Imperial College London
Faculty of Medicine
Department of Surgery & Cancer

Ph.D. Thesis

Epigenomic and genomic dissection of oestrogen receptor-positive breast cancer

Giacomo Corleone

CID:01155843

Supervisor
Dr. Luca Magnani

March 2019

Abstract

Genetic and Epigenetic mechanisms underlie the natural complexity of phenotypical and morphological aspects of cancer. Extensive phenotypical intra and inter-patient heterogeneity of cancer specimens is a relevant limiting factor in cancer studies. Besides, loss of cell identity and frequent phenotypic switches occurring during cancer progression are elusive and intrinsically challenging to study. Historically, significant efforts have been devoted to identifying the genetic alterations occurring within DNA coding regions of cancer cells. More recently, new studies point the attention on the strong interplay between the epigenetic layout of cells and non-coding regulatory network of DNA. Mutations and alterations of the epigenetic asset among these regions are frequent in cancer; however relevant questions about the different key players involved are still unsolved.

This dissertation aims to provide novel insights into the complex regulatory network of oestrogen receptor-positive breast cancer progression by mining the epigenetic and genetic landscape of patient tissues.

We depict the epigenomic landscape ERα-positive BC focusing on active promoters and enhancers. We describe a novel computational workflow which led to the identification of YY1 and SLC9A3R1 proteins. We show data supporting the concept of a phenotypical evolution of epigenetic traits during cancer progression and experimentally validate the findings.

Also, we introduce a new approach aiming to prioritise the investigation of sites prone to be critical players in breast cancer by coupling the systematic tracking of regulatory regions modulation during cancer progression with single nucleotide variants (SNV) discovery at high depth. This approach resulted with the identification of a recurrent mutation occurring in the core of a CTCF site lying at a border of a silent TAD in q25.1 of the chromosome 17.

Finally, we provide evidence of an epigenetic induced reprogramming occurring in a distinct subpopulation of cancer cell lines exhibiting high plasticity. These cells become quiescent upon oestrogen deprivation and resume proliferation after a deterministic epigenetic modulation of NOTCH target regions.

Declaration of Originality

I declare that I am the sole author of this dissertation and that all the work described is written in my own words. I fully observed the Imperial College Research Integrity Policy available at http://www.imperial.ac.uk/research-and-innovation/about-imperial-research/research-integrity.

Methods, data and processes described here are documented truthfully. I have not manipulated any result, and all persons and institutions who contributed and supported his work have been acknowledged.

Giacomo Corleone

January 2018

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non-Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

I would like to thank EU H2020 Marie Sklodowska-Curie Innovative Training Network under the Epipredict Consortium No 642691 and CRUK for having funded my research.

I would like to express my highest gratitude to my PI Dr Luca Magnani for having guided and supported me in the world of scientific research.

Special thanks to all the collaborators involved in this project Dr Darren K. Patten and Dr Sung Pil Hong for their work in providing me data.

Thank you to all my colleagues and the Magnani's group for their support, patience and encouragement.

I would like to thank all the patients who kindly support scientific research.

This thesis is dedicated to mum Giovanna, dad Antonio, Biagio, Stefano and Virginia for their endless love and support.

Table of Contents

ABSTRACT **2**

DECLARATION OF ORIGINALITY **3**

COPYRIGHT DECLARATION **4**

ACKNOWLEDGEMENTS **5**

LIST OF FIGURES **9**

LIST OF TABLES **12**

LIST OF ACRONYMS **12**

STRUCTURE OF THE THESIS **17**

CHAPTER 1 INTRODUCTION **19**

1.1 CANCER 19

1.2 THE HALLMARKS OF CANCER 22

1.3 BREAST CANCER 25

1.3.1 Epidemiology *25*

1.3.2 Risk Factors *26*

1.3.3 Mammary gland and oestrogens *29*

1.3.4 Breast Cancer classification *35*

1.3.5 Oestrogen Positive Breast Cancer therapy *36*

1.3.6 Genetics of Breast Cancer *39*

1.4 EPIGENETICS, ENHANCER AND CHROMATIN PACKAGING 44

1.5 MECHANISMS OF ENHANCERS REWIRING IN CANCER 50

1.6 TUMOUR HETEROGENEITY 52

CHAPTER 2 ..... **54**

EPIGENOMIC PROFILING OF ERA+ BREAST CANCER REVEALS NOVEL PRINCIPLES OF HETEROGENEITY AND PHENOTYPICAL EVOLUTION 54

AIMS OF THE STUDY 54

2.1 H3K27AC CHIP-SEQ OF 39 PRIMARIES AND 16 METASTATIC ERA POSITIVE SAMPLES 55

2.2 H3K27AC LANDSCAPE OF 47 ERA + BREAST CANCER TISSUES 61

2.3 EPIGENETIC INFORMATION AS A SURROGATE OF CELL PHENOTYPE OF ERA + BREAST CANCER TISSUES 63

2.4 A NOVEL METRIC FOR ASSESSING PHENOTYPICAL HETEROGENEITY 65

2.5 ASSESSMENT OF POTENTIAL SOURCE OF BIAS IN REGULATORY REGIONS MAPPING 67

2.5.1 CNA *68*

2.5.2 Dinucleotide artefacts *71*

2.5.3 H3K27ac marker during cell cycle *73*

2.5.4 Reproducibility of SI formulation *74*

2.5.3 Assessment of genetic alteration systematically affecting the results. *75*

2.6 PHENOTYPIC EVOLUTION DURING BC CANCER PROGRESSION AND THERAPY SELECTIVE PRESSURE 77

2.7 COMPREHENSIVE ASSESSMENT OF ERA + BREAST CANCER CISTROME 82

2.8 PAN-ANALYSIS OF YY1 IN NORMAL AND TUMOUR TISSUES 89

2.9 FUNCTIONAL CHARACTERISATION OF YY1 IN ERA+ BC 92

2.10 PROFILING OF YY1 BINDING AND ITS FUNCTIONAL EFFECT IN ERA+ BREAST CANCER 94

2.11 ROLE OF ERA, YY1 AND SLC9A3R1 GENE AS POTENTIAL PLAYERS IN ERA+ BC ENDOCRINE RESISTANCE 98

2.12 Experimental validation of YY1 and SLC9A3R1 phenotypic clonality in ERa+BC 103

CHAPTER 3 107

Non-coding genetic reprogramming during the development of endocrine resistance in ERa+BC 107

Aims of the study 107

3.1 SID-V1 109

FIGURE 74 THEORETICAL WORKFLOW OF SID-V1 DESIGN. 110

3.2 Investigation of the mutational landscape in SID-V1 panel 110

FIGURE 75. GOLDEN PATIENT CASE. 111

3.3 Investigation of the CTCF mutational landscape in SID-V1 panel 113

3.4 Functional investigation of the Somatic Regulatory Mutation number 5 118

3.5 SID-V2 120

CHAPTER 4 123

Aims of the study 123

4.1 Investigation of cell plasticity in cell lines and patients 124

4.2 Cell of origin model may explain the shift from drug sensitive to not sensitive condition 129

4.3 Computational assessment of CD44high and CD44low 132

CHAPTER 5 138

5.1 Computational methods chapter 2 138

5.1.1 H3K27ac ChIP-seq data processing 138

5.1.2 Assessment of heterogeneity 139

5.1.3 Functional characterisation of the peaks 139

5.1.4 Sharing Index (SI) 140

5.1.5 Ranking Index (RI) 141

5.1.5 Ranking Index in cancer cell lines and normal tissue epigenomes 141

5.1.6 Transcription factor profiling 142

5.1.7 Footprint analysis 143

5.1.8 Assessment of somatic copy number alteration (CNA) 144

5.1.9 Assessment of dinucleotide composition 144

5.1.10 Targeted capture essay processing 145

5.2 Brief experimental methods chapter 2 146

5.2.1 Tumour tissue processing 146

5.2.2 ChIP 146

5.2.3 ChIP–qPCR 147

5.2.4 Cell culture 147

5.2.5 siRNA 148

5.2.6 RNA extraction and real-time PCR 149

5.2.7 Luciferase reporter assay 149

5.3 Computational methods chapter 3 150

5.3.1 SID-V1 mutation profiling analysis 150

5.3.2 Mutation calling post-processing 150

5.3.3 CTCF mutation exploratory analysis 151

5.3.4 CTCF ChIP-seq peak calling 151

5.4 Brief experimental methods chapter 3 152

5.4.1 Targeted sequencing 152

5.4.2 CTCF ChIP-seq 152

5.4.3 CRISPR-Cas9 transfection 152

5.4.4 CRISPR Validation Using Sanger Sequencing 153

5.4.5 RT q-PCR *154*
5.5 Computational methods chapter 4 ..... *154*
5.5.1 ATAC-seq analysis *154*
5.5.2 Differential peaks analysis *155*
5.5.3 Data normalisation and hierarchical clustering analysis *155*
5.5.4 Identification of Differential Pathways and Transcription-Factor-Bound Regions *156*
5.6 BRIEF EXPERIMENTAL METHODS CHAPTER 4 156
5.6.1 Cell lines *156*
5.6.2 Plasmids *157*
5.6.3 FACS analysis *157*
5.6.4 Neo-adjuvant patient selection *158*
5.6.5 FAIRE-qPCR *158*
5.6.6 ATAC-seq *158*
CHAPTER 6 159
DISCUSSION AND CONCLUSIONS 159
CHAPTER2 159
CHAPTER 3 163
CHAPTER 4 165
CONCLUSIONS 167
REFERENCES 168
APPENDIX 182

List of Figures

Figure 1 Worldwide incidence of all cancer cases from 2005 to 2015. 19
Figure 2. The mortality rate for Breast Cancer in the USA from 1975 to 2015 by Ethnicity. 22
Figure 3. The hallmarks of cancer proposed in 2000. 23
Figure 4. The hallmarks of cancer proposed in 2011. 24
Figure 5. Age-standardised Breast Cancer death rates in UE and US. 25
Figure 6. The incidence of Breast Cancer diagnosis among relatives. 27
Figure 7. Sagittal section of a mammary gland. 29
Figure 8. Biochemical structure of the 4 major oestrogens. 30
Figure 9. Oestrogens biosynthesis. 31
Figure 10. Changes in concentration of oestradiol and progesterone during menstrual cycle. 32
Figure 11. Overview of ERα function. 33
Figure 12. Expression and distribution of ERα and ERβ in mammary gland. 34
Figure 13. Cluster diagram of Breast Cancer subtypes. 35
Figure 14. The absolute risk of Breast Cancer Relapse with and without radiotherapy following lumpectomy. 37
Figure 15. 15 years of Recurrence and Breast cancer mortality rate with and without Tamoxifen administration. 38
Figure 16. The effects of 10 years consecutive administration of AI. 39
Figure 17. Risk of Breast Cancer among BRCA1 and BRCA2 mutations. 40
Figure 18. Genetic portrait of tumour samples by mRNA subtype. 41
Figure 19 Copy number analysis of 560 Breast Cancer samples. 41
Figure 20. Example of evolutionary trajectories in Breast Cancer. 43
Figure 21. Example of the effects of histone modifications to chromatin accessibility. 45
Figure 22. . Association of histone post-translational modifications to cell regulatory activity. 45
Figure 23. Differential epigenomic reprogramming example. 46
Figure 24. FOXA1 is a key determinant of ERα binding. 47
Figure 25. . Contact probability of DNA intra and intrachromosomic regions. 48
Figure 26. Conservation of DNA sequences of promoters and enhancers across 20 species. 48
Figure 27. TAD borders are enriched of CTC and cohesion. 49
Figure 28. Risk-associated SNPs in Breast Cancer. 51
Figure 29. FOXA1 binding to heterozygous SNPs. 51
Figure 30. Tumour heterogeneity. 52
Figure 31. Significant H3K27ac peaks by sample. 56
Figure 32. Number of peaks in primary and metastatic sample group. 56
Figure 33 Saturation analysis of Enhancer and Promoters and sharing distribution in N=47 ERα Breast Cancer. 62
Figure 34. EGR3 transcriptomics, epigenomics and spike in-experiment. 64
Figure 35. EGR3 transcriptomics, epigenomics and spike in-experiment. 65
Figure 36. Enhancer clonality and SI relationship. 67
Figure 37. CNA profiles of samples. 69

Figure 38. CNA profile of chromosome 1. ..... **70**
Figure 39. Dinucleotide composition of the hg19 reference genome and sequenced genome. **72**
Figure 40. Dinucleotide content of each SI bin. **72**
Figure 41. H3K27ac marker during the cell cycle. **73**
Figure 42 Targeted sequencing for common genes affecting BC. **77**
Figure 43. Reproducibility of SI formulation. ..... Errore. Il segnalibro non è definito.
Figure 44. Theoretical framework of the analysis. **78**
Figure 45. Usage of enhancers and promoters in primary and metastatic patients. **79**
Figure 46. RI of differential enhancers in primary and metastatic patients. **79**
Figure 47. Survival analysis of PE and ME sets. **80**
Figure 48. Pathway analysis for gene association to PE and ME. **81**
Figure 49. Computational framework of the NFR identification. **83**
Figure 50 Overlap of DHS imputation with Honey Badger dataset. **84**
Figure 51. Imputed transcription factor analysis in promoter and metastatic. **84**
Figure 52. Transcription factor motif analysis heatmap applied to sharing index (SI) bins of enhancers and promoter. **86**
Figure 53. Footprint enrichment of enhancer in MCF7. ..... **87**
Figure 54. Overlap of imputed DHS regions with in vivo derived ERα binding sites. **87**
Figure 55. YY1 motif is the top enriched in clonal enhancers. **88**
Figure 56. Pan-cancer analysis of YY1 expression in healthy tissues. **90**
Figure 57.YY1 is higher express in ER+ ERα BC then normal tissues. **91**
Figure 58. IHC analysis of YY1 in Normal lobule, Duct andERα-positive invasive ductal carcinomas. **91**
Figure 59. Regulatory landscape of YY1 locus **92**
Figure 60. eKO YY1 experiment. **93**
Figure 61. YY1 RI in the patient dataset and normal tissues. **94**
Figure 62. YY1 ChIP-seq in MCF7 upon oestradiol induction **95**
Figure 63. Comparative ChIP-seq analysis between YY1+E2, ERα E2 and FOXA1+E2. ..... **96**
Figure 64. YY1 depletion experiments confirm a transcriptional shutdown **97**
Figure 65.SLC9A3R1 is the top prognostic gene for ERα+ERα BC relapse. **99**
Figure 66. SLC9A3R1 regulatory landscape. **99**
Figure 67. Pan-cancer analysis of SLC9A3R1 expression coupled with expression in cancer cell lines. **100**
Figure 68. SLC9A3R1 silencing experiments lead to proliferation arrest. **101**
Figure 69.SLC9A3R1 expression analysis in Metabric and TCGA and relative enhancer activity in ENCODE. **102**
Figure 70. IHC staining of YY1 in primary and metastatic ERα+BC tissues. **104**
Figure 71. IHC staining of SLC9A3R1 in primary and metastatic ERα+BC tissues. **104**
Figure 72. Phenotypical evolution of enhancers in primary and metastatic samples. **105**
Figure 73: Staining of YY1 and SLC9A3R1 in an independent matched longitudinal cohort of 22 ERα+ ERα BC. **106**
Figure 74 Theoretical workflow of SID-V1 design **110**
Figure 75. . Golden patient case **111**

Figure 76. Mutation profile of L1 and L2. ..... **112**
Figure 77. Sub clonal distribution. Of SNVs between L1 and L2 samples. Of SNVs. **112**
Figure 78. Pan-cancer analysis investigating the mutational profile of CTCF binding sites. **114**
Figure 79 CTCF window mutation profiling. **115**
Figure 80. CTCF conservation amongst vertebrates. **116**
Figure 81. Regulatory landscape of SRM5. ..... **117**
Figure 82. CTCF-ChIP-seq of the three mutated clones **118**
Figure 83. Transcriptional exploration by modulating stochastic reprogramming of 3D interactions **119**
Figure 84.SID-V2 selection strategy. **122**
Figure 85. MCF7 and LTED plasticity assays. **125**
Figure 86 qRT-PCR analysis of EMT genes. **125**
Figure 87 Clonogenic assay **126**
Figure 88. T47D and T47D-LTED plasticity assays. **127**
Figure 89. Cell plasticity is common in drug resistant ERα+ BC cells. **128**
Figure 90. Differential cell response to oestradiol. **130**
Figure 91 CD44high cells persist in patients with endocrine treatments. **131**
Figure 92 The computational strategy developed to identify differential DNA accessibility sites between CD44 high and low. **132**
Figure 93. CD44high cells are epigenetically distinct **133**
Figure 94. Downstream analysis of differential accessibility of CD44+ and CD44. ..... **134**
Figure 95 qRT-PCR analysis of Notch targets in sorted CD44high and CD44low from MCF7 ..... **135**
Figure 96. ELF1 Meta-analysis integrated with ENCODE ChIP-Seq of NOTCH targets. **136**
Figure 97 Experimental validation of siELF1. ..... **137**

List of Tables

Table 1: Cancer incidence rate ranking in the most 50 populous countries in 2015..... 21
Table 2. Estimates of new Breast Cancer cases and deaths in US. 26
Table 3. Proportions of families with disease due to BRCA1 and BRCA2. Estimated proportion of Breast Cancer cases linked to BRCA1, BRCA2 and other genes. 28
Table 4 Clinicopathological information for primary breast tumour samples collected and processed for H3K27ac ChIP-Seq. 57
Table 5 Clinicopathological information for metastatic breast tumour samples collected and processed for H3K27ac ChIP-Seq. 58
Table 6 Alignment statistics and peak calling processing statistics in primary tissues 59
Table 7. Alignment statistics and peak calling processing statistics in metastatic tissues. 60

List of Acronyms

15-L1 = First Biopsy year 2015 from left breast

15-L2 = Second Biopsy year 2015 from left breast

15-R1 = First Biopsy year 2015 from right breast

ATCC = American Tissue Type Culture Collection

AI = Aromatase inhibitor

bp = Base pair

BC = Breast cancer

ERα BC = Breast cancer

BCN = Breast Cancer Now

BCS = Breast conserving surgery

BMI = Body Mass Index

BRCA1 = Breast Cancer Type 1 susceptibility protein

BRCA2 = Breast Cancer Type 2 susceptibility protein

BSA = Bovine serum albumin

CB = Cholesterol biosynthesis

CD44 = Cell-surface glycoprotein

CE = Enhancer active in both primary and metastatic tissues

CGH = Comparative genomic hybridization assay

CGHFBC = Collaborative Group on Hormonal Factors in Breast Cancer

ChIA-PET = Chromatin interaction analysis by paired-end tag sequencing

ChIP = Chromatin immunoprecipitation

CNA = Copy Number Aberration

CpG = A Cytosine followed by a Guanine

CRUK = Cancer Research United Kingdom

CSC = Cancer stem cell

CTCF = CCCTC-Binding factor

DCFCS = Charcoal-stripped fetal calf serum

DCIS = Ductal carcinoma in situ

ddH2O = Double distilled water

DHS = DNase I hypersensitivity sites

DMEM = Dulbecco's Modified Eagle's Medium

DNMT = DNA methyl-transferase

E1 = Oestrone

E2 = Oestradiol

E3 = Oestriol

E4 = Estetrol

EAS = European age-standardised

EBCTCG = Early Breast Cancer Trialists Collaborative Group

ECACC = European Collection of Cell Cultures

ECM = Extracellular matrix

EMT = Endothelial Mesenchymal Transition

ENCODE = The Encyclopaedia of DNA Elements project

ER = Oestrogen receptor

ER+ = Oestrogen receptor positive

ER- = Oestrogen receptor negative

ERα = Oestrogen receptor alpha

ERβ = Oestrogen receptor beta

ERE = Oestrogen responsive element

ESR1 = Oestrogen receptor 1

ESR2 = Oestrogen receptor 2

ETOH = Ethanol

ETR = Endocrine Therapy Resistance

eRNA = Enhancer RNA

ET = Endocrine therapy

EU = European Union

FANTOM= Functional Annotation of the Mammalian genome project

FAIRE-qPCR = Formaldehyde-Assisted Isolation of Regulatory Elements

FCS = Fetal calf serum

FISH = Fluorescent in situ hybridization

FOXA-1 = Forkhead box A-1

GBD = Global Burden of Disease Study

GC = Guanine followed by a Cytosine

GFP = Green Fluorescent Protein

GREAT = Genomic Regions Enrichment of Annotations Tool

GRN = Gene-regulatory network

GTEx = Genotype-Tissue Expression project

H3K27ac = Acetylation of lysine 27 of histone H3

H3K4me2 = Dimethylation of lysine 4 of histone H3

H3K4me1 = Monomethylation of lysine 4 of histone H3

H3K4me3 = Trimethylation of lysine 4 of histone H3

H3K27me3 = Trimethylation of lysine 27 of histone H3

HAT = Histone acetylases

HDAK = Histone deacetylases

HDI = Human Development Index

HER2 = Human epidermal growth factor receptor 2

HPA = Human Protein Atlas

ICHTB = Imperial College Healthcare NHS trust Tissue Bank

ICR = Institute of Cancer Research

IDC = Invasive ductal carcinoma

IHC = Immunohistochemistry

ILC = Invasive lobular carcinoma

ITH = Inter-tumour heterogeneity

LR = Local recurrence

LTED = Long-term oestrogen deprived MCF7 cells

M = molar

ME = Enhancers active in metastatic tissues

μM = Micromolar

mM = Millimolar

Mb = Megabase

Mbase = Million bases

MCF7FulvR = MCF7 fulvestrant resistant cells

mESC = Murine embryonic stem cell

METABRIC = Molecular Taxonomy of Breast Cancer International Consortium

NCSM = Non-coding somatic mutation

NFR = Nucleosome-free region

O.S./OS = Overall survival

PBS = Phosphate buffered saline

PBST = Phosphate buffered saline tween

PBX1 = Pre-B cell leukaemia transcription factor 1

PE = Enhancer active in primary tissues

PF = Pioneer factor

POL II = RNA polymerase 2/II

PR/PGR = Progesterone receptor

qPCR = Quantitative polymerase chain reaction

PSG = 100 U penicillin/0.1 mg mL-1 streptomycin, 2mM L-glutamine

RI = Ranking index

RPKM = Reads per kilobase of transcript per million mapped reads

RT = Reverse transcriptase

RT-qPCR = Real time quantitative polymerase chain reaction

SERD = Selective oestrogen receptor degrader

SERM = Selective oestrogen receptor modulator

SI = Sharing index

SID-V1 = Systematic Identification of Variants 1

SID-V2 = Systematic Identification of Variants 2

SNP = Single nucleotide polymorphism

SNV = Single nucleotide variant

SRM = Somatic Regulatory Mutation

TAD = Topologically associated domain

TAM-R = Tamoxifen resistant

TCGA = The Cancer Genome Atlas

TF = Transcription factor

TNBC = Triple negative breast cancer

TNM = TNM (Tumour, Node, Metastasis) Classification of Malignant Tumours

TSS = Transcription start site

US = United States of America

Structure of the thesis

This paragraph is built with the purpose to facilitate the reader's understanding of the conceptual workflow described in the chapters of this thesis.

Chapter 1 recalls the current state of the art in the field of Breast Cancer research. An epidemiological overview of cancer is followed by a broader description of ERa+ Breast Cancer. Particular details are used to introduce important concepts of cancer genomics and epigenomics. The description of the factors contributing to cancer heterogeneity and the impact on cancer studies conclude this introductory section.

In **Chapter 2**, I present a novel computational method developed to dissect the complex epigenome of a large dataset of patient material. I outline the principles underlying data mining, explain the main advances over conventional analysis. The methods and results presented in this chapter were published in 2018:

Patten DK, Corleone G*, Gyorffy B, Perone Y, Slaven N, Barozzi I, Erdos E, Saiakhova A, Goddard K, Vingiani A, Shousha S, Pongor LS, Hadjiminas DJ, Schiavon G, Barry P, Palmieri C, Coombes RC, Scacheri P, Pruneri G, Magnani, 2018, Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer, Nature Medicine, Vol:24, ISSN:1078-8956, Pages:1469*

Patten, D. K., Corleone, G. & Magnani, L. Epigenome Editing. Methods Mol Biology Clifton N J 1767, 271–288 (2018).

Of note, the patient dataset (N=55 H3K27ac ChIP-seq) and the central experimental validation of the findings have been produced by Dr Darren K. Patten.

In **Chapter 3**, I present the analytical framework of a new experimental and computational application developed for the investigation of the existing links between genetic and epigenetic alterations. Also, I described the examination of the mutational profile on a rare case of ERa+ Breast Cancer using data obtained from biopsies at different stages of the disease.

The experimental procedures of this work have been mainly conducted by Dr Darren K. Patten and Neil Slaven.

In **Chapter 4**, I explore more in detail the role of epigenetic plasticity in supporting cancer seeding cells in driving drug resistance and metastatic progression. This work contains a robust experimental design coupled with a computational component for data processing and data reduction.

This work has been conducted with the strong support of Dr Sung Pil Hong, who performed all the experimental procedures and validation.

Finally, in **Chapter 5** I discuss the main findings of each chapter and provide an outlook of future possibilities for implementing the described approaches and methods.

Chapter 1 Introduction

1.1 Cancer

Cancer is a major healthcare problem. In 2016, 8.9 million people died of cancer, and 17.5 million new cases were diagnosed[1]. The incidence rate of new cases between 2005 and 2016 rose in nearly the totality of the world countries with peaks of more than more 20% new cases in Japan, France, Italy and Australia (Fig.1). GLOBOCAN2018 estimates suggest a 75-100% increase in incidence in low and medium HDI index countries by 2040[2].

Socioeconomic factors and gender differences strongly impact the rate of cancer diagnoses and clinical outcome. The odds of developing cancer during a lifetime are very high. In the highest developed countries, 50 % of men and 33% of women will develop cancer (Table1). This statistic changes significantly in low-income countries where the paucity of cancer screening and weaker observational intensities contribute to a more reduced number of cancer diagnoses[1,3].

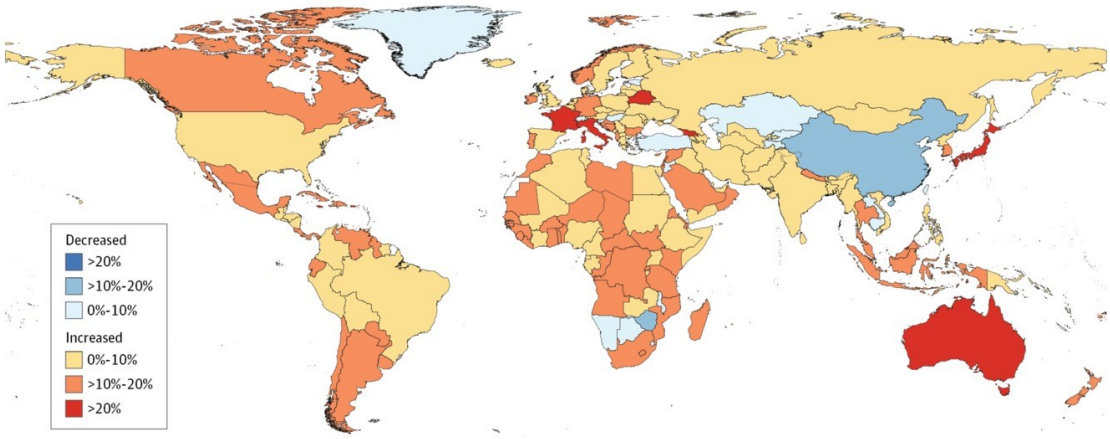


Figure 1 Worldwide incidence of all cancer cases from 2005 to 2015.

Data reflect relative changes of age-standardised incidence rates in both sexes in each country. Colour scheme reflects cancer incidence between 2005 to 2015 with blue representing >20% decrease and red >20% increase. [Figure adapted from GBDCC, Jama Oncology 2017]

Globally, the most common cancers are in order female breast, lung, colorectal and prostate cancer (Table1). Of note, many regional cancer burdens affect the world cancer incidence atlas. As an

example, South Asian countries (India, Pakistan, Bangladesh) are affected by an enormous number of lip and oral cavity cancer cases which make it the second in the ranking while is uncommon in other countries. In Australasia, malignant melanoma is the fourth most common cancer while is 16th in the global ranking[3,4] (Jama. Oncology ,2017, Torre CEPB 2016).

Mortality data show that cancer is the second leading cause of death worldwide. Lung cancer is by far the deadliest worldwide followed by colon-rectal, stomach, liver and breast cancer. Lung cancer accounts for 20% of cancer-related deaths in both sexes. Lung cancer is doubling the people dying of colorectal cancer, and notable, prostate cancer is the third-ranked in men while breast cancer the third in women[1–3].

Although these numbers depict a dire situation of cancer care, higher awareness in preventive medicine, new drug development and cure accessibility are contributing to lower the mortality rate worldwide. A remarkable example is the induction of tobacco control policies in western world countries in reducing the smoking rate among the population which resulted in a rapid drop of lung cancer hazard and improvement of life expectancy[5] . Besides, just in the United States, between 1989 and 2015 the mortality rate for breast cancer decreased by roughly 40% due to scientific research, improved screening and new therapies (Fig.2)[6].

Overall, these data suggest that cancer is the primary responsible for death worldwide but progresses in research, treatment and awareness are strongly ameliorating clinical outcome.

Country	Breast cancer	Tracheal, bronchus, and lung cancer	Colon and rectum cancer	Prostate cancer	Stomach cancer	Liver cancer	Non-Hodgkin lymphoma	Leukemia	Bladder cancer	Cervical cancer	Esophageal cancer	Uterine cancer	Pancreatic cancer	Kidney cancer	Lip and oral cavity cancer	Malignant skin melanoma	Thyroid cancer	Brain and nervous system cancer	Ovarian cancer	Larynx cancer	Chronic lymphoid leukemia	Acute myeloid leukemia	Gallbladder and biliary tract cancer	Other pharynx cancer	Acute lymphoid leukemia	Multiple myeloma	Nasopharynx cancer	Hodgkin lymphoma	Testicular cancer	Chronic myeloid leukemia	Mesothelioma
Global	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
High SDI	2	4	3	1	5	10	6	13	8	22	15	11	12	7	18	9	14	17	16	24	21	23	20	25	30	19	31	27	26	28	29
High-middle SDI	2	1	4	5	3	6	9	7	10	11	8	12	13	16	15	20	18	14	19	17	23	24	21	26	22	27	25	28	29	30	31
Middle SDI	1	2	5	9	3	4	10	8	12	7	6	13	15	23	11	26	17	14	20	16	21	22	25	24	18	27	19	28	30	29	31
Low-middle SDI	1	2	7	10	3	9	8	6	12	4	11	18	17	23	5	26	21	14	20	13	25	19	24	15	16	28	22	27	30	29	31
Low SDI	1	10	8	7	3	4	5	6	13	2	9	20	15	23	11	21	19	16	18	17	27	14	24	22	12	25	26	29	30	28	31
South Asia	1	3	5	10	4	14	8	6	12	7	9	19	16	28	2	27	22	15	18	11	23	17	24	13	20	25	21	26	30	29	31
India	1	4	6	10	3	14	8	7	13	5	9	20	15	28	2	25	22	16	18	11	23	17	24	12	19	26	21	27	30	29	31
Pakistan	1	3	5	16	9	13	4	6	8	17	7	11	20	22	2	29	21	12	14	10	27	19	25	15	18	23	24	28	26	30	31
Bangladesh	1	2	4	12	7	11	8	6	14	5	10	22	18	29	3	26	21	17	20	9	23	16	25	13	15	28	19	24	30	27	31
Nepal	1	5	9	15	3	11	8	4	16	2	12	22	17	29	6	27	19	18	20	7	23	14	25	10	13	28	21	24	30	26	31
East Asia	4	1	5	12	2	3	8	7	9	11	6	10	14	19	15	24	20	13	22	17	18	25	23	27	21	26	16	28	31	29	30
China	4	1	5	12	2	3	8	7	9	11	6	10	14	20	15	24	19	13	23	17	18	25	22	27	21	26	16	28	31	29	30
Southeast Asia	1	2	3	9	7	4	8	6	15	5	20	13	16	19	10	26	12	17	14	21	25	23	24	22	11	29	18	27	30	28	31
Indonesia	1	9	7	5	4	8	6	3	13	2	20	10	16	21	15	26	14	17	12	19	24	22	25	23	11	29	18	27	31	28	30
Philippines	1	2	3	6	11	7	10	4	21	5	25	12	19	20	9	22	8	16	15	18	24	14	27	26	13	28	17	29	30	23	31
Vietnam	2	1	3	10	4	5	6	8	17	11	12	16	14	22	7	27	15	20	13	21	29	23	24	9	19	28	18	26	25	30	31
Thailand	2	1	4	8	9	3	11	5	14	7	18	21	17	13	10	28	15	16	20	19	25	23	12	22	6	29	24	27	30	26	31
Myanmar	1	3	5	14	4	8	9	6	16	2	25	12	15	24	7	26	11	17	13	21	23	20	22	19	10	29	18	27	30	28	31
Malaysia	1	2	3	6	7	9	5	4	15	8	20	16	17	19	13	24	12	21	14	18	23	22	28	25	11	26	10	29	30	27	31
North Africa and Middle East	1	2	5	6	3	10	8	4	7	11	22	18	14	20	19	24	16	9	21	17	13	12	23	29	15	27	28	26	30	25	31
Egypt	1	5	7	9	8	4	13	2	3	18	23	17	16	20	14	26	15	6	21	19	10	12	24	25	11	27	30	29	31	22	28
Iran	2	5	6	3	1	14	12	4	7	17	9	25	19	21	23	16	20	8	24	10	13	11	22	30	15	27	28	26	29	18	31
Turkey	2	1	4	5	3	14	7	6	8	21	26	13	10	17	22	19	15	9	18	16	12	11	24	31	20	23	28	27	25	29	30
Sudan	1	4	6	12	2	7	5	3	14	8	19	23	15	25	16	27	13	9	24	18	17	11	22	26	10	29	28	21	30	20	31
Algeria	1	2	4	8	5	15	7	3	9	6	28	20	14	25	23	26	18	10	22	19	17	13	11	29	16	24	12	21	31	27	30
Iraq	1	3	8	18	6	11	10	2	7	14	25	17	15	19	16	24	13	4	20	22	12	5	26	23	9	27	28	29	30	21	31
Morocco	1	2	7	5	3	10	9	4	8	6	27	12	14	25	15	24	17	11	21	18	16	13	19	23	20	26	22	29	31	28	30
Afghanistan	1	5	8	17	2	6	7	4	12	3	15	19	23	28	20	27	10	9	25	21	18	14	16	24	11	29	26	13	31	22	30
Saudi Arabia	1	4	2	6	8	7	5	3	10	20	23	15	12	11	17	28	13	9	19	16	18	14	25	29	21	24	26	27	30	22	31
Yemen	1	4	7	14	2	8	6	3	13	5	20	23	17	28	16	26	12	9	24	21	15	11	22	25	10	29	27	18	31	19	30
Western Europe	2	4	3	1	9	12	6	10	5	22	18	13	11	7	17	8	14	15	16	25	19	24	21	23	31	20	30	29	26	28	27
Germany	1	4	3	2	7	13	6	11	9	23	20	12	10	8	18	5	15	16	14	25	21	24	19	22	30	17	31	29	26	28	27
France	2	4	3	1	10	12	5	8	9	24	15	16	13	6	14	7	11	18	20	22	17	23	25	19	31	21	28	29	26	27	30
United Kingdom	1	3	2	4	9	15	7	8	11	23	12	13	10	5	19	6	22	17	14	26	16	20	24	25	31	18	30	28	27	29	21
Italy	2	4	3	1	6	8	7	12	5	22	24	13	11	10	20	14	9	15	18	21	17	23	19	25	31	16	30	29	27	26	28
England	1	4	2	3	9	15	7	8	11	23	12	13	10	5	19	6	22	17	14	27	16	20	24	25	31	18	30	29	26	28	21
Spain	3	4	2	1	7	9	8	10	5	20	24	12	13	6	15	11	16	14	17	19	18	25	22	23	30	21	29	28	27	26	31
Western Sub-Saharan Africa	1	8	9	6	3	4	5	7	10	2	17	20	11	15	16	21	23	14	22	18	27	12	19	24	13	25	29	26	30	28	31
Nigeria	1	8	9	6	3	5	2	7	12	4	17	22	11	10	16	21	24	15	20	18	26	13	19	23	14	25	29	27	30	28	31
Ghana	1	12	8	4	6	5	2	7	10	3	19	16	9	17	14	24	25	11	20	18	26	15	23	21	13	22	29	28	30	27	31
Eastern Sub-Saharan Africa	1	11	9	7	4	5	3	8	13	2	6	19	16	25	10	21	20	15	14	18	27	17	26	24	12	23	22	30	29	28	31
Ethiopia	1	11	6	8	4	3	5	7	15	2	9	19	14	26	10	21	18	17	13	20	27	16	25	23	12	24	22	30	29	28	31
Tanzania	1	10	8	6	7	4	3	5	14	2	9	21	17	23	12	20	19	13	15	18	27	16	26	25	11	22	24	29	31	28	30
Kenya	2	17	8	4	3	7	6	5	19	1	10	20	14	26	9	24	22	13	12	11	27	16	23	25	15	21	18	30	29	28	31
Uganda	1	10	6	4	9	7	3	8	17	2	5	18	15	22	13	21	20	14	11	23	27	16	26	25	12	24	19	30	29	28	31
Mozambique	1	9	7	12	4	3	8	5	15	2	6	20	17	24	10	22	19	14	16	18	26	13	25	21	11	23	28	29	31	27	30
High-income North America	2	3	4	1	13	14	5	11	7	22	21	10	12	9	15	6	8	17	19	23	18	20	26	24	29	16	31	27	25	30	28
United States	2	3	4	1	14	13	5	11	7	23	21	9	12	10	15	6	8	17	19	22	18	20	26	24	30	16	31	27	25	29	28
Canada	2	4	3	1	7	17	5	10	9	23	21	11	12	6	20	8	13	14	18	24	16	19	22	25	29	15	31	27	26	30	28
Central Latin America	1	6	5	2	3	9	8	7	16	4	25	15	12	10	19	18	13	17	14	22	23	21	20	29	11	26	30	27	24	28	31
Mexico	1	7	4	2	5	9	8	6	16	3	27	15	14	10	21	18	11	17	12	24	23	22	19	29	13	26	30	25	20	28	31
Colombia	1	4	5	3	2	9	6	8	17	7	24	18	10	20	16	13	11	12	14	21	22	23	19	29	15	25	30	27	28	26	31
Venezuela	1	4	5	2	6	14	7	8	11	3	26	9	12	10	16	20	17	18	13	15	23	21	22	29	19	25	30	24	28	27	31
Eastern Europe	2	3	1	5	4	15	12	10	9	14	19	6	8	7	17	13	24	16	11	18	20	26	22	21	29	23	30	25	27	28	31
Russia	2	3	1	6	5	14	9	10	11	15	19	4	8	7	17	12	24	16	13	20	18	26	25	23	28	22	30	21	27	29	31
Ukraine	3	2	1	5	4	17	18	12	8	11	19	6	7	10	14	13	23	16	9	15	30	22	20	21	28	24	31	27	29	26	25
Tropical Latin America	1	4	3	2	5	16	7	8	10	6	11	18	13	19	9	14	17	12	21	15	25	22	23	20	26	24	30	27	29	28	31
Brazil	1	4	3	2	5	16	7	9	10	6	11	18	13	19	8	14	17	12	21	15	25	22	23	20	26	24	30	27	29	28	31
High-income Asia Pacific	5	3	1	4	2	6	9	14	13	16	12	15	7	8	18	28	10	20	17	23	25	19	11	22	24	21	30	31	27	26	29
Japan	5	3	1	4	2	6	9	14	12	18	11	15	7	8	16	29	13	20	17	23	27	19	10	22	24	21	30	31	25	28	26
South Korea	5	3	2	7	1	4	10	12	13	14	15	16	8	11	19	27	6	18	17	22	25	20	9	24	23	21	28	29	30	26	31
Central Europe	3	1	2	4	6	15	11	10	5	14	23	8	9	7	17	12	25	13	16	18	19	26	21	20	29	24	30	27	22	28	31
Poland	2	1	3	4	8	20	13	11	5	14	21	7	9	6	17	15	24	12	10	16	18	26	19	22	29	23	31	27	25	28	30
Central Sub-Saharan Africa	1	6	7	9	4	3	8	5	15	2	10	21	14	22	11	20	19	17	18	12	25	16	24	23	13	26	27	28	30	29	31
Democratic Republic of the Congo	1	6	8	10	3	4	7	5	16	2	9	20	14	25	11	21	18	17	19	12	24	15	23	22	13	26	27	28	30	29	31
Central Asia	1	3	4	7	2	9	13	6	11	5	10	8	16	17	14	19	23	12	18	15	22	20	24	25	21	29	30	26	27	28	31
Uzbekistan	1	3	6	11	2	13	8	4	14	5	7	15	20	18	12	19	26	9	22	10	21	16	24	25	17	30	27	23	29	28	31
Southern Sub-Saharan Africa	1	4	5	2	7	9	12	8	11	3	6	16	13	22	10	14	20	21	17	15	23	19	25	26	18	24	28	27	31	30	29
South Africa	1	4	5	2	7	14	13	8	10	3	6	16	12	21	9	11	18	22	17	15	24	19	25	26	20	23	28	27	31	30	29
Southern Latin America	1	4	3	2	5	17	13	11	10	6	15	14	8	9	19	16	12	20	18	22	23	21	7	29	26	25	31	28	24	27	30
Argentina	1	4	3	2	5	18	12	11	8	6	15	13	7	9	17	16	10	21	19	20	23	22	14	29	26	25	31	28	24	27	30
Andean Latin America	1	7	5	3	2	10	6	8	15	4	24	9	13	16	21	20	12	14	19	22	25	18	11	26	17	23	30	29	28	27	31
Peru	1	6	5	2	3	9	7	8	16	4	25	10	11	15	21	19	12	14	17	22	23	20	13	26	18	24	30	29	27	28	31
Caribbean	2	4	3	1	6	11	8	9	10	5	16	7	14	15	12	22	18	17	19	13	21	23	25	24	26	20	29	27	30	28	31
Australasia	3	5	2	1	8	15	6	10	12	25	17	9	11	7	23	4	13	14	18	26	19	20	22	21	28	16	31	30	27	29	24
Oceania	1	4	7	10	3	6	9	5	18	2	24	8	16	28	11	22	13	23	20	15	29	12	25	14	21	27	19	26	30	17	31

Table 1: Cancer incidence rate ranking in the most 50 populous countries in 2015.

Colour scheme from red to green where red is top-rank and green is low-rank value. [Table adapted from GBDCC, Jama Oncology 2017]

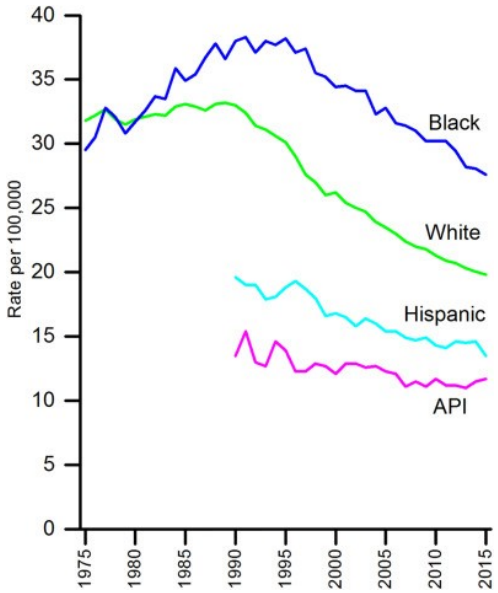


Figure 2. The mortality rate for Breast Cancer in the USA from 1975 to 2015 by Ethnicity.

Rates are per 100k females and age-adjusted. [Figure adapted from De Santis et al., Cancer J.Clin 2017]

1.2 The Hallmarks of Cancer

Cancer is a heterogeneous expansion of a population of cells which have acquired the ability to evolve independently from the tissue of origin. In 2000, Hanahan and Weinberg described the biological principles driving a cell from being normal into malignant[7] (Fig.3). They described cancer development as a multi-step process that requires distinct biological capabilities. Cancer needs a self-sufficient expression of growth factors which sustain hyper and limitless proliferation coupled with the ability to receive nutrients from blood with abnormal angiogenesis. Then, it requires the capacity to overcome self-defence anti-growth signals and cell death mechanisms. The combination of all these capabilities leads to the growth of benign tumour mass from which cancer cells may escape to colonise new tissues and then, to develop metastasis which drives the patient to death.

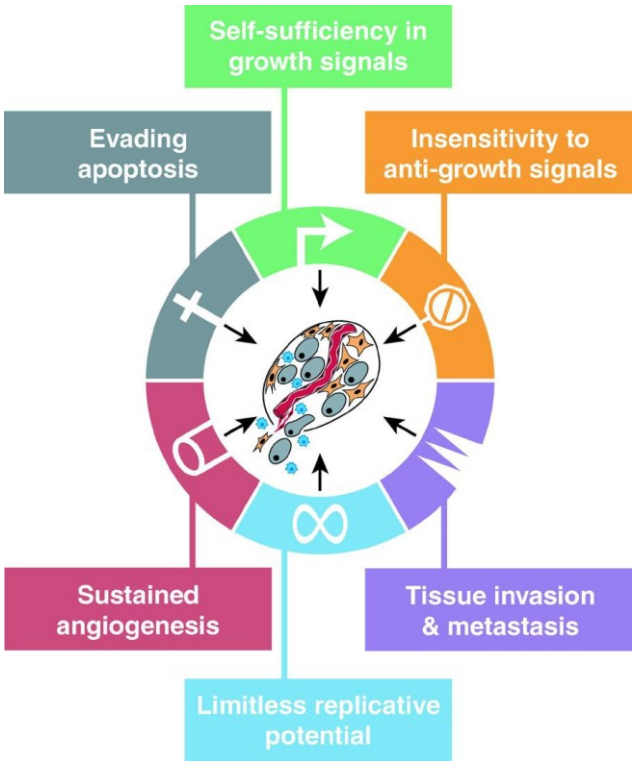


Figure 3. The hallmarks of cancer proposed in 2000.

The figure proposes the six hallmarks of cancer. [Figure adapted from Hanahan and Weinberg, Cell 2000]

This description was updated in 2011 followed by recent discoveries in the field of cancer metabolomics and cancer immunology[8] (Fig.4). Indeed, cancer cells show a progressive energy metabolism deregulation which sustains the metabolic demand for cell growth and proliferation. Then, although the immune system represents a barrier to tumorigenesis and actively contribute to tumour eradication, cancer cells are capable of evading immune surveillance.

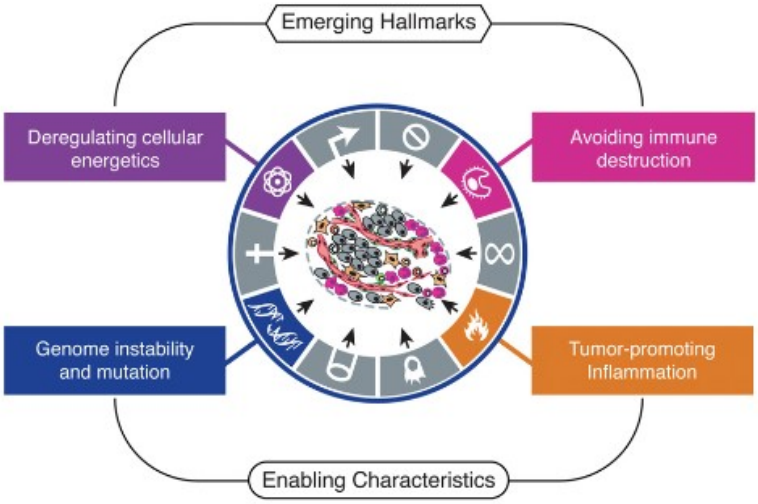


Figure 4. The hallmarks of cancer proposed in 2011.

Biological traits outlining the main characteristics shared in common by all human cancers. [Figure adapted from Hanahan and Weinberg, Cell 2011]

Overall, cancer is depicted by 10 hallmarks: 1. Self-sufficiency in growth signals 2. Unlimited replicative potential 3. Insensitivity to anti-growth signals 4. Opposition to cell death 5. Abnormal angiogenesis 6. Tissue invasion and metastasis 7. Deregulation of energy metabolism 8. Evasion of immune surveillance 9. Genomic instability 10. Tumour-specific inflammation.

Chronic inflammation and genome instability trigger the events associated with cancer development. The tumor-associated inflammatory response has the paradoxical effect of enhancing tumorigenesis and progression while genetic events such as non-silent mutations, copy number aberration and genomic doublings underlie each hallmark of cancer.

Besides, an emerging role underlying hallmark of cancer is played by Epigenomics. Accumulating evidence demonstrate the fundamental role of epigenetic events in cancer, and thus, it has been already proposed to add epigenetic plasticity to the hallmarks of cancer list[9] .

1.3 Breast Cancer

1.3.1 Epidemiology

BC is the leading malignancy and cause of cancer death among women worldwide. The European Cancer Observatory, in 2012, estimated 462000 women were diagnosed, and 131000 died from BC in Europe only. BC hit both sexes, but just 1% of the cases occur in men. Indeed, 1 out of 8 women will be diagnosed with BC during their lifetime while only 1 in 870 men. (Cancer Research UK data). BC accounts for roughly 30% of new cancer cases in females per year and 17% of cancer-related deaths. The incidence rate of BC cases is steadily increasing in all European countries. However, the death rate is decreasing. Indeed, in the EU, a drop of 11.8% of BC-related death cases (Fig.5) is expected compared to the observed number in 2012 with 32000 cases per year avoided. The same trend is expected in United States[10,11] .

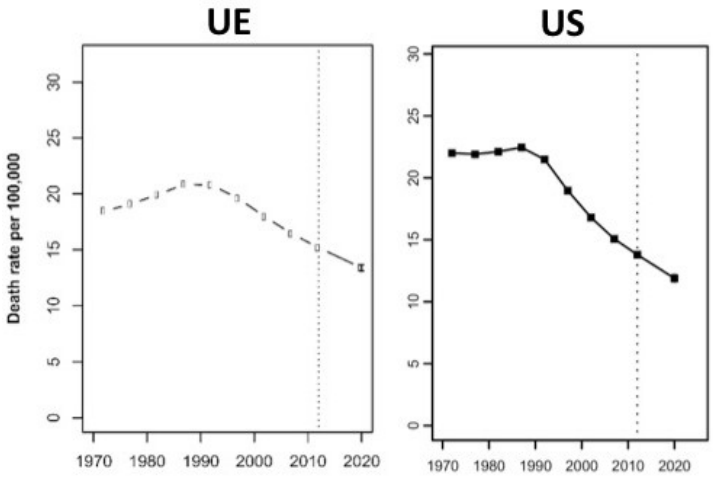


Figure 5. Age-standardised Breast Cancer death rates in EU and USA.

The illustration shows death rates from 1970 to 2012 and predicted rates for 2020. [Figure adapted from Carioli et al., The Breast 2018]

1.3.2 Risk Factors

Extensive epidemiological studies[12,13] and patient screening have been performed during recent years leading to the identification of BC risk factors routinely used to support the choice of screening in the general population. Established risk factors of BC are age, oestrogen exposure, family history, genetic predisposition, breast tissue density and lifestyle/environment.

Age is the primary risk for a BC diagnosis. Post-menopausal women have a higher chance to be diagnosed with BC while only 3% of cases affect woman less than 40 years old (Table2) (American cancer society 2017 Facts and Sheets).

	IN SITU CASES		INVASIVE CASES		DEATHS	
AGE, Y	NO.	%	NO.	%	NO.	%
<40	1,610	3%	11,160	4%	990	2%
40-49	12,440	20%	36,920	15%	3,480	9%
50-59	17,680	28%	58,620	23%	7,590	19%
60-69	17,550	28%	68,070	27%	9,420	23%
70-79	10,370	16%	47,860	19%	8,220	20%
80+	3,760	6%	30,080	12%	10,910	27%
All ages	**63,410**		**252,710**		**40,610**	

Table 2. Estimates of new Breast Cancer cases and deaths in US.

Data are reported by age. [Table adapted from American Cancer Society, 2017]

Since 80', a higher cancer risk has been associated with the exposure to oestrogens[14]. Early-age menarche, late age menopause, hormonal therapy, hormonal contraception exposes women to oestradiol and progestin. Extensive epidemiological studies[15] showed a modest absolute increase among women of these categories.

A higher risk of BC diagnoses is shown within woman appertaining to families with previous cases compared with women with no relatives affected. BC risk increases by the total number of cases within the family[16,17] (Fig.6).

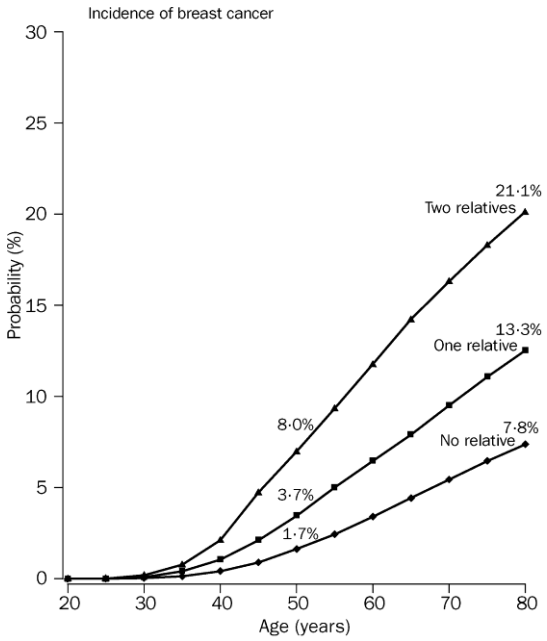


Figure 6. The incidence of Breast Cancer diagnosis among relatives.

Probability (%) that a disease-free woman develops the disease according to the number of relatives affected. [Figure adapted from CGHFBC, Lancet 2001]

However, only 10% of women developing BC have affected relatives. Genetic predisposition may explain an inevitable component of this moderated risk with inherited mutations and SNPs. Strong evidence proves that germline loss of function mutations in the coding region of BRCA1, BRCA2 genes explain BC in 60% of the cases within the same family[18] (Table3) while the germ-line loss of function mutation of PALB2 gives 14-30% increased risk[19]. Furthermore, numerous efforts have been deployed in identifying BC associated loci, but still, the genetic contribution to the disease is largely unknown[20].

Currently, many gene panel tests are available[12] to evaluate the patient risk of BC, and GWAS studies[21] are progressively revealing variants associated with BC risk. The most likely associated BC SNPs are on distal regulatory regions, and currently, 313 BC risk loci are known[22] however, this number is not yet commonly accepted and different risk SNPs are continuously proposed[23,24]. Further

functional annotation studies on identified loci and comprehensive understanding of the mechanism by which a given variant influence cancer risk are undoubtedly needed.

Heterogeneity Analysis under the CASH Model

Family Group	No. of Families	Proportion (95% CI) of Families with Disease Linked to: BRCA1	BRCA2	Other	LOD Score
All families:					
Four or five breast cancers, male or female	154	.55 (.40–.70)	.12 (.00–.29)	.33 (.14–.53)	19.28
Six or more breast cancers	83	.46 (.32–.60)	.50 (.35–.65)	.04 (.00–.17)	67.44
Overall	237	.52 (.42–.63)	.35 (.24–.46)	.13 (.03–.25)	84.08
All families, no males:					
Four or five breast cancers	140	.61 (.44–.76)	.08 (.00–.25)	.31 (.12–.53)	19.90
Six or more breast cancers	71	.50 (.34–.65)	.44 (.28–.61)	.06 (.00–.21)	50.40
Overall	211	.57 (.45–.68)	.28 (.18–.40)	.15 (.04–.28)	68.13
Male breast cancer	26	.19 (.01–.47)	.77 (.43–.97)	.04 (.00–.42)	18.03
Breast-ovarian cancer, no males:					
One ovarian cancer	42	.69 (.46–.89)	.19 (.04–.41)	.12 (.00–.33)	15.11
Two or more ovarian cancers	52	.88 (.69–.98)	.12 (.02–.31)	.00 (.00–.15)	44.07
Overall	94	.80 (.66–.92)	.15 (.05–.28)	.05 (.00–.17)	58.57
Female breast cancer only:					
Four or five breast cancers	83	.32 (.10–.55)	.09 (.00–.35)	.59 (.26–.89)	2.29
Six or more breast cancers	34	.19 (.04–.41)	.66 (.39–.88)	.15 (.00–.44)	14.18
Overall	117	.28 (.13–.45)	.37 (.20–.56)	.35 (.14–.57)	14.47

Table 3. Proportions of families with disease due to BRCA1 and BRCA2. Estimated proportion of Breast Cancer cases linked to BRCA1, BRCA2 and other genes.

[Table adapted from Ford et al., Am.J. Hum.Genet.1998]

An additional contribution to BC risk is provided by the density of epithelium and stroma of breasts. Dense breast tissue is associated with 4 to 6-fold higher risk BC compared to low-density tissue[25,26]. Besides, obesity, inadequate physical activity and poor diet typical of the urban lifestyle increase the incidence of BC[2]. Indeed, there is compelling evidence of reduction of incidence of BC in women who exercise at moderate intensity and with a regular BMI[27].

1.3.3 Mammary gland and oestrogens

The vast majority of BC cases arise from the epithelial cells of the mammary gland. The mammary gland is a branching network of epithelial cells which form ducts embedded in adipocytes, endothelial cells and fibroblasts that together constitute the mammary stroma (Fig.7). During late pregnancy, a rapid proliferation of epithelium terminates with the development of alveolus, aggregated in lobes. This process is then followed by the lactogenic switch[28]. The mammary epithelium has a layered structure. The inner layer consists of luminal epithelial cells and the outer layer of myoepithelial cells[29] .

