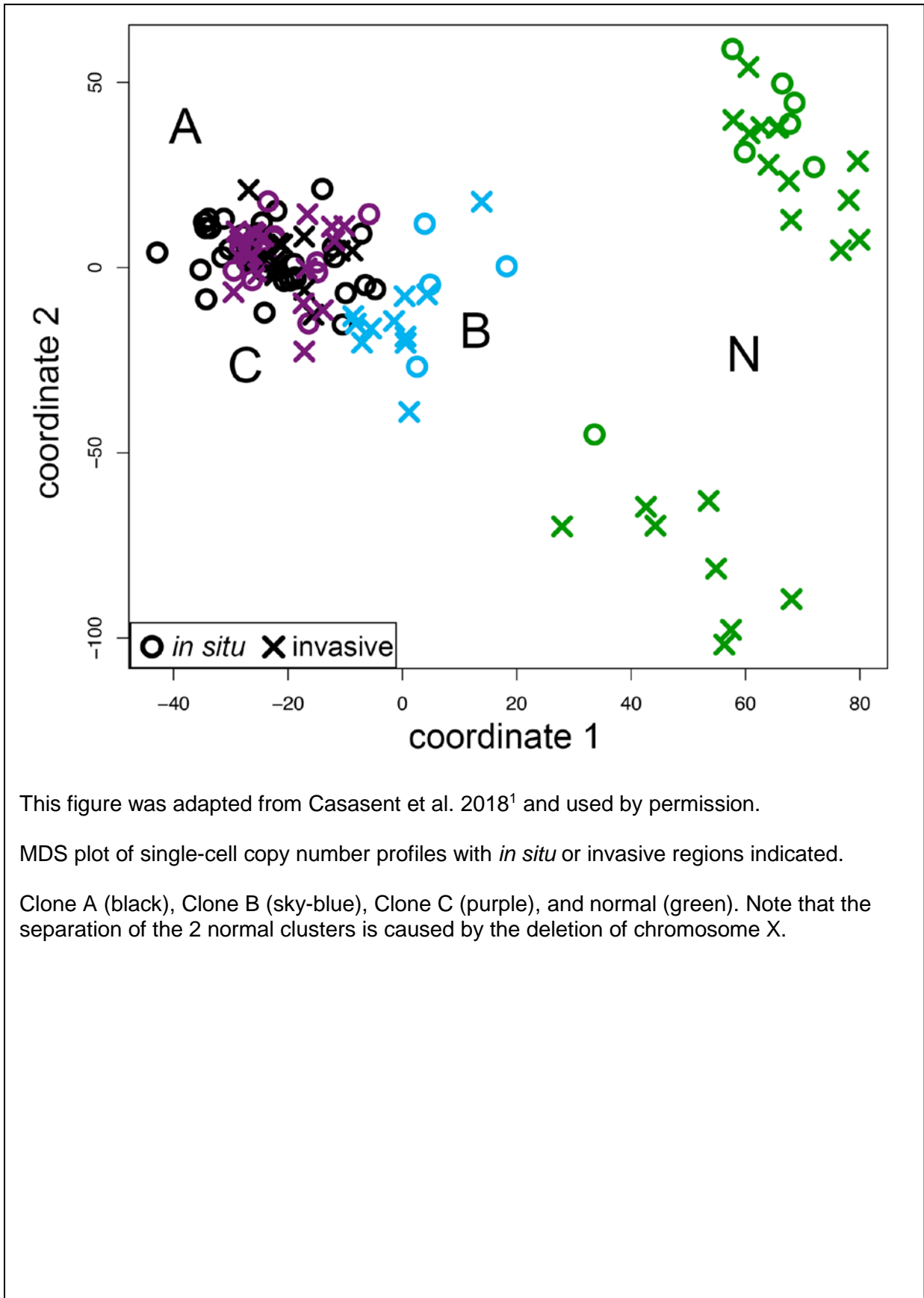
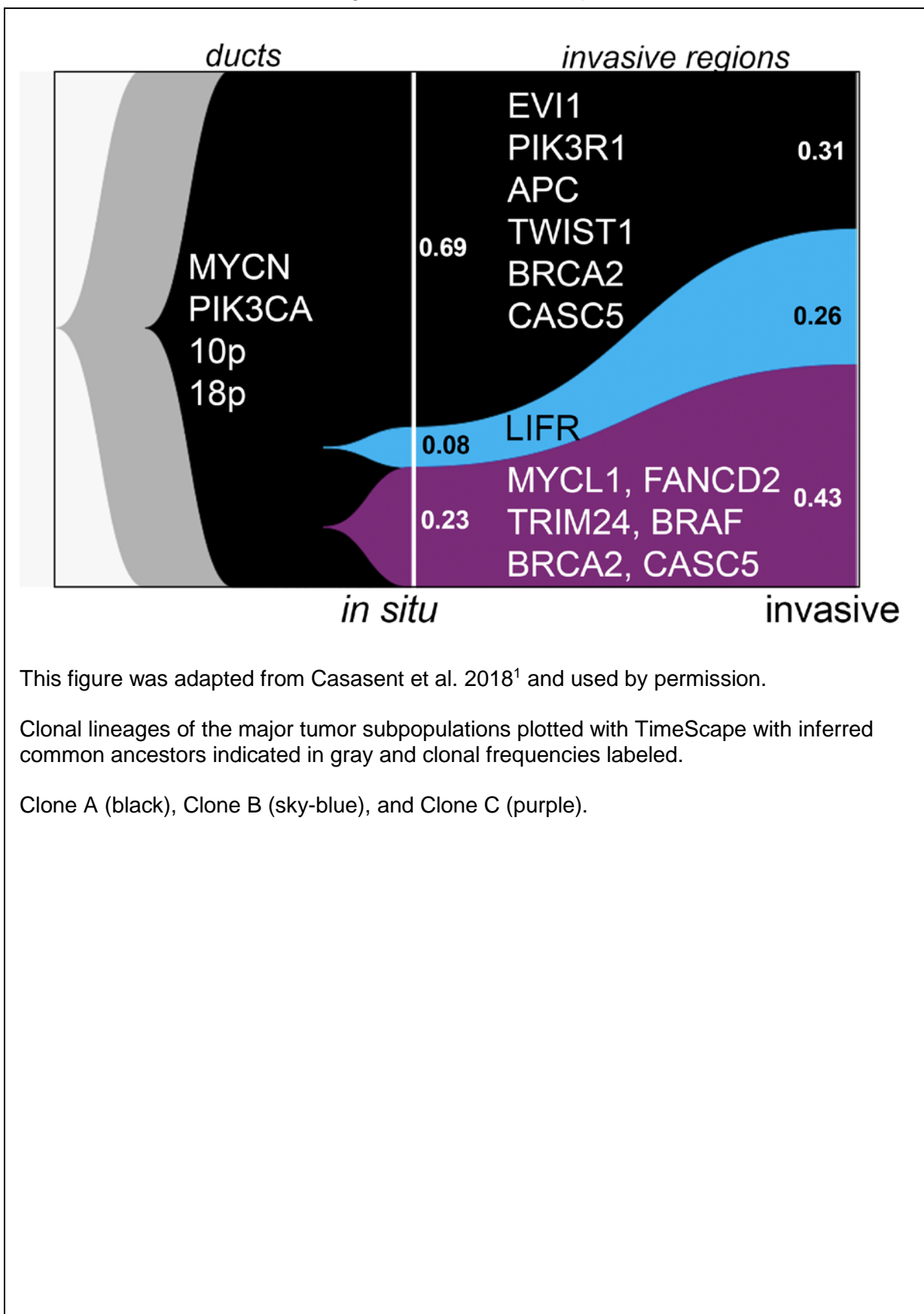


Figure 43 DC20 TimeScape



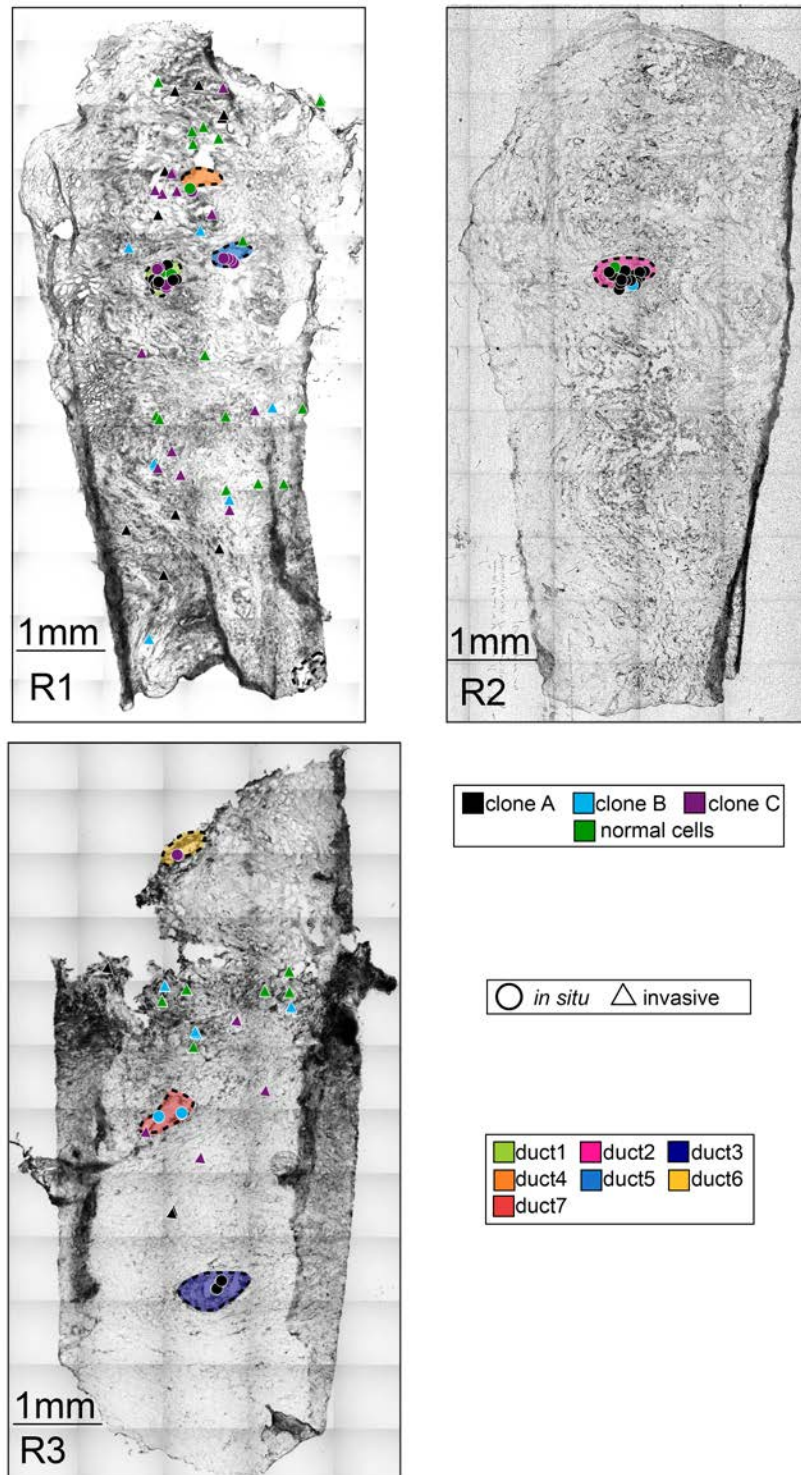
This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Clonal lineages of the major tumor subpopulations plotted with TimeScape with inferred common ancestors indicated in gray and clonal frequencies labeled.

Clone A (black), Clone B (sky-blue), and Clone C (purple).



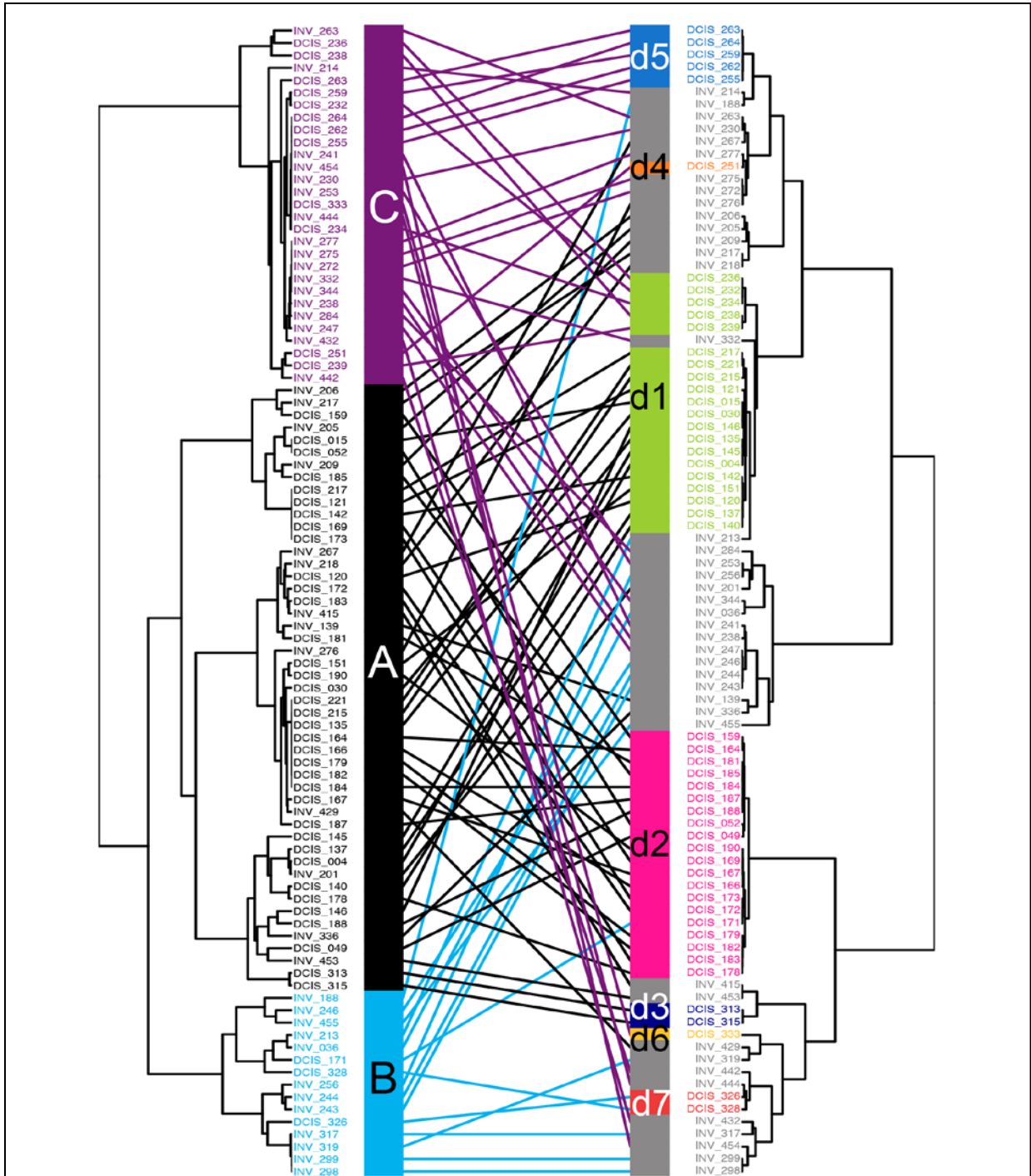
Figure 44 DC20 Image Maps



This figure was adapted from Casasent et al. 2018 <sup>1</sup> and used by permission.

Spatial maps of tissue sections from three different tumor sectors, with single cells marked as *in situ* or invasive. Tumor cells are color coded by their clonal genotypes or by diploid genomes, and ducts are annotated with different colors.

Figure 45 DC20 Tanglegram



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Genotype trees are located on the left side for each patient, with clonal subpopulations indicated by color. Spatial trees are located on the right side with different ducts indicated by colors and the invasive regions colored in gray. Mapping of cells coordinates and genotypes were performed by minimizing overlapping connections. Clone A (black), Clone B (sky-blue), and Clone C (purple).

### 3.2.2 Copy Number Evolution During Invasion Monoclonal Tumors

We observed 4 tumors that appeared to be monoclonal. These monoclonal results could represent a clonal expansion or under-sampling of highly diverse or noisy tumors. Monoclonal tumor sections will not have polyclonal-specific images.

#### 3.2.2.1 DC6

DC6 is a high-grade tumor (ER+, PR+, HER2+) from a very young patient. This tumor showed strong expansion of the ducts and separated into normal and tumor cells, with only 1 subclone being selected by k-means clustering, suggesting a possible strong clonal expansion of this clone within the ducts (Figure 49 DC6 Image Maps).

We noted clone A had a number of strong focal amplifications at 1-centromere(MCL1, SHC1), 6-centromere (FOXO3), 7p(JAZF1), 11q (PAK1), and 20q(AURKA), with lesser focal deletions of 8q (CSMD1, PPP2R2A, and FGFR1) and 17p (MAP2K4 and ERBB2/HER2) (See Figure 46 DC6 Copy Number Alteration Heatmap). Perhaps the most abnormal feature of this tumor was focal deletion of ERBB2/HER2, while the tumor classification was HER2+. Therefore, while we were expecting an alteration around the HER2 locus, we expected an amplification not a deletion (Table 2 Clinical Information).

The second peculiar feature of this tumor was, within the normal cell appears to be several clonal CNA, suggesting we might be measuring some additional subpopulations, possibly detectable if we changed our means of selecting our k for k-means clustering (Figure 46 DC6 Copy Number Alteration Heatmap), suggesting consideration of a more discrete or sensitive clustering algorithm. However, the cells are still intermixed, with both *in situ* and invasive cells being in both the pseudo-normal clone ( $N_p$ ) and the invasive clone across all four sectors sampled, supporting the multiple clonal invasion model (Figure 48 DC6 MDS and Figure 49 DC6 Image Maps). In addition, even with this result, the single clone of origin hypothesis is still supported because of the shared alterations on 1q, 6p, 7q, and 8 with the tumor subclones that we examined (Figure 46 DC6 Copy Number Alteration Heatmap).

Figure 46 DC6 Copy Number Alteration Heatmap

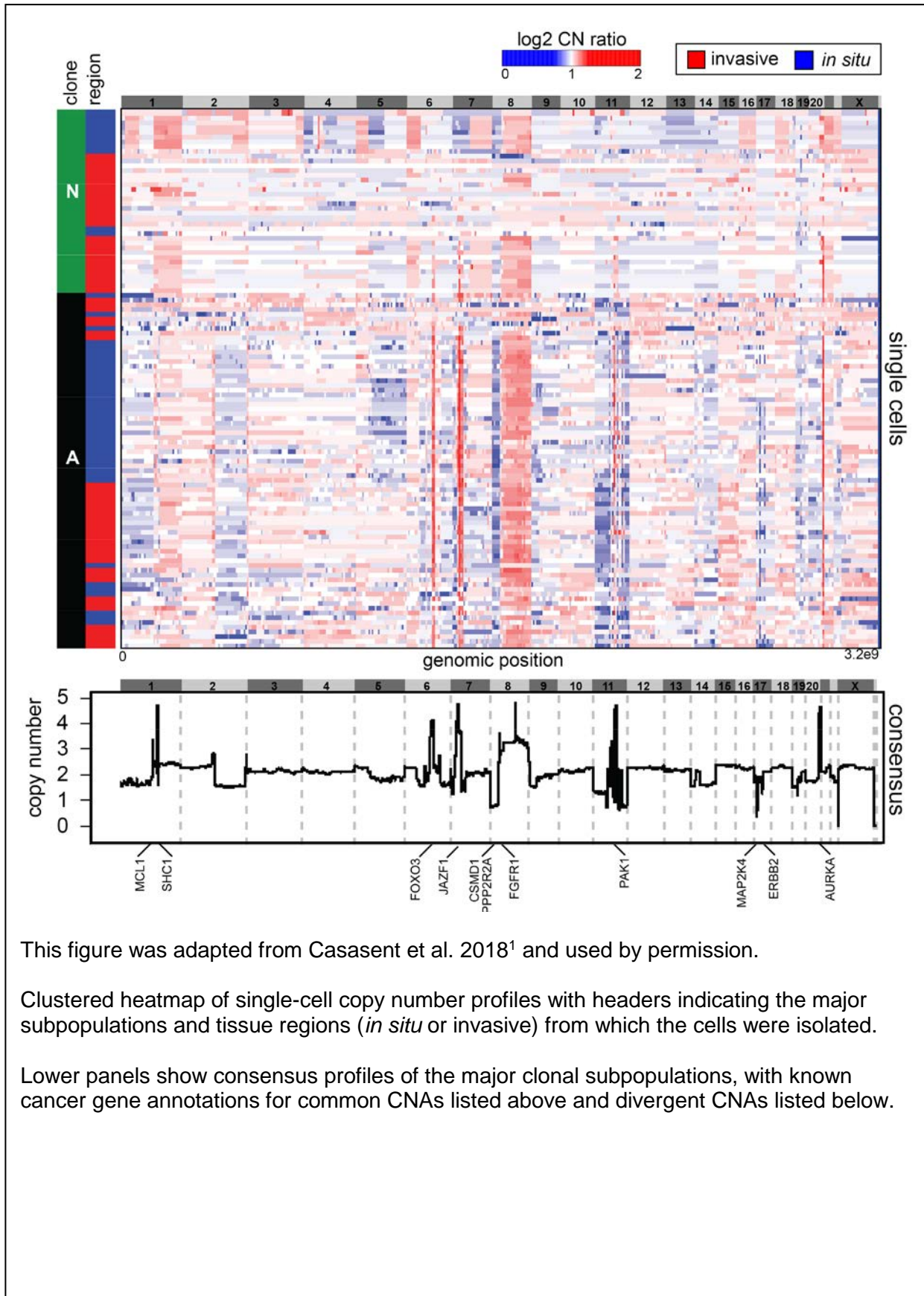




Figure 47 DC6 Saturation Curve

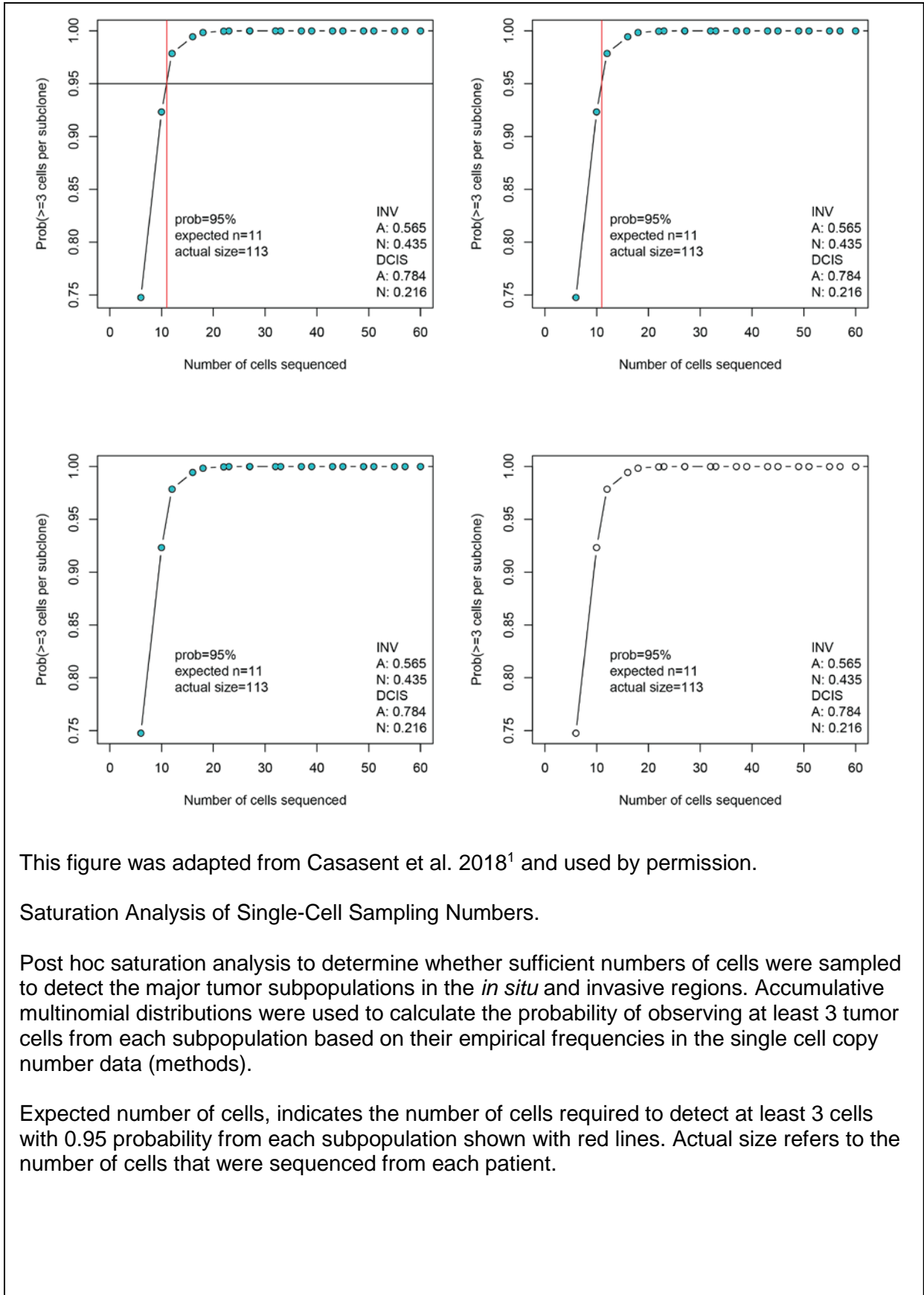


Figure 48 DC6 MDS

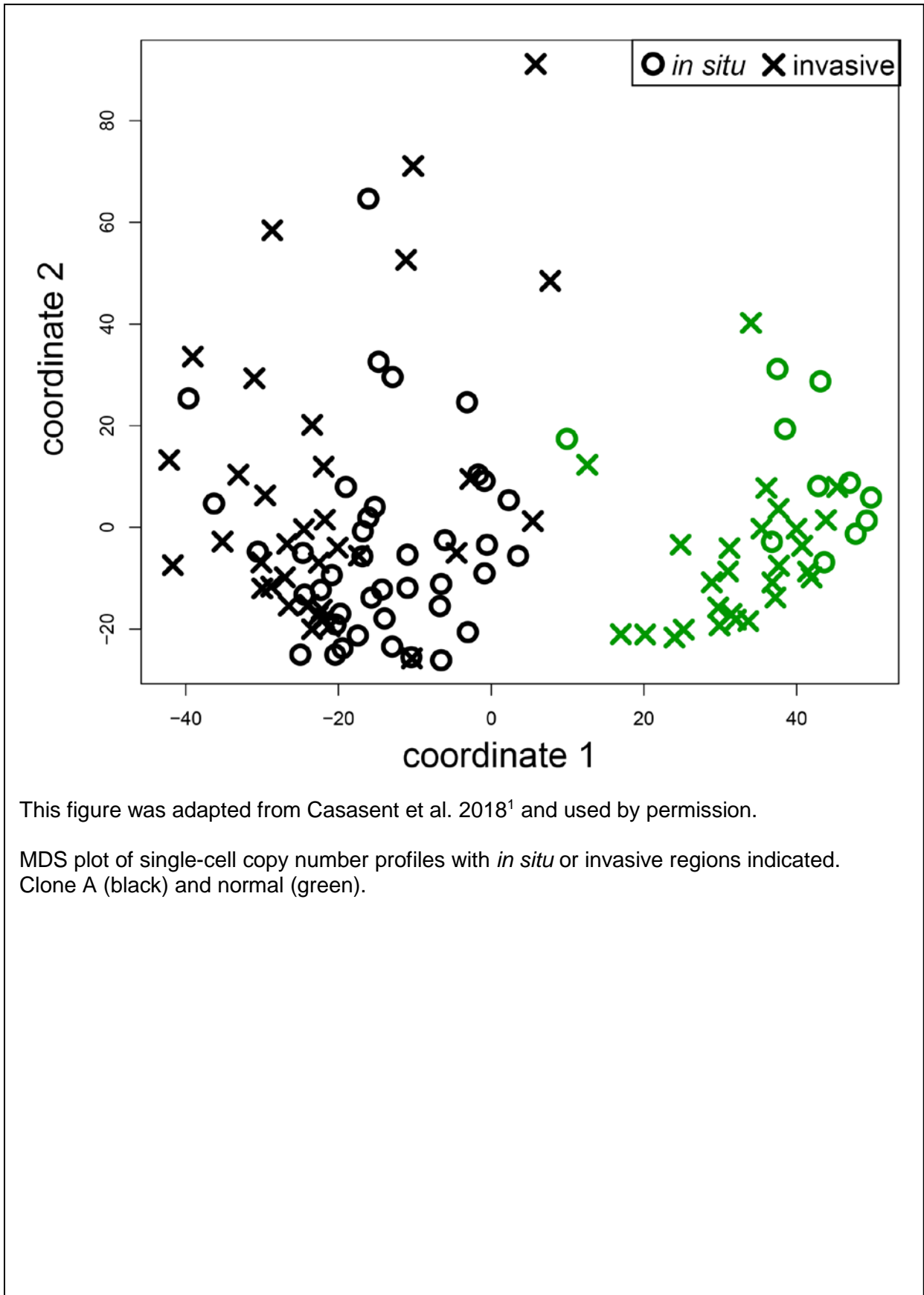
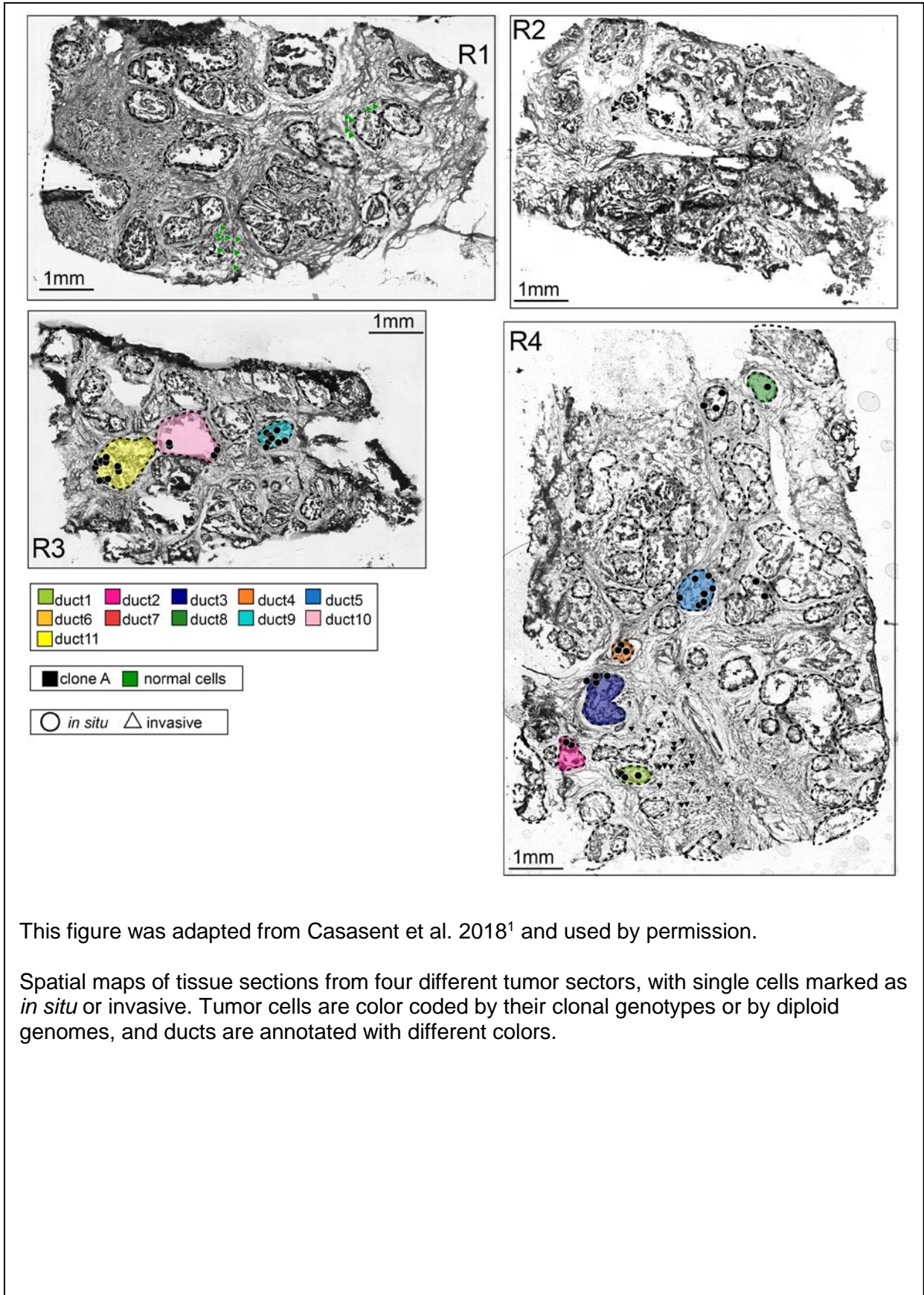


Figure 49 DC6 Image Maps



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Spatial maps of tissue sections from four different tumor sectors, with single cells marked as *in situ* or invasive. Tumor cells are color coded by their clonal genotypes or by diploid genomes, and ducts are annotated with different colors.

