



Characterisation of chromatin modifiers in endometrial cancer

Katarzyna Zofia Kedzierska
Brasenose College, University of Oxford



Thesis submitted for the degree of
Doctor of Philosophy
in Genomic Medicine and Statistics

March 26, 2024

Acknowledgements

Firstly, I would like to express my gratitude to my supervisors – David Church and Dan Woodcock, for their unwavering guidance, constructive feedback, and constant encouragement throughout my DPhil. My appreciation extends to the Wellcome Trust for their support and for providing the resources that allowed me to develop as a researcher.

Pursuing a DPhil amidst a pandemic proved particularly challenging. I am sincerely grateful to everyone who helped maintain my sanity through these trying times.

Matt, sharing the DPhil experience with you has made it so much better. I am grateful for our engaging scientific discussions during lunch walks in South Park, your insightful advice, and your empathetic ear when I needed to vent without the fear of judgment.

Thank you to Matt, Yannick, Chirstos, Luciana, Vassi, Laura, and Tijana for helping with the experimental sides of things and proving the power of collaboration. I also wish to thank the Church group members and the WHG for their support. Thank you to Ruth and Mark for supporting my *custom* IT requests and to John for helping me advance my Linux skills.

A heartfelt thank you goes to the wonderful people I met in Oxford who made this journey truly memorable, including Andrew, Natalia, Máire, Maria, and many more. To the GMS cohorts for their useful advice and for fostering a sense of community, the GEL users group in Oxford for commiserating over our shared experiences in GEL RE, and the R users for indulging our shared passion for R. To the lovely people I had the privilege of collaborating at the Novo Nordisk Research Center in Oxford.

To my dear friends Anamaria, Gienio, German, Maja, and Andrey. It has been an absolute privilege to get to know you. Our discussions about the big and small stuff were exactly what I needed, especially during trying times. I love that we can discuss the big world events and the difference between bioinformatics and computational biology with the same passion. I would also like to thank the NGSchool Society - seeing the impact of our work in organizing the summer schools, training, and building the community has brought me joy and fulfillment. And to all my friends outside of science for proving that there truly exists a world outside of academia.

My sincere gratitude goes out to everyone whom I guilted into providing feedback on my writing. Your input has been truly appreciated.

To my Mum, whose unwavering support has shaped me into who I am today, and to my brother, who has consistently cheered me on and teased me throughout this DPhil and our lives. Dziękuję.

To Kaspar - aitäh, mu kallis! I wouldn't have made it without your support. Meeting you was the best thing that happened to me, and I will forever be grateful for the events that led me here.

Lastly, to you, dear Reader. Here, I present the culmination of four and a half years of my DPhil journey. It took me over 483 hours, 6 minutes, and 51 seconds to write this thesis, and I sincerely thank you for taking the time to read it.

Cancer research would not be where it is today without the invaluable contributions of patients who selflessly provide their samples to advance our understanding of the disease. I want to express my heartfelt gratitude to the participants who generously contributed their samples to the 100,000 Genomes Genomics England (GEL) and the Cancer Genome Atlas (TCGA) projects. It is my sincere hope that my work, using those precious samples, contributes to the collective endeavor of advancing our understanding of endometrial cancer.

Abstract

Chromatin organization is a critical regulator of gene expression and cell phenotype, and is frequently dysregulated in cancer. Endometrial cancer (EC) is the most common gynecological malignancy, and casues significant morbidity and mortality. EC is notable for recurrent alterations in chromatin including the *ARID1A* gene – a key component of the SWI/SNF remodeling complex – has emerged as a prevalent driver in EC, along with other remodelers such as *CHD4* and *BCOR*. However, a systematic analysis of chromatin modifier alterations and their functional consequences in EC has not been done.

This thesis presents a comprehensive investigation of genomic alterations in chromatin modifiers using whole genome sequencing (WGS) data from the Genomics England (GEL) 100,000 Genomes Project, the largest EC cohort to date. I demonstrate that while mutational processes vary across molecular subtypes, numerous chromatin modifiers are consistently altered across all subtypes. These genomic alterations frequently occur in different subunits of the same complex, such as alterations in *CHD3*, *CHD4* and *MBD3*, subunits of the ATP-dependent chromatin remodeling complex NuRD. Additionally, I examine the correlation between driver mutations and patient survival, revealing that mutations in *PBRM1* and *CHD4* are associated with an increased risk of death after accounting for age, molecular subtype, and tumor mutation and copy number alteration burden.

To complement the correlative analysis, I employ CRISPR-Cas9 gene editing to study the functional consequences of perturbations in selected chromatin modifiers (*ARID1A*, *ARID1B*, *ARID5B*, *EP300*, *KMT2C*, and *SETD1B*) in normal and malignant endometrial cells using transcriptomic and chromatin accessibility data. Furthermore, I explore the implications of the N1459S *BCOR* mutation, a hotspot mutation near-unique to EC.

Considering the frequent occurrence of *ARID1A* mutations in malignant tissues and their absence in normal endometrium, I investigate tumor heterogeneity in endometrial cancer. I discuss the limitations of current methodologies and propose a deep learning approach to uncover the hidden evolutionary trajectories. With additional research, this approach could potentially facilitate understanding the sequence in which driver alterations occur.

In summary, this work presents a resource for investigating chromatin organization in EC. The functional analyses using gene editing techniques confirm that EC-associated drivers disrupt essential cellular processes involved in oncogenesis. By providing the first systematic correlative and functional analyses of chromatin modifiers in EC, this thesis offers novel insights into EC biology.

Preface

This thesis represents an exploration into the characterization of chromatin modifiers in endometrial cancer, undertaken at the Wellcome Centre for Human Genetics, under the guidance and supervision of David Church and Dan Woodcock at the University of Oxford. The body of work presented here is a testament to both individual effort and collaborative achievement, reflecting extensive research carried out between October 2018 and April 2023.

The research conducted is original and builds upon the foundations laid by numerous studies in the field, with due acknowledgment of these works through appropriate citations. Although no part of this thesis has been previously submitted for a degree, the collaborative nature of scientific inquiry has led to contributions that extend beyond the contents of this document.

Significant portions of this research were conducted in collaboration with the Endometrial Cancer Genomics England Clinical Interpretation Partnership (GEL EC) consortium, highlighting a concerted effort to analyze the GEL EC cohort. My role within this consortium spanned leading the research on alterations of chromatin modifiers, survival analysis, and heterogeneity analysis. I also generated visualizations for several analyses, including all visualizations included in this thesis. The full list of contributors to the GEL EC consortium is included in the Chapter 2.

The study of perturbations of chromatin modifiers described in the Chapter 3 was conducted in close collaboration with the members of the Church group at the Wellcome Center for Human Genetics. I designed the experiments and performed computational analyses for the study of chromatin modifiers, relying on the collaboration with the experimental researchers in the group to carry out the experimental work. The postdoc in the group, Yannick Comoglio, designed and led the N1459S *BCOR* study for which I performed the computational analyses. A detailed description of the contributions is provided in the Contributions section of the Chapter 3.

The final results chapter of this thesis, focusing on the Machine Learning modeling of heterogeneity and evolution of cancer, is the culmination of my independent research under the guidance and supervision of Dan Woodcock and David Church.

Throughout my degree, I have additionally contributed to the following publications that are not covered in this thesis:

- M Brown, A Leon, **K Kedzierska**, C Moore, HL Belnoue-Davis, S Flach, JP Lydon, FJ DeMayo, A Lewis, T Bosse, I Tomlinson, DN Church; Functional analysis reveals driver cooperativity and novel mechanisms in endometrial carcinogenesis; *EMBO Molecular Medicine* 15 (10), e17094

- MA Glaire, NAJ Ryan, ME Ijsselsteijn, **K Kedzierska**, S Obolenski, R Ali, EJ Crosbie, T Bosse, N Fcc De Miranda, DN Church; Discordant prognosis of mismatch repair deficiency in colorectal and endometrial cancer reflects variation in

antitumour immune response and immune escape; *The Journal of pathology* 257 (3), 340-351

- H Russell, **K Kedzierska**, DD Buchanan, R Thomas, E Tham, M Mints, A Keranen, GG Giles, MC Southey, RL Milne, I Tomlinson, D Church, AB Spurdle, TA O'Mara, A Lewis; The MLH1 polymorphism rs1800734 and risk of endometrial cancer with microsatellite instability; *Clinical Epigenetics* 13

- N Horeweg, M de Bruyn, RA Nout, E Stelloo, **K Kedzierska**, A León-Castillo, A Plat, KD Mertz, M Osse, IM Jürgenliemk-Schulz, LCHW Lutgens, JJ Jobsen, EM van der Steen-Banasik, VT Smit, CL Creutzberg, T Bosse, HW Nijman, VH Koelzer, DN Church; Prognostic Integrated Image-Based Immune and Molecular Profiling in Early-Stage Endometrial Cancer; *Cancer Immunology Research* 8 (12), 1508-1519