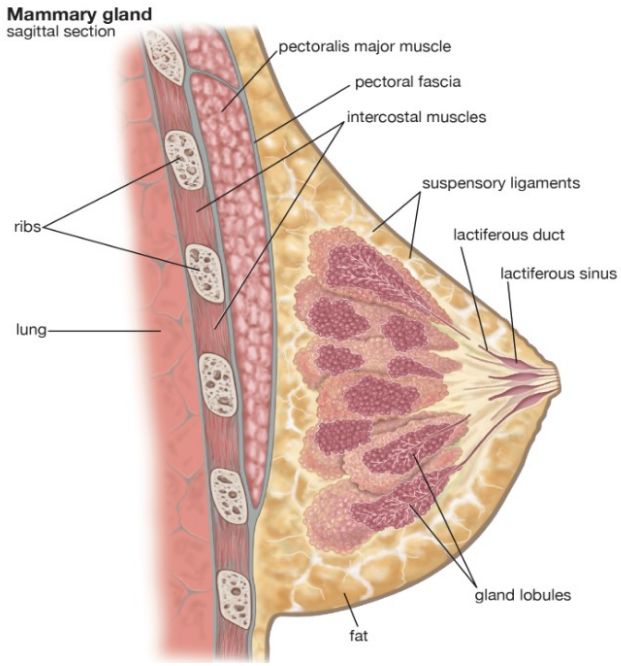


Figure 7. Sagittal section of a mammary gland.

The illustration shows the structure of the mammary gland. Gland lobules are embedded in an adipose structure and connected to the nipple with the lactiferous ducts. [Encyclopædia Britannica. Inc.]

Steroid hormones are primary regulators of biological processes and mammary gland development. They have a fundamental role in modulating most of the physiological events in vertebrates. They are secreted by the adrenal cortex, gonads, and placenta during pregnancy. Steroid hormones are

classified into five groups: glucocorticoids, mineralocorticoids, androgens, oestrogens and progesterone.

Here we describe the role of oestrogens cause their profound impact on the physiology of breast development and role in BC. Oestrogens belongs to the family of steroids hormone. Oestrogens regulate the development of secondary sexual female characteristics which include breasts, hips, distribution of adipose tissue and are the primary hormones in controlling the menstrual cycle. Furthermore, oestrogens control lactation, skeleton formation and affect the skin, hair and mucous membranes.

Four major oestrogens are produced in women (Fig. 8.): 1. estrone (E1) 2. 17β-oestradiol (E2) 3. estriol (E3) and estetrol (E4). Oestradiol is the main oestrogen of reproductive age and is synthesised mainly in ovaries. Estrone is produced in peripheral tissues and becomes the primary oestrogens during menopause. Estriol production is limited at the pregnancy while estetrol at the fetal life. Estriol is synthesised into the fetal placenta [30] while estetrol is produced in the fetal liver.

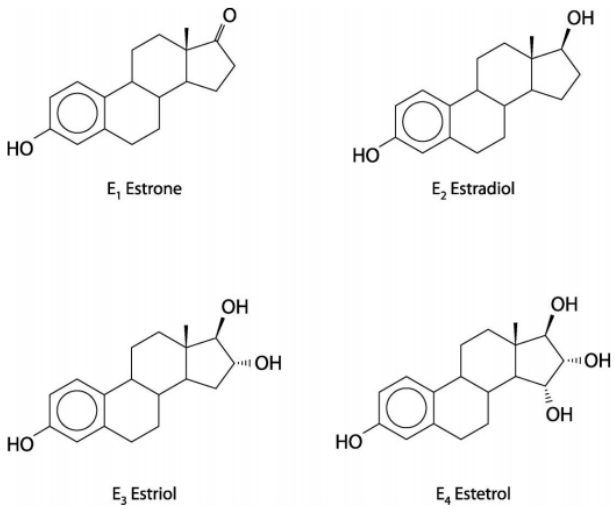


Figure 8. Biochemical structure of the 4 major oestrogens.

All oestrogens derive from cholesterol biosynthesis. [Encyclopædia Britannica. Inc.]

Regardless of the site of production, oestrogens are produced by a series of reactions having the androgens testosterone or androstenedione (both derived from cholesterol biosynthesis) as the

primary precursors. The enzyme aromatase P450 exerts a fundamental role in this process. P450 is derived from the gene CYP19 and catalyses the conversion of androgens in oestrogens[31] (Fig. 9).

Oestradiol (E2) is the most produced and active oestrogen in women. E2 is produced by the theca and granulosa cells of ovaries and circulates to distal tissues. The production varies cyclically during the menstrual cycle with the highest rate during the preovulatory phase while in the pre-menstrual period reach the lowest concentration (Fig.10). In the postmenopausal period, following the depletion of ovarian follicles, the E2 production is shallow (same level of childhood) and the site of production changes to extragonadal sites, such as adipose tissues, bone and nervous system, where it acts as a paracrine factor.

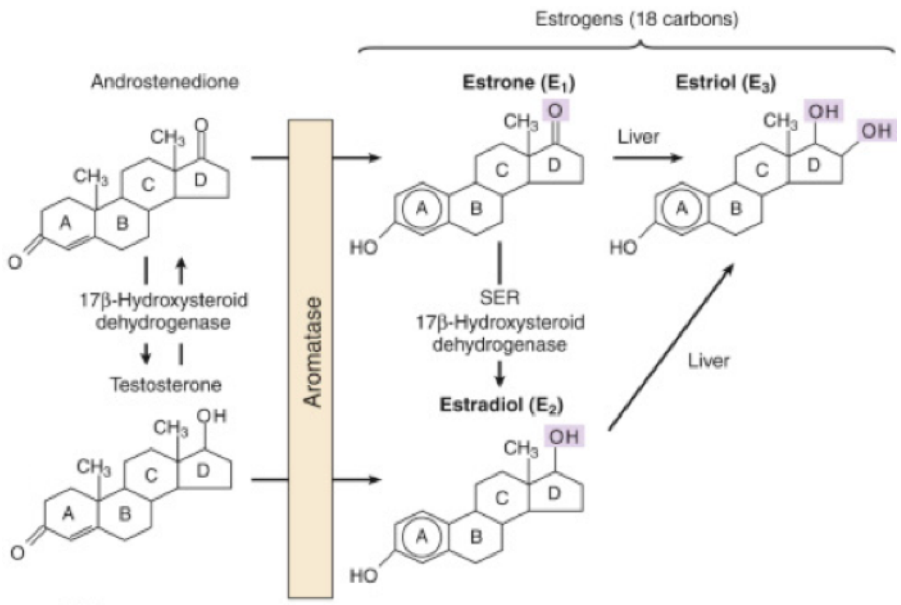


Figure 9. Oestrogens biosynthesis.

Oestrogens are obtained through a series of reactions catalysed by the enzyme aromatase P450. [Figure adapted from Wardell et al., Osteoporosis 2013]

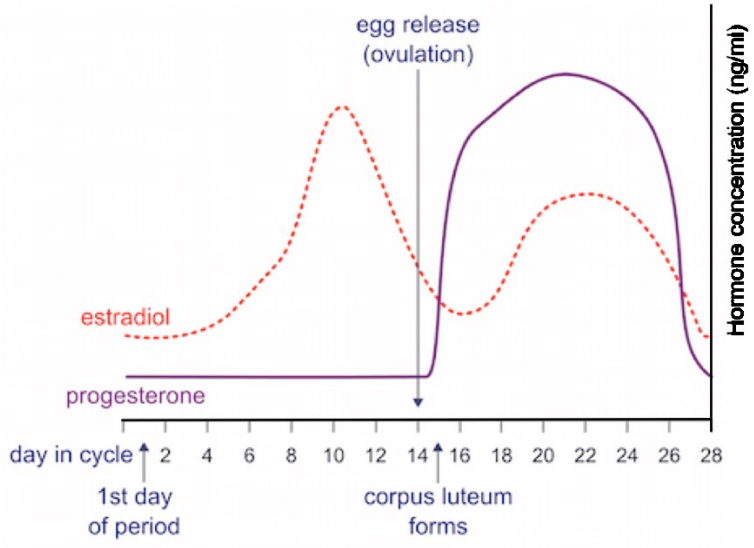


Figure 10.. Changes in concentration of oestradiol and progesterone during menstrual cycle.

Oestradiol production reaches his peak in the preovulatory phase of menstrual cycle followed by a drop and a further increase at day 20. [Figure adapted from Draper et al., Scientific Reports 2018]

Two distinct intracellular receptors, oestrogen receptors (ER)s, ERα (ER1) and ERβ (ER2), mediate the biological effects of E2. ERs are ligand-activated transcription regulators encoded by distinct genes, respectively ESR1 and ESR2. They mainly exert their function upon binding with oestrogens followed by dimerisation, migration to the nucleus and interaction with specific DNA sequences called oestrogen-response elements (EREs). 75-80k EREs elements are reported in the human genome[32]. They are cis-regulatory elements positioned all over the genome having the role in modulating transcriptional events. Around one-third of EREs are within 50kb to genes responsive to oestradiol. The majority of EREs are associated with enhancer elements which may be located over than 50kb to the TSS or in different chromosomes[33](Fig.11).

Interestingly a significant number of EREs are likely to be bound by ESRs without having any functional response to the interaction[33]. Besides, other different modes of transcriptional modulation have been reported[34]. The interaction with EREs modulate gene expression and ultimately results in the biological effects of oestrogens.

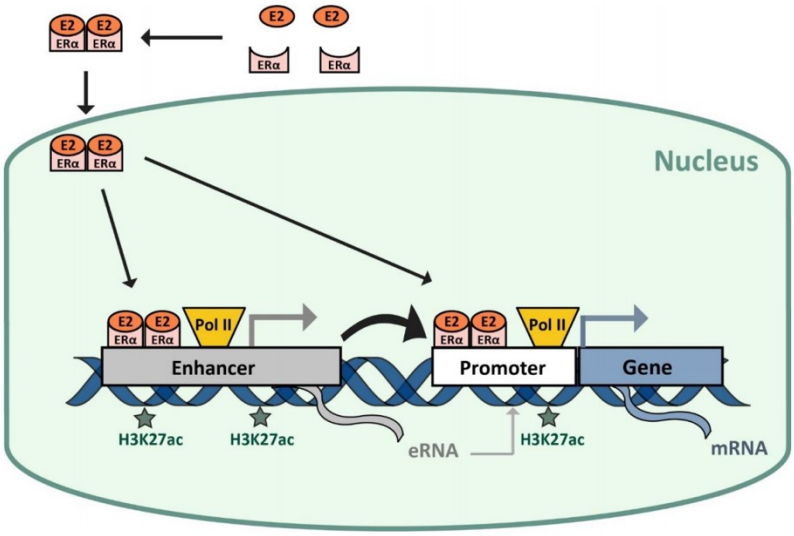


Figure 11.Overview of ERα function.

ERα binds with E2, dimerise and translocate to the nucleus. The complex ERα-E2 binds to the ERE elements and contribute to triggering RNA polymerase II (Pol II) action. H3K27ac marked histones typically flanks ERE elements. [Figure created by DK. Patten and N. Trasanidis]

ERα is predominantly expressed in the breast, uterus, cervix, while ERβ is primarily detected in the ovary, breast, prostate, testis, spleen, lung, hypothalamus, and thymus[35]. In the breast, ERα is the central mediator of ductal elongation, breast growth during puberty and menstrual cycle. ERα localises predominantly in the nuclei of epithelial cells embedding ducts and lobules. Instead, ERβ localises strongly in the nuclei of myoepithelial cells tissue[36] (Fig.12). Normal human and mouse mammary gland are more enriched of ERβ than ERα[37] however ERα KO mice are completely infertile and with atrophic mammary gland while, on the contrary, mice lacking ERβ show normal phenotype[38]. Furthermore, ERα is expressed in hyperproliferative BC cells, and almost the totality of BC cells is ERα positive[39] while ERβ expression decreases. Accumulating evidence report the pro-oncogenic function of ERα in BC while the role of ERβ is still under debate.

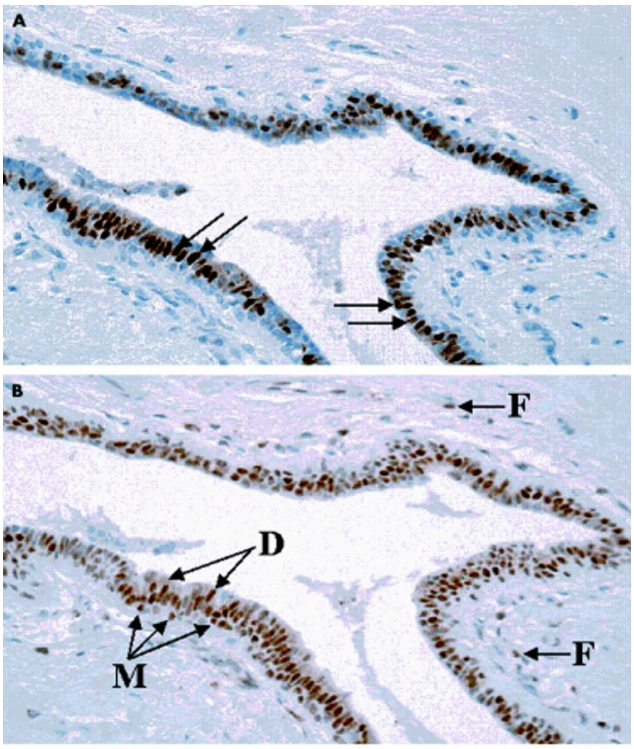


Figure 12.Expression and distribution of ERα and ERβ in mammary gland.

A) Expression of ERα in the normal mammary gland. The staining was restricted to breast ducts(arrows.) B) Expression of ERβ in serial section of A. Legend: D=Ductal epithelial, F=intratubular fibroblast M=myoepithelial cell. [Figure adapted from Speirs et al., J.Clin.Pathol., 2002]

1.3.4 Breast Cancer classification

BCs are classified using different schemes: hormone receptor status, histopathological type, grade, stage.

Receptor status: This scheme evaluates the molecular subtypes of the cancer specimens. Five molecular subtypes exist in BC[40] (Fig.13):

- Luminal A: ERα positive, HER2 negative, PR positive, low-levels of protein Ki-67
- Luminal B: ERα positive, HER2 positive or negative, PR negative, high levels of protein Ki-67
- HER2-enriched: ERα negative, HER2 positive, PR negative
- Basal-like: ERα negative, HER2 negative, PR negative
- Claudin-low: ERα negative, HER2 negative, PR negative and high frequency of cellular differentiation[41]

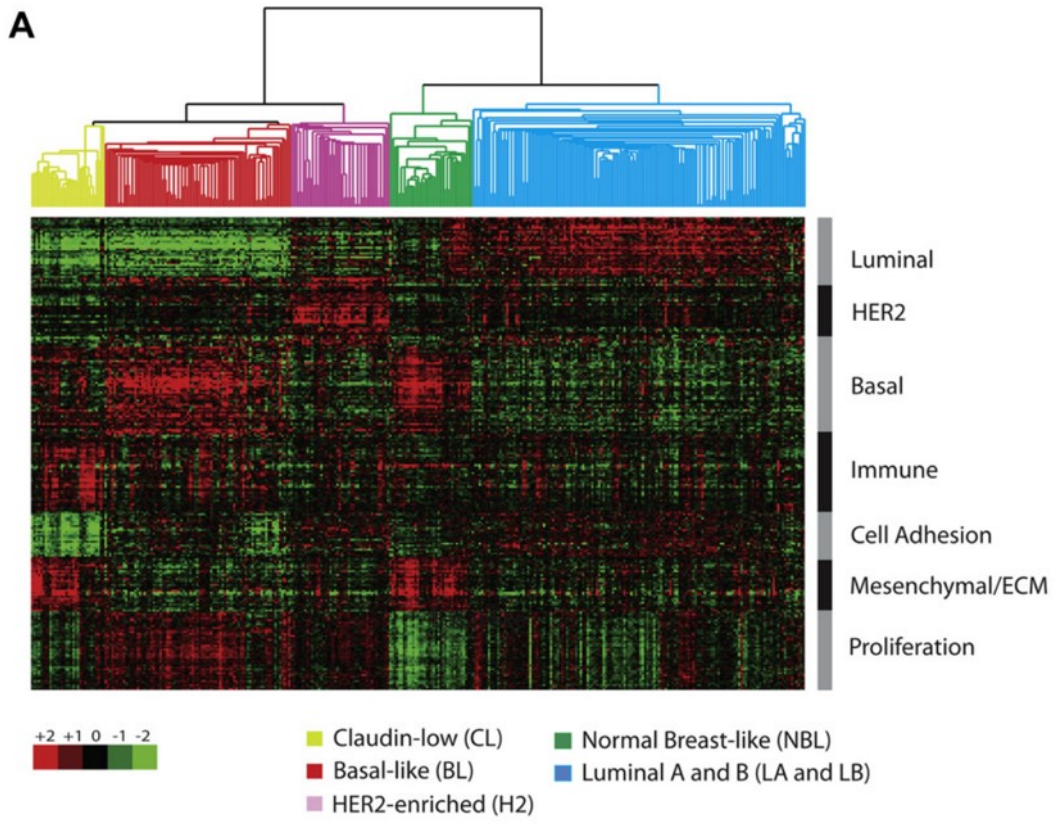


Figure 13. Cluster diagram of Breast Cancer Subtypes.

Genes upregulated are shown in red while genes downregulated are shown in green. [Figure adapted from Prat and Perou., Mol.Oncol., 2011]

Histopathological type: This scheme assesses the histology of the specimens. Approximately 80% of BCs are Invasive Ductal Carcinoma, followed by Invasive Lobular Carcinomas which account for approximately 9 -15% of cases[42] and finally, 5-15% of the cases are Ductal Carcinoma in Situ.

Grade: Grading classification asses the degree of differentiation of the pathological specimen compared to the neighbouring healthy tissue. Tumour grade is allocated to a scale of 3 grades where grade 1 is well differentiated, grade 2 moderately differentiated and grade 3 poorly differentiated. It is well reported a significant correlation between grade and prognosis. Patients with poorly differentiated tumours have a worse free survival rate compared to the good and moderate[43].

Stage: Stage or TNM classification evaluates the anatomical extent of a primary tumour (T), the presence of regional lymph node metastasis (N) and distal metastasis (M). It spans between stage 0 to 4 where 0 is a precancerous condition, and 4 is metastatic BC[44] .

1.3.5 Oestrogen Positive Breast Cancer therapy

Receptor status, grade, stage, the age of patient and size determine the patient treatment options. The primary objective of BC treatment is to improve the survival of the patient and reduce the local and distal recurrence. In the largest majority of BC patients (70%-80%) the growth of the primary tumours is regulated by oestrogens. This section briefly describes the state of art of the currently available therapy options for oestrogen positive BC cases. Detailed description therapies applied to different BC types is, therefore, beyond the scope of this dissertation.

The majority of women with early-stage BC receive breast-conserving surgery[45] followed by regional nodal or whole breast irradiation[46,47]. This combined approach allows the removal of the macroscopic neoplasia (surgery) and the microscopic tumour foci (radiotherapy)[48], significantly reducing the likelihood of recurrence (Fig.14)[46]. In addition, breast cancer surgery is hastened by neoadjuvant chemotherapy administered with the aim to shrink the tumour mass and reduce possible undetected micrometastases[49] .

After the surgical removal of a tumour, the patient receives adjuvant endocrine therapy which is the most successful approach to date. The main aim of systemic adjuvant endocrine treatment is to control any remaining deposits of disease, reduce the recurrence rate, and improve long-term survival by controlling oestradiol production (AI) or inhibiting ERα action (SERM).

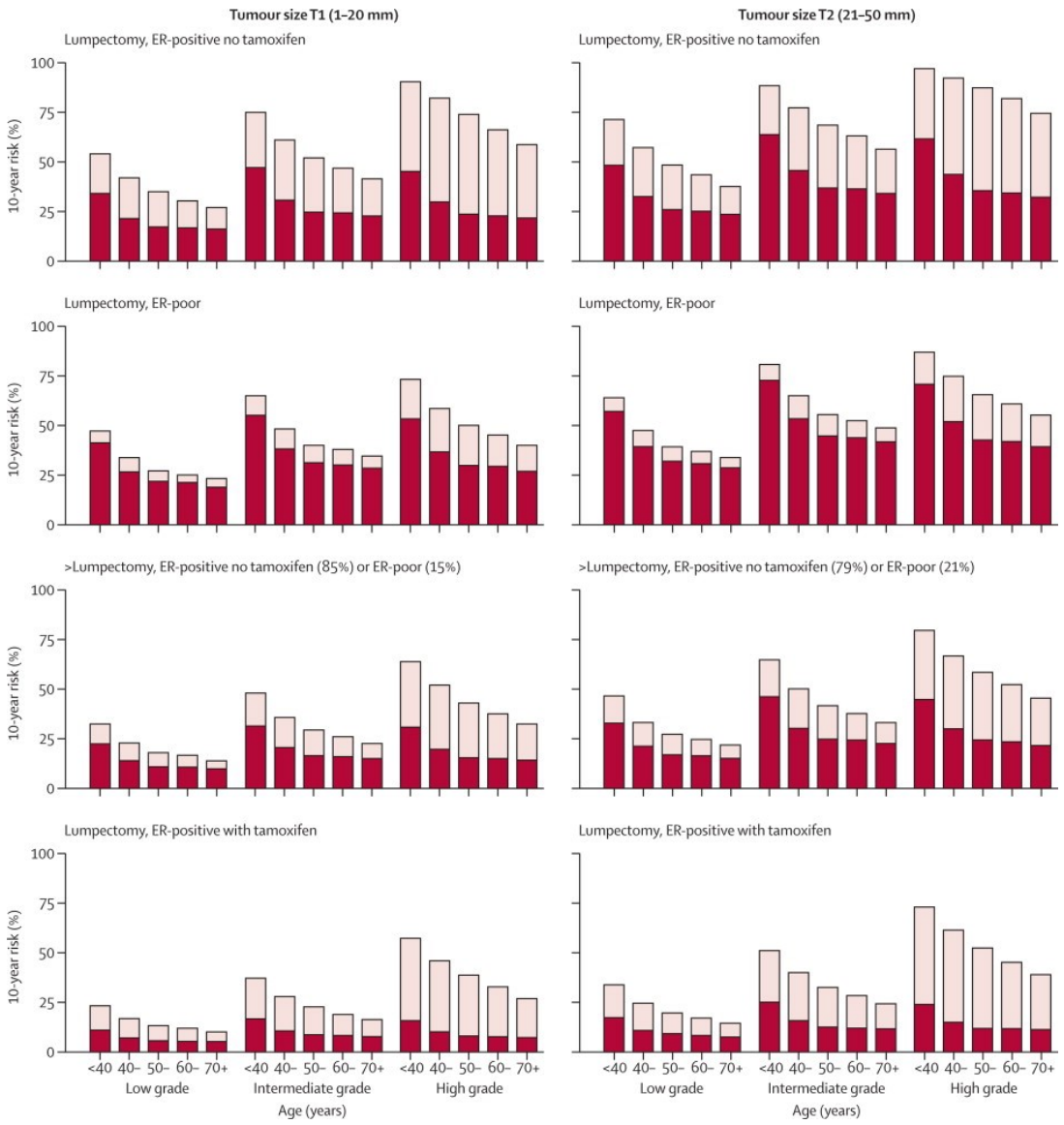


Figure 14. The absolute risk of Breast Cancer Relapse with and without radiotherapy following lumpectomy.

Absolute risk within 10 years on any relapse with (light red) and without (dark red) RT treatment after breast conservation surgery by tumour characteristics and patient type. [Figure adapted from EBCTCG., Lancet 2011].

Since 1970, Tamoxifen is the first-line therapy for ERα positive BC. Tamoxifen is a selective oestrogen modulator (SERM). It is similar in structure to oestradiol and binds ERα inhibiting the proliferation of Breast Cancer ERα cells. Tamoxifen is administered in pre and postmenopausal women for 5 to 10 years and randomised clinical trials demonstrated the efficacy in reducing the risk of tumour relapse by 44% to 69%[48,50] (Fig.15).

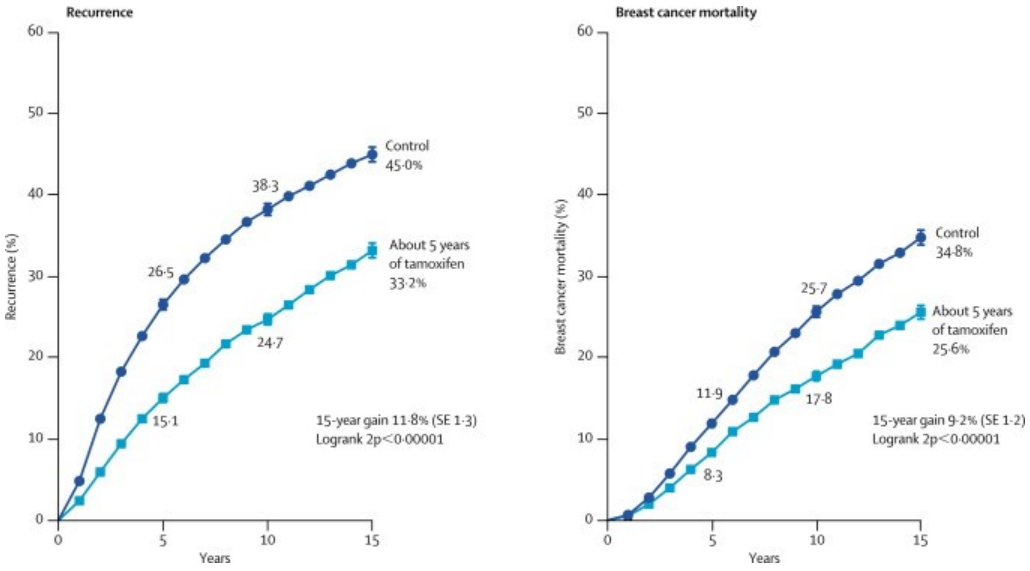


Figure 15. 15 years of Recurrence and Breast cancer mortality rate with and without Tamoxifen administration.

The illustration shows the % of recurrence and mortality in patients (N=10386) treated with (light blue) and without (dark blue) 5 consecutive years of tamoxifen. [Figure adapted from EBCTCG., Lancet 2011].

In the post-menopausal context, alternative endocrine therapy to Tamoxifen is preferred[51]. International guidelines recommend treating those patients with third generation AIs[52,53] such as exemestane, anastrozole and letrozole. AIs suppress the enzyme aromatase conversion of androgens to oestrogens. Adjuvant trials comparing tamoxifen and the new third-generation aromatase inhibitors (AI) indicate higher efficacy over tamoxifen by reducing contralateral breast cancer of approximately 70%. Also, the results from new trials[54] (Fig. 16) show a significant benefit of disease-free survival among women who were administered with AI beyond five years.

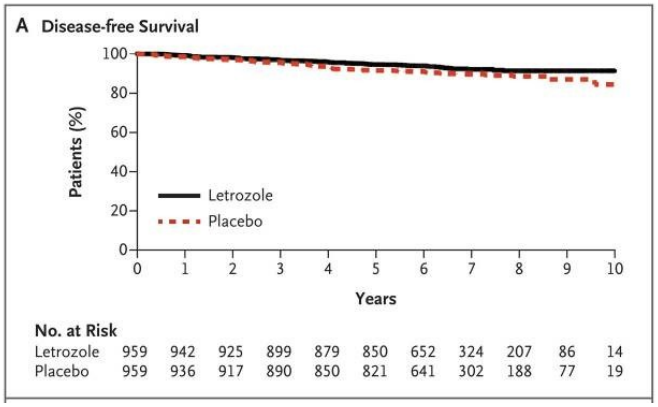


Figure 16. The effects of 10 years' consecutive administration of AI.

Disease free-survival rate of AI treated(N=67) patients is significantly low (95% CI, 0.48 to 0.91; P=0.01) compared to placebo (N=98). [Figure adapted from Goss., N. Engl. J. Med 2016].

Despite these advances, 30% of patient relapse, during or after completing adjuvant endocrine therapy administration, with a new aggressive tumour. These tumours commonly exhibit resistance to treatment and metastatic activity[55].

Furthermore, metastatic BC present sub clonal diversification of receptor status compared to the primary mass. Single-agent therapy in those patients provides a limited clinical benefit, and a combination of different drugs is the most common therapeutic choice. Tamoxifen and aromatase inhibitors (AIs) in conjunction with or without the cyclin-dependent kinase inhibitor such as palbociclib are proposed to treat metastatic BC followed a second-line therapy based on the administration of selective oestrogen receptor degrader (SERD)[56,57] .

1.3.6 Genetics of Breast Cancer

In the last few years, improvement of DNA sequencing technology enhanced the systematic characterisation of the genetic events during tumorigenesis and progression. Historically, genomics analysis of BC put the effort in the investigation of somatic mutation in coding regions[58]. However, little is known about alterations in the non-coding fraction although it represents 98% of the genome. In addition, the majority of the studies have highlighted the role of the recurrent mutations or

"drivers", which confer a clonal and selective advantage to cancer cells. By contrast, much less is known about the myriad of passenger processes occurring during oncogenesis and metastasis of BC.

A vast plethora of genetic events significantly contributes to the phenotype of cancer cells. The main event driving oncogenesis in 12% of BC cases is the mutation of the carriers BRCA1 and BRCA2.

Retrospective studies estimated the cumulative risk of BC for BRCA1 and BRCA2 carriers to be around 75%[59] (Fig.17). However, BRCA1 and BRCA2 mutation are sporadic in the population and an absolute minority of BC cases. Indeed, mutational screening of BRCA1 and BRCA2 are performed on the basis of family history and early age of diagnosis.

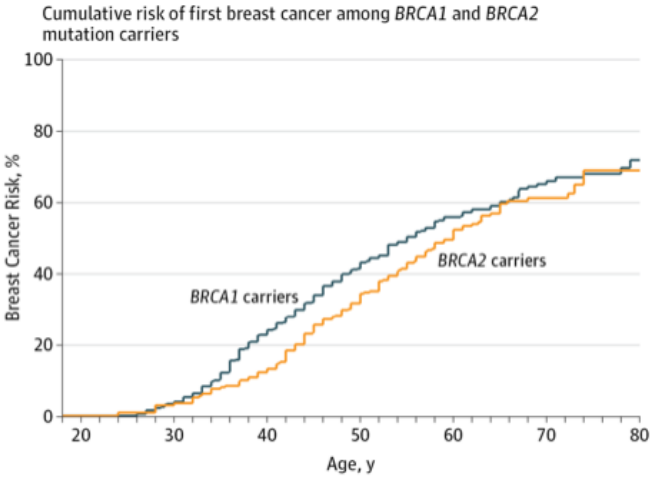


Figure 17. Risk of Breast Cancer among BRCA1 and BRCA2 mutations.

The cumulative risk of Breast Cancer diagnosis in function of age in the presence of BRCA1 mutation (blue) and BRCA2 (yellow). [Figure adapted from Kuchenbaecker et al., JAMA 2017]

Recent large-scale studies identified the events which have a profound impact on BC. The totality of BCs has at least one of the following genes mutated: TP53, PIK3CA, MYC, CCND1, PTEN, ERBB2, ZNF703/FGFR1locus, GATA3, RB1, PIK3CA[58] and MAP3K1 (Fig.18). Interestingly, other genes including MED23, FOXP1, MLLT4, XBP1, ZFP36L1 are now being recognised as cancer driver genes although previous reports did not implicate them in BC, and novel observations report recurrent rearranged regions involving ESR1 and ETV6.

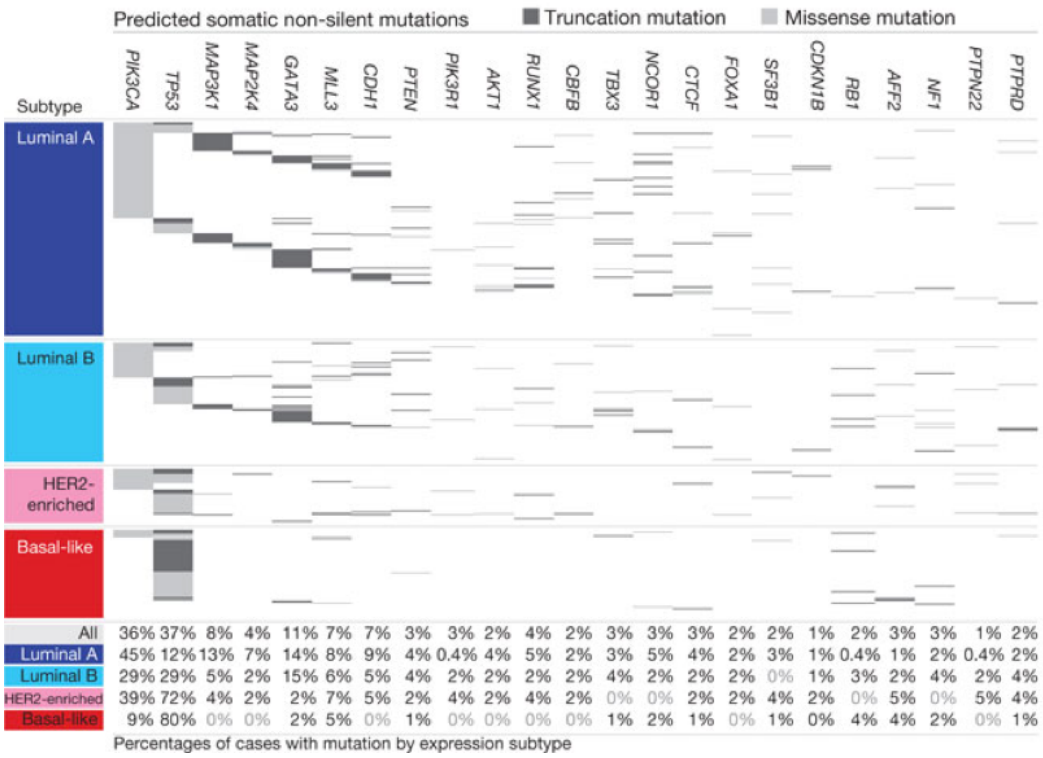


Figure 18. Genetic portrait of tumour samples by mRNA subtype.

This illustration shows pattern and frequencies of mutations for selected genes in Breast Cancer subtype. [Figure adapted from TCGA, Nature 2012]

BCs show numerous loci affected by hypermutation state defined as kataegis on chr 8, 17 and 22 and commonly, CGH investigation reveals frequent chromosome imbalance (Fig.19). Indeed, the vast majority of BC sample have substantial gains of 1q, 8q, losses of 16q[58],17p[60], focal amplification of PIK3CA, EGFR, FOXA1, HER2 and deletion of MLL3, PTEN, RB1 and MAP2K4 coding regions[61]. However, much of the genomic instability arises at the sub clonal level during the progression of the disease[60].

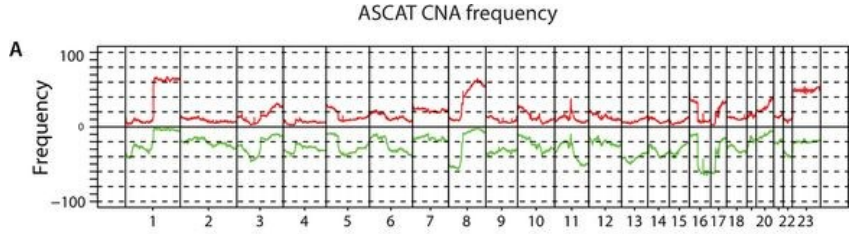


Figure 19 Copy number analysis of 560 Breast Cancer samples.

The frequency of copy number aberration in 560 Breast Cancer samples. X-axis represents chromosome position, Y-axis show gains (red) and losses (green). [Figure adapted from Nik-Zainal, Nature 2016]

It is believed that the distribution of somatic mutations is highly unbalanced in the genome[62,63]. However, our unpublished data reveal that coding regions harbour a similar mutational burden of non-coding regulatory regions. Although ENCODE[64] project estimates 80% of the genome is functional, the effect of mutations in those regions is poorly understood. Human population studies are reporting disease-associated genomic variation located in regulatory elements[65] such as enhancers, TAD boundaries and promoters. Indeed, genomic analysis of BC samples revealed recurrent mutational hot-spots harboured in the promoter of FOXA1, RMRP and NEAT1[66], oncogene activation associated to genome topological alterations caused by mutations[9] and transcriptional dysregulation generated by variation in enhancer regions. Furthermore, mutations of regulatory elements in tumours are associated with oncogenic cell phenotype[67].

Cancer is a disease which is accompanied and reinforced by an elusive evolution. The genetic events described above are the result of this evolutive aspects and contribute to the development of the cell phenotype toward different stages of the disease. Recently an investigation of the evolutionary trajectories led to intriguing results in the understanding of the temporal succession of events during the disease[68]. It is proposed that earlier events drive the oncogenesis, in particular TP53 mutations. TP53 is mutated in 20% of ERα+ BC. This mutation is considered the oncogenic driver of cancer and is typically followed by divergent mutational trajectories involving PIK3CA mutation,1q loss, 17p loss or 16q loss (Fig.20). New aberrations arise during the progression of the disease, and strikingly, studying the mutational trajectories of BC may give an advantage in the association of the patient to a distinct prognosis. Thus, the understanding of the genetic basis of tumour evolution may allow the clinical stratification of the patient to distinct personalised clinical decisions.

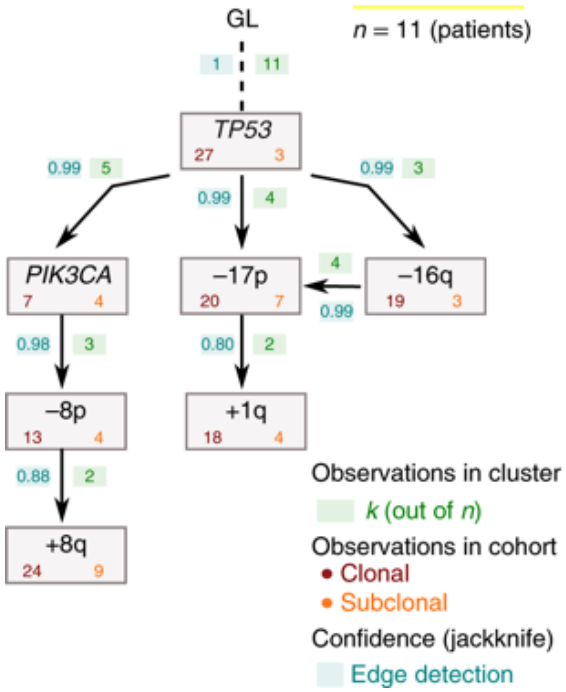


Figure 20. Example of evolutionary trajectories in Breast Cancer.

This figure shows the repeated mutations in a subset of BC samples (n=11). Each transition (arrows) is reported with the observed frequency of mutation (blue value). [Figure adapted from Caravagna et al., Nat. Methods 2018]

1.4 Epigenetics, enhancer and chromatin packaging

Aberrant gene function and altered patterns of gene expression are hallmarks of cancer[9] . Novel shreds of evidence reveal the contribution of epigenetic abnormalities in playing a fundamental role in this dysregulation[9,69,70]. Epigenetic alterations including aberrant methylation and chromatin remodelling modulate the events occurring from the earliest up to the final stages of the disease.

The word epigenetics refers to the whole set of information underlying gene expression not attributable to the DNA sequence. Epigenetics includes DNA methylation of CpG islands, post-translational modifications of nucleosomes and the high-order packaging of the chromatin.

DNA is wrapped around octamers of histone proteins (2 copies of H2A, H2B, H3 and H4) (Fig.21). This structure is the fundamental unit of DNA packaging and is termed nucleosome. Histones present an N-terminal tail which protrudes out through the nucleosome, and it is subjected to a diverse array of post-translational modifications. The combination of post-translational modifications directs nucleosome positioning (Fig.21), which is central to gene regulation and ultimately essential in cell development, identity and disease. Indeed, an extensive systematic rewiring of histone post-translation modification is associated with activation and inactivation of regulatory regions, and strong evidence suggests a fundamental role in functional dysregulation in cancer[9,69,70] (Fig.22).

A large number of epigenetic modifications occur on histone tails. As an example, acetylation, methylation events on histone tails and methylation of CpGs islands have a key role in the fine-tuning of DNA accessibility modulation. Trimethylation on the lysine 4 and 27 of H3 (H3K27me3 and H3K4me3) are markers of inactive promoters as well as methylation of the CpG islands. Mono, di and three methylations of lysine 4 of H3 (H3K4me, H3K4me2, H3K4me3) or cytosine hypomethylation are markers of active promoters[71]. For the scope of the dissertation, one particular covalent post-translation modification has a fundamental impact: the acetylation at the 27th lysine residue of the histone H3 protein (H3K27ac).

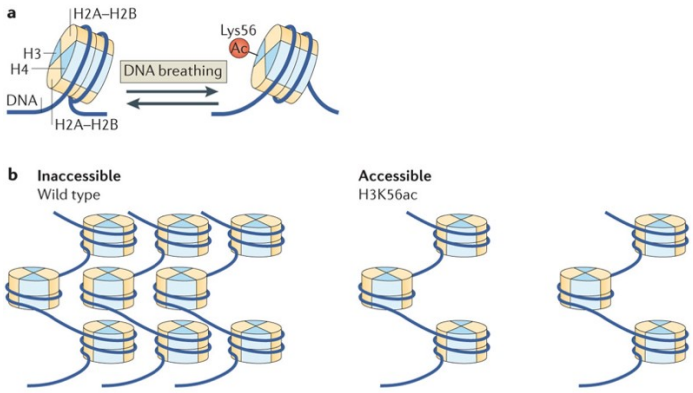


Figure 21. Example of the effects of histone modifications to chromatin accessibility.

A) H3k56ac is a well-known marker of DNA accessibility; B) H3K56ac regulates tertiary contacts to keep chromatin accessible to TFs. [Figure adapted from Tessarz & Kouzarides, Nat. Rev.Mol.Cell.Bio 2014)

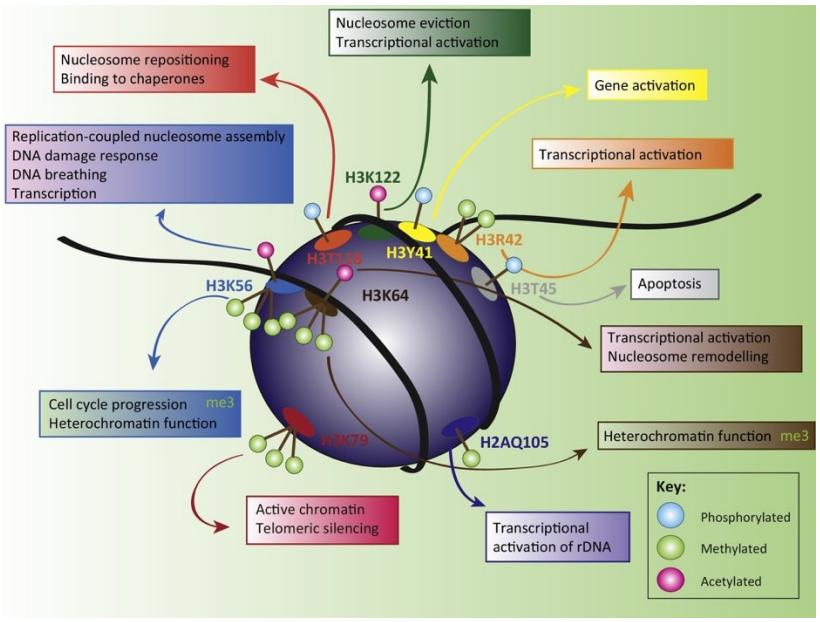


Figure 22. Association of histone post-translational modifications to cell regulatory activity.

Key modification of a nucleosome are linked to specific regulatory activity. Methyl marks are shown in light green, acetyl marks in pink, and phosphorylated residues in light blue. [Figure adapted from Lawrence et al., Trends Genet.2016]

Histone acetylation is regulated by two enzymatic families, histone acetylases (HAT) and deacetylases (HDAK). The acetylation of histones is associated with transcriptional activity[72,73].

In particular, H3K27ac strongly correlates with chromatin openness at the level of cis-regulatory elements such as enhancers and promoters. The mapping of H3K27ac status at genome-wide level is a standard approach in the evaluation of enhancers and promoter's activity. Recently, it has been shown that chromatin resetting is necessary to allow the recruitment of TFs. For example, Serandour et al. 2018 revealed TRSP1 to be an essential mediator of histone-deacetylation and thus preventing additional binding site for ER in mammalian tissues[74].

Modification of chromatin openness marked by H3K27ac[75] is linked to the activation and inhibition of several regulatory pathways able to confer a selective advantage to ERα BC subclones suggesting that phenotypic cancer clones are the result of genetic and epigenetic selection[76,77] (Fig.23). Among cis-regulatory regions, enhancers are of particular interest.

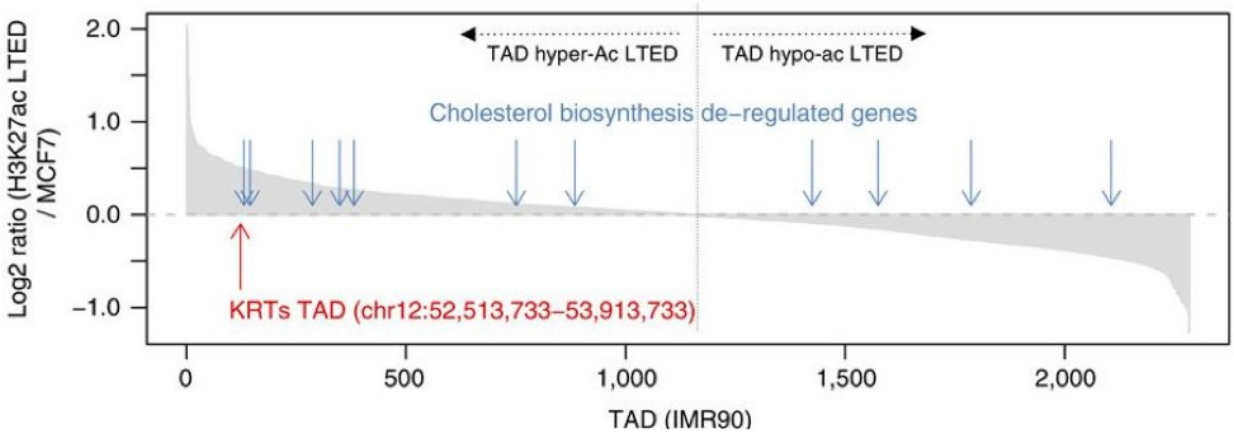


Figure 23. Differential epigenomic reprogramming in TAD context.

2283 Topologically Associated Domains (TADs) identified in IMR90 cell lines were sorted according to the ratio H3K27ac in LTED versus MCF7 cell lines. Left panel highlights differential chromatin openness in LTED while right panel show differential chromatin packaging. Red arrows show KRT genes activation while blue arrows deregulated genes involved in cholesterol biosynthesis. [Figure adapted from Van Nguyen & Barozzi et al., Nat.Comm., 2015)

Enhancers are key distal cis-regulatory elements that elevate the expression of nearby genes, independently by the distance to the target gene or orientation[78]. Enhancers are activated by the cis-binding of TFs and their associated co-factors to specific sequence motif. A specific class of TFs called pioneer factors have a prominent role in enhancer activation. As an example, the forkhead factor (FOXA1) is essential in enhancer response and facilitate the associations between ERa and ERE elements (Fig.24)[33].

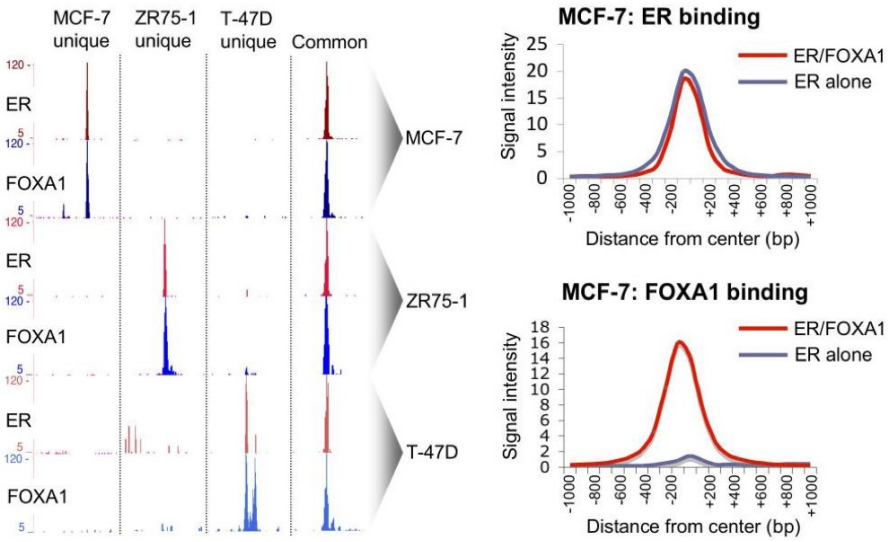


Figure 24. FOXA1 binding and ERα binding in Breast Cancer cell lines.

ChIP-seq profiles of ER, FOXA1 in different regions of breast cancer (left panel). Average signal intensity of ER binding sites that are shared with FOXA1 binding regions or ER binding events that are not shared with FOXA1. [Figure adapted from Hurtado et al., Nat.Gen. 2011]

Indeed, pioneer factors can 1) recognise DNA binding motif directly on compacted chromatin and then 2) trigger the activation of regulatory elements by recruiting other TFs and co-factors[79]. Pioneers factors ultimately regulate and predetermine TFs binding capabilities. The resulting complexes are enhancer specific, and the composition of the complex itself has a functional impact. In addition, growing data report the subsequent trans-binding recruitment of cofactors, resulting in mega trans complex[80] which mediate broad transcriptional programmes.

Sequencing-based DNA-DNA interaction studies showed enhancer-promoter interactions between 1kb to several hundred kilobases distant to the TSS[81] (Fig.25). Enhancer-promoter loopings are highly conserved across cell types and activated by lineage-specific transcription factors complexes[81,82] (Fig.26) .

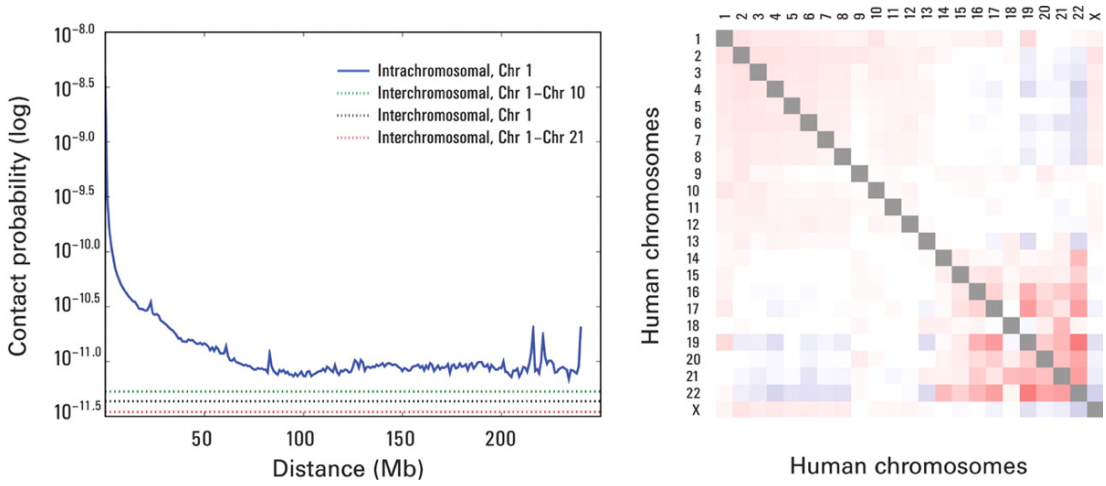


Figure 25. Contact probability of DNA intra and intrachromosomic regions.

The panel left: the probability of contacts within chromosome 1 decrease as a function of genomic distance. The probability of intrachromosomic interaction (dashed lines) is significantly low. Panel right: Heatmap showing the observed/expected intrachromosomic contacts between pairs of chromosomes. Red colour shows enrichment while blue shows depletion. [Figure adapted from Lieberman-Aiden et al., Science 2009].

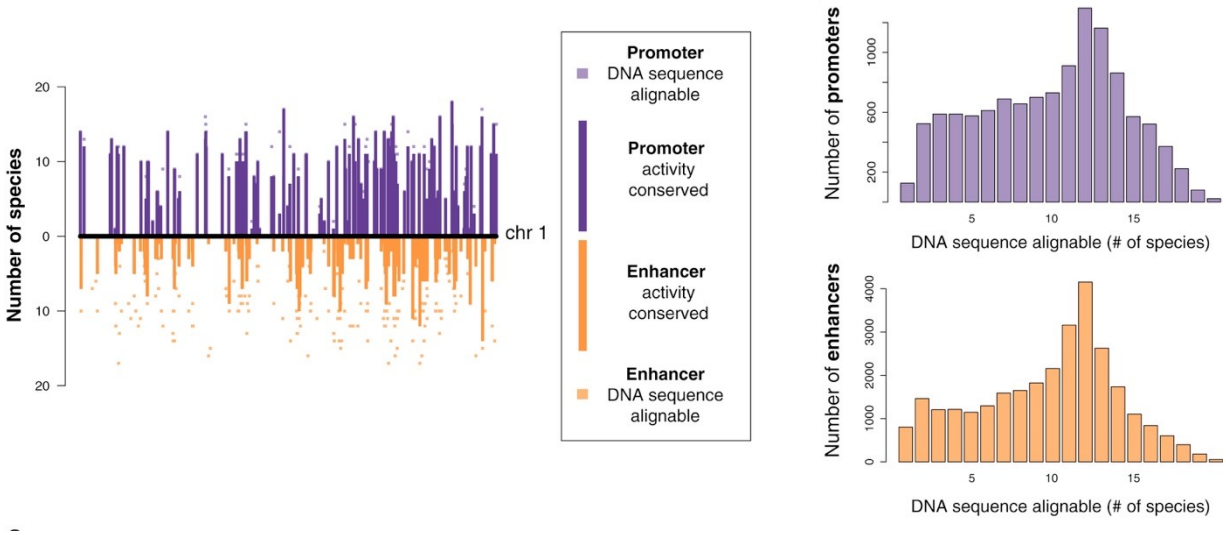


Figure 26. Conservation of DNA sequences of promoters and enhancers across 20 species.

Left panel: Conservation of regulatory sites between 20 species in chromosome1.Right panel: Alignment of DNA sequences proximal to promoters (purple) and enhancers (orange) to similar species. [Figure adapted from Villar et al., Cell 2015]

Lupianez et al. demonstrated that enhancer-promoter physical interaction is allowed by a predetermined cell type-specific genomic configuration[84]. Indeed, mammalian genomes are organised into submegabase compartments called associated topological domains (TADs) which are high order genomic structure having the role in restricting or facilitating DNA-DNA interaction within the same domain[85]. TADs are enriched by factors which establish and stabilise the architectural order of the genome. A genome-wide conserved signature of factors is enriched at TAD borders and delimit each TAD. This signature includes CTCF protein, ZNF143, cohesion complex, housekeeping genes which guarantee a physical barrier to inter-TAD interactions. These factors have a distinct role in TAD management (Fig.27). CTCF depletion leads to an abnormal inter TAD looping rewiring whereas deletion of cohesin complex seems to affect just intra TAD interactions. Disruption of TADs leads to an extensive rewiring of DNA long-range interaction which results in an inappropriate enhancer-promoter looping and thus new pathogenic phenotype[84].

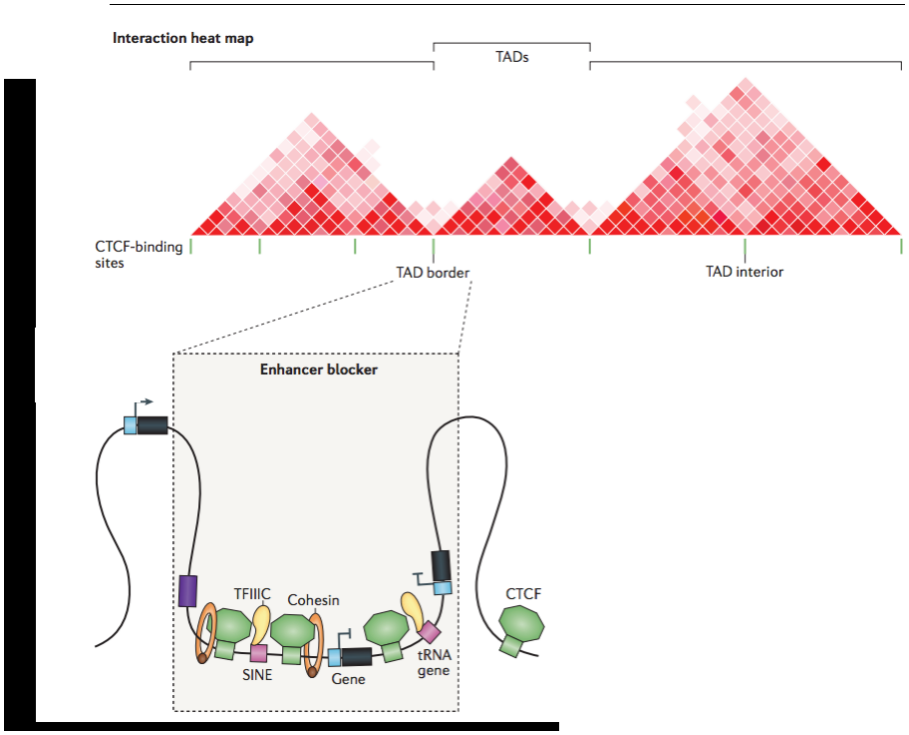


Figure 27. TAD borders are enriched of CTC and cohesion.