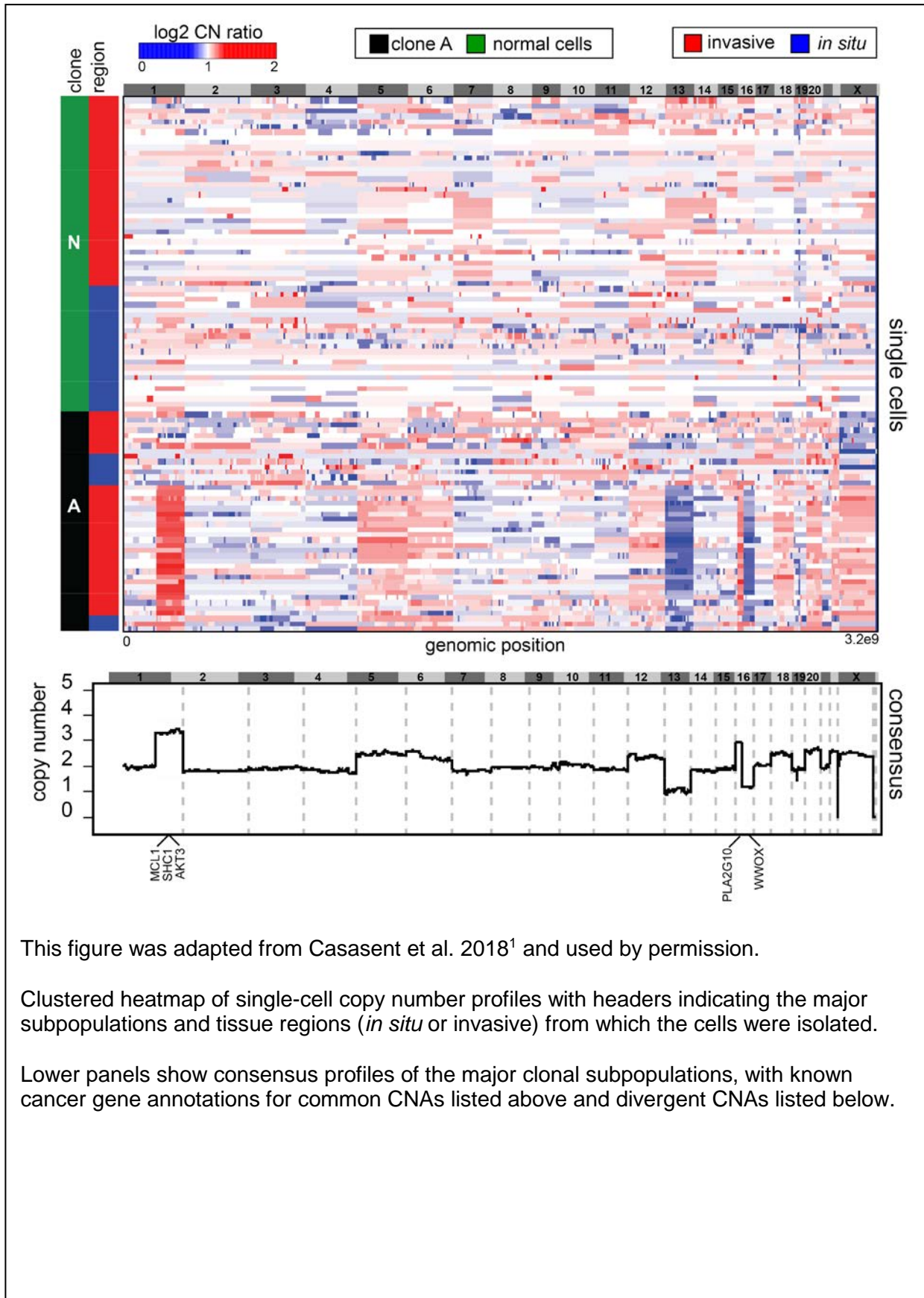
### 3.2.2.2 DC12

DC12 separated into normal and tumor cells. This tumor was perhaps the most difficult to dissect because it was grade 1. In the future, I strongly suggest that TSCS not be used on grade 1 tumors, unless another staining protocol, such as IHC, could be used to delineate normal cells from tumor cells. Grade 1 tumors have very small, more regularly shaped nuclei, making selection of tumor cells more difficult than in higher grade tumors (Figure 53 DC12 Image Maps).

Here we observed 1 clonal population (A) and one normal or noise population (N). I refer the normal population as possibly "noise" because while there are alterations, these alterations do not appear to be clonal events (Figure 50 DC12 Copy Number Alteration Heatmap). In addition, this tumor had a limited number of total cells that passed filtering: 33 *in situ* cells and 36 invasive cells, for a total of 69 cells (Figure 51 DC12 Saturation Curve). Most of the cells were removed during filtering, suggesting lower quality sample, possibly due to DNA degradation during freeze thaw cycles.

Even within clone A, there appeared to be 2 subclones  $A_N$  (which was more normal like) and  $A_T$  (which was more aneuploid) (Figure 50 DC12 Copy Number Alteration Heatmap and Figure 52 DC12 MDS). The  $A_N$  subclone appeared to very close to the pseudo-normal cells observed in DC16, with a strong deletion of chromosome X and few other common alterations across cells. While  $A_T$  did not share this deletion of chromosome X, it instead appeared to have an amplification of X. In addition to the amplification of chromosome X,  $A_T$  had a number of clonal amplification in 1q (MCL1, SHC1, and AKT3), 5, 6, 12, 16p (PLAG10), and 18, as well as deletions in 13, 16q(WWOX), 19, and 21. However, most of the alteration were very large whole chromosome amplifications or deletions, which fits with the overall copy number profiles observed in ER+, PR+ tumors like DC12 (Figure 50 DC12 Copy Number Alteration Heatmap).

Figure 50 DC12 Copy Number Alteration Heatmap



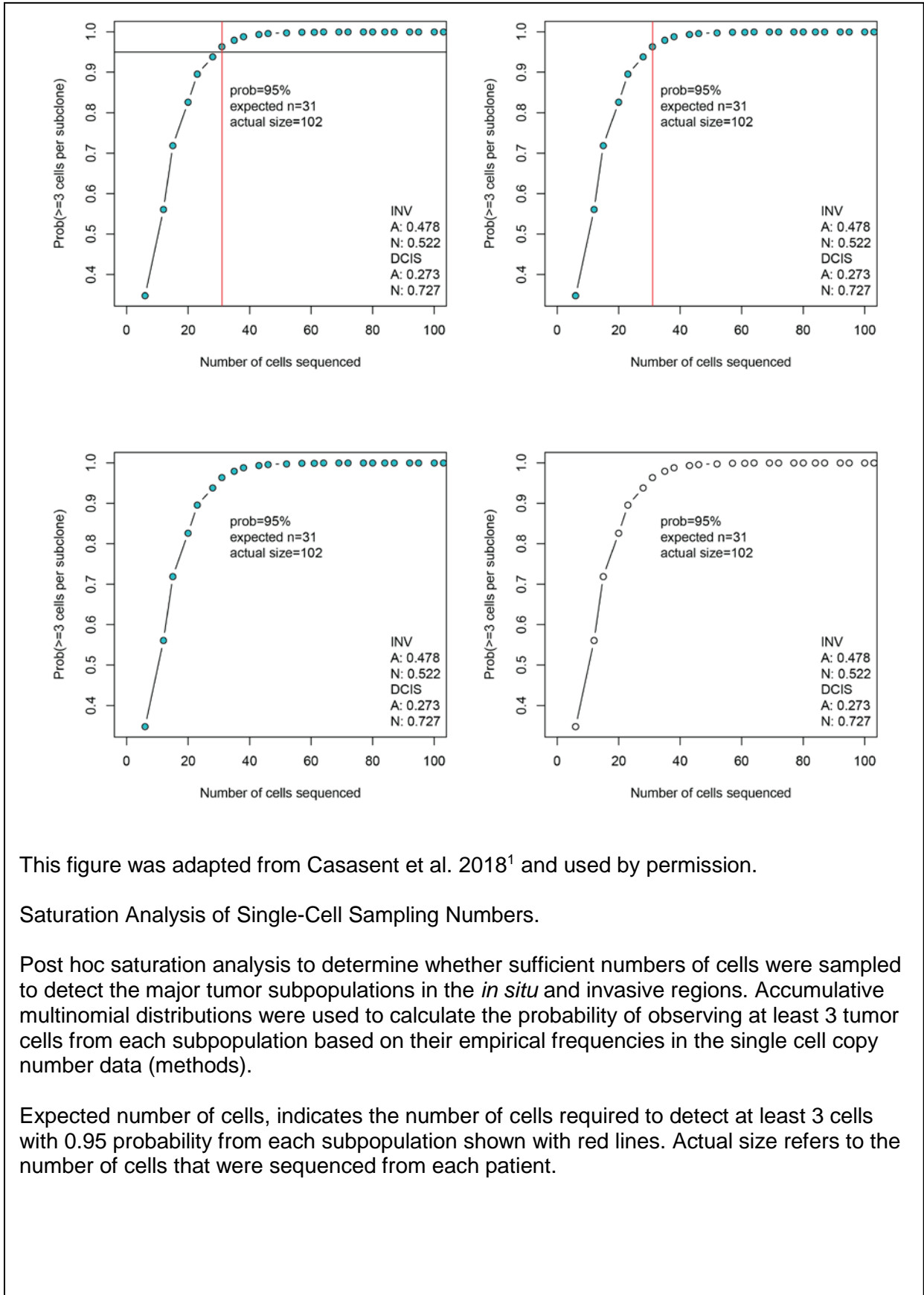
This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Clustered heatmap of single-cell copy number profiles with headers indicating the major subpopulations and tissue regions (*in situ* or invasive) from which the cells were isolated.

Lower panels show consensus profiles of the major clonal subpopulations, with known cancer gene annotations for common CNAs listed above and divergent CNAs listed below.



Figure 51 DC12 Saturation Curve



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Saturation Analysis of Single-Cell Sampling Numbers.

Post hoc saturation analysis to determine whether sufficient numbers of cells were sampled to detect the major tumor subpopulations in the *in situ* and invasive regions. Accumulative multinomial distributions were used to calculate the probability of observing at least 3 tumor cells from each subpopulation based on their empirical frequencies in the single cell copy number data (methods).

Expected number of cells, indicates the number of cells required to detect at least 3 cells with 0.95 probability from each subpopulation shown with red lines. Actual size refers to the number of cells that were sequenced from each patient.

Figure 52 DC12 MDS

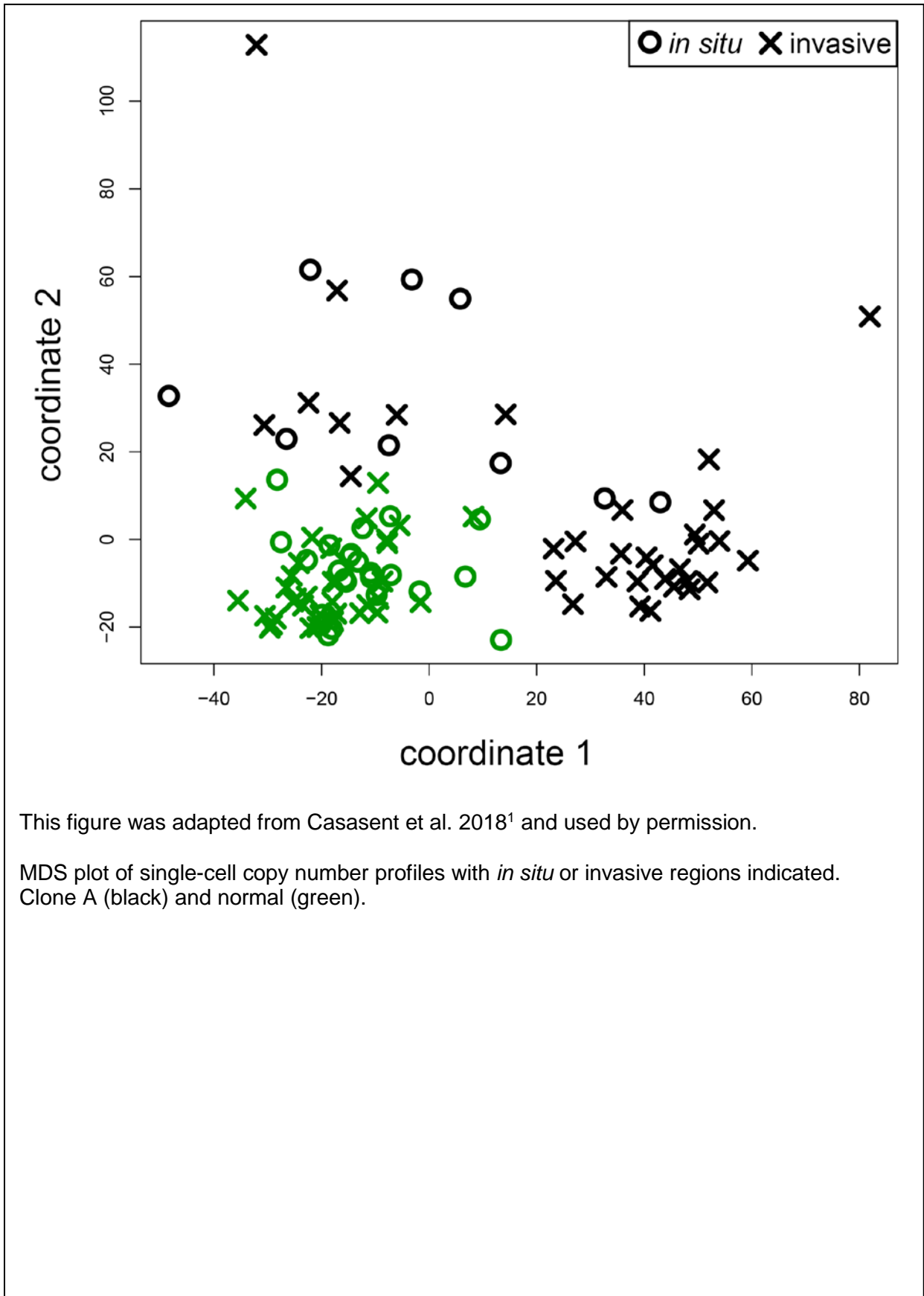
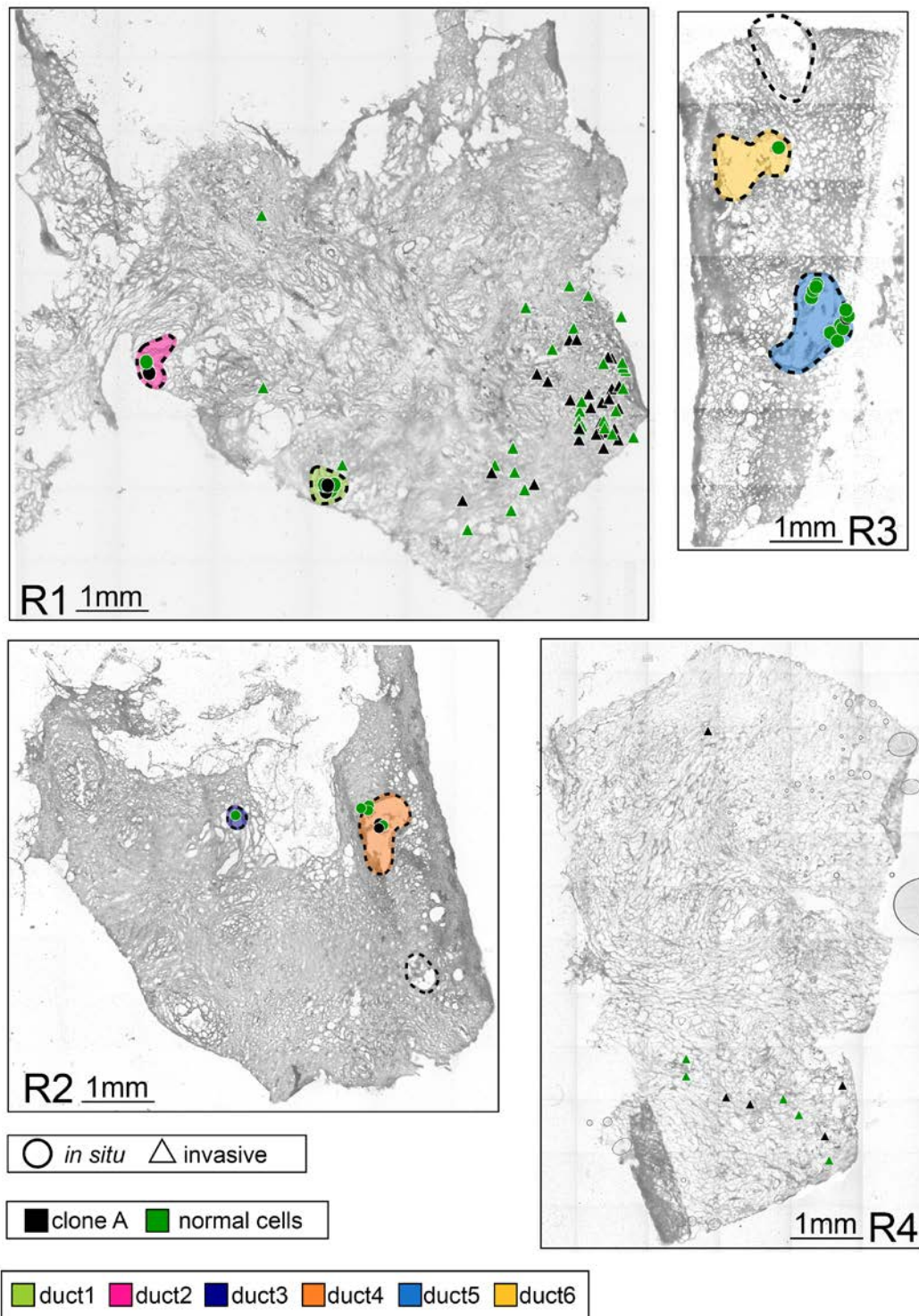


Figure 53 DC12 Image Maps



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Spatial maps of tissue sections from four different tumor vials, with single cells marked as *in situ* or invasive. Tumor cells are color coded by their clonal genotypes or by diploid genomes, and ducts are annotated with different colors.



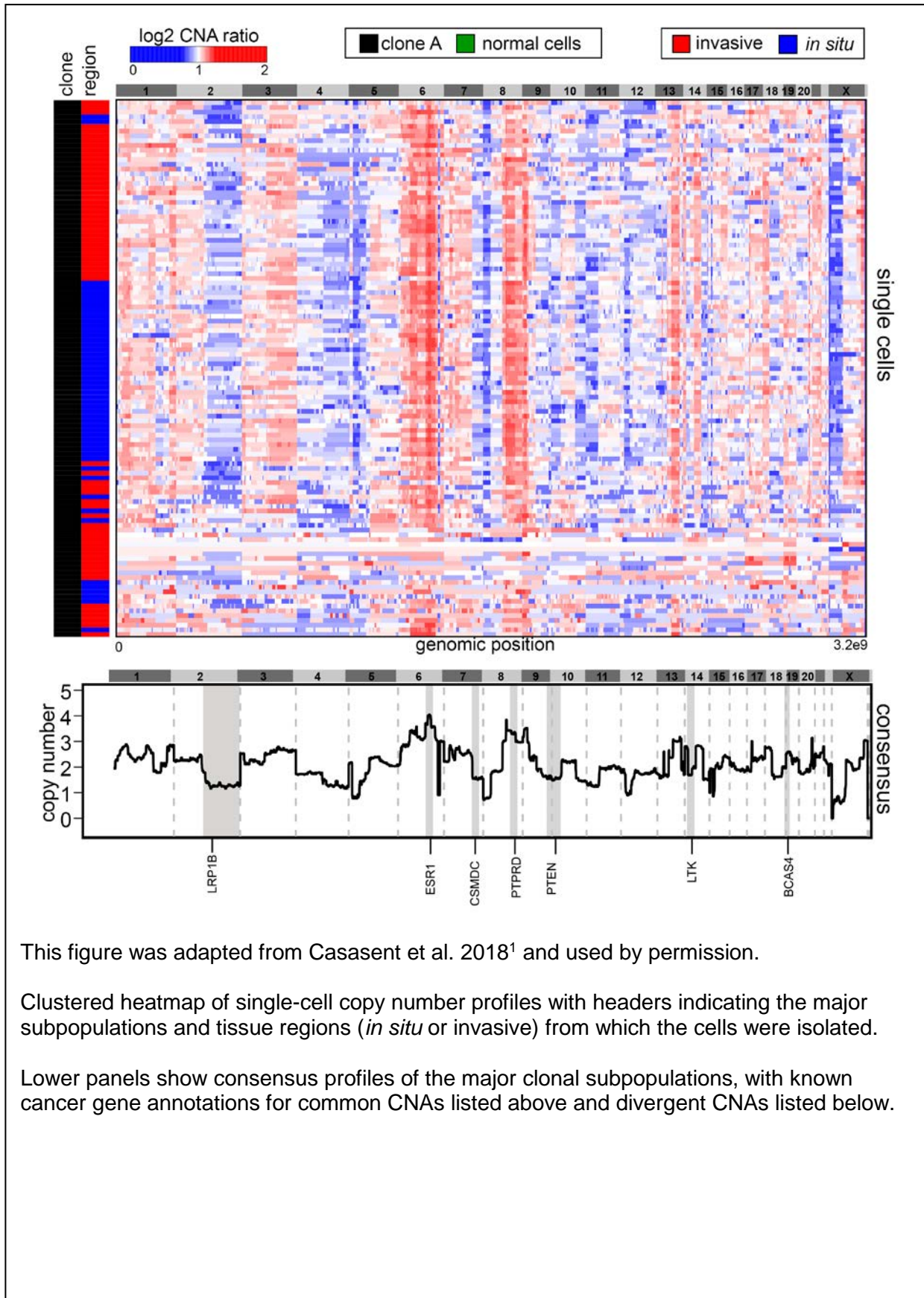
### 3.2.2.3 DC17

DC17 is an odd case. In DC17, we had a change in the receptor status between the *in situ* and invasive regions. The *in situ* regions were TNBC, while the invasion region was ER positive. This type of change in receptor status was previously considered evidence for the independent lineage model. However, this tumor, while following the zig-zag pattern for TNBC, contained a clonal amplification of ESR1. The focal amplification of ESR1 was a little strange in this TNBC tumor. This amplification occurred in both the *in situ* and invasive cells and might suggest a phenotype change turning the expression of ESR1 on and off in the invasive cells (See Figure 54 DC17 Copy Number Alteration Heatmap and Table 2 Clinical Information).

In the DC17 sample, we were unable to separate the normal cells from the invasive cells based on our k-means clustering algorithm. When examining the profiles, very few of cells appeared to have a normal like profile in DC17 (Figure 54 DC17 Copy Number Alteration Heatmap). Clone A was the only clone observed by k-means clustering and appears to be highly variable. While chromosome breakpoints did not appear to be as stable in this clone, the overall pattern of amplifications and deletions was consistent, suggesting the tumor might have been frozen and thawed too many times (Figure 54 DC17 Copy Number Alteration Heatmap). The supposition was also supported because the quality of the DNA appears to be lower, another side-effect of too many freeze-thaw cycles.

We see consistent amplifications on chromosome 3q, 6 (ESR1), 8q (PTPRD), and 13, as well as deletions on 2q(LRP1B), 4, 5p, 7q (CSMD1), 10p, 11p, 12p and Xp (Figure 54 DC17 Copy Number Alteration Heatmap). However, while these alterations appeared to be present in most of the cells, the breakpoints were not consistent and neither were the strength of the amplifications and deletions, resulting in these cells possibly erroneously being classified as one clone. Because only 1 clone was observed and no normal cells, we were unable to provide a saturation curve because the calculation requires at least two clones.

Figure 54 DC17 Copy Number Alteration Heatmap

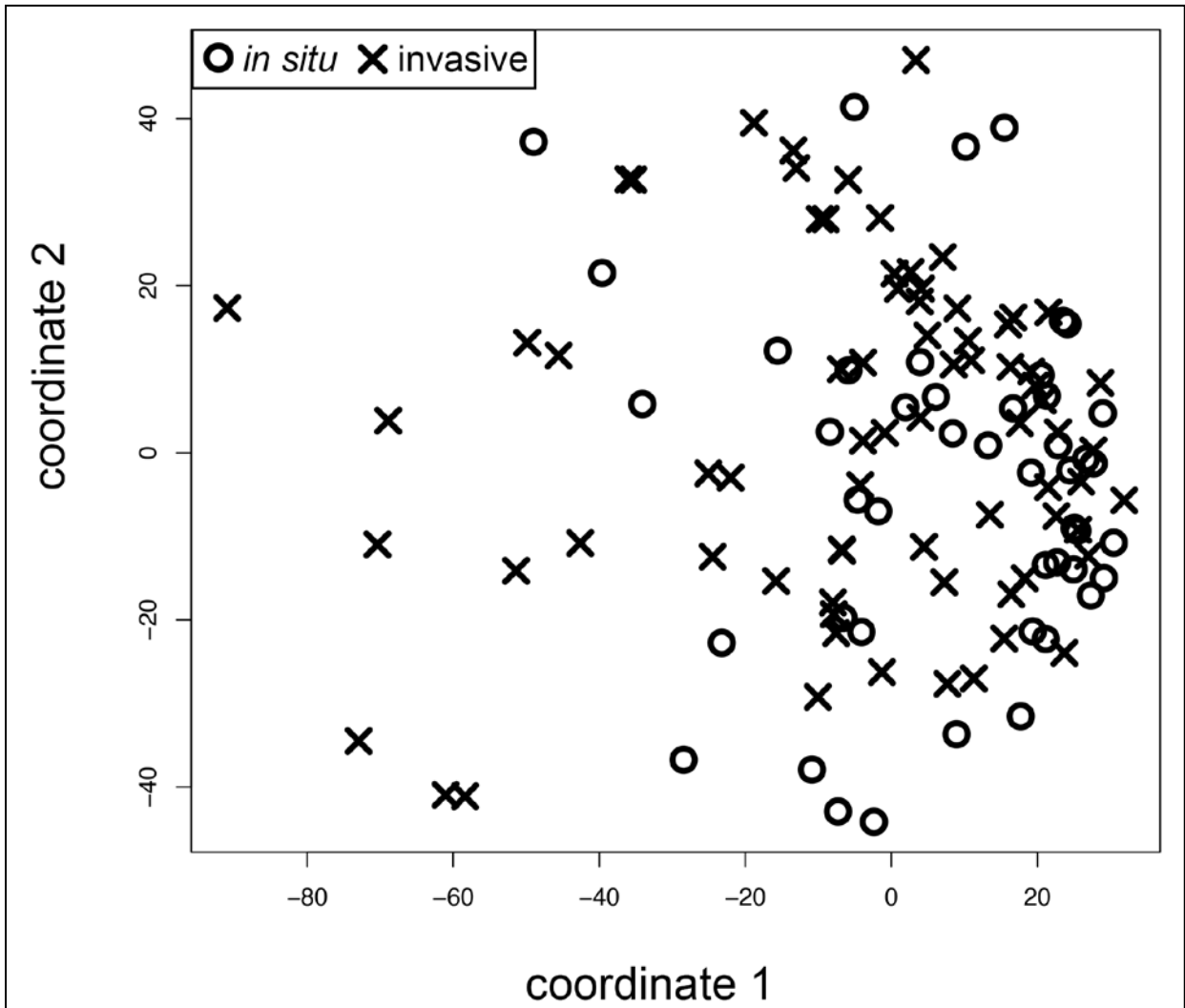


This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Clustered heatmap of single-cell copy number profiles with headers indicating the major subpopulations and tissue regions (*in situ* or invasive) from which the cells were isolated.