Contents

1 Introduction **1**
- 1.1 Endometrial cancer . . . 2
  - 1.1.1 Pathological classification . . . 5
- 1.2 Molecular features of EC . . . 7
  - 1.2.1 Heritability in endometrial cancer . . . 7
  - 1.2.2 Somatic mutations in normal endometrium . . . 8
  - 1.2.3 Role of mutations in endometrial cancer . . . 10
  - 1.2.4 Molecular subtypes of endometrial cancer . . . 12
  - 1.2.5 Molecular landscape of endometrial cancer . . . 15
- 1.3 Cancer driver mutations and genes . . . 19
  - 1.3.1 Driver and passenger alterations . . . 19
  - 1.3.2 Detection of positively selected genes . . . 20
  - 1.3.3 An overview of computational methods used for driver gene identification . . . 21
- 1.4 Packaging of DNA and chromatin organization . . . 24
  - 1.4.1 Nucleosome - a core structural element of chromatin . . . 24
  - 1.4.2 Regulation of chromatin accessibility . . . 25
- 1.5 Chromatin organization in cancer . . . 32
  - 1.5.1 DNA methylation . . . 33
  - 1.5.2 SWI/SNF complex . . . 33
  - 1.5.3 NuRD complex . . . 34
  - 1.5.4 Polycomb repressive complex . . . 34
  - 1.5.5 Dysregulation of chromatin in cancer initiation and progression . . . 35
- 1.6 Heterogeneity and cancer evolution . . . 35
  - 1.6.1 Tumor heterogeneity . . . 35
  - 1.6.2 Cancer evolution . . . 36
- 1.7 Motivation of this study . . . 38

2 Whole genome and chromatin landscape in endometrial cancer **41**
- 2.1 Background . . . 41
  - 2.1.1 Molecular subtypes of Endometrial Cancer . . . 42
  - 2.1.2 Molecular landscape of the coding EC genome . . . 44
  - 2.1.3 Chromatin modifiers in Endometrial Cancer . . . 44
  - 2.1.4 Unanswered questions . . . 46
- 2.2 Introduction . . . 47
  - 2.2.1 Genomics England endometrial cancer cohort . . . 47
  - 2.2.2 GEL EC cohort - overview . . . 48
  - 2.2.3 Contributions . . . 49

2.3 Results . . . 49
2.3.1 Molecular EC subgroups in the GEL EC cohort . . . 49
2.3.2 Chromatin modifiers present among genes identified as drivers . . . 65
2.3.3 Chromatin remodeling pathways are recurrently altered in endometrial cancer . . . 77
2.3.4 Relationship between molecular alterations and overall survival . . . 81
2.3.5 Temporal Heterogeneity . . . 90
2.4 Discussion . . . 96
2.5 Methods . . . 102
2.5.1 Genomics England data . . . 102
2.5.2 TCGA somatic variation . . . 109

3 Perturbations of chromatin modifiers in endometrial cancer 111
3.1 Background . . . 112
3.1.1 Contributions . . . 112
3.2 Results . . . 113
3.2.1 Selection of cell lines for CRISPR screen . . . 115
3.2.2 CRISPR validation . . . 116
3.2.3 Cell line and genomic perturbations are dominant sources of variation 118
3.2.4 *ARID1A* loss disrupts chromatin accessibility and gene expression in endometrium and EC models . . . 121
3.2.5 Loss of *ARID1B* reveals perturbed gene expression profiles . . . 136
3.2.6 *ARID5B* loss results in up-regulation of interferon response pathways 146
3.2.7 JUND motifs are more accessible in *EP300* knockouts . . . 154
3.2.8 Loss of *KMT2C* results in up-regulation of E2F targets . . . 160
3.2.9 Loss of *SETD1B* down-regulates E2F targets . . . 164
3.2.10 Common perturbations induced by the loss of chromatin modifiers . 169
3.2.11 Functional study of *BCOR* N1459S hotspot mutation . . . 175
3.3 Discussion . . . 182
3.3.1 Functional study of *BCOR* N1459S hotspot mutation . . . 187
3.4 Methods . . . 189
3.4.1 Chromatin modifiers in hEM3 and Ark2 cells . . . 189
3.4.2 *BCOR* N1459S experiment . . . 192

4 Heterogeneity and evolution in cancer 195
4.1 Background . . . 195
4.1.1 Relevance of cancer evolution and tumor heterogeneity . . . 197
4.2 Methodology for ordering genomic events . . . 197
4.2.1 Methodologies for reconstructing evolutionary trees . . . 198
4.2.2 Problem setup . . . 201
4.2.3 Existing approaches for reconstructing evolutionary trajectories from single-sample bulk sequencing . . . 201
4.2.4 Deep Learning approach . . . 203
4.3 Discussion and conclusions . . . 225

5 Conclusions and future work 229
5.1 Molecular landscape of endometrial cancer . . . 229
5.2 Mutations in drivers are predictive of survival . . . 231
5.3 Chromatin modifiers are frequently altered in endometrial cancer . . . 232

5.4 Endometrial cancer hetergoneity . . . 233
5.5 *ARID1A* loss in normal endometrium cell lines causes hormone response imbalance . . . 234
5.6 Functional consequences of perturbations of chromatin modifiers in EC . . . 234
5.7 *BCOR* N1459S hotspot mutation is near-exclusive to EC . . . 237
5.8 RNN-based approach to uncover cancer evolutionary trajectories . . . 238
5.9 Final remarks . . . 239

Appendices

A Software and packages **243**
A.1 Software . . . 244
A.2 R packages . . . 245
A.3 Python modules . . . 246

B Supplementary Information **247**
B.1 Chapter 3 . . . 248
B.1.1 *BCOR* N1459S hotspot mutation experiment . . . 248
B.1.2 Chromatin complexes definitiions . . . 250
B.2 Chapter 4 . . . 253
B.2.1 Statistical approach . . . 253
B.2.2 Deep Learning approach . . . 255

Bibliography **255**

Chapter 1

Introduction

Contents

1.1 Endometrial cancer 2
1.1.1 Pathological classification 5
1.2 Molecular features of EC 7
1.2.1 Heritability in endometrial cancer 7
1.2.2 Somatic mutations in normal endometrium 8
1.2.3 Role of mutations in endometrial cancer 10
1.2.4 Molecular subtypes of endometrial cancer 12
1.2.5 Molecular landscape of endometrial cancer 15
1.3 Cancer driver mutations and genes 19
1.3.1 Driver and passenger alterations 19
1.3.2 Detection of positively selected genes 20
1.3.3 An overview of computational methods used for driver gene identification 21
1.4 Packaging of DNA and chromatin organization 24
1.4.1 Nucleosome - a core structural element of chromatin 24
1.4.2 Regulation of chromatin accessibility 25
1.5 Chromatin organization in cancer 32
1.5.1 DNA methylation 33
1.5.2 SWI/SNF complex 33
1.5.3 NuRD complex 34
1.5.4 Polycomb repressive complex 34
1.5.5 Dysregulation of chromatin in cancer initiation and progression 35
1.6 Heterogeneity and cancer evolution 35
1.6.1 Tumor heterogeneity 35
1.6.2 Cancer evolution 36
1.7 Motivation of this study 38

1.1 Endometrial cancer

Endometrial cancer (EC) - cancer of the lining of the uterus (Fig. 1.1), is the most common gynecological cancer in the UK [1] and US [2]. In 2020, there were 417,000 new diagnoses made globally [3]. EC is the sixth leading cause of cancer-related deaths among women in the US [1] and eighth in Europe [3].

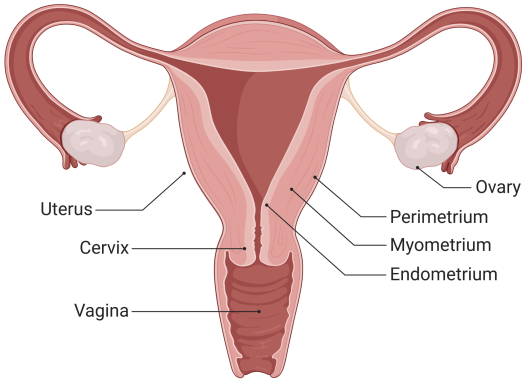


Figure 1.1: **Schematic of the uterus and female reproductive tract.** Endometrial cancer originates in the endometrium - the lining of the uterus and is considerably more common than other uterine malignancies that arise from the muscle or connective tissue. Created with BioRender.com

The incidence of EC is higher in countries with high and very high Human Development Index (Fig. 1.2) with the highest incidence in Europe found in Poland (Fig. 1.3). Cumulative EC risk is 5.3 and 2.4 times higher in very high and high HDI countries respectively than in regions with medium and low HDI [1, 3].

The incidence of EC increased globally by 132% between 1990 and 2019 with a year-on-year increase of age-standardized incidence of 0.69% [4]. This increase is thought to be due to the increase in unopposed estrogen as a consequence of rising population levels of obesity, discussed further in Section 1.2.5.