The figure illustrates the interaction heatmap of a chromosome segment depicted in Hi-C data. TAD borders are enriched of CTCF, cohesin and other factors. [Figure adapted from Ong & Corces, Nat.Rev. Genet. 2014]

1.5 Mechanisms of enhancers rewiring in cancer

Most cancers exhibit distinct enhancer activity profiles[86]. The investigation of causal enhancer associations with clinically actionable genes is a hot topic of cancer epigenomics. Indeed, growing data support the prominent role of the dynamics of these regulatory elements in development and disease. Furthermore, chromatin architecture rearrangement strongly affects the functioning of enhancers leading to aberrant cell phenotypes[87].

Enhancers precisely regulate gene expression networks and are essential in the proper human development. As an example, the MYC gene expression is finely regulated by enhancer modules exhibiting a selected activity. MYC is a proto-oncogene involved in essential biological functions which are deregulated in almost the totality of cancers. Typically, cancer cells exhibit MYC overexpression. This condition is not explained by coding mutations and gene amplification, which occur in about 20% of cancer patients, but by modulation and additive dynamics of neighbour MYC enhancers clusters[88].

Chen et al. in 2018 identified an enhancer located 140kb to the PD-L1 gene, a major target of immunotherapy in lung cancer and melanoma[86]. This enhancer is an upstream regulator of the PD-L1 gene, and it is targeted by NF-kB TF, which participate in the enhancer-promoter interaction and thus a potential target in cancer therapy.

Furthermore, Genome-wide association studies (GWAS) showed that enhancers are enriched of disease-associated variants and single nucleotide polymorphisms (SNPs)[90]. Mainly, BC risk associated SNPs are enriched for FOXA1 and ESR1 motif (Fig. 28, Fig. 29). It is postulated that those SNPs may disrupt breast cancer identity by modulating the differential binding of ESR1 and FOXA1[91].

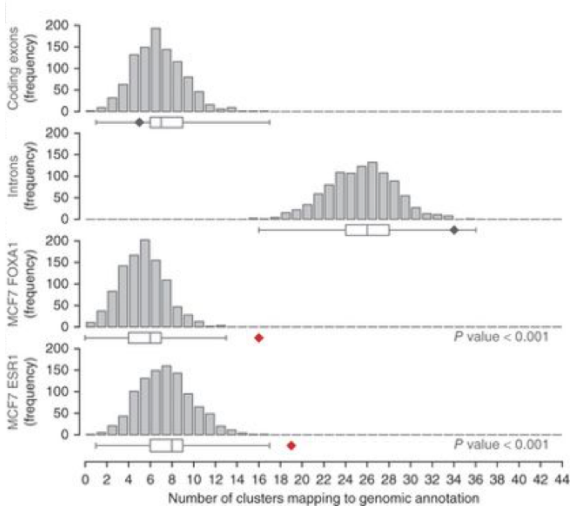


Figure 28. Risk-associated SNPs in Breast Cancer.

Variant set enrichment analysis shows that associated variant set is strongly enriched in FOXA1 and ESR1 sites. [Figure adapted from Cowper-Sallari et al., Nat. Gen 2012]

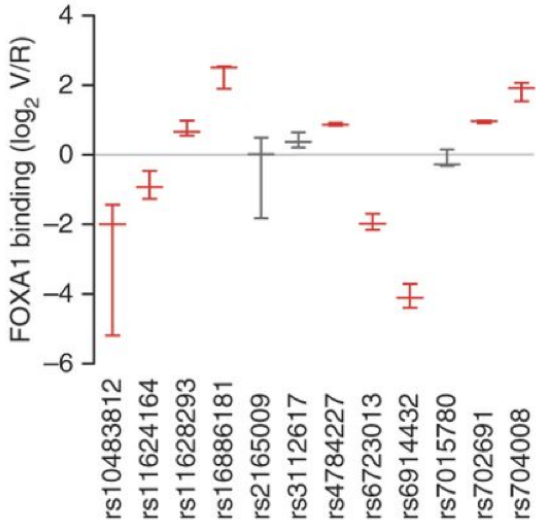


Figure 29. FOXA1 binding affinity at heterozygous SNPs.

Allele-specific ChIP-qPCR of 12 SNPs in Breast Cancer cell lines. Y-axis shows the log2 fold change of the FOXA1 binding to the variant and reference allele of each SNPs. Red shows differential binding enrichment to the alleles. [Figure adapted from Cowper-Sallari et al., Nat. Gen 2012]

Numerous studies focused on the interplay between copy number alteration, and cis-regulatory elements revealing a strict association between regulatory elements and genetic events[92,93]. A phenomenon recently discovered in genomic cancer studies is the "enhancer hijacking"[94,95]. Northcott et al., extensively described in brain tumours, the oncogenic activation of the GF11 gene occurring

after its physical translocation to high-density enhancer regions. The juxtaposition of GF11 proximity to active enhancers results in gene deregulation. Another mechanism of enhancers hijacking entails translocation of enhancers to nearby genes located on different chromosomes[94]. Drier et al. in 2016 described this mechanism in adenoid cystic carcinoma where commonly, enhancers translocate nearby MYB gene, ultimately driving its overexpression[95].

1.6 Tumour Heterogeneity

The extensive tumour heterogeneity is a major limitation in cancer studies. Modern single cell studies show that no two cells in a carcinoma specimen exhibit an identical genomic landscape and large histopathological differences within the same tumour specimens are commonly appreciated in surgery pathological practice (Fig.30). Furthermore, the clinical outcome of patients exhibiting tumours originating from the same organ and classified in the same type, grade and stage evolve differently. Unfortunately, our current understanding of tumour heterogeneity poses a challenge in the clinical management of the patient and cancer research. However rapid advances in technologies for genetic and non-genetic tumour profiling are rapidly providing new insights and proposing more accurate therapy intervention[96].

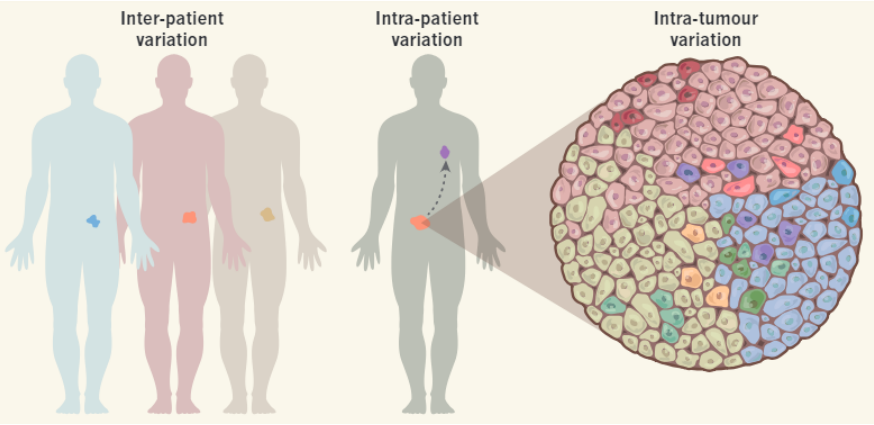


Figure 30. Tumour heterogeneity.

Inter-patient and intrapatient tumour heterogeneity are appreciated at genetic and phenotypical level. Different colours represent different cell subpopulation with the tissue. [Figure adapted from Fox & Loeb, Nature 2014]

Tumour bulks exhibit a vast array of phenotypical and genetic heterogeneity which become more prominent in the latest stages of the disease. New evidence from many cancer studies reports oncogenesis and metastasis as the result of a combination of neutral and Darwinian like evolution[97]. Novel theories postulate that cancer development is the result of a critical genomic event followed by neutral evolutionary dynamics[98,99]. Neutral evolution is an evolutionary process sustained by myriads of neutral genetic events occurring in cancer cells which support the development of an extraordinary number of cell subpopulations. Growing data highlights also detectable acquisition and depletion of genetic and epigenetic factors[100] at different stages of the disease. This may be explained by cancer treatment pressure which may favourite non – neutral evolution by applying an evolutionary selection on the cancer cell phenotypes.

Cancer evolution is elusive and does not follow a fixed course due to a vast number of neutral and not neutral stochastic genetic events[100] coupled with endogenous and exogenous environmental factors.

Yates et al. 2015, in their work, revealed solid tumour samples enriched of variable genetic subclonal diversity using multiregional sequencing. They showed that a minority of these subclones evolve during the disease surviving the selective pressure of cancer treatments and contributing to the marked aggressiveness of this condition. These evolutionary processes generally lead to drug resistance mechanisms which finally conduct the patient to the final stage of the disease. Practically, cells undergo a phenotypical switch which contributes to the lack of response to the therapy. Interestingly, genetic heterogeneity does not necessarily correlate with increased numbers of cell phenotypes, and further efforts need to be dedicated to understanding the molecular factors driving the development of cell phenotype which ultimately is the therapy target.

Chapter 2

Epigenomic profiling of ERα+ Breast cancer reveals novel principles of heterogeneity and phenotypical evolution

Aims of the study

Breast Cancer is the most prevalent type of cancer among women in Europe. In 70% of patients, BC manifests ERα positivity and is treated with endocrine therapy after surgical removal of the tumour. The vast majority of patient positively respond to the therapy. However, data show that 30% of patients relapse with a new, aggressive, drug insensitive BC within ten years from the surgical operation. The mechanisms sustaining drug-resistance in cancer are not yet identified mainly cause an extensive phenotypical intra and interpatient heterogeneity which limit the understanding of the genetic and epigenetic factors contributing to the disease progression. Indeed, the interplay between genetics and epigenetics and their link with cancer progression is still elusive. Extensive studies have been funded to uncover the genetic profile of ERα positive BC patient. The analysis of coding DNA in patients have identified the most common mutations, and genomic aberrations, however little is known about the events occurring in the non-coding genome. This may be explained by the difficulties in mining DNA regions of apparently no functional activity. To our knowledge, the epigenomic asset of ERα positive BC has never been described although it may represent an important source of knowledge for shrinking the field of investigation to the underlining non-coding DNA. Indeed, identifying active functional regulatory regions through epigenomic approaches would directly contribute to discovering essential critical players of phenotypic control still hidden in the vast non-coding sea.

In this interdisciplinary study, we propose to depict the landscape of active promoter and enhancers regions in ERα-positive BC followed by the identification principles of cancer heterogeneity. The development of computational approaches applied to epigenomic data aims to perform a systematic

and comprehensive deconvolution of the phenotypic heterogeneity and then successfully estimate key controllers of cancer progression.

This study for the first time, to our knowledge, describes the dynamics of regulatory regions activity in ERα-positive BC patients in primary and metastatic tissues at an unprecedented level of resolution and ultimately identify two novel main players in ERα+ BC oncogenesis, YY1 and SLC9A3R1.

2.1 H3K27ac ChIP-seq of 39 primaries and 16 metastatic ERα positive samples

The study was conducted by analysing a total of 55 fresh frozen patient samples (Table 4-5). 39 samples were primary ERα positive Breast Cancer, and 16 specimens were metastatic samples derived from patients in which a primary tumour was positive to ERα. Primary samples were obtained from Breast Cancer Now tissue bank (20 specimens) while other 19 from VERB (still ongoing study) and Imperial College Healthcare NHS Trust tissue Bank study. Metastatic samples collection was placed at Charing Cross Breast Unit and Oncology departments (Imperial College Healthcare NHS Trust, London, U.K.).

We processed all the samples using H3K27ac ChIP-seq to build a comprehensive map of regulatory regions activity. The average number of significant peaks called in primary and metastatic samples was respectively 64,042 and 63,691 (Table 6-7 and Fig.31-32). The numbers of peaks obtained in each sample among the two categories were not homogeneous. Peak numbers were spanning between 0 to 170k significant peaks after peak calling post-processing. Prior to the sequencing, all the tumours showed sufficient enrichment in qPCR. However, 8 in 55 samples generated a number of peaks lower than 2k and subsequently eliminated from the analysis.

Tissue heterogeneity, technical aspects in tissue collection and the multi-step ChIP-seq processing contribute to the variability of the final results. Tumours are per se highly heterogeneous. Primary

and metastatic samples exhibit profound histological, pathological, genomic differences and various level of contamination from other tissues which are ultimately highlighted in peaks calling. The histological characteristics of the specimen may represent a significant factor of ChIP-seq processing failure. All samples were macrodissected from the lipidic tissue; however, it is not possible to exclude partial contamination from fibroblast and immune cells which may decrease the efficiency of antibody in pulling down the enriched regions. Furthermore, tissue collection from surgical procedures is prone to many factors which may delay the standard protocol procedures resulting in DNA deterioration which finally produce a variable number of peaks.

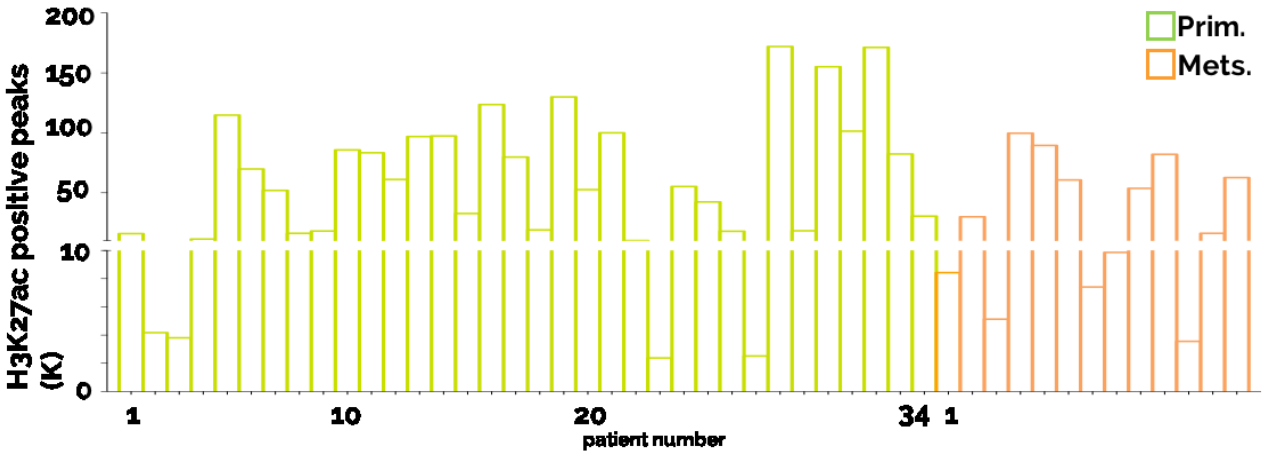


Figure 31. Significant H3K27ac peaks by sample.

Green are primary samples, orange are metastatic samples. Peaks generated with MACS2 with qval 0.001. [Analysis performed by Giacomo Corleone. Figure produced in collaboration with Darren Patten].

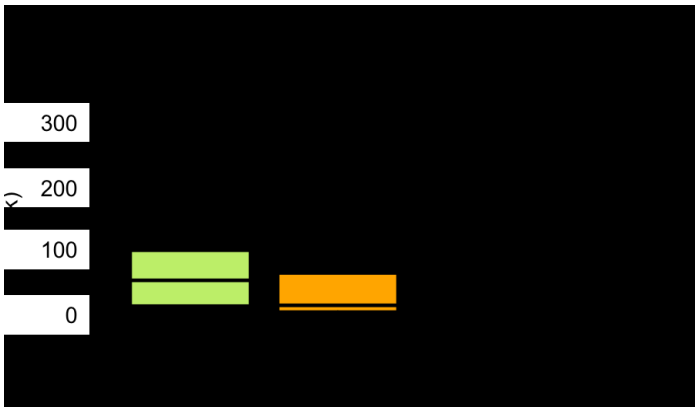


Figure 32. Number of peaks in primary and metastatic sample group.

Sample identifier	Column1	Size of tumour (mm)	Tumour grade	Histology details	Lymph nodes positive	Receptor status	Adjuvant(A) and/or Neoadjuvant (N) therapy	Recurrence type	Recurrence: Date	Date of diagnosis	Alive/Deceased (+Date)	Time to relapse (months)	Further Analysis
G_DKP_006	0702019A	38	3	IDC	1 of 23	ER +, PR -, HER2 -	A		None	10/01/2007	Alive (12/05/2016)		NA_P1
P_DKP_010	1244	17	2	IDC	0 of 8	ER +, PR +, HER2 -	A		None	15/06/2011	Alive (12/05/2016)		NA_P2
G_DKP_003	0402133C_Small	15	2	IDC	5 of 22	ER +, PR +, HER2 -	A		None	24/11/2004	Alive (12/05/2016)		NA_P3
P_DKP_013	2011_0028	15	2	IDC	0 of 1	ER +, PR -, HER2 -	A		None	07/07/2011	Alive (12/05/2016)		NA_P4
G_DKP_013	DKP_BCC_03_GO	4	2	IDC	0 of 1	ER +	No therapy	CR	38427	14/05/1993	Alive (12/05/2016)	142	NA_P5
G_DKP_001	302047	35	2	IDC	15 of 34	ER +, PR +, HER2 -	A		None	19/03/2003	Alive		P1
G_DKP_002	0402133C_Big	15	2	IDC	5 of 22	ER +, PR +, HER2 -	A		None	24/11/2004	Alive		P2
G_DKP_004	0502072A	30	2	IDC	6 of 15	ER +, PR +, HER2 -	A		None	24/05/2005	Alive (12/05/2016)		P3
G_DKP_012	DKP_BCC_91_GO	19	2	IDC	1 of 6	ER +, PR +, HER2 -	A		None	22/05/1991	Alive (12/05/2016)		P10
G_DKP_014	DKP_BCC_94_GO	20	2	IDC	3 of 7	ER +, PR +, HER2 -	Unknown		None	15/03/1994	Alive (12/05/2016)		P11
G_DKP_015	DKP_BCC_1991_034	25	2	IDC	4 of 6	ER +, PR +, HER2 -	A	LR	38869	22/05/1991	Alive (12/05/2016)	181	P12
G_DKP_016	DKP_BCC_1991_049	12	2	IDC	1 of 7	ER +, PR +, HER +	No therapy		None	13/09/1991	L		P13
G_DKP_017	DKP_BCC_1991_053	8	2	IDC	0 of 2	ER +	A+N	CR	38413	30/09/1991	Alive (12/05/2016)	162	P14
G_DKP_018	DKP_BCC_1992_004	20	2	IDC	0 of 8	ER +, HER2 -	No therapy		None	17/01/1992	L		P15
G_DKP_019	DKP_BCC_1994_009	15	2	IDC	2 of 5	ER +, PR +, HER2 -	A	LR	40113	12/01/1994	Alive (12/05/2016)	189	P16
G_DKP_020	DKP_BCC_1994_043	20	2	IDC	2 of 6	ER +, PR +, HER2 -	A		None	15/08/1994	Alive (12/05/2016)		P17
G_DKP_021	0502132A	16	2	IDC	0 of 18	ER +, PR +, HER2 -	A		None	15/10/2005	Alive (12/05/2016)		P18
P_DKP_001	982028A	43 * 8	1	SAR+IDC	0	ER +, PR +, HER2 -	A		None	11/01/2009	Deceased (04/12/2010)		P19
G_DKP_005	0502171A	50 * 10	2	IDC	1 of 8	ER +, PR -, HER2 +	A		None	27/09/2005	Alive (12/05/2016)		P4
G_DKP_007	CP1128	15	2	IDC	5 of 22	ER +, PR +, HER2 -	A		None	24/11/2007	Alive (12/05/2016)		P5
G_DKP_008	ONC_CP_2001	50 * 10	2	IDC	1 of 8	ER +, PR -, HER2 +	A		None	24/11/2004	Alive (12/05/2016)		P6
G_DKP_009	ONC_CP_002002006	71	2	ILC	0 of 10	ER +, PR -, HER2 -	N		None	27/09/2005	Alive (12/05/2016)		P7
G_DKP_010	ONC_CP_002002008	16	2	IDC	0 of 18	ER +, PR +, HER2 -	A		None	04/12/2007	Alive (12/05/2016)		P8
G_DKP_011	ONC_CP_002002010	20	2	ILC	0 of 6	ER +, PR +, HER2 -	A		None	13/08/2008	Alive (12/05/2016)		P9
P_DKP_002	502154	30	3	IDC	1 of 12	ER +, PR -, HER2 -	A		None	16/11/2005	Deceased (23/03/2008)		P20
P_DKP_003	0802012A	64	3	IDC	7 of 20	ER +, PR -, HER2 -	A	LR	40502	10/12/2007	Deceased (2012)	35	P21
P_DKP_004	1102006A1	20 * 16	2	IDC	0 of 20	ER +, PR +, HER2 -	No therapy	D	40477	17/01/2007	Deceased (12/07/2012)	45	P22
P_DKP_005	ONC_CP_11_002_0020140-2100	50	2	IDC	10 of 22	ER +, PR +, HER2 -	A	D	40288	20/04/2010	Deceased 25/01/2013	0	P23
P_DKP_006	ONC_CP_11_002002013-2254	30	3	IDC	1 of 12	ER +, PR -, HER2 -	A		None	16/11/2005	Deceased (23/03/2008)		P24
P_DKP_007	ONC_CP_11_002002035-2012	64	3	IDC	7 of 20	ER +, PR -, HER2 -	A	LR	40502	10/12/2007	Deceased (2012)	35	P25
P_DKP_008	201	31	2	IDC	0 of 4	ER +, PR +, HER2 -	A		None	25/09/2012	Alive (12/05/2016)		P26
P_DKP_009	1025	32	2	IDC	2 of 13	ER +, PR +, HER2 -	A		None	15/11/2010	Deceased (25/05/2011)		P27
P_DKP_011	1647	25	2	IDC	0 of 1	ER +, PR +, HER2 -	A		None	15/11/2012	Alive (12/05/2016)		P28
P_DKP_012	1994_002	25	2	IDC	1 of 6	ER +, PR +, HER2 -	No therapy	D	35117	22/12/1993	Deceased (28/07/1997)	26	P29
P_DKP_014	2012_0120	28	2	IDC	0 of 2	ER +, PR -, HER2 -	A		None	22/08/2012	Alive (12/05/2016)		P30
P_DKP_015	2012_0188	9	2	IDC	5 of 13	ER +, PR +, HER2 -	A		None	10/09/2012	Alive (12/05/2016)		P31
P_DKP_016	2012_221	39	2	IDC	6 of 21	ER +, PR -, HER2 -	A		None	10/09/2012	Alive (12/05/2016)		P32
P_DKP_017	2013_0150	18	2	IDC	0 of 4	ER +, PR -, HER2 +	A		None	17/05/2013	Alive (12/05/2016)		P33
P_DKP_018	2013_0272	29	2	IDC	0 of 2	ER +, PR +, HER2 -	A		None	19/06/2013	Alive (12/05/2016)		P34

Table 4 Clinicopathological information for primary breast tumour samples collected and processed for H3K27ac ChIP-Seq.

Sample Identifier	Column1	Recurrence site	Size of metastasis (mm)	Grade	Histology details	Lymph nodes positive	Receptor status	Primary disease diagnosis date	Date of metastatic diagnosis	Alive (+ Consor date) or Deceased (+ Date)	Time to relapse (months)	Further Analysis
M_DKP_003	DKP_001_002_METS	Axillary Lymph node	N/S	N/S	IDC	N/S	ER +, PR +, HER2 -	20/05/2005	03/06/2014	Alive (12/05/2016)	109	NA_M1
M_DKP_045	LUCA004_METS	Sternal mass	85	N/S	IDC	N/S	ER -, PR -, HER2 -	19/03/2010	11/04/2014	Deceased (21/03/2016)	49	NA_M2
M_DKP_016	Legacy-1D1-(Patient 1)	Liver	29	N/S	IDC	N/S	ER +, PR +, HER2 -	24/11/2013	21/11/2013	Deceased (13/08/2015)	0	NA_M3
M_DKP_001	CP0040030006	Liver	N/S	N/S	IDC	N/S	ER +, PR -, HER2 -	08/08/1999	04/03/2006	Deceased (04/03/2008)	191	M1
M_DKP_002	DKP_001_001_METS	Liver	42 x 35	N/S	IDC	N/S	ER -, PR -, HER2 -	04/11/2010	03/08/2012	Deceased (17/08/2014)	21	M2
M_DKP_004	DKP_004_040_METS	Chest wall mass	100 x 83 x 102	N/S	IMC	N/S	ER +, PR -, HER2 -	03/04/1999	19/08/2014	Alive (12/05/2016)	304	M3
M_DKP_005	L001A_METS	Breast recurrence	38	3	IDC	N/S	ER +, PR -, HER2 +	28/02/2014	28/02/2014	Alive (12/05/2016)	0	M4
M_DKP_006	L001B_METS	Breast recurrence	30	3	IDC	N/S	ER +, PR -, HER2 +	28/02/2014	28/02/2014	Alive (12/05/2016)	0	M5
M_DKP_007	L002_METS	Internal thoracic mass	N/S	2	IDC	N/S	ER -, PR -, HER2 +	30/01/2013	25/03/2014	Deceased (01/05/2014)	14	M6
M_DKP_008	LUCA009_METS	Liver	24	N/S	IDC	N/S	ER +, PR +, HER2 -	19/12/2001	27/03/2014	Alive (12/05/2016)	147	M7
M_DKP_009	LUCA004_METS	Liver	20	N/S	IDC	N/S	ER +, PR +, HER2 -	20/02/2007	11/04/2014	Deceased (26/10/2015)	86	M8
M_DKP_011	L008_METS	Groin lymph node	3x9	N/S	IDC	N/S	ER -, PR +, HER2 -	14/08/1998	15/08/2011	Alive (12/05/2016)	179	M9
M_DKP_012	L007_METS	Axillary lymph Node	12	N/S	IDC	1 of 27	ER +, PR +, HER2 -	25/07/2014	25/07/2014	Alive (12/05/2016)	0	M10
M_DKP_013	Legacy-1C3-(Patient 1)	Liver	40	N/S	IDC	N/S	ER +, PR +, HER2 -	21/11/2013	21/11/2013	Deceased (13/08/2015)	0	M11
M_DKP_014	Legacy-1C2-(Patient 1)	Liver	20	N/S	IDC	N/S	ER +, PR +, HER2 -	22/11/2013	22/11/2013	Deceased (13/08/2015)	0	M12
M_DKP_015	Legacy-1C1-(Patient 1)	Liver	24	N/S	IDC	N/S	ER +, PR +, HER2 -	22/11/2013	22/11/2013	Deceased (13/08/2015)	0	M13

Table 5 Clinicopathological information for metastatic breast tumour samples collected and processed for H3K27ac ChIP-Seq.

Table 6 Alignment statistics and peak calling processing statistics in primary tissues

Sample'Number'	Number'of'reads'in'Input'sample	Redundant'rate	Number'of'dedu'reads'in'Control	Number'of'reads'in'H3K27ac'ChIP'sample	Redundant'rate2	Number'of'rdedu'reads'in'treatment'	MACS2'Peaks	further'analysis
G_DKP_006	54,050,875	0.88	6,709,108	82,756,265	0.45	45,725,207	0	**NA_P1**
P_DKP_010	6,857,077	0.10	6,153,624	63,757,687	0.26	47,271,136	0	**NA_P2**
G_DKP_003	39,116,635	0.05	37,014,317	2,091,525	0.01	2,075,954	204	**NA_P3**
P_DKP_013	24,074,930	0.06	22,528,925	46,123,154	0.08	42,345,493	1,468	**NA_P4**
G_DKP_013	43,873,611	0.15	37,140,133	38,059,931	0.07	35,237,112	364,228	**NA_P5**
G_DKP_001	29,416,384	0.15	25,123,727	48,617,322	0.41	28,822,320	15,957	**P1**
G_DKP_002	45,513,399	0.05	43,348,259	26,915,513	0.09	24,404,518	4,175	**P2**
G_DKP_004	51,250,051	0.30	36,085,942	44,081,910	0.66	14,957,984	3,807	**P3**
G_DKP_005	46,998,364	0.14	40,483,700	47,732,047	0.12	42,051,904	11,441	**P4**
G_DKP_007	38,557,798	0.05	36,643,309	37,704,908	0.04	36,170,771	114,806	**P5**
G_DKP_008	46,998,364	0.14	40,483,700	7,604,832	0.03	7,392,475	69,859	**P6**
G_DKP_009	29,647,976	0.12	26,156,591	24,409,362	0.11	24,409,362	52,073	**P7**
G_DKP_010	35,527,333	0.10	31,848,417	16,132,966	0.08	14,801,314	16,207	**P8**
G_DKP_011	8,482,217	0.11	7,585,968	10,730,688	0.07	9,969,157	18,098	**P9**
G_DKP_012	59,423,819	0.15	51,155,594	59,423,819	0.14	51,155,594	85,829	**P10**
G_DKP_014	21,693,843	0.15	18,482,495	37,052,893	0.13	32,246,197	83,447	**P11**
G_DKP_015	28,242,679	0.09	25,822,782	37,386,385	0.10	33,811,074	61,201	**P12**
G_DKP_016	23,202,730	0.16	17,582,561	23,202,730	0.24	17,582,561	96,973	**P13**
G_DKP_017	36,681,717	0.29	24,717,941	49,067,405	0.33	34,973,407	97,466	**P14**
G_DKP_018	40,816,138	0.41	23,896,190	33,080,341	0.40	19,807,861	32,818	**P15**
G_DKP_019	32,033,225	0.09	29,127,585	33,331,400	0.11	29,824,074	123,661	**P16**
G_DKP_020	28,670,863	0.09	26,157,793	23,914,487	0.10	21,598,647	79,846	**P17**
G_DKP_021	41,584,558	0.17	34,432,211	41,584,558	0.22	34,432,211	19,134	**P18**
P_DKP_001	33,868,872	0.11	30,260,525	33,868,872	0.11	30,260,525	130,031	**P19**
P_DKP_002	26,270,448	0.15	21,408,480	23,002,643	0.07	21,408,480	52,672	**P20**
P_DKP_003	18,990,914	0.05	18,747,394	19,954,126	0.06	18,747,394	100,068	**P21**
P_DKP_004	24,755,410	0.13	21,589,384	31,318,859	0.27	43,116,870	10,202	**P22**
P_DKP_005	3,033,924	0.10	2,745,484	3,033,924	0.08	2,745,484	2,383	**P23**
P_DKP_006	26,270,448	0.15	22,222,658	31,013,027	0.17	25,605,492	55,272	**P24**
P_DKP_007	10,325,143	0.58	4,303,907	25,407,101	0.12	22,421,625	42,429	**P25**
P_DKP_008	6,954,287	0.03	6,742,122	61,660,871	0.14	52,984,880	17,986	**P26**
P_DKP_009	7,304,342	0.09	6,653,058	75,611,766	0.17	63,052,925	2,521	**P27**
P_DKP_011	49,675,611	0.12	44,726,242	35,312,910	0.10	31,111,580	172,092	**P28**
P_DKP_012	24,787,260	0.13	21,673,823	46,059,302	0.18	37,686,881	18,340	**P29**
P_DKP_014	57,577,050	0.19	46,357,312	73,139,872	0.17	60,360,449	155,186	**P30**
P_DKP_015	25,984,041	0.05	24,777,058	36,541,319	0.21	28,962,803	101,444	**P31**
P_DKP_016	51,860,134	0.12	45,390,661	41,575,023	0.22	32,249,718	171,338	**P32**
P_DKP_017	17,994,528	0.06	16,960,718	40,731,205	0.21	32,111,033	82,394	**P33**
P_DKP_018	38,222,773	0.12	33,794,658	62,871,556	0.30	44,085,078	30,615	**P34**

Average	31,707,430	0.16	26,077,804	37,842,680	0.18	30,717,373	64,043
standard◂deviation	15,102,022	0.16	12,943,206	19,016,919	0.13	14,460,596	70,892

Sample'Number'	Number'of'reads'in'Input'sample	Redundant'rate	Number'of'reads'in'control'sample'after'Filtering	Number'of'reads'in'H3K27ac'ChIP'sample	Redundant'rate2	Number'of'reads'in'treatment'sample'after'Filtering	MACS2'Peaks	further'analysis
M_DKP_003	9,463,740	0.68	3,016,791	30,311,922	0.74	7,944,040	1,040	NA_M1
M_DKP_010	23,859,790	0.18	19,624,489	27,669,040	0.14	23,867,566	84	NA_M2
M_DKP_016	18,633,770	0.14	15,998,892	21,114,290	0.16	17,636,173	825	NA_M3
M_DKP_001	23,851,177	0.05	22,756,985	22,526,970	0.04	21,681,987	8,431	M1
M_DKP_002	14,117,528	0.15	11,969,698	12,706,557	0.15	10,792,912	30,011	M2
M_DKP_004	34,440,294	0.58	14,625,360	48,687,811	0.58	20,663,350	5,133	M3
M_DKP_005	24,582,698	0.03	23,809,766	61,786,583	0.12	54,499,692	99,738	M4
M_DKP_006	26,186,664	0.04	25,169,712	25,958,127	0.06	24,314,944	89,530	M5
M_DKP_007	28,787,681	0.05	25,709,307	28,647,301	0.10	25,709,307	60,791	M6
M_DKP_008	29,260,547	0.42	16,831,215	26,329,506	0.47	13,990,704	7,416	M7
M_DKP_009	30,119,569	0.30	21,228,854	26,207,315	0.36	16,746,266	9,864	M8
M_DKP_011	46,279,779	0.04	44,568,755	31,207,234	0.03	30,243,502	53,772	M9
M_DKP_012	73,251,397	0.09	66,598,280	36,195,802	0.12	31,696,346	82,105	M10
M_DKP_013	38,574,636	0.12	34,118,135	24,677,684	0.17	20,574,946	3,559	M11
M_DKP_014	27,466,355	0.09	24,858,027	47,227,405	0.10	42,434,157	16,226	M12
M_DKP_015	15,713,576	0.19	13,262,804	15,713,576	0.16	13,262,804	62,707	M13

Average	29,036,825	0.20	24,009,192	30,435,445	0.22	23,503,669	33,202	
standarddeviation	14,942,657	0.20	14,780,839	12,678,235	0.21	11,934,287	35,603	

Table 7. Alignment statistics and peak calling processing statistics in metastatic tissues.

2.2 H3K27ac landscape of 47 ERα + Breast Cancer Tissues

Pre-processing analysis revealed that seven samples were not fitting the required standard of peak enrichment for further investigation. Consequently, the whole study was conducted using a dataset composed of 33 primaries and 14 metastatic (Fig.31) tissues processed using H3K27ac ChIP-seq.

H3K27ac is a marker of active promoter and enhancer, and subsequently, we assigned the peaks to the promoter and enhancer category. Promoters were the peaks or the portion of peaks occurring within the gene TSSs and 1kb upstream, while the enhancers were all the peaks or portion not overlapping with the list of promoters. This annotation method classified 23,976 active promoters and 326,729 active enhancers.

To assess the adequacy of the available tissue samples for achieving the study aims, we calculated the saturation curve of the active regulatory regions marked by significant peaks. A saturation curve of genomic data reveals the amount of new genomic information added by each sample. The plateau phase of the curve, if reached, show that saturation of information is obtained and adding further samples to the dataset may not provide any novel information.

Saturation plots in enhancer and promoter sets showed different curve growth patterns. Enhancers data covered 80% (Fig.33 left) of the potential regulatory activity without reaching the plateau while the same results in promoter occurred after profiling only four tumours. These results were in agreement with enhancers being more cell type-specific[75] and heterogeneous compared to promoters. Besides, the average size of peaks is similar between the two groups and not a potential source of bias in the genomic coverage calculation. Furthermore, we investigated the penetrance of each regulatory region in the dataset by calculating the number of patients sharing each specific location (Fig.33 right). This analysis revealed, as expected, promoters' activity is much more conserved compared to the enhancers. Indeed, just 14% of promoters are sample specific while the vast majority of enhancer regions, roughly 80%, are patient specific or shared only within two patients (Fig.33 right). These data demonstrate that enhancers are harbouring the greatest majority of potential epigenetic heterogeneity in ERα BC.

As a result of the assessment of the adequacy of genomic coverage within the H3K27ac marked regions, we proceeded in building a map of the active promoters and a map of active enhancers in ERα BC which has been used as reference annotation of potentially active regulatory regions in the analysis described in the following paragraphs.

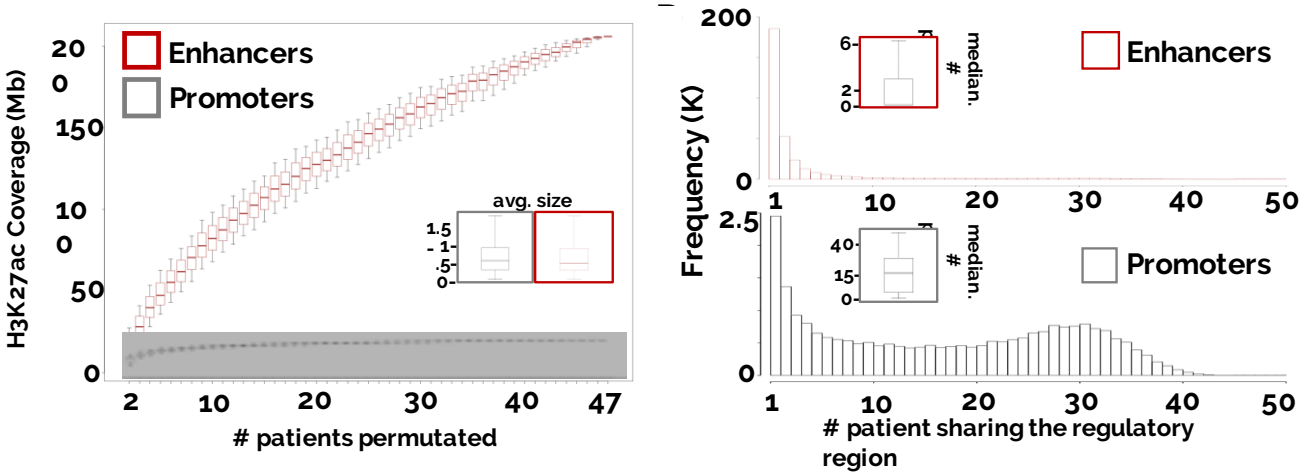


Figure 33 Saturation analysis of Enhancer and Promoters and sharing distribution in N=47 ERα Breast Cancer.

Left figure: Enhancers (red) and Promoter (grey) regions are illustrated as the coverage (y-axis) achieved at each permutation (x-axis). The average peak size in promoters and enhancers is 517 and 463bp, respectively. Right figure: Sharing of H3K27ac peaks at enriched enhancers (red plot, top graph) and promoters (grey plot, bottom graph) across 47 patient samples. The median (represented by the black lines in the small inset diagrams) number of patient samples sharing the same enhancer peak = 1 and the median number of patient samples sharing the same promoter peak = 15. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone].

2.3 Epigenetic information as a surrogate of cell phenotype of ERα + Breast Cancer Tissues

Numerous efforts have been devoted to understanding the consequences of genetic variability in cancer cell phenotypes. But still, the impact of genetic changes to phenotypical characteristics is far from being understood.

Recent studies suggest that epigenetic mechanisms may prevail on genetic information in contributing to cell phenotype in normal cell development[101,102] . This would suggest not only that the interplay of genetics and epigenetics maintains phenotypical cell characteristics, but also that epigenetics can confer a stronger impact on cell phenotypic state. This paradigm shift opens new strategies in dissecting the phenotypical landscape of tumour tissues.

Inferring phenotypical heterogeneity using epigenetic signals from H3K27ac ChIP-seq may not be a challenge. A recent study proposes new principles of interpretation of epigenetic signals. Buenrostro et al. suggest that ATAC-seq signal is linearly proportional to the number of cells contributing to it[103]. This implies the possibility to qualitatively assess the proportion of cells, within the examined tissue, carrying each given nucleosome-free region (NFR). This finding can easily translate to histone ChIP-seq where the signal would be the representation of nucleosome positioning. Thus, we tested this hypothesis by performing a spike in experiments in which known numbers of cells with characterised active regulatory regions and the same genomic background were admixed in incremental proportions before H3K27ac ChIP-seq (Fig.34). The results confirmed the hypothesis with a strong positive correlation between the signal intensity and the number of cells analysed with H3K27ac qPCR. To corroborate this finding, we also repeated the same experiment performing ChIP-qPCR at well-known ERα target sites (EGR3 enhancer, TFF1 promoter, SLC7A5) and a not target (PTRF promoter) in MCF7 upon induction of oestradiol for 45m (Fig. 35). The experiment was performed in triplicate and evidences a positive correlation between signal intensity and the number of cells in the plate.

Collectively these experiments suggest that a qualitative assessment of the ChIP-seq signal might correlate with the clonality of the population of cells for a particular phenotype such an active enhancer. Based on this data, we developed a novel computational method that for the first time relate tumour bulk epigenomic data to phenotypical clonality. We propose a new metric the Ranking (RI) and Sharing indexes (SI), extensively described in the following paragraphs.

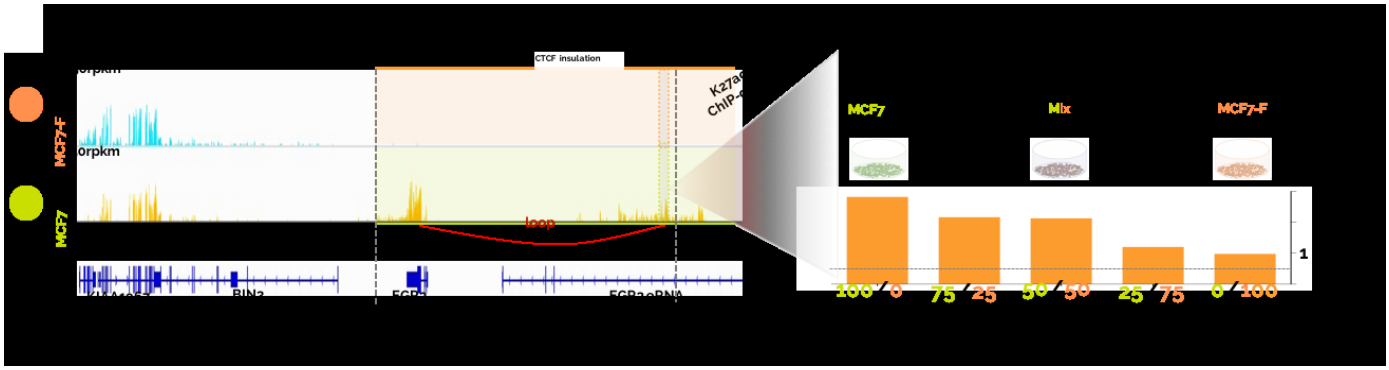


Figure 34. EGR3 transcriptomics, epigenomics and spike in-experiment.

EGR3 mRNA is expressed in MCF7 but not in MCF7-F cells. eRNA and Pol-II ChIA-PET show enhancer activity in MC7 but not MCF7-F. The spike-in experiment (right side) shows that observed ChIP–qPCR signal for H3K27ac at EGR3 enhancers is proportional to the number of MCF7 cells mixed in the sample. [Figure produced by Luca Magnani].

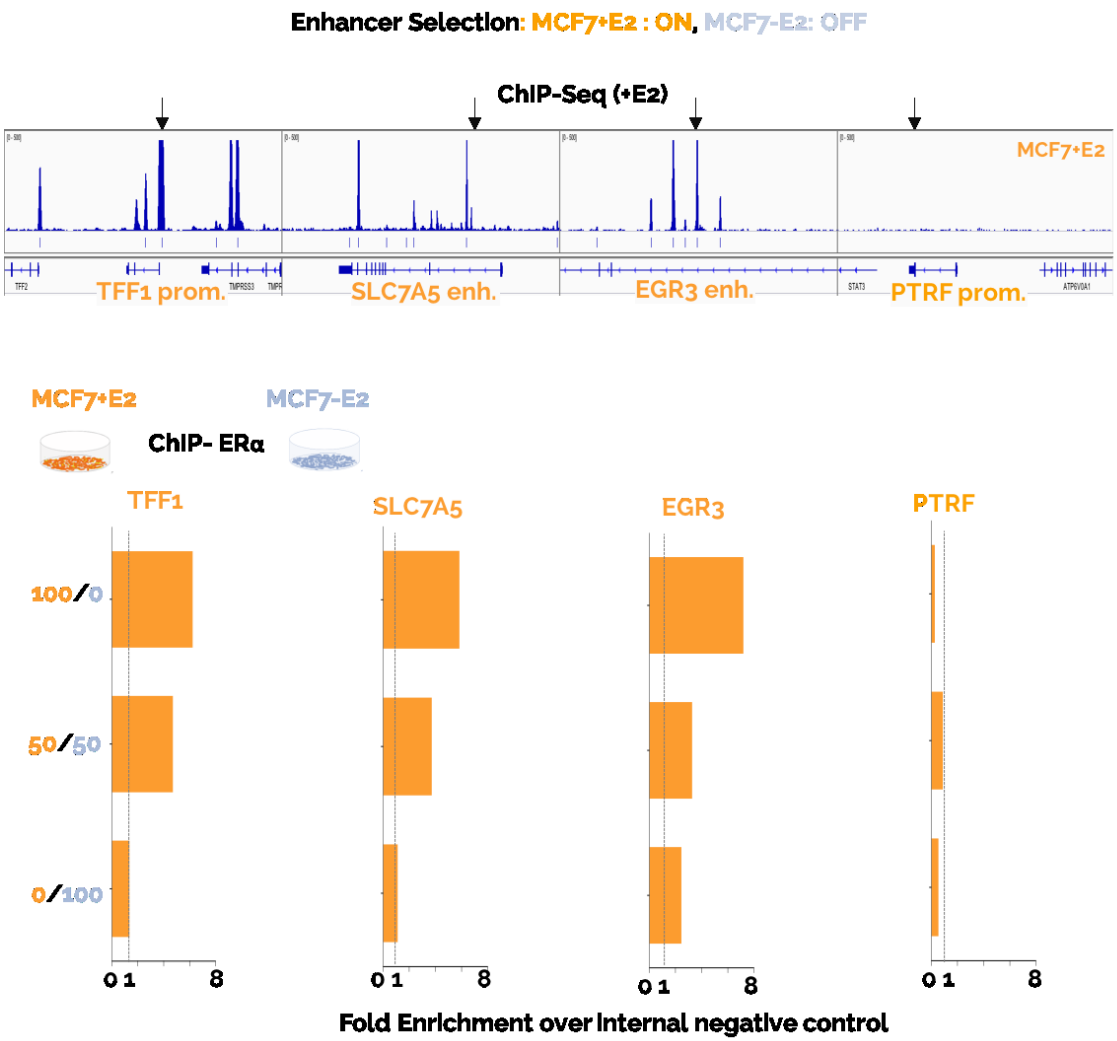


Figure 35. EGR3 transcriptomics, epigenomics and spike in-experiment.

Upper panel: ChIP-seq profiles of Era in the 4 loci under examination. Lower panel: Spike-in experiment results, X-axis=fold enrichment over negative control, Y-axis=mixture of cell in the plate. [Analysis performed by Giacomo Corleone. Figure produced in collaboration with Luca Magnani].

2.4 A Novel metric for assessing phenotypical heterogeneity

A comprehensive assessment of phenotypical heterogeneity of ERα+ BC has never been performed to our knowledge. The previous paragraphs of this dissertation provided the theoretical concepts underlying this study and highlighted epigenetic signals, obtained from H3K27ac ChIP-seq and ATAC-seq, as a proxy of cell phenotypical clonality. We demonstrated with low throughput ChIP qPCR that signal intensity of each active regulatory regions correlates with the number of cells

contributing to it. These findings led us to develop a computational framework which systematically extrapolates phenotypic heterogeneity in tumour samples by analysing epigenomic data.

This framework consists of the assignment of a bidimensional scoring to each regulatory region being part of the enhancers and promoters map. The first score is an index of peak enrichment (Ranking Index RI) and the second an index of penetrance in the dataset (Sharing Index SI). RI is a calculated individually in each sample by 1) counting the number of deduplicated reads contributing to each signal 2) normalising them by the number of deduplicated reads of the sample and then 3) ranking them within the sample where RI 1 is the strongest while RI 100 is the weakest enrichment. Ultimately, RI is associated with regulatory regions active in most cells while RI 100 marks regions active in an only small fraction of the cell population. The mean of RI of any given regulatory regions among the samples is then assigned to the correspondent region in the reference map of activity.

The second score assigned to the regulatory regions is the SI. SI is the number of samples sharing at least 60% of a given region. SI spans 1 to 47 (the total number of samples in the dataset). The peak overlap is defined with a static threshold and peaks shared among many samples have a higher value. Summing up, RI metric qualitatively predicts the clonality of enhancer and promoter in ERα+ BC while SI measures the usage of those active regulatory regions across the tumour samples.

Next, we investigated the relationship between SIs (x-axis) and RIs (y-axis) (Fig.36). RIs were binned according to the associated 47 SI, and linear regression analysis reported an R^2 of 0.97 between SI and the median of binned RIs. High penetrant enhancers and promoters are significantly more enriched compared the rest of the dataset. This finding indicates that higher shared regions (High SI) are more prone to be clonal (Low RI) whereas regions active in a modest number of patients tend to be sub-clonal. Of note, we report a small proportion of regions associated with high RI /low SI which may be explained by a clonal phenotype specific to a limited number of patients.

This investigation provided principles to deconvolute the complex phenotypical landscape of enhancer and promoter activity in ERα+ BC and allowed the prioritisation of regulatory region subsets for further investigations.

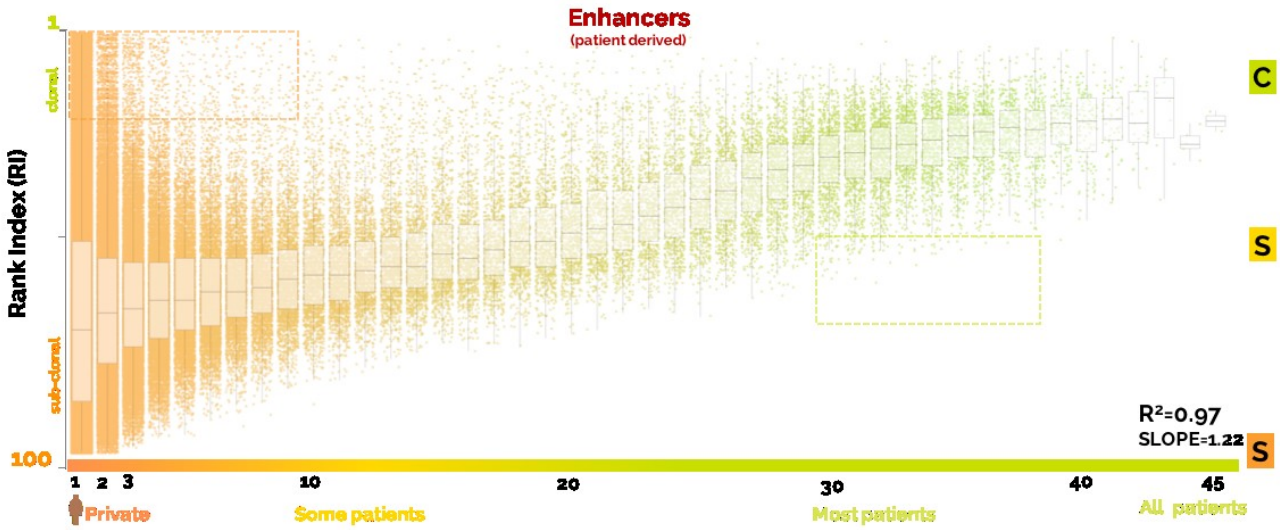


Figure 36. Enhancer clonality and SI relationship.

Linear regression shows clonal enhancers (Upper right) are commonly shared across BC patients. A discrete proportion of enhancers have extremely low RI despite being private (dashed areas orange) or higher RI while being shared (dashed areas yellow). Y-axis = RI, x-axis = SI. Each dot represents the median RI for each single enhancer. Boxplots show the median RI value and interquartile ranges of enhancers with the same SI. Dashed areas represent regions in the map which may be considered for further investigations. Colour scheme: Green= clonal regions (C), RI from 1 to 20; Yellow=subclonal regions (S), RI from 21 to 80; Orange= Subclonal regions (S), RI from 81 to 100. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone].

2.5 Assessment of potential source of bias in regulatory regions mapping

We assessed these results for possible artefacts, biases and data processing inaccuracies. Potential copy number aberrations (CNAs), dinucleotide artefacts, cell cycle effects and peak calling may significantly affect RI and SI calculation. Recurrent CNA in a small fraction of patients would be sufficient to explain the relation low-SI/low-RI peaks (patient-specific clonal regulatory regions) whereas dinucleotide artefacts are a known source of enrichment inflation in ChIP-seq. Then, ChIP signals may be influenced by the phase of the cell cycle during which the sample is collected and thus affecting the logistic relationship between RI / SI. Lastly, detection biases across the samples and the

SI formulation, which relies on a static threshold and peak calling calibration, are also potentially a source of bias.

2.5.1 CNA

ChIP-seq call was run using the matching input. This approach would normalise H3K27ac signals to the patient genetic background with the consequence of removing any chromosome amplification or depletion which may affect the peak calling. Despite the safety of the used approach, we identified the CNA regions of the genetic background of each sample to gain further insight into the relationship with the new metric system. CNAs were detected from the input of each sample and then matched with the relative enhancer clonality.

This analysis showed that CNA positioning was in line with published datasets[60] (Fig. 37). Indeed, accordingly with previous[104] CGH, FISH assessment of genetic profiling of ERα+ BC we report common focal aberrations (gains and losses) of restricted and large chromosomal regions. Gains of chr1q and chr16p and chr20 coupled with the loss of large portions of chr 13, chr18 and chr22 are well-known signatures of ERα+BC genomes. In addition, we report profound patient-specific variabilities such as amplification at chr7 in 30% of the patients and large deletions of chr2 in a small subset. This genomic mapping was performed using the input DNA of each patient by shallow whole genome sequencing, a robust approach for copy number profiling[105]. Of note, the samples previously deleted from the analysis were included in the copy number profiling and resulted in having the DNA too fragmented for Chip-seq purposes (Fig. 37).

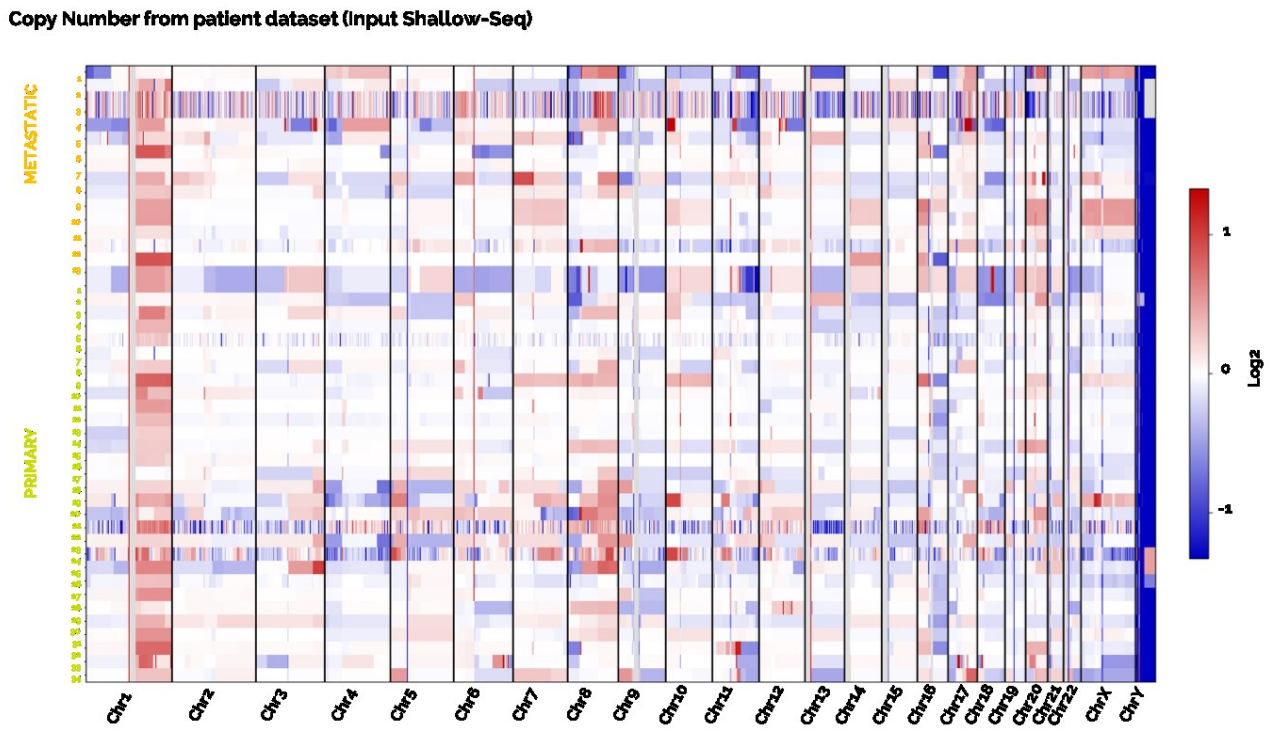


Figure 37. CNA profiles of samples.

Each row of the heatmap represents a sample and each column a chromosome. Colour scheme: Red show amplifications while blue shows deletions. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

We then assessed whether copy number aberrations were affecting the RI/SI calculation of the enhancers in each chromosome. Theoretically, higher SI (a proxy of clonality) would match amplified regions while low SI the region's loss. We performed a systematic analysis to assess this possibility by matching the SI of the enhancers with their genomic positioning in function of copy numbers and then calculating the prevalence of any significant relationship.