Lower panels show consensus profiles of the major clonal subpopulations, with known cancer gene annotations for common CNAs listed above and divergent CNAs listed below.

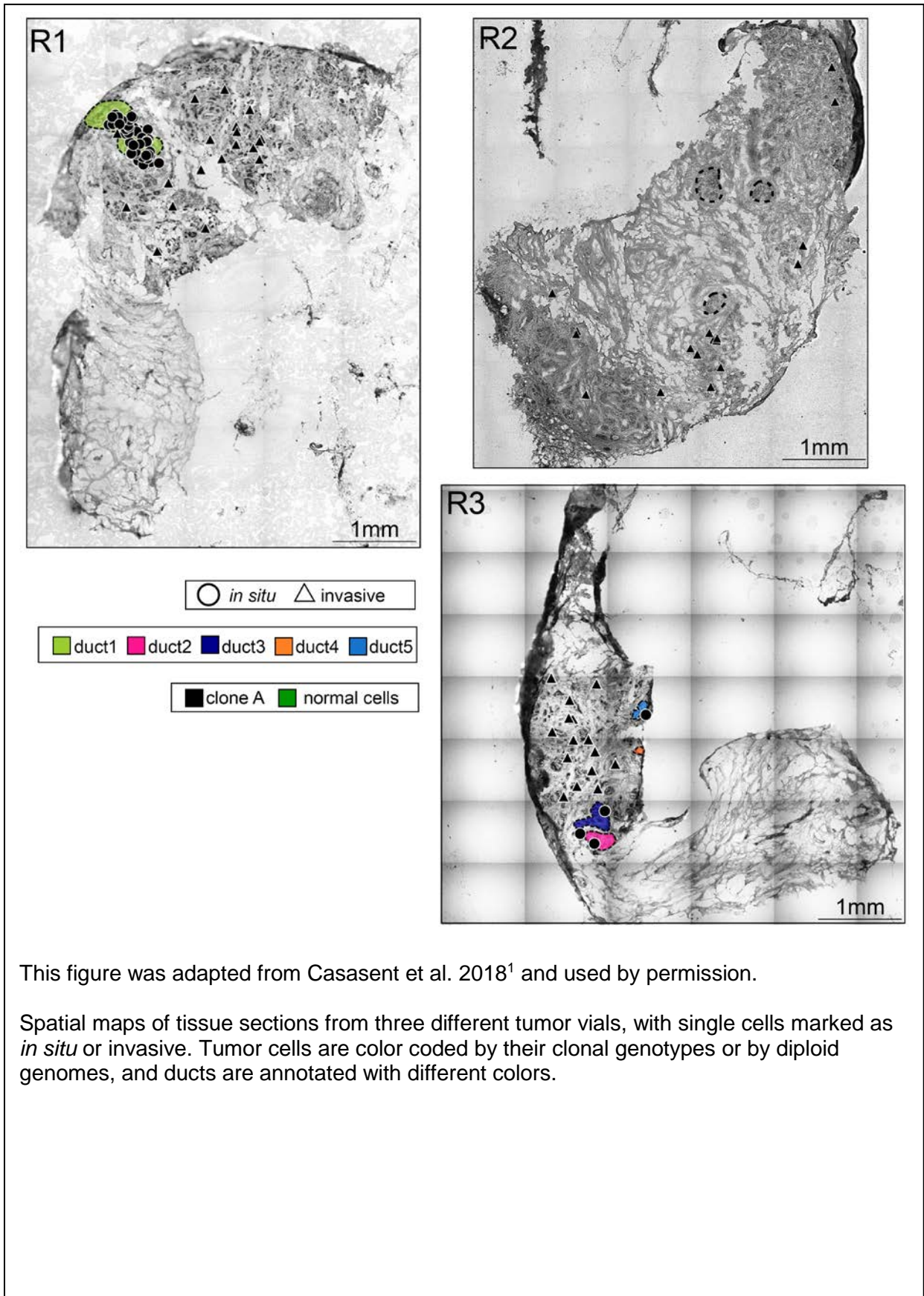
Figure 55 DC17 MDS



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

MDS plot of single-cell copy number profiles with *in situ* or invasive regions indicated.

Figure 56 DC17 Image Maps



#### 3.2.2.4 DC19

DC19, a grade 3 TNBC sample, was classified as monoclonal but could be separated into normal and tumor cells. Like DC17, the chromosome breakpoints and strength of amplifications and deletions are inconsistent, which possibly resulted in the tumor classification of only 2 clones, normal and clone A (Figure 57 DC19 Copy Number Alteration Heatmap). With tumor clone A, visually there appeared to be a least 4 pseudo-subclones ( $A_N$ ,  $A_1$ ,  $A_2$ ,  $A_3$ , going from top to bottom of the single cell heatmap). These pseudo-subclones are visually distinct but have not been distinguished mathematically. Pseudo-subclone  $A_N$  is the noisiest of these profiles, with very few consistent amplification and deletions, suggesting noise, possibly from the tumor being frozen and thawed too many times or the necrosis we observed in the center of the ducts (Figure 60 DC19 Image Maps). Pseudo-subclone  $A_1$ , was distinguished by the deletion of chromosome 4 and X, as well as the amplifications of chromosome 5, 7, and 8. Pseudo-subclone  $A_2$  is almost opposite of  $A_1$ , having a deletion in 5q and amplification in X. Lastly, Pseudo-subclone  $A_3$  has the amplification of 2p, 6, 7, 8, and 9 (Figure 57 DC19 Copy Number Alteration Heatmap).

The inconsistency in DC19 resulted in a flat profile, except for amplifications on 1-centermere (MCL1 and SHC1), 2p, 5p(MYO10 and ANKH), 16q (WWOX), and 18 (SMAD4), as well as focal deletions on 8p (PPP2R2A) and X. We visually observed the pseudo-subclone and the major clone in DC19 were intermixed between *in situ* and invasive regions. The inconsistency suggests one of the following: (1) the tumor might have been frozen and thawed too many times, (2) the necrosis or apoptosis of the cells resulted in no-single cell DNA being collected, or (3) the tumor is highly diversity.



Figure 57 DC19 Copy Number Alteration Heatmap

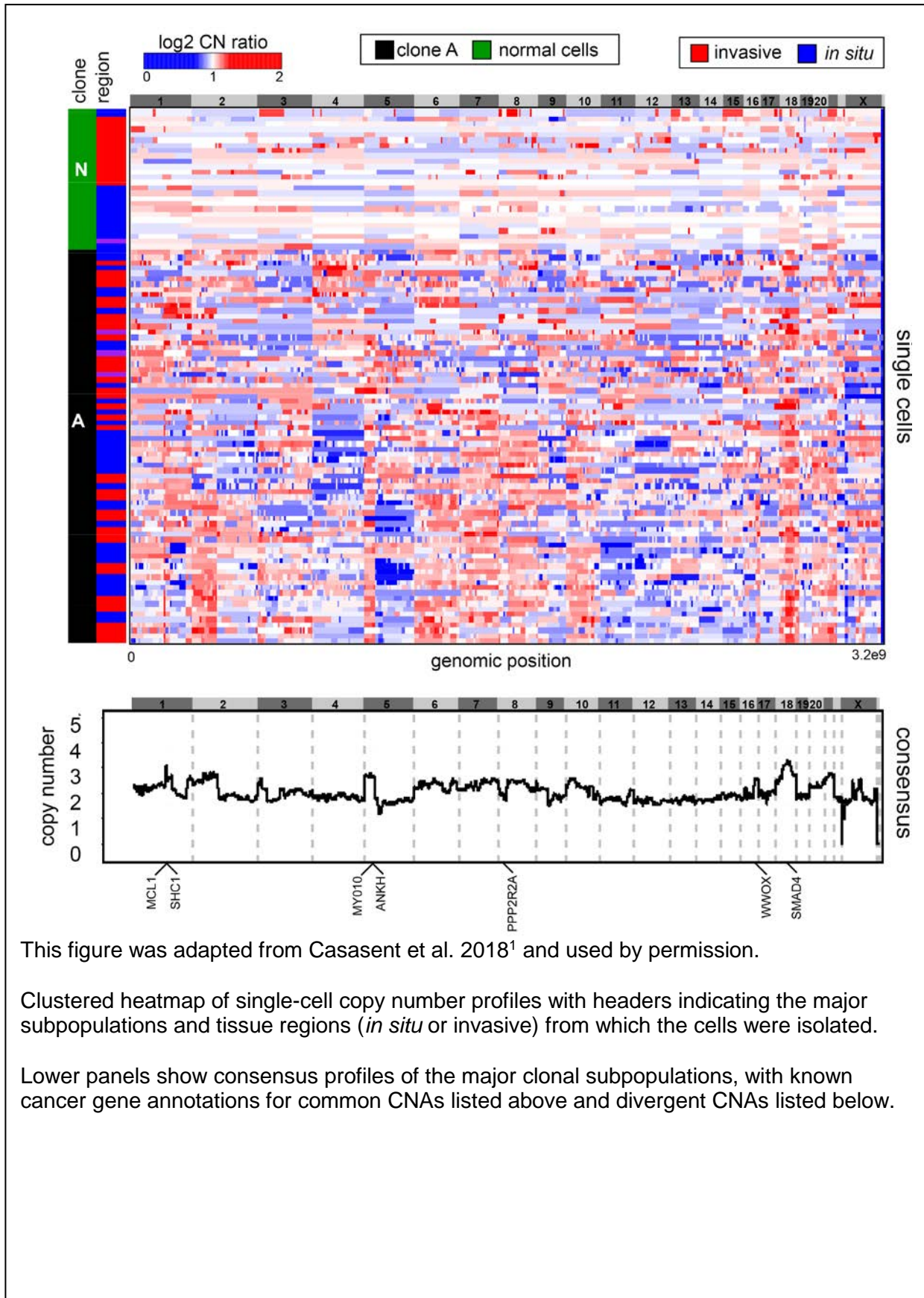


Figure 58 DC19 Saturation Curve

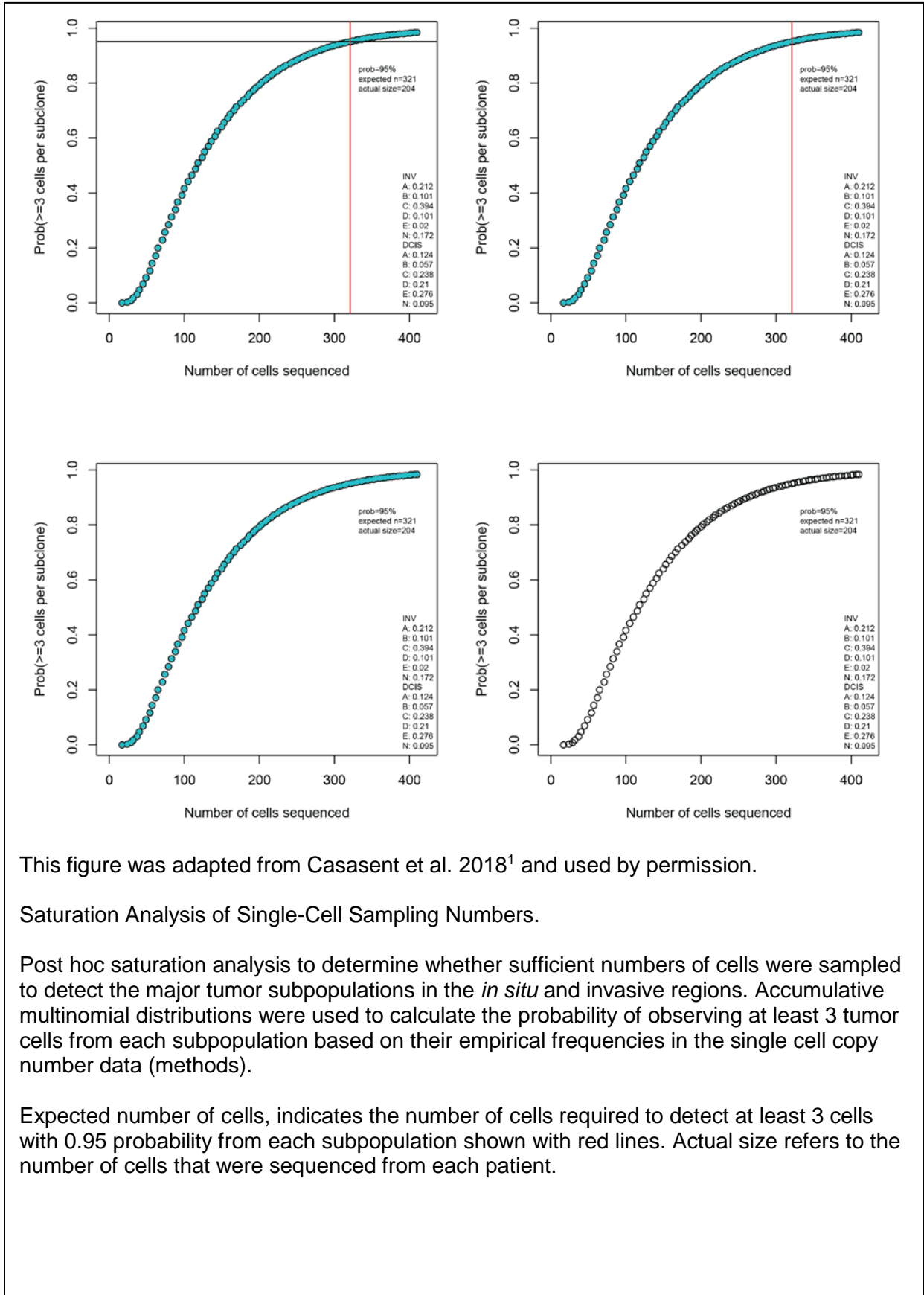


Figure 59 DC19 MDS

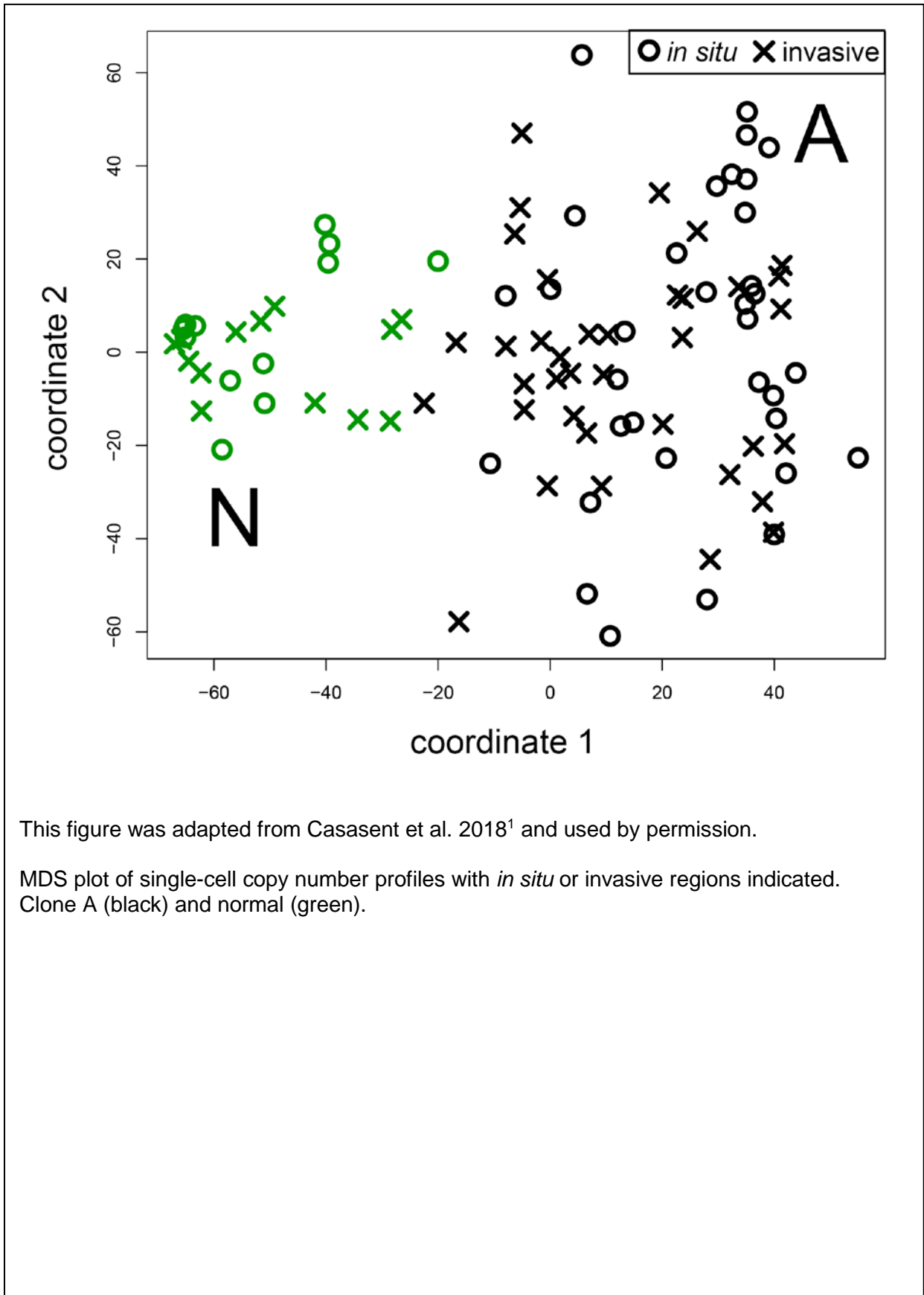
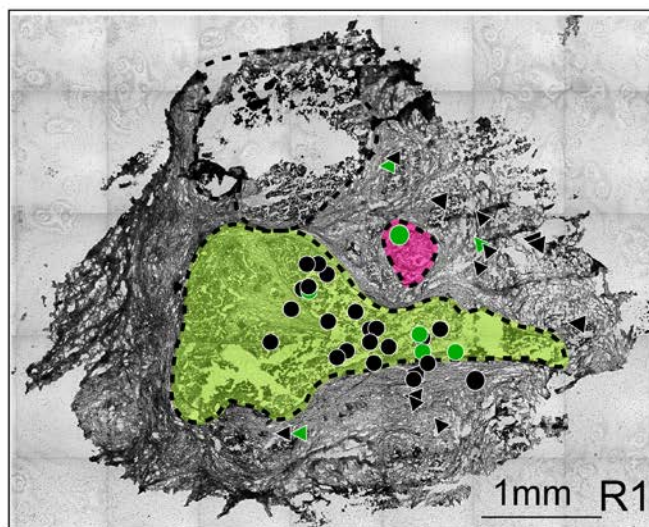
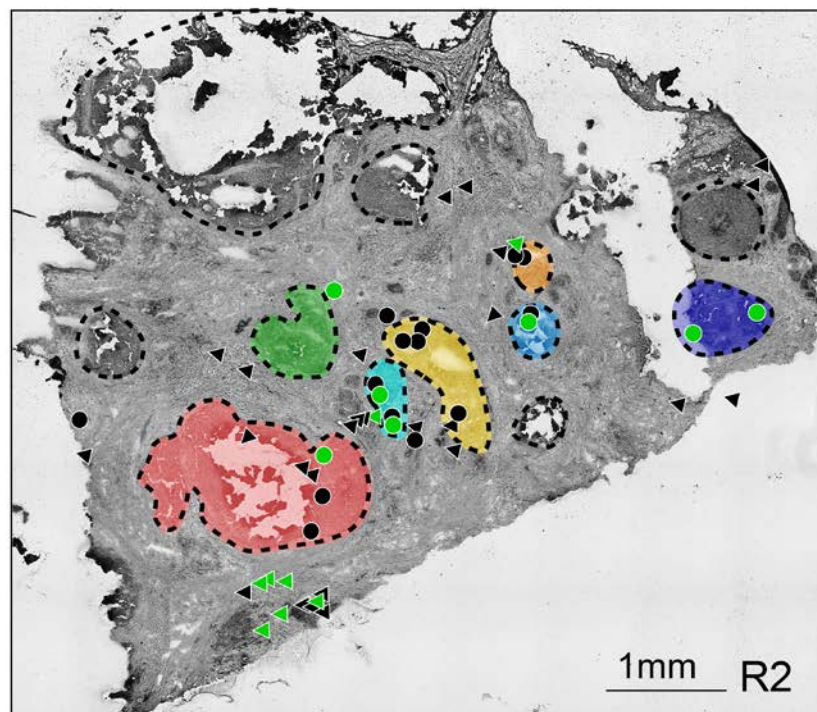




Figure 60 DC19 Image Maps



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

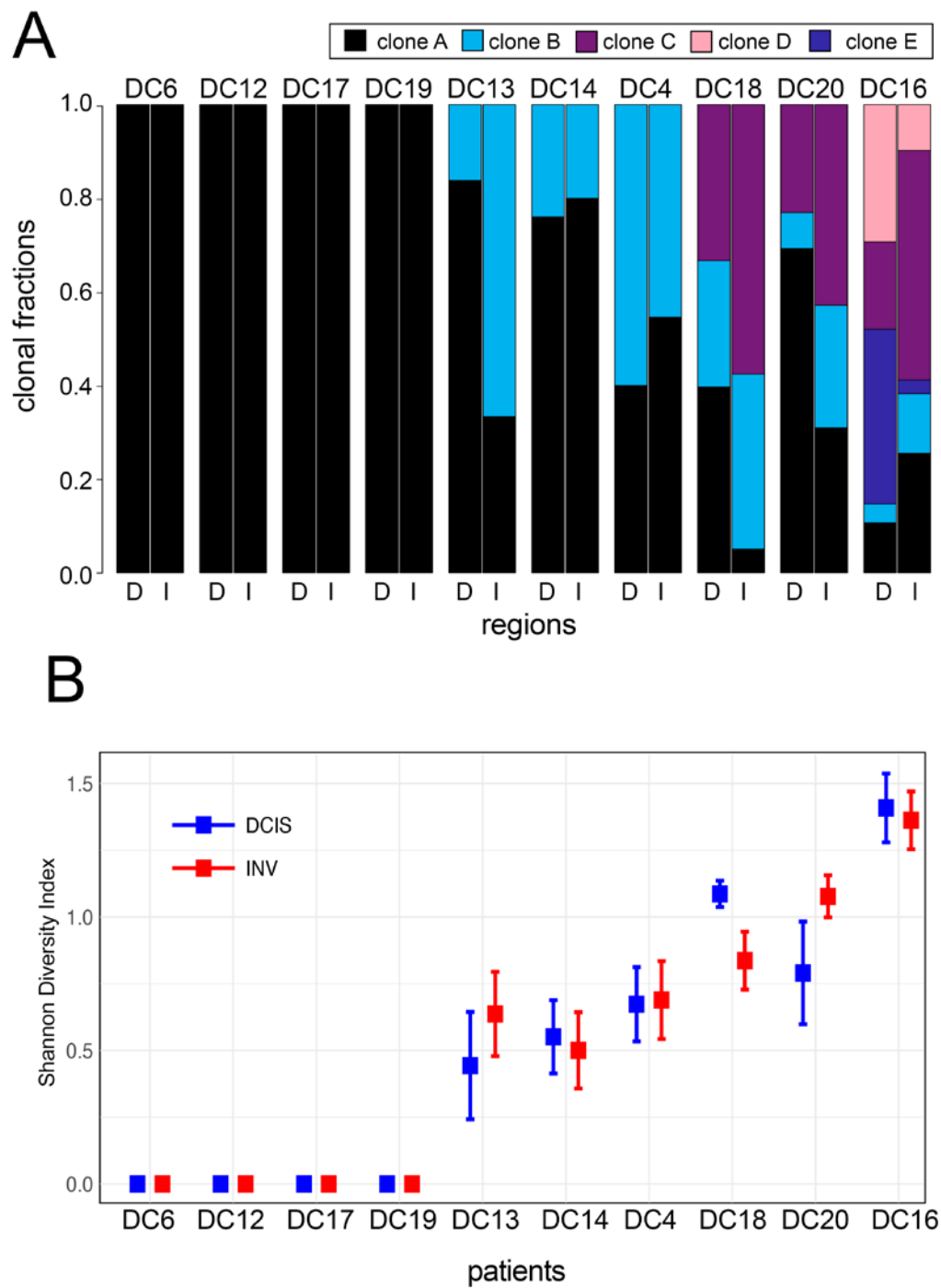
Spatial maps of tissue sections from two different tumor regions, with single cells marked as *in situ* or invasive. Tumor cells are color coded by their clonal genotypes or by diploid genomes, and ducts are annotated with different colors.

### 3.2.3 Copy Number Evolution Summary

We applied TSCS to a total of 10 synchronous DCIS-IDC patients to study copy number evolution during invasion (Figure 61 Copy Number Summary). Whole-tissue scanning of H&E tissue sections from each patient was performed to identify *in situ* and invasive regions for single cell isolation. In total, 425 *in situ* and 503 invasive cells were sequenced from the 10 patients, as well as 365 stromal diploid cells. The data was analyzed to delineate clonal substructure and copy number evolution during invasion. Clustering of single cell CNA profiles showed most patients harbored 1-5 major tumor subpopulations and these subpopulations were located in both *in situ* and invasive regions.

We found four tumors to be monoclonal (DC6,12,17,19). However, this could be due to inconsistencies in breakpoints resulting in under-clustering. Six tumors were polyclonal (DC4, 13,14, 16,18, 20), harboring multiple clonal subclones in both the *in situ* and invasive regions (Copy Number Evolution During Invasion Polyclonal Tumors). Shannon Diversity indexes calculated from the single cell CNA profiles showed the amount of clonal diversity did not show major changes during invasion in most patients (Figure 61 Copy Number Summary). These data showed the amount of genomic diversity correlated with the number of subpopulations detected in the *in situ* or invasive regions and was inconsistent with a population bottleneck, in which a decreased in clonal diversity is expected (due to the selection of a specific clonal genotype). MDS analysis of all 10 DCIS patients identified 1-6 major clusters in each patient, including the normal cells (N) and 1-5 major tumor subpopulations (A-E), often separated in high-dimensional space (see MDS plots for individual tumors). Moreover, the MDS plots showed that within each genotype cluster, tumor cells were localized to both *in situ* and invasive regions.

Figure 61 Copy Number Summary



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Copy Number Substructure and Clonal Evolution in 10 DCIS Patients

(A) Bar plots of clonal frequencies calculated from single-cell copy number profiles in the *in situ* (labeled D) or invasive (labeled I) regions. (B) Shannon diversity indexes calculated from single-cell copy number profiles from the *in situ* and invasive regions of each patient with confidence intervals

Clonal lineages were inferred in the 6 polyclonal DCIS patients and plotted with TimeScape<sup>227</sup> (see TimeScape figures for individual tumors). These data showed in all patients, the subpopulations shared a common evolutionary origin with shared truncal CNAs, suggesting the tumors evolved from a single cell in the duct. These data are inconsistent with an independent lineage model, in which different initiating cells give rise to the *in situ* and invasive subpopulations separately.

In every patient, we found that the same clonal subpopulations present in the ducts and invasive regions. However, we did observe shifts in clonal frequencies in some patients (DC13, DC16, DC18), suggesting some genotypes may be more invasive than others. For example, in DC13, clone B increased from 16% to 67% during invasion, while in DC16, clone C increased from 19% to 49%. This change suggests genome evolution initiated from a single cell in the ducts and gave rise to one or more clonal subpopulations that migrated into the adjacent tissues to establish the invasive tumor mass.

### **3.2.4 Spatial Topography and Clonal Copy Number Genotypes**

To understand the distribution of clonal genotypes and their spatial organization in the polyclonal tumors, we constructed tanglegrams<sup>213</sup>. We calculated genetic distance trees from single cell copy number profiles and mapped to spatial trees (X, Y coordinates) with minimal overlapping connections (see Tanglegrams for individual tumors). In patient DC13, clone A (81.5%) localized mainly to the ducts, with only a few cells (N=7) in the invasive regions, while clone B showed a higher frequency in the invasive regions. In patient DC14, the two major clones (A, B) mapped to all three ducts and the invasive regions; however, clone B was restricted more to ducts 2 and 3. In patient DC16, we identified 5 clonal subpopulations, in which clones A, B, and C mapped more frequently to the invasive regions, while clones D and E were found mainly in the ductal regions (ducts 1, 2, and 5). In patient DC18, we identified 3 clonal subpopulations, in which clones B and C each mapped to 8 of the 10 ducts, while clone A localized mainly to two ducts (d1 and d2). In other cases (DC20 and DC4), we found the

clones equally distributed to the *in situ* and invasive regions. These data show while all clones were detected in both the *in situ* and invasive regions, specific subclones were more restricted to the ducts, while others were more prevalent in the invasive regions, suggesting a more invasive or migratory phenotype.