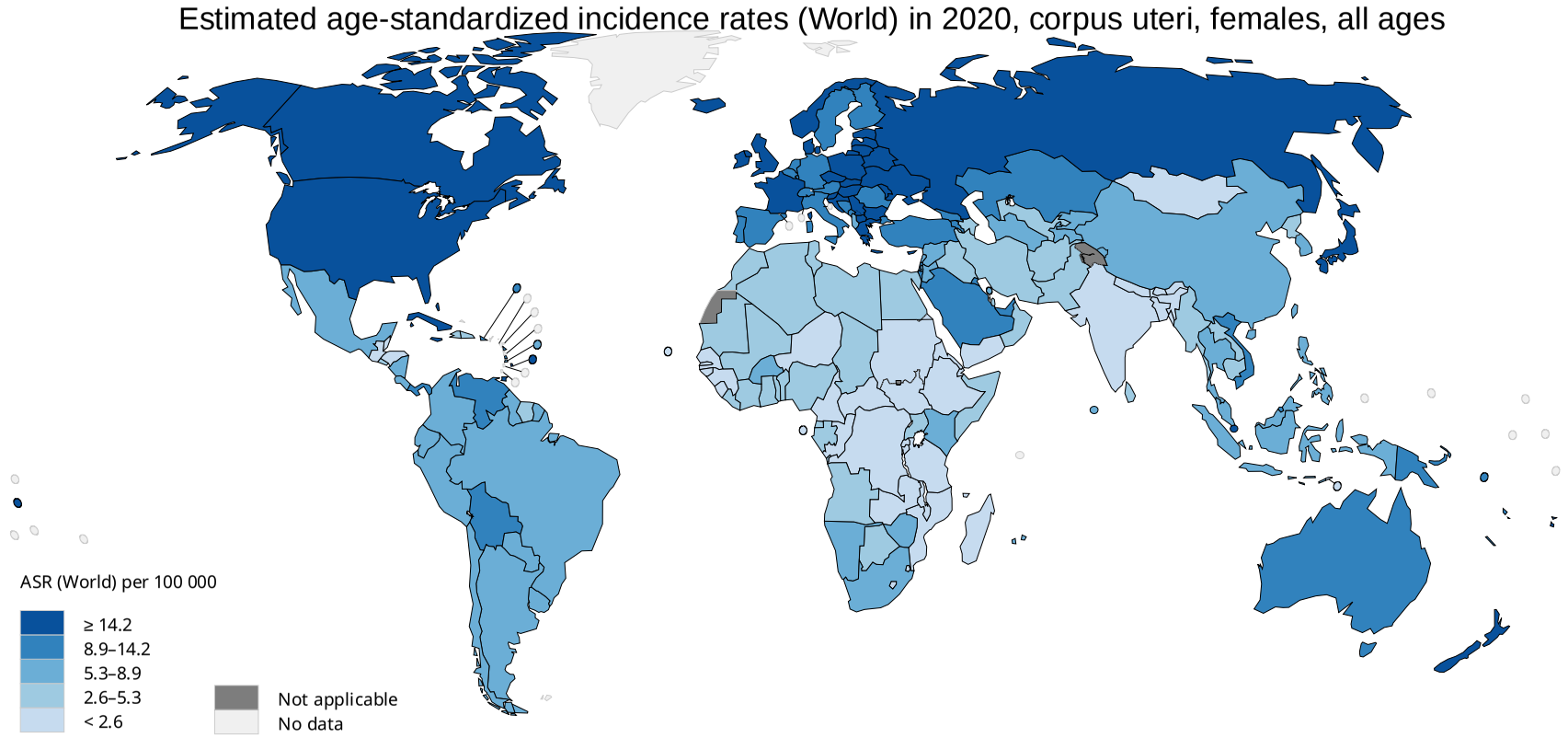


Figure 1.2: **Incidence of Endometrial Cancer Worldwide.** Figure from [3].

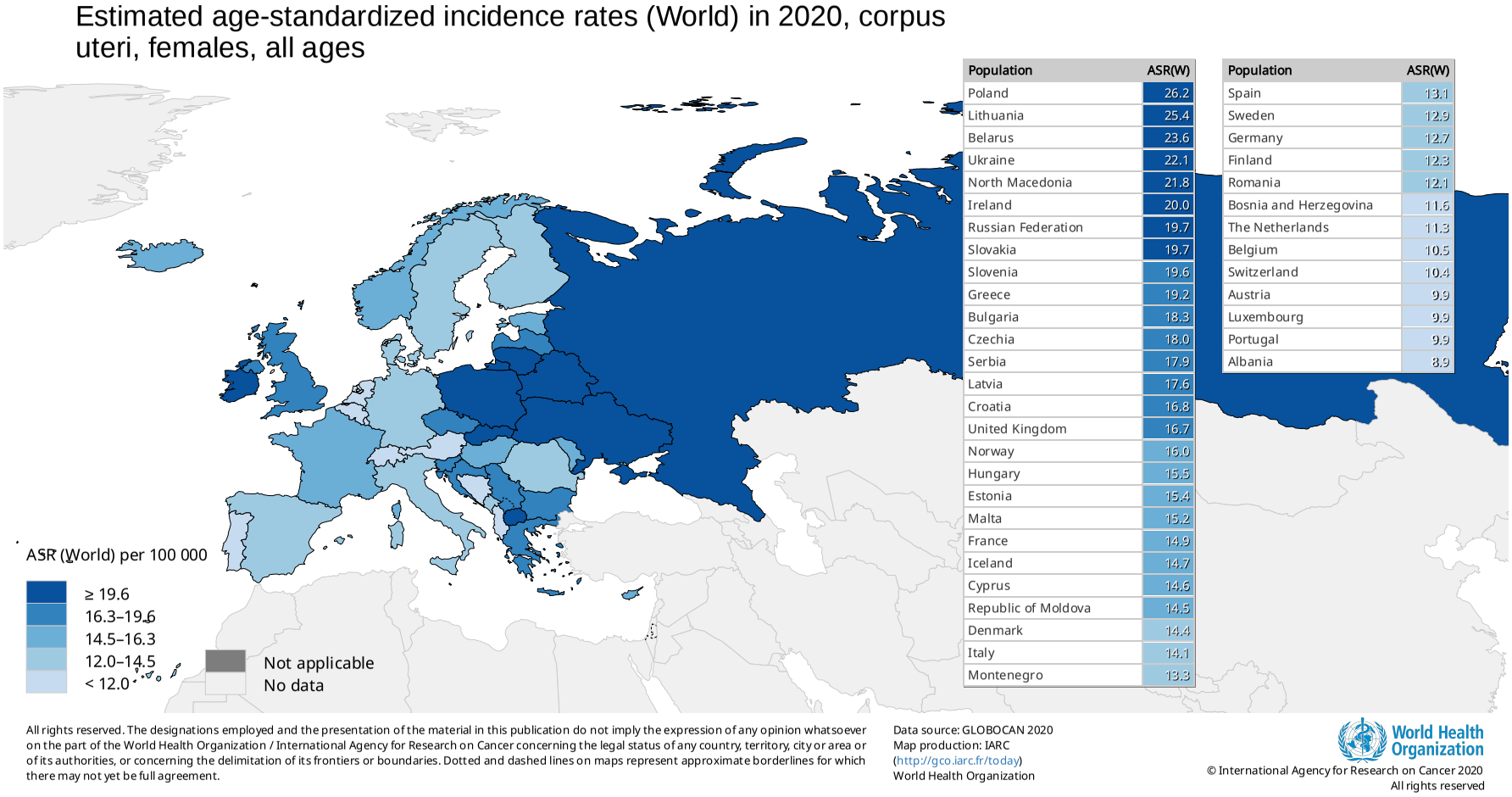


Figure 1.3: **Incidence of Endometrial Cancer in Europe.** Figure from [3]

1.1.1 Pathological classification

Cancer is often classified based on easily observable features to aid the clinical investigation, assess the risk, and choose the best treatment. EC can be classified according to Bokhman's classification based on clinical and epidemiological characteristics [5]. EC can also be classified based on its histopathology into separate types, as discussed in Section 1.1.1.2. Furthermore, the International Federation of Gynecology and Obstetrics (FIGO) staging system is commonly used to classify EC based on the anatomical extent of the disease - with stage I referring to disease confined to the uterus, progressively increased reach of the malignancy with stage IV referring to cancer spread beyond the pelvis to other parts of the body [6].

1.1.1.1 Bokhman classification

Bokhman's dual classification system, introduced in 1983 [5], categorizes EC into two types based on clinical and epidemiological characteristics. Type I EC, representing 60-70% of cases, is hormone receptor-positive and often develops from hyperplastic endometrium, indicating estrogen excess. It typically has a favorable prognosis with a 5-year survival rate of 86%. In contrast, Type II EC, constituting 30-40% of cases, is hormone receptor-negative, arises in atrophic endometrium, and has a less favorable prognosis [5, 7].

1.1.1.2 Histomorphological subtyping

Histopathologically, EC is classified into endometrioid (EEC) and non-endometrioid (NEEC) types. EEC, the most prevalent form, accounts for about 80% of cases and is typically linked to estrogen exposure. It aligns with Type I in Bokhman's classification. EECs are further categorized by a grading system, with high grade indicating increased levels of aggressiveness and abnormality [8–10].