Here is reported the result of the analysis on the enhancers set in chr1. A large amplification on the q arm of this chromosome is prevalent in all the samples. If the hypothesis of any bias was correct, then we would observe higher SI enhancer in the q arm and lower SI clustering in the p arm. As shown in Fig.38, only 1 out of 9 high clonal enhancers dwell in the q arm while the majority in the p arm.

Besides we performed this estimation throughout each chromosome not finding evidence of any source of bias.

Overall, not significant shreds of evidence were found to accept the hypothesis of copy numbers affecting the SI/RI calculation.

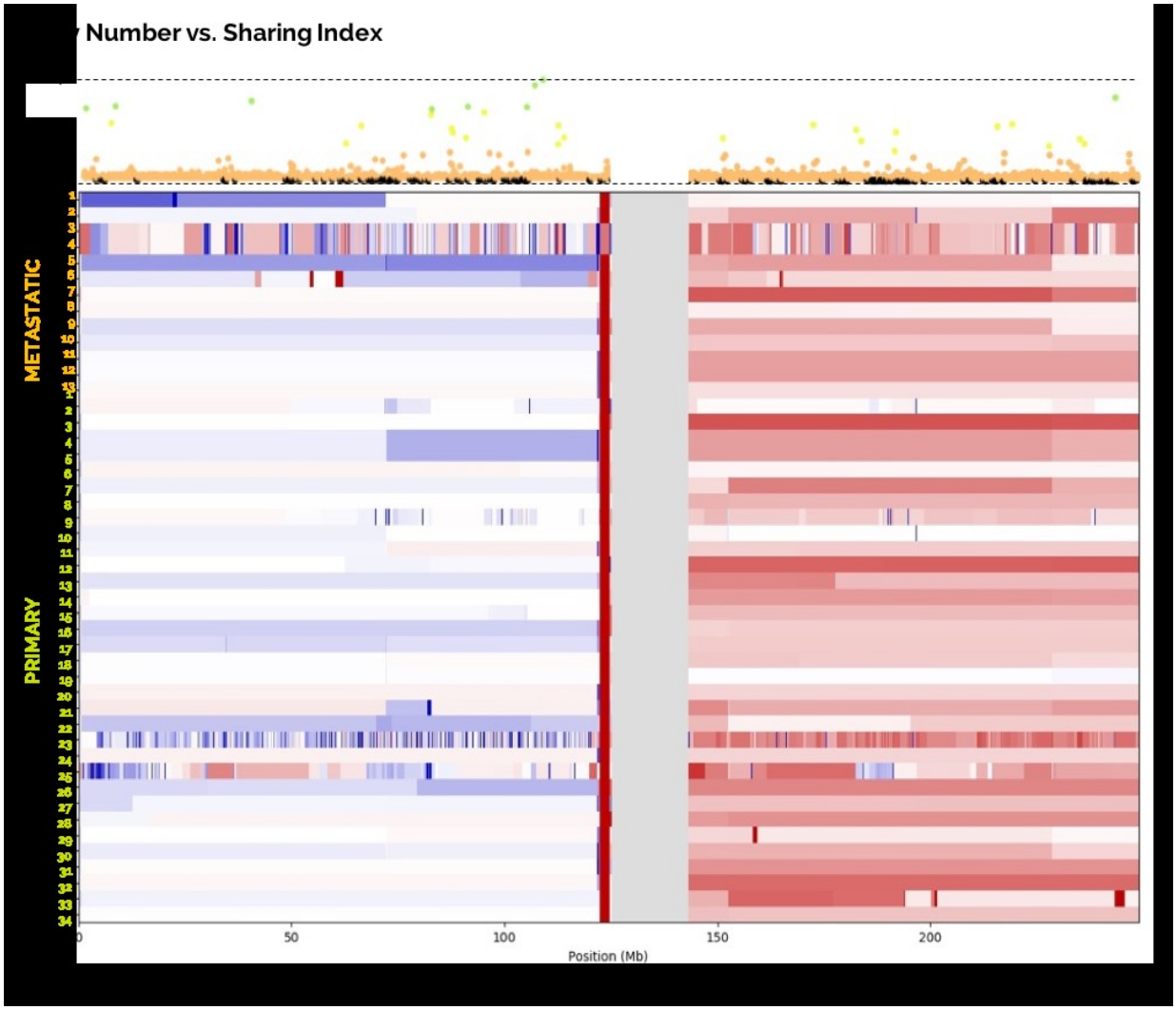


Figure 38. CNA profile of chromosome 1.
Each row of the heatmap represents a sample. Red show amplifications while blue shows deletions. Above the heatmap: Each dot represents an active enhancer in function of the SI. Colour scheme of dots: Orange: SI<20; Yellow: SI>20 & SI <=30; Green:SI>30. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

2.5.2 Dinucleotide artefacts

It is known[106,107] that the DNA fragmentation and amplification, two necessary procedures in DNA sequencing technologies, produce a non-random bias in the nucleotide composition which leads to over or under-sampling specific regions in the genome. The majority of artefacts do not map to the reference genomes and are discarded in further analysis steps. However, a fraction of them efficiently aligns to the reference genome thus influencing the biological interpretation of the data. Elevated GC bias is ordinarily reported in sequencing applications[106] and a significant source of variability large enough to result in a different interpretation of the data.

In this context, we assessed the eventuality of a dinucleotide composition affecting the SI/RI relationship by calculating the frequencies of dinucleotide content in the reference genome hg19 and the pooled input of our dataset and then manually comparing it with the dinucleotide rates in the binned SI enhancers (Fig. 39, Fig. 40).

Dinucleotide frequencies in the pooled input of our dataset were similar to the reference genome, and in particular, GC and the complement CG content did not result qualitatively different to the expected rate. Then, sonication impact on the sequenced genome did not show any particular distribution suggesting the reliability of the genomic fragmentation (Fig.39).

We further calculated the frequencies of the di-nucleotide content of the total enhancers in function of the SIs (Fig.40) and no trend emerged as a potential source of bias.

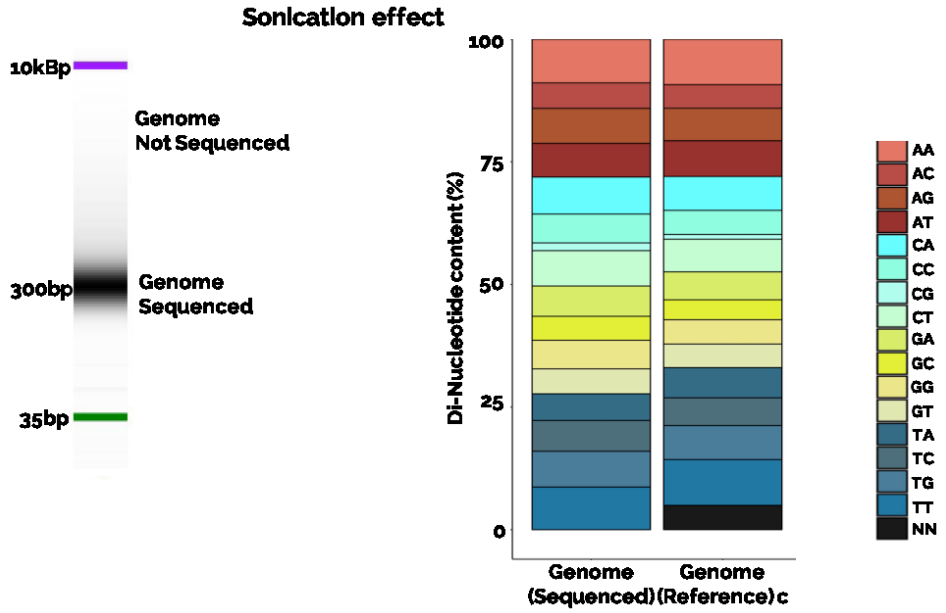


Figure 39. Dinucleotide composition of the hg19 reference genome and sequenced genome.

Left: Fragment distribution of the 55 ChIP-seq libraries sequenced. Right: Dinucleotide content in the reference genome and the pooled inputs. Each colour represents a different dinucleotide [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

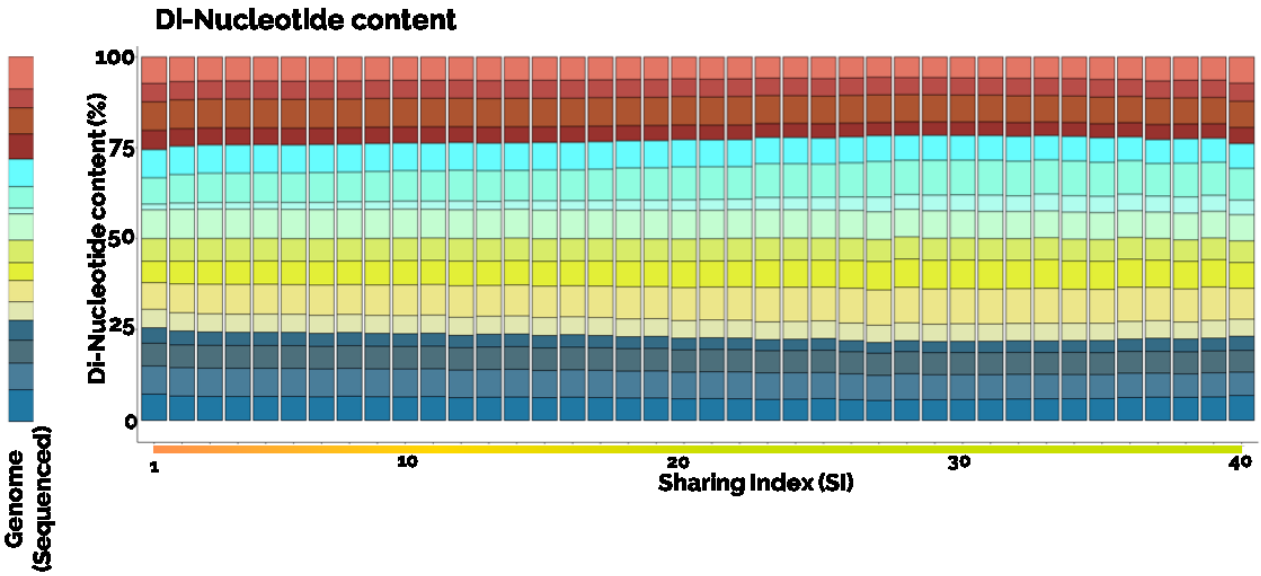


Figure 40. Dinucleotide content of each SI bin.

The illustration shows the dinucleotide content with the same colour scheme of figure 39. Trends show dinucleotide content across each SI bin in enhancers. X-axis=SI binning, y-axis= percentage of dinucleotide content. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

2.5.3 H3K27ac marker during the cell cycle

Besides, we assessed the stability of H3k27ac marker during the cell cycle to exclude the hypothesis of turnover dynamics of H3K27ac may alter SI and RI formulation. Theoretically, RI may reflect the time nucleosomes exhibit H3K27ac marker during the cell cycle. This would result in SIs low regions being a representation of loci having H3K27ac with a fast turnover while SIs high slow turnover.

Thus, we investigated the turnover dynamics of H3K27ac among the cell cycle from a pool of MCF7 cell lines. We sorted MCF7 at G1 and G2-M phase separately by FACS after DAPI staining in order to measure the amount of DNA (experimental procedure performed by Laura Jilderda), then performed H3K27ac ChIP-seq and finally observed a strong linear relationship between the two population (r^2=0.82) (Fig.41). These results demonstrated that H3K27ac signal is stable across the cell cycle in the large proportion of enhancers. However, it is not possible to completely exclude an effect upon a minority of a subclonal population of more dynamic enhancers.

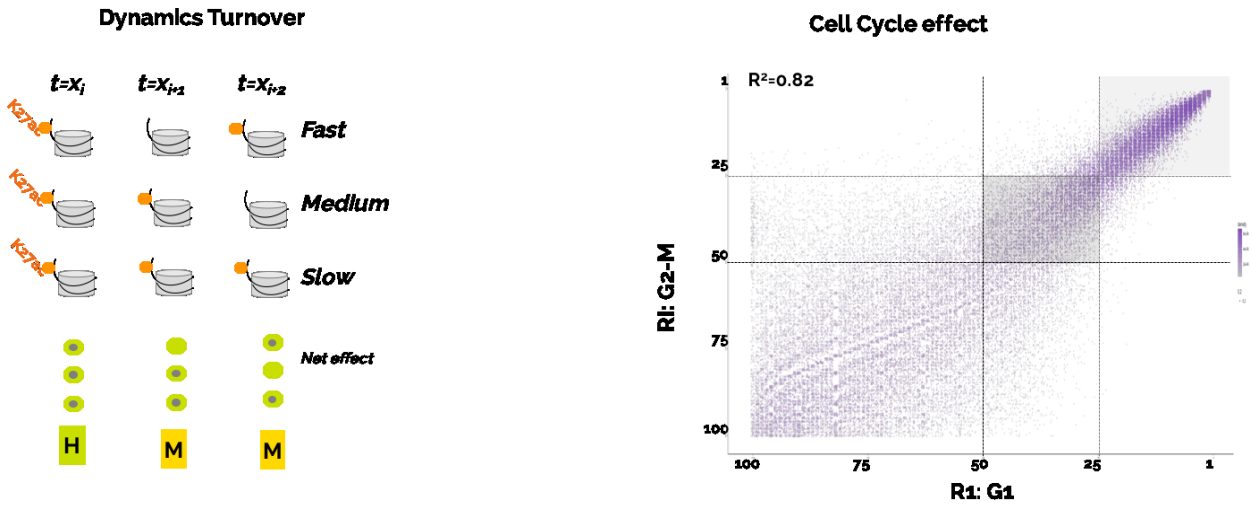


Figure 41. H3K27ac marker during the cell cycle.

Left: Theoretical turn-over dynamics of H3K27ac marker may affect peak detection. Right: Linear regression of synchronised MCF7 at G1 (x-axis) and G2 (y-axis) phase. Colour scheme: Dark purple indicates strong correlation data of the two conditions. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

2.5.4 Reproducibility of SI formulation

The main result of the SI/RI analysis is the finding of a strongly linear relationship between enhancers SIs in the dataset and relative enrichment RI.

To show the robustness and reproducibility of the R-squared observed between SI and RI, we run a parallel analysis in which the SI was reformulated by using a different method of calculation. This method was based on continuous measurements of signals to appraise the consequences of the SI distribution amongst our data compared to the discrete metric described in the Methods section.

Indeed, a discrete formulation, on which is based the current SIs, may cause issues dependent by 1) the significance of peak detection value and 2) the static threshold of 40% overlap coverage used to define shared peaks amongst the samples. Thus, the discrete metrics may be over or down estimated by peak calling detections bias, ultimately strongly affecting the SI/RI analysis.

Thus, to avoid any misinterpretation of the data, we applied a novel quantile normalisation method on the signal population after relaxing the threshold of peak detection. Briefly, (see Methods for an extensive description) we recalled all the peaks relaxing the threshold of detection. We created a master list of all the peaks which was used to build a matrix (Fig.42 A) having as columns all the samples and as rows all the master peak enhancers. The matrix was populated with the enrichment (the number of reads per sample matching each enhancer), and subsequently, all the data were normalised and compressed. Then, we iteratively applied 26 thresholds to the matrix (Fig.42 B-C), plotting the number of enhancers for each interaction in the function of the R^2 obtained with the relative enrichment (Fig.42 D).

Interestingly, this method allowed to define a range of threshold spanning 11 iterations of selection supporting a linear regression with the enrichment. However, a dramatic drop in the peak population was following the increase of the threshold leading to an over-conservative selection.

Comparable results were obtained when we applied the same computational workflow on the enhancers selected with the discrete method (Fig.42 E).

Collectively these data demonstrated that the RI/SI formulation obtained with different methods (described in the Methods) robustly show a linear relationship between SI and RI. However, adopting cut-offs that are outside what is typically practised in the field, either calling every single tag a peak or increase the stringency to extremely high values result in losing this correlation.

In conclusion, the results described in this chapter are accurate and not prone to independent macro artefacts potentially affecting the linear relationship of enhancer penetrance with the enhancer enrichments and are highly reproducible amongst different methods of analysis.

2.5.3 Assessment of genetic alteration systematically affecting the results.

Genetic alterations may affect the DNA accessibility and ultimately being a source of patient-specific bias. Theoretically, the genetic mutations in genes previously associated with endocrine therapy can drive the rewiring of the epigenome. Thus, we gained background on the potential synergy between epigenetic changes and genetic drivers by developing a cancer hotspot panel which was including ESR1, TP53 mutations (Fig.43). We performed targeted capture sequencing of this panel using input material of the dataset described in this chapter and demonstrated that PIK3CA mutation affects 13/55 patients, while ESR1 only in 4 out of 13 metastatic patients. Overall, this analysis supports that mutations may be effectively a sample specific source of bias. However, we did not find consistent mutations in our dataset. Undoubtedly, future studies may exploit these data by linking the genetic and epigenetic background of the patient in the dataset to identify evidence of genetic/epigenetic interplay.

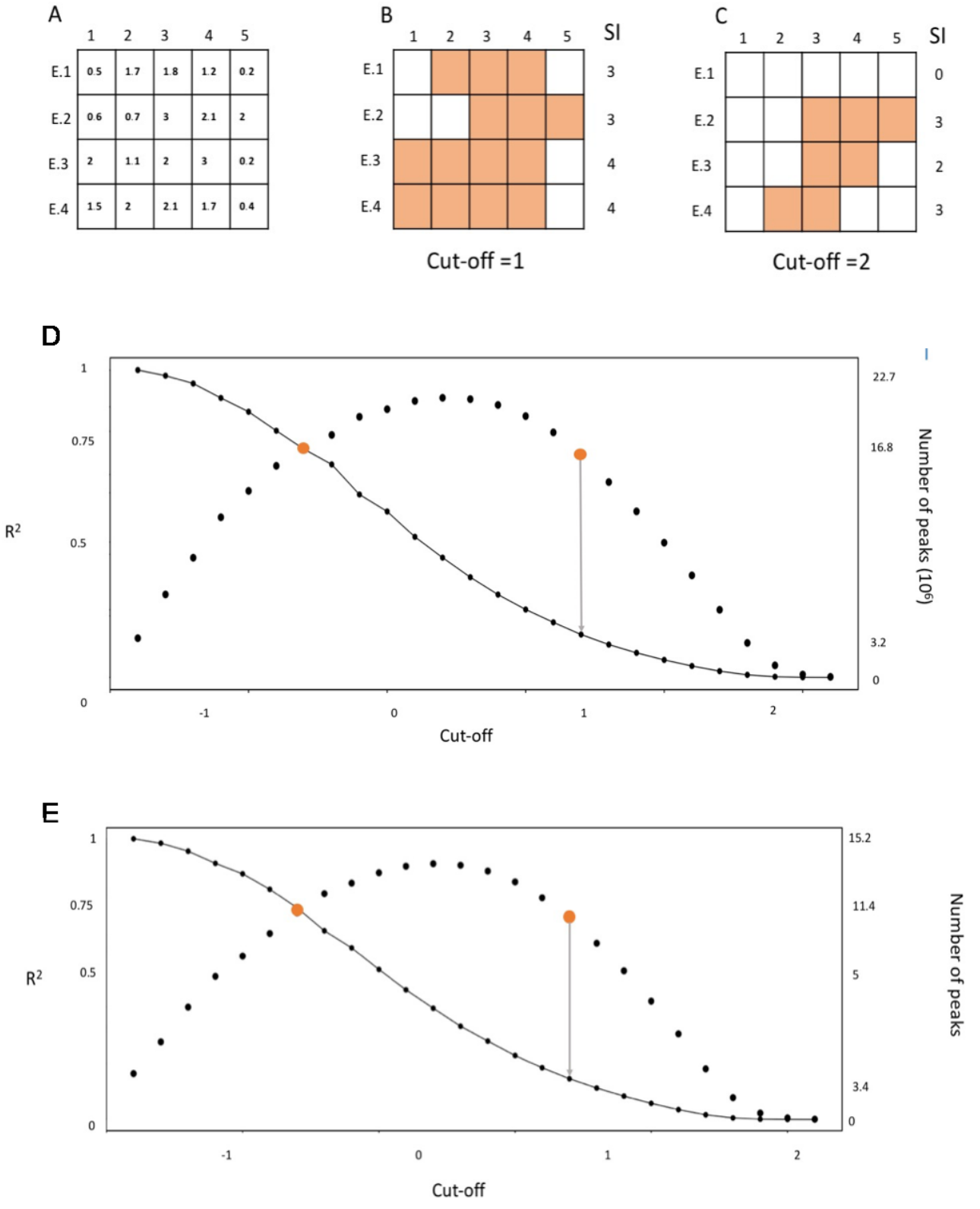


Figure 42. Reproducibility of SI formulation.

A) The matrix of all peak enrichment scores. Columns: sample; Row: enhancers B-C) Example of iteration with two thresholds applied. Orange cells represent the selected enhancer in the given samples when the threshold is applied. D)R squared (y-axis left) of each linear regression in function of the cut-off applied in the continuous method. Smaller dots represent the number of selected peaks (y-axis right) at each given cut-off E) R squared (y-axis left) of each linear regression in function of the cut-off applied in the discrete method. Smaller dots represent the number of selected peaks (y-axis right) at each given cut-off. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

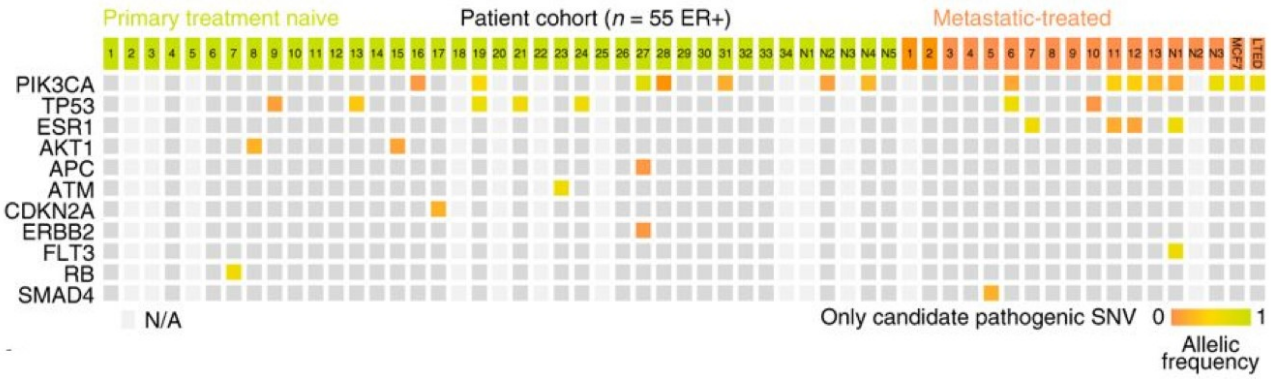


Figure 43. Targeted sequencing for common genes affecting BC.

Colours show the allele frequency of identified mutations. Colour scheme: from green=100% allele frequency to orange=0% allele frequency [Experimental procedures performed by Ylenia Perone. Analysis performed by Iros Barozzi. Figure produced by Luca Magnani]

2.6 Phenotypic evolution during BC cancer progression and therapy selective pressure

Tumours exhibit a high level of heterogeneity and myriad of different genetic subclones are present in the tissue. The investigation of evolutionary trajectories of tumours has primarily focused on the identification of the genetic events occurring during the progression of the disease. The primary outcome of these analyses is the definition of clonal and subclonal events by assessing the genetic code of the cell population.

Generally, genetic clones are directly linked to cell phenotype; however recent studies[107] suggest that genetics does not uniquely drive cell phenotype and a key role in this aspect may be occupied by epigenetics. To our knowledge, there has not been a systematic study that focuses on the epigenetic events occurring in cancer between primary and metastatic stage of BC.

We, thus, assigned SI and RI indexes to each regulatory region in primary and metastatic tissues available in our dataset to describe potentially evolutive trends. We hypothesised that the evolution of phenotypic clones might be monitored by inferring the clonality of each regulatory region during BC progression (Fig.44). Theoretically, phenotypical evolution may result in the activation and

deactivation of regulatory regions that can be estimated by assigning SI and RI to primaries (treatment naïve) and metastatic (treatment-resistant) sample.

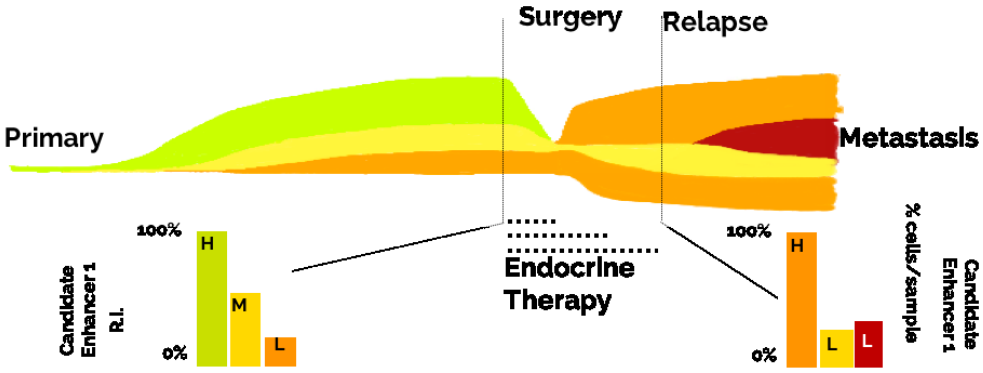


Figure 42. The theoretical framework of the analysis.

The relative size of phenotypic clones can be tracked using enhancer ranking (RIs). RIs and SI scores can be leveraged to trace phenotypical clones during Breast Cancer progression in response to ET. [Figure produced by Luca Magnani].

We have demonstrated that SIs is a surrogate of phenotypical clonality in the above paragraphs. We took advantage of this notion to systematically assign primary and metastatic specific SIs to the active regulatory regions and evaluate the differential bias between the two states (Fig.45).

Most regions did not significantly diverge between primary and metastatic samples (Fig.45). In a population of 320K enhancers, only 0.1% were differentially active in primaries and roughly the same number in metastatic patients ($P < 0.05$). The promoter subset showed very similar results with 239 regions differentially active in primary samples and only 20 biased towards the metastatic samples. Then, we checked the RIs assigned to the three identified subsets of enhancers (Fig.46). As expected, RIs were significantly different in the primary and metastatic subset and concordant in the non-differential enhancers. In addition, to corroborate these results we permutated 50 sets of 320 non-differential enhancers to finally demonstrate that the patterns of differential clonality are not affected by any artefacts or stochastic events.

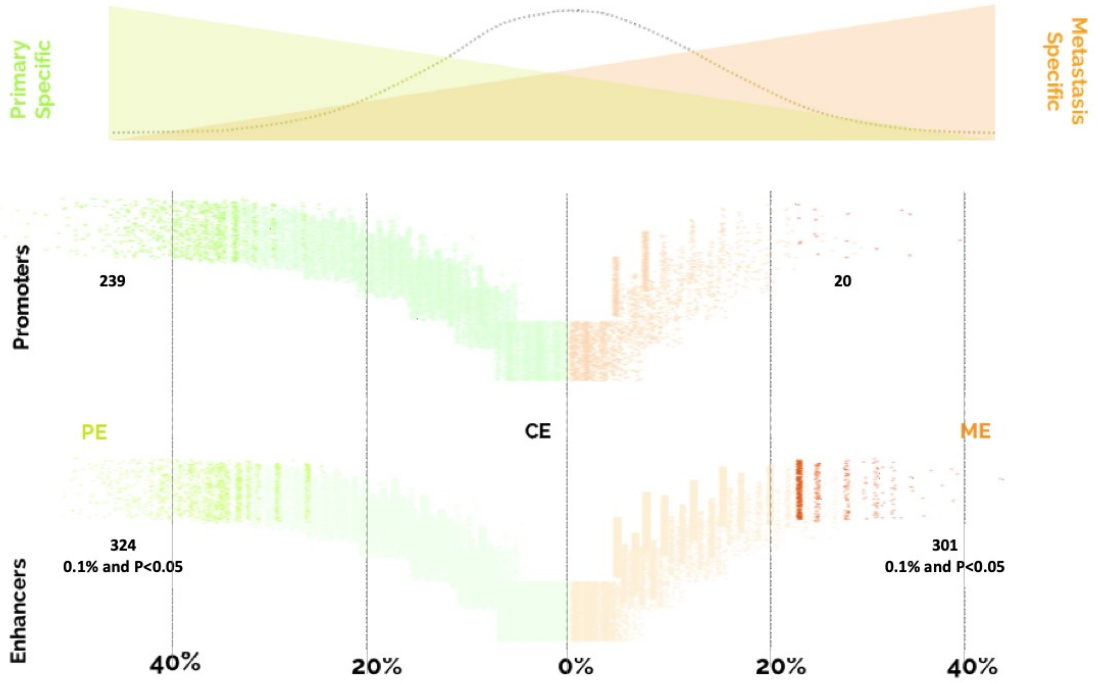


Figure 43. Volcano plot showing the usage of enhancers and promoters in primary and metastatic patients.

Enhancer and promoters' stratification based on SI in primary and metastatic patients. Percentages (x-axis) are calculated for each regulatory region in primary and metastatic. Y-axis represents the significance of the regulatory regions in the distribution satisfying the Fisher exact test P<0.05. PE=primary specific enhancers, ME=metastatic specific enhancers, CE=not differential enhancers between primary and metastatic. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

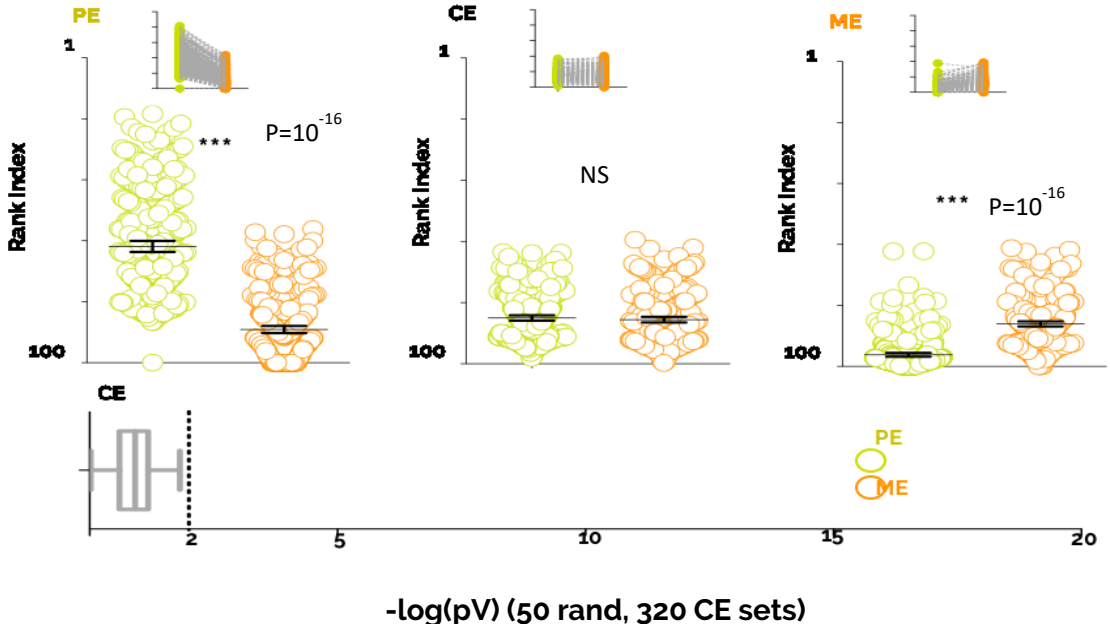
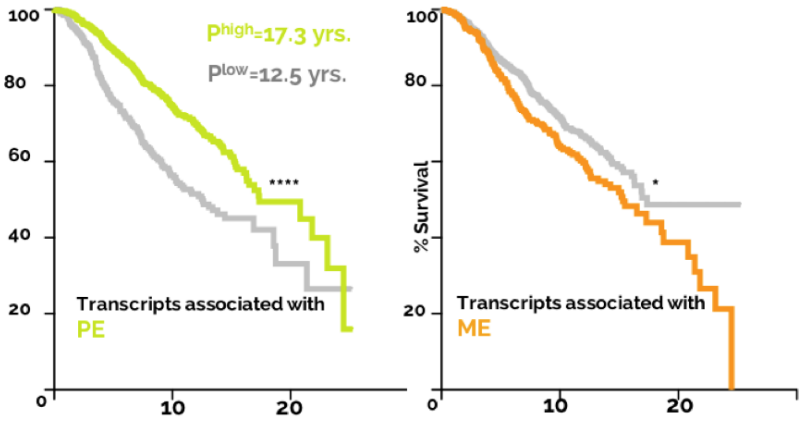


Figure 44. RI of differential enhancers in primary and metastatic patients.

This dot-plot represent RIs of all PE, ME and CE enhancer sets in primary and metastatic samples. Bottom: permutation used to assess the significance of RI variation in both conditions. 50 permutations of 320 randomly selected CE. Box and whiskers represent median and 1–99th percentile for the P distribution. A Wilcoxon matched-pairs signed rank test was used to test for statistical significance. NS, not significant. [Analysis produced by Giacomo Corleone. Figure produced in collaboration with Luca Magnani]

We further explored the functional role of the 2 subsets of differential enhancers by performing survival analysis from a panel of 1,427 ERα -positive patients (Fig. 47). Expression data from these patients were averaged to the genes associated with each primary and metastatic specific enhancer group. Strikingly, the 2 set of enhancers resulted in having significant divergent results. Primary specific enhancers correlate with better survival outcome compared to the metastatic enhancers. Also, enhancers differentially clonal in metastatic patients correlate with poorer prognosis.

The interrogation of pathways databases provided further functional insights into the analysis. The 2 set of enhancers were associated with largely different pathways enrichment data. Primary specific enhancers enriched for vascularisation and proliferation while metastatic specific enhancers supported pathways commonly raising in the latest stages of the disease such as "xenograft TAM-resistant" and "up in acquired endocrine therapy resistance (ETR)" (Fig.48).



OS	Pval	HR	Pval	HR
PE	1.5E-05	0.64	0.047	1.21
Lymph	3.8E-07	1.92	7.3E-07	1.95
Grade	0.01	1.22	1.5E-03	1.28
Size	8.4E-07	1.02	6.0E-07	1.02

Figure 45. Survival analysis of primary specific (PE) and metastatic specific (ME) enhancers sets.

Kaplan–Meier (KM) analysis using 1,427 ERα-positive patients and averaged RNA expression of genes associated with PE or ME regulatory regions. CI for PE: 0.39–0.61. CI for ME: 1.1–1.67. Comparison of survival curves was performed using a log-rank (Mantel–Cox) test. Genes were assigned considering CTCF insulated perimeters. Multivariate correction for the comparisons is also shown. P, primary; M, metastatic. [Figure produced by Luca Magnani]

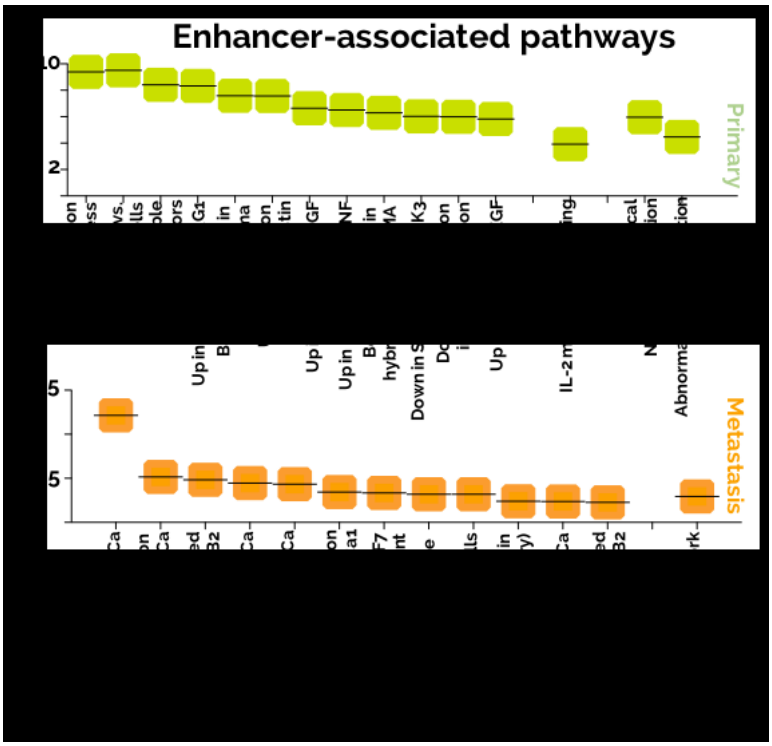


Figure 46.Pathway analysis for gene association to PE and ME.
Pathways were identified with GREAT tool and ranked by the significance. Y-axis: q-value. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

Collectively, this data add support to the hypothesis of a phenotypical evolution of epigenetic traits during cancer progression and confirm the reliability of our stratification model of tumour heterogeneity. Endocrine treatment may shape the epigenome of cancer cells by potentially applying a selection to cells with a set of regulatory traits and fostering the activation of molecular mechanisms specific to drugs. Our ability to dissect the entire epigenome into two main groups, primary specific and metastatic specific enhancers, confers a useful tool to the scientific community, which may encourage more in-depth investigation on the role of regulatory regions dynamics in oncogenesis and metastasis.

2.7 Comprehensive assessment of ERα + Breast Cancer cistrome

Transcription factors are critical regulators of gene expression. They bind sequence-specific DNA motifs which are instructive in controlling regulatory processes and ultimately defining cell phenotype. To exert their function, TFs require direct access to the cognate DNA motifs. The regulatory information stored in enhancers and promoters' relays on the availability of motifs for TFs. Identifying the TFs and the cognate motifs which direct the activation of enhancers and promoters is a crucial step in understanding the mechanistic basis of their activity.

The mapping of H3K27ac marked regions in our dataset of 34 primaries, and 13 metastatic tissues enabled the identification of the whole set of ERα + Breast Cancer nucleosome-free regions (NFRs). NFRs are DNA sequences transiently occupied by TFs and other cofactors. NFR exposure is the result of nucleosome positioning. This phenomenon is highly variable across cell types and is determined by the fine tuning exerted by chromatin modifiers.

Histone ChIP-seq data are a powerful source for consistently determining the NFRs. Indeed, the output of Histone ChIP analysis is a broad spectrum of signals which cover the entire genome. In the case of H3K27ac, these signals cover the position of all the nucleosomes carrying the H3K27ac marker. These signals consist of peaks directly proportional to the number of reads matching each specific genomic region. Peaks may be broad up to several kb and contain patterns of summits and valleys. The summits of the peaks represent the position of the marked nucleosome while the valleys match the NFRs. The interpretation of ChIP-signals allows us to perform a genome-wide characterisation of NFRs profiles associated with the regulatory regions and consequently to predict the putative TFs at each site.

We inferred the TF landscape by performing an integrative analysis using our data and DNase-I signal (DHS) from 129 cell lines. This allowed precise and more accurate identification of NFRs at very high resolution. NFRs were considered the regions intersecting the peak valleys and the DHS patterns (Fig.49). Quality control work established DHS regions from cell lines were consistent with our

definition of NFR from H3K27ac ChIP-seq. Furthermore, we compared the master NFRs list generated with the HoneyBadger dataset from ENCODE (Fig.50) available at:(http://www.broadinstitute.org/~meuleman/reg2map/HoneyBadger2).

This comparative analysis resulted in the overlap between the two datasets for the vast majority of the site and furthermore, showed that the strategy undertook by downloading all the available cell line accessibility assays comprise 20% more putative DHS loci.

We then applied TF analysis separately to NFRs specific to enhancers and promoters and obtained two distinct populations of significantly enriched TFs, concordant with the expected enhancer/promoter bias. Enhancers were enriched of common BC TF activators like the pioneer factors FOXA1 and GATA3 or well-known triggers of differentiation BACH2, FOS, FOSL2[108]. Promoter set was enriched of know TFs linked with BC as ISRE, JUND, PBX1 (Fig.51).

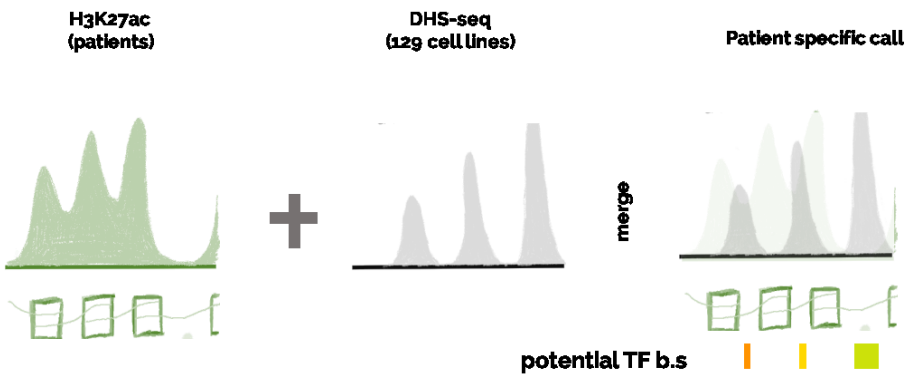


Figure 47. The computational framework of the NFR identification.

H3K27ac regions in patients are split at nucleosome levels to identifying the effective binding loci. DHS-seq were integrated into the patient NFRs and only the overlapping regions DHS-NFR were retained in the analysis. [Analysis performed by Giacomo Corleone. Figure produced in collaboration with Luca Magnani.]

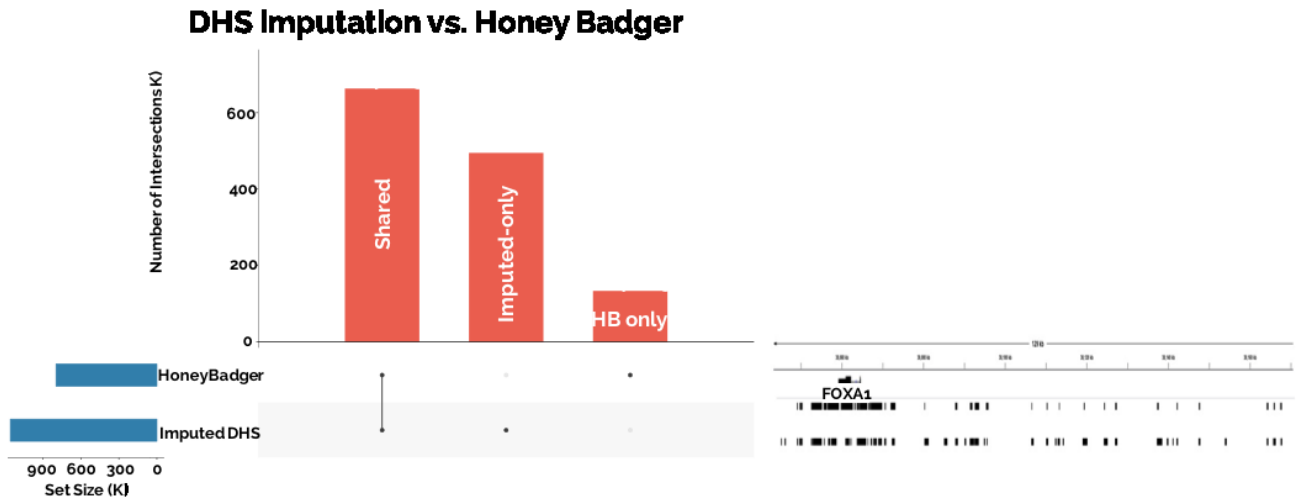


Figure 48 Overlap of DHS imputation with HoneyBadger dataset.

This figure illustrates the overlap between HoneyBadger and our imputed DHS. On the right the imputed DHS and HoneyBadger data among FOXA1 window. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

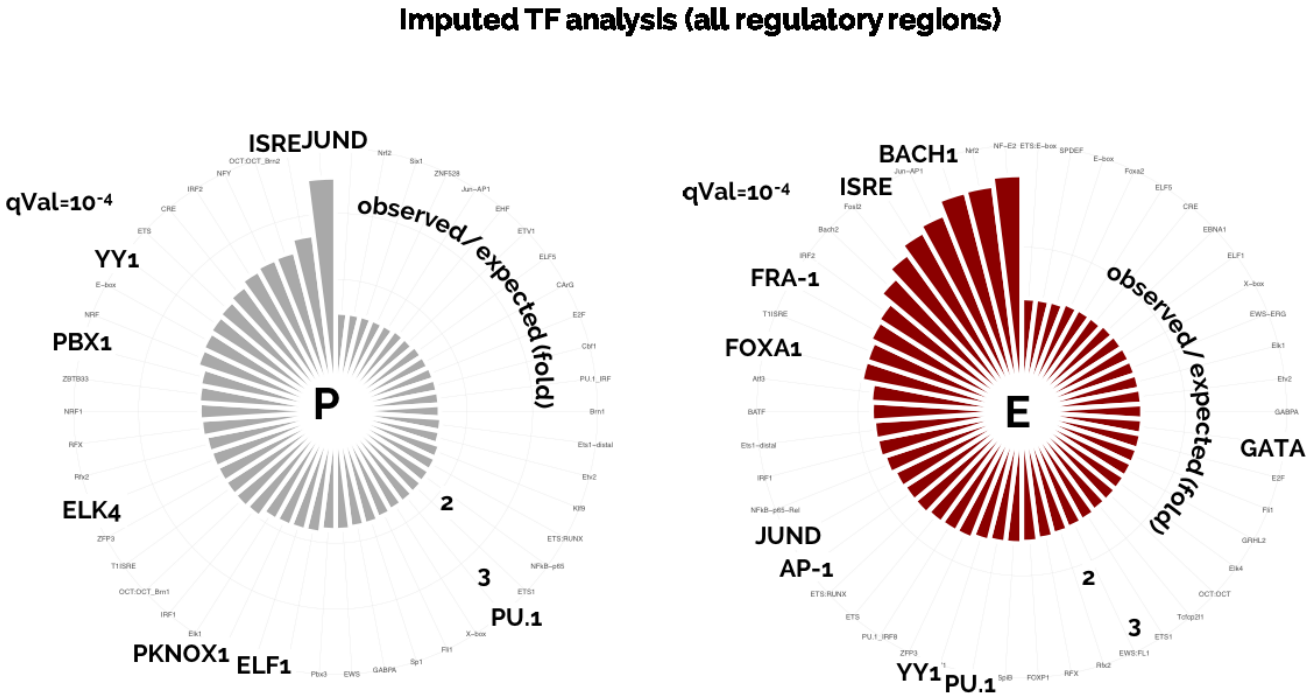


Figure 49.Imputed transcription factor analysis in the promoter and metastatic.

The 2 polar histograms show the results of the transcription factor analysis. Results show the different landscape of promoter (grey) and enhancers (red) cistrome. Length of the bars is proportional to the ration observed/expected of the motif discovery. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

We then explored whether any TFs were associated with the clonality of the NFRs. This analysis was performed by firstly associating the SIs to each NFRs and then running the motif enrichment analysis specific for promoter and enhancers. Interestingly, this analysis produced 2 clusters matching high SI (20<SI>40) and low SI (1<SI>19) (Fig.52). This may suggest a difference in the information stored in clonal enhancers and promoters compared to the sub-clonal regions. Clonal promoters were highly enriched of ZNF143, JUND, YY1, while non-clonal promoters did not exhibit any significant TFs. This may imply clonal promoters regulates the BC specific transcriptome, and subclonal promoters regulate non-specific BC signalling. On the other side, the enhancer specific TFs population exhibited a different trend of data stratification. Sub-clonal enhancers were generally more enriched of significant TFs compared the clonal enhancers. Interestingly, footprint analysis (Fig.53) as a function of RI revealed clonal enhancer accumulates more footprints than the expected suggesting a higher residency of TF in their cognate motifs.

Interestingly, oestrogen receptor elements (ERE) which are the direct target of ERα and supposed to be highly BC specific were significantly enriched only in sub clonal enhancers. Thus, we investigated the binding sites of ERα in our dataset. As a result, ERα preferentially binds clonal enhancers but not specifically EREs elements (Fig.54). Also, a small fraction of ERα binding sites (n=384) is shared across primary and metastatic patients while the vast majority is patient specific.

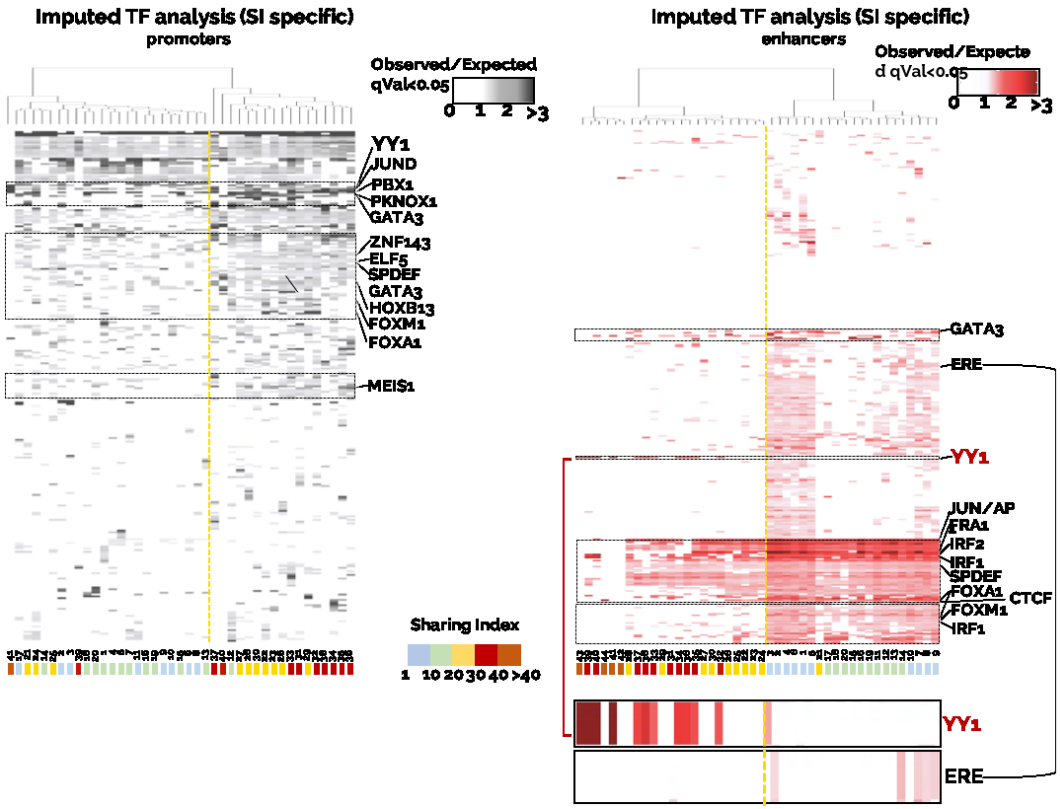


Figure 50. Transcription factor motif analysis heatmap applied to sharing index (SI) bins of promoter and enhancers.

Transcription factor motif analysis of individual bins (SI) followed by unsupervised clustering. This analysis called enriched motif for each bin separately in promoters (grey) and enhancers (red). Unsupervised clustering was then applied to bin groups contained similar motifs. High and low SI regulatory regions cluster separately in two distinct clades. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

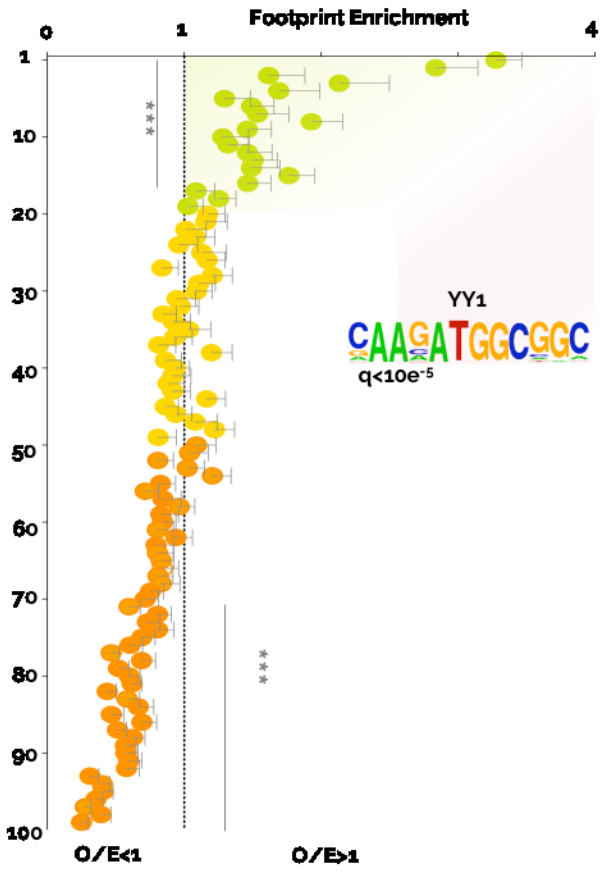


Figure 51. Footprint enrichment of enhancer in MCF7.

Clonal enhancers (green) RI<20 show a higher number of TF footprints. Subclonal enhancers (orange) have less footprint than expected. The number of footprints is normalised per number site. The observed/ expected ratio was computed by dividing the number of normalised footprints of each enhancer by the overall average. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

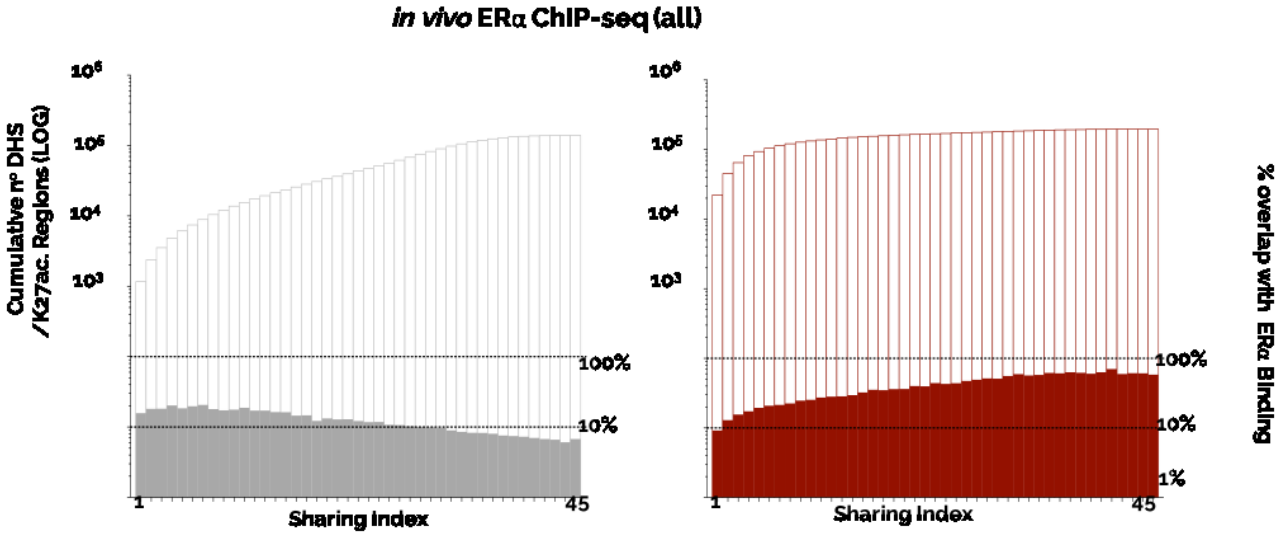


Figure 52. Overlap of imputed DHS regions with in vivo derived ERα binding sites.

Left=promoter, right: enhancers. Left y-axis: cumulative DHS regions. Right y-axes: percentage of overlap based on total DHS in each SI bin. ERa ChIP has been obtained from Ross-Innes et al., Cell 2011. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

Collectively, these data imply that ERα binding is highly heterogeneous in BC and each patient exhibit a specific ERα repertoire. The functional effect of this duality between a large number of ERα binding motifs and the sparse number of bound functional sites (EREs) may be explained by the potential capacity of ERα of scanning the genome for hunting co-factors such as pioneer factors which may stabilise the binding for a long-term residency and ultimately leading to a functional effect.

The imputed TFs analysis of the enhancer set highlighted the role of TFs such as FRA1, IRF1, IRF2, CTCF, and FOXA1. However, the examination of the high and low SI bias resulted in the identification of YY1 being the top TF significantly enriched in clonal enhancers compared the subclonal enhancers (Fig.55). These results were confirmed in MCF7 cell by performing RI assignment to the regulatory repertoire and footprint analysis.

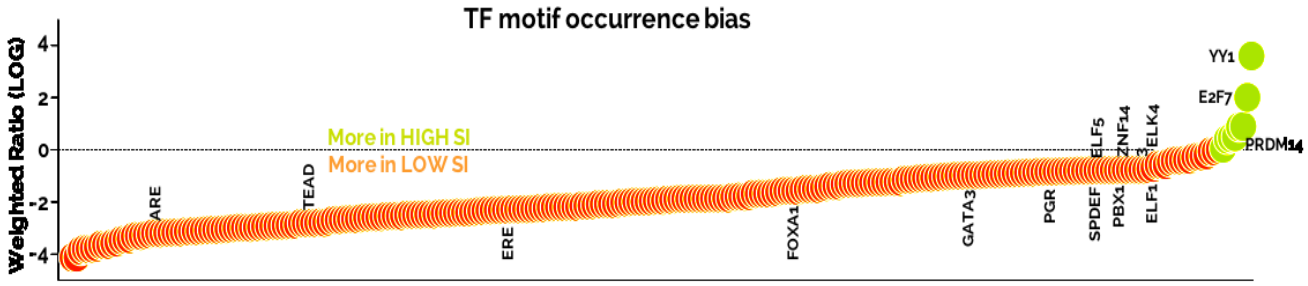


Figure 53. Putative TF rank by SI bin

Y-axis: the weighted ratio of the log(observed/expected) in high shared and not shared enhancers. TFs were ranked from the lowest score (left) to the highest (Right). TF more enriched in the high SI have been coloured green. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

Taken together, these observations suggest that promoter and enhancer in ERα BC exhibit a differential regulatory potential which reflects the clonality of each epigenetic region within the tissues. The information stored in promoters is more uniform than enhancers. Intriguingly, the large

majority of enhancers exhibit a spurious potential of TF binding while clonal enhancers are highly specific to certain TFs. In particular, YY1 TF is by far the most significantly detected element binding NFRs associated with clonal enhancers.