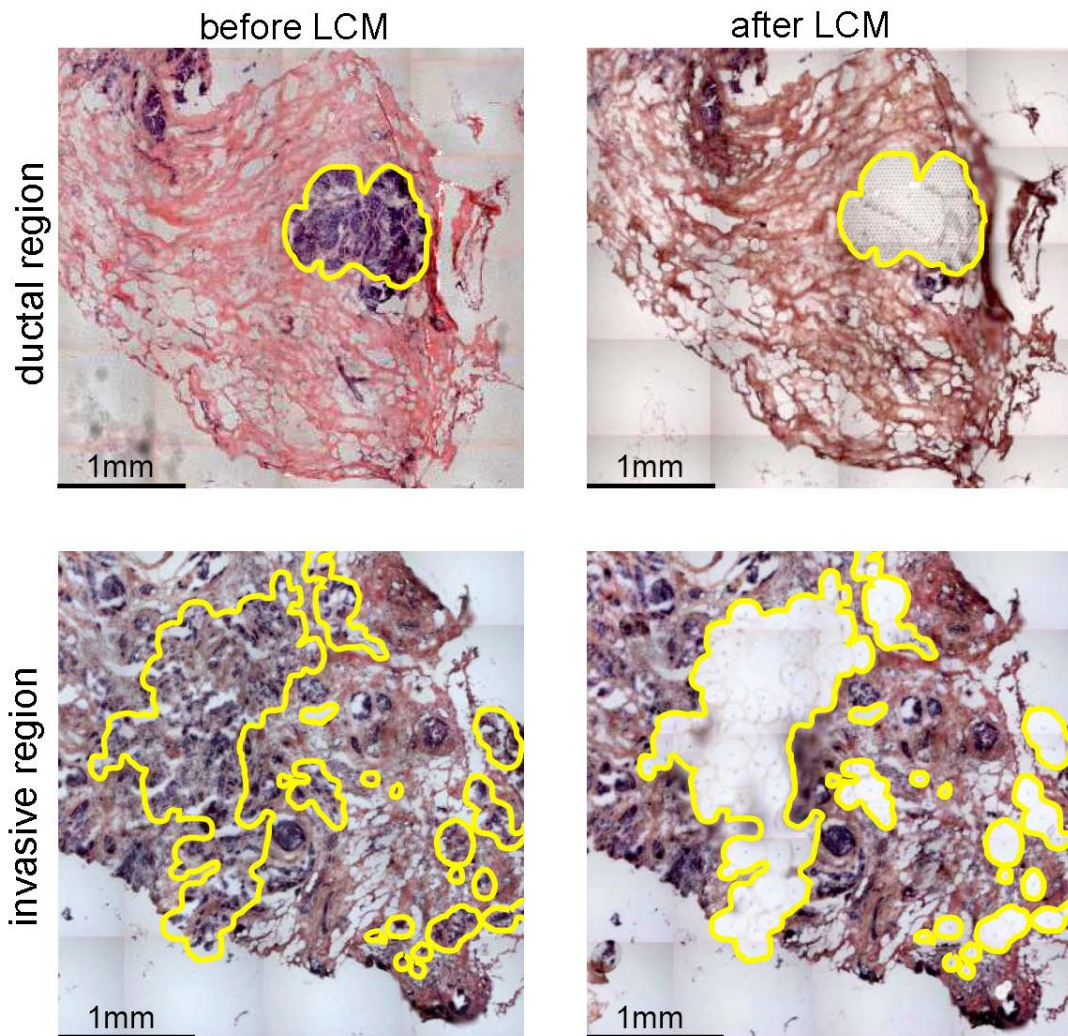
### **3.2.5 Regional Exome**

In this Section I discuss results for the Regional Exome Sequencing for our 10 Synchronous DCIS-IDC.

#### **3.2.5.1 Mutational Evolution During Invasion**

To investigate mutational evolution during invasion, we used LCM to microdissect thousands of tumor cells from the *in situ* and invasive regions for deep-exome sequencing (mean=162.8X, SEM=18.9, Figure 62 Regional Microdissection). Matched normal breast tissue (mean=144.1X, SEM=20.3) was sequenced in parallel to distinguish germline variants from somatic mutations. From this data we detected point mutations showing the total number of exonic mutations (mean=23, SEM=3.3) were highly consistent between the *in situ* and invasive regions (t-test, p=0.868) (Figure 63 Regional Exome Oncomap). To identify specific discordant mutations, we constructed oncomaps using nonsynonymous mutations (Figure 63 Regional Exome Oncomap). Most nonsynonymous mutations (mean 87.4%) were concordant in the ducts and invasive regions, including mutations in known breast cancer genes such as *TP53*, *PIK3CA*, *NCOA2*, *ABL2*, *PDE4DIP*, *AHNAK*, and *RUNX1*, suggesting they were acquired in the ducts prior to invasion. However, a few mutations were *in situ*-specific (N=12) or invasive-specific (N=11) in 4 patients (DC12, DC13, DC17, DC18) and were not recurrent among the patients (Table 4 Regional Invasive-Specific Mutations).

Figure 62 Regional Microdissection

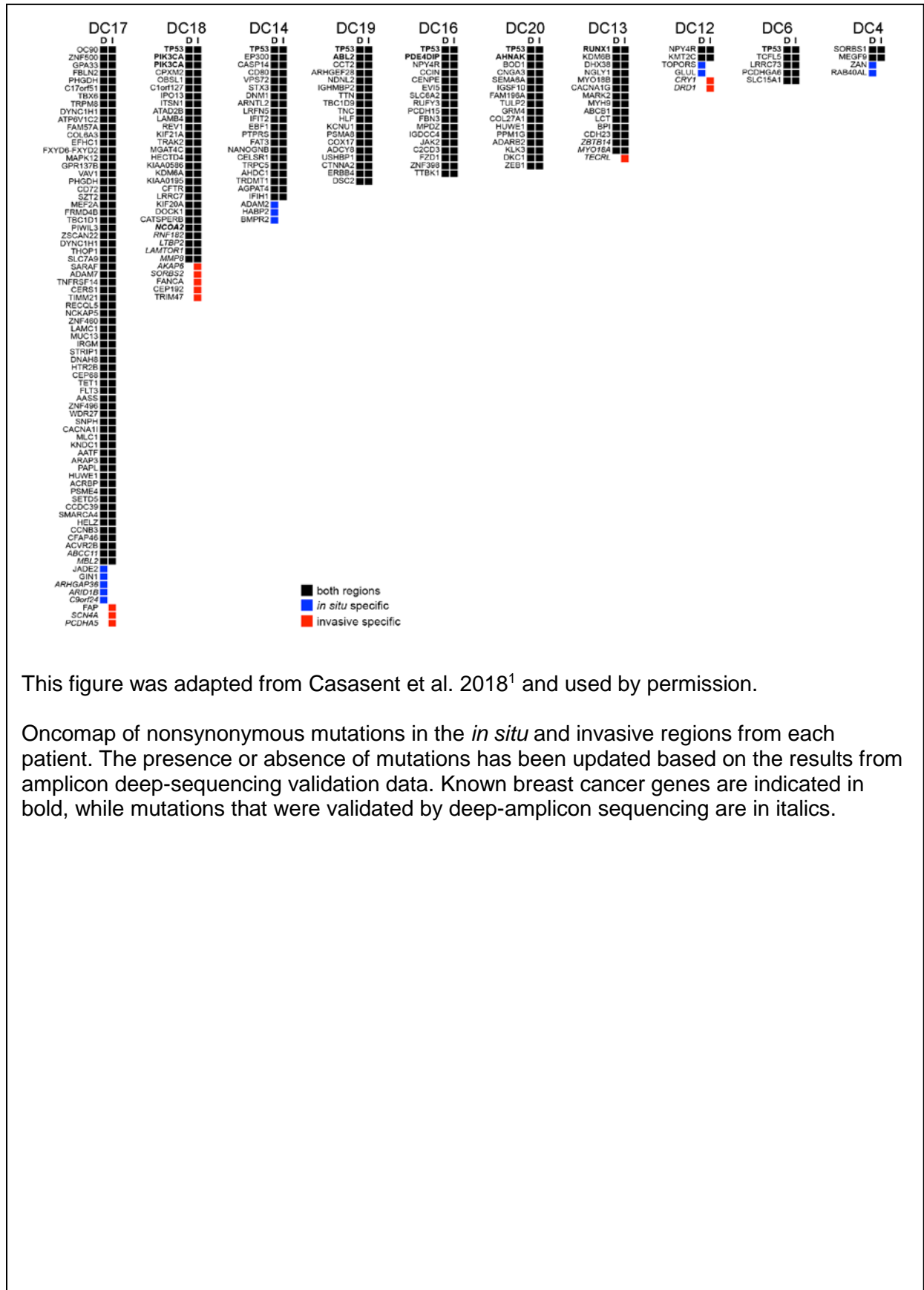


This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Exome sequencing of laser-capture microdissected *in situ* (top) and invasive (bottom) regions.



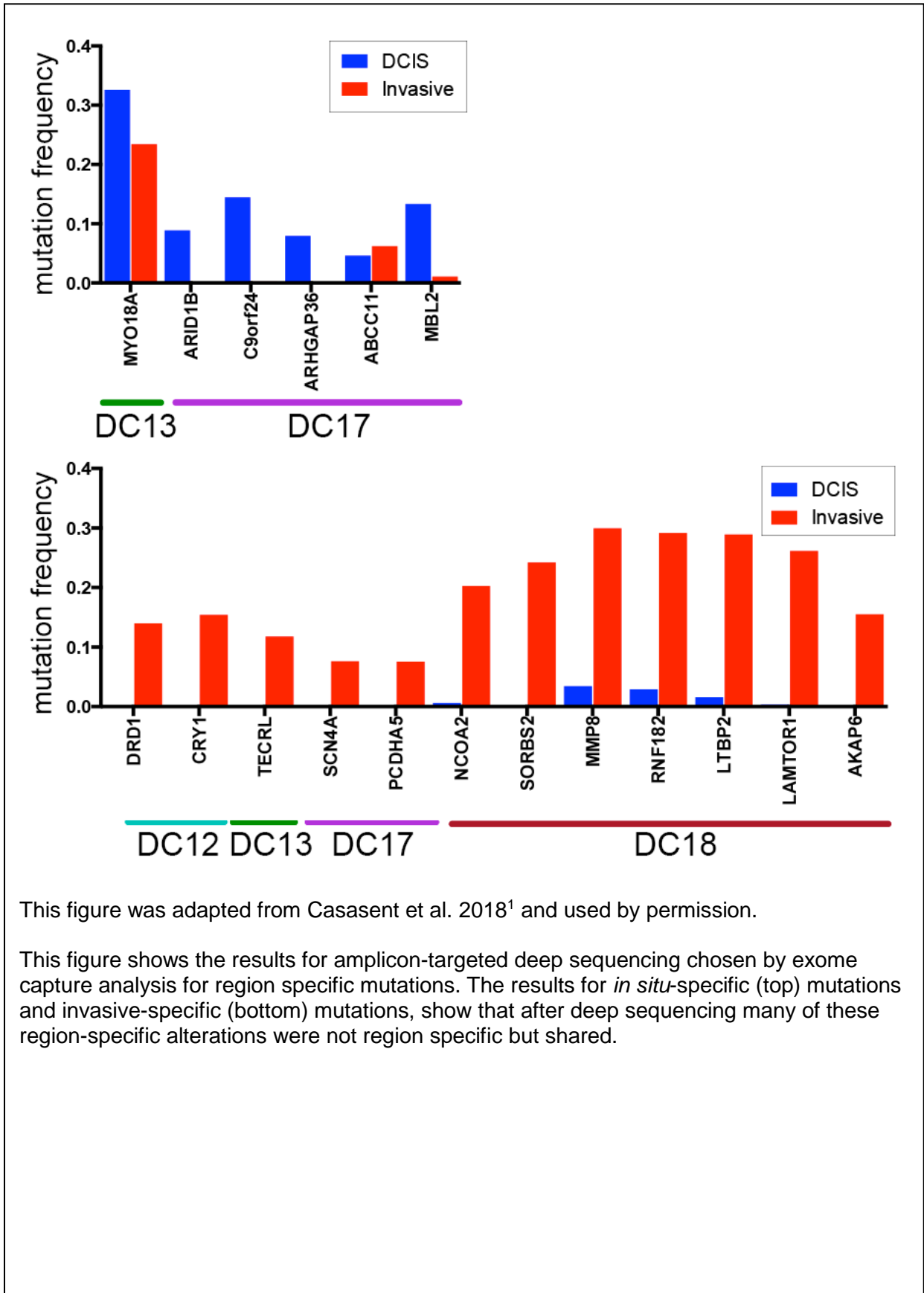
Figure 63 Regional Exome Oncomap



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Oncomap of nonsynonymous mutations in the *in situ* and invasive regions from each patient. The presence or absence of mutations has been updated based on the results from amplicon deep-sequencing validation data. Known breast cancer genes are indicated in bold, while mutations that were validated by deep-amplicon sequencing are in italics.

Figure 64 Regional Amplicon Validation



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

This figure shows the results for amplicon-targeted deep sequencing chosen by exome capture analysis for region specific mutations. The results for *in situ*-specific (top) mutations and invasive-specific (bottom) mutations, show that after deep sequencing many of these region-specific alterations were not region specific but shared.



Table 4 Regional Invasive-Specific Mutations

Gene Annotations for Invasive-Specific Mutations									
Patient	Gene	Chr	Position	Ref	Var	SITU_FREQ	INV_FREQ	POLYPHEN	SIFT
DC12	CRY1	chr12	107395083	C	T	0.01	0.52	0.912	0
DC12	DRD1	chr5	174869715	G	A	0.00	0.47	0.981	0
DC13	TECRL	chr4	65274856	G	A	0.00	0.27	0.735468	0.66
DC17	FAP	chr2	163070565	C	G	0.00	0.22	1	0.09
DC17	SCN4A	chr17	62025343	G	C	0.00	0.08	NA	0
DC17	PCDHA5	chr5	140201856	C	G	0.00	0.08	0.523167	0.02
DC18	AKAP6	chr14	33242950	C	G	0.01	0.46	0.867	0
DC18	SORBS2	chr4	186541289	T	A	0.01	0.71	0.705202	0
DC18	FANCA	chr16	89815132	A	T	0.00	0.38	0.955	0.42
DC18	CEP192	chr18	13071040	T	A	0.00	0.38	0.924	NA
DC18	TRIM47	chr17	73872438	C	A	0.00	0.62	0.983	0.02

This table was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

This table lists the invasive-specific nonsynonymous mutations that were identified by exome sequencing of laser-capture-microdissected tissue regions.

The mutations listed were detected in the invasive regions and were not detected in the *in situ* region, after filtering by matched normal germline variants.

The table lists the following columns in order: patient identifiers, gene names, chromosome and position, reference nucleotide, variant nucleotide, *in situ* mutation frequencies (SITU\_FREQ) normalized by tumor purity, invasive mutation frequencies (INV\_FREQ) normalized by tumor purity, polyphen2 damaging impact scores (POLY), and SIFT functional impact prediction scores.

Table 5 Regional Deep SNVs Genomics and Reads

Patient	Gene	Chr	Position	Ref	Var	Type	A	C	G	T
DC12	DRD1	chr5	174869715	G	A	in situ	72	9	52208	123
						Invasive	1666	1	10190	45
						Normal	75	4	14579	35
DC12	CRY1	chr12	107395083	C	T	in situ	503	128175	11	245
						Invasive	59	11798	4	2161
						Normal	76	10319	1	38
DC13	TECRL	chr4	65274856	G	A	in situ	9770	171	4214428	1993
						Invasive	62	0	464	0
						Normal	0	1	556	0
DC13	MYO18A	chr17	27423866	T	A	in situ	276279	541	1033	569662
						Invasive	14452	26	93	47125
						Normal	114	13	99	59998
DC13	ZBTB14	chr18	5291971	T	A	in situ	2815	19	90	89006
						Invasive	25	0	4	121
						Normal	1	0	9	5430
DC13	MYO18B	chr22	26239717	G	A	in situ	480489	780	925921	754
						Invasive	7491	48	81899	148
						Normal	113	43	128978	37
DC17	PCDHA5	chr5	140201856	C	G	in situ	575	929093	81	191
						Invasive	111	28932	2373	31
						Normal	126	86789	28	18
DC17	SCN4A	chr17	62025343	G	C	in situ	63	194	404521	567
						Invasive	37	12095	145897	442
						Normal	12	246	100712	80
DC17	ARHGAP36	chrX	130217875	C	T	in situ	32	14378	16	1244
						Invasive	7	11738	13	20
						Normal	1	2693	3	9
DC17	ABCC11	chr16	48226526	G	T	in situ	39	20	81322	3933
						Invasive	76	24	231953	15359
						Normal	54	23	185477	502
DC17	MBL2	chr10	54530499	G	A	in situ	18229	58	117956	423
						Invasive	2372	48	212951	538
						Normal	191	4	16598	28
DC17	ARID1B	chr6	157431633	G	A	in situ	1736	22	17703	49
						Invasive	405	145	234196	403
						Normal	95	133	246785	240
DC17	C9orf24	chr9	34382807	G	C	in situ	11	3258	19190	92
						Invasive	2	32	14489	49
						Normal	9	334	77171	74
DC18	NCOA2	chr8	71075008	T	C	in situ	7	413	139	65000
						Invasive	4	10052	66	39499
						Normal	8	83	44	23264
DC18	MMP8	chr11	102589262	A	T	in situ	28146	38	6	1012
						Invasive	139769	544	141	60075
						Normal	171278	606	48	536
DC18	RNF182	chr6	13977826	A	T	in situ	43564	104	21	1320
						Invasive	18089	96	16	7499
						Normal	16286	109	17	174
DC18	LAMTOR1	chr11	71809862	C	A	in situ	442	125801	47	88
						Invasive	38570	108833	42	73
						Normal	337	48816	15	28
DC18	LTBP2	chr14	75019600	C	T	in situ	817	549025	111	8784
						Invasive	2187	1148556	896	468608
						Normal	80	21492	23	3765
DC18	SORBS2	chr4	186541289	T	A	in situ	32	10	99	42162
						Invasive	2986	7	24	9309
						Normal	54	33	230	21864
DC18	AKAP6	chr14	33242950	C	G	in situ	215	242691	297	30
						Invasive	483	238533	43940	52
						Normal	72	76655	331	8
DC18	HDAC4	chr2	240078423	C	T	in situ	1	335	0	0
						Invasive	1	469	0	3
						Normal	162	82716	4	30

This table was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

This table shows the results of targeted amplicon deep sequencing selected region-specific mutations detected by exome sequencing. The table columns include patient number, gene name, chromosome number, chromosome position, reference base, variant base, region, read counts for A, C, G, and T.

Table 6 Regional DeepSNVs Result Details

Patient	Gene	SITU_MF	INV_MF	NORM_MF	P-value DN	P-value IN	Validated
DC12	DRD1	0.00	0.14	0.01	2.7211E-02	1.7892E-03	INV ONLY
DC12	CRY1	0.00	0.15	0.00	Not Significant	1.3216E-03	INV ONLY
DC13	TECRL	0.00	0.12	0.00	1.9430E-18	3.2333E-03	INV ONLY
DC13	MYO18A	0.33	0.23	0.00	1.9430E-18	2.0994E-04	Pre-existing
DC13	ZBTB14	0.03	0.17	0.00	2.1105E-06	5.1415E-12	Pre-existing
DC13	MYO18B	0.34	0.08	0.00	3.3535E-21	4.2214E-04	Pre-existing
DC17	PCDHA5	0.00	0.08	0.00	2.0250E-02	1.4055E-04	INV ONLY
DC17	SCN4A	0.00	0.08	0.00	2.4231E-05	1.4316E-03	INV ONLY
DC17	ARHGAP36	0.08	0.00	0.00	1.0944E-04	Not Significant	DCIS ONLY
DC17	ABCC11	0.05	0.06	0.00	1.1563E-11	1.4828E-03	Pre-existing
DC17	MBL2	0.13	0.01	0.01	3.3084E-08	Not Significant	DCIS ONLY
DC17	ARID1B	0.09	0.00	0.00	7.6277E-76	Not Significant	DCIS ONLY
DC17	C9orf24	0.14	0.00	0.00	6.8761E-18	5.2801E-02	DCIS ONLY
DC18	NCOA2	0.01	0.20	0.00	Not Significant	1.4828E-03	Pre-existing
DC18	MMP8	0.03	0.30	0.00	2.3582E-13	1.8028E-05	Pre-existing
DC18	RNF182	0.03	0.29	0.01	1.5575E-02	3.9687E-04	Pre-existing
DC18	LAMTOR1	0.00	0.26	0.01	Not Significant	2.3879E-04	INV ONLY
DC18	LTBP2	0.02	0.29	0.15	2.1247E-28	Not Significant	Pre-existing
DC18	SORBS2	0.00	0.24	0.00	Not Significant	6.4445E-44	INV ONLY
DC18	AKAP6	0.00	0.16	0.00	2.3346E-02	4.6470E-04	INV ONLY
DC18	HDAC4	0.00	0.01	0.00	Not Significant	Not Significant	False Postive

This table was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

This tables shows the results of targeted amplicon deep sequencing for a subset of the *in situ*-specific and invasive-specific mutations detected by exome sequencing.

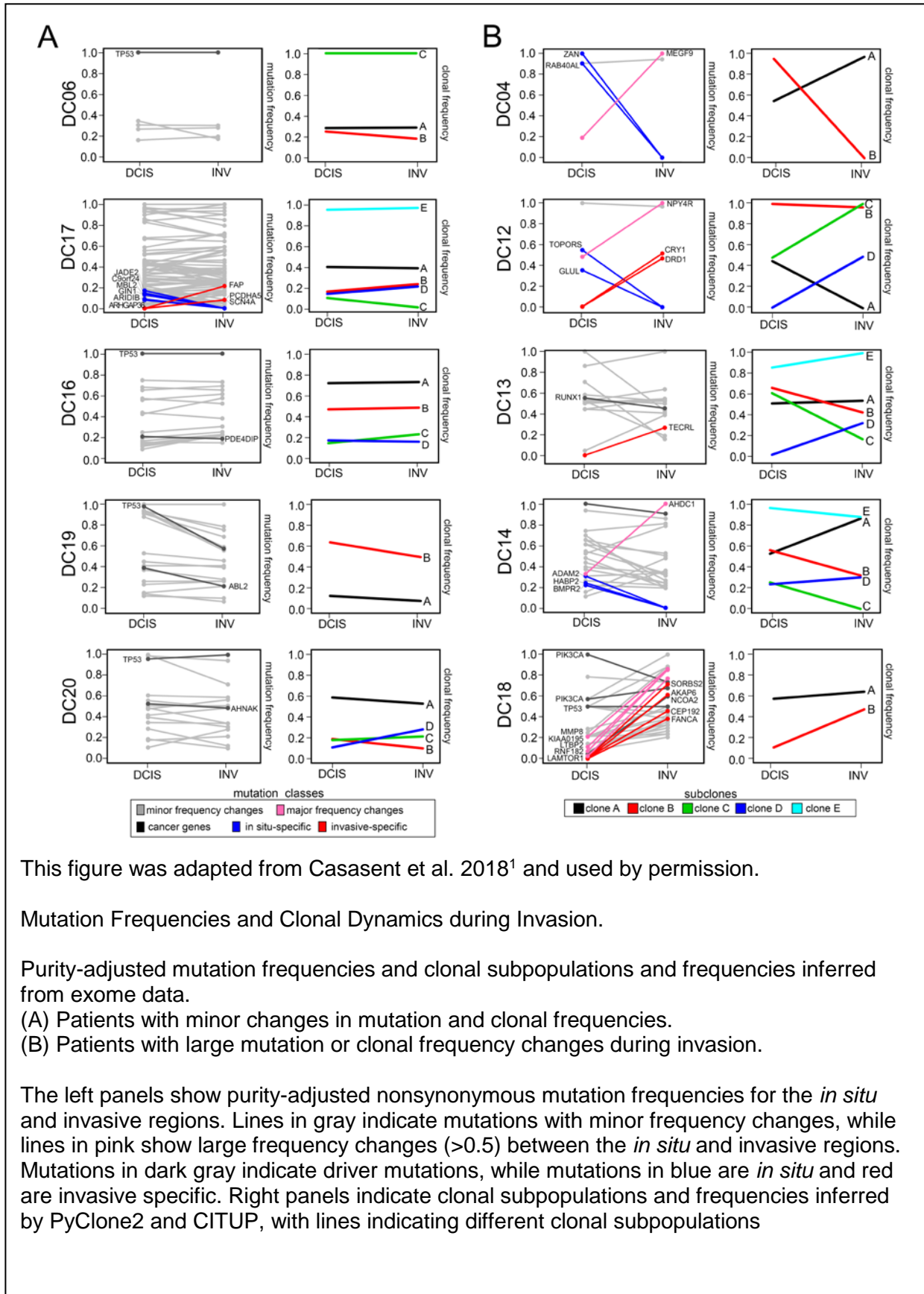
DeepSNV was used to determine the statistical significance of each mutation relative to the site-specific background error rate in matched normal tissues.

The table columns include patient number, gene name, *in situ* mutation frequencies (SITU\_MF), invasive mutation frequencies (INV\_MF), p-value for DeepSNV of *in situ* to normal comparison (p-value DN), p-value for DeepSNV invasive to normal comparison (p-value), and finally the validation results.

The invasive-specific mutations may have occurred at low frequencies in the ducts prior to invasion (below the exome sensitivity level) or alternatively after invasion, during the expansion of the invasive tumor mass. Another possibility is they were sampled from different geographical regions; however, this is unlikely in synchronous DCIS-IDC tissue since cells were collected from adjacent regions in the same tissue sections. To determine if the invasive-specific mutations were acquired in the ducts or after invasion, we performed targeted deep-amplicon sequencing at high coverage depth (mean=453,446X) for a subset of the *in situ*-specific and invasive-specific mutations (Figure 64 Regional Amplicon Validation). In parallel, we performed targeted deep-amplicon sequencing of matched normal breast tissues to establish site-specific background error rates and identified significant mutations using DeepSNV<sup>228</sup> (Table 5 Regional Deep SNVs Genomics and Reads and Table 6 Regional DeepSNVs Result Details).

The amplicon data at higher coverage depth (226,000X) showed many of the *in situ*-specific mutations were present at low frequencies in the invasive regions. However, most of the invasive-specific mutations (8/12) were found to be exclusive to the invasive tissues as shown in Table 4 Regional Invasive-Specific Mutations. These mutations are unlikely to play an important role in invasion, since they were acquired after the tumor cells escaped the basement membrane, during the expansion of the invasive carcinoma. However, in one patient (DC18), we identified a few mutations (*NCOA2*, *MMP8*, *RNF182*, *LTBP2*) that were pre-existing at low frequencies and increased in frequency during invasion (shown in pink in Figure 65 Regional Frequency Changes and Table 7 Regional Pre-Existing Mutations). These mutations included *MMP8*, a matrix metalloproteinase that plays a role in breaking down the extracellular matrix<sup>229</sup>, and *LTBP2* that interacts with TGF-beta to regulate cell adhesion<sup>230</sup>.

Figure 65 Regional Frequency Changes



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

#### Mutation Frequencies and Clonal Dynamics during Invasion.

Purity-adjusted mutation frequencies and clonal subpopulations and frequencies inferred from exome data.

(A) Patients with minor changes in mutation and clonal frequencies.

(B) Patients with large mutation or clonal frequency changes during invasion.

The left panels show purity-adjusted nonsynonymous mutation frequencies for the *in situ* and invasive regions. Lines in gray indicate mutations with minor frequency changes, while lines in pink show large frequency changes (>0.5) between the *in situ* and invasive regions. Mutations in dark gray indicate driver mutations, while mutations in blue are *in situ* and red are invasive specific. Right panels indicate clonal subpopulations and frequencies inferred by PyClone2 and CITUP, with lines indicating different clonal subpopulations

Table 7 Regional Pre-Existing Mutations

Patient	Gene	Chr	Position	Ref	Var	SITU_FREQ	INV_FREQ	POLY	SIFT
DC04	MEGF9	chr9	123476336	C	A	0.19	1.00	NA	0
DC12	NPY4R	chr10	47087501	C	T	0.48	1.00	0.332	0.16
DC14	AHDC1	chr1	27874569	G	A	0.33	1.00	0.983	0
DC18	KIAA0195	chr17	73489016	C	A	0.07	0.85	0.732	0.02
DC18	NCOA2	chr8	71075008	T	C	0.05	0.60	NA	0.05
DC18	RNF182	chr6	13977826	A	T	0.21	0.86	0.996	NA
DC18	LTBP2	chr14	75019600	C	T	0.11	0.85	0.603	0
DC18	LAMTOR1	chr11	71809862	C	A	0.03	0.77	0.334	0.03
DC18	MMP8	chr11	102589262	A	T	0.25	0.88	0.732	0

This table was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

This table lists nonsynonymous mutations with increased frequencies (>0.5) in the exome data of the laser-microdissected *in situ* and invasive regions.

The table columns list patient identifiers, gene names, chromosome and position, reference nucleotide, variant nucleotide, nucleotide position in gene, amino acid (AA) positions, amplicon deep-sequencing validation status, *in situ* mutation frequency adjusted by tumor purity (SITU\_FREQ), invasive mutation frequency adjusted by tumor purity (INV\_FREQ), polyphen2 damaging impact scores (POLY), and SIFT functional impact prediction scores.

We further investigated concordant mutations for large changes in mutation frequencies by constructing tumor-purity normalized line plots (Figure 65 Regional Frequency Changes). This analysis showed only minor changes in mutation frequencies during invasion in five patients, while the other five patients had at least one mutation with a large ( $>0.5$ ) frequency change. From these data, we identified 7 mutations that underwent large ( $>0.5$ ) mutation frequency changes during invasion, including *MEGF9* in DC4 (19% to 100%), *NPY4R* in DC12 (48% to 100%), *AHDC1* in DC14 (33% to 100%), and 4 mutations in DC18 (Table 7 Regional Pre-Existing Mutations). However, most patients (DC4, DC12, DC14) had only a single concordant mutation that underwent a large frequency shift during invasion.

To infer clonal dynamics during invasion, we applied PyClone2<sup>231</sup> and CITUP<sup>232</sup> to cluster mutation frequencies and estimate clonal subpopulations after purity and copy number normalization (Figure 65 Regional Frequency Changes). This analysis identified 2-5 major subpopulations in each patient, which was higher than the number of subpopulations detected by single cell copy number profiling. We found several tumors to be monoclonal by single cell copy number profiling (DC6, DC12, DC17, and DC19) but showed 2-5 subpopulations based on inferred mutation clusters. This data suggested an ongoing mutational evolution in the ducts after copy number evolution, leading to further subclonal diversification prior to invasion into adjacent tissues. While some of the clonal frequencies shifted during invasion (Figure 65 Regional Frequency Changes), the total number of subpopulations estimated from exome mutations remained consistent in most patients.