Nonendometrioid carcinomas, including serous, clear-cell, and mixed EC, as well as uterine carcinosarcoma, are high-grade and align with Type II in Bokhman's system. They represent a smaller fraction of EC cases but are more aggressive [7, 11].

Nonendometrioid endometrial carcinomas include serous endometrial carcinoma (SEC),

clear-cell carcinoma (CCEC), mixed EC, and uterine carcinosarcoma (UCS). SEC is an aggressive type of endometrial cancer - it accounts for about 10% of all EC diagnoses and 40% of EC-related deaths. It is high grade by definition and tends to grow and spread quickly; it is more common in older women [11]. Serous carcinomas have a background of endometrial atrophy and show abnormal p53 staining [7], falling under the group of Type II tumors, according to Bohkman. CCEC is a rare type of endometrial cancer that accounts for less than 5% of all cases [12]. It arises from the glandular cells of the endometrium and is typically associated with endometriosis. CCEC is always high-grade and has a poor prognosis compared to other types [13, 14]. This poor prognosis can be explained by its distinct mutational profile [14]. UCS, cancer with both epithelial, typically serous, and mesenchymal components [7], represents less than five percent of all EC diagnoses [15]. Though initially, UCS was classified with uterine sarcomas and treated as a distinct entity (including by The Cancer Genome Atlas as discussed below), it is now considered an epithelial dedifferentiated tumor [16]. Other rare subtypes of endometrial cancer include mucinous carcinoma, mixed epithelial and mesenchymal tumors, and carcinosarcoma (also known as malignant mixed Mullerian tumor).

1.1.1.3 Beyond Bokhman classification of EC

However, while associated with some changes in survival rate, the Bokhman classification, together with the grading system, require further stratification to allow for better risk assessment. One way to further classify EC that can convey prognostic value is based on its molecular characteristic [13, 17]. I discuss the molecular subtypes and their associated clinical value in Section 1.2.4. Ongoing work is carried out, including research described in this thesis, to further describe molecular features of EC, stratify the cases and then use this knowledge to develop clinically relevant strategies to assign best therapeutic strategies [18].

1.2 Molecular features of EC

Cancer's etiology is fundamentally rooted in the accumulation of genomic alterations [19]. Within this framework, germline mutations often play a pivotal role by increasing the risk of disease by predisposing cells to further genetic instability. As tissues accrue somatic mutations, a subset may undergo malignant transformation, highlighting the significance of understanding the molecular underpinnings of cancers. Specifically, in endometrial cancer, understanding of molecular landscape allowed for stratification into molecular subtypes.

1.2.1 Heritability in endometrial cancer

1.2.1.1 Lynch syndrome

Approximately 2-6% of EC can be attributed to Lynch syndrome [20]. Lynch syndrome (LS), an autosomal dominant disorder with high penetrance, is marked by germline mutations in DNA mismatch repair (MMR) genes, *MLH1*, *MSH2*, *MSH6*, and *PMS2*. LS causes increased risk of intestinal and non-intestinal malignancies, particularly colorectal and endometrial cancers. Germline *MSH6* mutations are linked to an increased risk of developing endometrial carcinomas [21]. In LS, the pathogenesis of associated cancers, including endometrial cancer, involves the inactivation of the second MMR gene allele, causing an increased mutation rate. LS-associated endometrial cancers share molecular similarities with those arising from somatic MMR deficiencies but notably differ in the younger age of onset [22, 23]. Those similarities can pose challenges in properly diagnosing patients with LS-associated EC. Given the median timeframe of 11 years for LS patients with endometrial carcinoma to develop a secondary tumor, early detection and proactive management are crucial [20].

1.2.1.2 Polymerase proofreading-associated syndrome

Germline variants in the exonuclease domain of the DNA polymerases *POLE* and *POLD1* have been show to predispose to cancer, including EC [24]. EC cases with mutations in the

exonuclease domain are characterized by an increased mutation rate and relative excess of G:C>T:A transversions [25]. Mutations in *POLE* proofreading domain have been shown to be associated with reduced risk of recurrence and death [26].

1.2.1.3 Cowden syndrome

Cowden syndrome (CS), an autosomal dominant disorder, is characterized by a spectrum of both malignant and benign growths affecting diverse tissues such as the breast, thyroid, uterus, brain, and mucocutaneous tissue [27]. This condition primarily arises from germline mutations in the *PTEN* gene, which plays a critical role in cellular growth and proliferation. Individuals with CS are at an increased risk of developing carcinomas, particularly in the breast, thyroid, and endometrium [28]. In the context of EC, *PTEN* mutations emerge as one of the most prevalent genetic alterations, frequently co-occurring with disruptions in the Pi3K signaling pathway.

1.2.2 Somatic mutations in normal endometrium

Human tissues accumulate somatic mutations at varying rates, with skin, lung, and esophagus exhibiting the highest mutation burdens [29]. Some mutations, despite being associated with cancer, are present in normal tissues without leading to oncogenesis. Notably, mutations in *NOTCH* genes show signs of positive selection in normal skin [30] and esophagus [31] tissues, with *NOTCH1* mutations occurring more frequently in normal esophageal tissue than in esophageal cancer. Understanding the rates and patterns of somatic mutations in normal tissues is crucial for elucidating their role in the onset and progression of cancer.

Extensive research has been conducted on mutations in normal and endometriotic[1] endometrium, utilizing gene panels and whole exome sequencing (WES) [33–36]. These studies have identified recurrent mutations in several cancer-associated genes, including *KRAS*, *PIK3CA*, *FBXW7*, *PPP2R1A*, and *PIK3R1*. Specifically, mutations in *KRAS* were observed at higher mutation frequencies in endometriotic epithelium compared to

[1]Endometriosis is a disease characterized by the presence of endometrium-like tissue outside of the uterus [32].

normal endometrial epithelium, indicating clonal expansion of epithelial cells harboring these mutations in endometriosis [34]. Interestingly, *ARID1A* mutations, prevalent in the endometriotic epithelium, were either absent or significantly less frequent in normal endometrium [34, 35]. Conversely, mutations in *PIK3R1* and *FBXW7* were more commonly found in normal endometrial tissue [34]. Moreover, over 50% of normal endometrial samples analyzed were found to contain somatic driver mutations, including hotspot mutations in *KRAS*, *PIK3CA*, and *FGFR2*, along with PTEN-loss, as identified through immunohistochemistry [33].

A pivotal study employing whole genome sequencing (WGS) on single endometrial glands revealed that normal endometrial glands consist predominantly of clonal cell populations, likely originating from a single progenitor cell [37]. This study also established a correlation between age and mutation burden in normal endometrial glands, with an average accumulation rate of approximately 29 base substitutions per gland per year. The mutation rate in normal tissues was found to be five times lower than that in cancerous tissues. Analysis of the mutations identified 12 genes under positive selection (*PIK3CA*, *PIK3R1*, *ARHGAP35*, *FBXW7*, *ZFHX3*, *FOXA2*, *ERBB2*, *CHD4*, *KRAS*, *SPOP*, *PPP2R1A*, and *ERBB3*) that were also reported as known cancer genes.

The study of somatic mutations in normal endometrial tissue reveals that certain mutations, undergoing positive selection, target genes commonly associated with cancer. Studied glands were healthy and showed no signs of malignancy, thereby suggesting that a combination of such mutations, rather than a single key mutation, might be needed for cancer progression. This finding aligns with research indicating that EC carries more than ten coding substitutions under positive selection, typically necessitating 3-4 hits in key cancer genes per patient [38]. The authors of the single endometrial glands study suggest that lack of *ARID1A* mutations might be symptomatic of the requirement of biallelic mutations to confer a growth advantage and that this could explain the lower prevalence in normal cells [37]. Alternatively, the mutation in *ARID1A* and potentially other chromatin modifiers might not be enough to confer the growth advantage and might require an additional mutation in the cancer driver to result in cancer progression. Further

Somatic mutations	Frequency of alterations [%] EEC	SEC	CCEC
PTEN	52-90 [13, 39–42] 11-33 (UCS) [15, 43]	2-3 [39, 44]	0-21 [39, 45–48]
PIK3CA	22-59 [13, 39–42, 49] 22-40 (UCS) [15, 43, 50]	15-35 [44, 50–53]	24-36 [46, 47]
PIK3R1	9-43 [13, 40, 54] 6-20 (UCS) [15, 43]	5-8 [44, 54]	7-18 [46–48, 54]
CTNNB1	19-40 [13, 41, 42, 49] 0-5 (UCS) [15, 43]	0-3 [42, 43]	0 [46]
ARID1A	39-60 [13, 41, 43] 10-24 (UCS) [15, 43]	7-11 [43]	14-21 [47, 48]
POLE	7-20 [25, 55, 56] 3-4 (UCS) [43, 56]	0-2 [25, 43, 50]	2-7 [46, 47]
FBXW7	0-15 [13, 43] 11-39 (UCS) [15, 43]	15-29 [44, 50–53]	13-25 [46]
TP53	5-35 [13, 40–42] 44-91 (UCS) [15, 43]	59-93 [44, 47, 51]	28-46 [46–48]

Table 1.1: **Frequencies of somatic mutations frequently reported in EC.** Adapted from [57].

investigations, including the heterogeneity and timing analyses of endometrial cancer mutations, are needed to explore the role of *ARID1A* and other chromatin modifiers in EC and its progression.