2.8 Pan-analysis of YY1 in normal and tumour tissues

Previous studies have assigned to YY1 the role of ubiquitous TF necessary in embryogenesis, tissue differentiation and physiological cellular processes with the dual role of activator and inhibitor[108,109]. Just recently, YY1 has been associated with the regulatory network of cells. Specifically, Weintraub et al. investigation proposes YY1 having a structural function in stabilising the looping between enhancers and promoters[110].

In this study, YY1 TF resulted in being associated with dominant enhancers clones in ER+ BC regulatory regions. We consequently investigated the expression of YY1 using independent public available tumour and healthy tissues dataset. Analysis from HPA[111], GTEx[112] and FANTOM5[112] dataset confirmed, as expected, YY1 is ubiquitously expressed in most of the tissues (Fig.56). YY1 exhibits higher expression in thyroid and parathyroid, lower expression in the hearth, pancreas while is stable in other tissues. The interrogation of YY1 expression in ERα-positive BC (n=1486) and normal mammary epithelial set (n=144) stored in METABRIC[113] dataset reported a significant YY1 increase of expression in tumour tissues (data not shown). These results were confirmed by interrogation of TCGA dataset.

Interestingly, YY1 overexpression was observed in ERα+ BC but not in basal-like BC. In addition, predicted survival in patients exhibiting higher YY1 expression was shorter compared YY1 low expression levels (Fig.57). We, then, experimentally corroborated these observations by performing immunohistochemistry (IHC) analysis in ERα+ ductal carcinomas, ERα- carcinomas and normal lobules and ducts, which strongly highlighted YY1 overexpression in ERα+ BC cells (Fig.58). In conclusion, YY1 positive cells are a dominant clonal population in ERα-positive BC invasive carcinoma.

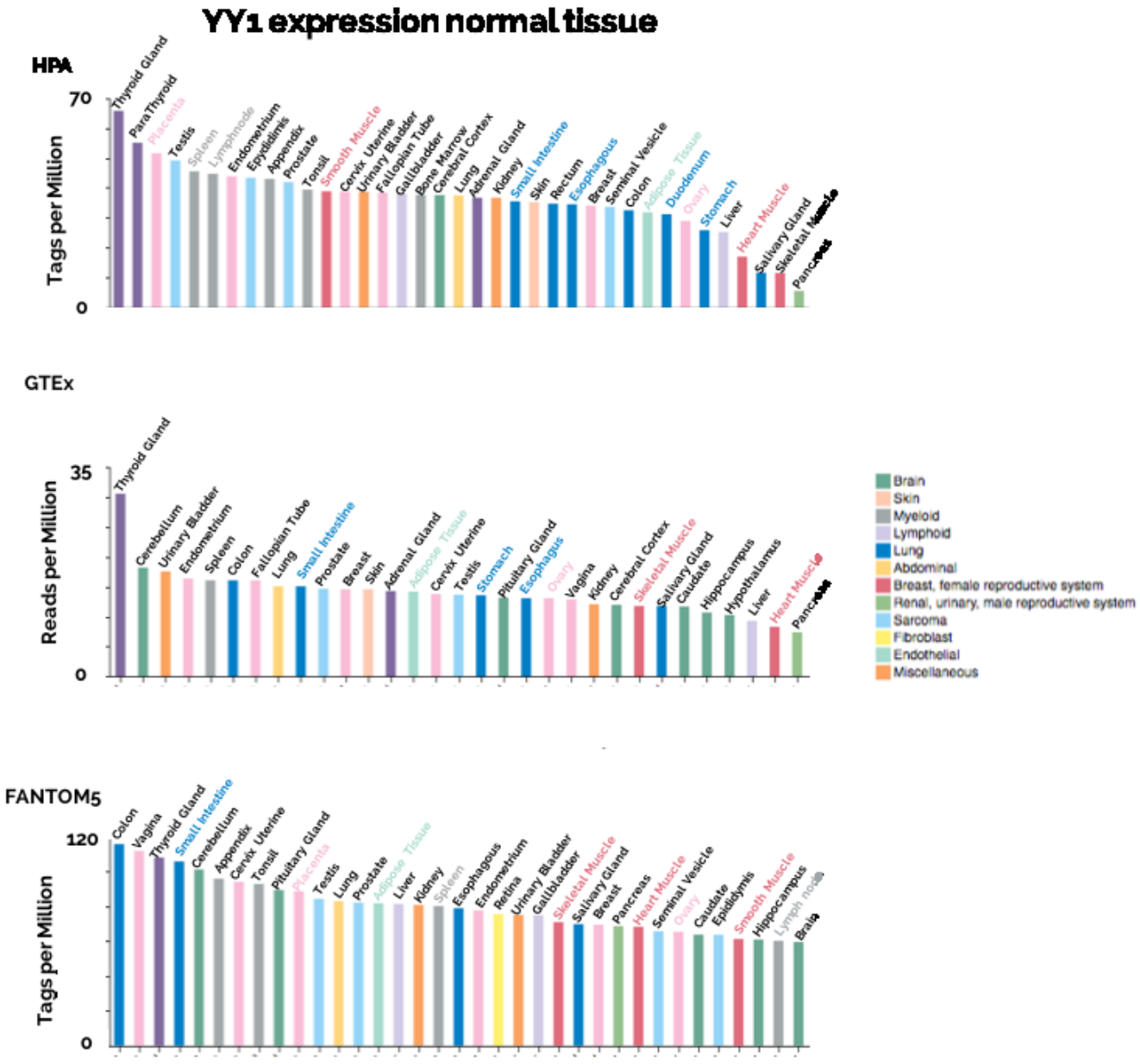


Figure 54. Pan-cancer analysis of YY1 expression in healthy tissues.

YY1 RNA levels from three independent RNA-seq datasets of normal tissues. [Figure produced in collaboration with Luca Magnani.]

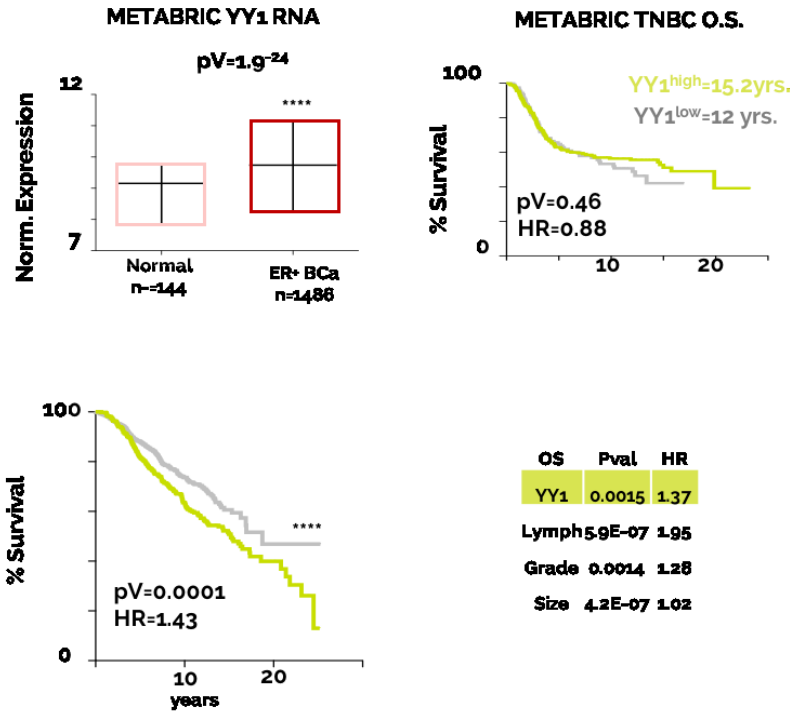


Figure 55.YY1 RNA-seq data and YY1 survival analysis.

Top-left=YY1 expression of ER+ERα BC and normal tissues (expression data obtained from METABRIC database). Top-right: Survival analysis of YY1 in triple negative BC. Bottom: Survival analysis of YY1 expression in ER+ERα BC accompanied by multivariate correction for luminal BC dataset. Comparison of survival curves performed using log-rank (Mantel-Cox). [Analysis and figure produced by Luca Magnani.]

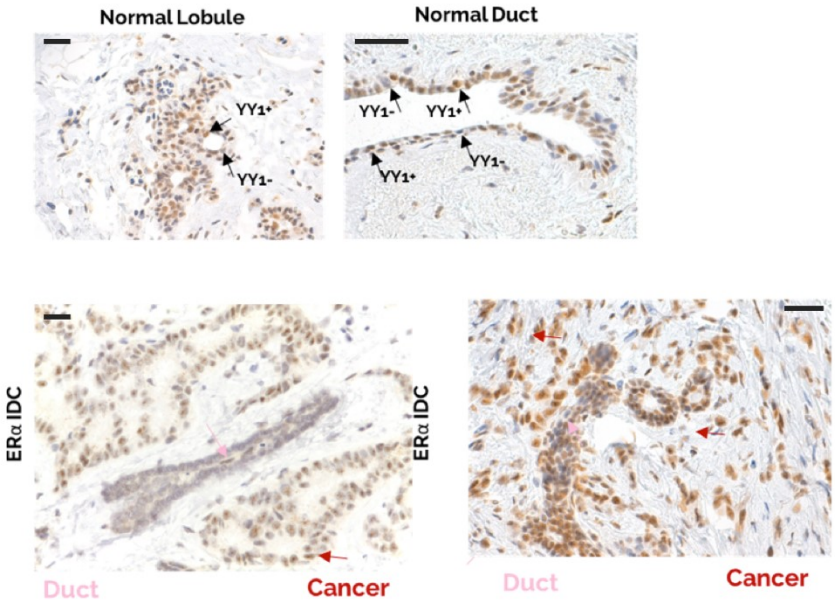


Figure 56. IHC analysis of YY1 in Normal lobule, Duct and ERα-positive invasive ductal carcinomas.

This figure illustrates YY1-positive clones in intra ductal carcinoma (bottom) and functional sub clones in normal breast. Scale bars: 50 µm. [Figure produced by Luca Magnani.]

2.9 Functional characterisation of YY1 in ERα+ BC

The global mapping of H3K27ac ChIP-seq revealed YY1 gene located downstream to 3 active enhancers (A, B, C). All three enhancers are associated with high clonality (low RI) and respectively shared amongst 41, 33 and 26 patients. Chromatin looping essays (ChIA-PET) available in ENCODE report evidence of looping between A >C, A>B and between enhancers A, B, C with YY1 promoter (Fig.59)

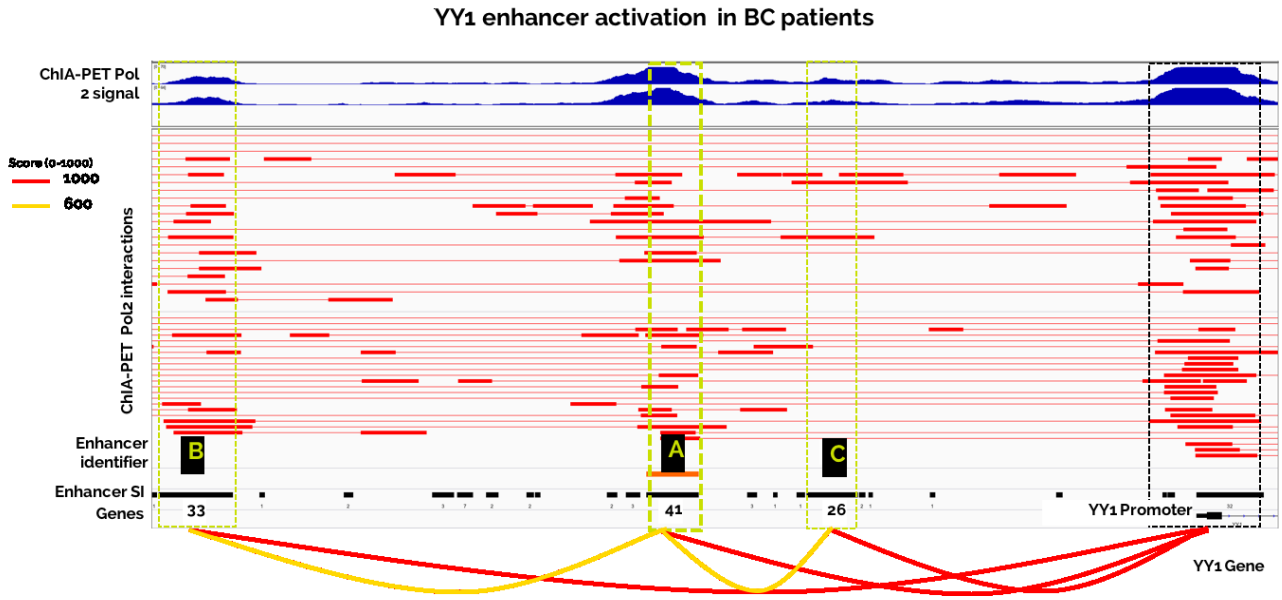


Figure 57.Regulatory landscape of YY1 locus.

Chromatin interaction analysis by paired-end tag sequencing (CHIA-PET) and RNA polymerase 2 (Pol 2) signals (top) and with CHIA-PET Pol interactions at identified enhancers (A, B and C) for YY1 and their corresponding sharing index below. 3 enhancers loop with YY1 promoter (bottom left). Enhancer A possesses the highest sharing index (41) and is shown to interact with enhancers B and C as well as the YY1 promoter. [Figure produced in collaboration with Luca Magnani.]

We subsequently tested the regulatory potential of enhancer A (higher in SI and lower in RI compared to B and C) by performing a CRISPR-Cas9 mediated deletion in MCF7 cell lines (Fig.60). This essay produced cKO MCF7 exhibiting a reduction of 30% YY1 expression levels. This result implies that enhancer A activity directly supports YY1 gene expression and a cumulative reinforcing effect is obtained via enhancer B and C interaction (Fig.59).

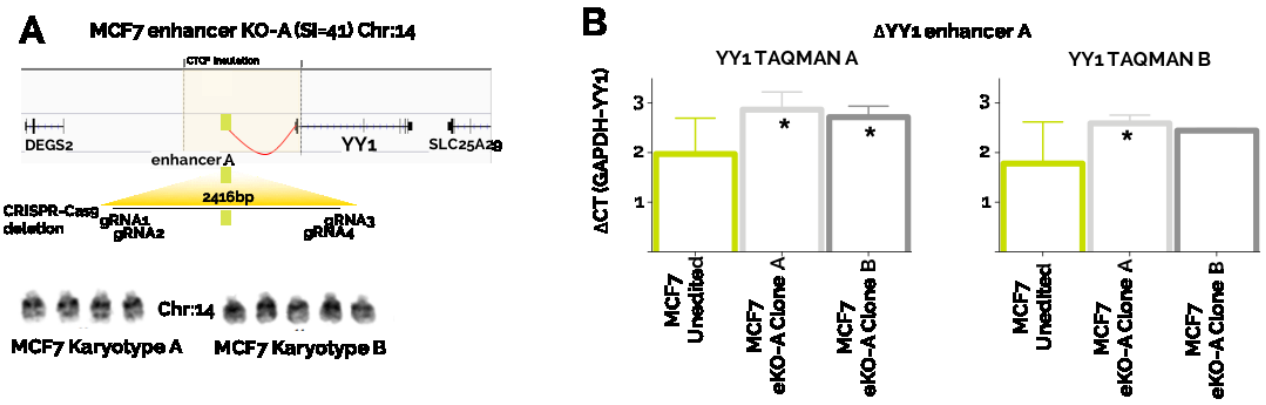


Figure 58.eKO YY1 experiment.

Left: MCF7 enhancer KO was generated by deleting 2.4kb matching YY1 enhancer A. The karyotyping in the bottom was performed from 10 individual cells. Right: YY1 expression in control and eKO cell lines using RT-qPCR. 5 independent experiments contributed to the results. Significance was calculated with a one-way ANOVA followed by Tukey's test. [Experiment performed by Neil Slaven, Figure produced in collaboration with Luca Magnani.]

RI assigned to regulatory regions within all individual patients included in the study confirmed enhancer A being a dominant clone with 80% of patients reporting RI < 20. To assess the hypothesis that enhancer A was clonal in other tissues, we performed a comprehensive assignment of RIs to all the available normal tissues from Epigenomic RoadMap dataset[113] (Fig. 61). This analysis showed that enhancer A is widely active in the ovary, placenta, digestive system and heart tissues while is consistently undetected in embryonic cell lines.

Collectively, our observation report that YY1 is a TF globally expressed in differentiated tissues and overexpressed in ERα+ BC. YY1 transcription is regulated by three enhancers which are dominant in our patient dataset. Suppression of enhancer A is sufficient to reduce proliferation in MCF7 cell lines significantly. In conclusion, YY1 is a crucial player of ERα+ BC development.

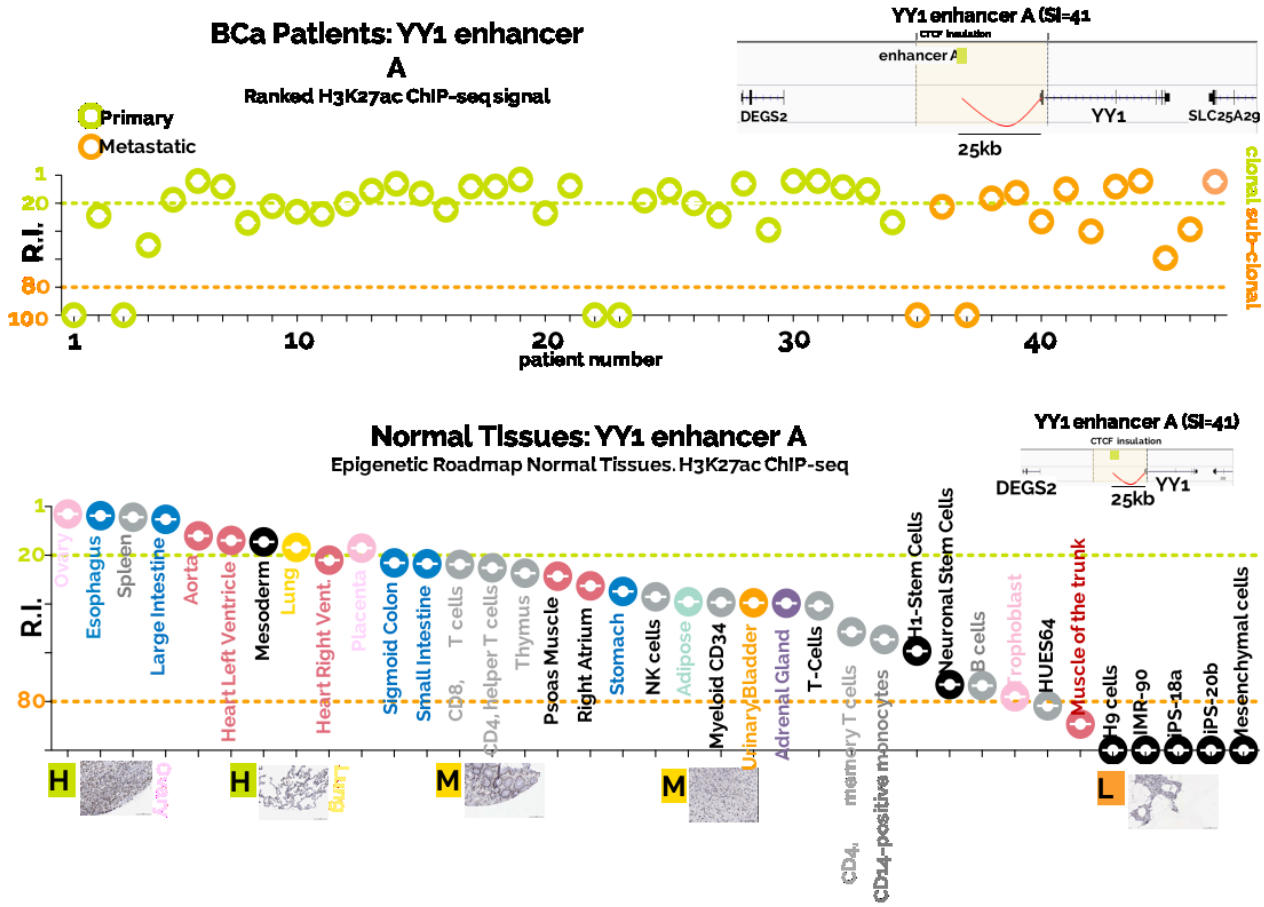


Figure 59. YY1 RI in the patient dataset and normal tissues.

Top: RIs for the YY1 enhancer within all individual patients included in the current study. YY1 enhancer A locations is shown in the top right. Bottom: YY1 enhancer ranking analysis of available Epigenome Roadmap H3K27ac data sets. Tissues are displayed from the strongest to weakest YY1 enhancer A activity based on RI. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone in collaboration with Luca Magnani.]

2.10 Profiling of YY1 binding and its functional effect in ERα+ Breast Cancer

The functional relationship between YY1 and ERα+ BC phenotypes was assessed by using reverse genetics and genome editing approaches in addition to the integration of H3K27ac global enrichment profiles at different conditions and YY1 ChIP-seq.

We investigated the binding profile of YY1 TF in MCF7 cell lines in oestrogen-stimulated and oestrogen-deprived culture (Fig. 62). YY1 TF was found binding 2223 sites in MCF7+ETOH (oestrogen deprived) while in 48029 sites under oestrogen stimulation. The increase of 23-fold of the binding repertoire was not expected and provided further evidence that YY1 activity in ERα+ BC,

but not in triple negative BC, is dependent by oestrogen. More interestingly, 81% of binding sites are occurring in enhancer regions, and 93% of YY1 binding sites in MCF7+ETOH are conserved upon stimulation with oestradiol. This suggests those regions may support the regulation of housekeeping genes or networks while the regions enriched upon oestradiol induction may support cancer specific regulation. We thus performed enrichment analysis to identify pathways associated with the YY1 binding sites in the two conditions. Not surprising, MCF7+ ETOH sites enrich for cell cycle, mRNA splicing and processing while ERα BC signatures were strongly emerging (q-value 10^{-260}) after oestradiol induction.

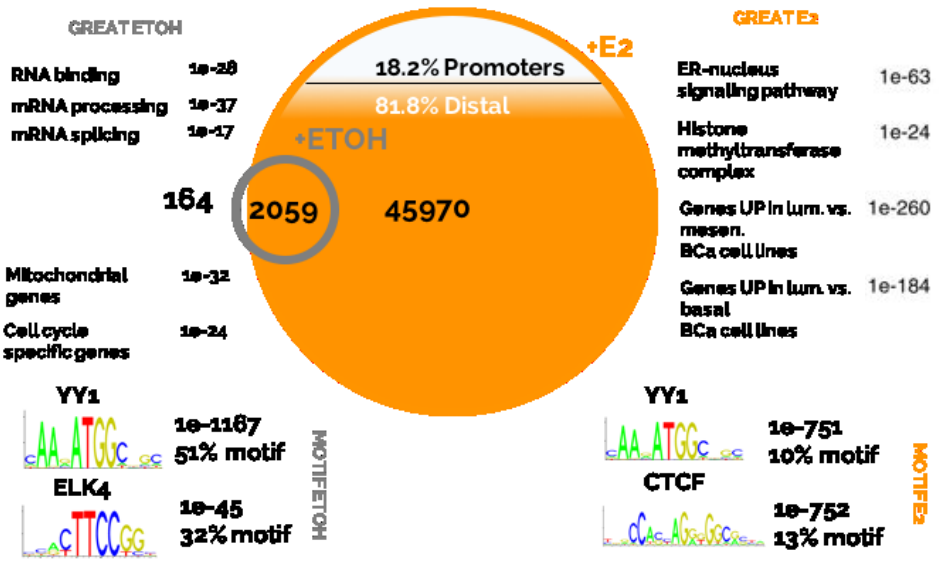


Figure 60. YY1 ChIP-seq in MCF7 upon oestradiol induction.

The figure shows the number of YY1 binding sites before (grey) and upon (orange) oestradiol induction with the relative enrichment analysis on the right. [ChIP-seq performed by DK. Patten. Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone in collaboration with Luca Magnani.]

Comparative ChIP-seq analysis between YY1+E2, ERα E2 and FOXA1+E2 show only a marginal overlap, however the binding of YY1+E2 and the top shared ERα binding motifs (n=484, ERα core) reaches 62% and nevertheless it is strongly associated with genes involved in TCGA luminal ERα+ BC signature suggesting a functional cooperative role of YY1 and ERα in controlling ERα specific depending pathways in cancer cells (Fig. 63).

Thus, we examined the association of YY1 and ERα performing in vitro studies which showed that YY1 deletion abrogates the proliferation of MCF7, ZR75 and T47D cell lines and the expression of ERα associated genes such as TFF1 and GREB1 with and without E2 administration (Fig.64).

Taken together, these observations strongly support the hypothesis of multiple roles played by YY1. YY1 TF is essential to cell viability and has a key role in response to intracellular oestradiol. A subset of ERα binding sites highly shared among patients, putatively co-bind with YY1. Moreover, YY1 seems to be involved in the complex scenario of acquired drug resistance, which commonly arises in the late onset of ERα BC. In conclusion, these novel observations strongly imply that YY1 acts as a functional ERα binding modulator during the different stages of the disease and certainly a pivotal factor of BC pathogenesis.

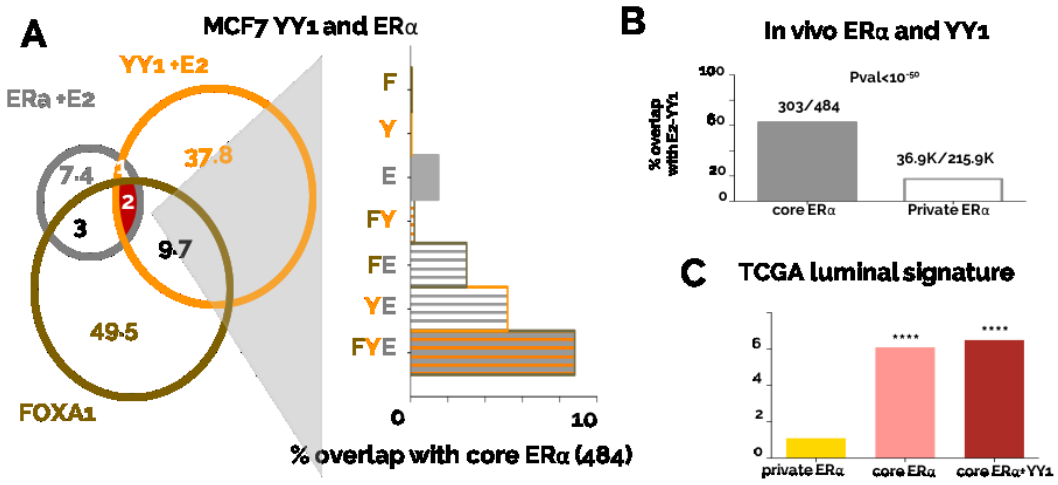


Figure 61. Comparative ChIP-seq analysis between YY1+E2, ERα E2 and FOXA1+E2.

YY1 ChIP-seq were produced in house while the FOXA1 and Era have been published by Ross-Innes, Nature 2012. A: overlap between ChIP-seq analysis of YY1+E2 (Y), ERα+E2 (E), FOXA1(F) in MCF7. On right potential overlap with ERα binding sites. B: ERα core binding sites strongly enrich for YY1 binding in MCF7 cells. Patient-specific ERα bindings are generally YY1-free. Proportions were compared using Fisher's exact test. C: Genes generally used to classify ERα+ BC patients strongly enrich for ERα-YY1 binding sites. P < 1 × 10−5 in Fisher's exact test versus private ERα. [ChIP-seq performed by DK. Patten. Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone in collaboration with Luca Magnani.]

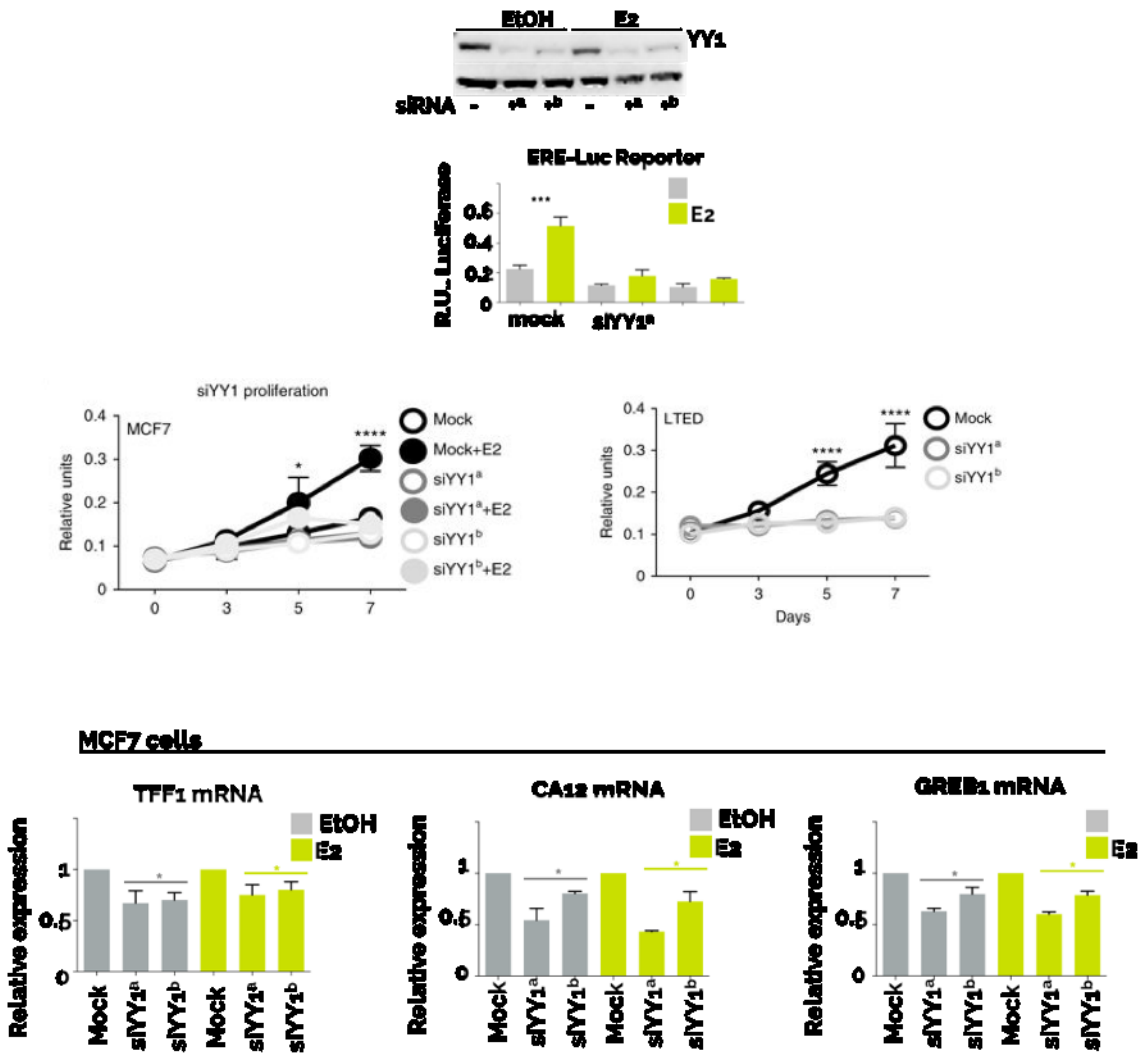


Figure 62. siYY1 and proliferation essay in MCF7 and LTED

Top: YY1 depletion leads to down-regulation confirmed by ER-driven luciferase reporter. Error bars represent the data obtained from4 independent experiments. Middle: Same results obtained from LTED and MCF7 and. Significance calculated two-way ANOVA with Bonferroni correction. Bottom: YY1 depletion leads to reduced transcription of common ER target genes. Each experiment was performed in biological triplicates. Column and bars represent the average and SEM of all experiment. Asterisks represent significance at P<0.05 after one-way ANOVA. [Experimental procedures executed by DK. Patten. Analysis performed by Giacomo Corleone. Figure produced by Luca Magnani.]

2.11 Role of ERα, YY1 and SLC9A3R1 gene as potential players in ERα+ BC endocrine resistance

Data from in vivo ERα core sites and YY1 binding profile in LTED cell lines suggested a potential role of the ERα-YY1 complex in supporting therapy resistance mechanisms commonly occurring in 30% of ERα BC patients.

We thus examined the sites of putative ERα-YY1 co-bind in our set of 47 ERα+ BC tissues and the association with their neighbouring genes. Data from a previous study[114] reporting a screening of the genes not responsive to endocrine therapy in ERα BC were used in this analysis and led to the identification of an interesting subset of potential genes involved in drug resistance. These genes were ranked by their predictive power for ERα+ BC relapse. This workflow led us to identify the gene SLC9A3R1 as a potential driver of resistance associated with the ERα-YY1 binding (Fig. 65).

SLC9A3R1 gene encodes a Na/H exchanger previously associated with resistance. By examining its regulatory landscape, SLC9A3R1 activity is supported by three enhancers. Enhancer 1 and two are localised in the intron while enhancer three lies downstream the exon at the 3'. A comprehensive analysis of our ChIP-seq profile of ERα, FOXA1, DHS, H3K27ac, YY1 in presence and absence of oestradiol provided evidence supporting the hypothesis that SLC9A3R1 enhancer activity was dependent on ERα binding. Interestingly, SLC9A3R1 enhancer 1 and 2 are active in MCF7 and LTED cell lines and are enriched of ERα and YY1 binding at the same genomic location (Fig.66).

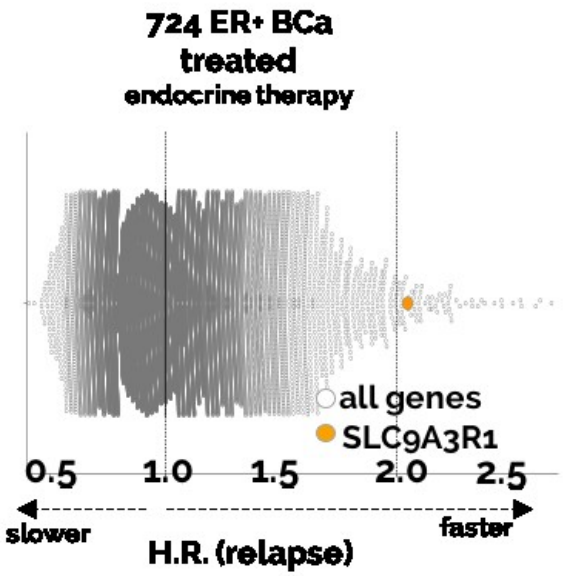


Figure 63.SLC9A3R1 is the top prognostic gene for ERα+ERα BC relapse.

Kaplan–Meier analysis summarises univariate analysis for 22,278 genes included in the Affymetrix microarray platform. X-axis: hazard ratio. [Analysis performed by Luca Magnani. Figure produced by Luca Magnani.]

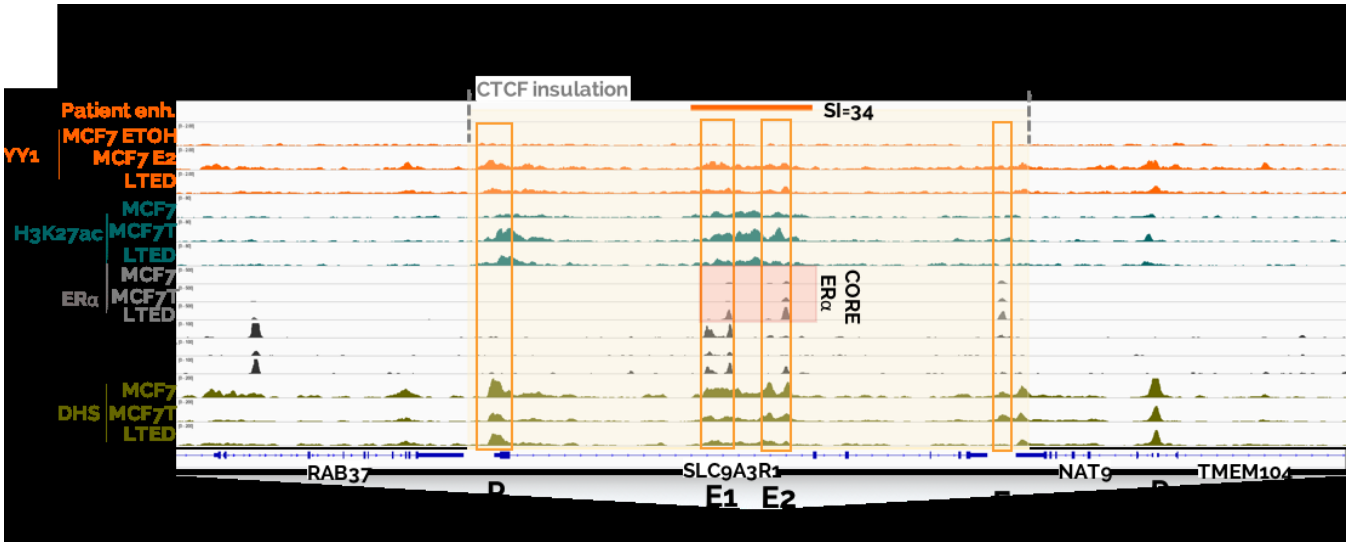


Figure 64. SLC9A3R1 regulatory landscape.

Chromatin landscape at the SLC9A3R1 enhancer locus in breast cancer cell lines with data integration of H3K27ac, ERα, FOXA1, DHS. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone in collaboration with Luca Magnani.]

Pan-cancer gene expression profiling significantly links SLC9A3R1 to higher activity in tumour samples compared to normal. Also, cancer cell lines data show that the ERα BC cell lines are the most associated with high expression (Fig.67).

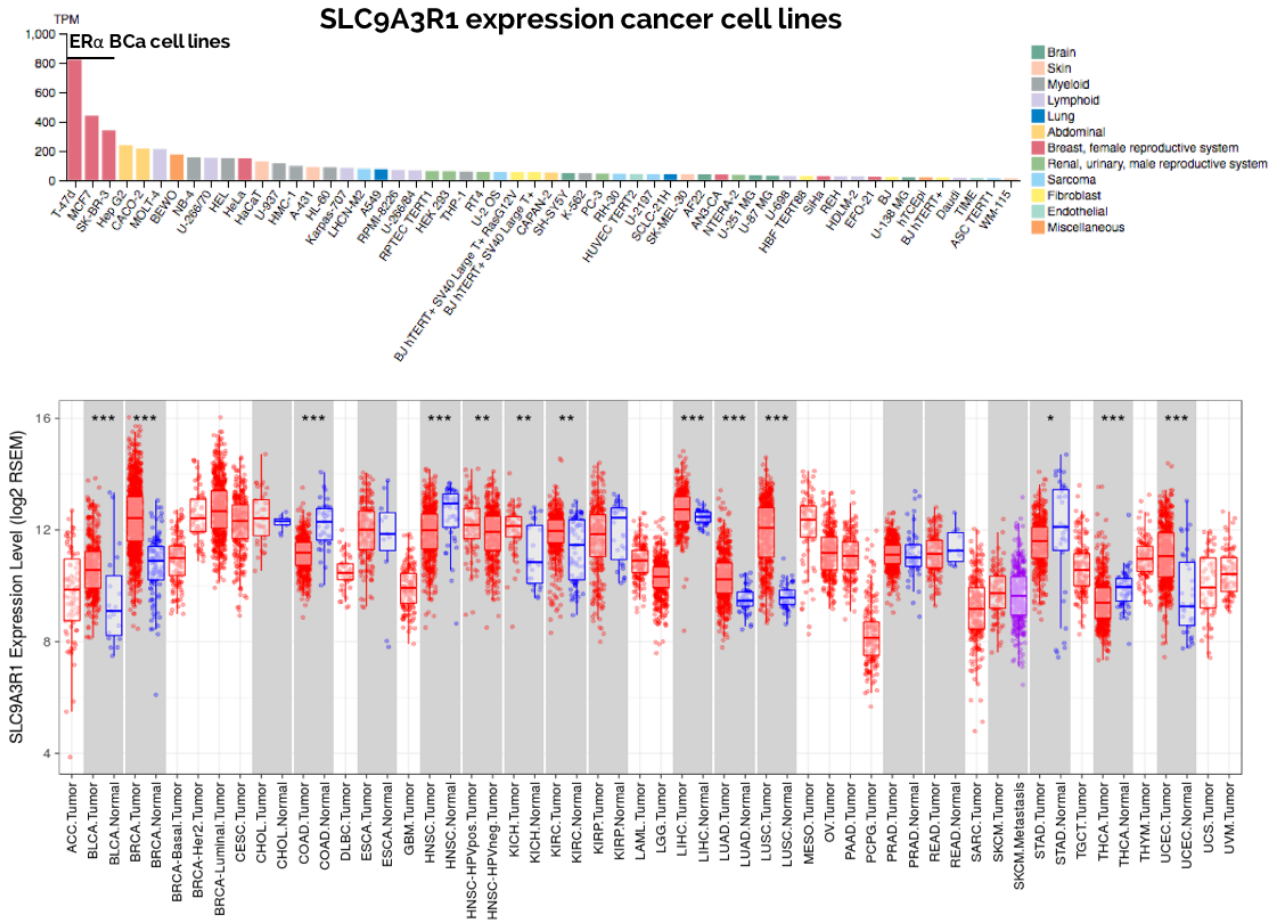


Figure 65. Pan-cancer analysis of SLC9A3R1 expression coupled with expression in cancer cell lines.

Top: Expression analysis of SLC9A3R1 in cancer cell lines. Data obtained from Protein Atlas archive. Bottom: SLC9A3R1 expression comparing normal tissues and cancer tissues from TCGA data. Figure generated using TIMER (https://cistrome.shinyapps.io/timer/) with the Diff. Exp. Module. Lines and boxplots represent median and interquartile ranges. [Analysis performed by Giacomo Corleone and Luca Magnani. Figure produced by Luca Magnani.]

Interestingly, in vitro analysis showed SLC9A3R1 depletion is sufficient to reduce proliferation in MCF7, and LTED cell lines (Fig.68 A, B). Also, the SLC9A3R1 expression is maintained by YY1 (Fig.68 C). These results link YY1 and SLC9A3R1 to a fundamental role in sustaining ERα BC pathogenesis.

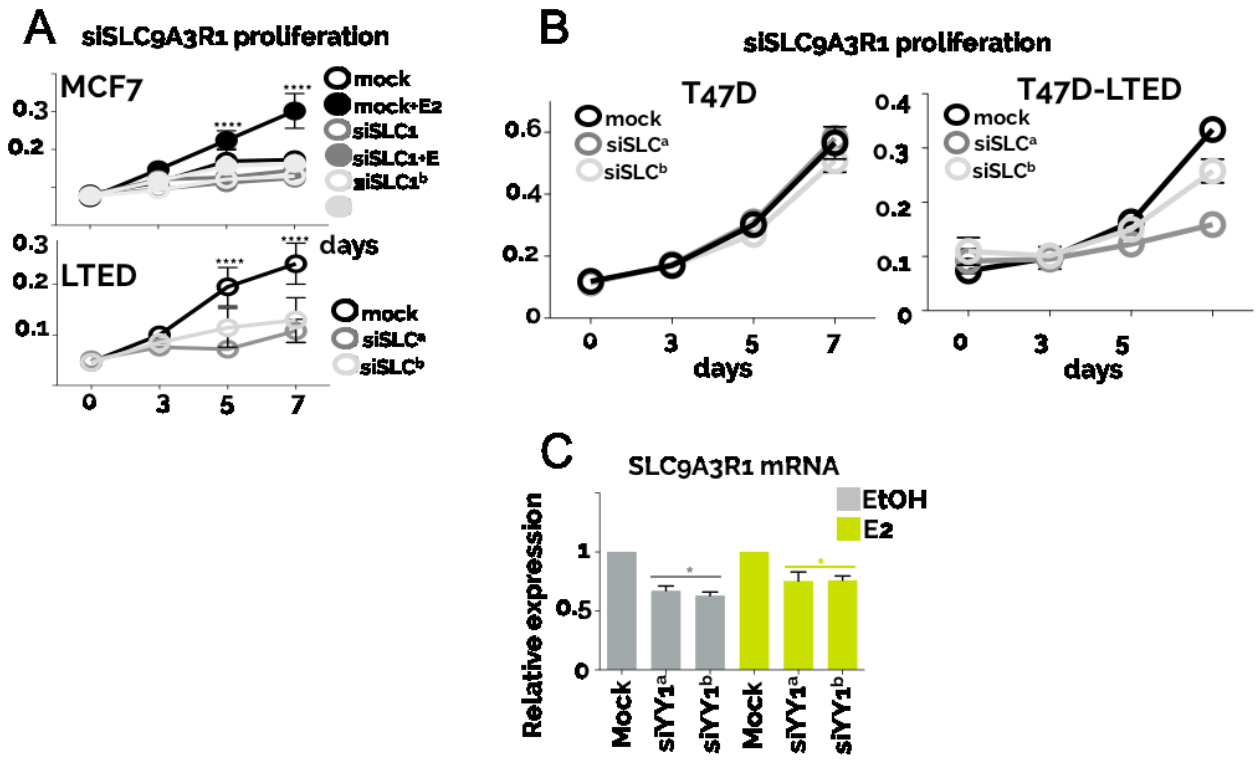


Figure 66. SLC9A3R1 silencing and proliferation essays

*A) silencing of SLC9A3R1 in MCF7 and LTED with and without E2. Proliferation essay conducted in triplicate. Symbols and error bars indicate average and 95% CIs. ****P < 0.0001; two-way ANOVA with Bonferroni's correction. B) silencing of SLC9A3R1 in T47D and T47-LTED with and without E2. Proliferation essay conducted in biological triplicate. Symbols and error bars indicate average and 95% CIs. ****P < 0.0001; two-way ANOVA with Bonferroni's correction. C)YY1 silencing decreases SLC9A3R1 expression. Each experiment was performed in biological triplicates. Column and bars represent the average and SEM of all experiment. Asterisks represent significance at P <0.05 after ANOVA test. [Experimental procedures executed by DK. Patten. Figure produced by Luca Magnani.]*

We further assessed the role of SLC9A3R1 enhancer 1 and 2 by evaluating its role in supporting the activity of other genes within 300 kb. Expression analysis from TCGA panel of BC strongly highlighted that SLC9A3R1 is the only gene in the genomic window being significantly higher expressed in cancer (Fig.69 top). Furthermore, the meta-analysis of enhancer 1 activity in the available cell lines from ENCODE, supports the central role in MCF7 cell lines. Indeed, this analysis reports MCF7 the only cell lines having enhancer 1 as a dominant phenotypical clone (Fig.69 bottom).

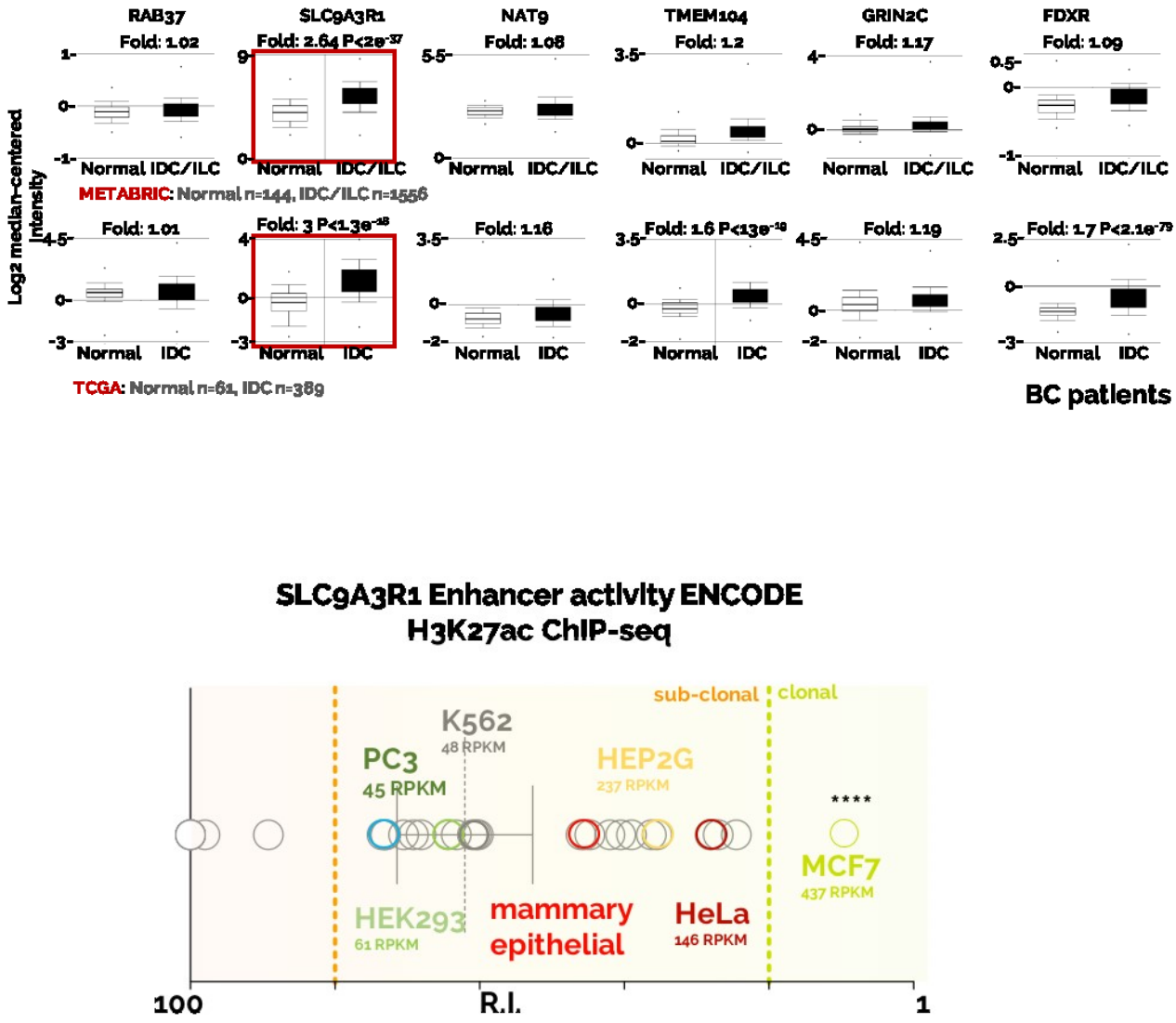


Figure 67.SLC9A3R1 expression analysis in Metabric and TCGA and relative enhancer activity in ENCODE.

Top: Gene expression of the closest genes to SLC9A3R1 enhancer. Y-axis=Log2 Median centred intensity. Each box represents normal and intraductal cancer. Data obtained from METABRIC and TCGA. Bottom: SLC9A3R1 enhancer RIs are shown for all available ENCODE cell lines. RI and relative RNA levels (RPKM) are shown when available for the same cell type. Asterisks represent significance at P<0.0001 after two-tailed Student's T-test. [Analysis performed by Giacomo Corleone and Luca Magnani. Figure produced by Luca Magnani.]

Collectively, these data strongly support a role of SLC9A3R1 in ERα-positive BC pathogenesis. Expression data demonstrate that despite being ERα dependent, SLC9A3R1 expression increases to endocrine therapy in ERα + BC patients. Here, we propose that SLC9A3R1 activity is supported by the activation of enhancer 1 and 2 which is driven by the preferential binding of YY1 and ERα at the same genomic loci.

2.12 Experimental validation of YY1 and SLC9A3R1 phenotypic clonality in ERα+BC

The last few paragraphs reported our findings regarding the new computational workflow based on the SI/RI, the role of YY1 and SLC9A3R1 in ERα+BC. SI-RI metric allowed the dissection of the complex epigenetic landscape of ERα+BC by assigning a score of enrichment and penetrance to each regulatory region in the sample dataset.

Data mining and functional studies highlighted the role of YY1 and SLC9A3R1. YY1 is a crucial player in ERα+BC oncogenesis while SLC9A3R1 is strongly associated with the arising of drug resistance mechanisms commonly developed in the latest stages of the disease.

To prove the efficacy and reliability of SI-RI method in identifying phenotypic clones and sub-clones, we performed immunohistochemistry (IHC) of YY1 and SLC9A3R1 in normal, primary and metastatic samples.

Exploratory analysis of ERα+ BC genome predicted YY1 sustained by the activity of 3 enhancers. Specifically, enhancer A resulted in having assigned RI associated with high clonality in almost the totality of the samples. Further investigations of enhancer A activity in all the available EPIGENOMIC ROADMAP tissues and cell lines showed enhancer A being sub clonal in the majority of tissue repertoire. These observations supported the hypothesis of YY1 TF being expressed in most of the duct and lobules cells under ER+BC condition.

We, thus, stained normal breast tissues and ERα+BC tissues for YY1. This essay demonstrated with high significance (12 different staining from independent patients) that YY1 production is much higher in cancer cells compared to normal (IHC slides not showed, data not showed). Further YY1 staining of the available primary and metastatic ERα+BC samples previously profiled for H3K27ac ChIP-seq (N primaries=10; N metastatic=2) resulted in being concordant to the assigned RI and ultimately demonstrated that large clonal YY1 positive population are present in both tissues category (data summarised in FIG.70, IHC slides not showed).

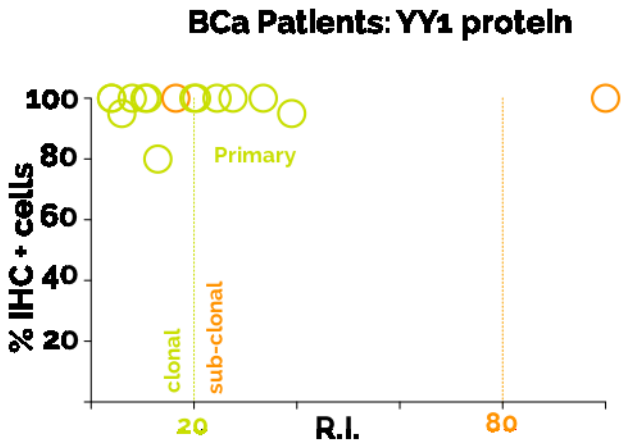


Figure 68. IHC staining of YY1 in primary and metastatic ERα+BC tissues.

X-axis: Predicted activity (RI) of YY1. Y-axis: the number of cells positively stained with YY1 IHC per samples. [Analysis performed by Giacomo Corleone. IHC performed by Giancarlo Pruneri, Figure produced by Luca Magnani.]

SLC9A3R1 gene expression is highly heterogeneous amongst healthy tissues and remarkably higher in different cancer tissues. The analysis of enhancers 1 and 2, which support SLC9A3R1 expression confirmed the variability of phenotypical clonality in ERα+BC tissues. Analogous to YY1, although the higher clonal variability, IHC staining (N=19) measurement of SLC9A3R1 in primary and metastatic tissues were directly proportional to the matching RIs derived from the previous H3K27ac ChIP-seq (R^2=0.74) (Fig. 71).

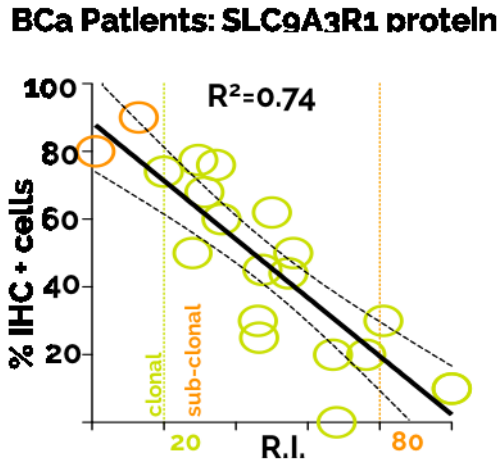


Figure 69. IHC staining of SLC9A3R1 in primary and metastatic ERα+BC tissues.

X-axis: Predicted activity (RI) of SLC9A3R1. Y-axis: the number of cells positively stained with YY1. [Analysis performed by Giacomo Corleone. IHC performed by Dr Pruneri, Figure produced by Luca Magnani.]

To ultimately demonstrate the relevance of our functional studies regarding the clonality of YY1 and SLC9A3R1 during ERα+BC cancer progression we carried IHC staining for both proteins in matched primary and metastatic samples.

Our previous results suggested that YY1 enhancer is not subjected to any phenotypical evolution during cancer progression while SLC9A3R1 positive clones expand from primary to metastatic tissues (Fig.72). YY1- and SLC9A3R1 were stained in an independent matched longitudinal cohort of 22 ERα BC patients (primary and metastatic) undergoing endocrine therapy with tamoxifen or aromatase inhibitors. This staining confirmed the large clonality of YY1 in both condition and an undeniable expansion of SLC9A3R1 clones in metastatic samples compared to primary. Remarkably, the SLC9A3R1 staining reports the protein contraction (from clonal to low abundance) in just one case which was explained by the loss of ERα positivity of the tissue (Fig. 73).