### 3.3 Study Limitations

This study has a few notable limitations. First, the cohort size was limited to 10 patients, providing only a small snap shot of synchronous DCIS-IDC patients. In addition, the cohort receptor status was mixed. Since, we did not have just one receptor status, like TNBC, expanding our conclusions was difficult because the variation observed between tumors could be caused by receptor status, which has been observed to provide significantly divergent clinical outcomes<sup>41</sup>. Also, we studied only synchronous DCIS-IDC, which has already progressed to invasive carcinoma, preventing us from examining ITH as a method to discover useful prognostic markers. These issues with our cohort and cohort size can be addressed by examining more DCIS-IDC and DCIS-only tumors using TSCS. Thus, we cannot exclude the possibility some early breast cancer patients follow alternate evolutionary models, particularly in low-grade tumors.

Second, we profiled a limited number of cells in each patient, which may lead to sampling bias. We profiled approximately 50 cells per region and 100 cells per patient. While we did calculate posterior saturation curves<sup>16</sup>, these curves are based on the number of subclones detected, and our subclone method is based on the number of cells we profiled, providing a self-referential infinite loop. Our posterior saturation curves suggest we sampled sufficient cells to detect the subclones defined for each patient (see Saturation Curves for individual tumors). For future analyses, we suggest a different method for defining subclones be used, to prevent the infinite loop of interdependent conditional variables. By using chromosomal events, with each unique event set defining a subclone, the number of subclones defined should not be determined by a clustering algorithms dependence on number of cells to define clusters. While this method might significantly increase the number of subclones defined, and therefore require more clustering, it should be more reproducible than the k-means methods of clustering.

Third, our study was limited to DNA alterations, copy number in single cells, and regional mutation analysis of synchronous DCIS-IDC. Other follow up studies could investigate



the regional transcriptome (using Geo-Seq<sup>12, 13</sup> or improved FISSEQ<sup>5</sup> protocols like barcode *in situ* targeted sequencing BaristaSeq<sup>233</sup>), epigenetic modifications (for which no single cell spatially resolved methods are currently available), and even spatially resolved global protein (Imaging Mass Cytometry<sup>234, 235</sup>) expression of single cells in synchronous-DCIS.

Fourth, we examined copy number profiles of single tumor cells and did not examine mutation or expression profiles of the surrounding stromal cell types. Stroma have been shown to assist in the invasion and migration of tumor cells *in vivo* and could also modulate the ability of tumor cells to invade surrounding tissues<sup>236, 237</sup>. These represent important future directions addressable with single cell RNA, epigenomic profiling, and protein expression methods.

### **3.4 Conclusions**

In this study we developed a spatially-resolved single cell DNA sequencing method and applied it to study genome evolution during invasion in 10 synchronous DCIS-IDC breast cancer patients. We created a new method combining Single Cell Laser Capture Microdissection with Single Cell DOP-PCR and Single Cell Sequencing to preserve spatial information when investigating single cell heterogeneity. We called this method TSCS and used it to examine the differences in single cell copy number heterogeneity in synchronous breast cancer. Our results from TSCS data from synchronous DCIS-IDC strongly supports three major biological conclusions: (1) the subclones observed all arose from a single cell of origin, (2) that all the clones were able to escape the ducts (multiclonal invasion), and (3) that these copy number clones were created in bursts of CIN within the ducts prior to invasion.

#### **3.4.1 Using Topographical Single Cell Sequencing (TSCS)**

Our first conclusion is Single Cell Laser Capture Microdissection with Single Cell DOP-PCR and Single Cell Sequencing provided quality single cell copy number data. The Zeiss Robo PALM LCM system enabled selection of single cells within about 1-micron of another cell and used a touchless approach to transfer cells. TSCS generated high-resolution single cell copy number profiles with spatial X-Y coordinates by mapping back to the original tissue scan. For multiple slides per tumor, we estimated the Z-axis based on the number of sections between slides. The single cell genomic data was mapped to the spatial coordinates to delineate the topographic organization of different clonal genotypes in the tissue sections. This method allowed us to examine single cell genomics, copy number clones, and spatial information.

Our current protocol should be expandable to other types of data such as DNA mutations and RNA. Using this method, while time consuming, can produce very detailed maps of changes in aneuploidy within a tumor, which gives us more information about changes in aneuploidy than FISH (which can only examine a few alterations at a time) while still retaining

the spatial information. While Fluorescent *in situ* sequencing (FISSEQ<sup>5, 238</sup>) can provide very detailed spatial and RNA data, it has not been expanded to DNA. TSCS, on the other hand, provides very detailed spatial and DNA copy number data. Therefore, TSCS is a powerful tool to define the differences in subclonal intermixture within a tumor and suggests a model of clonal progression based on this data.

In Navin and Hick's review "Tracing the Tumor Lineage" from 2010, the authors suggested 5 different models of tumor progression, differentiated by combining the information from a phylogenetic tree and tumor subclones spatial relationships<sup>9</sup>. Our technique provides a way to examine both copy number (used to build phylogenies) and to examine the spatial relationship of these phylogenies. In addition to being able to distinguish between these different models of progression, we propose that TSCS and other spatially-resolved single cell methods will be able to differentiate polyclonal models of invasion, such as bottlenecks and multiclonal invasion.

### **3.4.2 Single Cell of Origin**

The first major biological conclusions from our TSCS data is that synchronous DCIS-IDC subclones arose from the same cell, a single common ancestor, fitting the single cell of origin hypothesis. Since single cells shared almost all copy number events, we propose all our cases arose from common ancestors. Only one case, DC13, identified a very limited set of shared copy number events. We proposed the DC13 subclones diverged very early, producing two major clones with few shared copy number events. Although, for DC13, we must acknowledge the possibility of two distinct cells of origin. However, the two subclones were closely related to each other and highly correlated. Therefore, we conclude that we observed a single clone of origin because highly similar copy number profiles are more likely to occur from one single cell of origin rather than convergent evolution.

### 3.4.3 Multiclonal Invasion

The next major biological conclusion is that we have multiple clones and all clones were present in the ducts and the invasive tumor. From this data, we suggested a model of invasion where all clones escape and survive outside the ducts called *multiclonal invasion*. *Multiclonal invasion* is distinguished from evolutionary bottleneck because all the clones escape into the surrounding tissue. Our model is consistent with studies using flow sorting and single cell copy number profiling in a single DCIS patient, which also reported evidence that multiple clones crossed the basement membrane <sup>6</sup>.

Our model challenges previous work which posited DCIS invasion occurred via a population bottleneck<sup>114, 239, 170</sup> or through independent cell lineages<sup>139</sup>. In our data, we show the same subclones were present in both the *in situ* and invasive regions in all 10 patients, with no additional CNA events acquired during invasion and few invasive-specific mutations found though regional exome capture. These data suggested a single clone was not selected during invasion through an evolutionary bottleneck. Furthermore, our data does not support an independent lineage model<sup>160, 161</sup>, since we identified a large number of shared truncal mutations and CNAs in all tumor cells, suggesting a field effect did not give rise to two clones that formed the *in situ* and invasive regions independently.

The *multiclonal invasion* model suggested selection of clones observed in metastasis occurs later, possibly during dissemination into the blood stream<sup>240</sup>. This model suggests that differences between the *in situ* and invasive regions are likely caused by larger spatial effects, but on a more local level there are few differences between paired *in situ* and invasive regions. It is possible the further from the ducts cells migrate, the more selection occurs, either from changes in the microenvironment or a more stochastic bottleneck due to regional separation. If the first case is true, finding a genetic marker to define these super-adaptor clones could help cancer treatment, because it addresses the more deadly or aggressive clones. Separating super-adaptors, highly migratory, or highly invasive clones is like the difficult effort and analysis required to separate driver and passenger mutations. However, if instead the model of invasion

and progression is more stochastic, like the localized multiclonal invasion we observed, treatments for all clones is necessary, because any of them could eventually invade.

#### **3.4.4 Intraductal Punctuated Evolution**

There are two well-known models of punctuated evolution: the branching expansions model<sup>241, 242</sup> and the big bang model<sup>243, 244</sup>. Both models could explain the observed punctuated copy number or mutation profiles observed in cancer. The branching model postulates that bursts of alteration occur during the initial stages of tumor development followed by expansion of a few clones<sup>241, 242</sup>. On the other hand, the big bang model suggests a major event occurred early in tumor progression and created many diverse clones, while only a select few of these clones may survive and continue to gradually accumulate additional mutations, the initial clones will be very distinct<sup>243, 244</sup>. Our data fits the branching expansions model for all tumors, except for DC13, which appears to have extreme divergence between clones and therefore might fit the big bang model better.

Previous single cell genomic studies of IDC cancers from the Navin lab<sup>16, 4, 17</sup> suggested a model of punctuated evolution, which also matches the branching expansions model. These papers examined IDC and found in single cell copy number, bursts of genomic alterations instead of a gradual accumulation. Therefore, the Gao et al<sup>16</sup> paper suggested copy number evolution was occurring in a sudden burst of genomic instability, followed by clonal expansion. The distinct copy number clones observed suggested intermediate copy number states were selected against, not present, or present for a short time.

While earlier papers used gated diploid and aneuploid tumor peaks<sup>134, 8, 17</sup> and could have missed intermediates based on sorting, the Gao et al paper<sup>16</sup> examined both ungated and gated single cell tumor copy number profiles<sup>16</sup>. We observed 5-7% of "intermediate" copy number profiles, suggesting intermediates were present. Gao et al suggested these intermediates did not enhance fitness, and were outcompeted by both predecessors and descendants, resulting in the tumor mass consisting of what appears to be punctuated copy

number clones, because of a series of discrete clonal expansions. Oddly, in the Gao et al paper, the somatic mutations observed in single cells appeared to show a more gradual accumulation instead of discrete clones<sup>16</sup>.

However, a series of discrete clonal expansions can still explain the gradual accumulation of mutations. If the somatic mutations are passenger mutations which do not affect the fitness of the clones, then the mutations could be carried between two copy number clones if a direct descendant was part of a later clonal expansion. It is even possible for a driver mutation, one that initiates tumorigenesis, invasion, or tumor progression, to also be observed as a gradual accumulation of mutations, because the direct descendants of this mutation would expand. This is especially true if the driver mutation is related to genome dysregulation, such as chromosome segregation, which could cause sudden bursts of chromosomal alterations of which only some would be viable.

The model of single cell of origin where a burst of genome instability (copy number) expands, explains the clonal relationships, low number of intermediates, and both diverse and monoclonal tumors. Although the mechanisms of punctuated copy number evolution need additional research, we speculate that telomere crisis<sup>112</sup> is a plausible model<sup>1, 16</sup>. While it is possible for a tumor to undergo many bursts of genome instability, it is not required. For example, monoclonal tumors can be observed and still fit the punctuated copy number model. Monoclonal tumor could arise from (1) only 1 burst of CIN at the time of surgery, (2) the first clone to be the most fit, (3) one clone greatly out competing the others, or (4) more clones existing but not observed because (a) they were too small in number or (b) there was a strong regional effect or (c) our method selected against them. TSCS would select against cells or clones that have (1) smaller nuclei – possibly due to more genome packaging or less aneuploidy or (2) cells/clones that adhere closely together or closely with other cells<sup>1</sup>. Our study did not gate cells, increasing the possibility for us to observe intermediates as per Gao et al's ungated populations<sup>16</sup>. While we did not specifically see intermediates, we did see our populations were usually closely related, suggesting succession of discrete clonal expansions.

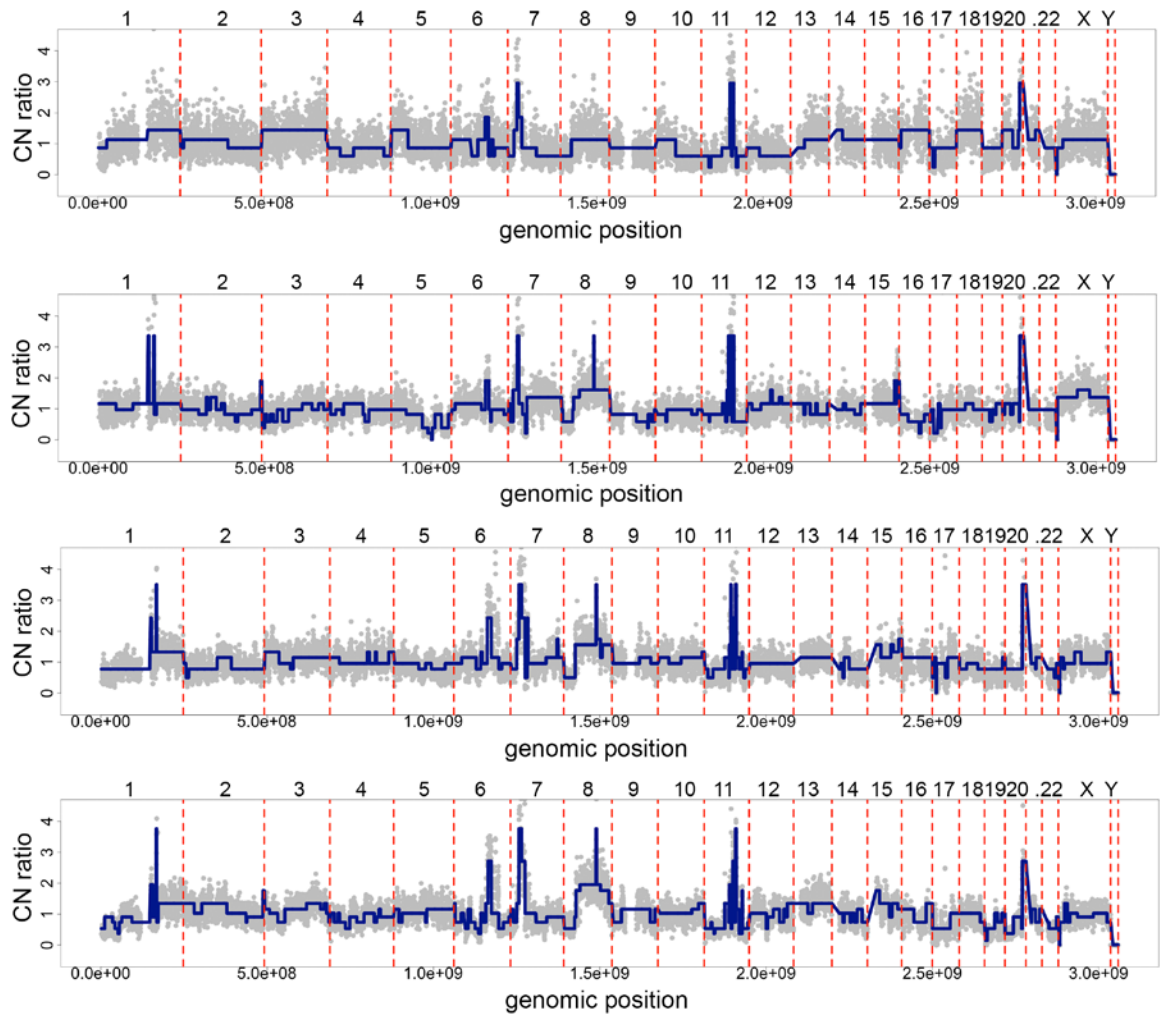
Due to the spatial information from TSCS, we can go a step further and suggest punctuated evolution most likely occurs within the ducts prior to invasion. We propose this because we observed the same number of clones and often similar frequencies of clones in both the *in situ* and invasive regions, suggesting clonal instability and clonal expansion occurred in the ducts prior to invasion. An alternative model of invasion that could account for all clones observed in both *in situ* and invasive regions is self-seeding<sup>245, 246</sup>. However, while self-seeding has been observed in model organisms in metastatic cases, where tumor cells are shedding into the blood stream<sup>27</sup>, it has not been observed in localized invasion, especially into nutrient low places like ducts, which often have necrotic centers due to lack of nutrients (specifically oxygen)<sup>247</sup>.

While, we proposed a series of discrete clonal expansions, there are a few bioinformatic technical reasons we could have missed observing intermediates including (1) we could be filtering our intermediates as noise, (2) k-means clustering grouped intermediates in a discrete number of k clusters making them appear more discrete, or (3) the CBS<sup>205</sup> parameters could be over smoothing the data.

For each of these statements, we ask if this is logical and probable. Filtering intermediates is unlikely because we observed intermediate copy number profiles in the form of pseudodiploid cells, specifically in DC6 which included the loss of the X chromosome in otherwise diploid cells. Loss during clustering could happen if we did not observe 3 cells with similar profiles, but this is unlikely because the intermediates should be very similar to other clusters and hence be unaffected by filtering via dbscan, which requires correlation within a cluster or number of break points.

While k-means clustering could group together intermediates with a clonal expansion, or even a mutator phenotype<sup>9</sup>, when we examine the consensus profiles as means, we observed consistent discrete differences between the clusters, causing a mutator phenotype to be discountable, since it would create a number of unshared events with many similar profiles.

Figure 66 Segmentation Data



This figure was adapted from Casasent et al. 2018<sup>1</sup> and used by permission.

Example of TSCS Copy Number Profiles from DC6.

Genomic copy number profiles corresponding to single cells in brightfield images that were isolated from tissue sections by laser-catapulting.



A clustering algorithm without sufficient sensitivity could cluster two or more clones together and therefore under-sample the tumor; however, I point specifically to DC18 where clones B and C are closely related but distinct, suggesting our clustering algorithm is sensitive enough, since DC18 clones B and C share almost all chromosomal alterations.

The last reason we could be missing intermediates is that CBS is over smoothing the data. While our TSCS data is noisier than data generated by FACS, since we used the same parameters, it is highly unlikely we are missing intermediates based on under-segmentation. In addition, when we examine raw bin data overlaid with segmentation, we do not see over smoothing (Figure 66 Segmentation Data). Over smoothing would result in the segmentation line (blue) appearing uncentered over the area of the bins and continuing through areas where the bins (grey) have moved up or down.

Of all possible reasons for missing intermediates, missing them based on under-clustering appears the most likely. It is possible we are sampling the intermediates and miss classifying them because we are not ordering them clearly enough to see the sequential progression. Alternatively, we could be missing intermediates due to sampling bias because it is not feasible to sample every cell within the tumor and therefore are stating that the data supports sequential clonal progression because we do not observe each stage. Instead we observed cells with many of the same alterations and another set with some of the same shared mutations, many of them unique single CNAs.

As for sampling, our calculations for sample size are based on the assumption that we are calculating the number of clones correctly, which produces a logical loop. If we are undercalculating the number of clones, then we are also under calculating the number of cells needed to evaluate the heterogeneity and evolution within a tumor and need to increase the number of clones used in our ad hoc saturation model, which would in turn increase number of cells required.

However, the total number of clonal subpopulations we identified is similar to previous reports in IDC <sup>16, 8, 110, 85</sup> and is consistent with a punctuated model of copy number evolution<sup>16</sup>,

in which short bursts of genome instability give rise to multiple clones that stably expand to form the tumor mass <sup>16, 4</sup>, suggesting that we are not underestimating the major subclones, even if we are missing rare intermediates. Unlike the previous studies which could not resolve where the punctuated bursts of genomic instability occurred, we were able to demonstrate that these events most likely took place inside the ducts prior to invasion.

In addition to our single cell copy number data, our regional exome data also suggest most somatic mutations, including driver mutations in *TP53* and *PIK3CA*, were acquired in the ducts prior to invasion, at the earliest stages of tumor progression. While it might be possible for the differences we observed between the *in situ* and invasion region to be driving invasion, we find this unlikely because (1) most of the mutations were not found to be invasive specific, (even if the mutation first appeared to be invasive specific, amplicon validation indicated the majority of invasion specific mutations appeared in the ducts prior to invasion) and (2) if a mutation occurs and is driving invasion, we would expect this mutation to arise in one clone and cause a bottleneck of invasive subclones, with the subclone containing the driver mutation being more frequent in the invasive regions.

Since we do not see a bottleneck of copy number clones, it is highly improbable mutation is driving invasion since it would need to occur multiple different times in separate subclones. Instead if the driver mutation exists, it is likely this mutation is allowing a clone to break through the basal membrane and other clones are following it out, making it difficult to separate the more aggressive subclones. Sampling for much larger distances from the ducts might make it possible to observe a bottleneck, but it would still be difficult to determine if this bottleneck is occurring because the clones are migrating or if it is a stochastic bottleneck. To examine this alternate bottleneck question, it would be necessary to observe the migration patterns of the different clones. While there is some modeling work that tries to examine this, it is still difficult to address this question using current resources.

## 4 Discussion and Future Directions

This section is based on the research paper "Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing" published in the *Cell* in 2018, by Casasent et al<sup>1</sup> and the review paper "Genome evolution in ductal carcinoma *in situ*: invasion of the clones" published in the *Journal of Pathology* in 2017, by Casasent, Edgerton, and Navin<sup>2</sup>. This section expands from the discussion of the future of TSCS and other single cells technologies. Portions of this section are adapted from the Casasent et al 2016<sup>2</sup>. Figures and text from this Casasent et al 2016<sup>2</sup> have been reused or modified under the journal's academic copyright license with permission from John Wiley & Sons, Ltd for Pathological Society of Great Britain and Ireland.

Here, I discuss the implications of this work. I explain what the TSCS protocol brings in terms of advancement to science and to what protocols it can be applied. I will also discuss the importance of the multiclonal evolution in terms of basic research and clinical implications.

### 4.1 Single Cells and Topographical Information

Being able to examine the morphology, location, and genetics of single cells allows more direct association of the relationships of genotypes and phenotypes. While other techniques have examined a specific genetic change *in situ* (for example, FISH<sup>248, 249</sup> and specific-to-allele PCR–FISH (STAR-FISH)<sup>108</sup>), these are usually limited to 1 to 5 genetic markers. In contrast, TSCS allows the examination of single cell full genome CNA with morphology and location. The first method to examine full genome RNA with micron level spatial resolution is fluorescent *in situ* sequencing (FISSEQ<sup>5</sup>), which is limited to RNA and encounters issues across different cell types and even different subcellular regions, making FISSEQ data noisy and somewhat unreliable<sup>5</sup>. However, there have been multiple improvements to FISSEQ, including BaristaSeq, which increased the barcode length of unique cells and changed the chemistry from SOLiD to Illumina, increasing the output by about 30%<sup>233</sup>.

Combining spatial and morphologic information with genomic or transcriptomic information is important to allow observation of changes occurring within an organism. In synchronous DCIS, these observations revealed the clones were intermixing and, while there might be gradient changes across a larger region, two or more distinct clones were often found side by side, raising questions of possible symbiotic relationships between the clones.

TSCS is currently limited to working with copy number profiles and has a tedious collection process. In TSCS, each cell is collected one at a time, allowing recording of the precise location of every cell, but reducing the throughput. Future directions to improve this technique should include increasing throughput and expanding to single cell DNA mutations and RNA expression.

#### **4.1.1 Technical Barrier: Scalability**

Previous studies by Martelotto et al showed an ability to hand-microdissect thick FFPE tissue sections and to use 4',6-diamidino-2-phenylindole (DAPI) staining and flow sorting to separate aneuploid cells from the dissected *in situ* and invasive regions<sup>6</sup>. This method provides higher throughput compared to TSCS but loses spatial resolution. However, the loss of spatial resolution might be an adequate tradeoff for a higher throughput technique depending on the scientific question. In addition, the Martelotto et al technique required pairing with sorting methods, which will only work for cases where good nuclear markers can separate the cells of interest.

To improve the throughput of our method, we examined pairing different stains with DAPI, so that microdissection can be used on small regions of tumor. We tested hemotoxalin, eosin, "hemotoxalin and eosin" (the standard H&E used for previous methods in our lab), methylene blue, methylene green, and cresyl violet. After staining, we manually removed tissue from the slide and flow sorted the results similar to Martelotto et al<sup>6</sup>. We found hemotoxalin, eosin, H&E, and cresyl violet all seemed to inhibit DAPI results for flow sorting. Ploidy peaks appeared less disrupted in the methylene blue and green stains. However, more data is

needed. Methylene blue and green paired with the direct tagmentation method (described in Haowei Du's Master's thesis) should allow us to collect and amplify 384 cells from one collection step and produce libraries with fewer steps, and thus increase the throughput of this process.

Our next step is to apply methylene blue and methylene green on LCM samples for *in situ* and invasive regions. However, even with these changes we know that we will lose the very specific and exact spatial resolution we currently have with TSCS, since locations will become groups of cells instead of individual cells. The Navin lab plans to examine consecutive sections to flow sort sufficient cells from specific ducts. Because sorting is dependent on ploidy, it is necessary to evaluate the ploidy of the tumor before any microdissection occurs, to increase the likelihood that sorting will work.

We have done some calculations of the requirements for using consecutive sections.

Based on our estimates, we expect to require  $10^5$  cells after filtering for a good flow sort (so we require at least  $1.5 \times 10^5$  before filtering based on an estimated 30% loss doing nanopore filtering). We estimate cells to be 12 microns in diameter (about 2x the size of a lymphocyte) and expect, based on pathology, cells would not be tightly packed, making up about 75% of the selected area. We calculated the volume for each cell as 16 cubic microns. Using this estimate, if we are cutting 50-micron thick layers, we estimate we would get about three 16-micron layers, and for each 16-micron thick layer, we expect several potentially usable tumor sectors. Therefore, we expect to use 1mm cubes (1000 microns), to cut into 16-micron cubes, giving 62 16-micron cubes. We expect to see 3,844 cells in a 16-micron 1mm layer and 11,532 from a 50-micron section. Therefore, we would need at least 13 sections of 50 microns thick which are at least 1mm square, after microdissection for each region of interest.

Since the ducts and invasive regions will only make up so much space in each section, we expect to need even more sections. However, it is possible to record the size of the areas being microdissected, and I suggest adding these regions for the *in situ* and invasive regions

separately when determining if sufficient cells have been collected. The result of this will be very broad spatial information of regions within 0.65 mm to 1mm cubic regions.

We expect a duct to be mappable through multiple sections and, for invasive regions, mappable either using the duct networks to break up the regions or by breaking the section into quadrants. While this would increase the number of 50-micon layers required, it allows more complete tracing of clones and ducts, with enough power to examine the changes in frequency of subclones between ducts or invasive regions, allowing examination of regional effects more thoroughly.

While some scalability is possible, much of the unique precise location information of TSCS will be lost if we pair LCM with a flow sorting method, since the only spatial information left will be very broad 0.65mm or larger cubic regions.

#### **4.1.2 Technical Barrier: DNA Mutations**

A major advantage sought in LCM methods is one in which nuclei are damaged less. When we first started this project, I tested 3 different LCM machines: Arcturus, Lecia and Zeiss.

The Arcturus system is a touch-based UV and Infrared (IR) laser system. When we tested it, we were able to isolate single cells and amplify using both DOP-PCR and MDA-PCR methods. We think the two reasons both methods worked were (1) the IR is gentler on the DNA and does not cause double stranded breaks and (2) the touch-based system might be unintentionally capturing more than one cell.

The Lecia system by comparison uses a UV laser and gravity to isolate microdissected sections. In the Lecia system, the cell or region of interest is separated from the mass by UV laser and the stage is vibrated slightly to cause the membrane with the cell(s) of interest to detach and fall into the tubes or caps. In theory, this system should work very well for isolation of single cells; however, when the cells fall, they have so much air resistance that they usually float more than fall. This causes them to not always fall directly into the tubes or caps below. We only tested this system once and it appeared that none of the captured cells were able to

be amplified. We think this was because (1) the system was out on a bench and therefore affected by drafts and (2) the system had been unused for some time and might not have been correctly aligned. Since we only tested this once, we cannot determine for sure what issues kept the collection from working.

The last system was the Zeiss PALM system, used for this project. The Zeiss system uses one UV laser to do fine cuts and a different setting of the same laser to catapult the cells into a cap above the system. This set up reduces the chances of contamination (one of the major issues with the Arcturus system), but it can cause fragmentation of the DNA. To reduce the chances of fragmentation, we tested the lowest energy useable to consistently cut and catapult the selected cells.

One of the major issues with mutation data from microdissected single cells is that MDA-PCR relies on long DNA strands. If the DNA is too short, then the displacement does not work to amplify the strand. The Zeiss system appears to cause the DNA to fragment more than the Arcturus system. However, since contamination is such an issue with the Arcturus system, we tried varying the energy levels of the UV laser when catapulting the cells with the Zeiss system. However, even with lower energy settings, the nuclei seem to be slightly damaged and produce shorter fragments than those captured by the Arcturus System.

An interesting next step would be to test collecting and flow sorting a group of cells as described in the consecutive sections discussion. This should allow filtering damaged cells and allow more cells to be collected at one time from adjacent tumor sections therefore increasing throughput. The Navin lab is currently working on a method to provide spatially resolved single cell DNA mutation sequencing with higher throughput and retaining general location information.

Another major issue with collection of single cells from tissues for mutation analysis is related to fixation methods to prevent degradation. Previous experiments in the Navin lab have demonstrated the MDA-PCR reaction to be severely inhibited by even the tiniest bit of fixatives

such as ethanol or methanol. This is an unresolved issue for spatially resolved SCS with mutation calling.

Current technology for spatially resolved single cell mutation calling is specific-to-allele PCR–FISH (STAR-FISH)<sup>108</sup>. STAR-FISH uses competitive probes to measure point mutations in thousands of single cells directly in tissue sections by *in situ* hybridization. STAR-FISH has been used to examine single cell alterations in *Her2* breast cancer when treated with trastuzumab. Currently STAR-FISH is limited by the number of mutations examined at one time, usually 2-3, with increases being limited by fluorophore combinations, a far cry from spatially resolved whole transcriptome methods like FISSEQ<sup>5</sup>, which uses fluorophore labeled nucleotides for spatially resolved sequencing, but is restricted to RNA.

#### **4.1.3 Technical Barrier: TSCS and RNA**

One of the next major steps is to expand from DNA to RNA with a TSCS-like method. Understanding the RNA expression can provide significant information about tumor functions, cell-cell interactions, and phenotypic heterogeneity, making RNA examination desirable. Encouragingly, Geo-Seq<sup>12</sup> was able to combined LCM with single cell RNA sequencing to compare single cell and bulk transcriptome profiles of motor neurons and dopamine neurons, suggesting technical feasibility for RNA profiling<sup>14</sup>. However, one of the major issues with this is timing. RNA unprotected by ribosomes degrades very quickly. Therefore, to examine RNA expression, it is necessary to keep the samples as cold as possible for as long as possible to prevent the degradation.