1.2.3 Role of mutations in endometrial cancer

Before the advent of whole exome and genome assays, the study of mutations in cancer was limited to a panel of genes, often analyzed in small cohorts. Despite these limitations, initial studies successfully identified several genes that were later confirmed in large-scale investigations as being frequently mutated in EC. These include *PTEN*, *PIK3CA*, *PIK3R1*, *KRAS*, *ARID1A*, *CTNNB1*, *TP53* and *FBXW7*, as summarized with their respective frequencies in Table 1.1.

1.2.3.1 *PTEN* mutations in EC

PTEN is the most frequently mutated gene in endometrial cancer, with mutations present in up to 90% of cases [13, 41]. *PTEN* mutations often co-occur with mutations in *PIK3CA* and *PIK3R1* in tumors of endometroid histotype [40]. There is a high prevalence of *PTEN* mutations in both endometrioid endometrial adenocarcinomas and their precursors [58]. However, following the acquisition of *PTEN* mutation, further accumulation of alterations is required for a precancerous to cancerous transition [59].

1.2.3.2 *ARID1A* mutations in EC

ARID1A, gene encoding a member of the SWI-SNF complex (described in more detail in Section 1.4.2.3), is one of the highly mutated genes in EC. *ARID1A* mutations occur in 39-60% of EEC cancers and 7-24% of other EC types [13, 15, 41]. *ARID1A* was also identified as a causative gene in microsatellite instability through the silencing of the *MLH1* promoter [60].

While *ARID1A* mutations are prevalent in ovarian clear-cell carcinoma, they are not common in high-grade serous ovarian carcinoma [61]. This led to speculation that *ARID1A* might play a crucial role in the transformation of endometriosis into cancer [61]. However, mutations in *ARID1A* alone may not be sufficient to drive progression into malignancy, as evidenced by high rates of *ARID1A* mutation in endometrium affected by endometriosis, which rarely evolves into cancer [35], and *Arid1a* loss not being sufficient to cause hyperplasia or cancer in mouse models [62, 63].

Another study found a progressive loss of *ARID1A* expression from benign endometrium to complex atypical hyperplasia to low-grade and high-grade endometrioid carcinoma, suggesting that ARID1A loss, presumably due to mutation, plays a crucial role in tumor progression of uterine endometrioid carcinoma [64].

1.2.3.3 *TP53* mutations in EC

TP53, a crucial tumor suppressor gene, plays a significant role in EC, particularly in the SEC and high-grade EEC. The *TP53* gene is responsible for encoding the p53 protein,

which acts as a critical regulator of cell cycle arrest, apoptosis, and DNA repair. Mutations in *TP53* are associated with aggressive tumor behavior and poor prognosis in EC [65, 66]. These mutations are found in approximately 21-35% high-grade EEC [13, 40–42] and up to 93% of SEC cases [44, 47, 51].

1.2.3.4 *CTNNB1* mutations in EC

CTNNB1, the gene encoding β-catenin, a key regulator in the Wnt signaling pathway, is mutated in 19-40% EC [13, 41, 42, 49]. Pathogenic *CTNNB1* mutations block degradation and accumulation of β-catenin in the nucleus and cytoplasm, resulting in the activation of Wnt target genes that promote cell proliferation and tumorigenesis [67]. Inhibiting Wnt/*beta*-catenin has been shown effective in EC with *CTNNB1* mutations [68].

1.2.4 Molecular subtypes of endometrial cancer

Following the advancements in sequencing technology, a more holistic analysis of the molecular landscape of EC became possible.

The Cancer Genome Atlas (TCGA) was a landmark project supported by the US government that characterized over 20,000 primary cancers across 33 cancer types undertaken with the ambition of improving understanding of cancer biology. Among the multi-omic assays TCGA generated and made publicly available are whole exome sequencing (WES), single nucleotide polymorphisms (SNP) arrays, methylation arrays, and RNA-seq data. As mentioned in Section 1.1.1.2, uterine carcinosarcomas were historically classified as uterine sarcomas. Following this classification, TCGA split EC into two cohorts - Uterine Corpus Endometrial Carcinoma (UCEC) consisting of EEC and NEEC, and Uterine Carcinosarcoma (UCS). However, the subsequent demonstration of shared etiology has led investigators to analyze UCEC and UCS in combination when studying this disease.

TCGA proposed four molecular subtypes of EC classified according to the genomic architecture of analyzed samples (Fig. 1.4A). Using whole exome sequencing (WES) and array-based copy number profiling of 232 tumors, they separated cases into four groups based on mutations in coding DNA and somatic copy number alterations (CNAs). First,

samples with a fraction of nucleotide substitutions C to A above 20%, C to G below 3%, and with more than 500 single nucleotide variants were assigned to *POLE* - ultra-mutated subtype. Those tumors, with exceptionally high mutation rates, were found to carry missense mutations in the proofreading exonuclease domain of the DNA polymerase POLE [13]. The second subgroup comprised a larger group of hypermutated tumors with microsatellite instability (MSI) due to defects in DNA mismatch repair (MMR). The remaining samples were split into copy number high subtype and copy number low based on hierarchical clustering of copy number alterations. The samples assigned to copy number high subtype belonged to the cluster characterized by a high degree of somatic copy number alterations (SCNAs). Nearly all of these were found to carry pathogenic *TP53* mutations. The remaining samples were classified as copy number low. The TCGA molecular subtypes were associated with different survival profiles. CN high samples had the poorest outcome, while POLE samples had significantly better progression-free survival with no recurrences in the group analyzed in the initial study (despite frequent high grade). The remaining two molecular subgroups' clinical outcome was somewhere between the CN high and POLE (Fig. 1.4B).

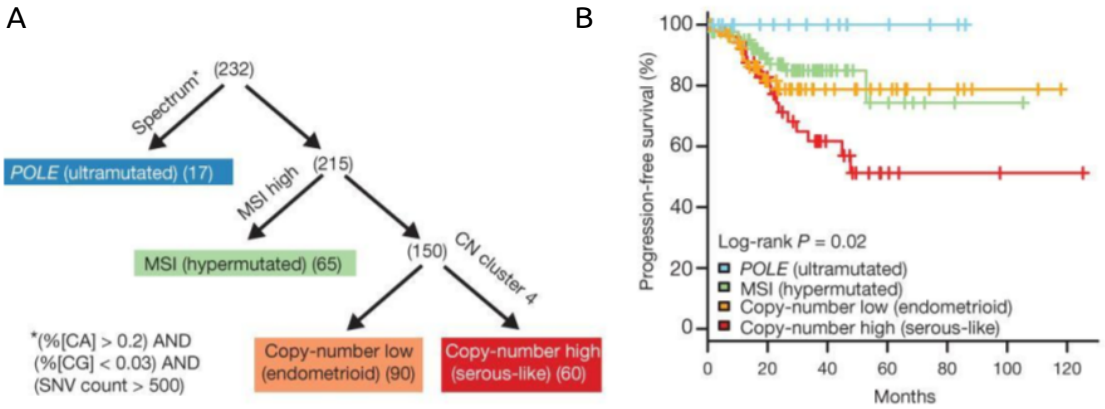


Figure 1.4: **TCGA defined molecular subtypes. A** - flowchart of the assignment of samples to molecular subtypes. **B** - progression-free survival with respect to molecular subtype assignment. Figure from [13].

The TCGA molecular subtypes are distributed unevenly across different EC histotypes (Fig. 1.5). More than half of low-grade EEC samples belong to CN low subtype, while MSI subtype accounts for the majority of high-grade EEC. 98% of serous samples are assigned to CN high subtype. Between 5 and 17% of EEC samples are assigned to POLE

subtype.

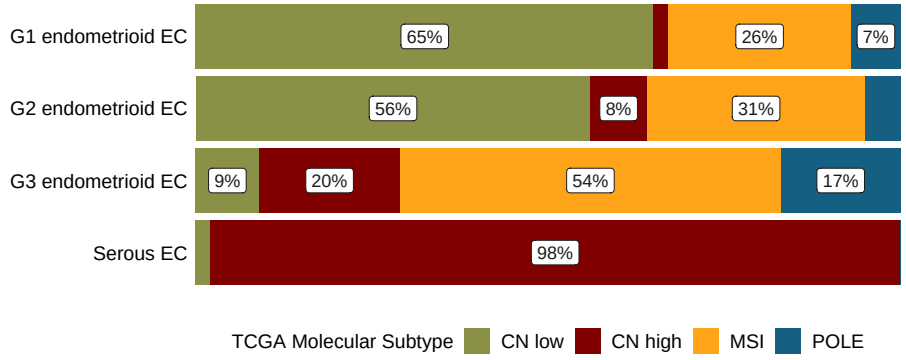


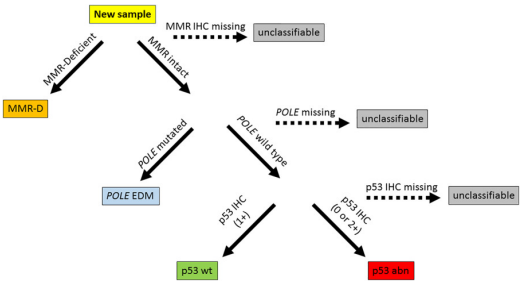
Figure 1.5: **Distribution of molecular subtypes across EC histotypes.** Reproduced from [57].