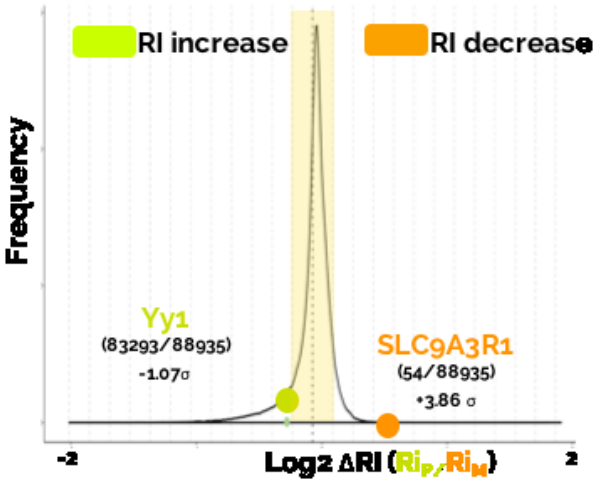


Figure 70.Phenotypical evolution of enhancers in primary and metastatic samples.

Phenotypic evolution of clones was defined based on the RI ratio of primary (P) and metastatic (M) samples (RIP/RIM). Y-axis: frequency of the ratio, X-axis: Log2 delta RI (RIm/RIp). Vertical bars represent 1σ(s.d.) increments from the population median. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone.]

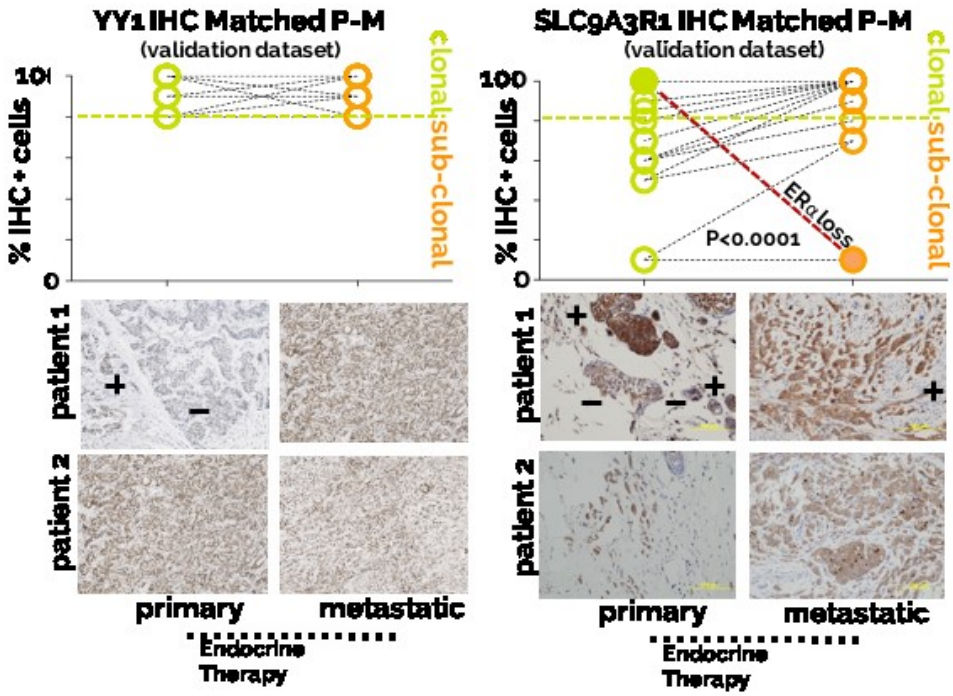


Figure 71: Staining of YY1 and SLC9A3R1 in an independent matched longitudinal cohort of 22 ERα+ ERα BC.

Scale bars, 100 μm. Statistical significance was calculated using a pair-wise, two-tailed t-test. Representative images are also shown. [IHC performed by Giancarlo Pruneri. Figure produced by Luca Magnani.]

Taken together these data ultimately demonstrated that enhancer activity of YY1 and SL9A3R1 predict protein expression. RI-SI metric is highly reliable in qualitatively assessing the clonality of regulatory regions which strongly affect the expression outcome of their targets. Ultimately, this integrative approach demonstrated a phenotypical evolution of a regulatory region during cancer progression and lays the foundation for further studies in the field.

Chapter 3

Non-coding genetic reprogramming during the development of endocrine resistance in ERα+BC

Aims of the study

The spatial organisation of chromosomes and the accessibility of transcription factors to DNA is designed by the cell epigenomic configuration. Epigenetic states are dynamical and respond to numerous endogenous and exogenous factors with the final effect of a new nucleosome organisation. Epigenetic states are heritable and persist through cell generations exerting a primary effect on cellular phenotype.

Results described in Chapter 2 of this dissertation revealed that modifications of chromatin openness at regulatory elements such as enhancers and promoters had been linked to the activation and inhibition of several regulatory pathways able to confer selective advantages to the drug-resistant ERα+BC subclones. This suggests that cancer progression is not only entirely directed by selection of pre-existing cells within an original tumour but also by phenotypical clones which become dominant during the progression of the disease.

Links between genetic variations and epigenomic architecture have recently been shown[115]. Mutations occurring among regulatory regions may have a functional effect which ultimately leads to a differential phenotype. Indeed, non-coding regulatory regions store information efficiently accessed by transcription factors. Interestingly, non-coding regulatory regions exhibit mutational burden similar to the coding[115], and the understanding of the associated functional role is relevant for cancer studies along with coding regions.

Although whole genome studies demonstrated that the non-coding DNA accounts around 98% of the entire genome and harbours 99% of the mutations, cancer genomic profiles have devoted more considerable effort to characterise the coding fraction efficiently. These studies revealed mutations

including single nucleotide variants (SNVs), small insertions and deletions that span from 2bp to 50bp in length, and extensive structural rearrangement depicted by inversions, duplications and translocations. The functional implications of these variants affecting the coding regions have mainly been addressed in the disease context. However, there is still a lack of understanding of the non-coding somatic mutations and their impact on diseases.

The primary limitations for studies aiming to investigate the role of non-coding somatic mutations (NCSMs) relate to the vastness, poor annotation, and the difficulties in addressing the functional role of these regions. Only recently biochemical evidence revealed[116] the regulatory nature of a large fraction of the non-coding genome which is still mostly unexplored.

Here, I report the way I approached the investigation of NCSM in ERα+BC context in primary and metastatic samples. I will describe two different strategies named Systematical Identification of Variants 1 and 2 (SID-V1 and SID-V2). I will discuss the main results obtained so far and the significant advances these approaches provide to ERα+ BC research.

The primary goal of this study is to define a strategy which systematically targets the investigation on the functional regions in the disease context (ERα+BC) and to identify candidate NCSMs involved in the development of drug resistance and metastatic progression coupled with the functional effects to the regulatory network.

3.1 SID-V1

It is believed that in cancer, mutations may occur at TF binding motif cores of epigenetically defined regulatory regions. These aberrations may induce a differential modulation both at TFs binding sites and epigenetic states configuration, strongly affecting the gene expression and ultimately cell phenotypes. The disruption of the binding site of essential proteins involved in the chromosomal organisation, such as CTCF, result in the disruptions of the physical boundaries of DNA architecture, potentially affecting the transcriptome of the cell. CTCF proteins display unusually high nucleotide conservation within its DNA binding domains among all vertebrates[83]. This high degree of sequence conservation supports an evolutionarily conserved function for CTCF that might be easily lost due to SNVs.

The dataset discussed in chapter 2 composed with 47 ER+BC tissue chipped for H3K27ac mark, were used to annotate the epigenome and leveraged for the identification of active enhancers to be used as targeted deep sequencing (Fig.74). 300k enhancers were identified, a number largely overwhelming the capture-sequencing assay (Agilent SureSelect XT) coverage availability accounting to 12MB. So, a further selection was applied to narrow down the target area for the capture sequencing essay. Mutations were called from H3K27ac patients' profiles at a depth ranging between 20x and 100x and known CTCF binding sites were selected by interrogating Hi-C maps and CTCF ChIP-Seq in ENCODE cell lines.

56866 potential NCSMs were identified among the H3K27ac profiles and capture probes 120bp long were designed to span genomic location targeted by the potential NCSMs. CTCF sites were selected from over 2400 TAD boundaries present in the genome. CTCF binding sites which were clustering within 20Kb from the potential boundary (resolution of Hi-C maps=40kb) were identified and then filtered based on conservation score in different cell types and high affinity to the optimal motif. 120bp long probes were designed to span genomic location containing the CTCF binding site (Fig.74).

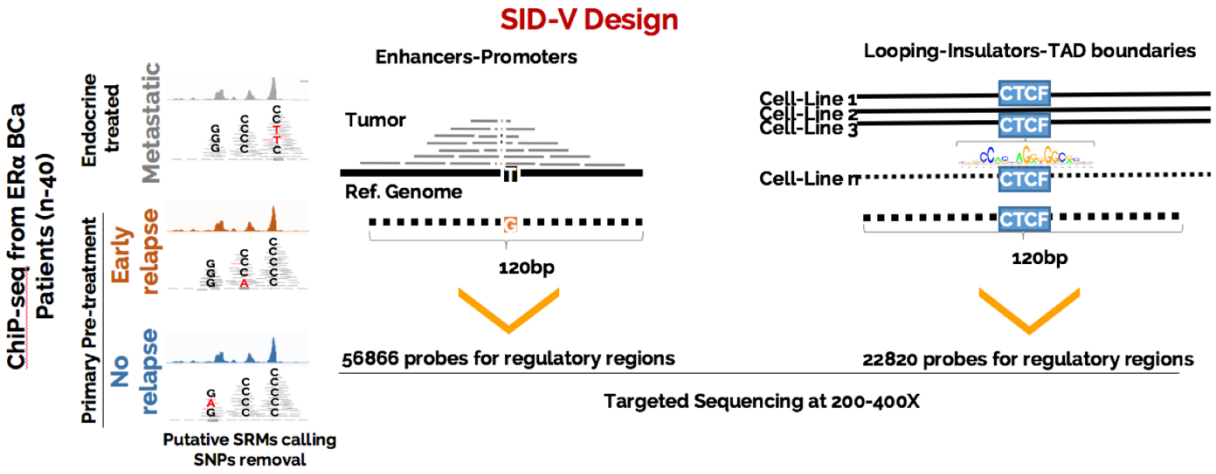


Figure 72 Theoretical workflow of SID-V1 design.

From Left of Right: We performed ChIP-seq in 40 ER-positive breast cancer patients using H3K27ac antibodies. H3K27ac positive regions were analysed, and we found 56866 potential NCSMs contained in the enhancer elements. Capture probes (120bp) were designed to span genomic location targeted by the potential NCSMs. Starting from over 2400 TAD boundaries present in the genome we looked for CTCF binding sites clustering within 20Kb from the potential boundary (resolution of Hi-C maps=40kb). We filtered CTCF binding sites for a) conservation in several cell types b) high affinity as based on the optimal motif. Probes (120bp) were designed to span genomic location containing the CTCF binding site. [Figure produced by Luca Magnani.]

Overall, SID-V1 is an application specifically developed to investigate at high resolution the genetic profile of regulatory regions in ERα+BC samples. SID-V1 relies on a targeted sequence capture essay built to survey 79686 genomic loci simultaneously, including potential NCSMs sites and CTCF binding sites at high coverage (400X).

3.2 Investigation of the mutational landscape in SID-V1 panel

SID-V1 has been applied for the investigation of a rare clinical case of primary bilateral ERα+BC in which deposits were removed from the same patient. The dataset comprises two biopsies from left breast (15-L1 and 15-L2), and one from right breast (15-R1). These are fresh frozen sections belonging to tumour mass excised in 2015 (Fig.75). Of note, the patient was diagnosed with contralateral primary breast cancer in 2013. The patient initially refused surgical treatment for a primary tumour, and then, she was administrated with AI (letrozole) treatment for two years. Initially, the patient presented an excellent response to the therapy reducing the tumour masses and stabilising the disease; however after one year of exposure to AI, she developed drug resistance. This new

condition resulted in a new disease progression and growth of primary cancers which obligated the surgical removal of the masses.

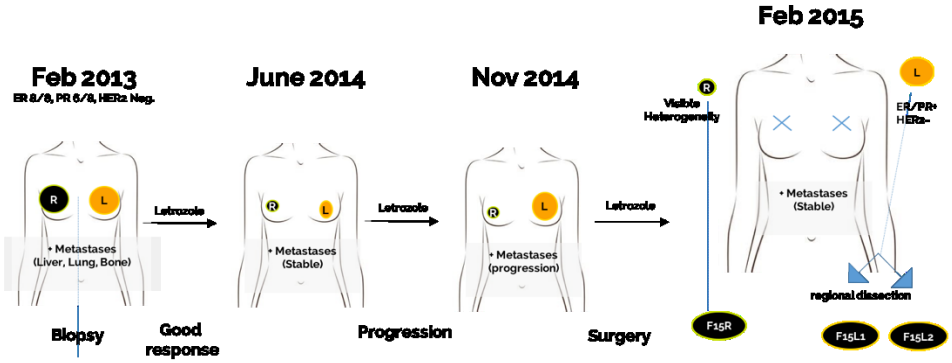


Figure 73. Golden patient case.

The patient was diagnosed with contralateral ERα+ BC and several metastases in 2013. She refused surgery, and she was treated to Letrozole with a good response. In 2014 tumours was reduced and metastases stabilised. Between June and November 2014, cancer progressed aggressively and then in February she underwent surgery. Clinical biopsies were taken at the surgery stage. [Figure produced by Luca Magnani.]

SID-V1 is a powerful application which allows the exploration of the mutational landscape within non-coding regulatory regions of ERα + BC. The variant calling of 15-L1/Blood, 15-L2/blood, 15-R1/Blood pairs were performed. The assay covered 79686 targets comprising 12 megabases (Mbps) of which 3 Mbps were covering CTCF binding sites and 9Mbps the selected enhancers. Variants calling in non-coding regions from high throughput sequencing in cancer samples is challenging[116] as non-coding regions include high-frequency repetitive DNA and cancer genomes contain a high fraction of structural variants.

We analysed SNVs profile of 15-L1, 15- L2 and 15- R1 samples against the normal blood. Mutational analysis of these samples reported a similar number of SNVs in the pair biopsies of the left and a significantly lower number from the mass on the right breast. 513 and 532 SNV events were identified respectively in 15-L1 and 15-L2 while only 15 in 15-R1. As expected due to the different location of the biopsies, different allele frequencies were found. The average allele frequencies of the mutational events in 15-L1 were considerably higher compared to 15-L2, respectively 31% and 22%. The average allele frequency of 15-R1 was 20%. Of note, all mutational events showing allele

frequency below 5% were removed from the analysis since they were prone to experimental and technical bias (Fig.76).

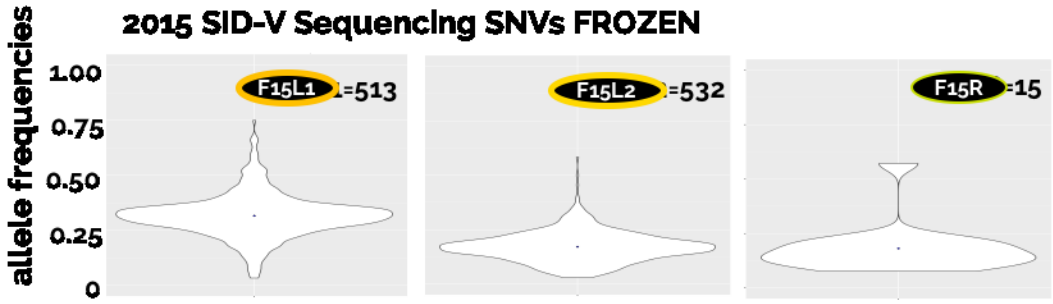


Figure 74. Mutation profile of L1 and L2.

Violin plot illustrating the number and allele frequency of the mutations identified from the sequencing. Y-axis shows allele frequency; X-axis shows population density. [Analysis performed by Giacomo Corleone. Figure produced Giacomo Corleone.]

We tested whether mutational events were shared amongst the samples. Interestingly, 80% of SNVs occurred both in 15-L1 and 15-L2 while none event with 15-R1, which exhibits only private mutations. We then tested the allele frequency of the shared occurrences in 15-L1 and 15-L2. 395 SNVs showed higher allele frequency in 15-L1 compared to 15-L2; only 13 exhibits the same allele frequency and only 3 increased from 15-L1 to 15-L2 (Fig.77).

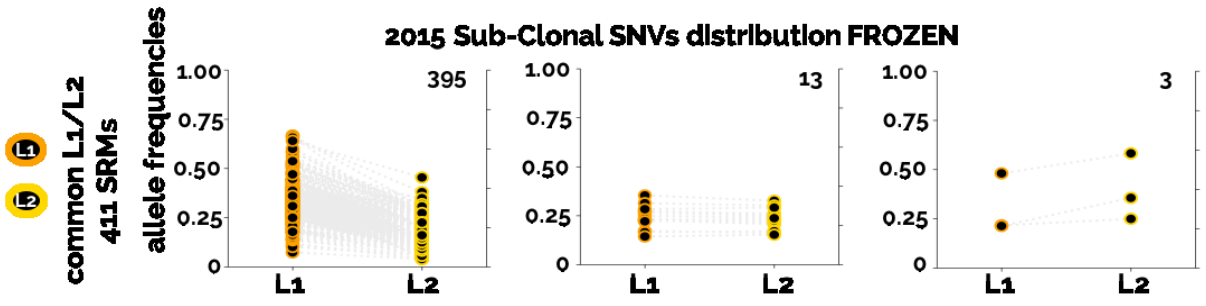


Figure 75. Subclonal distribution of SNVs between L1 and L2 samples. Of SNVs.

The dot-plot illustrates the mutational trend using allele frequency (Y-axis) in L1 and L2. Left-panel: mutations exhibiting higher clonality in L1 compared to L2. Central panel: mutations exhibiting similar between L1 and L2. Right panel: mutations showing higher allele frequency in L2 compared to L1.

The exploration of the genetic profile of the available tumour specimens highlighted the expected genetic heterogeneity of a tumour. Although the large majority of mutations obtained from the same tumour mass is shared amongst the two biopsies, the differential allele frequency suggests 1) different phenotypic subclones in a same tumour 2) probably evolutive events occurring in an expanding tumour mass. Also, the 15-R1 tumour does not exhibit any similarity with the one occurred on the left breast and its biological and functional role is still a matter of controversy. 15-R1 may be an entirely new different tumour or a metastasis generated from the left tumour. A broader evaluation of the genetic characteristics with the whole genome profiling is necessary.

3.3 Investigation of the CTCF mutational landscape in SID-V1 panel

We focused on SNVs mapping to CTCF binding regions with the hypothesis that NCSMs could alter gene expression by disrupting the binding sites of CTCF proteins. An exploratory analysis of different cancer publicly available data assessed the frequency of mutations amongst the window spanning -+60bps to the centre of the CTCF motif (Fig.78). Strikingly, the frequency of mutations occurring in the CTCF core (-+ 6bps) resulted in being higher on average compared the other genomic areas within the window.

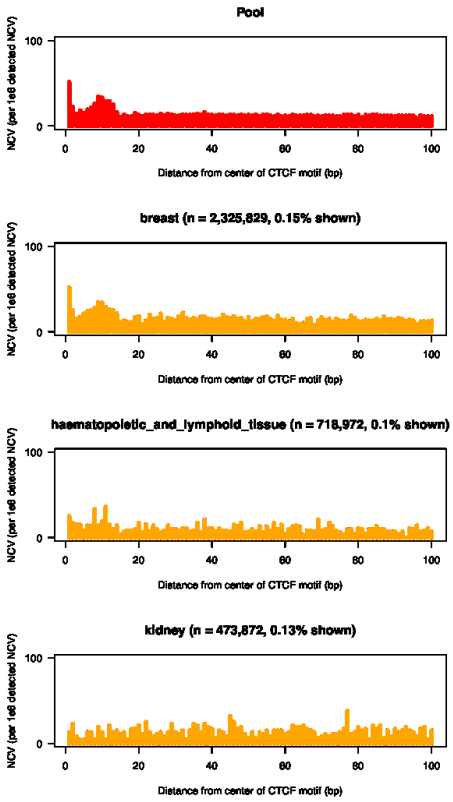


Figure 76 . Pan-cancer analysis investigating the mutational profile of CTCF binding sites.

The figure illustrates the number of normalised mutations according to a window of 100 bps to the centre of CTCF binding motifs in 3 different cancers. On top of the mutation profile from a pool of cancer. [Analysis performed by Iros Barozzi. Figure produced by Iros Barozzi.]

We thus reduced the region of interest to +- 6bps from the centre of the CTCF motif and observed a significant enrichment in the centre of the motif compared to the other nucleotides. We identified 15 Somatic Regulatory Mutations (SRMs) of which nine were occurring precisely on the central guanine of the binding motif (Fig. 79A) Then we computational predicted the effect of each variant in the local genomic context. The variants occurring in the +-6 bp window from the centre of the CTCF motif were significantly impacting the binding affinity of the CTCF protein (Fig.79 B).

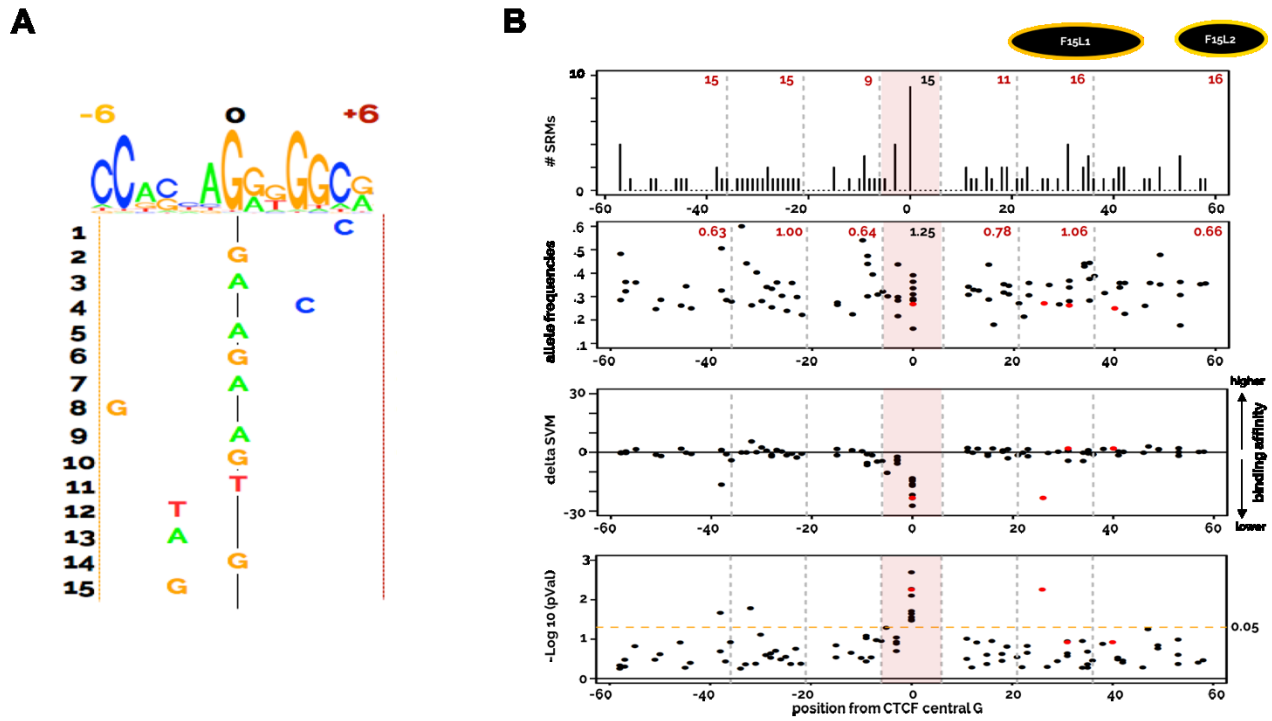


Figure 77 .CTCF window mutation profiling.

A: CTCF mutations in different cancers occur in the centre of the binding motif window size +- 6bps from the centre of the motif. Rows: SNV occurring at the site; Columns=motif position at 1bp resolution. B: Delta SVM analysis to measure the binding affinity variation at a resolution of 1 bp. Central red dot represents SRM5. [Analysis performed by Giacomo Corleone and Iros Barozzi. Figure produced by Giacomo Corleone and Iros Barozzi.]

It has been demonstrated[158] that 8% of the human genome is conserved among vertebrate and these regions are likely to be functional. The 15 SRMs effect was quantified by measuring the evolutionary score in the vertebrate placental context. 5 of the 15 SRMs were predicted to be regions of extremely high conservation in vertebrates and subsequently an optimal target for functional studies (Fig.80).

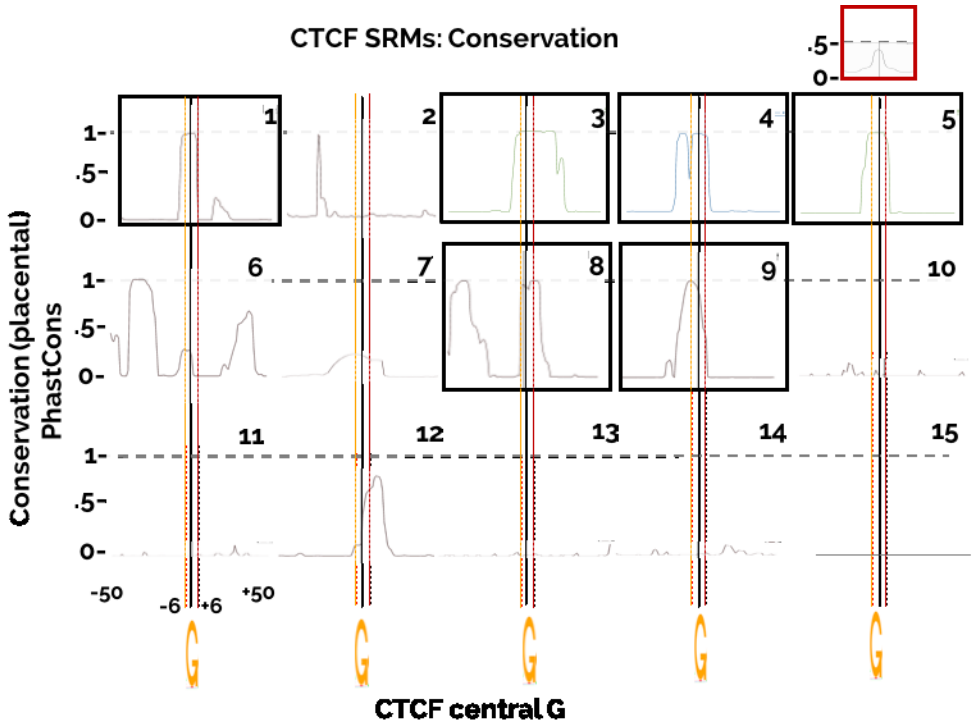


Figure 78. CTCF conservation amongst vertebrates.

Conservation scores amongst vertebrates at the centre of each identified SRM (N=15). 5 mutations hit highly conserved regions. [Analysis performed by Giacomo Corleone. Figure produced by Luca Magnani.]

Then, we explored the genomic context of the 5 SRMs by leveraging bona fide TAD data at +- 20kb of resolution from Hi-C data. SRM5 was occurring precisely in the bona fide TAD associated boundary in the q25.1 of the chromosome 17 (Fig.81).

An integrative analysis comprising RNA-seq, H3K27ac, DHS seq, ERα and FOXA1 in MCF7 cell lines supported the hypothesis that SRM5 lies on the border of a TAD containing eight genes (RPL38, DNAI2, GPRC5C, CD300C, RAB37, NAT9, SLC9A3R1, TMEM104) which exhibit a low level of functional enrichment (silent TAD). On the contrary, the TAD downstream SRM5 contains thorough highly active regions, and in addition, no physical interaction between regions in the 2 TADs is reported (Fig.81).

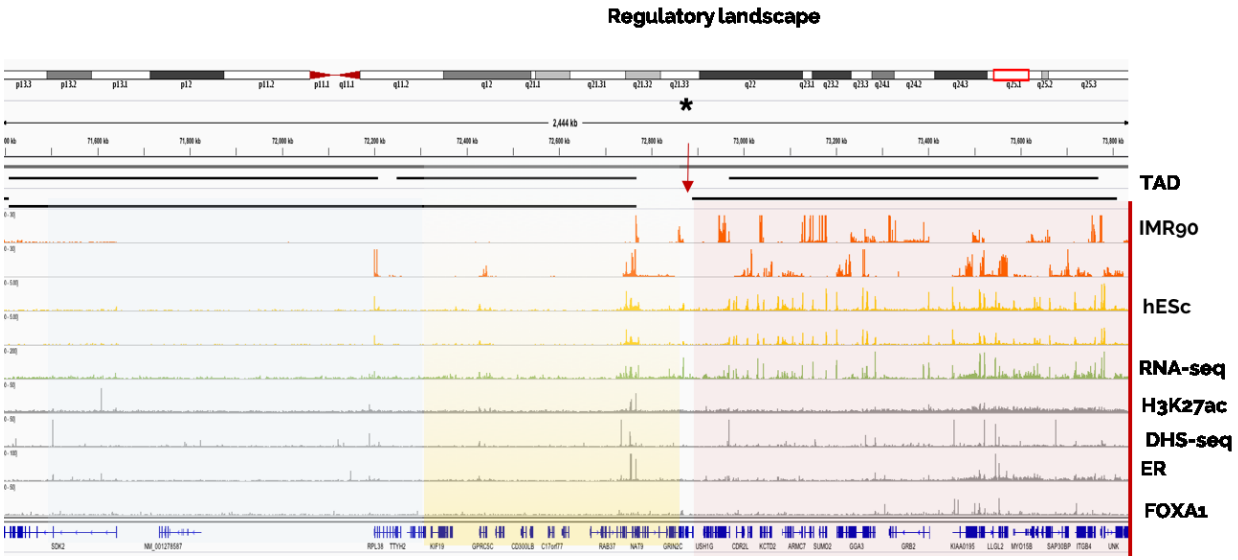


Figure 79. Regulatory landscape of SRM5.

Visualization of the SRM5 regulatory landscape and data integration which includes transcriptomic, DNA accessibility and ChIP seq of ERα, FOXA1. Shadow colours represent the 3 different TADs identified using Hi-C maps. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone and Luca Magnani]

Collectively, these data suggest that SNVs identified in the dataset commonly target critical region of the CTCF motif. CTCF are highly conserved regions across species and when binding to chromatin insulator sustain the structural configuration of the genome. Of 15 SRMs identified from our dataset, we have applied a computational procedure to narrow the search for targets more likely to have a dramatic functional impact to the gene expression network which may arise during cancer progression. This method highlighted SRM5 lying at a border of a silent TAD in MCF7.

3.4 Functional investigation of the Somatic Regulatory Mutation number 5

The results described in the preceding paragraphs suggested a potentially dramatic effect of mutations occurring in the highly conserved CTCF core motif laying in the TAD boundaries. Our computational workflow combined with the SID-V1 approach highlighted a potential role of SRM5 in driving gene deregulation in the neighbour genomic context. SRM5 lays in a critical position which theoretically may affect the expression of a dozen genes located within a window of 500 Kb. Regulatory mapping of the chr17 q25 genomic context suggested SRM5 holding at a border of a silent TAD. We hypothesised that the disruption of CTCF binding caused by SRM5 mutation might lead to a significant effect within the TAD context. Thus, we developed a CRISPR-Cas9 approach to engineer ERα-positive breast cancer cell lines. To test if SRM5 could effectively alter CTCF binding, we generated three individual clones containing unique indels centring at the CTCF locus. We successfully engineered three distinct clones, each with complete abrogation of CTCF binding at this site as demonstrated through ChIP-seq (Fig.82). Identification of the exact mutations proved difficult with one clone presenting three separate deletions led us to discover that in some cases MCF7 cells have three copies of chromosome 17. The mutations that were successfully identified consisted of 6-11 base pair deletions on both copies of the site. The engineered indels were not affecting the binding of the other 2 CTCFs lying in the same genomic context.

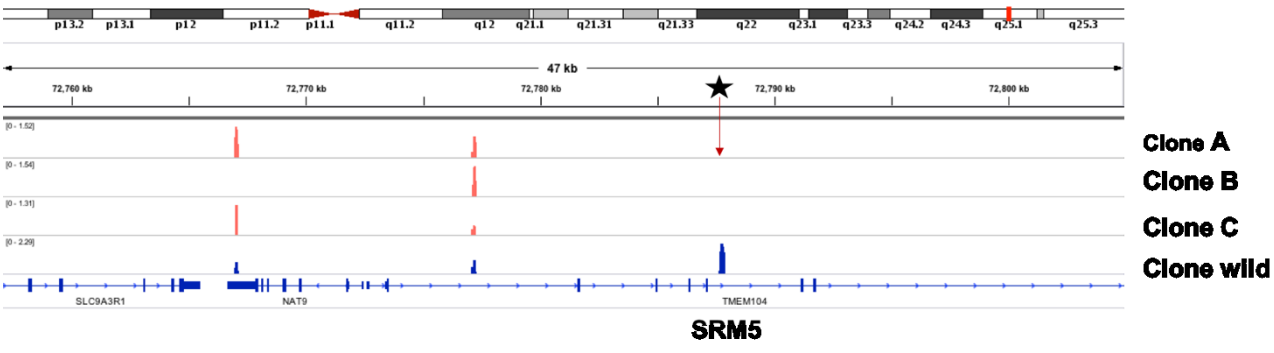


Figure 80 . CTCF-ChIP-seq of the three mutated clones

The crispr-cas9 approach was performed to engineer SRM5 loci. The figure illustrates CTCF binding in the engineered cell lines (clone A-B-C) and wild type. [Experimental procedure performed by Neil Slaven. Analysis performed Giacomo Corleone. Figure produced by Giacomo Corleone.]

We next evaluated the functional effect of the SRM5 in vitro by measuring the relative expression of a panel of genes occupying a window of 500kb upstream and downstream the SRM5. RT-qPCR measured transcriptional changes in 3 engineered clones (Clone A, Clone B and Clone C). Five biological replicates represented each clone (Fig.83).

Interestingly, each clone showed a unique transcriptional phenotype which may indicate a stochastic transcriptional effect induced by the loss of CTCF at the border of the TAD. Of particular interest was the result obtained in the clone A analysis which underpins the role of SLC9A3R1. Strikingly, SLC9A3R1 gene seems significantly affected by a potential disruption of the CTCF binding at the SRM5 loci. In Clone A its expression increases of 2.5-fold as average. This striking result, in combination on what described in chapter 2, further elevates the role of SLC9A3R1 in sustaining ERα+ BC progression.

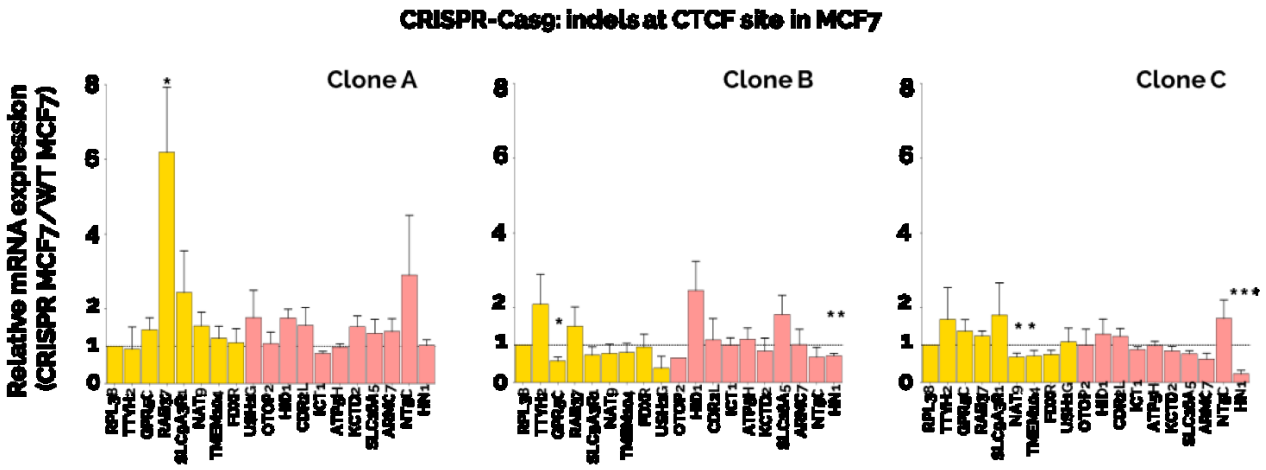


Figure 81. RT-qPCR of genes sitting in SRM5 genomic neighbourhood

Transcriptomic profile of gene potentially involved in CTCF activity at TAD boundary. Y-axis: Relative mRNA expression (CRISPR-Cas9 / WT MCF7) in the 3 different clones under examination. .[Experimental procedure performed by Neil Slaven. Figure produced by Luca Magnani.]

3.5 SID-V2

The epigenomic annotation of ERα+BC tissues described in chapter 2 coupled with the RI and SI was further leveraged to improve SID-V1 in a second version called SID-V2. As described in the previous paragraph the enhancer tagging of SID-V1 was performed by calling mutations from the ChIP-seq and the matching input file of each sample resulting in 56866 potential NCSMs. The capture probes were designed +- 60bps to each point mutation. The availability of an improved annotation of the ERα+BC regulatory regions encouraged to develop a more reliable and performing application covering the regulatory regions exceeding a multi-step decision process based on functional criteria.

The epigenomic annotation of ERα+BC tissues led to the identification 326,729 enhancers active in the ERα+BC context. The vast majority of the enhancers were private or poorly shared amongst the patient in the dataset. We speculated they were the result of cell composition heterogeneity or experimental artefacts. However, the RI analysis associated with each regulatory region demonstrated a proportional enrichment of enhancer activity in highly shared patients as the main trend. Thus, we focused our investigation on the enhancers having assigned high SI and low RI. We speculated that the association of high SI low RI implies functionality and subsequently demonstrated that SI/RI ratio was qualitatively associated with the average cell phenotypical clonality.

Based on these results, we investigated whether the regions covered in SID-V1 were predictive of any activity in vivo. We integrated the data described in chapter 2 with SID-V1 annotation and observed that only a small number of sites (14.6% (SID-V1s / RI<40 & SI>20) x100) covered in SID-V1 were showing high SI low RI. This result encouraged us to further improve the site selection of SID-V1 with a more accurate methodology which precisely targets active enhancers and CTCF significantly involved in ERα+ BC oncogenesis.

Tag selection of SID-V2 was performed by developing a multistep procedure in which priority to regulatory regions showing a significant evolutionary trend of activation during cancer progression. Starting from a total of 326K enhancers, we filtered all the private enhancer with assigned SI <2 and

enhancers not showing any particular enrichment of activity (RI>=80) (Fig. 84A). 19482 enhancers passed the filtering procedure and were leveraged for the evaluation of their activity in primary and metastatic samples as described in Chapter 2 Fig.72. We computed the average RI of each enhancer in primary and metastatic sample separately followed by the calculation of the log2(RI_Prim/RI_Mets). In the final selection, priority was assigned to enhancers showing higher enrichment to metastatic samples covering a total of 8.05 Mbps. We gave space also to a selection of enhancers with RI higher in primaries with RI <=50 in both primary and metastatic samples accounting to 3.7Mbps coverage. Finally, 2.5Mbps coverage was assigned to private enhancers being clonal in only 1 or 2 samples. As a control, SIDV2 includes 800 enhancers randomly selected from a pool of SI==1 with RI==100 enhancers (Fig. 72 B and C). As described in Chapter 2 section 2.7, DHS data available in ENCODE are highly conserved across cell lines and tissues and can be used to restrict the area of investigation within regulatory regions precisely toward the functional regions which are typically the NFR. We took advantage of this notion by shrinking the area of investigation to all the well-conserved DHSs within each enhancer thus being able to low the coverage respectively from a total of 14.65Mbps to 8.93Mbps.

SIDV- 2 includes 3MB of CTCF sites which have been selected through a multistep process. All the complete set of available CTCF, ZNF143 and cohesin ChIP-seq profiles from human tissues have been clustered (N=100). Each cluster was scored according to the binding strength and pervasiveness across multiple cell types. The clusters were ranked according to the scores and then selected up to covering ~3Mbp (>95% of the boundaries of topologically-associating domains or TADs from Dixon et al., Nature 2012) available in SID-V2.

In conclusion, SID-V2 is the direct upgrade of SID-V1 application specifically designed to investigate the genetic profile of enhancer and CTCF regions in ERα+BC samples. SID-V2 covers a total of 13Mbs of highly functional regions active in the different oncogenic stages of ERα+BC evolution in humans. SID-V2 represents the first known attempt to precisely investigate NCSMs in ERα+BC at

high resolution relying on an accurate target selection of non-coding genome from in-vivo data. This application will contribute to an immense number of unsolved questions whether 1) preferential NCSMs occur at specific TF motifs or enhancers participating to specific pathways 2) enhancers gain of function and effectiveness after point NCSM at specific sites 3) NCSMs point to specific non-coding loci which aid the phenotypic alteration to a more aggressive potential 4) NCSMs disrupt CTCF binding at specific TADs.

A

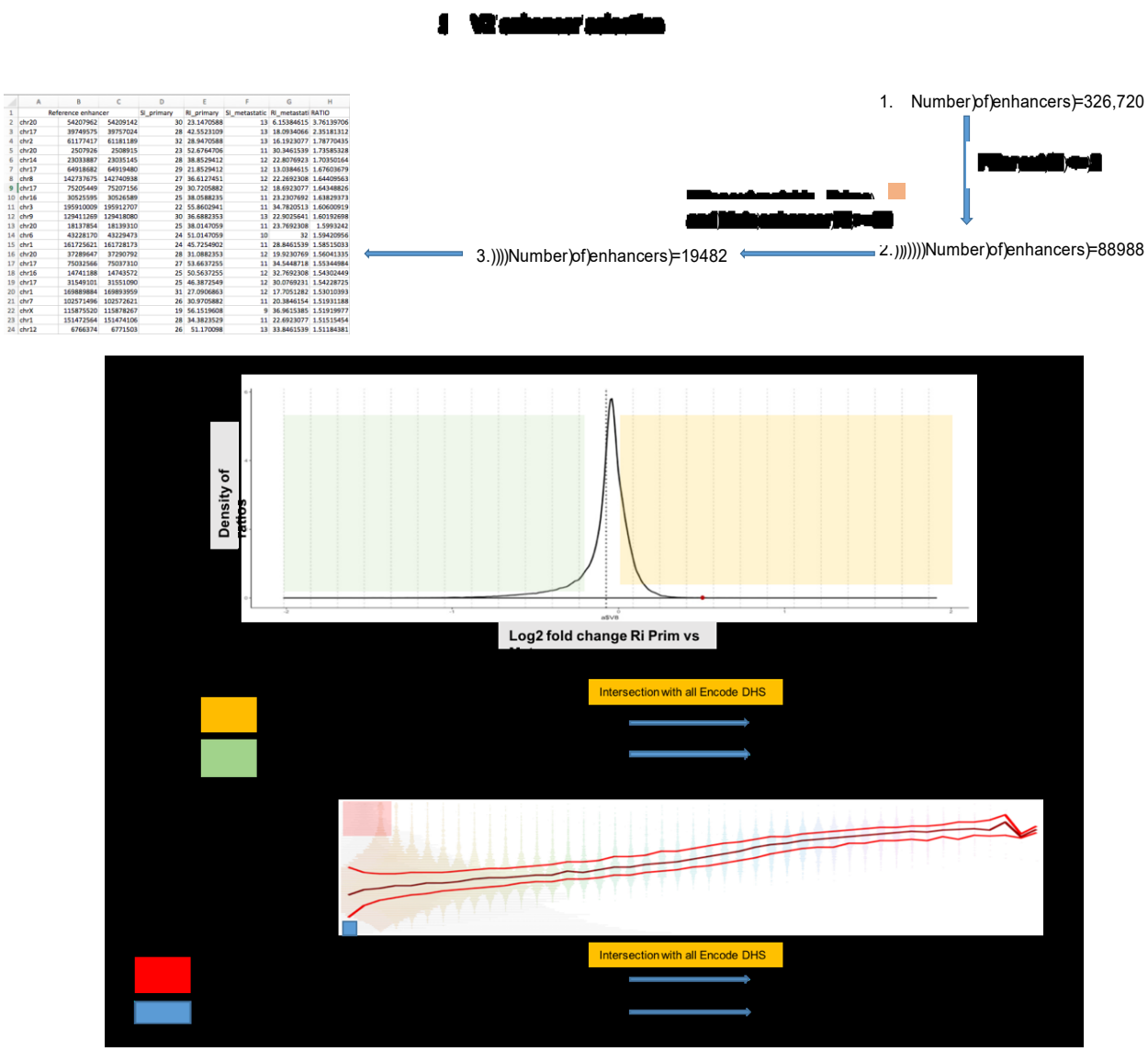


Figure 82.SID-V2 selection strategy.

Overview of the enhancer selection strategy applied to build SID-V2. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone.]

Chapter 4

Aims of the study

Adjuvant ET includes selective ERα modulators (i.e. Tamoxifen), selective ERα down-regulators (i.e. Fulvestrant) or systemic aromatase inhibitors (AI) designed to lower circulating oestrogens. Although the high efficacy of these drugs, 30% of patients relapse within ten years from the first surgical removal of primary BC. Drug-resistant mechanisms remain elusive at the molecular level, and it is not clear if ET-resistant cells originate through a stochastic reprogramming process or emerge via deterministic selection. It has been shown that ERα BC cell lines typically exhibit a distinct subpopulation of cancer stem-like cells (CSCs) with mesenchymal properties capable of driving tumour initiation[117]. Indeed, minor subpopulations of cells exhibiting stem-like properties are capable to seed new tumours in many cancers[118] and epigenetic reprogramming has been proposed to be intimately involved in these mechanisms[119]. CSCs cells can be identified using cell markers including CD44 protein. CD44 protein is a transmembrane cell adhesion which is commonly trusted[120,121] as a marker of stem-like potential linked to poor prognosis in different cancers. Recent observations[122] support the hypothesis that CD44 protein is overexpressed in cells conferring metastatic potential in ERα+BC. In this study, we aim to investigate whether epigenetics support putative CSCs to switch from quiescent to proliferative state and subsequently driving aromatase inhibitor resistance and metastatic progression. We provide evidence that subpopulation of cancer cells are highly plastic cells which 1) become quiescent upon oestrogen deprivation and 2) require additional reprogramming to resume proliferation. On the contrary, non-plastic cells are efficiently cleared by oestrogen deprivation.

By using experimental and computational approaches on cell lines and patient data, we hypothesise that loss of plasticity is irreversible in MCF7 CD44- and this condition is maintained by the loss of regulatory potential near Notch target genes. On the contrary, MCF7 CD44+ exhibit accessible profiles within the NOTCH target genes. In particular, the transcription factor ELF1 modulates the

expression of a-secretase subunits, such as PSEN1 and APH1A to regulate the Notch targets. Notch signalling is a crucial feature of breast luminal progenitor cells[123] and AI resistant cells have Notch dependencies[124]. We find evidence in cell lines that Notch inhibition significantly incentivises the switch of plastic cells into an ET sensitive phenotype. Indeed, anti-Notch drug promotes loss of plasticity by reducing the number of metastatic cells and significantly expanding the subpopulation sensitive to aromatase inhibitors.

In conclusion, the activity of Notch target genes represents a potential marker of CSC like potential. Interfering with the Notch pathway would force the conversion into ET sensitive cells.

4.1 Investigation of cell plasticity in cell lines and patients

We investigated whether the switch from non-CSC to-CSC occurs in MCF7 and LTED cell lines by monitoring the expression of CD44 protein on the cell membrane. We sorted MCF7-CD44high and CD44low subpopulations from a pool of MCF7 cells using FACS and then measured the re-population rate of CD44high and CD44low a week after the initial sorting. Our data fortified the notion that at time 0 only 4% of cells in a pool of MCF7 exhibit CD44 high protein levels. After monitoring cells for a week, a population of 100% CD44high cells reverted into a mixed population (Fig.85).

We also confirmed that MCF7-CD44low are non-plastic as they cannot convert to MCF7-CD44high (Fig.85). As expected, MCF7-CD44high cells were phenotypically different from MCF7-CD44low and possessed several CSC-like traits like higher motility (data not shown), higher expression of EMT genes (CDH2, Vimentin, Twist) (Fig.86) and proliferative capacity (Fig.87). Same experiments were repeated in MCF7-derived, aromatase inhibitor (AI) resistant long-term oestrogen-deprived (LTED) (Fig. 85,86,87). LTED cells mimic resistant metastatic cells and represent a powerful tool to understand mechanisms inducing drug resistance.

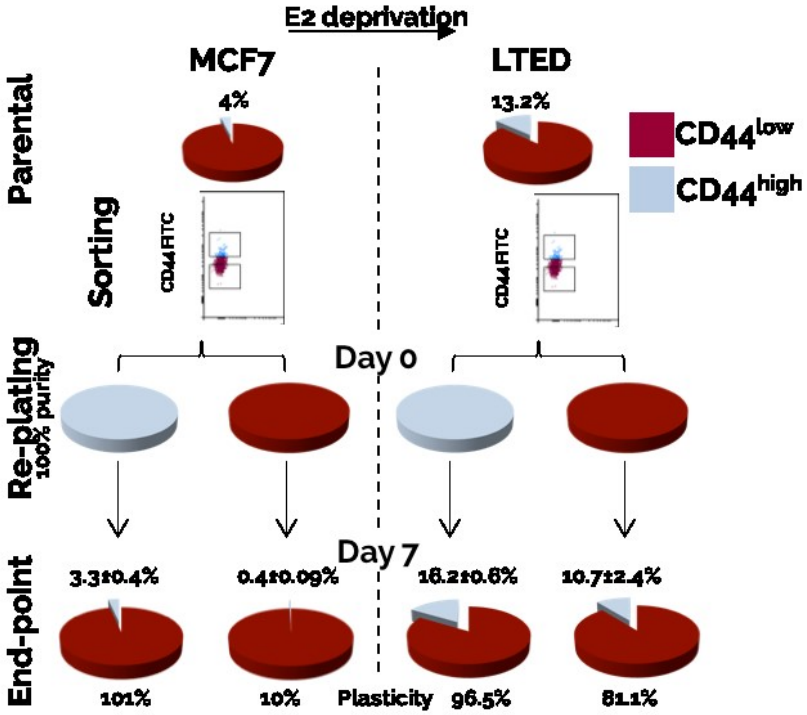


Figure 83. MCF7 and LTED plasticity assays.

MCF7 and LTED cells were sorted using CD44-FITC antibodies into CD44high and CD44low cells. FACS analyses to re-assess the relative percentage of both subpopulations were conducted one week after sorting. [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]

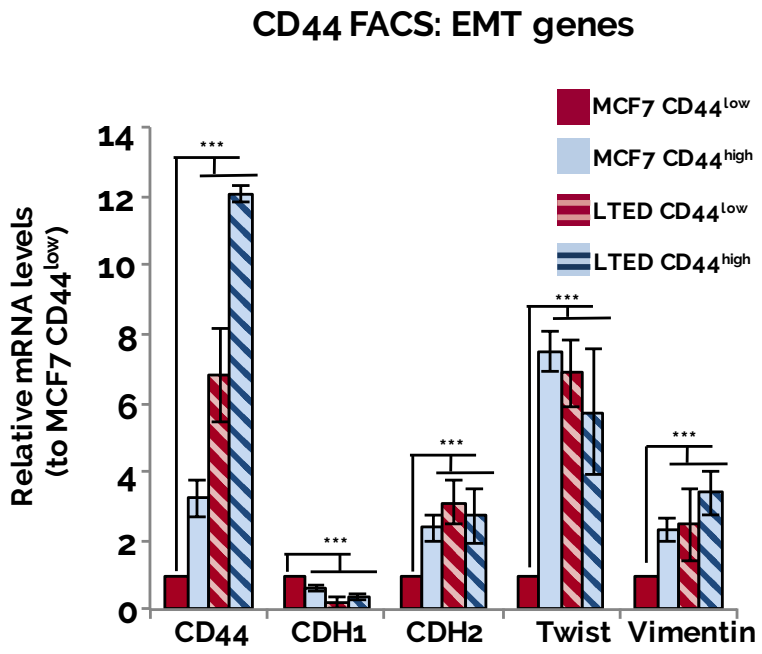


Figure 84 qRT-PCR analysis of EMT genes.

qRT-PCR analysis of EMT genes in sorted CD44high and CD44low from MCF7 and LTED cells. Asterisks represent significant differences after ANOVA analysis with Bonferroni's correction. [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]

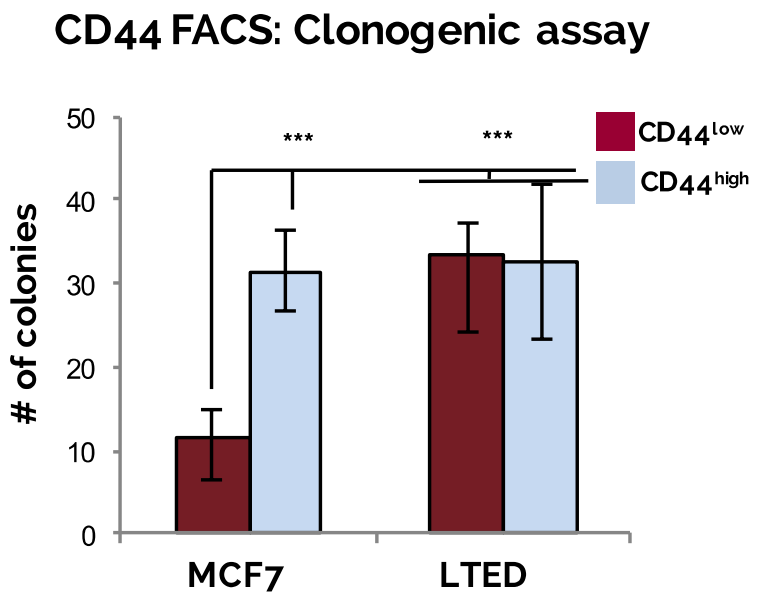


Figure 85. Clonogenic assay

Asterisks represent significant differences after ANOVA analysis with Dunnett's correction [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]

Strikingly, both LTED CD44low and CD44high manifested plastic capabilities in reverting into a mixed population. In agreement with these observations, both LTED-CD44low and LTED-CD44high were characterised by CSCs traits. To corroborate these data, we developed a CD44 knock-in cell reporter to monitor CD44 high status in the MCF7 population. We reached consistent results with the endogenous experiments and 1) confirmed the hypothesis that asymmetric division occurs in MCF7 CD44high 2) showed LTED CD44high revert to CD44low and, more importantly, 3) LTED CD44low can divide into LTED CD44 high population. To generalise these findings, we employed a second AI-resistant model (T47D and T47D-LTED) observing results in agreement with MCF7 and LTEDs observations (Fig.88 *migration essay and proliferative capacity not shown*).

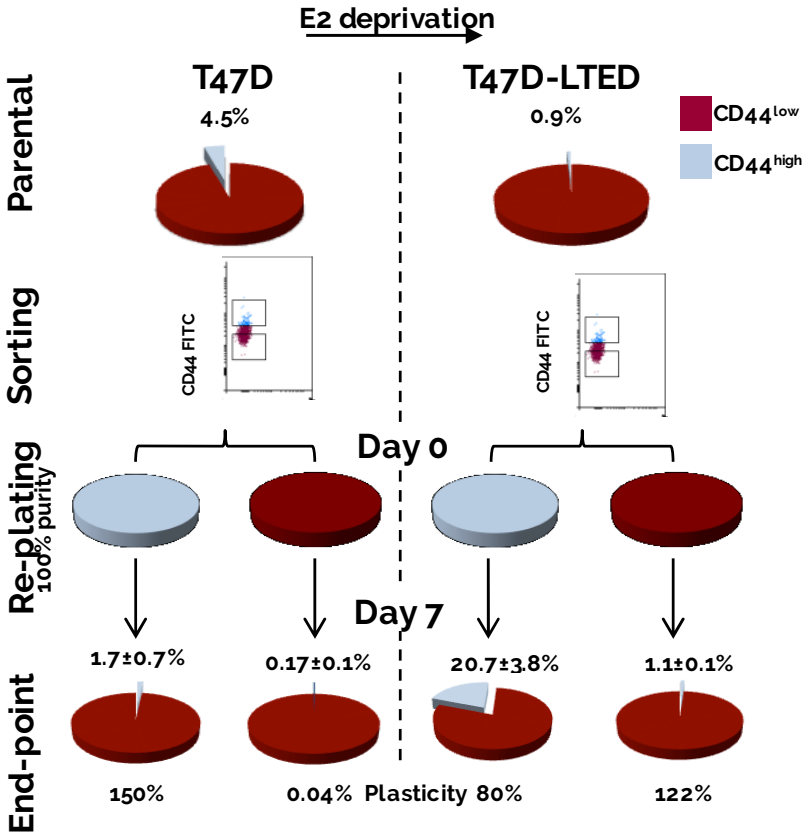


Figure 86. T47D and T47D-LTED plasticity assays.

T47D and T47D_LTED cells were sorted using CD44-FITC antibodies into CD44high and CD44low cells. FACS analyses to re-assess the relative percentage of both subpopulations were conducted one week after sorting. [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]

Our investigations in cell lines were extended in metastatic pleural cells derived from AI resistant patients and as expected, CD44low cells from all the patients could efficiently re-convert to CD44high cells (Fig.89). This finding supports the principle that oestrogen-deprived cells acquire stem-like capabilities. These data were corroborated by evaluating the relative mRNA levels of EMT genes and the GFP knock-in reporter essay.