We have not tried to use single cell collection of RNA from LCM slides. However, one of the major issues is the time required for collection and staining. The collection via Zeiss LCM machine occurs at room temperature. To collect ~24 cells using the TSCS method, the slide would be at room temperature for over 4 hours. Methods to increase throughput will extend time spent at room temperature and might increase RNA degradation. For example, the LCM collection of enough regions to flow sort requires about ~5 days of LCM collection, with the



LCM running for 8 or more hours at a time, meaning that the slides would be at room temperature for at least 40 hours.

While there are methods like Fluorescent *in situ* sequencing (FISSEQ)<sup>5</sup> and Geo-Seq<sup>12</sup> that work on single cell or small cellular clumps, these methods have drawbacks.

Perhaps the most spatially resolved high throughput sequencing method currently is FISSEQ<sup>5</sup>. FISSEQ allows one to examine the full transcriptome of single cells *in situ*, allowing spatial visualizations of gene expression. In theory, FISSEQ allows visualization of not just the RNA but where the RNA is within a cell by sequencing the RNA directly from microscope images. Random hexamer-primer reverse transcriptase is used to transform cellular RNA into short cDNAs, which have an additional sequence adapter to allow the RNAs to be converted into loops, and then amplified *in situ* using rolling circle amplification. To generate a stable matrix, the amplicons are crosslinked for NGS, allowing the amplicons to be like a cluster in NGS, while retaining the RNA molecules original location. A major issue is a lack of consistency across different cellular subunits and across different cell types. However, this technology has the possibility of producing highly-spatially resolved RNA transcriptomes. While improvements to the FISSEQ protocol for RNA spatial sequencing are underway (such as BaristaSeq<sup>233</sup>), it would require a major effort to alter a FISSEQ-like protocol to examine spatially resolved whole genome DNA alterations.

Geo-Seq<sup>12</sup> by comparison is a medium range spatially resolved method of RNA sequencing. The Geo-Seq papers suggest the possibility of SCS using Geo-Seq<sup>12, 13</sup>; however, most of the data currently available is from clumps of 10 to 20 cells. To map the cells back to their location, each clump is given a "zip code" identity. This zip code is used to track the expression of any gene within the clump in an illustrative way, including corn plots to show location and expression levels. Expanding this method to select single cells will be dependent on the LCM technology used. Geo-Seq design occurred on the Leica system but should be transferable to the Zeiss system.

Just as with TSCS, the distance between two or more cells will have a large impact on the need for single cells or clumps of cells. Both Geo-Seq<sup>12, 13</sup> and TSCS methods encounter technical issues such as UV lasers damaging DNA and RNA prior to amplification or cells cut in half during tissue sectioning via cryomicrotome leading to a substantial loss of DNA or RNA. While it would be interesting to expand the TSCS method to RNA, it is probably more practical to use a FISSEQ<sup>5</sup> method and to improve the sequencing consistency across cell types for RNA.

#### **4.1.4 Technical Barrier: TSCS and Genome and Transcriptome Protocols**

The most major and perhaps important improvement to the TSCS protocol would be concurrent Genome and Transcriptome (G&T) protocols<sup>250</sup>, something currently not feasible in FISSEQ<sup>5</sup>, which is limited to RNA. A G&T TSCS protocol would allow examination of morphology, spatial location, DNA-mutation or copy number, and RNA transcriptome. While G&T methods for SCS do exist, they are not commonly used because they are more difficult and time consuming and produce slightly lower quality data<sup>250</sup>. Pairing of G&T single cell protocols with TSCS would be a large advancement and allow connections between genomics and transcriptomics as well as genomics to morphology of a single cell. Since the genome and transcriptome data from SCS-G&T are from the same cell, the data intergradation is more streamlined than for bulk multiple platform studies like TCGA<sup>101</sup>. In these bulk multiple platform studies, DNA and RNA isolation was from separate regions and direct relationships between genetics and expression was inhibited by the purity of the sample, allowing for only association between genetics and expression to be observed.

I think that the best "next step" for TSCS protocol improvements is to expand to G&T methods, which provide a unified analysis of the genome, transcriptome, location, and morphology of a single cell. Implementation in tissue sections would be extremely technically challenging and would likely require over collection of cells, double or more, in order to have enough cells to pass the quality control steps.

#### **4.1.5 Technical Barrier: Spatial Genomics and Clinical Tools**

Spatial genomics is currently a hot topic in science, and the effects of ITH could be crucial to personalized medicine. In the case of invasion and progression of cancer, we often want to consider if the finding (1) could affect treatment for a specific subset of patients and (2) if so how our findings translate into a clinical test.

##### **4.1.5.1 Clinical Impact of Single Cells of Origin**

Our findings indicate *in situ* and invasive tumors arose from the same cell of origin, suggesting that finding common targetable mutations could allow a single treatment for both *in situ* and invasive regions. The single cell of origin hypothesis provides fundamental support that treating common mutations is probably the best way to eliminate the tumor, by targeting the foundation or truncal mutations instead of rare subclone specific mutations.

##### **4.1.5.2 Clinical Impact of Multiple Clonal Invasion**

Multiclonal invasion goes a step further than the single cell of origin or direct lineage hypothesis. Multiclonal invasion allows a single tumor cell to initiate tumorigenesis, but integral to multiclonal invasion is the concept that all subclones can escape the duct during invasion. The clinical impact of this will depend on how invasion occurred.

The two scenarios we discussed earlier had different implications for clinical research and treatment.

If the clones cooperate to escape the ducts, then inhibiting cooperative interactions could be a clinical target. The cooperative scenarios use different mechanisms, each suggests a different prospective on how to prevent invasion. For example, if secreted factors caused the breakdown of the basal membrane, using antibodies or drugs that can interfere with the secretion is a possible treatment. Alternatively, in a leader clone scenario, it becomes necessary to either understand which clone is the "leader clone" to provide targeted intervention or to target alterations that occur in all clones. However, to gain such knowledge about each patient tumor would require mechanistic knowledge currently unavailable and

would probably be patient specific. This would require detailed studies using *in vitro* or *in vivo* systems, such as xenografts, for each patient which is impractical, expensive, and too time-consuming to usefully direct therapy.

Further understanding of the mechanisms of multiclonal invasion are needed and require the use of further functional studies using *in vivo* systems.

#### **4.1.5.3 Clinical Impact of Punctuated Evolution**

There are two well-known models of punctuated evolution: the branching expansions<sup>251, 16</sup> and the big bang models<sup>244</sup>. Our data supports the model of punctuated evolution via branching expansions for polyclonal tumors and clonal sweeps for monoclonal tumors. Branching expansions tumors and clonal sweep tumors should have a set of common alterations which should be examined to treatment targetability. If targetable alterations are not found in these truncal alterations then these tumors should be treated similarly to big-bang tumors. Big bang tumors have a critical event that leads to tumor progression and multiple diverse clones, which should share very few mutations, and therefore require broader-based treatment such as radiation or chemotherapy.

#### **4.1.5.4 Clinical Impact of Intratumor Heterogeneity**

Many papers have postulated that increased clonal diversity or ITH will be linked to resistance and progression of tumors<sup>252-254</sup>. Current studies have shown that, similar to bacteria<sup>255</sup>, resistant clones are more likely to be present before treatment and arise from the original clonal subpopulations due to the selective pressures of treatment<sup>256</sup>. This model of progression favors a stochastic or neutral model of evolution, where more unique alterations correlate with a higher probability of invasion. High diversity indexes correlated with potential to metastasize or with poor response to therapy<sup>107, 194, 127</sup>. However, ITH measurements such as using diversity indexes<sup>107, 16</sup> have not yet been proven to be good prognostic markers.

If this is the case, a low genomic diversity is expected to correlate with a lower risk of invasion in DCIS patients. While overall diversity might be correlated with progression, it is also

possible for there to be mutational diversity, which is greater than that observed by copy number markers. These mutations might be responsible for resistance and progression and add another layer of genetic diversity. This mutational diversity could be hidden within what appears to be a homogeneous tumor<sup>200, 257</sup>, if diversity is calculated based on copy number subclones. This genetic mutational diversity might be enough to preserve tumor cells under the pressure of an anaerobic environment or chemotherapy, resulting in the eradication of non-resistant clones and an expansion of the "hidden" but resistant clones.

#### **4.1.5.5 Technical Barrier of Clinical Tools**

When creating clinical tools, some aspects to consider are robustness of results, speed, expense, and analysis complexity. Currently, SCS methods are not robust. SCS methods are strongly affected by sampling bias and require significantly more time for meticulous data analysis in addition to the expensive of sequencing multiple cells. This makes translating SCS to the clinical environment impractical at the current time. This is especially true of TSCS which requires multiple days for collection. Methods like TSCS, while a powerful tool for research, are currently impractical for translational work.

However, there is some promise for clinical tools in the future, with the advent of powerful new multiplexing single cell methods like 10xGenomics, which allows for the evaluation of 10,000 or more single cells at a time<sup>258</sup>. Topographical whole genome RNA methods like FISSEQ<sup>5</sup> are limited in consistency and robustness, but ReadCor is working to expand FISSEQ and possibly make evaluating single cell spatial RNA data in the clinic possible.

## **4.2 Future Research Directions**

TSCS and other spatially resolved SCS methods have many uses in the biological sciences. A natural extension is to examine other early cancer CNA, mutations, and RNA expression.

#### 4.2.1 ITH in Pure DCIS and Earlier Cancers

In this thesis I primarily discuss the single cell CNA observed in DCIS-IDC. However, a natural extension of this research is examining even earlier stages of breast cancer for patterns of progression on the single cell level. It is possible to use TSCS on earlier stages of breast cancers like ADH and pure DCIS, because TSCS requires far less tissue and cells than other isolation methods like flow sorting.

A hypothesis frequently heard is that more diverse tumors are more likely to progress. TSCS is well suited to examining breast cancers or other aneuploidy driven cancers. Two basic experiments with TSCS to examine possible relationships between CNA diversity and cancer progression are described below.

First is directly comparing clonal CNA diversity of pure DCIS and DCIS-IDC. This would be best with tumors which matched grade and receptor statuses to reduce possible confounding factors. By comparing the clonal CNA diversity of pure DCIS and DCIS-IDC, we can appraise clonal diversity at the two-time points to determine if clonal diversity is a driving factor in progression. If clonal diversity is a driving factor, we expect the DCIS-IDC tumors to have higher clonal diversity than the pure DCIS tumors. However, a limit on this experiment could be a lack of pairing of samples (pure DCIS and DCIS-IDC) and the confounding possibility that "pure" DCIS might progress to DCIS-IDC at some point. The first issue cannot be address without a longitudinal study to would allow pairing of samples. The second issue is addressed by making sure pure DCIS has not progressed within a specific timeframe, for example 5 years for high grade DCIS. Therefore, if CNA clonal diversity was a key feature of progression differences between pure DCIS and DCIS-IDC should be observed.

Second is to examine the progression of pure DCIS to DCIS-IDC. While this might be considered impractical due to the large number of samples needed to get 10 DCIS-IDC samples (Figure 5 Estimates of Number of Samples Needs for Longitudinal Studies), this is the most direct experiment. By examining a large cohort of pure DCIS patients, determining their CNA diversity, following these patients for 15+ years, and then comparing CNA diversity

between patients that progressed and those that did not, this experiment would answer the question of any relationship between diversity and progression the most directly. However, as discussed above in 2.1 Sample Selection, longitudinal studies are considered impractical due to the large number of patients required.

Alternatively, TSCS can easily be utilized to investigate similar questions about the nature of cancer progression in other early cancers. These early stage progression studies best suited for TSCS would require the early cancers to have well-defined histopathologies. TSCS works best when visualizing the difference between tumor and normal cells is easier. Future directions to using TSCS in early cancer, as we discussed in our review paper<sup>2</sup> can be in cancer types like colorectal adenomas, lobular carcinoma *in situ* (LCIS), prostatic intraepithelial neoplasias (PIN), and pancreatic intraepithelial neoplasias (PanINs).

#### **4.2.2 Examining Spatial and DNA Mutations**

For DNA mutations, we can examine the possible progression of gradual accumulations of mutations, as seen in colon cancer<sup>257, 78</sup>. SCS theoretically can separate driver mutations from passenger mutations based on the frequency of mutations within single cells. With spatially resolved SCS, we can take this a step further, and examine mutation differences between the *in situ* and invasive regions of other cancers.

With a TSCS-like method, one can look at the shape and placement of the single cell to select a cell for sequencing. Therefore, one can examine morphology changes associated with subclones or mutations. While, this type of method would reduce the number of cells captured per region, it does allow more spatial and morphologic data to be collected for each cell. Therefore, there will be a trade off in increasing the throughput or increasing the spatial and morphology information.

However, it seems feasible that one would prefer to use a high throughput spatially resolved single cell method that would allow one to sort the diploid and aneuploid cells, estimate the frequency of different mutations, and then use a different type of method, like

STAR-FISH<sup>108</sup>, to examine morphology and spatial intermixtures of clones. The most notable limitations of STAR-FISH are the requirement of prior knowledge of the mutation of interest and the limited number of genes measured in a single section.

Other possible uses for this technique include other cancer research, developmental science, neuroscience, and exploring mosaicism. In other cancer research, we often encounter one type of cancer cell morphology adjacent to another (such as with the morphology of micropapillary breast cancer). Using TSCS, we can take a closer look at the genetic alterations of copy number changes between distinct neighboring cancer morphologies. Methods like FISSEQ<sup>5</sup> could provide fast, morphologic, spatial, and transcriptomic data, but it cannot provide the mutational data so important to cancer analysis and mosaicism.

#### **4.2.3 Profiling Geographic Heterogeneity**

We developed TSCS to examine geographic ITH and to specifically examine ITH in synchronous cancers where *in situ* and invasive regions are closely paired to understand invasion better. Diverse expression of proteins<sup>15</sup>, RNA expression, DNA mutations<sup>108, 8, 17</sup> and CNA<sup>16, 17</sup> have been observed in breast cancer multiple times. With single cell DNA sequencing, regional heterogeneity and the spatial relationships of ITH became an important avenue of research, providing insight to tumor evolution and invasion.

However, the importance of geographic heterogeneity goes beyond cancer. Spatial copy number heterogeneity has been observed in neurons and other brain tissue, suggesting DNA copy number and mutations might play an important role in organ development. Another developmental question concerns background mosaicism in any given tissue. Understanding mosaicism is important for both development and cancer, providing empirical evaluations of the levels of background mutations to help provide better estimates about the accumulation of mutations during development and aging.

TSCS and other spatially-resolved SCS methods obviously hold great potential for new avenues of investigation in early stage cancers. For more advanced cancers with large invasive



components or where metastasis spread throughout the body, micron-level resolution of spatial genomics is less important than in early stage cancers. By studying the spatially resolved genomics in early cancer, we can examine tumor initiation and invasion, giving us more insight into genomic alterations needed for each cancer to progress. By first studying early cancers with well-defined histopathologies, such as colorectal adenomas, LCIS, PIN, and PanINs, we can examine the spatial effects of ITH.

In these early cancers, spatial resolution can provide new insights into the context, organization, and migration of tumor clones as they escape the basement membranes and invade the surrounding tissues. Spatially resolved profiling methods of DNA, RNA and proteins can examine why some premalignant cancers remain indolent for the lifetime of the patient, while others progress to invasive disease and ultimately cause morbidity in patients.

These concepts cross many fields, as witness by the development of use of Geo-Seq first in developmental biology, while LCM-Seq and FISSEQ were both developed for neurology.

### **4.3 Closing Remarks**

My closing remarks address issues related specifically to synchronous DCIS-IDC and the technology involved with making spatially resolved SCS possible.

#### **4.3.1 Synchronous DCIS-IDC**

Collectively, genomic studies of synchronous DCIS-IDC patients suggest an independent lineage model is uncommon, since most synchronous DCIS patients share many concordant CNAs and mutations. Instead genomic data support a direct lineage model and indicate evolution of both DCIS and IDC subpopulations from a common origin. Up until now, the concordance and discordance of mutations in DCIS and IDC regions of synchronous patients have provided circumstantial evidence for distinguishing between the different direct lineage hypothesis<sup>114, 163, 86, 84, 119, 165, 144, 166, 85</sup>, *evolutionary bottleneck* and *multiclonal invasion*.

Without SCS methods, it has been difficult to distinguish between *evolutionary bottleneck* and *multiclonal invasion* models of the direct lineage hypothesis. Our spatially

resolved SCS data provided insight into the progression of DCIS to IDC and demonstrated that all clones escaped the ducts and expanded to invasive regions. This data strongly suggests a model of *multiclonal invasion*. While more data is necessary to confirm multiclonal invasion, recently developed spatially resolved DNA, RNA, and protein technologies should help overcome technical obstacles such as tumor purity. These methods are likely to have important applications for studying direct models of invasion, to determine if invasive subclones have any distinguishing mutations<sup>108</sup>.

#### **4.3.2 Technology**

Technology development often occurs because current technologies are unable to provide the resolution of detail needed to fully examine an issue of interest. Several of the innovative technologies I have highlighted here were developed for other research questions but have the potential to provide new insights into cancer invasion. While STAR-FISH<sup>108</sup> and Solid-phase Imaging Mass Cytometry (IMC)<sup>235</sup> were developed for cancer; Geo-Seq was developed to examine RNA transcription changes during development<sup>12, 13</sup> and LCM-Seq<sup>14</sup> and FISSEQ was developed to examine RNA transcription changes in the brain<sup>5</sup>.

Multiple improvements to FISSEQ have been developed to increase the throughput and improve output per cell. For example BaristaSeq<sup>233</sup>, which was developed in baby hamster kidney (BHK) cells, was developed to barcode single cells increase data quality and resolution. In the BaristaSeq, however, they suggest that this technique could be used to map neuronal projections or lineages.

All have the potential to enhance our studies of spatially resolved genomes and transcriptomes.

## BIBLIOGRAPHY

### References

1. Casasent, A. K., A. Schalck, R. Gao, E. Sei, A. Long, W. Pangburn, T. Casasent, F. Meric-Bernstam, M. E. Edgerton, and N. E. Navin. 2018. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* 172: 205-217 e212.
2. Casasent, A. K., M. Edgerton, and N. E. Navin. 2017. Genome evolution in ductal carcinoma *in situ*: invasion of the clones. *J Pathol* 241: 208-218.
3. Russnes, H. G., N. Navin, J. Hicks, and A. L. Borresen-Dale. 2011. Insight into the heterogeneity of breast cancer through next-generation sequencing. *J Clin Invest* 121: 3810-3818.
4. Navin, N., J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* 472: 90-94.
5. Lee, J. H., E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, B. M. Turczyk, J. L. Yang, H. S. Lee, J. Aach, K. Zhang, and G. M. Church. 2015. Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc* 10: 442-458.
6. Martelotto, L. G., T. Baslan, J. Kendall, F. C. Geyer, K. A. Burke, L. Spraggon, S. Piscuoglio, K. Chadavada, G. Nanjangud, C. K. Ng, P. Moody, S. D'Italia, L. Rodgers, H. Cox, A. da Cruz Paula, A. Stepansky, M. Schizas, H. Y. Wen, T. A. King, L. Norton, B. Weigelt, J. B. Hicks, and J. S. Reis-Filho. 2017. Whole-genome single-cell copy number profiling from formalin-fixed paraffin-embedded samples. *Nat Med* 23: 376-385.
7. Baslan, T., J. Kendall, L. Rodgers, H. Cox, M. Riggs, A. Stepansky, J. Troge, K. Ravi, D. Esposito, B. Lakshmi, M. Wigler, N. Navin, and J. Hicks. 2012. Genome-wide copy number analysis of single cells. *Nat Protoc* 7: 1024-1041.
8. Navin, N., A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor, D. Levy, P. Lundin, S. Maner, A. Zetterberg, J. Hicks, and M. Wigler. 2010. Inferring tumor progression from genomic heterogeneity. *Genome Res* 20: 68-80.
9. Navin, N. E., and J. Hicks. 2010. Tracing the tumor lineage. *Mol Oncol* 4: 267-283.
10. Wang, Y., and N. E. Navin. 2015. Advances and applications of single-cell sequencing technologies. *Mol Cell* 58: 598-609.
11. Rudkin, G. T., and B. D. Stollar. 1977. High resolution detection of DNA-RNA hybrids *in situ* by indirect immunofluorescence. *Nature* 265: 472-473.
12. Chen, J., S. Suo, P. P. Tam, J. J. Han, G. Peng, and N. Jing. 2017. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat Protoc* 12: 566-580.
13. Peng, G., S. Suo, J. Chen, W. Chen, C. Liu, F. Yu, R. Wang, S. Chen, N. Sun, G. Cui, L. Song, P. P. Tam, J. D. Han, and N. Jing. 2016. Spatial Transcriptome for the Molecular Annotation of Lineage Fates and Cell Identity in Mid-gastrula Mouse Embryo. *Dev Cell* 36: 681-697.
14. Nichterwitz, S., G. Chen, J. Aguila Benitez, M. Yilmaz, H. Storrval, M. Cao, R. Sandberg, Q. Deng, and E. Hedlund. 2016. Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. *Nat Commun* 7: 12139.

15. Allred, D. C., Y. Wu, S. Mao, I. D. Nagtegaal, S. Lee, C. M. Perou, S. K. Mohsin, P. O'Connell, A. Tsimelzon, and D. Medina. 2008. Ductal carcinoma *in situ* and the emergence of diversity during breast cancer evolution. *Clin Cancer Res* 14: 370-378.
16. Gao, R., A. Davis, T. O. McDonald, E. Sei, X. Shi, Y. Wang, P. C. Tsai, A. Casasent, J. Waters, H. Zhang, F. Meric-Bernstam, F. Michor, and N. E. Navin. 2016. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet*.
17. Wang, Y., J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, A. Multani, H. Zhang, R. Zhao, F. Michor, F. Meric-Bernstam, and N. E. Navin. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512: 155-160.
18. Turashvili, G., and E. Brogi. 2017. Tumor Heterogeneity in Breast Cancer. *Front Med (Lausanne)* 4: 227.
19. 2018. U.S. Breast Cancer Statistics breastcancer.org.
20. Siegel, R. L., K. D. Miller, and A. Jemal. 2018. Cancer statistics, 2018. *CA Cancer J Clin* 68: 7-30.
21. Hanahan, D., and R. A. Weinberg. 2000. The hallmarks of cancer. *Cell* 100: 57-70.
22. Hanahan, D., and R. A. Weinberg. 2011. Hallmarks of cancer: the next generation. *Cell* 144: 646-674.
23. Kwei, K. A., Y. Kung, K. Salari, I. N. Holcomb, and J. R. Pollack. 2010. Genomic instability in breast cancer: pathogenesis and clinical implications. *Mol Oncol* 4: 255-266.
24. Guirouilh-Barbat, J. K., T. Wilhelm, and B. S. Lopez. 2010. AKT1/BRCA1 in the control of homologous recombination and genetic stability: the missing link between hereditary and sporadic breast cancers. *Oncotarget* 1: 691-699.
25. Gerashchenko, B. I., A. Huna, and J. Erenpreisa. 2014. Characterization of breast cancer DNA content profiles as a prognostic tool. *Exp Oncol* 36: 219-225.
26. Allred, D. C. 2010. Ductal carcinoma *in situ*: terminology, classification, and natural history. *J Natl Cancer Inst Monogr* 2010: 134-138.
27. Fearon, E. R., and B. Vogelstein. 1990. A genetic model for colorectal tumorigenesis. *Cell* 61: 759-767.
28. Alvarado, M., D. L. Carter, J. M. Guenther, J. Hagans, R. Y. Lei, C. E. Leonard, J. Manders, A. P. Sing, M. S. Broder, D. Cherepanov, E. Chang, M. Eagan, W. Hsiao, and M. J. Schultz. 2015. The impact of genomic testing on the recommendation for radiation therapy in patients with ductal carcinoma *in situ*: A prospective clinical utility assessment of the 12-gene DCIS score result. *J Surg Oncol* 111: 935-940.
29. Baxter, N. N., B. A. Virnig, S. B. Durham, and T. M. Tuttle. 2004. Trends in the treatment of ductal carcinoma *in situ* of the breast. *J Natl Cancer Inst* 96: 443-448.
30. Ernster, V. L., and J. Barclay. 1997. Increases in ductal carcinoma *in situ* (DCIS) of the breast in relation to mammography: a dilemma. *J Natl Cancer Inst Monogr*: 151-156.
31. Jones, J. L. 2006. Overdiagnosis and overtreatment of breast cancer: progression of ductal carcinoma *in situ*: the pathological perspective. *Breast Cancer Res* 8: 204.
32. Zujewski, J. A., L. C. Harlan, D. M. Morrell, and J. L. Stevens. 2011. Ductal carcinoma *in situ*: trends in treatment over time in the US. *Breast Cancer Res Treat* 127: 251-257.
33. Bartlett, J. M., S. Nofech-Moses, and E. Rakovitch. 2014. Ductal carcinoma *in situ* of the breast: can biomarkers improve current management? *Clin Chem* 60: 60-67.
34. Leonard, G. D., and S. M. Swain. 2004. Ductal carcinoma *in situ*, complexities and challenges. *J Natl Cancer Inst* 96: 906-920.
35. Thompson, A., K. Brennan, A. Cox, J. Gee, D. Harcourt, A. Harris, M. Harvie, I. Holen, A. Howell, R. Nicholson, M. Steel, and C. Streuli. 2008. Evaluation of the current

- knowledge limitations in breast cancer research: a gap analysis. *Breast Cancer Res* 10: R26.
36. Pinder, S. E., C. Duggan, I. O. Ellis, J. Cuzick, J. F. Forbes, H. Bishop, I. S. Fentiman, W. D. George, and U. K. C. C. o. C. R. D. C. I. S. W. Party. 2010. A new pathological system for grading DCIS with improved prediction of local recurrence: results from the UKCCCR/ANZ DCIS trial. *Br J Cancer* 103: 94-100.
  37. Arpino, G., V. J. Bardou, G. M. Clark, and R. M. Elledge. 2004. Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome. *Breast Cancer Res* 6: R149-156.
  38. Vo, T., Y. Xing, F. Meric-Bernstam, N. Mirza, G. Vlastos, W. F. Symmans, G. H. Perkins, T. A. Buchholz, G. V. Babiera, H. M. Kuerer, I. Bedrosian, J. S. Akins, and K. K. Hunt. 2007. Long-term outcomes in patients with mucinous, medullary, tubular, and invasive ductal carcinomas after lumpectomy. *Am J Surg* 194: 527-531.
  39. Rakha, E. A., A. H. Lee, A. J. Evans, S. Menon, N. Y. Assad, Z. Hodi, D. Macmillan, R. W. Blamey, and I. O. Ellis. 2010. Tubular carcinoma of the breast: further evidence to support its excellent prognosis. *J Clin Oncol* 28: 99-104.
  40. Dawood, S., and M. Cristofanilli. 2011. Inflammatory breast cancer: what progress have we made? *Oncology (Williston Park)* 25: 264-270, 273.
  41. Li, J., Y. Xia, Q. Wu, S. Zhu, C. Chen, W. Yang, W. Wei, and S. Sun. 2017. Outcomes of patients with inflammatory breast cancer by hormone receptor- and HER2-defined molecular subtypes: A population-based study from the SEER program. *Oncotarget* 8: 49370-49379.
  42. Horlings, H. M., B. Weigelt, E. M. Anderson, M. B. Lambros, A. Mackay, R. Natrajan, C. K. Ng, F. C. Geyer, M. J. van de Vijver, and J. S. Reis-Filho. 2013. Genomic profiling of histological special types of breast cancer. *Breast Cancer Res Treat* 142: 257-269.
  43. Weigelt, B., F. C. Geyer, and J. S. Reis-Filho. 2010. Histological types of breast cancer: how special are they? *Mol Oncol* 4: 192-208.
  44. Harris, L., H. Fritzsche, R. Mennel, L. Norton, P. Ravdin, S. Taube, M. R. Somerfield, D. F. Hayes, R. C. Bast, Jr., and O. American Society of Clinical. 2007. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol* 25: 5287-5312.
  45. Pinder, S. E. 2010. Ductal carcinoma *in situ* (DCIS): pathological features, differential diagnosis, prognostic factors and specimen evaluation. *Mod Pathol* 23 Suppl 2: S8-13.
  46. Bauer, K., C. Parise, and V. Caggiano. 2010. Use of ER/PR/HER2 subtypes in conjunction with the 2007 St Gallen Consensus Statement for early breast cancer. *BMC Cancer* 10: 228.
  47. Paik, S., S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 351: 2817-2826.
  48. Kennecke, H., R. Yerushalmi, R. Woods, M. C. Cheang, D. Voduc, C. H. Speers, T. O. Nielsen, and K. Gelmon. 2010. Metastatic behavior of breast cancer subtypes. *J Clin Oncol* 28: 3271-3277.
  49. McGuire, A., O. Kalinina, E. Holian, C. Curran, C. A. Malone, R. McLaughlin, A. Lowery, J. A. L. Brown, and M. J. Kerin. 2017. Differential impact of hormone receptor status on survival and recurrence for HER2 receptor-positive breast cancers treated with Trastuzumab. *Breast Cancer Res Treat* 164: 221-229.