1.2.4.1 Clinical significance of molecular subtypes

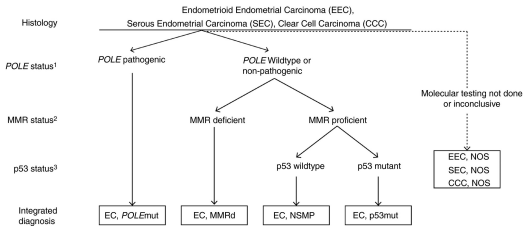
Given that molecular subgroups showed significant prognostic value in the initial study, the subtypes were adapted for clinical use. The authors of Proactive Molecular Risk Classifier for Endometrial Cancer (ProMisE) - a molecular classification system based on TCGA molecular subtypes, assign samples to molecular subtypes based on mismatch repair (MMR) and p53 immunohistochemistry (IHC) and sequencing of polymerase-ϵ (*POLE*) exonuclease domain [69]. The first assessment is based on IHC for the presence of MMR proteins - cases with DNA mismatch repair deficiency are subject to further analysis to determine whether they require investigation for Lynch syndrome (a germline predisposition to cancer). MMR intact samples are next assessed for *POLE* mutations. The presence of polymerase-ϵ exonuclease domain mutations (EDMs) results in tumors classification as *POLE* EDM - subgroup corresponding to TCGA POLE ultramutated subtype. The remaining samples are assessed for levels of TP53 protein and split into presumed p53 wild type and p53 abn (with aberrant p53 IHC staining) corresponding to CN low and CN high TCGA subgroups, respectively (Fig. 1.6a). An alternative diagnostic algorithm, proposed in PORTEC clinical trials [70], follows an assignment to (1) POLE-mut subtype (POLE) based on pathogenic *POLE* EDMs, (2) mismatch repair deficient subgroup (MMRd, equivalent of MSI) defined by the loss of one or more MMR-proteins (MLH1, PMS2, MSH2 and MSH6), (3) p53 mut subtype (CN high) based on abnormal

p53 IHC. The remaining samples are assigned to a non-specific molecular profile (NSMP) subtype - an equivalent of CN low TCGA molecular subtype (Fig. 1.6b).

A favorable prognosis for POLE subtype samples was independently reported in clinical studies [26]. Separately, the ProMisE study found that p53 abn (CN high) tumors had the worst prognosis, while tumors with *POLE* EDM (POLE) had the most favorable outcome [71].



(a) **Proactive Molecular Risk Classifier for Endometrial Cancer (ProMisE)** Figure from [69].



(b) **Post-Operative Radiation Therapy in Endometrial Cancer trials (PORTEC)** Figure from [72].

Figure 1.6: **Clinical adaptation of TCGA molecular subtypes.**

1.2.4.2 Future directions

The current assignment of cases to molecular subtypes is costly and labor-intensive. An improved PORTEC classification, used by Translational Research in Post-Operative Radiation Therapy in EC (TransPORTEC) initiative, requires sequencing of *POLE*, *TP53* and *CTNNB1*, identification of MMRd, IHC of p53, MLH1, MSK2, MSH6, PMS2, L1-CAM, ER/PR and (lymphovascular space invasion) LVSI quantification. The ability to have a one-step DNA-based test for molecular classification would allow a broader application of molecular subtyping in clinical practice [73]. The research carried out in Genomics England 100,000 Genomes Project endometrial cancer cohort, described in the next chapter, contributes to addressing this challenge.

1.2.5 Molecular landscape of endometrial cancer

As discussed above, the TCGA described the molecular landscape of the coding EC genome (Fig. 1.7) [13]. POLE (ultramutated) subtype is characterized by 232 mutations

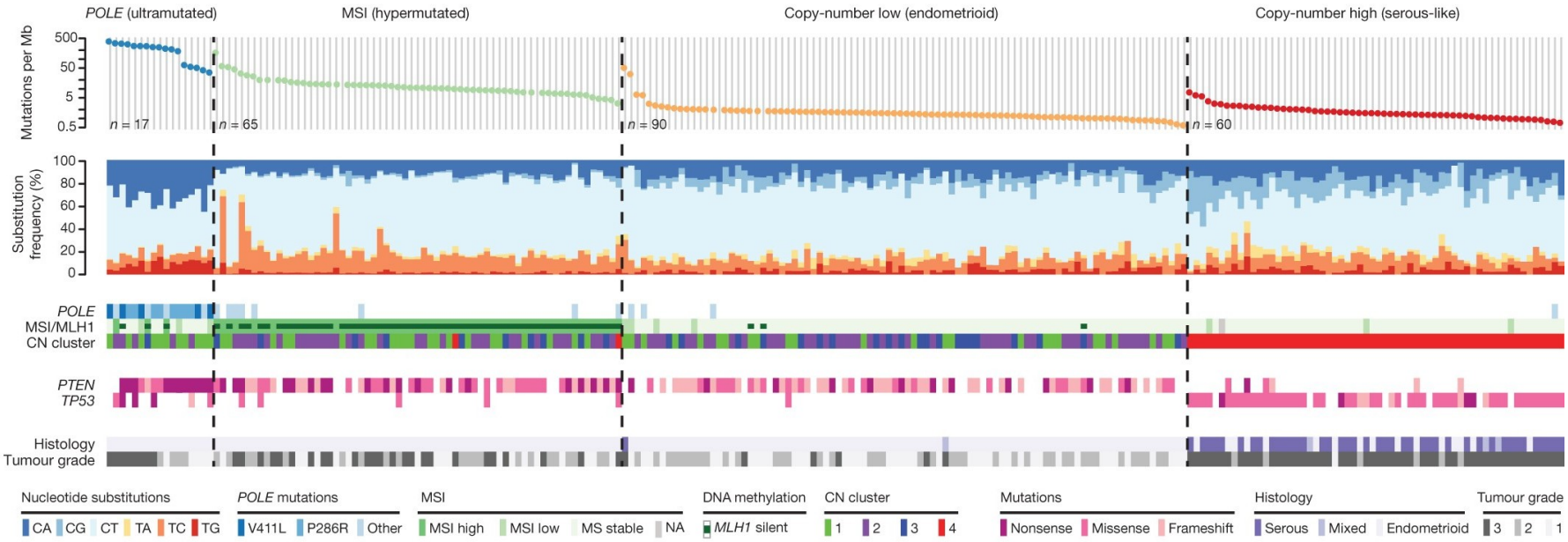


Figure 1.7: **Mutational landscape of TCGA UCEC cohort.** Figure from [13].

per Mb with a higher frequency of C>A substitutions than other subtypes and frequent nonsense mutations in *PTEN* and other tumor suppressors. MSI (hypermutated) samples, in which loss of DNA mismatch repair function was mostly due to somatic methylation of the MLH1 promoter, had a mutation frequency of 18 mutations per Mb and a higher frequency of T>C substitutions - in over a third of samples had 20% mutations consisting of T>C mutations. The CN high subtype, consisting of mostly serous tumors, has high numbers of SCNAs, a lower mutation frequency of 2.3 mutations per Mb, an increased proportion of C>G substitutions, near-universal mutations in *TP53* gene and a lower frequency of mutations in *PTEN*. The CN low subtype has a mutation rate of 2.9 mutations per Mb, with most samples carrying *PTEN* mutation and other alterations in the PI3K pathway.

β-catenin pathway is altered through mutually exclusive mechanisms in most EEC samples (Fig. 1.8a). 70% of all samples from MSI, CN low, and CN high subtypes had at least one alteration of a pathway member. In MSI samples *KRAS*, activating mutations were present in 36% of samples. In CN low subtype, 53% of samples carried an activating mutation in *CTNNB1*. In CN high samples were harboring either *ERBB2* amplifications (25%) or oncogenic mutations in *FBXW7* (22%).

Like most other cancer types, EC is characterized by recurrent alterations in cellular signaling pathways responsible for regulating cellular homeostasis. The PI3K pathway, mentioned above, is the most commonly altered, owing to mutations in *PTEN*, *PIK3CA*, *PIK3R1*, and multiple other components. The MEK-ERK-MAPK pathway is also commonly altered, with frequent amplifications in the receptor tyrosine kinases (RTKs) of the ERBB (HER2) family, less commonly activating mutations affecting these RTKs and, frequently, recurrent mutations in KRAS. The third most commonly altered pathway is the Wnt pathway, with common *CTNNB1* mutations in CN low tumors.