Overall, these data demonstrate that metastatic patients with AI resistant ERα BC contain highly plastic cells, strongly confirming the in vitro models results.

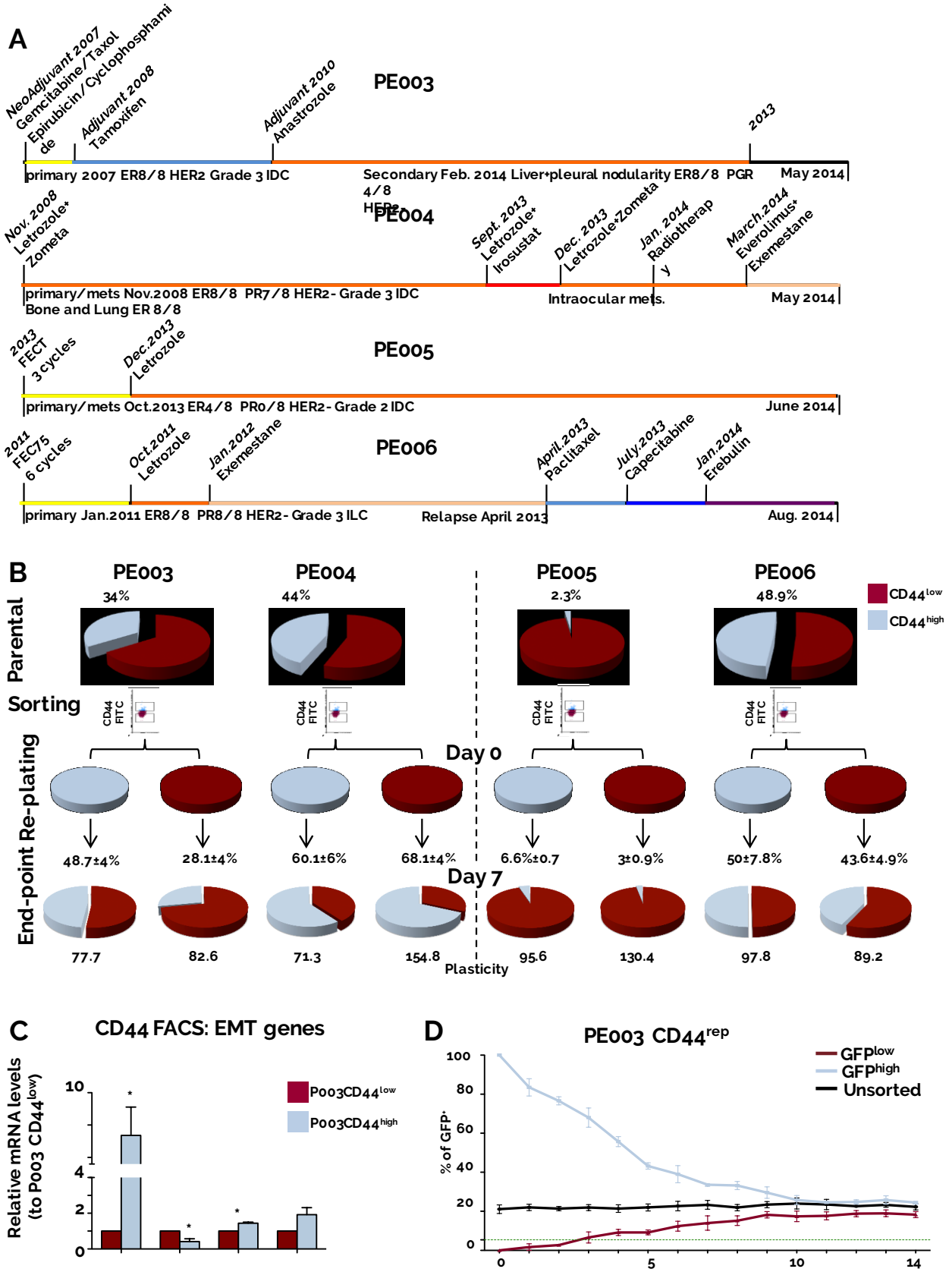


Figure 87. Cell plasticity is common in drug-resistant ERα+ BC cells.

A) Treatment history for all patients that donated the pleural effusion (PE) cells described in this dissertation. B) PE cells plasticity assay. Luminal metastatic breast cancer cells were sorted using CD44-FITC antibodies into CD44high and CD44low cell populations.

*FACS analyses to re-assess the relative percentage of both subpopulations were conducted one week after sorting. C) qRT-PCR analysis of EMT genes in sorted CD44high and CD44low population PE003 cells. D) Plasticity assays using PE003-CD44GFP cells. PE003 were sorted for GFP, and the ratio GFP high/total cells were monitored for 14 days. Percentages were calculated using Image J. Asterisks represent levels of significance as follow: * $p<0.05$. All experiments are reported as the average of at least in three independent replicates. [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]*

4.2 Cell of origin model may explain the shift from drug sensitive to a not sensitive condition

The presence of plastic cells in AI resistant models might also be explained by a cell-of-origin model in which plasticity is carried over from MCF7-CD44high to LTED cells via de novo resistance. However, clinical data and in vitro models demonstrate that the development of resistance is slow and can take months in vitro and years in vivo. One model that would reconcile these two scenarios might include semi-stochastic epigenetic reprogramming within a pre-defined subpopulation.

It is already showed in previous papers of our group that epigenetic reprogramming contributes to resistance[124]. Epigenetic reprogramming in these systems is not immediate and requires months from in-vivo studies. We then hypothesised that MCF7-CD44high might be the cell-of-origin from which LTEDs emerge following long-term epigenetic reprogramming.

We observed that MCF7 CD44 high and low behave differently upon oestradiol deprivation using proliferation assay. In the presence of oestradiol, both populations proliferate, however in the absence of oestradiol, only CD44high cells showed marginal proliferation properties (Fig.90). On the contrary, CD44 low completely stopped any proliferative activity, and their numbers were actually reduced by the treatment. This suggested that CD44high became cytostatic in an oestrogen deprived environment and potentially are the only subgroup of cells undergoing to an epigenetic rewiring during the time.

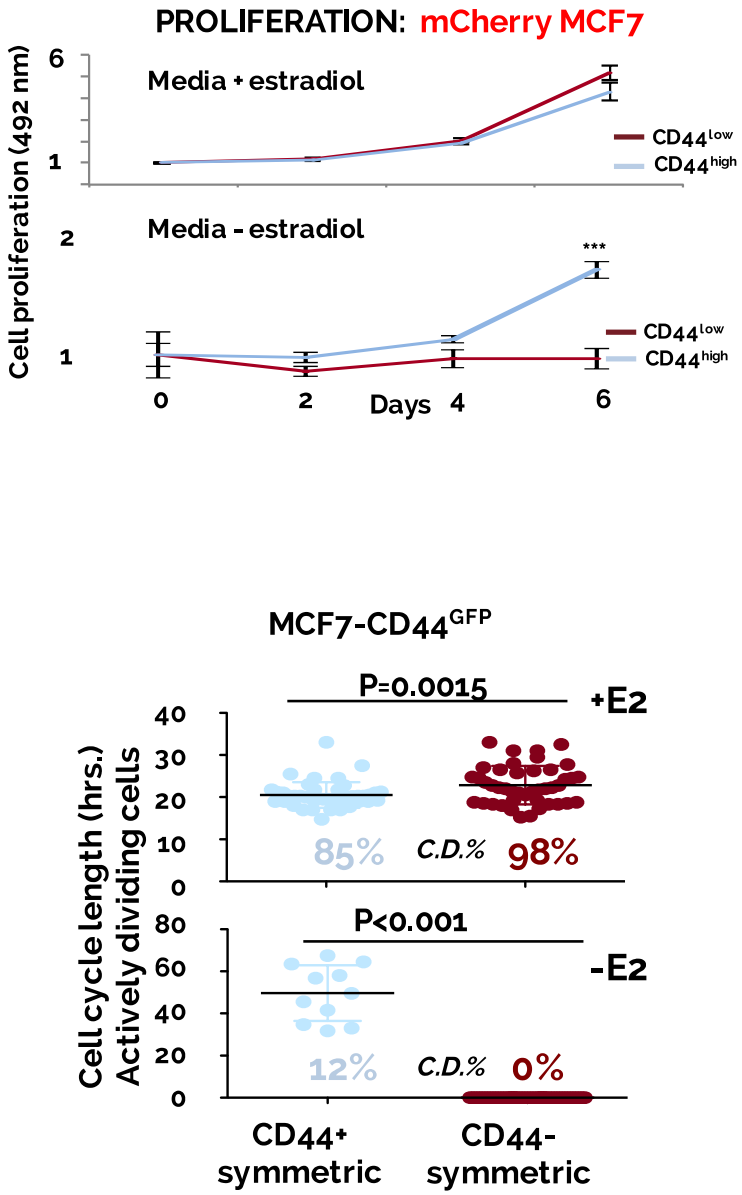


Figure 88.Differential cell response to oestradiol.

Top: Cell proliferation assay of sorted MCF7-CD44high and CD44low in oestrogen-supplemented or oestrogen-depleted conditions. Asterisks represent significant differences after Student's T-test. Bottom: cell cycle length analyses of MCF7-CD44GFP cell cultured in oestradiol or oestradiol-deprived conditions. Cell cycle length was calculated based on the first 2-3 cell divisions. [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]

We translated these findings by examining ERα+ BC tissues from 2 independent datasets using immunohistochemistry (IHC) in detecting the ratio of CD44positive and negative cells.

After neoadjuvant therapy, patients with ERα BC undergo primary surgery to remove the remaining local cancer cells. However, a possible source of relapse is represented by distal micrometastases[126]

entering a quiescent state until progression occurring years later. We examined a pool of sample obtained from a neo-adjuvant trial based on five months of AI which result in an expansion of CD44+ cells in the tissue (Fig.91).

We then examined a cohort of matched primary and metastatic lesions (all ER-positive, all treated with AI). Interestingly, all metastasis of this cohort were still mostly ERα positive however CD44 positive cells have an increment of 3.8-fold in patients undergoing five years of AI therapy.

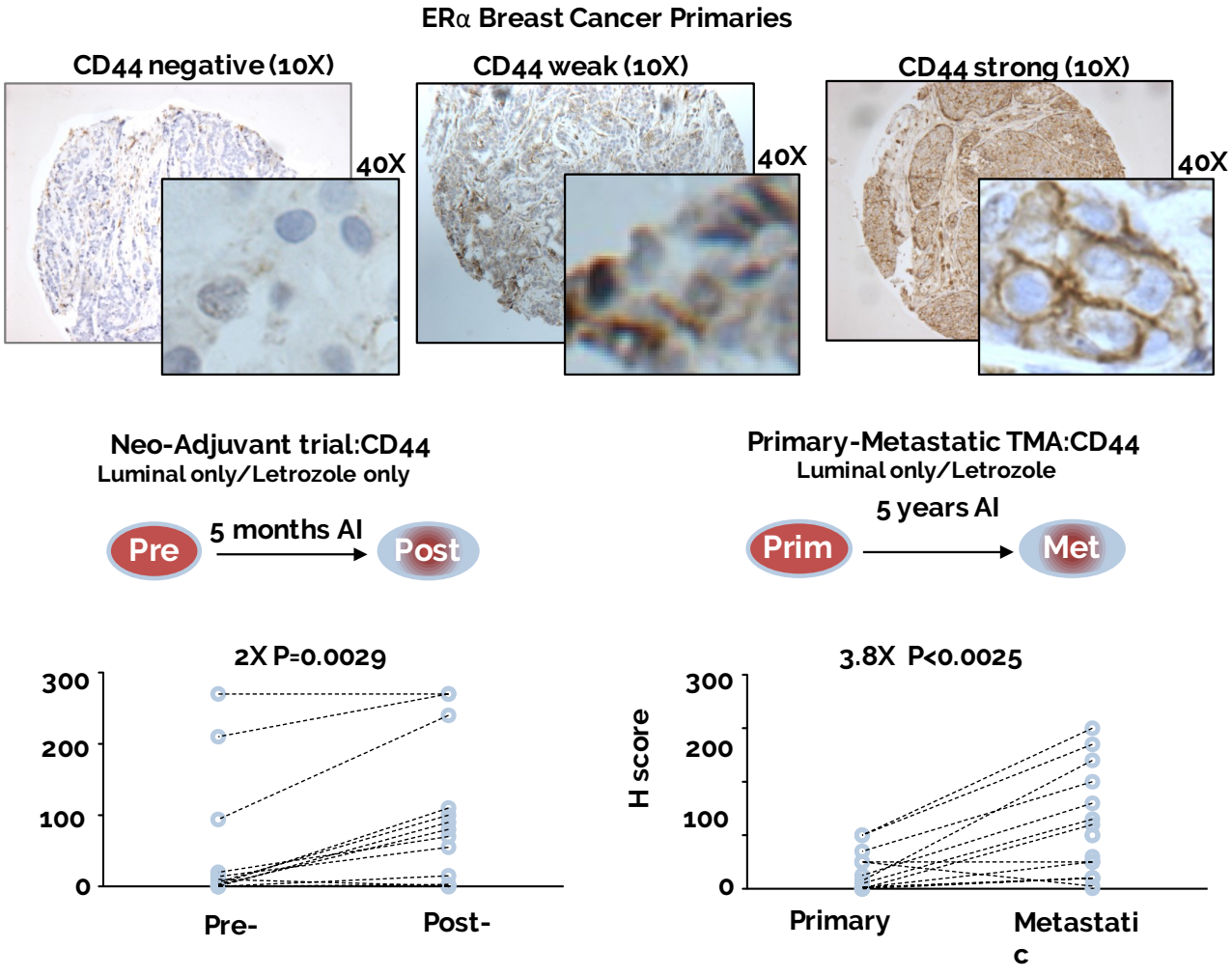


Figure 89 CD44high cells persist in patients with endocrine treatments.

Top: immunohistochemistry (IHC) staining of tissue micro-array cores presenting various degrees of CD44 staining Bottom left: IHC analysis of matched pre-treatment and post-treatment samples. Statistical significance was determined using a paired T-Test. All patients were classified as ERα positivity and received with neo-adjuvant therapy administering Letrozole for 5 months. Bottom right: IHC analysis of matched primary and metastatic samples. Statistical significance was determined using a paired T-Test. All patients received one cycle of adjuvant AI. [Experiments performed by Sung-Pil Hong. Figure produced by Luca Magnani]

Overall, these data lead us to speculate that ER-positive ERα BC subpopulations have a differential response to acute oestrogen deprivation potentially involving changes in cell cycle speed, the frequency of cell cycle entry and interconversion dynamics. An activation of unknown mechanism supports the quiescence of the CD44+ cell population which expand during oestrogen deprivation condition.

4.3 Computational assessment of CD44high and CD44low

Conversion from MCF7-CD44high to MCF7-CD44low is irreversible during our 7 days examination suggesting the existence of locking mechanisms. Chromatin accessibility plays a central role in bookmarking regulatory regions during normal cell differentiation[127]. Thus, we used ATAC-seq to annotate the accessible landscape of the individual CD44 high and low subpopulations focusing on MCF7 cells. MCF7-CD44high resulted in having a more significant number of accessible regulatory regions when compared to MCF7-CD44low (33864 vs 24558). Using a rigorous computational approach (Fig.92), described in chapter 5 section 5.5.3, we mapped the open landscape of both cell lines and investigated the differentially accessible regions.

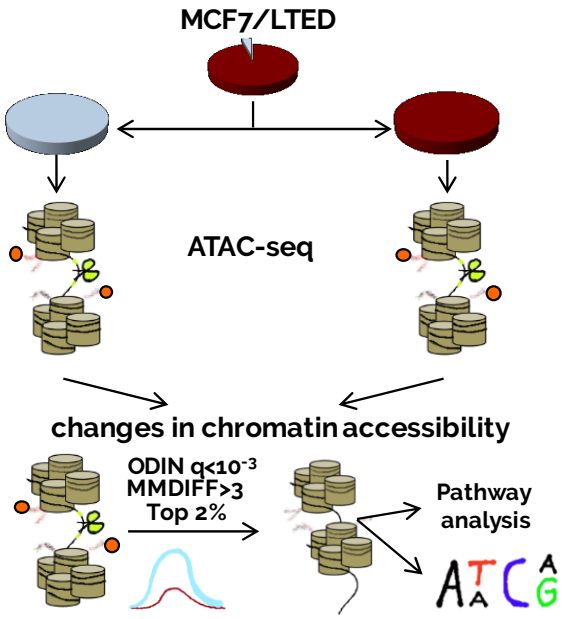


Figure 90 The computational strategy developed to identify differential DNA accessibility sites between CD44 high and low.

CD44 high and low were compared to identify changes in chromatin accessibility. ODIN and MMDIFF were used to identify differential peaks. Different cut-offs were applied to the data. The significant accessible sites were deployed for further analysis. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone and Luca Magnani]

We found 466 regions lost in MCF7-CD44low and only nine regions gained providing principles that differentiation is enforced through a loss of regulatory potential from CD44high to CD44low. As expected, these regions remained accessible in LTED model, regardless of CD44 status (Fig.93).

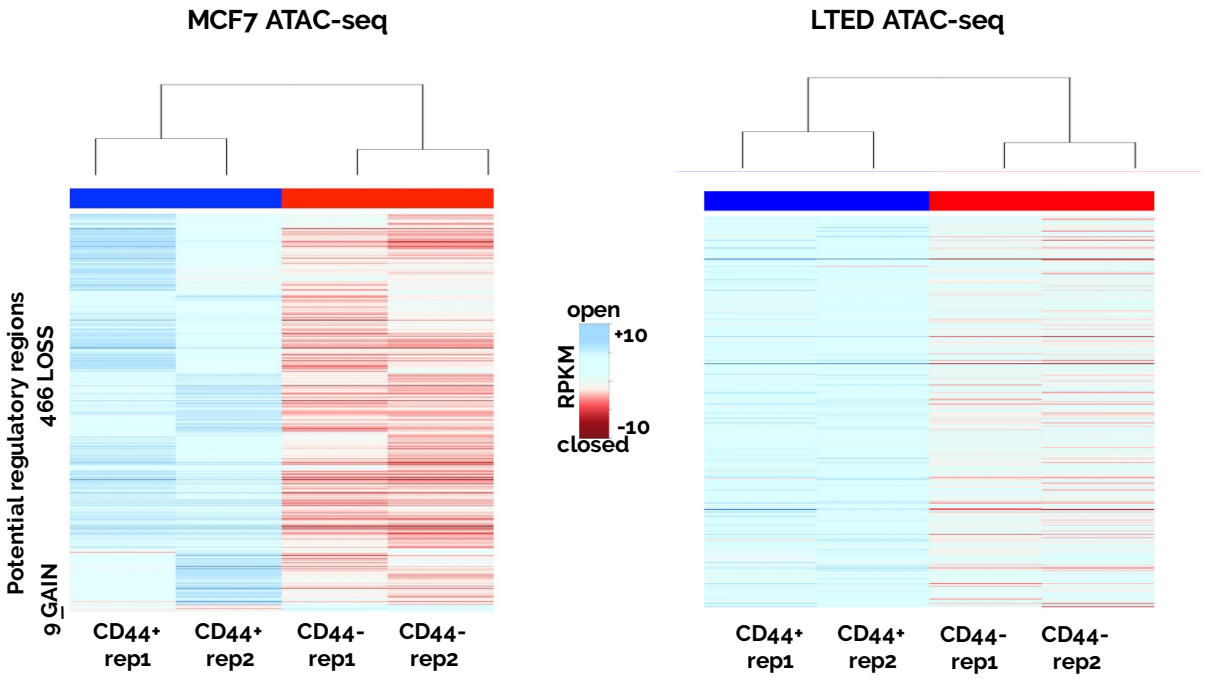


Figure 91. CD44high cells are epigenetically distinct

Heatmaps illustrating normalized ATAC-seq read counts at loci lost and gained during differentiation. Colours represent reads differentials. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone]

We next explored the biological features of these regulatory regions. MCF7-CD44high-specific loci are significantly associated with genes expressed explicitly in the invasive subpopulation of cancer cells from mammary carcinoma. Motif analysis identifies critical transcription factors (TFs) involved in ERα BC progression including ELF1, ELK4, and KLF5[122](Fig.94). These TFs were highly expressed in MCF7-CD44high. Using pathway clustering, we found that regions were losing accessibility in MCF7-CD44low map with a significant frequency near genes involved in Notch

signalling such as PSEN1 and APHA1A[128]. We then validated chromatin accessibility data with quantitative ATAC-qPCR (not shown) and FAIRE-qPCR (Fig.94 D) further demonstrating that these regions are differentially accessible in the two subpopulations.

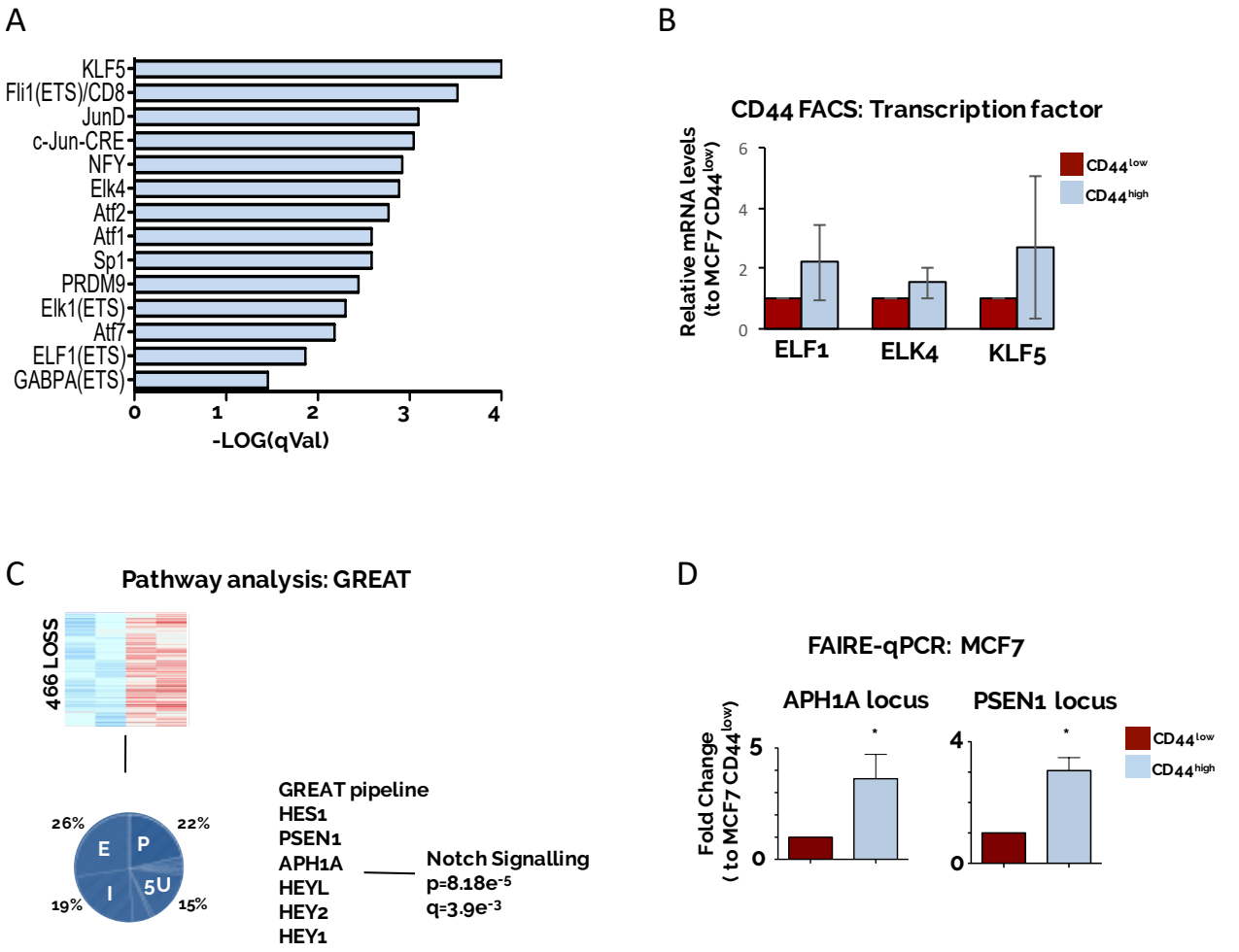


Figure 92. Downstream analysis of differential accessibility of CD44+ and CD44.

A) Motif analysis performed on the differential DNA accessible sites identified in the computational workflow (N=466). B)-Relative mRNA levels of the top transcription factors identified in the analysis to MCF7 CD44low. C) Pathway analysis performed using the 466 DNA accessible regions associated to the closest gene reveal NOTCH signalling. D) FAIRE essay to visualise the enrichment of regulatory loci identified in the pathway analysis. All experiments are reported as the average of at least in three independent replicates. [Analysis performed by Giacomo Corleone. Experimental procedure performed by Sung-Pil Hong. Figure produced by Giacomo Corleone and Luca Magnani]

We next validated that Notch targets were up-regulated in MCF7-CD44high than CD44low (Fig.95). Notch signalling has been implicated in the maintenance of cancer stem cells in breast cancer, glioblastoma[129,130] and represents a key survival pathway for AI resistant LTED cells[131].

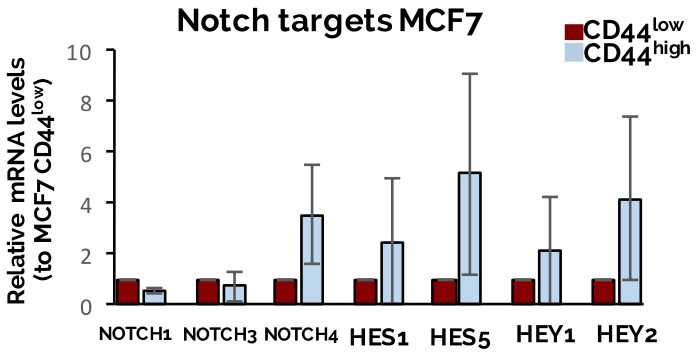


Figure 93 qRT-PCR analysis of Notch targets in sorted CD44high and CD44low from MCF7.

Asterisks represent significant differences after ANOVA analysis with Bonferroni's correction. [Experimental procedures performed by Sung Pil Hong. Figure produced by Giacomo Corleone and Luca Magnani]

We then hypothesised that there might be a link between the TFs enriched in MCF7-CD44high accessible regions and Notch signalling. To investigate this, we initially analysed ELF1 ChIP-seq data in several cell lines focusing on the promoters identified by differential ATAC-seq. We performed ELF1 ChIP-qPCR in the two MCF7 subpopulations and found increased ELF1 binding in plastic MCF7-CD44GFP-high cells (Fig.96).

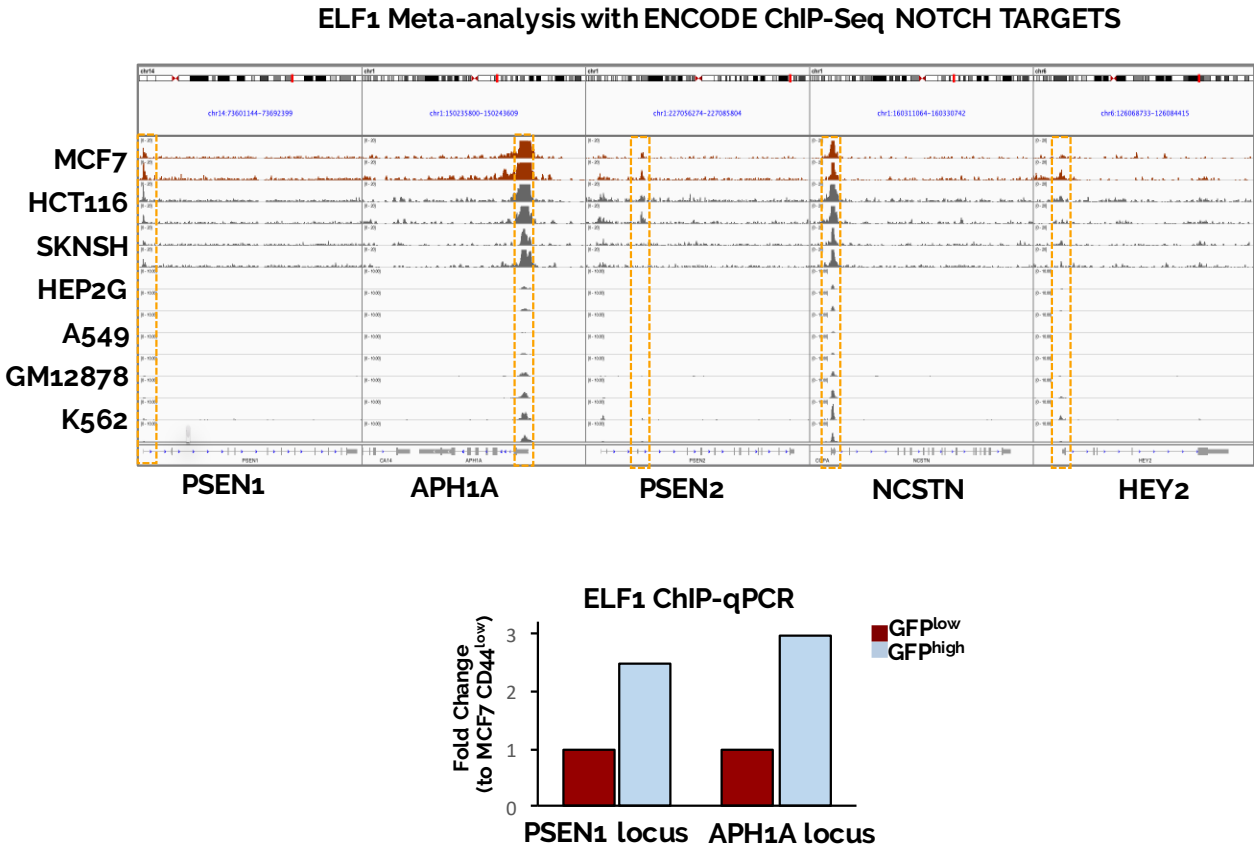


Figure 94.ELF1 Meta-analysis integrated with ENCODE ChIP-Seq of NOTCH targets.

Top: DNA accessibility profile of the common NOTCH target in MCF7 and other cell lines. Bottom: ELF1 ChIP-qPCR analysis with primers designed within the ATAC-seq peak. [Analysis performed by Giacomo Corleone. Figure produced by Giacomo Corleone and Luca Magnani]

Silencing ELF1 significantly reduced the transcripts of PSEN1 and APH1A. ELF1 depletion led to a significant increase in MCF7-CD44GFP-high conversion to MCF7-CD44GFP-low, and a decrease of Notch target genes expression. On the other side, silencing Notch3 and Notch4 resulted in the loss of plasticity and rapid conversion of MCF7-CD44GFP-high to MCF7-CD44GFP-low. These data suggest that ELF1 regulates Notch targets and that Notch signalling is fundamental to maintain a plastic phenotype.

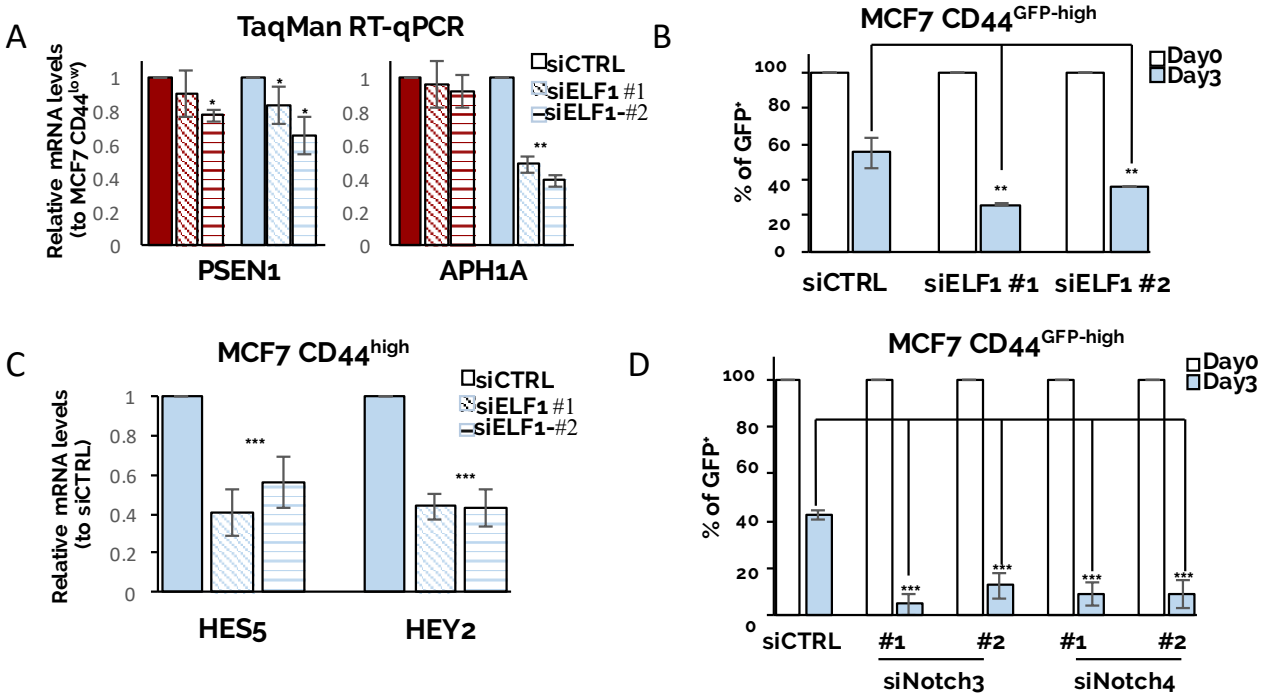


Figure 95 Experimental validation of siELF1.

*A) TaqMan RT-qPCR analysis of PSEN1 and APH1A in sorted CD44high and CD44low from MCF7 after silencing ELF1. Asterisks represent significant differences after ANOVA analysis with Bonferroni's correction. B) Plasticity assay using MCF7-CD44GFP cells after silencing ELF1. MCF7 were silenced for ELF1 for 48 hours and then sorted for GFP. The ratio of GFPhigh/total-cells was monitored on day0 and day3. Percentages were calculated C)) qRT-PCR analysis of Notch targets in sorted CD44high and CD44low from MCF7 after silencing ELF1. Asterisks represent significant differences after ANOVA analysis with Bonferroni's correction. D) Plasticity assay using MCF7-CD44GFP cells after silencing Notch3 or Notch4. MCF7 were silenced for Notch3 or Notch4 for 48 hours and then sorted for GFP. The ratio of GFPhigh/total-cells was monitored on day0 and day3. Percentages were calculated using Image J. Statistical significance was calculated using a 2-way ANOVA (vs CD44GFP-high siCTRL). All experiments are reported as the average of at least in three independent replicates. Asterisks represent levels of significance as follow: * $p<0.05$, ** $p<0.01$ *** $p<0.001$. [Experimental procedure performed by Sung Pil Hong. Figure produced by Giacomo Corleone and Luca Magnani]*

Collectively, these data support the notion that deterministic DNA accessibility dynamics is occurring between CD44low to CD44high state. Specifically, CD44high cells are epigenetically into a stem-like state. Indeed, CD44high cells are enriched of around 10k more accessible sites compared to CD44low however only 466 sites are significantly contributing to the phenotypical switch. Our data demonstrated that Notch signalling is of fundamental importance in CD44high activity. Also, interfering with Notch target force conversion to an ET sensitive subpopulation.

Chapter 5

5.1 Computational methods chapter 2

5.1.1 H3K27ac ChIP-seq data processing

Reads have been quality controlled using FastQC[132] v0.11.5, aligned to the human genome reference hg38 using bowtie[133] v1.1.2.5 with default parameters, converted into BAM and then sorted and indexed with SAMtools[134] v1.3. ChIP-seq peaks were generated with MACS2[135] v2.1.16 with the following parameters set: -callpeak --format AUTO -B --SPMR --call-summits -q 0.01. Each peak call was normalised with the sample-specific input. We did not consider samples showing either less than 2K or more than 200K H3K27ac peaks for further analysis. A total of 337 epigenomic profiles including 33 cell lines from ENCODE[136] and 37 tissues from Epigenomic Roadmap[137] were re-analysed with the following procedure:

1. Matching. bam and. bed profiles from ENCODE and matching raw reads of input and H3k27ac ChIP from Epigenomic Roadmap were downloaded
2. The epigenomic profiles of ENCODE cell lines from the human hg19 reference genome were lifted to the human hg38 assembly using CrossMap[138] v0.2.312.
3. Epigenetic profiles of the samples downloaded from Epigenomic Roadmap were called following the procedure described above.
4. Active promoters and enhancer of BC samples were intersected with all the epigenomic profiles, and then RI calculation of each peak was repeated as described in the RI section.

5.1.2 Assessment of heterogeneity

The level of heterogeneity of promoter and enhancer dataset was assessed to reveal whether the available dataset achieves high genomic coverage. ACT[139] SaturationPlotCreator with default parameters was used to perform saturation analysis. Frequency distributions of enhancer and promoter set coupled with average peak size distribution were calculated by intersecting individual peaks with the associated master list and then plotted using in-house BASH and R scripting. Peak sizes were extracted from MACS2[139] output files and the peaks binned by sharing index.

5.1.3 Functional characterisation of the peaks

After peaks generation, an in-house pipeline based on BEDTOOLS[140] v2.25.0 and custom BASH scripts were developed to assign the annotation "enhancer", "promoter" to each annotated peak. A master annotation from promoters was generated from UCSC[141] table browser (PMID 27899642) (assembly: hg38; groups: Genes and Gene predictions; track: GENCODE[142] v24. This master list of hg38 promoters classifies the promoter as the region 1kb upstream of the transcription start site (TSS). We intersected the peaks using BEDTOOLS intersect (default parameters) to identify the promoter-specific peaks active in BC. All the peaks not overlapping with the patient signal but annotated in hg38 genome reference were considered inactive.

The master list of active core promoters was produced with a multistep procedure:

1. To identify the common overlapping signal amongst all the samples, promoter peaks were intersected using BEDTOOLS multiinter.
2. The book-ended regions from the core signal file were merged using BEDTOOLS merge, then intersected with the original peak calls and sorted.
3. All those peaks showing no overlap with the promoter annotation were considered enhancers.

The same procedure described above was applied to build a master list of active enhancers with the only exception of using non-promoter peaks.

5.1.4 Sharing Index (SI)

SI is assigned to enhancers and promoters in the master list. 2 different methodologies have been applied in SI calculation to add further significance to the accuracy of this metric. Generally, SI is a discrete metric introduced to measuring the level of penetrance of regulatory regions across the tumour dataset. In particular, SI calculation relies on the regulatory region activity in the patient dataset.

The first methodology consists in assigning SI to each regulatory region in the reference master list by counting the number of patients in which each promoter or enhancer overlap with at least 40% of genomic coverage annotated in the master list.

The second methodology is a quantile normalised continuous equivalent of SI calculated. This workflow requires the computation of a matrix populated of the read count of each regulatory region in each sample as follows:

1. A matrix having individual samples ad column and regulatory regions as rows were built
2. The matrix was populated with the number of deduplicated reads matching each regulatory region coordinate in any given sample. BEDTOOLS Multicov with default parameters was used to calculate the matching deduplicated reads at each region.
3. The matrix is quantile normalised after Voom[143] transformation included in LIMMA[144] package available in Bioconductor[145].
4. Data were scaled (z-score) and further compressed with (arcsinh) transformation

5.1.5 Ranking Index (RI)

RI is the intra-sample quantile normalised measure of enrichment calculated at each regulatory region. RI is calculated as follows:

1. Duplicated reads form ChIP-Seq treatment files were filtered out using PICARD v2.1.1 MarkDuplicates with parameter (REMOVE_DUPLICATES=true). Further analysis was conducted using only uniquely mapped reads.
2. BEDTOOLS Multicov function was used to compute the peak read counting, and these values were further normalised using the following equation:
3. Nscore=((peak read count / peak size) · 106)) · 103 / total mapped reads (FPKM).
4. Peaks calls in each sample were split in promoter and enhancer set according to what described in the paragraph "H3K27ac ChIP-seq data processing".
5. Split peak calls were sorted according to their relative FPKM and assigned to their respective intra-sample percentile score.
6. This final intra-sample percentile score is the RI where 1 is the highest enrichment value, and 100 is the lowest.

5.1.5 Ranking Index in cancer cell lines and normal tissue epigenomes

Ranking index approach was applied to all ChIP-seq data of H3K27ac profile available in ENCODE and Epigenomic Roadmap on October 2016. We re-analysed a total of 337 epigenomic profiles. We downloaded matching. bam and. bed from ENCODE and matching input and ChIP from Epigenomic Roadmap. All the epigenomic data annotated in the hg19 reference genome were lifted to the human hg38 assembly using CrossMap v0.2.312.

5.1.6 Transcription factor profiling

We profiled the BC cistrome by considering all the potential accessible sites encoded in the active promoters and enhancers. TF proteins bind in the so-called nucleosome-free regions (NFR) which are the regions of DNA dynamically exhibited between 2 nucleosomes. NFRs were computationally characterised by selecting 1) the regions among two subpeak summits at a distance of +- 71bps and 2) the regions between 2 broad peaks distant at most 500bps.

To perform a more accurate analysis pointed on the DNA regions effectively bound by TF, we matched the selected NFR (with the method showed above) with a panel of all the putative DHS sites. This DHS panel was obtained by downloading 220 different ENCODE DHS-seq profile from all the available cell lines and tissues at: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwDnase/ and http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromDnase/; DHS profiles were generated using MACS2 with the following parameters: --format AUTO --nomodel -shift -100 --extsize 200 -B --SPMR --call-summits -q 0.01 and lifted to the human hg38 assembly. After that, all the DHS peaks were concatenated into one sorted BED file.

We then matched the NFR with the DHS panel and only the intersected DHS site were retained for further data processing. These selected DHS regions were sorted and stretched using BEDTOOLS merge function in order to have a unique DHS flag for all the NFR. This approach resulted in a high-reliability panel of putative DHS specific to promoters and enhancers of ERα+ BC patients

This panel of high integrity DHS was used to carry motif enrichment analysis specific promoters and enhancers. HOMER function findMotifsGenome.pl with parameters: -size given -preparse was run to perform the motif analysis and the highest 50 ranked TFs in the two groups (enhancer and promoters) were selected and represented in polar histograms by developing a custom R script.

Moreover, we performed an SI-based motif analysis specific to promoters and enhancers SIs bins by separately binning the high integrity DHS to SI annotation. The motif analysis was performed using the HOMER function findMotifsGenome.pl with parameters: -size given -preparse. The results relative to each bin were further filtered for their statistical significance setting q-val <=0.05. The filtered data were integrated with the observed/expected ratio (OEr) of each TF with a custom R script. OEr results amongst the bins were exploited to create two heatmaps (one for promoters, the other for enhancers). Heatmaps were generated using the function heatmap.2 available in the package ggplot2[146] from R library.

The most significant results from enhancer heatmap were highlighted by computing a differential analysis between the two clades of the heatmap (SI 1-21 and SI 22-44) and showed in a separate figure. We calculated the mean of OEr assigned to each TF in both clades and counted the number of significant enrichments in each clade. A weighted score specific to each TF was computed by multiplying the relative TF clade mean x the number of significant clade enrichment. We further computed the log of the ratio, followed by ranking and plotting.

5.1.7 Footprint analysis

Analysis of footprints was performed by exploiting ATAC-seq data from a panel of MCF7 cell lines available in the lab. Footprints from the NFR regions were identified using Wellington[147] with parameters -fdr0.01 -pv -5, -10, -20, -30, -50, -100. We then intersected the patient's regulatory regions and the associated RI with the cell lines NFR. We then calculated the number of footprints in each active regulatory region in patients and normalised the number of footprints by region size. Finally, we plot the normalised number of footprints of cell lines regulatory region in the function of the relative RI.

5.1.8 Assessment of somatic copy number alteration (CNA)

We estimated the somatic copy number alterations of each tumour in the dataset and available cell lines. Input BAM files from ChIP-seq experiment were exploited to estimate the chromosomal losses and gains in a shallow whole genome sequencing manner. Duplicate reads were removed from each input raw bam and were processed using two methods, QDNAseq[148] and CNVkit tools[149].

QDNAseq workflow includes genome binning, GC content and mappability correction followed by data normalization. We processed the data as described in the QDNAseq package manual.

The CNVkit analysis was performed following the standard procedure described in https://cnvkit.readthedocs.io/en/stable/pipeline.html. Flat reference genome was created as suggested in the manual and CAN profiles were obtained running the batch command.

5.1.9 Assessment of dinucleotide composition

Dinucleotide artefacts are a known source of bias in ChIP-seq studies. We assessed the dinucleotide composition of the samples genomic profiles to exclude the possibility SI scores were affected by unexpected dinucleotide frequencies. Furthermore, we evaluated also bin specific dinucleotide composition. The expected dinucleotide frequency was computed by merging all the available deduplicated Input bam files using SAMtools function “merge”. The merged BAM files were then converted to FASTA and used as input to”compseq” module of EMBOSS[150] with parameter “word -2” to calculate the frequencies of the 16 possible dinucleotides. Frequencies of bin specific dinucleotides were obtained by converting the regulatory regions coordinates from bed to FASTA and then computing the frequencies. BEDTOOLS function “getFasta” was used to convert the bed into fasta and EMBOSS compseq with parameters -word2 to calculate the actual dinucleotide frequencies within the bins.

5.1.10 Targeted capture essay processing

NEB Cancer Hotspot panel was used for performing the Targeted capture sequencing essay. The panel was modified to include the ESR1 ligand binding domain (NEB E7000X). Input material used for ChIP-seq analysis described in Chapter 2 was used and processed as specified by the manufacturer.

24 samples per lane in 2 lanes (Single End 75bp flow cell) we multiplexed in a NextSeq Illumina machine. The obtained Single-end reads long 75base -pairs we aligned to the human genome hg38 using bwa1 version 0.7.15 setting the -q 0 and default parameters. SAM files generated by the alignment procedure were converted in indexed BAM files then sorted and indexed using Samtools version 1.3.1 The aligned reads from each captured sample were processed following GATK2[151] v3.6 best practises for variant identification:

1. Picard (http://broadinstitute.github.io/picard) version 2.6.0, applying functions AddOrReplaceReadGroups (parameters RGID=1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=1) and sortSam (parameters: SORT_ORDER=coordinate).
2. Read duplicates were marked using the MarkDuplicates function from Picard (parameters REMOVE_DUPLICATES=False AS=True).
3. RealignerTargetCreator and IndelRealigner from GATK (known indels from the GATK bundle: Mills_and_1000G_gold_standard.indels.hg38.vcf) were used to re-align reads around indels.
4. Quality scores of each aligned base were recalibrated using GTAK BaseRecalibrator
5. Finally, variants were called separately on each capture using Mutect2 (part of GATK v3.6) without control samples.
6. The identified variants were then annotated to known SNPs using 1000G_phase1.snps.high_confidence.hg38.vcf downloaded from GATK bundle and then filtered

7. Further annotation was performed matching the variants with the broadest mutational database at time COSMIC[152] 3 version 34 (hg38).

Retained variants which were showing allele frequency lower than 1% were excluded from further analysis. Variants covered by ten or more reads and supported by evidence on both alleles were retained. All the variants predicted as "Neutral" by FATHMM[153] were removed, and we finally kept only the variants previously reported in COSMIC.

5.2 Brief experimental methods chapter 2

5.2.1 Tumour tissue processing

Tissue collection was conducted by the Imperial Tissue Bank (project ethic approval R15021) and from Breast Cancer Now Tissue Bank (BCNTB- TR000053-MTA and TR000040). Tissues were selected by high tumour burden (tumour burden > 70%). Patient fresh frozen tissues were macrodissected, sectioned into 2 mm x 2mm fragments, placed in a Petri dish and covered with dry ice. Besides, tumours were fixed using 1% formaldehyde solution for ten minutes followed by Cold glycine for 10min. The fragments were ground by using pestle and mortar and homogenised using liquid nitrogen

5.2.2 ChIP

ChIP protocol was conducted as described in Patten., Corleone., and Magnani in Meth.Mol.Bio. 2018 with few modifications:

1. After fixation, chromatin was extracted and sonicated using a Bioruptor Pico sonication device (Diagenode, B01060001) using 20 cycles (30s on and 30 s off) at maximum intensity.
2. Chromatin immunoprecipitation was conducted using 4 μ g of H3k27ac antibodies (Abcam; ab4729) per ChIP experiment

3. YY1 ChIP-seq conducted in biological duplicates before and after 45min oestrogen stimulation using 4 µ g of YY1 antibodies (Santa Cruz; sc-281X)
4. Non-immunoprecipitated chromatin was used as input control
5. Sonication efficiencies were assessed using 1% agarose gel.
6. Before construction of ChIP-seq libraries (NEB Ultra II kit, Supplementary Methods), enrichment of the immunoprecipitated sample was ascertained using positive and negative controls for ChIP-qPCR.
7. Library preparation was conducted using 10–50 ng of immunoprecipitated and Input samples.
8. Libraries were retested to confirm enrichment before sequencing.

5.2.3 ChIP–qPCR

1. Reactions were carried out in 10 µ l volume containing 5 µ of SYBR-green mix (ABI; 4472918), 0.5 µ l of primer (5 µ M final concentration), 2.5 µ l of genomic DNA and 2 µ l of DNASE/RNase-free water.
2. Three step cycle programme was run as follows: 10 s at 95 °C, 30 s at 60 °C and 30 s at 72 °C, repeated 40 times followed by melting analysis.

5.2.4 Cell culture

MCF7 were routinely maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS) and 100 U penicillin/0.1 mg ml−1streptomycin, 2 mM l-glutamine plus 10−8 17-β -oestradiol (SIGMA E8875).

MCF7 long-term oestrogen-deprived (MCF7-LTED) cells were routinely grown in phenol-free DMEM with 10% charcoal-stripped FCS (DCFCS) and 100 U penicillin/0.1 mg ml−1 streptomycin and 2 mM l-glutamine.

T47D and T47D-LTED cells were passaged using DMEM containing 10% FCS and 100 U penicillin/0.1 mg ml−1 streptomycin, 2 mM l-glutamine and phenol-free DMEM with 10% DCFCS and 100 Upenicillin/0.1 mg ml−1 streptomycin and 2 mM l-glutamine, respectively.

ZR75-1 cells were grown in DMEM containing 10% FCS and 100 U penicillin/0.1 mg ml−1 streptomycin, 2 mM l-glutamine.

5.2.5 siRNA

siRNA was used to interfere with SLC9A3R1 (gene ID 9368: Ambion s17919, s17920), YY1 (gene ID 7528: Ambion s14958, s14959, s14960) and Silencer negative control (Ambion AM4611).

siRNA was conducted in MCF7, T47D and ZR75-1cells. Generally, six-well plates were used for seeding cells (1.5 x 105 cells per well).

MCF7 cells were seeded in phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml−1 streptomycin and 2 mM l-glutamine. After 24 h, the cells were transfected with siRNA using Lipofectamine 3000 (Invitrogen L3000015).

T47D and ZR75-1 cells were seeded in DMEM containing 10% FCS and 100 U penicillin/0.1 mg ml−1 streptomycin, 2 mM l-glutamine. After 24 h, the cells were transfected with siRNA using Lipofectamine 3000 (Invitrogen L3000015).

Cells were settled for analysis after 48h transfection.

5.2.6 RNA extraction and real-time PCR

RNeasy Mini Kit (Qiagen 74106) was used to purify and extract the total RNA. The cDNA was then reverse transcribed from 1 µ g of RNA using iScript cDNA synthesis kit (Bio-Rad 1708891).

Real-time qPCR (RT–qPCR) reactions were carried out in a 10 µ l volume containing 5 µ l sybergreen mix (ABI 4472918), 0.5 µ l primer (2.5 µ M final concentration), 2.5 µ l genomic DNA and 2 µ l DNASE/RNAse-free water. A three-step cycle programme and a melting analysis were applied. The cycling steps were 10 s at 95 °C, 30 s at 60 °C and 30 s at 72 °C, repeated 40 times.

5.2.7 Luciferase reporter assay

1. MCF7 cells were seeded in a 24-well plate at 5 × 104 cells per well in phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml−1 streptomycin and 2 mM l-glutamine.
2. DNA plasmid transfection was performed using Lipofectamine 3000 (Invitrogen L3000015) after 24 hours of incubation.
3. Cells were transfected with 100 ng of ERE Luciferase reporter, 10 ng of the renilla luciferase control plasmid (pRL-CMV), 10 ng of pSG5_ER-α, 15 nm of siRNA and 280 ng of Bluescribe DNA (BSM) per well, totalling 400 ng of DNA per well.
4. After 12 h of transfection, the medium was replaced with fresh phenol-free DMEM with 10% DCFCS and 100 U penicillin/0.1 mg ml−1 streptomycin and 2 mM l-glutamine.
5. Treatment with 10−8 17-β -oestradiol (Sigma E8875) or control treatment was administered, and the cells incubated for 24 h.
6. Cell lysates were then obtained using Passive lysis 5× buffer (Promega E1941).
7. According to the manufacturer's protocol, DualGlo luciferase assay kit (Promega E2920) was used to determine Firefly and renilla luciferase activity. The renilla luciferase activity measurement was utilized as a control for transfection efficiency, so the ERE Luciferase activity was normalized to the reading obtained for renilla luciferase activity.

5.3 Computational methods chapter 3

5.3.1 SID-V1 mutation profiling analysis

Raw reads were quality checked using FASTQC v.0.11.5 and trimmed with Trimmomatic[154]. Trimmed reads were aligned to the reference genome (hg19 assembly) with BWA[155] mem v.0.7.17 software with -v 1 -t 12 parameters. Aligned reads were converted to bam file then sorted and indexed using samtools v1.9. BAMs header were adjusted with processed with addOrReplaceReadGroups tool available in Picard suite v2.2.1 AddOrReplaceReadGroups setting GID=1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=1 parameters. The resulting files where then sorted and indexed. All duplicates in the files were removed using MarkDuplicates available in Picard suite v2.2.1. Data processing for mutation calling followed the GATK best Practises v.01-09-2018 which includes Indel Realignment and BaseRecalibration available in GATK4 v4.0.12. Somatic mutation calling was performed using Mutect2 available in GATK4[156] with default parameters. Normal/Blood bam files were used as reference input in all our analysis. SNPs annotated in 1000G_omni2.5.hg38.vcf were removed from the analysis and only the mutations tagged with "PASS" were retained.

5.3.2 Mutation calling post-processing

Mutations tagged with "PASS" were exploited for further down-processing. The allele frequency profile of each sample against the normal was evaluated by plotting the allele frequency score assigned as output by Mutect2 under the tag AF. All the AFs of each sample were plotted with a custom R script in a density plot with a violin shape using ggplot2.

Comparison of the mutation sharing was obtained with a custom bash script, and the AF trend between L1 and L2 were plot using GraphPad Prism v.6.

5.3.3 CTCF mutation exploratory analysis

The exploratory analysis of CTCF mutational landscape in different cancer was performed by exploiting the data freely available in the Catalog of Somatic Mutations COSMIC v65. All the mutations matching a window size of 100bps from the centre of a CTCF motif were retained and plot as 1bp resolution histogram. The number of the mutations at each bp was then normalised to 1000 detected mutation in the sample.

We analysed SNV profiles of 15-L1, 15- L2 and 15- R1 pointing the analysis to the CTCF investigation. CTCF mutations were selected by performing an intersection of all the "PASS" mutations list and SID-V1 CTCF specific panel using bedtools v2.27.1 intersect. All the mutations according to +- 6bps to the centre of the CTCF binding sites were plot using a custom R script.

To accurately predict the impact of the identified SRM within the CTCF core (+-60bps) we run the deltaSVM[157] algorithm using all the weighted matrix of each kmer available at http://www.beerlab.org/deltasvm/. The scores were plotted in function of their relative position using a custom R script.

The conservation profiles of all the 15 SRMs identified were obtained by precisely interrogating the conservation track of PhastCons[158] available in table browser section of UCSC Genome Browser named Cons.100.Verts. The PhastCons score of each bp in each SRM5 CTCF core. The data of each bps were then plotted with an R custom script.

5.3.4 CTCF ChIP-seq peak calling

Paired 75nt long reads were obtained on a HiSeq 2000 instrument (Illumina) and aligned using Bowtie2 in --local mode. IPs run against their respective cell type-specific input using MACS 2.1.1 with default settings (p-value<=1e-3). To visualize the raw profiles on Genome Browser, bedgraph files were converted to with bedGraphToBigWig v.4.

5.4 Brief experimental methods chapter 3

5.4.1 Targeted sequencing

Whole genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen). DNA quality and size were assessed using the Agilent Genomic DNA ScreenTape Assay (Agilent) run on the Agilent Tapestation 2200 system (Agilent) both before and after sonication. This ensured adequate sonication cycles and DNA fragment size for library preparation. Library preparation was performed using the NEBNext Ultra 2 DNA Library Prep Kit for Illumina (NEB) with custom unique 8 bp dual indices, designed by Dr Paolo Piazza, Head of Imperial BRC Genomics Facility, for each sample. Library molarity and fragment size were confirmed through the Agilent High Sensitivity D5000 ScreenTape Assay. Libraries were pooled and captured using a custom targeted capture panel, produced by Twist Biosciences, at a ratio of 1 ug of capture probes to 1.5 ug of pooled libraries. Post capture amplification was carried out using the KAPA HiFi HotStart ReadyMixPCR Kit. The post-capture library pool molarity and fragment size were again confirmed through the Agilent High Sensitivity D5000 ScreenTape Assay.