50. Perez, E. A., E. H. Romond, V. J. Suman, J. H. Jeong, N. E. Davidson, C. E. Geyer, Jr., S. Martino, E. P. Mamounas, P. A. Kaufman, and N. Wolmark. 2011. Four-year follow-up of trastuzumab plus adjuvant chemotherapy for operable human epidermal growth factor receptor 2-positive breast cancer: joint analysis of data from NCCTG N9831 and NSABP B-31. *J Clin Oncol* 29: 3366-3373.
51. Perez, E. A., E. H. Romond, V. J. Suman, J. H. Jeong, G. Sledge, C. E. Geyer, Jr., S. Martino, P. Rastogi, J. Gralow, S. M. Swain, E. P. Winer, G. Colon-Otero, N. E. Davidson, E. Mamounas, J. A. Zujewski, and N. Wolmark. 2014. Trastuzumab plus adjuvant chemotherapy for human epidermal growth factor receptor 2-positive breast cancer: planned joint analysis of overall survival from NSABP B-31 and NCCTG N9831. *J Clin Oncol* 32: 3744-3752.
52. Voduc, K. D., M. C. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen, and H. Kennecke. 2010. Breast cancer subtypes and the risk of local and regional relapse. *J Clin Oncol* 28: 1684-1691.
53. Bergamaschi, A., Y. H. Kim, P. Wang, T. Sorlie, T. Hernandez-Boussard, P. E. Lonning, R. Tibshirani, A. L. Borresen-Dale, and J. R. Pollack. 2006. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes Cancer* 45: 1033-1040.
54. Chin, K., S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W. L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. M. Ljung, L. Esserman, D. G. Albertson, F. M. Waldman, and J. W. Gray. 2006. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 10: 529-541.
55. Bauer, K. R., M. Brown, R. D. Cress, C. A. Parise, and V. Caggiano. 2007. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. *Cancer* 109: 1721-1728.
56. Muller, K. E., J. D. Marotti, V. A. Memoli, W. A. Wells, and L. J. Tafe. 2015. Impact of the 2013 ASCO/CAP HER2 Guideline Updates at an Academic Medical Center That Performs Primary HER2 FISH Testing: Increase in Equivocal Results and Utility of Reflex Immunohistochemistry. *Am J Clin Pathol* 144: 247-252.
57. Perou, C. M. 2010. Molecular stratification of triple-negative breast cancers. *Oncologist* 15 Suppl 5: 39-48.
58. Perou, C. M. 2011. Molecular stratification of triple-negative breast cancers. *Oncologist* 16 Suppl 1: 61-70.
59. Carey, L. A., C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston, S. L. Deming, J. Geradts, M. C. Cheang, T. O. Nielsen, P. G. Moorman, H. S. Earp, and R. C. Millikan. 2006. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* 295: 2492-2502.
60. Simpson, P. T., J. S. Reis-Filho, T. Gale, and S. R. Lakhani. 2005. Molecular evolution of breast cancer. *J Pathol* 205: 248-254.
61. Simpson, P. T., J. S. Reis-Filho, and S. R. Lakhani. 2010. Breast pathology: beyond morphology. *Semin Diagn Pathol* 27: 91-96.
62. Buerger, H., E. C. Mommers, R. Littmann, R. Simon, R. Diallo, C. Poremba, B. Dockhorn-Dworniczak, P. J. van Diest, and W. Boecker. 2001. Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. *J Pathol* 194: 165-170.

63. Fujii, H., R. Szumel, C. Marsh, W. Zhou, and E. Gabrielson. 1996. Genetic progression, histological grade, and allelic loss in ductal carcinoma *in situ* of the breast. *Cancer Res* 56: 5260-5265.
64. Roylance, R., P. Gorman, W. Harris, R. Liebmann, D. Barnes, A. Hanby, and D. Sheer. 1999. Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. *Cancer Res* 59: 1433-1436.
65. Simpson, P. T., T. Gale, J. S. Reis-Filho, C. Jones, S. Parry, J. P. Sloane, A. Hanby, S. E. Pinder, A. H. Lee, S. Humphreys, I. O. Ellis, and S. R. Lakhani. 2005. Columnar cell lesions of the breast: the missing link in breast cancer progression? A morphological and molecular analysis. *Am J Surg Pathol* 29: 734-746.
66. Ellis, I. O., M. Galea, N. Broughton, A. Locker, R. W. Blamey, and C. W. Elston. 1992. Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. *Histopathology* 20: 479-489.
67. Ignatiadis, M., and C. Sotiriou. 2008. Understanding the molecular basis of histologic grade. *Pathobiology* 75: 104-111.
68. Rakha, E. A., J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani, J. Palacios, A. L. Richardson, S. J. Schnitt, F. C. Schmitt, P. H. Tan, G. M. Tse, S. Badve, and I. O. Ellis. 2010. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* 12: 207.
69. Carter, C. L., C. Allen, and D. E. Henson. 1989. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* 63: 181-187.
70. Lee, S. H., Y. S. Kim, W. Han, H. S. Ryu, J. M. Chang, N. Cho, and W. K. Moon. 2016. Tumor growth rate of invasive breast cancers during wait times for surgery assessed by ultrasonography. *Medicine (Baltimore)* 95: e4874.
71. Zheng, Y. Z., L. Wang, X. Hu, and Z. M. Shao. 2015. Effect of tumor size on breast cancer-specific survival stratified by joint hormone receptor status in a SEER population-based study. *Oncotarget* 6: 22985-22995.
72. Waldman, F. M., S. DeVries, K. L. Chew, D. H. Moore, 2nd, K. Kerlikowske, and B. M. Ljung. 2000. Chromosomal alterations in ductal carcinomas *in situ* and their *in situ* recurrences. *J Natl Cancer Inst* 92: 313-320.
73. Sakorafas, G. H., and D. R. Farley. 2003. Optimal management of ductal carcinoma *in situ* of the breast. *Surg Oncol* 12: 221-240.
74. Van Cleef, A., S. Altintas, M. Huizing, K. Papadimitriou, P. Van Dam, and W. Tjalma. 2014. Current view on ductal carcinoma *in situ* and importance of the margin thresholds: A review. *Facts Views Vis Obgyn* 6: 210-218.
75. Brandt, J., J. P. Garne, I. Tengrup, and J. Manjer. 2015. Age at diagnosis in relation to survival following breast cancer: a cohort study. *World J Surg Oncol* 13: 33.
76. Collins, L. C., R. M. Tamimi, H. J. Baer, J. L. Connolly, G. A. Colditz, and S. J. Schnitt. 2005. Outcome of patients with ductal carcinoma *in situ* untreated after diagnostic biopsy: results from the Nurses' Health Study. *Cancer* 103: 1778-1784.
77. Sanders, M. E., P. A. Schuyler, W. D. Dupont, and D. L. Page. 2005. The natural history of low-grade ductal carcinoma *in situ* of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* 103: 2481-2484.
78. Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, Jr., and K. W. Kinzler. 2013. Cancer genome landscapes. *Science* 339: 1546-1558.

79. Welch, H. G., and W. C. Black. 1997. Using autopsy series to estimate the disease "reservoir" for ductal carcinoma *in situ* of the breast: how much more breast cancer can we find? *Ann Intern Med* 127: 1023-1028.
80. Afghahi, A., E. Forgo, A. A. Mitani, M. Desai, S. Varma, T. Seto, J. Rigdon, K. C. Jensen, M. L. Troxell, S. L. Gomez, A. K. Das, A. H. Beck, A. W. Kurian, and R. B. West. 2015. Chromosomal copy number alterations for associations of ductal carcinoma *in situ* with invasive breast cancer. *Breast Cancer Res* 17: 108.
81. Buerger, H., F. Otterbach, R. Simon, C. Poremba, R. Diallo, T. Decker, L. Riethdorf, C. Brinkschmidt, B. Dockhorn-Dworniczak, and W. Boecker. 1999. Comparative genomic hybridization of ductal carcinoma *in situ* of the breast-evidence of multiple genetic pathways. *J Pathol* 187: 396-402.
82. Reis-Filho, J. S., and S. R. Lakhani. 2003. The diagnosis and management of pre-invasive breast disease: genetic alterations in pre-invasive lesions. *Breast Cancer Res* 5: 313-319.
83. Shackney, S. E., and J. F. Silverman. 2003. Molecular evolutionary patterns in breast cancer. *Adv Anat Pathol* 10: 278-290.
84. Kim, S. Y., S. H. Jung, M. S. Kim, I. P. Baek, S. H. Lee, T. M. Kim, Y. J. Chung, and S. H. Lee. 2015. Genomic differences between pure ductal carcinoma *in situ* and synchronous ductal carcinoma *in situ* with invasive breast cancer. *Oncotarget* 6: 7597-7607.
85. Yates, L. R., M. Gerstung, S. Knappskog, C. Desmedt, G. Gundem, P. Van Loo, T. Aas, L. B. Alexandrov, D. Larsimont, H. Davies, Y. Li, Y. S. Ju, M. Ramakrishna, H. K. Haugland, P. K. Lilleng, S. Nik-Zainal, S. McLaren, A. Butler, S. Martin, D. Glodzik, A. Menzies, K. Raine, J. Hinton, D. Jones, L. J. Mudie, B. Jiang, D. Vincent, A. Greene-Colozzi, P. Y. Adnet, A. Fatima, M. Maetens, M. Ignatiadis, M. R. Stratton, C. Sotiriou, A. L. Richardson, P. E. Lonning, D. C. Wedge, and P. J. Campbell. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* 21: 751-759.
86. Johnson, C. E., K. L. Gorringer, E. R. Thompson, K. Oakes, S. E. Boyle, Y. Wang, P. Hill, G. B. Mann, and I. G. Campbell. 2012. Identification of copy number alterations associated with the progression of DCIS to invasive ductal carcinoma. *Breast Cancer Res Treat* 133: 889-898.
87. Aubele, M., M. Cummings, A. Walsch, H. Zitzelsberger, J. Nahrig, H. Hofler, and M. Werner. 2000. Heterogeneous chromosomal aberrations in intraductal breast lesions adjacent to invasive carcinoma. *Anal Cell Pathol* 20: 17-24.
88. Friedlander, M. L., D. W. Hedley, and I. W. Taylor. 1984. Clinical and biological significance of aneuploidy in human tumours. *J Clin Pathol* 37: 961-974.
89. Moore, E., H. Magee, J. Coyne, T. Gorey, and P. A. Dervan. 1999. Widespread chromosomal abnormalities in high-grade ductal carcinoma *in situ* of the breast. Comparative genomic hybridization study of pure high-grade DCIS. *J Pathol* 187: 403-409.
90. Aubele, M., A. Mattis, H. Zitzelsberger, A. Walch, M. Kremer, P. Hutzler, H. Hofler, and M. Werner. 1999. Intratumoral heterogeneity in breast carcinoma revealed by laser-microdissection and comparative genomic hybridization. *Cancer Genet Cytogenet* 110: 94-102.
91. Dutrillaux, B., M. Gerbault-Seureau, and B. Zafrani. 1990. Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiplod cases with few chromosome changes. *Cancer Genet Cytogenet* 49: 203-217.



92. Pandis, N., S. Heim, G. Bardi, I. Idvall, N. Mandahl, and F. Mitelman. 1993. Chromosome analysis of 20 breast carcinomas: cytogenetic multiclonality and karyotypic-pathologic correlations. *Genes Chromosomes Cancer* 6: 51-57.
93. Pandis, N., Y. Jin, L. Gorunova, C. Petersson, G. Bardi, I. Idvall, B. Johansson, C. Ingvar, N. Mandahl, F. Mitelman, and et al. 1995. Chromosome analysis of 97 primary breast carcinomas: identification of eight karyotypic subgroups. *Genes Chromosomes Cancer* 12: 173-185.
94. Pandis, N., Y. Jin, J. Limon, G. Bardi, I. Idvall, N. Mandahl, F. Mitelman, and S. Heim. 1993. Interstitial deletion of the short arm of chromosome 3 as a primary chromosome abnormality in carcinomas of the breast. *Genes Chromosomes Cancer* 6: 151-155.
95. Teixeira, M. R., N. Pandis, G. Bardi, J. A. Andersen, and S. Heim. 1996. Karyotypic comparisons of multiple tumorous and macroscopically normal surrounding tissue samples from patients with breast cancer. *Cancer Res* 56: 855-859.
96. Thompson, F., J. Emerson, W. Dalton, J. M. Yang, D. McGee, H. Villar, S. Knox, K. Massey, R. Weinstein, A. Bhattacharyya, and et al. 1993. Clonal chromosome abnormalities in human breast carcinomas. I. Twenty-eight cases with primary disease. *Genes Chromosomes Cancer* 7: 185-193.
97. Watters, A. D., J. J. Going, T. G. Cooke, and J. M. Bartlett. 2003. Chromosome 17 aneusomy is associated with poor prognostic factors in invasive breast carcinoma. *Breast Cancer Res Treat* 77: 109-114.
98. Reinholz, M. M., A. K. Bruzek, D. W. Visscher, W. L. Lingle, M. J. Schroeder, E. A. Perez, and R. B. Jenkins. 2009. Breast cancer and aneusomy 17: implications for carcinogenesis and therapeutic response. *Lancet Oncol* 10: 267-277.
99. Jonsdottir, A. B., O. A. Stefansson, J. Bjornsson, J. G. Jonasson, H. M. Ogmundsdottir, and J. E. Eyfjord. 2012. Tetraploidy in BRCA2 breast tumours. *Eur J Cancer* 48: 305-310.
100. Min, J., E. S. Choi, K. Hwang, J. Kim, S. Sampath, A. R. Venkitaraman, and H. Lee. 2012. The breast cancer susceptibility gene BRCA2 is required for the maintenance of telomere homeostasis. *J Biol Chem* 287: 5091-5101.
101. Cancer Genome Atlas, N. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490: 61-70.
102. Chuaqui, R. F., Z. Zhuang, M. R. Emmert-Buck, L. A. Liotta, and M. J. Merino. 1997. Analysis of loss of heterozygosity on chromosome 11q13 in atypical ductal hyperplasia and *in situ* carcinoma of the breast. *Am J Pathol* 150: 297-303.
103. Lakhani, S. R., N. Collins, M. R. Stratton, and J. P. Sloane. 1995. Atypical ductal hyperplasia of the breast: clonal proliferation with loss of heterozygosity on chromosomes 16q and 17p. *J Clin Pathol* 48: 611-615.
104. Chappell, S. A., T. Walsh, R. A. Walker, and J. A. Shaw. 1997. Loss of heterozygosity at chromosome 6q in preinvasive and early invasive breast carcinomas. *Br J Cancer* 75: 1324-1329.
105. Devilee, P., M. van Vliet, P. van Sloun, N. Kuipers Dijkshoorn, J. Hermans, P. L. Pearson, and C. J. Cornelisse. 1991. Allelotype of human breast carcinoma: a second major site for loss of heterozygosity is on chromosome 6q. *Oncogene* 6: 1705-1711.
106. Orphanos, V., G. McGown, Y. Hey, J. M. Boyle, and M. Santibanez-Koref. 1995. Proximal 6q, a region showing allele loss in primary breast cancer. *Br J Cancer* 71: 290-293.
107. Almendro, V., Y. K. Cheng, A. Randles, S. Itzkovitz, A. Marusyk, E. Ametller, X. Gonzalez-Farre, M. Munoz, H. G. Russnes, A. Helland, I. H. Rye, A. L. Borresen-Dale, R. Maruyama, A. van Oudenaarden, M. Dowsett, R. L. Jones, J. Reis-Filho, P. Gascon,

- M. Gonen, F. Michor, and K. Polyak. 2014. Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Rep* 6: 514-527.
108. Janiszewska, M., L. Liu, V. Almendro, Y. Kuang, C. Paweletz, R. A. Sakr, B. Weigelt, A. B. Hanker, S. Chandarlapaty, T. A. King, J. S. Reis-Filho, C. L. Arteaga, S. Y. Park, F. Michor, and K. Polyak. 2015. *In situ* single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. *Nat Genet* 47: 1212-1219.
  109. Nguyen, A., M. Yoshida, H. Goodarzi, and S. F. Tavazoie. 2016. Highly variable cancer subpopulations that exhibit enhanced transcriptome variability and metastatic fitness. *Nat Commun* 7: 11246.
  110. Shah, S. P., A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, A. Bashashati, L. M. Prentice, J. Khattra, A. Burleigh, D. Yap, V. Bernard, A. McPherson, K. Shumansky, A. Crisan, R. Giuliany, A. Heravi-Moussavi, J. Rosner, D. Lai, I. Birol, R. Varhol, A. Tam, N. Dhalla, T. Zeng, K. Ma, S. K. Chan, M. Griffith, A. Moradian, S. W. Cheng, G. B. Morin, P. Watson, K. Gelmon, S. Chia, S. F. Chin, C. Curtis, O. M. Rueda, P. D. Pharoah, S. Damaraju, J. Mackey, K. Hoon, T. Harkins, V. Tadiotla, M. Sigaroudinia, P. Gascard, T. Tlsty, J. F. Costello, I. M. Meyer, C. J. Eaves, W. W. Wasserman, S. Jones, D. Huntsman, M. Hirst, C. Caldas, M. A. Marra, and S. Aparicio. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 486: 395-399.
  111. Aubele, M., and M. Werner. 1999. Heterogeneity in breast cancer and the problem of relevance of findings. *Anal Cell Pathol* 19: 53-58.
  112. Chin, K., C. O. de Solorzano, D. Knowles, A. Jones, W. Chou, E. G. Rodriguez, W. L. Kuo, B. M. Ljung, K. Chew, K. Myambo, M. Miranda, S. Krig, J. Garbe, M. Stampfer, P. Yaswen, J. W. Gray, and S. J. Lockett. 2004. *In situ* analyses of genome instability in breast cancer. *Nat Genet* 36: 984-988.
  113. Desmedt, C., D. Fumagalli, E. Pietri, G. Zoppoli, D. Brown, S. Nik-Zainal, G. Gundem, F. Rothe, S. Majjaj, A. Garuti, E. Carminati, S. Loi, T. Van Brussel, B. Boeckx, M. Maetens, L. Mudie, D. Vincent, N. Kheddoumi, L. Serra, I. Massa, A. Ballestrero, D. Amadori, R. Salgado, A. de Wind, D. Lambrechts, M. Piccart, D. Larsimont, P. J. Campbell, and C. Sotiriou. 2015. Uncovering the genomic heterogeneity of multifocal breast cancer. *J Pathol* 236: 457-466.
  114. Hernandez, L., P. M. Wilkerson, M. B. Lambros, A. Champion-Flora, D. N. Rodrigues, A. Gauthier, C. Cabral, V. Pawar, A. Mackay, R. A'Hern, C. Marchio, J. Palacios, R. Natrajan, B. Weigelt, and J. S. Reis-Filho. 2012. Genomic and mutational profiling of ductal carcinomas *in situ* and matched adjacent invasive breast cancers reveals intra-tumour genetic heterogeneity and clonal selection. *J Pathol* 227: 42-52.
  115. Hicks, J., A. Krasnitz, B. Lakshmi, N. E. Navin, M. Riggs, E. Leib, D. Esposito, J. Alexander, J. Troge, V. Grubor, S. Yoon, M. Wigler, K. Ye, A. L. Borresen-Dale, B. Naume, E. Schlicting, L. Norton, T. Hagerstrom, L. Skoog, G. Auer, S. Maner, P. Lundin, and A. Zetterberg. 2006. Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* 16: 1465-1479.
  116. Park, S. Y., M. Gonen, H. J. Kim, F. Michor, and K. Polyak. 2010. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J Clin Invest* 120: 636-644.
  117. Radford, D. M., N. J. Phillips, K. L. Fair, J. H. Ritter, M. Holt, and H. Donis-Keller. 1995. Allelic loss and the progression of breast cancer. *Cancer Res* 55: 5180-5183.

118. Stratton, M. R., N. Collins, S. R. Lakhani, and J. P. Sloane. 1995. Loss of heterozygosity in ductal carcinoma *in situ* of the breast. *J Pathol* 175: 195-201.
119. Kroigard, A. B., M. J. Larsen, A. V. Laenkholtm, A. S. Knoop, J. D. Jensen, M. Bak, J. Mollenhauer, T. A. Kruse, and M. Thomassen. 2015. Clonal expansion and linear genome evolution through breast cancer progression from pre-invasive stages to asynchronous metastasis. *Oncotarget* 6: 5634-5649.
120. Aubele, M., A. Mattis, H. Zitzelsberger, A. Walch, M. Kremer, G. Welzl, H. Hofler, and M. Werner. 2000. Extensive ductal carcinoma *In situ* with small foci of invasive ductal carcinoma: evidence of genetic resemblance by CGH. *Int J Cancer* 85: 82-86.
121. Burkhardt, L., T. J. Grob, I. Hermann, E. Burandt, M. Choschzick, F. Janicke, V. Muller, C. Bokemeyer, R. Simon, G. Sauter, W. Wilczak, and A. Lebeau. 2010. Gene amplification in ductal carcinoma *in situ* of the breast. *Breast Cancer Res Treat* 123: 757-765.
122. Chen, T., A. Sahin, and C. M. Aldaz. 1996. Deletion map of chromosome 16q in ductal carcinoma *in situ* of the breast: refining a putative tumor suppressor gene region. *Cancer Res* 56: 5605-5609.
123. Munn, K. E., R. A. Walker, and J. M. Varley. 1995. Frequent alterations of chromosome 1 in ductal carcinoma *in situ* of the breast. *Oncogene* 10: 1653-1657.
124. Murphy, D. S., P. McHardy, J. Coutts, E. A. Mallon, W. D. George, S. B. Kaye, R. Brown, and W. N. Keith. 1995. Interphase cytogenetic analysis of erbB2 and topoII alpha co-amplification in invasive breast cancer and polysomy of chromosome 17 in ductal carcinoma *in situ*. *Int J Cancer* 64: 18-26.
125. Park, K., S. Han, H. J. Kim, J. Kim, and E. Shin. 2006. HER2 status in pure ductal carcinoma *in situ* and in the intraductal and invasive components of invasive ductal carcinoma determined by fluorescence *in situ* hybridization and immunohistochemistry. *Histopathology* 48: 702-707.
126. Werner, M., A. Mattis, M. Aubele, M. Cummings, H. Zitzelsberger, P. Hutzler, and H. Hofler. 1999. 20q13.2 amplification in intraductal hyperplasia adjacent to *in situ* and invasive ductal carcinoma of the breast. *Virchows Arch* 435: 469-472.
127. Kurozumi, S., M. Padilla, M. Kurosumi, H. Matsumoto, K. Inoue, J. Horiguchi, I. Takeyoshi, T. Oyama, J. Ranger-Moore, D. C. Allred, E. Dennis, and H. Nitta. 2016. HER2 intratumoral heterogeneity analyses by concurrent HER2 gene and protein assessment for the prognosis of HER2 negative invasive breast cancer patients. *Breast Cancer Res Treat* 158: 99-111.
128. Ellsworth, R. E., A. Vertrees, B. Love, J. A. Hooke, D. L. Ellsworth, and C. D. Shriver. 2008. Chromosomal alterations associated with the transition from *in situ* to invasive breast cancer. *Ann Surg Oncol* 15: 2519-2525.
129. Navin, N., and J. Hicks. 2011. Future medical applications of single-cell sequencing in cancer. *Genome Med* 3: 31.
130. Eirew, P., A. Steif, J. Khattra, G. Ha, D. Yap, H. Farahani, K. Gelmon, S. Chia, C. Mar, A. Wan, E. Laks, J. Biele, K. Shumansky, J. Rosner, A. McPherson, C. Nielsen, A. J. Roth, C. Lefebvre, A. Bashashati, C. de Souza, C. Siu, R. Aniba, J. Brimhall, A. Oloumi, T. Osako, A. Bruna, J. L. Sandoval, T. Algara, W. Greenwood, K. Leung, H. Cheng, H. Xue, Y. Wang, D. Lin, A. J. Mungall, R. Moore, Y. Zhao, J. Lorette, L. Nguyen, D. Huntsman, C. J. Eaves, C. Hansen, M. A. Marra, C. Caldas, S. P. Shah, and S. Aparicio. 2014. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*.
131. Nik-Zainal, S., P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton,

- A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jonsson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerod, S. A. Aparicio, A. Tutt, A. M. Sieuwerts, A. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A. L. Borresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell, and C. Breast Cancer Working Group of the International Cancer Genome. 2012. The life history of 21 breast cancers. *Cell* 149: 994-1007.
132. Gerlinger, M., S. Horswell, J. Larkin, A. J. Rowan, M. P. Salm, I. Varela, R. Fisher, N. McGranahan, N. Matthews, C. R. Santos, P. Martinez, B. Phillimore, S. Begum, A. Rabinowitz, B. Spencer-Dene, S. Gulati, P. A. Bates, G. Stamp, L. Pickering, M. Gore, D. L. Nicol, S. Hazell, P. A. Futreal, A. Stewart, and C. Swanton. 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 46: 225-233.
  133. Zhang, J., J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C. W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, Wistuba, II, and P. A. Futreal. 2014. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* 346: 256-259.
  134. Leung, M. L., Y. Wang, J. Waters, and N. E. Navin. 2015. SNES: single nucleus exome sequencing. *Genome Biol* 16: 55.
  135. Xu, X., Y. Hou, X. Yin, L. Bao, A. Tang, L. Song, F. Li, S. Tsang, K. Wu, H. Wu, W. He, L. Zeng, M. Xing, R. Wu, H. Jiang, X. Liu, D. Cao, G. Guo, X. Hu, Y. Gui, Z. Li, W. Xie, X. Sun, M. Shi, Z. Cai, B. Wang, M. Zhong, J. Li, Z. Lu, N. Gu, X. Zhang, L. Goodman, L. Bolund, J. Wang, H. Yang, K. Kristiansen, M. Dean, and Y. Li. 2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148: 886-895.
  136. Zong, C., S. Lu, A. R. Chapman, and X. S. Xie. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338: 1622-1626.
  137. Leung, M. L., Y. Wang, C. Kim, R. Gao, J. Jiang, E. Sei, and N. E. Navin. 2016. Highly multiplexed targeted DNA sequencing from single nuclei. *Nat Protoc* 11: 214-235.
  138. Navin, N. E. 2014. Cancer genomics: one cell at a time. *Genome Biol* 15: 452.
  139. Foschini, M. P., L. Morandi, E. Leonardi, F. Flamminio, Y. Ishikawa, R. Masetti, and V. Eusebi. 2013. Genetic clonal mapping of *in situ* and invasive ductal carcinoma indicates the field cancerization phenomenon in the breast. *Hum Pathol* 44: 1310-1319.
  140. Luzzi, V., V. Holtschlag, and M. A. Watson. 2001. Expression profiling of ductal carcinoma *in situ* by laser capture microdissection and high-density oligonucleotide arrays. *Am J Pathol* 158: 2005-2010.
  141. Westbury, C. B., J. S. Reis-Filho, T. Dexter, B. Mahler-Araujo, K. Fenwick, M. Iravani, A. Grigoriadis, S. Parry, D. Robertson, A. Mackay, A. Ashworth, J. R. Yarnold, and C. M. Isacke. 2009. Genome-wide transcriptomic profiling of microdissected human breast tissue reveals differential expression of KIT (c-Kit, CD117) and oestrogen receptor-alpha (ERalpha) in response to therapeutic radiation. *J Pathol* 219: 131-140.
  142. Ghazani, A. A., N. Arneson, K. Warren, M. Pintilie, J. Bayani, J. A. Squire, and S. J. Done. 2007. Genomic alterations in sporadic synchronous primary breast cancer using array and metaphase comparative genomic hybridization. *Neoplasia* 9: 511-520.
  143. Koboldt, D. C., K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* 155: 27-38.