CCEC The TCGA project did not analyze CCEC cases, and their molecular profiles remain underexplored in comparison to EEC and SEC. CCEC mutation profile was reported to resemble that of SEC rather than EEC, with no mutations in *PTEN* and *CTNNB1* reported [48]. Further research has shown that CCEC might, in fact, be a heterogeneous

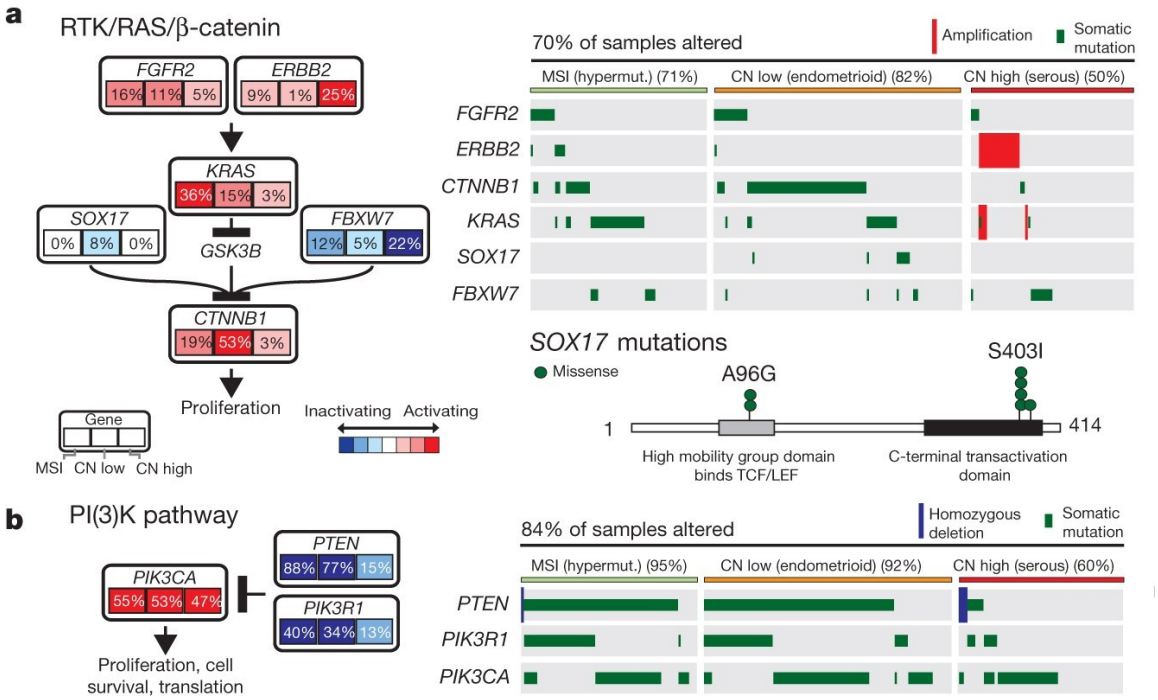


Figure 1.8: (a) **RTK/RAS/β-catenin** and (b) **PI(3)K pathways are altered in EC.** Figure from [13]

group with molecular profiles similar to those of SEC and, in other cases, to EEC [47].

1.2.5.1 Therapeutic strategies informed by molecular landscape

An increasing understanding of EC molecular biology has guided rational therapeutic strategies. The PI3K-AKT-mTOR signaling pathway, a key regulator of cellular growth and survival, has been identified as the most frequently dysregulated pathway in endometrial cancer, which prompted the advancement of PI3K inhibitors as targeted therapeutic agents. Despite these efforts, clinical trials evaluating the efficacy of mTOR inhibitors, either as monotherapy or in combination, have yielded modest results, with response rates hovering around a mere 10% [74]. Combining hormonal therapy with mTOR inhibitors, such as everolimus and letrozole, has shown promising response rates of 28-32%, increasing to 45% in PR-positive patients [75, 76]. At the same time, adding metformin did not enhance outcomes [76].

Hypermutated and ultramutated endometrial tumors, characterized by an exceptionally high mutational burden, are attractive candidates for immunotherapeutic interventions. Immune checkpoint inhibitors, such as pembrolizumab and dostarlimab, have

shown promise in treating advanced and metastatic endometrial cancer, with response rates ranging from 42% to 57% in MMR-deficient tumors [77, 78]. However, despite the encouraging results, tolerability remains a concern due to a high incidence of grade 3-4 treatment-related toxicity in two-thirds of patients [79]. Tumors with high chromosomal instability, such as those harboring *TP53* mutations or frequent CNAs, often exhibit increased sensitivity to PARP inhibition. Preliminary data from a study of olaparib in high-grade serous EC suggests that PARP inhibitors might potentially improve survival [80].

However, despite these efforts, there remains an unmet need to identify other recurrently altered pathways and processes in EC that can be targeted for therapeutic gain. Chromatin organization, one of the fundamental pathways for cell functioning, is frequently dysregulated in endometrial cancer. I discuss my work in understanding the role of chromatin in endometrial cancer in Chapter 2 and Chapter 3.

1.3 Cancer driver mutations and genes

In early studies and the initial TCGA publications, important cancer genes were defined primarily based on recurrent mutations and the frequency of mutations in particular genes. However, to identify less common but key alterations, more sophisticated methods are needed.

1.3.1 Driver and passenger alterations

Cancer, fundamentally a genetic disease, evolves through the accumulation of mutations, among which "driver" mutations play a pivotal role by conferring a selective growth advantage to cells, directly contributing to the cancer phenotype. These alterations promote cell proliferation, evade apoptosis, and sustain angiogenesis, among other hallmarks of cancer [81]. In contrast, "passenger" mutations, while arising due to the high mutation rates in cancer cells, do not confer a selective advantage and are considered neutral in the context of cancer progression.

1.3.2 Detection of positively selected genes

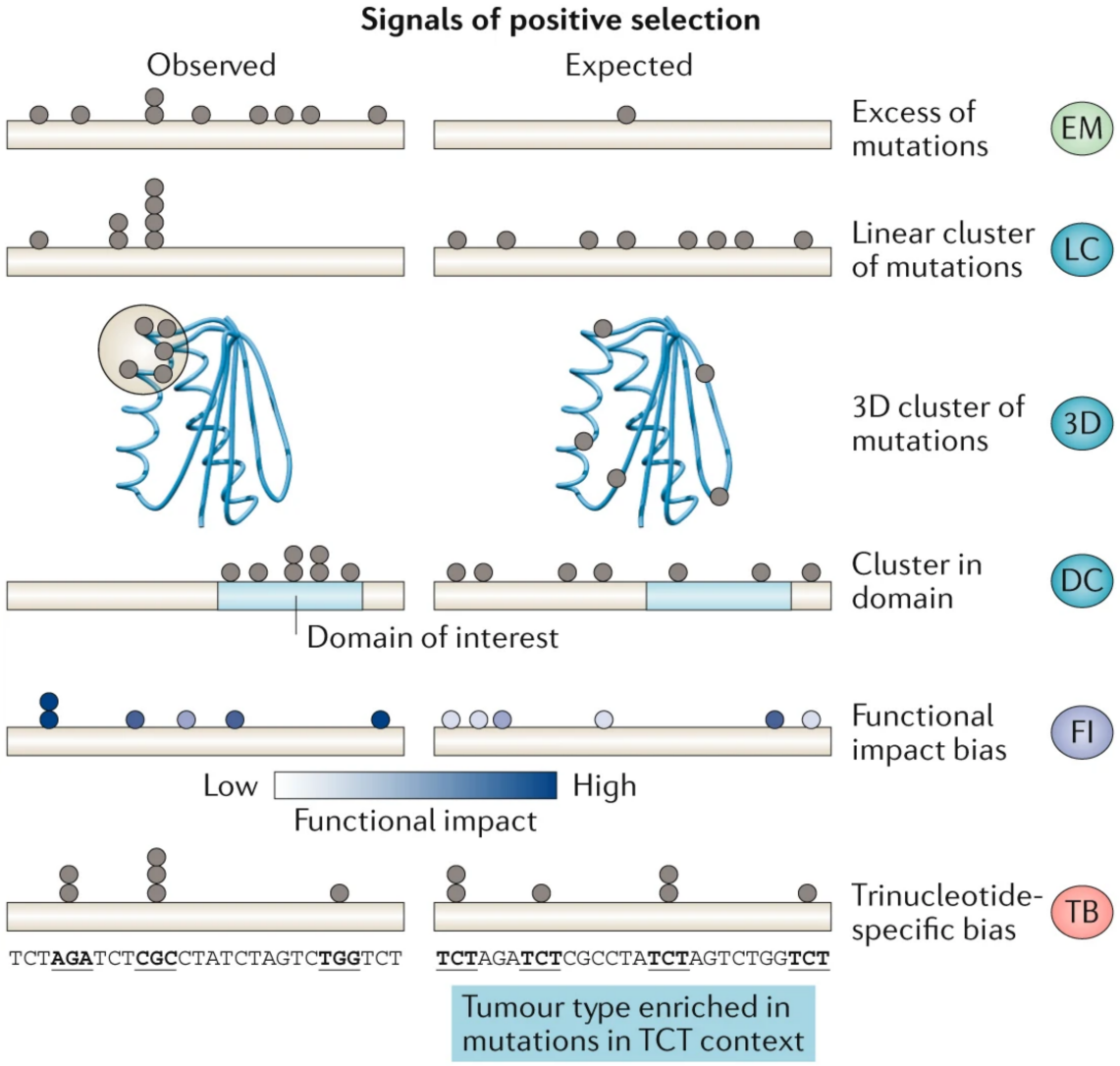


Figure 1.9: **Signals of positive selection.** The overview presents multiple ways in which positive selection of the mutations in genes can be identified. This can be done by observing an excess of mutations (EM) over the expected amount. Additionally, an increased number of mutations in clusters can be observed - either linear clusters of mutations at a given location (LC), 3D clusters in protein structure (3D), or clusters equivalent to functional domains (DC). Another sign of positive selection includes observing mutations of higher functional impact (FI) and a specific trinucleotide bias (TB). Depending on the underlying molecular processes, only one or multiple signals of the positive selection can be observed. Figure from [82].