5.4.2 CTCF ChIP-seq

CTCF ChIP seq was performed according to what described in brief experimental methods chapter2. Magnetic beads (Dynabeads, Life technologies, 20 µl per ChIP) were coated with anti-CTCF antibody (Merck Millipore, 07-729, 4 µg per ChIP) through incubation at 4°C rotating for at least 6 hours.

5.4.3 CRISPR-Cas9 transfection

CRISPR gRNA vectors were co-transfected with a pCas9_GFP plasmid (Addgene plasmid # 44719) into MCF7 cells using the Amaxa Cell Line Nucleofector Kit V (Lonza, Switzerland). 2 x106 MCF7 cells were suspended in the Nucleofector Solution and a 1:1 ratio solution (2 µg: 2 µg) of the gRNA

and Cas9-GFP plasmids were added before transfer to a provided cuvette. The cuvette was placed in a 4D-Nucleofector System (Lonza, Basel, Switzerland) to electroporate the cells using program P-020. Cells were replated and left 6 hours before the media was replaced. 48 hours after transfection cells were detached using 1 mM EDTA (Ethylenediaminetetraacetic acid), centrifuged and resuspended in FACS (phosphate-buffered saline (PBS) containing 1% FCS and 5 mM EDTA). Cell suspensions were then sorted at the Imperial College London MRC Flow Cytometry Facility using GFP expression. Cells were first sorted into 96 well plates using single cell sorting.

5.4.4 CRISPR Validation Using Sanger Sequencing

CRISPR cells were lysed using a series of lysis buffers. First, 10 minutes rotating at 4°C in Lysis Buffer Cell lysates were sonicated (12 high cycles of 30 seconds on and 30 seconds off) to fracture the chromatin using a Biorupter Pico sonicator (Diagenode). DNA was then extracted using Phenol/ chloroform extraction; Phenol/chloroform was added to samples at a 1:1 ratio vortexed and centrifuged, following this DNA is precipitated out using 5 M NaCl and 100% ethanol and finally eluted in water. The DNA concentration was measured by a fluorometric quantitation assay using a Qubit 3.0 Fluorometer (Thermo Fischer Scientific) using its standard protocol. The area of interest was subsequently amplified through a Polymerase Chain Reaction (PCR) using 100 ng of DNA and the Phusion High-Fidelity PCR Master Mix with High Fidelity buffer (New England Biolabs, Massachusetts, USA). Cloning primers (forward and reverse) were designed for SRM5, CTCF7 and CTCF9 (Supplementary figure 1) to clone a ~300 bp fragment spanning the area. PCR cycles were set; denaturation 98°C (30 seconds), 30 cycles of 98°C (5 seconds), 60°C (30 seconds), 72°C (30 seconds), final extension 72°C (5 minutes). PCR products were cleaned up using a QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) as per the manufacturer's standard protocol and sent for DNA Sanger Sequencing at the Imperial College-MRC Genomics Centre.

5.4.5 RT q-PCR

Lysed CRISPR cells were homogenised using the QIAshredder columns (Qiagen, Hilden, Germany). RNeasy columns (Qiagen, Hilden, Germany) were used to isolate total RNA by following the manufacturers standard protocol. RNA was subsequently eluted in DEPC water, and concentration was quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, USA). In order to synthesise the complementary DNA (cDNA), 1µg of RNA was used with a High Capacity cDNA Reverse Transcription kit (Applied Biosystems Carlsbad, USA). The thermal cycler (Veriti, ThermoFisher Scientific) setup was as follows: 25°C (10 min), 37°C (120 min), 85°C (5 min). RT-qPCR was set up and performed in 96 well plates using 1x SYBR Green PCR Master Mix and primers for the relevant genes. The PCR was performed in a StepOne Plus Real-Time PCR System (Applied Biosystems, Carlsbad, USA) with the thermocycler setup as previously described. Relative mRNA levels expressed were calculated using the comparative threshold cycle (Ct) method with the 2-ΔΔCt method normalising against our housekeeping gene RPL38.

5.5 Computational methods chapter 4

5.5.1 ATAC-seq analysis

Single-end 50bp long reads were obtained on an HISeq 2000 instrument (Illumina) and aligned to Hg19 using Bowtie v1.1.2 with default settings. The aligned reads were passed to Samtools v1.9 to be converted in bam format the sorted and indexed. Single-end reads were pooled in 4 groups called MCF7 CD44+, MCF7 CD44-, LTED CD44+ and LTED CD44- using the “cat” command available in the Samtools suite. Peaks were called by MACS2 v2.1.1 with parameters ‘--nomodel --shift -100 --extsize 200 -B --SPMR --call-summits’. Peaks with a –log10(q-value) lower than 2.0 were discarded.

5.5.2 Differential peaks analysis

We performed 2 pairwise analysis: MCF7 CD44+ vs MCF7 CD44-, LTED CD44+ vs LTED CD44-. To capture meaningful biologic changes between the groups we used the Bioconductor package MMDiff[159] v1.2.1. We started our analysis by creating a set of consensus peaks from a region overlapping both our two datasets. We summarised peaks in bins corresponding to 30 bp and applied the 'Generalized Mover Distance (GMD)method to determine the distance score k between single peaks in a pair present in the two groups. We accepted differential scores over then k=1. To detect the statistical significance of the differential peaks, we applied the method proposed by ODIN[160] v0.4 It identifies differential peaks between 2 conditions and then counts the number of reads for each candidate peak performing a signal normalisation and applying statistical tests assuming differential count model like DESeq[161]. We performed differential peak detection and statistical test following the ODIN framework using the concatenated bam files and filtering differential peaks setting a threshold of q-value 10-3. The identification of common peaks between the 2 previously described procedures was carried out running intersect command available in bedtools v2.27.1. In this way, we were able to obtain a table of common peaks presenting associated GMD score and q-value.

5.5.3 Data normalisation and hierarchical clustering analysis

We created a 2x2 contingency table comprising ODIN peaks over thresholds of 10-3 and 10-9 and MMDiff peaks over the threshold of k=1 and k=3. The lower stringent q-value paired with the most stringent GMD score were used to provide evidence of the difference between the two groups by the raw reads belonging to the single replicates. Picard MarkDuplicates v2.2.1 was used to mark and remove duplicates from the Bam files. Then, we sorted and indexed the deduplicated bam files generated above. Deduplicated bam files were processed to count the number of aligned reads per genomic range per sample using Bedtools multicov from Bedtools suite. We, then, normalised each genomic range count giving the score X, where X is the number of mapped reads to the single genomic location per million divided by the length of the genomic coordinate per kb.

$$\frac{(Genomic\ range\ read\ count / Mapped\ reads\ ) x\ 10^6}{Genomic\ range\ length\ x\ 10^3}$$

To determine whether normalised read count differences observed between MCF7 CD44+ and MCF7 CD44- are stochastic or indicative of biological meaning, we developed an R script to build a heatmap showing differences between our groups. We centred the data to the row mean, fix the colour palette from the lowest to the highest value. Centred data were hierarchically clustered (Pearson distance, complete linkage) using the hclust package. Results were visualised using heatmap.2 available in the gplot R package.

5.5.4 Identification of Differential Pathways and Transcription-Factor-Bound Regions

We then computationally predicted the functions of the differential genomic coordinates with GREAT[162]. GREAT associates cis-regulatory element (in this case NFR) with a putative target gene(s) (in this case the Single nearest gene) and the uses annotation from various databases to associate the target gene(s) with a function. GREAT calculates the statistical enrichment for the associations with the given genomic coordinates and the annotations. We utilised GREAT to associate the genomic coordinate to the PANTHER[163] database to identify pathways and genes significantly enriched. For motif finding, we called the peak pileup of the above genomic coordinates considered, and extended the region to +- 10 bp. The 21bp sequences were input for HOMER package motif finder algorithm findMotifGenome.pl.

5.6 Brief experimental methods chapter 4

5.6.1 Cell lines

Cells were originated as previously described in Nguyen et al. Nat. Comm 2015. MCF7 cells were maintained in Dulbecco's modified Eagle's medium (DMEM) containing 10% fetal calf serum (FCS). Long-term oestrogen-deprived cells (LTED) were derived from MCF7 after six-month oestrogen

deprivation as previously described and were maintained in phenol-red free DMEM containing 10% charcoal-stripped fetal calf serum (SFCS). Both mediums were supplemented with 2 mM L-glutamine, 100 units/mL penicillin, 0.1 mg/mL. 10−8 M oestradiol (E2758 Sigma) was added routinely to MCF7. 10− tamoxifen (H7904 Sigma) and fulvestrant (I4409 Sigma) were added to tamoxifen-resistant (TAMR) and fulvestrant-resistant (FULVR) cells. Primary metastatic breast cancer cells were derived from pleural effusions of patients with metastatic breast cancers. The pleural effusion (PE) cells were maintained in DMEM containing 10% fetal calf serum (FCS) and 2 mM L-glutamine, 100 units/mL penicillin, 0.1 mg/mL.

5.6.2 Plasmids

pLVX-IRES-mCherry-puro lentiviral vector (Cambridge Bioscience, Cambridge, UK) was used to infect MCF7 and LTED cells. CD44 reporter GFP cells were established with CD44CR1-IRES-GFP-puro lentiviral vector (Tebu-Bioscience). Stable and polyclonal cell populations were established after puromycin selection (0.5 μg/ml). H2B-mCherry-puro lentiviral vector was used to infect stable CD44 reporter GFP cells. Stable and polyclonal cell populations were established after sorting.

5.6.3 FACS analysis

Cells were cultured to 70 to 80% confluence and detached from the cell culture flasks using EDTA. Cell pellets were obtained and washed with cold phosphate-buffered saline (PBS) containing 1% FCS, 1% penicillin/streptomycin, 5mM EDTA. All further steps were performed on ice and all centrifugation steps at 4°C. Fluorochrome-conjugated monoclonal antibodies against human CD44 (FITC, BD Pharmingen), CD24 (PE, BD Pharmingen), and their isotype controls were added to the cell suspension at concentrations recommended by the manufacturer (BD Biosciences) and incubated at 4 °C in the dark for 30 min. The labelled cells were washed in PBS and then were analysed on a FACSAria (BD Biosciences). Gating was set to relevant isotype control (IgG-FITC)-labelled cells for each cell line.

5.6.4 Neo-adjuvant patient selection

All clinical data from patients operated at the European Institute of Oncology (IEO) were prospectively entered in an Institutional data base. For the present study we retrieved data from patients with a neoadjuvant treatment from 1999 to 2014. We selected patients having had pre-surgical biopsy and surgery at IEO hospital, in order to have their sample analysed in the same laboratory. We randomly selected 20 patients with luminal tumours, treated by neoadjuvant hormone only (aromatase inhibitors), 10 responders and 10 non-responders.

5.6.5 FAIRE-qPCR

Formaldehyde-assisted isolation of regulatory elements (FAIRE) was performed as previously described with minor modifications. Briefly, chromatin was extracted and subjected to 5 rounds of phenol-chloroform extraction. The resulting DNA was purified using SPRI beads (Beckman-Coulter) and used for qPCR analysis. For differential chromatin accessibility analysis, cells were pre-treated for 3 days with an indicated dose of GSI, and then DNA was analysed using FAIRE-qPCR.

5.6.6 ATAC-seq

ATAC seq was performed as previously described by Buenrostro et al in 2013. Briefly, after sorting with CD44 using FACS Aria III (BD Biosciences), 5x104 cells of CD44high and CD44lowwere resuspended in 200ul of cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630), and immediately centrifuged at 500 xg for 10 minutes in 4oC. Then, the nuclei were resuspended in transposition reaction mix, including transposase (Illumina), and incubated at 37oC for 30 min. The DNA was purified using Qiagen MinElute Kit (Qiagen). The library preparation for next-generation sequencing was performed according to the manufacture's instruction (New England Labs).

Chapter 6

Discussion and Conclusions

Chapter2

The study described in chapter 2 applies the concept of epigenetic signals, such H3K27ac mark as a digital signal. The computational work I performed proposes that the epigenetic signal (i.e. “on or off”) can be employed as a qualitative measure for inferring phenotypical clonality to regulatory regions. This approach allowed the dissection of phenotypical heterogeneity of ERα + Breast Cancer followed by the first evidence of principles of evolutive activity of regulatory regions at different stages of the disease.

RI/SI metric relies on the fact that histone-based ChIP-seq correlates with the number of cells in the population carrying specific epigenetic information, similarly to what described by Buenrostro 2015 for single cell ATAC seq case[107]. Interesting patterns of activity have been identified, however the study described in this dissertation focused on the high SI – low RI enhancer (clonal regions). For example, further effort may be conducted to characterise the enhancers exhibiting high activity in private patients. In this context, early results show that this pool of enhancer lays in the so-called open sea region between 500kb to 1MB to the closest promoter. These enhancers are privately active but geographically close. It is tempting to hypothesise the existence of generally packaged DNA with the potential to be activated and thus to produce regulatory activity operating on genes distant at the scale of MBs. Chromosome conformation capture techniques would be used to define the targets of these privately active enhancers. Besides, a formal investigation of the motif characterising these enhancers may reveal some interesting protein players and contribute "super-"long-range DNA interactions.

The main applications of RI/SI method span between new data generation and publicly available datasets. Certainly, RI/SI metric can be efficiently applied to large publicly available datasets of histone and ATAC –seq experiments of bulk DNA. We provided strong evidence that clonal regulatory regions are also highly shared across patients. This concept can be translated to

characterise the regulatory pattern specific to cell tissues and eventually leveraged to identify patterns of contamination in bulk tissues. Theoretically, RI/SI when applied to a large dataset of normal tissues, which is currently available in Epigenomic Roadmap / ENCODE project, would contribute to classifying clusters of enhancers exhibiting distinct patterns of activity. High SI and low RI (high enrichment) across the same tissue type can potentially highlight cell-specific enhancers. High SI and low RI across all tissues would reveal housekeeping enhancers. High variable RI SI scores across the different tissue types would support the association of enhancers, highly dynamical and possibly specific, to active regulatory pathways. The identification of the cell-specific enhancers would support a tagging strategy of epigenetic activity in tumour studies. If correct, the epigenetic landscape of a bulk tumour would be dissected using a computational tool developed with the aim to assign a cell specificity of the enhancer activity. For example, immune and adipose infiltration is a common source of heterogeneity in breast cancer. Currently, to my knowledge, no methods are available to deconvolute phenotypic heterogeneity by assigning to each regulatory region the likelihood to be active cause of tissue contamination based on epigenetic information. This application may support further studies applied on publicly available datasets having the aim to maximise data processing with more accurate and precise results.

Further efforts contributed to the identification, and experimental validation of YY1 and SLC9A3R1 proteins. YY1 protein is described as a novel driver of BC oncogenesis, SLC9A3R1 as a critical protein implied in the arising and maintenance of drug resistance. We propose that YY1 acts as global co-activator involved in orchestrating the cancer epigenome in BC. Our data reveal that YY1 collaborates with ERα, as a co-activator by stabilising ERα binding and increasing transcription at a subset of enhancers specific to BC. Interestingly, ERα, FOXA1 and YY1 bind specifically to SLC9A3R1 enhancer. Tamoxifen or aromatase inhibitors do not antagonise SLC9A3R1 enhancer but gain its activity (patient and cell lines data) during cancer progression possibly because ERα and YY1 binding provides structural stability to his mechanism of looping and thus gene regulation.

Understanding the mechanism of action of YY1 is crucial to facilitate our view of gene regulation in healthy and cancer tissue. Weintraub et al. describe YY1 as a critical regulator of enhancer-promoter looping and thus a facilitator of gene expression[110]. Our data show a systematic binding of YY1 in a broad spectrum of enhancers and promoters in both patient tissues and cell lines. A formal computational analysis developed with the aim to identify significant covariates of promoter and enhancer enrichment coupled with YY1 ChIP-seq may outline YY1 involvement only in the maintenance of a subset of enhancer-promoter looping, probably the clonal ones. Also, adding transcriptomic data to this analysis may provide further insights into gene regulation. Recently, Furlong group described that enhancer-promoter loops are required but not sufficient to gene expression since loopings are established hours before the construction of mRNA in cell development[164]. I propose to monitor looping formation and gene expression coupled with YY1 binding essays to investigate whether YY1 is involved only transcriptionally productive looping. Plausibly, YY1 is generally crucial in looping, but it may drive transcriptionally productive looping only when part of a complex with FOXA1 and other factors.

Novel studies are also required to investigate the potential interplay of YY1 with specific epigenetic marks. In our work YY1 has been identified by deconvoluting data from H3K27ac mark. To gain more insights in the mechanism of action, exploring the interplay between different histone modification and YY1 during short and prolonged exposure of oestradiol may reveal novel regulatory mechanisms by which YY1 is recruited at specific loci. I hypothesise that YY1 binding may have an interplay with specific epigenetic markers. A screening of the most promising epigenetic markers involved in the mechanism followed by Chip-CO IP (YY1, epigenetic marker) might provide a foundation for further research.

In this study, we assessed the presence of genetic alteration previously associated with endocrine resistance to gain a bit of background on the potential synergy between epigenetic changes and more conventional genetic drivers such as ESR1. We have performed targeted capture sequencing using

the input material from the actual epigenetic analysis wherever we had sufficient DNA to build reliable libraries. The analysis revealed that our patients' tissues were not commonly affected by mutation to specific driver genes. In particular, only metastatic samples were affected by point mutation within the coding part of the gene ESR1. This data excluded the possibility of genetic background significantly affecting the epigenetic landscape throughout the entire analysis. However, we did not investigate the effect of ESR1 mutations in the epigenomic landscape of the metastatic patient carrying it. It would be feasible to perform a study which explores the differential ERα binding in patients affected by ESR1 mutation compared to patients with wild type ESR1. It is possible to speculate that the mutation of the ESR1 genes may potentially affect the ESR1 binding followed with an epigenetic rewiring with a new DNA accessibility at specific sites.

Unexpectedly, the motif analysis of enhancers showed that oestrogen responsive elements (EREs) motifs, which are supposed to be the predominant target of ERα were significantly enriched only in subclonal enhancers. The integration of in vivo ERα data from Ross-Innes et al. 2015 shows that clonal enhancers have high propensity to be bound by ERα although they do not enrich for ERE elements[114]. This discrepancy is still unexplained. Probably, different TFs collaborate with ERα inducing transcriptional activity in clonal enhancers without binding with known ERE elements.

Chapter 3

In chapter 3, I described the approach undertook to investigate NCSMs in ERα+BC. It is well known the difficulty to investigate the NCSM due to a lack of annotation and regulatory asset of the non-coding genome. SID-V1 and SID-V2 are applications developed with the aim to investigate non-coding somatic mutations in cancer tissues. SID-V1 covers regions matching active enhancers of ERα+ BC coupled with highly conserved CTCF regions. Targeted sequencing of 3 samples from a single patient (Golden patient) has been performed. The postprocessing analysis of the matched Blood- L1 -L2 and R1 resulted in the identification of a mutation (SRM5) hitting precisely the binding motif of a CTCF involved in TAD maintenance. Indeed, this CTCF has a structural role in a well-defined TAD border localised at the q25.1 of chromosome 17. Computational methods predicted that SRM5 mutation strongly affects the binding affinity of CTCF protein and consequently, potentially contribute to the stability of the chromatin conformation within the TAD. We modelled the phenotypical consequences of SRM5 by developing CRISPR-Cas9 approach to engineer ERα-positive breast cancer cells MCF7. To test if CTCF-SRM-chr17 can effectively alter CTCF binding, we generated three individual clones containing unique indels centring at the CTCF locus, and strikingly, each clone shows a unique and significant transcriptional pattern suggesting that targeting CTCF could induce "stochastic" transcriptional reprogramming. Interestingly, one of the deregulated genes is SLC9A3R1, a gene described in chapter 2 of this dissertation for being in the top one per cent of the highest predictors of ERα BC relapse.

SID-V1 has been further developed with a new version called SID-V2. The selection strategy of the regulatory regions covered by the latter represents a shift in the field which has never been applied in previous studies. Indeed, the main limitation of SID-V1 consists of the selection procedure applied to enhancer regions. The enhancers selection lies on the mutation calling performed on tumour samples at generally low depth (20X) without a healthy matching control but to the more general hg19 reference genome. We overcome this drawback by leveraging the epigenomic mapping

performed in the same tumour dataset (described in chapter 2 and used to build SID-V1) focusing on the accessible DNA regions significantly relevant to the oncogenesis and progression of ERα+ BC.

On the other side, the selection strategy of CTCF sites between SID-V1 and SID-V2 has not produced any significant shift. Indeed, SID-V2 includes almost the totality of SID-V1 CTCF sites. This incentivised us to endeavour to the analysis of only mutations occurring in CTCF sites from the sequenced tissues based on SID-V1. However, further exploitation of these data may potentially reveal novel targets for experimental validation and new hypothesis.

SID-V2 is explicitly built for the broad and in-depth investigation of the genetic layer of active enhancers in ERα+ BC. The results which will be obtained after sequencing the first dataset of sample coupled with the epigenomic profiling described in chapter 2 would consistently provide new paradigms for tumour progression and drug-resistance studies. Early results, (not described in this dissertation) obtained from a panel of patients matching blood- primary-metastatic tissues patients exposed to 1 drug endocrine therapy (Tamoxifen or AI) and subsequently processed with SID-V2 application revealed potential high impact genomic alterations strongly linked to epigenetic shifts. These results are anticipating the main study which will follow after a systematic quality check of the SID-V2 efficiency.

Indeed, it is known that not all the regions of the human genome show same capture and sequencing efficiency. Substantial portions of the human genome are intrinsically hard to assemble,which makes them hard to map to the reference genome. On top of this, problematic regions such as segmental duplication and highly repetitive regions pose a further challenge to the sequencing. All these concerns are being thoroughly evaluated and prompt for a final restructuring of SID-V2 with a new version, SID-V3.

Chapter 4

The study described in chapter 4 was designed to investigate the role of epigenetics in supporting phenotypical plasticity in breast cancer cells lines before and after oestrogen deprivation. We provide evidence with computational and experimental procedures that the majority of cancer cells are not plastic and efficiently cleared upon oestrogen deprivation. Nevertheless, subpopulations of cancer cells are highly plastic and become quiescent in the absence of oestrogen, to then resuming proliferation probably after epigenetic reprogramming. Notably, our data have identified a subpopulation of plastic cells characterised by several cancer-stem cell-like (CSCs) traits. Using this plastic subpopulation, we have uncovered a series of novel paradigms regarding cancer drug resistance using epigenomic approaches. We provide evidence that subpopulation of plastic cells responds to ET by stacking the cell cycle and maintaining specific activity within certain loci to resume proliferation and support tumour relapse. We have demonstrated that epigenetic reprogramming is unidirectional in MCF7 and involves the loss of regulatory potential, possibly mimicking what might happen during differentiation. More specifically, our data suggest that ELF1-Notch signalling axis plays a central role in the modulation of epigenetic locks. Notch signalling is a crucial feature of breast luminal progenitor cells, and as shown by the Magnani lab in the past, AI resistant cells have Notch dependencies[122]. Here we show that Notch inhibition disrupt plasticity locking cells into an ET sensitive subpopulation.

The main limitation of this study is the use of in vitro model which do not represent a reliable model to study subpopulation of cells due to the high chromatin instability and metabolism rewiring as showed by Cantor et al. in 2017 commonly occurring in cell lines[165]. Nevertheless, the computational approach used for this analysis relies on tools available in 2015 and a revision of the workflow may provide additional results. The results described in the chapter are highly reproducible, validated and represent an excellent achievement considering the immense amount of bias occurring in cell cultures. In addition, these results have been further corroborated using the principles applied for SI/RI methodology to MCF CD44+ and MCF CD44 – ATAC-seq available in the lab. Data from well

annotated patient tissue at different stages the disease would further ameliorate this investigation. Indeed, identifying small subpopulations of cells that, as described from our in vitro model, represent only 1-4% of the cell population, is an immense challenge in bulk studies.

Our results set the principles which can be used for future studies aiming to investigate epigenetic plasticity of tumour cells in cancer patient tissues. It is tempting to design a study in which cancer tissues data are exploited at high resolution with new technologies such as single-cell ATAC seq. Single cell transcriptomics and DNA accessibility essays would immensely improve this investigation by providing further insights into the quantitative characterisation of cell subpopulation phenotypes by outcoming the issue of tumour heterogeneity. Also, data generated with these approaches may actively contribute to the identification of features of plasticity in the transcriptomic or accessibility context beyond the described NOTCH pathway.

Intriguing data from single cell transcriptomics suggest the use of trajectory inference from different time points. This methodology with the availability of cancer data from matching sample at a different stage of the disease as described in the case of Golden patient in chapter 3 will highlight the dynamical pattern involved in cancer cell selection mixed with adaptation, as the results in chapter 4 suggest. Recently different assays have been proposed to model chromatin accessibility[166,167.] However, several challenges exist such as the inexistence of standard experimental procedure and data processing. Inferring the factors contributing to epistatic interaction among essential genes or regulatory regions will contribute to determining the underlying fitness landscape within cancer evolution occurs. Understanding this fitness landscape and the related cell response would be the ultimate goal of this investigation.

In conclusion, this study reveals fascinating patterns and principles of CSCs features in a subpopulation of cells marked by the protein CD44. Although the various experimental and computational limitation of studies in-vitro models, we have been able to identify and validate that epigenetic reprogramming involve loss of regulatory potential. Indeed, Notch inhibition disrupts

plastic cells into ET sensitive population. Taken together these results set the stage for cutting edges studies and give principles for future pre-clinical and clinical studies based on NOTCH inhibitors targeting plasticity as a strategy to delay the arising of drug resistance.

Conclusions

The evolutionary process of cancer is poorly understood. The high magnitude of inter and intra-tumour heterogeneity exhibited by patient tissues is a major limiting factor in cancer studies. The structural complexity of phenotypical and morphological characteristics of cancer specimens rely on underlying genetic and epigenetic mechanisms. Recent findings point the awareness on the non-coding portion of the genome and its interplay with the numerous epigenetic mechanisms mostly underexplored. New data shows how non-coding somatic mutations and alteration of epigenetic mechanisms may disrupt chromatin conformation heading to aberrant cell phenotypes. Both changes in genetic and epigenetic levels may be cause and consequence of each other. Understanding the causality of aberrant gene expression in cancer coupled with a systematic tracking of genetic and epigenetic changes occurring in subpopulations of cells from the carcinogenesis to the metastasis represents a powerful tool for prioritising therapy targets.

In this thesis, I have discussed a new experimental workflow which has the main aim to prioritise and link epigenetic and genetic alterations to the cell phenotype in the context of luminal breast cancer. This approach consists of deep sequencing technologies and novel bioinformatics framework followed by experimental validations. A new scoring system assigned to epigenetic features and the profiling of non-coding mutations at sub-clonal level allows the parsing of heterogeneity and the prioritisation of investigation of sites prone to be critical players in breast cancer. In addition, the availability of tumour samples at different stages of the disease allows the determination of genetic and epigenetic locus-specific changes during the disease development. This provides a precise

estimation of the features involved in the functional evolution during cancer progression and potentially set the stage for clinical studies in precision medicine.

References

1. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*. [Online] 2017;390(10100): 1151–1210. Available from: doi:10.1016/s0140-6736(17)32152-9

2. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries - Bray - 2018 - CA: A Cancer Journal for Clinicians - Wiley Online Library. [Online] Available from: https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21492

3. Global Burden of Disease Cancer Collaboration, G, Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncology*. [Online] 2016; Available from: doi:10.1001/jamaoncol.2016.5688

4. Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends—An Update. *Cancer Epidemiology and Prevention Biomarkers*. [Online] 2016;25(1): 16–27. Available from: doi:10.1158/1055-9965.epi-15-0578

5. Jha P, Peto R. Global Effects of Smoking, of Quitting, and of Taxing Tobacco. *The New England Journal of Medicine*. [Online] 2014;370(1): 60–68. Available from: doi:10.1056/nejmra1308383

6. DeSantis CE, Ma J, Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: A Cancer Journal for Clinicians*. [Online] 2017;67(6): 439–448. Available from: doi:10.3322/caac.21412

7. Hanahan D, Weinberg RA. The Hallmarks of Cancer. *Cell*. [Online] 2000;100(1): 57–70. Available from: doi:10.1016/s0092-8674(00)81683-9

8. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. [Online] 2011;144(5): 646–674. Available from: doi:10.1016/j.cell.2011.02.013

9. Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. *Science*. [Online] 2017;357(6348): eaal2380. Available from: doi:10.1126/science.aal2380

10. Carioli G, Malvezzi M, Rodriguez T, Bertuccio P, Negri E, Vecchia C. Trends and predictions

to 2020 in breast cancer mortality in Europe. *The Breast*. [Online] 2017;36: 89–95. Available from: doi:10.1016/j.breast.2017.06.003

11. Malvezzi M, Carioli G, Bertuccio P, Negri E, Vecchia LC. Relation between mortality trends of cardiovascular diseases and selected cancers in the European Union, in 1970–2017. Focus on cohort and period effects. *European journal of cancer (Oxford, England : 1990)*. [Online] 2018; Available from: doi:10.1016/j.ejca.2018.06.018

12. Easton DF, Pharoah P, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *The New England Journal of Medicine*. [Online] 2015;372(23): 2243–2257. Available from: doi:10.1056/nejmsr1501341

13. King M-C, Marks JH, Mandell JB. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science*. [Online] 2003;302(5645): 643–646. Available from: doi:10.1126/science.1088759

14. Bergkvist L, Adami H-O, Persson I, Hoover R, Schairer C. The Risk of Breast Cancer after Estrogen and Estrogen–Progestin Replacement. *The New England Journal of Medicine*. [Online] 1989;321(5): 293–297. Available from: doi:10.1056/nejm198908033210505

15.Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology*. [Online] 2012;13(11): 1141–1151. Available from: doi:10.1016/s1470-2045(12)70425-4

16. Pharoah P, Day NE, Duffy S, Easton DF, Ponder B. Family history and the risk of breast cancer: A systematic review and meta-analysis. *International Journal of Cancer*. [Online] 1997;71(5): 800–809. Available from: doi:10.1002/(sici)1097-0215(19970529)71:5<800::aid-ijc18>3.0.co;2-b

17. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *The Lancet*. [Online] 2001;358(9291): 1389–1399. Available from: doi:10.1016/s0140-6736(01)06524-2

18. Ford D, Easton D, ratton, Narod S, Goldgar D, Devilee P, et al. Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. *The American Journal of Human Genetics*. [Online] 1998;62(3): 676–689. Available from: doi:10.1086/301749

19. Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkäs K, Roberts J, et al. Breast-Cancer Risk in Families with Mutations in PALB2. *The New England Journal of Medicine*. [Online] 2014;371(6): 497–506. Available from: doi:10.1056/nejmoa1400382

20. Li J, Humphreys K, Heikkinen T, Aittomäki K, Blomqvist C, Pharoah PD, et al. A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Research and Treatment*. [Online] 2011;126(3): 717–727. Available from: doi:10.1007/s10549-010-1172-9

21. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C

identifies putative target genes at 33 breast cancer risk loci. *Nature Communications*. [Online] 2018;9(1): 1028. Available from: doi:10.1038/s41467-018-03411-9

22. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American journal of human genetics*. [Online] 2018;104(1): 21–34. Available from: doi:10.1016/j.ajhg.2018.11.002

23. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. [Online] 2017; Available from: doi:10.1038/nature24284

24. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature Genetics*. [Online] 2018;50(7): 968–978. Available from: doi:10.1038/s41588-018-0132-x

25. Vachon CM, Pankratz SV, Scott CG, Haeberle L, Ziv E, Jensen MR, et al. The Contributions of Breast Density and Common Genetic Variation to Breast Cancer Risk. *JNCI: Journal of the National Cancer Institute*. [Online] 2015;107(5). Available from: doi:10.1093/jnci/dju397

26. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic Density and the Risk and Detection of Breast Cancer. *The New England Journal of Medicine*. [Online] 2007;356(3): 227–236. Available from: doi:10.1056/nejmoa062790

27. Warburton D, Nicol C, Bredin S. Health benefits of physical activity: the evidence. *Canadian Medical Association Journal*. [Online] 2006;174(6): 801–809. Available from: doi:10.1503/cmaj.051351

28. Watson CJ, Khaled WT. Mammary development in the embryo and adult: a journey of morphogenesis and commitment. *Development*. [Online] 2008;135(6): 995–1003. Available from: doi:10.1242/dev.005439

29. Keymeulen A, Rocha A, Ousset M, Beck B, Bouvencourt G, Rock J, et al. Distinct stem cells contribute to mammary gland development and maintenance. *Nature*. [Online] 2011;479(7372): 189. Available from: doi:10.1038/nature10573

30. Siiteri P, MacDonald P. Placental estrogen biosynthesis during human pregnancy. *The Journal of clinical endocrinology and metabolism*. [Online] 1966;26(7): 751–761. Available from: doi:10.1210/jcem-26-7-751

31. SIN ER, MAHENDROO MS, MEANS GD, KILGORE MW, HINSHELWOOD MM, GRAHAM-LORENCE S, et al. Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis. *Endocrine reviews*. [Online] 1994;15(3): 342–355. Available from: doi:10.1210/edrv-15-3-342

32. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*. [Online] 2011;43(1): 27. Available from: doi:10.1038/ng.730

33. Carroll JS, Liu SX, Brodsky AS, Li W, Meyer CA, Szary AJ, et al. Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1. *Cell*. [Online] 2005;122(1): 33–43. Available from: doi:10.1016/j.cell.2005.05.008

34. Galien R, Garcia T. Estrogen receptor impairs interleukin-6 expression by preventing protein binding on the NF-κB site. *Nucleic Acids Research*. [Online] 1997;25(12): 2424–2429. Available from: doi:10.1093/nar/25.12.2424

35. Couse J, Lindzey J, Grandien K, Gustafsson J, Korach K. Tissue distribution and quantitative analysis of estrogen receptor-alpha (ERalpha) and estrogen receptor-beta (ERbeta) messenger ribonucleic acid in the wild-type and ERalpha-knockout mouse. *Endocrinology*. [Online] 1997;138(11): 4613–4621. Available from: doi:10.1210/en.138.11.4613

36. Speirs V, Skliris G, Burdall S, Carder P. Distinct expression patterns of ERα and ERβ in normal human mammary gland. *Journal of Clinical Pathology*. [Online] 2002;55(5): 371–374. Available from: doi:10.1136/jcp.55.5.371

37. Huang B, Omoto Y, Iwase H, Yamashita H, Toyama T, Coombes R, et al. Differential expression of estrogen receptor α, β1, and β2 in lobular and ductal breast cancer. *Proceedings of the National Academy of Sciences*. [Online] 2014;111(5): 1933–1938. Available from: doi:10.1073/pnas.1323719111

38. Krege JH, Hodgin JB, Couse JF, Enmark E, Warner M, Mahler JF, et al. Generation and reproductive phenotypes of mice lacking estrogen receptor β. *Proceedings of the National Academy of Sciences*. [Online] 1998;95(26): 15677–15682. Available from: doi:10.1073/pnas.95.26.15677

39. Shoker BS, Jarvis C, Clarke RB, Anderson E, Hewlett J, Davies M, et al. Estrogen Receptor-Positive Proliferating Cells in the Normal and Precancerous Breast. *The American Journal of Pathology*. [Online] 1999;155(6): 1811–1815. Available from: doi:10.1016/s0002-9440(10)65498-3

40. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*. [Online] 2011;5(1): 5–23. Available from: doi:10.1016/j.molonc.2010.11.003

41. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research*. [Online] 2010;12(5): 1–18. Available from: doi:10.1186/bcr2635

42. Li CI, Anderson BO, Daling JR, Moe RE. Trends in Incidence Rates of Invasive Lobular and Ductal Breast Carcinoma. *JAMA*. [Online] 2003;289(11): 1421–1424. Available from: doi:10.1001/jama.289.11.1421

43. ELSTON CW, ELLIS IO. pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*. [Online] 1991;19(5): 403–410. Available from: doi:10.1111/j.1365-2559.1991.tb00229.x

44. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. *Annals of Surgical Oncology*. [Online] 2010;17(6):

1471–1474. Available from: doi:10.1245/s10434-010-0985-4

45. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, et al. Twenty-Year Follow-up of a Randomized Study Comparing Breast-Conserving Surgery with Radical Mastectomy for Early Breast Cancer. *The New England Journal of Medicine*. [Online] 2002;347(16): 1227–1232. Available from: doi:10.1056/nejmoa020989

46.Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Darby S,McGale P, Taylor C,Arriagada R, Clarke M, Cutter D, Davies C, Ewertz M, Godwin J, Gray R, Pierce L, Whelan T, Wang Y, Peto R. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials. *The Lancet*. [Online] 2011;378(9804): 1707–1716. Available from: doi:10.1016/s0140-6736(11)61629-2

47. Whelan TJ, Olivotto IA, Parulekar WR, Ackerman I, Chua BH, Nabid A, et al. Regional Nodal Irradiation in Early-Stage Breast Cancer. *The New England Journal of Medicine*. [Online] 2015;373(4): 307–316. Available from: doi:10.1056/nejmoa1415340

48. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *The Lancet*. [Online] 2005;365(9472): 1687–1717. Available from: doi:10.1016/s0140-6736(05)66544-0

49. Jacquillat C, Weil M, Baillet F, Borel C, Auclerc G. Results of neoadjuvant chemotherapy and radiation therapy in the breast-conserving treatment of 250 patients with all stages of infiltrative breast cancer. *Cancer*. [Online] 1990; Available from: doi:10.1002/1097-0142(19900701)66:1%3c119::aid-cncr2820660122%3e3.0.co;2-3

50. Fisher B, Costantino JP, Wickerham LD, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *JNCI: Journal of the National Cancer Institute*. [Online] 1998;90(18): 1371–1388. Available from: doi:10.1093/jnci/90.18.1371

51. Jakesz R, Jonat W, Gnant M, Mittlboeck M, Greil R, Tausch C, et al. Switching of postmenopausal women with endocrine-responsive early breast cancer to anastrozole after 2 years' adjuvant tamoxifen: combined results of ABCSG trial 8 and ARNO 95 trial. *The Lancet*. [Online] 2005;366(9484): 455–462. Available from: doi:10.1016/s0140-6736(05)67059-6

52. Dombernowsky P, Smith I, Falkson G, Leonard R, Panasci L, Bellmunt J, et al. Letrozole, a new oral aromatase inhibitor for advanced breast cancer: double-blind randomized trial showing a dose effect and improved efficacy and tolerability compared with megestrol acetate. *Journal of Clinical Oncology*. [Online] 1998;16(2): 453–461. Available from: doi:10.1200/jco.1998.16.2.453

53. Coombes CR, Hall E, Gibson LJ, Paridaens R, Jassem J, Delozier T, et al. A Randomized Trial of Exemestane after Two to Three Years of Tamoxifen Therapy in Postmenopausal Women with Primary Breast Cancer. *The New England Journal of Medicine*. [Online] 2004;350(11): 1081–1092. Available from: doi:10.1056/nejmoa040331

54. Goss PE, Ingle JN, Pritchard KI, Robert NJ, Muss H, Gralow J, et al. Extending Aromatase-Inhibitor Adjuvant Therapy to 10 Years. *The New England Journal of Medicine*. [Online] 2016;375(3): 209–219. Available from: doi:10.1056/nejmoa1604700

55. (EBCTCG) E, wsett, Forbes J, Bradley R, Ingle J, Aihara T, et al. Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials. *The Lancet*. [Online] 2015;386(10001): 1341–1352. Available from: doi:10.1016/s0140-6736(15)61074-1

56. Cristofanilli M, Turner NC, Bondarenko I, Ro J, Im S-A, Masuda N, et al. Fulvestrant plus palbociclib versus fulvestrant plus placebo for treatment of hormone-receptor-positive, HER2-negative metastatic breast cancer that progressed on previous endocrine therapy (PALOMA-3): final analysis of the multicentre, double-blind, phase 3 randomised controlled trial. *The Lancet Oncology*. [Online] 2016;17(4): 425–439. Available from: doi:10.1016/s1470-2045(15)00613-0

57. Rugo HS, Rumble BR, Macrae E, Barton DL, Connolly H, Dickler MN, et al. Endocrine Therapy for Hormone Receptor–Positive Metastatic Breast Cancer: American Society of Clinical Oncology Guideline. *Journal of Clinical Oncology*. [Online] 2016;34(25): 3069–3103. Available from: doi:10.1200/jco.2016.67.1487

58. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490.

59. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips K-A, Mooij TM, Roos-Blom M-J, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*. [Online] 2017;317(23): 2402–2416. Available from: doi:10.1001/jama.2017.7112

60. Nik-Zainal S, Alexandrov L, Wedge DC, Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012;149.

61. Stephens PJ, Tarpey PS, Davies H, Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*. 2012;486.

62. Martincorena I, Roshan A, Gerstung M, Ellis P, Loo P, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*. 2015;348.

63. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. [Online] 2016;534(7605): 47. Available from: doi:10.1038/nature17676

64. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. [Online] 2012;489(7414): 57–74. Available from: doi:10.1038/nature11247

65. Khurana E, Fu Y, Colonna V, Mu X, Kang H, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013;342.

66. Rheinbay E, Parasuraman P, Grimsby J, Tiao G, Engreitz JM, Kim J, et al. Recurrent and functional regulatory mutations in breast cancer. *Nature*. [Online] 2017; Available from: doi:10.1038/nature22992

67. D′Antonio M, Weghorn D, D′Antonio-Chronowska A, Coulet F, Olson KM, DeBoever C, et al. Identifying DNase I hypersensitive sites as driver distal regulatory elements in breast cancer. *Nature Communications*. [Online] 2017;8(1): 436. Available from: doi:10.1038/s41467-017-00100-x

68. Caravagna G, Giarratano Y, Ramazzotti D, Tomlinson I, Graham TA, Sanguinetti G, et al. Detecting repeated cancer evolution from multi-region tumour sequencing data. *Nature Methods*. [Online] 2018;15(9): 707–714. Available from: doi:10.1038/s41592-018-0108-x

69. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee T, et al. Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. *Molecular cell*. [Online] 2015;58(2): 362–370. Available from: doi:10.1016/j.molcel.2015.02.014

70. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome Research*. [Online] 2017;27(5): 757–767. Available from: doi:10.1101/gr.214874.116

71. Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, Jaenisch R, et al. Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell stem cell*. 2010;7.

72. Struhl K. Histone acetylation and transcriptional regulatory mechanisms. *Genes & Development*. [Online] 1998;12(5): 599–606. Available from: doi:10.1101/gad.12.5.599

73. Gosselin D, Link VM, Romanoski CE, Fonseca GJ, Eichenfield DZ, Spann NJ, et al. Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. *Cell*. 2014;159.

74. Serandour A, Mohammed H, Miremadi A, Mulder K, Carroll J. TRPS1 regulates oestrogen receptor binding and histone acetylation at enhancers. *Oncogene*. [Online] 2018; 1–11. Available from: doi:10.1038/s41388-018-0312-2

75. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011;473.

76. Nguyen VT, Barozzi I, Faronato M, Lombardo Y, Steel JH, Patel N, et al. Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. *Nature Communications*. [Online] 2015;6: 10044. Available from: doi:10.1038/ncomms10044

77. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*. 2016;351.

78. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. *Nature*. 2009;461.

79. Magnani L, Eeckhoute J, Lupien M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics*. [Online] 2011;27(11): 465–474. Available from:

doi:10.1016/j.tig.2011.07.002

80. Liu Z, Merkurjev D, Yang F, Li W, Oh S, Friedman MJ, et al. Enhancer Activation Requires trans-Recruitment of a Mega Transcription Factor Complex. *Cell*. [Online] 2014;159(2): 358–373. Available from: doi:10.1016/j.cell.2014.08.027

81. Lieberman-Aiden E, Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*. 2009;326.

82. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*. [Online] 2010;38(4): 576–589. Available from: doi:10.1016/j.molcel.2010.05.004

83. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. *Cell*. [Online] 2015;160(3): 554–566. Available from: doi:10.1016/j.cell.2015.01.006

84. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*. [Online] 2015;161(5): 1012–1025. Available from: doi:10.1016/j.cell.2015.04.004

85. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485.

86. Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Network T, et al. A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell*. [Online] 2018;173(2): 386-399.e12. Available from: doi:10.1016/j.cell.2018.03.027

87. Karnuta JM, Scacheri PC. OUP accepted manuscript. *Human Molecular Genetics*. [Online] 2018;27(R2): R219–R227. Available from: doi:10.1093/hmg/ddy167

88. Bahr C, von Paleske L, Uslu VV, Remeseiro S, Takayama N, Ng SW, et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature*. [Online] 2018;553(7689): 515. Available from: doi:10.1038/nature25193

89. Topalian SL, Taube JM, Anders RA, Pardoll DM. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nature Reviews Cancer*. [Online] 2016;16(5): 275–287. Available from: doi:10.1038/nrc.2016.36

90. Map of disease gene regulation by noncoding DNA. *Science-Business eXchange*. [Online] 2012;5(36): 969–969. Available from: doi:10.1038/scibx.2012.969

91. Cowper-Sal·lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, et al. Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics*. [Online] 2012;44(11): 1191. Available from: doi:10.1038/ng.2416

92. Katainen R, Dave K, Pitkänen E, Palin K, Kivioja T, Välimäki N, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*. [Online] 2015;47(7): 818–821. Available from: doi:10.1038/ng.3335

93. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*. [Online] 2015;162(4): 900–910. Available from: doi:10.1016/j.cell.2015.07.038

94. Northcott PA, Lee C, Zichner T, ütz A, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*. 2014;511.

95. Drier Y, Cotton MJ, Williamson KE, Gillespie SM, Ryan RJ, Kluk MJ, et al. An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. *Nature Genetics*. [Online] 2016;48(3): 265–272. Available from: doi:10.1038/ng.3502

96. Collins FS, Varmus H. A New Initiative on Precision Medicine. *The New England Journal of Medicine*. [Online] 2015;372(9): 793–795. Available from: doi:10.1056/nejmp1500523

97. McGranahan N, Swanton C. Clonal Heterogeneity and Tumour Evolution: Past, Present, and the Future. *Cell*. [Online] 2017;168(4): 613–628. Available from: doi:10.1016/j.cell.2017.01.018

98. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumour evolution across cancer types. *Nature Genetics*. [Online] 2016;48(3): 238–244. Available from: doi:10.1038/ng.3489

99. Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, et al. Extremely high genetic diversity in a single tumour points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences*. 2015;112.

100. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*. 2015;347.

101. Domcke S, Bardet A, Ginno P, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*. [Online] 2015;528(7583): 575. Available from: doi:10.1038/nature16462

102. Tang W, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. *Cell*. [Online] 2015;161(6): 1453–1467. Available from: doi:10.1016/j.cell.2015.04.053

103. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. [Online] 2015;523(7561): 486–490. Available from: doi:10.1038/nature14590

104. Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumours. *Proceedings of the National Academy of Sciences*. [Online] 2002;99(20): 12963–12968. Available from: doi:10.1073/pnas.162471999

105. Chin S-F, Santonja A, Grzelak M, Ahn S, Sammut S-J, Clifford H, et al. Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers. *Experimental and molecular pathology*. [Online] 2018;104(3): 161–169. Available from: doi:10.1016/j.yexmp.2018.03.006

106. Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, Giuffre A, et al. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biology*. [Online] 2009;10(10): R116. Available from: doi:10.1186/gb-2009-10-10-r116

107. Atlasi Y, Stunnenberg HG. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*. [Online] 2017;18(11): 643–658. Available from: doi:10.1038/nrg.2017.57

108. Brown JL, Mucci D, Whiteley M, Dirksen M-L, Kassis JA. The Drosophila Polycomb Group Gene pleiohomeotic Encodes a DNA Binding Protein with Homology to the Transcription Factor YY1. *Molecular Cell*. [Online] 1998;1(7): 1057–1064. Available from: doi:10.1016/s1097-2765(00)80106-9

109. Vella P, Barozzi I, Cuomo A, Bonaldi T, Pasini D. Yin Yang 1 extends the Myc-related transcription factors network in embryonic stem cells. *Nucleic Acids Research*. [Online] 2012;40(8): 3403–3418. Available from: doi:10.1093/nar/gkr1290

110. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*. [Online] 2017;171(7): 1573-1588.e28. Available from: doi:10.1016/j.cell.2017.11.008

111. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology*. [Online] 2010;28(12): 1248. Available from: doi:10.1038/nbt1210-1248

112. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*. [Online] 2015;16(1): 22. Available from: doi:10.1186/s13059-014-0560-6

113. Zhou X, Li D, Zhang B, Lowdon RF, Rockweiler NB, Sears RL, et al. Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nature Biotechnology*. [Online] 2015;33(4): 345–346. Available from: doi:10.1038/nbt.3158

114. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali RH, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. [Online] 2012;481(7381): 389–393. Available from: doi:10.1038/nature10730

115. Chan CW, Gu Z, Bieg M, Eils R, Herrmann C. Impact of cancer mutational signatures on transcription factor motifs in the human genome. *bioRxiv*. [Online] 2018; 422386. Available from: doi:10.1101/422386

116. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive

detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*. [Online] 2013;31(3): 213–219. Available from: doi:10.1038/nbt.2514

117. Mani SA, Guo W, Liao M-J, Eaton E, Ayyanan A, Zhou AY, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell*. 2008;133.

118. Shibue T, Weinberg RA. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nature Reviews Clinical Oncology*. [Online] 2017;14(10): 611–629. Available from: doi:10.1038/nrclinonc.2017.44

119. Tam W, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature Medicine*. [Online] 2013;19(11): 1438–1449. Available from: doi:10.1038/nm.3336

120. Todaro M, Francipane M, Medema J, Stassi G. Colon Cancer Stem Cells: Promise of Targeted Therapy. *Gastroenterology*. [Online] 2010;138(6): 2151–2162. Available from: doi:10.1053/j.gastro.2009.12.063

121. Takaishi S, Okumura T, Tu S, Wang S, Shibata W, Vigneshwaran R, et al. Identification of Gastric Cancer Stem Cells Using the Cell Surface Marker CD44. *STEM CELLS*. [Online] 2009;27(5): 1006–1020. Available from: doi:10.1002/stem.30

122. Lombardo Y, Filipović A, Molyneux G, Periyasamy M, Giamas G, Hu Y, et al. Nicastrin regulates breast cancer stem cell properties and tumour growth in vitro and in vivo. Proceedings of the National Academy of Sciences. [Online] 2012;109(41): 16558–16563. Available from: doi:10.1073/pnas.1206268109

122. Magnani L, Stoeck A, Zhang X, Lánczky A, Mirabella AC, Wang T-L, et al. Genome-wide reprogramming of the chromatin landscape underlies endocrine therapy resistance in breast cancer. Proceedings of the National Academy of Sciences. [Online] 2013;110(16): E1490–E1499. Available from: doi:10.1073/pnas.1219992110

124. Nguyen VT, Barozzi I, Faronato M, Lombardo Y, Steel JH, Patel N, et al. Differential epigenetic reprogramming in response to specific endocrine therapies promotes cholesterol biosynthesis and cellular invasion. Nature Communications. [Online] 2015;6: 10044. Available from: doi:10.1038/ncomms10044

125. Shaw T, Martin P. Epigenetic reprogramming during wound healing: loss of polycomb-mediated silencing may enable upregulation of repair genes. EMBO reports. [Online] 2009;10(8): 881–886. Available from: doi:10.1038/embor.2009.102

126. Karnoub AE, Dash AB, Vo AP, Sullivan A, Brooks MW, Bell GW, et al. Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. Nature. [Online] 2007;449(7162): 557. Available from: doi:10.1038/nature06188

127. Jia L, Zhou Z, Liang H, Wu J, Shi P, Li F, et al. KLF5 promotes breast cancer proliferation, migration and invasion in part by upregulating the transcription of TNFAIP2. Oncogene. [Online] 2015;35(16): 2040–2051. Available from: doi:10.1038/onc.2015.263

128. Andersson ER, Lendahl U. Therapeutic modulation of Notch signalling — are we there yet? Nature Reviews Drug Discovery. [Online] 2014;13(5): nrd4252. Available from: doi:10.1038/nrd4252

129. Zhu TS, Costello MA, Talsma CE, Flack CG, Crowley JG, Hamm LL, et al. Endothelial Cells Create a Stem Cell Niche in Glioblastoma by Providing NOTCH Ligands That Nurture Self-Renewal of Cancer Stem-Like Cells. Cancer Research. [Online] 2011;71(18): 6061–6072. Available from: doi:10.1158/0008-5472.can-10-4269

130. Stylianou S, Clarke RB, Brennan K. Aberrant Activation of Notch Signaling in Human Breast Cancer. Cancer Research. [Online] 2006;66(3): 1517–1525. Available from: doi:10.1158/0008-5472.can-05-3054

131. Singh A, Settleman J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. Oncogene. [Online] 2010;29(34): 4741. Available from: doi:10.1038/onc.2010.215

132. Bioinformatics B. FastQC. 2014;

133. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4): 357–359.

134. Langmead B. Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics. [Online] 2010;Chapter 11: Unit 11.7. Available from: doi:10.1002/0471250953.bi1107s32

135. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. [Online] 2008;9(9): R137. Available from: doi:10.1186/gb-2008-9-9-r137

136. Consortium E. An integrated encyclopedia of DNA elements in the human genome. Nature. [Online] 2012;489(7414): 57–74. Available from: doi:10.1038/nature11247

137. Consortium R, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. [Online] 2015;518(7539): 317. Available from: doi:10.1038/nature14248

138. Zhao H, Sun Z, Wang J, Huang H, Kocher J-PP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics (Oxford, England). [Online] 2014;30(7): 1006–1007. Available from: doi:10.1093/bioinformatics/btt730

139. Jee J, Rozowsky J, Yip KY, Lochovsky L, Bjornson R, Zhong G, et al. ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics. [Online] 2011;27(8): 1152–1154. Available from: doi:10.1093/bioinformatics/btr092

140. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.

Bioinformatics (Oxford, England). [Online] 2010;26(6): 841–842. Available from: doi:10.1093/bioinformatics/btq033

141. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. Nucleic acids research. [Online] 2017;45(D1): D626–D634. Available from: doi:10.1093/nar/gkw1134

142. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. [Online] 2012;22(9): 1760–1774. Available from: doi:10.1101/gr.135350.111

143. voom: precision weights unlock linear model analysis tools for RNA-seq read counts | Genome Biology | Full Text. [Online] Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29

144. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. [Online] 2004;5(10): R80. Available from: doi:10.1186/gb-2004-5-10-r80

145. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell. [Online] 2010;38(4): 576–589. Available from: doi:10.1016/j.molcel.2010.05.004

146. Valero-Mora PM. ggplot2:Elegant Graphics for Data Analysis. Journal of Statistical Software. [Online] 2010; Available from: doi:10.18637/jss.v035.b01

147. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Research. [Online] 2013; e201–e201. Available from: doi:10.1093/nar/gkt850

148. Sie D, Scheinin I, Lieshout S, Cordes M, Pinkel D, Albertson DG, et al. QDNAseq: A bioinformatics pipeline for DNA copy number analysis from shallow whole genome sequencing with noise levels near the probabilistic lower limit imposed by read counting. Clinical Cancer Research. 2016;22.

149. Talevich E, Shain HA, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Computational Biology. [Online] 2016;12(4): e1004873. Available from: doi:10.1371/journal.pcbi.1004873

150. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. [Online] 2000;16(6): 276–277. Available from: doi:10.1016/s0168-9525(00)02024-2

151. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. [Online] 2010;20(9): 1297–1303. Available from: doi:10.1101/gr.107524.110

152. COSMIC: somatic cancer genetics at high-resolution | Nucleic Acids Research | Oxford Academic. [Online] Available from: https://academic.oup.com/nar/article/45/D1/D777/2605743

153. Shihab HA, Gough J, Mort M, Cooper DN, Ny I, Gaunt TR. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. Human Genomics. [Online] 2014;8(1): 11. Available from: doi:10.1186/1479-7364-8-11

154. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. 2014;30.

155. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). [Online] 2009;25(14): 1754–1760. Available from: doi:10.1093/bioinformatics/btp324

156. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. [Online] 2013;31(3): 213–219. Available from: doi:10.1038/nbt.2514

157. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nature Genetics. [Online] 2015;47(8): 955–961. Available from: doi:10.1038/ng.3331

158. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. [Online] 2005;15(8): 1034–1050. Available from: doi:10.1101/gr.3715005

159. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. BMC Genomics. [Online] 2013;14(1): 826. Available from: doi:10.1186/1471-2164-14-826

160. Allhoff M, Seré K, Chauvistré H, Lin Q, Zenke M, Costa IG. Detecting differential peaks in ChIP-seq signals with ODIN. Bioinformatics. [Online] 2014;30(24): 3467–3475. Available from: doi:10.1093/bioinformatics/btu722

161. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. [Online] 2010;11(10): R106. Available from: doi:10.1186/gb-2010-11-10-r106

162. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology. [Online] 2010;28(5): 495–501. Available from: doi:10.1038/nbt.1630

163. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research. [Online]. Available from: doi:10.1093/nar/gky1038

164.Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature.

[Online] 2014;512(7512): 96. Available from: doi:10.1038/nature13417

165.Cantor JR, Abu-Remaileh M, Kanarek N, Freinkman E, Gao X, Louissaint A, et al. Physiologic Medium Rewires Cellular Metabolism and Reveals Uric Acid as an Endogenous Inhibitor of UMP Synthase. Cell. [Online] 2017;169(2): 258-272.e17. Available from: doi:10.1016/j.cell.2017.03.023

166.Corces RM, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. [Online] 2018;362(6413): eaav1898. Available from: doi:10.1126/science.aav1898

167.Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. [Online] 2018;361(6409): eaau0730. Available from: doi:10.1126/science.aau0730

Appendix