144. Newburger, D. E., D. Kashef-Haghighi, Z. Weng, R. Salari, R. T. Sweeney, A. L. Brunner, S. X. Zhu, X. Guo, S. Varma, M. L. Troxell, R. B. West, S. Batzoglou, and A. Sidow. 2013. Genome evolution during progression to breast cancer. *Genome Res* 23: 1097-1108.
145. Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6: 377-382.
146. Biezuner, T., Raz, O., Amir, S., Milo, L., Adar, R., Fried, Y., Ainbinder, E., Shapiro, E. 2017. Comparison of seven single cell Whole Genome Amplification commercial kits using targeted sequencing *BioRxiv*.
147. Huang, L., F. Ma, A. Chapman, S. Lu, and X. S. Xie. 2015. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu Rev Genomics Hum Genet* 16: 79-102.
148. Biezuner, T., A. Spiro, O. Raz, S. Amir, L. Milo, R. Adar, N. Chapal-Ilani, V. Berman, Y. Fried, E. Ainbinder, G. Cohen, H. M. Barr, R. Halaban, and E. Shapiro. 2016. A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res* 26: 1588-1599.
149. Shibata, D., D. Hawes, Z. H. Li, A. M. Hernandez, C. H. Spruck, and P. W. Nichols. 1992. Specific genetic analysis of microscopic tissue after selective ultraviolet radiation fractionation and the polymerase chain reaction. *Am J Pathol* 141: 539-543.
150. Emmert-Buck, M. R., R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta. 1996. Laser capture microdissection. *Science* 274: 998-1001.
151. Lampejo, O. T., D. M. Barnes, P. Smith, and R. R. Millis. 1994. Evaluation of infiltrating ductal carcinomas with a DCIS component: correlation of the histologic type of the *in situ* component with grade of the infiltrating component. *Semin Diagn Pathol* 11: 215-222.
152. Buerger, H., F. Otterbach, R. Simon, K. L. Schafer, C. Poremba, R. Diallo, C. Brinkschmidt, B. Dockhorn-Dworniczak, and W. Boecker. 1999. Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes. *J Pathol* 189: 521-526.
153. Fujii, H., C. Marsh, P. Cairns, D. Sidransky, and E. Gabrielson. 1996. Genetic divergence in the clonal evolution of breast cancer. *Cancer Res* 56: 1493-1497.
154. Martincorena, I., A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, A. Fullam, L. B. Alexandrov, J. M. Tubio, L. Stebbings, A. Menzies, S. Widaa, M. R. Stratton, P. H. Jones, and P. J. Campbell. 2015. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348: 880-886.
155. Jakubek, Y., W. Lang, S. Vattathil, M. Garcia, L. Xu, L. Huang, S. Y. Yoo, L. Shen, W. Lu, C. W. Chow, Z. Weber, G. Davies, J. Huang, C. Behrens, N. Kalhor, C. Moran, J. Fujimoto, R. Mehran, R. El-Zein, S. G. Swisher, J. Wang, J. Fowler, A. E. Spira, E. A. Ehli, Wistuba, II, P. Scheet, and H. Kadara. 2016. Genomic Landscape Established by Allelic Imbalance in the Cancerization Field of a Normal Appearing Airway. *Cancer Res* 76: 3676-3683.
156. Forsberg, L. A., C. Rasi, G. Pekar, H. Davies, A. Piotrowski, D. Absher, H. R. Razzaghian, A. Ambicka, K. Halaszka, M. Przewoznik, A. Kruczak, G. Mandava, S. Pasupulati, J. Hacker, K. R. Prakash, R. C. Dasari, J. Lau, N. Penagos-Tafurt, H. M. Olofsson, G. Hallberg, P. Skotnicki, J. Mitus, J. Skokowski, M. Jankowski, E. Srutek, W. Zegarski, E. Tiensuu Janson, J. Rys, T. Tot, and J. P. Dumanski. 2015. Signatures of

- post-zygotic structural genetic aberrations in the cells of histologically normal breast tissue that can predispose to sporadic breast cancer. *Genome Res* 25: 1521-1535.
157. Abdel-Fatah, T. M., C. Perry, A. Arora, N. Thompson, R. Doherty, P. M. Moseley, A. R. Green, S. Y. Chan, I. O. Ellis, and S. Madhusudan. 2014. Is there a role for base excision repair in estrogen/estrogen receptor-driven breast cancers? *Antioxid Redox Signal* 21: 2262-2268.
  158. Crook, T., L. A. Brooks, S. Crossland, P. Osin, K. T. Barker, J. Waller, E. Philp, P. D. Smith, I. Yulug, J. Peto, G. Parker, M. J. Allday, M. R. Crompton, and B. A. Gusterson. 1998. p53 mutation with frequent novel condons but not a mutator phenotype in BRCA1- and BRCA2-associated breast tumours. *Oncogene* 17: 1681-1689.
  159. Yang, D., S. Khan, Y. Sun, K. Hess, I. Shmulevich, A. K. Sood, and W. Zhang. 2011. Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA* 306: 1557-1565.
  160. Miron, A., M. Varadi, D. Carrasco, H. Li, L. Luongo, H. J. Kim, S. Y. Park, E. Y. Cho, G. Lewis, S. Kehoe, J. D. Iglehart, D. Dillon, D. C. Allred, L. Macconail, R. Gelman, and K. Polyak. 2010. PIK3CA mutations in *in situ* and invasive breast carcinomas. *Cancer Res* 70: 5674-5678.
  161. Sontag, L., and D. E. Axelrod. 2005. Evaluation of pathways for progression of heterogeneous breast tumors. *J Theor Biol* 232: 179-189.
  162. Hwang, E. S., A. Lal, Y. Y. Chen, S. DeVries, R. Swain, J. Anderson, R. Roy, and F. M. Waldman. 2011. Genomic alterations and phenotype of large compared to small high-grade ductal carcinoma *in situ*. *Hum Pathol* 42: 1467-1475.
  163. Iakovlev, V. V., N. C. Arneson, V. Wong, C. Wang, S. Leung, G. Iakovleva, K. Warren, M. Pintilie, and S. J. Done. 2008. Genomic differences between pure ductal carcinoma *in situ* of the breast and that associated with invasive disease: a calibrated aCGH study. *Clin Cancer Res* 14: 4446-4454.
  164. Liao, S., M. M. Desouki, D. P. Gaile, L. Shepherd, N. J. Nowak, J. Conroy, W. T. Barry, and J. Geradts. 2012. Differential copy number aberrations in novel candidate genes associated with progression from *in situ* to invasive ductal carcinoma of the breast. *Genes Chromosomes Cancer* 51: 1067-1078.
  165. Muggerud, A. A., M. Hallett, H. Johnsen, K. Kleivi, W. Zhou, S. Tahmasebpour, R. M. Amini, J. Botling, A. L. Borresen-Dale, T. Sorlie, and F. Warnberg. 2010. Molecular diversity in ductal carcinoma *in situ* (DCIS) and early invasive breast cancer. *Mol Oncol* 4: 357-368.
  166. Oikawa, M., H. Yano, M. Matsumoto, R. Otsubo, K. Shibata, T. Hayashi, K. Abe, N. Kinoshita, K. Yoshiura, and T. Nagayasu. 2015. A novel diagnostic method targeting genomic instability in intracystic tumors of the breast. *Breast Cancer* 22: 529-535.
  167. Yao, J., S. Weremowicz, B. Feng, R. C. Gentleman, J. R. Marks, R. Gelman, C. Brennan, and K. Polyak. 2006. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res* 66: 4065-4078.
  168. Cummings, M. C., M. Aubele, A. Mattis, D. Purdie, P. Hutzler, H. Hofler, and M. Werner. 2000. Increasing chromosome 1 copy number parallels histological progression in breast carcinogenesis. *Br J Cancer* 82: 1204-1210.
  169. Hwang, E. S., S. DeVries, K. L. Chew, D. H. Moore, 2nd, K. Kerlikowske, A. Thor, B. M. Ljung, and F. M. Waldman. 2004. Patterns of chromosomal alterations in breast ductal carcinoma *in situ*. *Clin Cancer Res* 10: 5160-5167.
  170. Sakr, R. A., B. Weigelt, S. Chandarlapaty, V. P. Andrade, E. Guerini-Rocco, D. Giri, C. K. Ng, C. F. Cowell, N. Rosen, J. S. Reis-Filho, and T. A. King. 2014. PI3K pathway

- activation in high-grade ductal carcinoma *in situ*--implications for progression to invasive breast carcinoma. *Clin Cancer Res* 20: 2326-2337.
171. Petridis, C., M. N. Brook, V. Shah, K. Kohut, P. Gorman, M. Caneppele, D. Levi, E. Papouli, N. Orr, A. Cox, S. S. Cross, I. Dos-Santos-Silva, J. Peto, A. Swerdlow, M. J. Schoemaker, M. K. Bolla, Q. Wang, J. Dennis, K. Michailidou, J. Benitez, A. Gonzalez-Neira, D. C. Tessier, D. Vincent, J. Li, J. Figueroa, V. Kristensen, A. L. Borresen-Dale, P. Soucy, J. Simard, R. L. Milne, G. G. Giles, S. Margolin, A. Lindblom, T. Bruning, H. Brauch, M. C. Southey, J. L. Hopper, T. Dork, N. V. Bogdanova, M. Kabisch, U. Hamann, R. K. Schmutzler, A. Meindl, H. Brenner, V. Arndt, R. Winqvist, K. Pylkas, P. A. Fasching, M. W. Beckmann, J. Lubinski, A. Jakubowska, A. M. Mulligan, I. L. Andrulis, R. A. Tollenaar, P. Devilee, L. Le Marchand, C. A. Haiman, A. Mannermaa, V. M. Kosma, P. Radice, P. Peterlongo, F. Marme, B. Burwinkel, C. H. van Deurzen, A. Hollestelle, N. Miller, M. J. Kerin, D. Lambrechts, G. Floris, J. Wesseling, H. Flyger, S. E. Bojesen, S. Yao, C. B. Ambrosone, G. Chenevix-Trench, T. Truong, P. Guenel, A. Rudolph, J. Chang-Claude, H. Nevanlinna, C. Blomqvist, K. Czene, J. S. Brand, J. E. Olson, F. J. Couch, A. M. Dunning, P. Hall, D. F. Easton, P. D. Pharoah, S. E. Pinder, M. K. Schmidt, I. Tomlinson, R. Roylance, M. Garcia-Closas, and E. J. Sawyer. 2016. Genetic predisposition to ductal carcinoma *in situ* of the breast. *Breast Cancer Res* 18: 22.
  172. Gorringer, K. L., S. M. Hunter, J. M. Pang, K. Opekin, P. Hill, S. M. Rowley, D. Y. Choong, E. R. Thompson, A. Dobrovic, S. B. Fox, G. B. Mann, and I. G. Campbell. 2015. Copy number analysis of ductal carcinoma *in situ* with and without recurrence. *Mod Pathol* 28: 1174-1184.
  173. Wu, C. I., and C. T. Ting. 2004. Genes and speciation. *Nat Rev Genet* 5: 114-122.
  174. Gundem, G., P. Van Loo, B. Kremeyer, L. B. Alexandrov, J. M. Tubio, E. Papaemmanuil, D. S. Brewer, H. M. Kallio, G. Hognas, M. Annala, K. Kivinummi, V. Goody, C. Latimer, S. O'Meara, K. J. Dawson, W. Isaacs, M. R. Emmert-Buck, M. Nykter, C. Foster, Z. Kote-Jarai, D. Easton, H. C. Whitaker, I. P. U. Group, D. E. Neal, C. S. Cooper, R. A. Eeles, T. Visakorpi, P. J. Campbell, U. McDermott, D. C. Wedge, and G. S. Bova. 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature* 520: 353-357.
  175. Ding, L., T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan, J. F. McMichael, J. W. Wallis, C. Lu, D. Shen, C. C. Harris, D. J. Dooling, R. S. Fulton, L. L. Fulton, K. Chen, H. Schmidt, J. Kalicki-Veizer, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, M. C. Wendl, S. Heath, M. A. Watson, D. C. Link, M. H. Tomasson, W. D. Shannon, J. E. Payton, S. Kulkarni, P. Westervelt, M. J. Walter, T. A. Graubert, E. R. Mardis, R. K. Wilson, and J. F. DiPersio. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481: 506-510.
  176. Kim, H., S. Zheng, S. S. Amini, S. M. Virk, T. Mikkelsen, D. J. Brat, J. Grimsby, C. Sougnez, F. Muller, J. Hu, A. E. Sloan, M. L. Cohen, E. G. Van Meir, L. Scarpace, P. W. Laird, J. N. Weinstein, E. S. Lander, S. Gabriel, G. Getz, M. Meyerson, L. Chin, J. S. Barnholtz-Sloan, and R. G. Verhaak. 2015. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res* 25: 316-327.
  177. Landau, D. A., K. Clement, M. J. Ziller, P. Boyle, J. Fan, H. Gu, K. Stevenson, C. Sougnez, L. Wang, S. Li, D. Kotliar, W. Zhang, M. Ghandi, L. Garraway, S. M. Fernandes, K. J. Livak, S. Gabriel, A. Gnirke, E. S. Lander, J. R. Brown, D. Neuberg, P. V. Kharchenko, N. Hacohen, G. Getz, A. Meissner, and C. J. Wu. 2014. Locally

- disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* 26: 813-825.
178. Patch, A. M., E. L. Christie, D. Etemadmoghadam, D. W. Garsed, J. George, S. Fereday, K. Nones, P. Cowin, K. Alsop, P. J. Bailey, K. S. Kassahn, F. Newell, M. C. Quinn, S. Kazakoff, K. Quek, C. Wilhelm-Benartzi, E. Curry, H. S. Leong, G. Australian Ovarian Cancer Study, A. Hamilton, L. Mileschkin, G. Au-Yeung, C. Kennedy, J. Hung, Y. E. Chiew, P. Harnett, M. Friedlander, M. Quinn, J. Pyman, S. Cordner, P. O'Brien, J. Leditschke, G. Young, K. Strachan, P. Waring, W. Azar, C. Mitchell, N. Traficante, J. Hendley, H. Thorne, M. Shackleton, D. K. Miller, G. M. Arnau, R. W. Tothill, T. P. Holloway, T. Semple, I. Harliwong, C. Nourse, E. Nourbakhsh, S. Manning, S. Idrisoglu, T. J. Bruxner, A. N. Christ, B. Poudel, O. Holmes, M. Anderson, C. Leonard, A. Lonie, N. Hall, S. Wood, D. F. Taylor, Q. Xu, J. L. Fink, N. Waddell, R. Drapkin, E. Stronach, H. Gabra, R. Brown, A. Jewell, S. H. Nagaraj, E. Markham, P. J. Wilson, J. Ellul, O. McNally, M. A. Doyle, R. Vedururu, C. Stewart, E. Lengyel, J. V. Pearson, N. Waddell, A. deFazio, S. M. Grimmond, and D. D. Bowtell. 2015. Whole-genome characterization of chemoresistant ovarian cancer. *Nature* 521: 489-494.
  179. Weng, Z., N. Spies, S. X. Zhu, D. E. Newburger, D. Kashef-Haghighi, S. Batzoglou, A. Sidow, and R. B. West. 2015. Cell-lineage heterogeneity and driver mutation recurrence in pre-invasive breast neoplasia. *Genome Med* 7: 28.
  180. Cowell, C. F., B. Weigelt, R. A. Sakr, C. K. Ng, J. Hicks, T. A. King, and J. S. Reis-Filho. 2013. Progression from ductal carcinoma *in situ* to invasive breast cancer: revisited. *Mol Oncol* 7: 859-869.
  181. Nyante, S. J., S. Devries, Y. Y. Chen, and E. S. Hwang. 2004. Array-based comparative genomic hybridization of ductal carcinoma *in situ* and synchronous invasive lobular cancer. *Hum Pathol* 35: 759-763.
  182. Aubele, M. M., M. C. Cummings, A. E. Mattis, H. F. Zitzelsberger, A. K. Walch, M. Kremer, H. Hofler, and M. Werner. 2000. Accumulation of chromosomal imbalances from intraductal proliferative lesions to adjacent *in situ* and invasive ductal breast cancer. *Diagn Mol Pathol* 9: 14-19.
  183. Marusyk, A., D. P. Tabassum, P. M. Altrock, V. Almendro, F. Michor, and K. Polyak. 2014. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* 514: 54-58.
  184. Cleary, A. S., T. L. Leonard, S. A. Gestl, and E. J. Gunther. 2014. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* 508: 113-117.
  185. Alizadeh, A. A., V. Aranda, A. Bardelli, C. Blanpain, C. Bock, C. Borowski, C. Caldas, A. Califano, M. Doherty, M. Elsner, M. Esteller, R. Fitzgerald, J. O. Korbel, P. Lichter, C. E. Mason, N. Navin, D. Pe'er, K. Polyak, C. W. Roberts, L. Siu, A. Snyder, H. Stower, C. Swanton, R. G. Verhaak, J. C. Zenklusen, J. Zuber, and J. Zucman-Rossi. 2015. Toward understanding and exploiting tumor heterogeneity. *Nat Med* 21: 846-853.
  186. Fisher, R., L. Pusztai, and C. Swanton. 2013. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer* 108: 479-485.
  187. Jamal-Hanjani, M., A. Hackshaw, Y. Ngai, J. Shaw, C. Dive, S. Quezada, G. Middleton, E. de Bruin, J. Le Quesne, S. Shafi, M. Falzon, S. Horswell, F. Blackhall, I. Khan, S. Janes, M. Nicolson, D. Lawrence, M. Forster, D. Fennell, S. M. Lee, J. Lester, K. Kerr, S. Muller, N. Iles, S. Smith, N. Murugaesu, R. Mitter, M. Salm, A. Stuart, N. Matthews, H. Adams, T. Ahmad, R. Attanoos, J. Bennett, N. J. Birkbak, R. Booton, G. Brady, K. Buchan, A. Capitano, M. Chetty, M. Cobbold, P. Crosbie, H. Davies, A. Denison, M. Djeerman, J. Goldman, T. Haswell, L. Joseph, M. Kornaszewska, M. Krebs, G.



- Langman, M. MacKenzie, J. Millar, B. Morgan, B. Naidu, D. Nonaka, K. Peggs, C. Pritchard, H. Remmen, A. Rowan, R. Shah, E. Smith, Y. Summers, M. Taylor, S. Veeriah, D. Waller, B. Wilcox, M. Wilcox, I. Woolhouse, N. McGranahan, and C. Swanton. 2014. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol* 12: e1001906.
188. Dillon, L. M., and T. W. Miller. 2014. Therapeutic targeting of cancers with loss of PTEN function. *Curr Drug Targets* 15: 65-79.
189. Grabchak, M., E. Marcon, G. Lang, and Z. Y. Zhang. 2017. The generalized Simpson's entropy is a measure of biodiversity. *PLoS One* 12.
190. Morris, E. K., T. Caruso, F. Buscot, M. Fischer, C. Hancock, T. S. Maier, T. Meiners, C. Muller, E. Obermaier, D. Prati, S. A. Socher, I. Sonnemann, N. Waschke, T. Wubet, S. Wurst, and M. C. Rillig. 2014. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution* 4: 3514-3524.
191. Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 623-656.
192. Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379-423.
193. Simpson, E. H. 1949. Measurement of Diversity. *Nature* 163: 688-688.
194. Almendro, V., H. J. Kim, Y. K. Cheng, M. Gonen, S. Itzkovitz, P. Argani, A. van Oudenaarden, S. Sukumar, F. Michor, and K. Polyak. 2014. Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res* 74: 1338-1348.
195. Erbas, B., E. Provenzano, J. Armes, and D. Gertig. 2006. The natural history of ductal carcinoma *in situ* of the breast: a review. *Breast Cancer Res Treat* 97: 135-144.
196. Ozanne, E. M., Y. Shieh, J. Barnes, C. Bouzan, E. S. Hwang, and L. J. Esserman. 2011. Characterizing the impact of 25 years of DCIS treatment. *Breast Cancer Res Treat* 129: 165-173.
197. Edgerton, M. 2017. Progression of DCIS at MDA. Email ed. A. Casasent, and N. Navin, eds.
198. Hammond, M. E., D. F. Hayes, A. C. Wolff, P. B. Mangu, and S. Temin. 2010. American society of clinical oncology/college of american pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Oncol Pract* 6: 195-197.
199. PALM Protocols - DNA Handling. Zeiss, ed.
200. Gao, R., A. Davis, T. O. McDonald, E. Sei, X. Shi, Y. Wang, P. C. Tsai, A. Casasent, J. Waters, H. Zhang, F. Meric-Bernstam, F. Michor, and N. E. Navin. 2016. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* 48: 1119-1130.
201. Langdon, W. B. 2015. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min* 8: 1.
202. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.
203. Baslan, T., J. Kendall, B. Ward, H. Cox, A. Leotta, L. Rodgers, M. Riggs, S. D'Italia, G. Sun, M. Yong, K. Miskimen, H. Gilmore, M. Saborowski, N. Dimitrova, A. Krasnitz, L. Harris, M. Wigler, and J. Hicks. 2015. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Res* 25: 714-724.

204. Shah, S. P., X. Xuan, R. J. DeLeeuw, M. Khojasteh, W. L. Lam, R. Ng, and K. P. Murphy. 2006. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* 22: e431-439.
205. Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5: 557-572.
206. Dudik, J. M., A. Kurosu, J. L. Coyle, and E. Sejdic. 2015. A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals. *Comput Biol Med* 59: 10-18.
207. Martin Ester, H.-P. K., Jorg Sander, Xiaowei Xu 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*.
208. Yue, S. H., P. Li, J. D. Guo, and S. G. Zhou. 2004. Using Greedy algorithm: DBSCAN revisited II. *J Zhejiang Univ Sci* 5: 1405-1412.
209. Piekenbrock, M. H. a. M. 2017. dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms., R package version 1.1-1 ed. CRAN.
210. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and K. Hornik. 2012. cluster: Cluster Analysis Basics and Extensions., R package version 2.0.6. ed.
211. Tibshirani, R., Walther, G. and Hastie, T. 2001. Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B*: 411-423.
212. Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. 2004. A census of human cancer genes. *Nat Rev Cancer* 4: 177-183.
213. Scornavacca, C., F. Zickmann, and D. H. Huson. 2011. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics* 27: i248-256.
214. Institute, B. 2016. Picard Tools.
215. Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38: e164.
216. Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132: 365-386.
217. Gerstung, M., C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch, and N. Beerenwinkel. 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* 3: 811.
218. Patel, R., A. Filer, F. Barone, and C. D. Buckley. 2014. Stroma: fertile soil for inflammation. *Best Pract Res Clin Rheumatol* 28: 565-576.
219. Tirosh, I., B. Izar, S. M. Prakadan, M. H. Wadsworth, 2nd, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J. R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A. C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jane-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, and L. A. Garraway. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352: 189-196.
220. Aceto, N., A. Bardia, D. T. Miyamoto, M. C. Donaldson, B. S. Wittner, J. A. Spencer, M. Yu, A. Pely, A. Engstrom, H. Zhu, B. W. Brannigan, R. Kapur, S. L. Stott, T. Shioda, S. Ramaswamy, D. T. Ting, C. P. Lin, M. Toner, D. A. Haber, and S. Maheswaran. 2014. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* 158: 1110-1122.
221. Lohr, J. G., V. A. Adalsteinsson, K. Cibulskis, A. D. Choudhury, M. Rosenberg, P. Cruz-Gordillo, J. M. Francis, C. Z. Zhang, A. K. Shalek, R. Satija, J. J. Trombetta, D.

- Lu, N. Tallapragada, N. Tahirova, S. Kim, B. Blumenstiel, C. Sougne, A. Lowe, B. Wong, D. Auclair, E. M. Van Allen, M. Nakabayashi, R. T. Lis, G. S. Lee, T. Li, M. S. Chabot, A. Ly, M. E. Taplin, T. E. Clancy, M. Loda, A. Regev, M. Meyerson, W. C. Hahn, P. W. Kantoff, T. R. Golub, G. Getz, J. S. Boehm, and J. C. Love. 2014. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat Biotechnol* 32: 479-484.
222. Datta, S., L. Malhotra, R. Dickerson, S. Chaffee, C. K. Sen, and S. Roy. 2015. Laser capture microdissection: Big data from small samples. *Histol Histopathol* 30: 1255-1269.
  223. Chen, W. Y., C. S. Chen, H. C. Chen, Y. J. Hung, and J. S. Chu. 2004. Mucinous cystadenocarcinoma of the breast coexisting with infiltrating ductal carcinoma. *Pathol Int* 54: 781-786.
  224. Wong, H., S. Lau, R. Leung, J. Chiu, P. Cheung, T. T. Wong, R. Liang, R. J. Epstein, and T. Yau. 2012. Coexisting ductal carcinoma *in situ* independently predicts lower tumor aggressiveness in node-positive luminal breast cancer. *Med Oncol* 29: 1536-1542.
  225. Andl, C. D., T. Mizushima, K. Oyama, M. Bowser, H. Nakagawa, and A. K. Rustgi. 2004. EGFR-induced cell migration is mediated predominantly by the JAK-STAT pathway in primary esophageal keratinocytes. *Am J Physiol Gastrointest Liver Physiol* 287: G1227-1237.
  226. Suh, S. S., J. Y. Yoo, R. Cui, B. Kaur, K. Huebner, T. K. Lee, R. I. Aqeilan, and C. M. Croce. 2014. FHIT suppresses epithelial-mesenchymal transition (EMT) and metastasis in lung cancer through modulation of microRNAs. *PLoS Genet* 10: e1004652.
  227. Smith, M. A., C. B. Nielsen, F. C. Chan, A. McPherson, A. Roth, H. Farahani, D. Machev, A. Steif, and S. P. Shah. 2017. E-scape: interactive visualization of single-cell phylogenetics and cancer evolution. *Nat Methods* 14: 549-550.
  228. Gerstung, M., E. Papaemmanuil, and P. J. Campbell. 2014. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* 30: 1198-1204.
  229. Sarper, M., M. D. Allen, J. Gomm, L. Haywood, J. Decock, S. Thirkettle, A. Ustaoglu, S. J. Sarker, J. Marshall, D. R. Edwards, and J. L. Jones. 2017. Loss of MMP-8 in ductal carcinoma *in situ* (DCIS)-associated myoepithelial cells contributes to tumour promotion through altered adhesive and proteolytic function. *Breast Cancer Res* 19: 33.
  230. Vehvilainen, P., M. Hyytiainen, and J. Keski-Oja. 2003. Latent transforming growth factor-beta-binding protein 2 is an adhesion protein for melanoma cells. *J Biol Chem* 278: 24705-24713.
  231. Roth, A., J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Cote, and S. P. Shah. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 11: 396-398.
  232. Malikic, S., A. W. McPherson, N. Donmez, and C. S. Sahinalp. 2015. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* 31: 1349-1356.
  233. Chen, X., Y. C. Sun, G. M. Church, J. H. Lee, and A. M. Zador. 2018. Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res* 46: e22.
  234. Chang, Q., O. I. Ornatsky, I. Siddiqui, A. Loboda, V. I. Baranov, and D. W. Hedley. 2017. Imaging Mass Cytometry. *Cytometry A* 91: 160-169.
  235. Giesen, C., H. A. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schuffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Gunther, and B. Bodenmiller. 2014. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* 11: 417-422.
  236. Hu, M., J. Yao, D. K. Carroll, S. Weremowicz, H. Chen, D. Carrasco, A. Richardson, S. Violette, T. Nikolskaya, Y. Nikolsky, E. L. Bauerlein, W. C. Hahn, R. S. Gelman, C.

- Allred, M. J. Bissell, S. Schnitt, and K. Polyak. 2008. Regulation of *in situ* to invasive breast carcinoma transition. *Cancer Cell* 13: 394-406.
237. Sharma, M., A. H. Beck, J. A. Webster, I. Espinosa, K. Montgomery, S. Varma, M. van de Rijn, K. C. Jensen, and R. B. West. 2010. Analysis of stromal signatures in the tumor microenvironment of ductal carcinoma *in situ*. *Breast Cancer Res Treat* 123: 397-404.
238. Lee, J. H., E. R. Daugharthy, J. Scheiman, R. Kallhor, J. L. Yang, T. C. Ferrante, R. Terry, S. S. Jeanty, C. Li, R. Amamoto, D. T. Peters, B. M. Turczyk, A. H. Marblestone, S. A. Inverso, A. Bernard, P. Mali, X. Rios, J. Aach, and G. M. Church. 2014. Highly multiplexed subcellular RNA sequencing *in situ*. *Science* 343: 1360-1363.
239. Heselmeyer-Haddad, K., L. Y. Berroa Garcia, A. Bradley, C. Ortiz-Melendez, W. J. Lee, R. Christensen, S. A. Prindiville, K. A. Calzone, P. W. Soballe, Y. Hu, S. A. Chowdhury, R. Schwartz, A. A. Schaffer, and T. Ried. 2012. Single-cell genetic analysis of ductal carcinoma *in situ* and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. *Am J Pathol* 181: 1807-1822.
240. Fidler, I. J., and G. Poste. 2008. The "seed and soil" hypothesis revisited. *Lancet Oncol* 9: 808.
241. Baines, H. L., J. B. Stewart, C. Stamp, A. Zupanic, T. B. Kirkwood, N. G. Larsson, D. M. Turnbull, and L. C. Greaves. 2014. Similar patterns of clonally expanded somatic mtDNA mutations in the colon of heterozygous mtDNA mutator mice and ageing humans. *Mech Ageing Dev* 139: 22-30.
242. Greaves, M., and C. C. Maley. 2012. Clonal evolution in cancer. *Nature* 481: 306-313.
243. Kang, H., M. P. Salomon, A. Sottoriva, J. Zhao, M. Toy, M. F. Press, C. Curtis, P. Marjoram, K. Siegmund, and D. Shibata. 2015. Many private mutations originate from the first few divisions of a human colorectal adenoma. *J Pathol* 237: 355-362.
244. Sottoriva, A., H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, D. Shibata, and C. Curtis. 2015. A Big Bang model of human colorectal tumor growth. *Nat Genet* 47: 209-216.
245. Kim, M. Y., T. Oskarsson, S. Acharyya, D. X. Nguyen, X. H. Zhang, L. Norton, and J. Massague. 2009. Tumor self-seeding by circulating cancer cells. *Cell* 139: 1315-1326.
246. Leung, C. T., and J. S. Brugge. 2009. Tumor self-seeding: bidirectional flow of tumor cells. *Cell* 139: 1226-1228.
247. Edgerton, M. E., Y. L. Chuang, P. Macklin, W. Yang, E. L. Bearer, and V. Cristini. 2011. A novel, patient-specific mathematical pathology approach for assessment of surgical volume: application to ductal carcinoma *in situ* of the breast. *Anal Cell Pathol (Amst)* 34: 247-263.
248. Gerdes, M. J., C. J. Sevinsky, A. Sood, S. Adak, M. O. Bello, A. Bordwell, A. Can, A. Corwin, S. Dinn, R. J. Filkins, D. Hollman, V. Kamath, S. Kaanumalle, K. Kenny, M. Larsen, M. Lazare, Q. Li, C. Lowes, C. C. McCulloch, E. McDonough, M. C. Montalto, Z. Pang, J. Rittscher, A. Santamaria-Pang, B. D. Sarachan, M. L. Seel, A. Seppo, K. Shaikh, Y. Sui, J. Zhang, and F. Ginty. 2013. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc Natl Acad Sci U S A* 110: 11982-11987.
249. Meaburn, K. J. 2010. Fluorescence *in situ* hybridization on 3D cultures of tumor cells. *Methods Mol Biol* 659: 323-336.
250. Dey, S. S., L. Kester, B. Spanjaard, M. Bienko, and A. van Oudenaarden. 2015. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* 33: 285-289.

251. Davis, A., R. Gao, and N. Navin. 2017. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim Biophys Acta* 1867: 151-161.
252. Andor, N., T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 22: 105-113.
253. Cooke, S. L., J. Temple, S. Macarthur, M. A. Zahra, L. T. Tan, R. A. Crawford, C. K. Ng, M. Jimenez-Linan, E. Sala, and J. D. Brenton. 2011. Intra-tumour genetic heterogeneity and poor chemoradiotherapy response in cervical cancer. *Br J Cancer* 104: 361-368.
254. Cross, W., T. A. Graham, and N. A. Wright. 2016. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J Pathol* 240: 126-136.
255. Luria, S. E., and M. Delbruck. 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* 28: 491-511.
256. Navin, N. E. 2014. Tumor evolution in response to chemotherapy: phenotype versus genotype. *Cell Rep* 6: 417-419.
257. Leung, M. L., A. Davis, R. Gao, A. Casasent, Y. Wang, E. Sei, E. Vilar, D. Maru, S. Kopetz, and N. E. Navin. 2017. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res* 27: 1287-1299.
258. Li, W. V., and J. J. Li. 2018. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 9: 997.

## VITA

Anna K. Casasent received her high school diploma from Eisenhower High School in Yakima, Washington in 2006. Whereupon, she entered St. Edward's University in Austin, Texas in 2006 and received the degree of Bachelor of Science with a major in Bioinformatics on the Biomathematics Track in May 2010. For the next two years, she worked as an Associate Statistical Analyst in the Department of Bioinformatics and Computational Biology at University of Texas MD Anderson Cancer Center in Houston, Texas. In August of 2012 she entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.