Positive selection in cancer can be identified through several key signals, one of which includes an elevated mutation count in a given gene, surpassing the estimated background mutation rate, making genes with typically low background mutation rates, such as *BRCA1* and *BRCA2* in ovarian cancer, more discernible as drivers [83]. Another

signal involves mutational hotspots, which represent a concentrated cluster of mutations deviating from a uniform distribution, exemplified by the *PIK3CA* R88Q mutation in endometrial and colorectal cancers [84]. Additional indicators include clustering of mutations at specific 3D protein positions [85], within functional domains [86], or at sites significantly impacting protein function [87], as well as aggregations in regions showing specific nucleotide contexts [88] (Fig. 1.9).

1.3.2.1 Challenges in identification of cancer drivers in EC

Endometrial cancer, like many cancers, exhibits unique genetic characteristics that add complexity to the identification of driver mutations. These include molecular characteristics discussed in Section 1.2.4 such as high prevalence of specific mutation types (e.g., microsatellite instability, *POLE* signatures), increased mutation rates in some subtypes in comparison to others, and various intermingled alterations (for example, high numbers of copy number alterations, particularly in CN-high subtypes that highlight the need for identification of interplay between different types of alterations).

1.3.3 An overview of computational methods used for driver gene identification

Identifying driver genes in cancer requires precise computational approaches to estimate the heterogeneous background mutation rate effectively [89]. These approaches use statistical models to differentiate driver mutations, which are critical for cancer development, from passenger mutations that do not contribute to the disease process. A prominent framework in this field includes IntOGen [82]. The IntOGen pipeline starts with careful preprocessing of input data and filtering of sample sets, followed by the framework that integrates seven different methods, including dNdScv [38], OncodriveFML [87], CBaSE [90], OncodriveCLUSTL [91], a modified version of HotMAPS [85], smRegions [86], and Mutpanning [88].

These methods are specifically designed to compare observed mutation distributions to expected distributions, based on detailed models of background mutation rates and func-

tional impacts. This approach allows for the identification of mutations in specific genomic regions, considering both the mutation's impact and its nucleotide context (Fig. 1.9).

By aggregating the findings from each method, IntOGen provides a comprehensive analysis, combining the candidate driver genes identified into a unified list. This approach ensures increased sensitivity in the detection of driver genes while maintaining specificity [82]. IntOGen's integrated suite of computational methods offers a robust framework for the identification of cancer driver genes, contributing to a deeper understanding of cancer genetics and potentially informing the development of targeted therapies.

1.3.3.1 Positive selection

Nearly all somatic mutations are passengers - they are tolerated and escape negative selection. Cancer drivers are subject to positive selection. To discern driver mutations, a refined model of the non-synonymous to synonymous mutations ratio (dN/dS model) was developed, accommodating context-dependent mutational processes among various mutation types. This model, known as dNdScv [38], integrates local observed mutation rates with a regression model that leverages genome-wide data to anticipate variable mutation rates. With their analysis of the TCGA data, authors identified between 1-3.9% of all genes to be under positive selection with 1-10+ drivers per tumor, across cancer types [38].

1.3.3.2 Functional impact bias

In [87], researchers developed a method to analyze the pattern of somatic mutations in both coding and non-coding genomic regions. The aim of the method is to identify regions with an increased functional impact (FI). FI is defined by the impact of the mutation on protein function, RNA structure, transcription binding site, microRNA target, and other factors. The method calculates the accumulated FI for each region and models the expected FI based on a local simulation of the mutational processes affecting it. This analysis includes coding and non-coding genomic regions and uses different functional impact scoring metrics. The observed FI is then compared to the distribution of expected

FIs under the null hypothesis to obtain a p-value.

1.3.3.3 Bayesian inference of negative and positive selection

CBaSE [90] employs a hierarchical probabilistic framework that allows modeling of coding point mutations. The method assumes neutral evolution of synonymous mutations and, using the count of synonymous mutations identified in each gene, establishes a baseline for detecting selection by comparing observed nonsynonymous mutation counts. Deviations from the modeled baseline allow the identification of genes under positive and negative selection.

1.3.3.4 Sequence-based clustering

OncodriveCLUSTL [91] is a sequence-based clustering algorithm designed to identify significant clustering signals across genomic regions and elements indicative of cancer drivers. This method utilizes a local background model derived from the simulation of mutations that account for the tri- or penta-nucleotide context substitutions observed in a given cohort. OncodriveCLUSTL is also shown to identify the mutation clusters in human promoter regions, as well as shows a possible extension to non-human data by identifying driver clusters in mouse gene data [91].

1.3.3.5 3D clusters of mutations

HotMAPS [85] identifies clusters of amino acid residues with significantly increased local mutation density compared to an empirical null distribution, based on 3D spatial proximity rather than linear sequence. In this method, proximity is defined by the 3D confirmation according to the experimentally resolved protein structure or from high-fidelity modeling of that structure. This approach allows for increased sensitivity of driver identification. Furthermore, it allows the identification of the hotspot driver mutations in a small region, often representing binding sites or other experimentally informed smaller regions, and allows researchers to hypothesize the fine-grained mechanistic understanding of the consequences of those mutations.

1.3.3.6 Enrichment of mutations in functional domains

The smRegions method [86] is a machine learning approach that identifies the enrichment of mutations in Pfam domains of the genes. The method was used to identify the new degron (protein sequences recognized by E3 ligases) instances, with evidence of positive selection suggesting their role in cancer development. For example, mutations in degrons identified with smRegions in *CTNNB1* were shown to increase the stability of the protein across cancer types and cancer-derived cell lines [86].

1.3.3.7 Nucleotide context

The MutPanning method [88] capitalizes on the distinct nucleotide context of passenger mutations, resulting from cancer cell mutational processes, to identify driver genes. By contrasting the nucleotide context of driver mutations with that of passenger mutations and integrating this analysis with conventional signals for driver identification, this method offers a nuanced approach to discovering cancer driver genes [88].

1.4 Packaging of DNA and chromatin organization

Chromatin accessibility is a crucial factor in regulating gene expression, dictating the ability of the transcriptional machinery to access specific genomic loci. As shown in Fig. 1.10, the compact nature of heterochromatin typically impedes access to DNA sequences, whereas the more open conformation of euchromatin allows for the binding of transcription factors and other regulatory proteins. Consequently, chromatin accessibility directly influences the level of transcriptional activity at specific loci, playing a vital role in cellular processes and organismal development [92].

1.4.1 Nucleosome - a core structural element of chromatin

Packaging of DNA within the cell nucleus is critical to eukaryotic genome organization and regulation. At the most basic level, DNA is tightly wrapped around a core of histone proteins to form nucleosomes, the fundamental structural units of chromatin [93]. Comprised

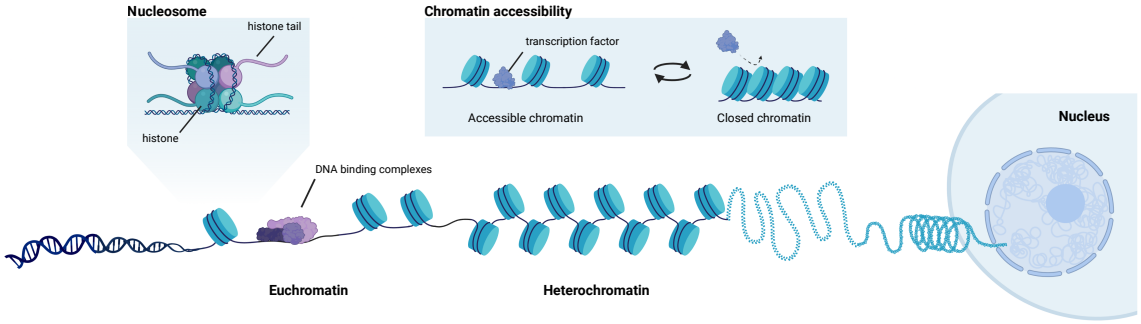


Figure 1.10: **Chromatin organization.** From left to right: DNA packaged around histones is spaced out in the regions of euchromatin (open) chromatin and tightly packaged in heterochromatin chromatin regions. The chromatin is then further compacted and stored in chromosome territories in the nucleus. Nucleosome (presented in the left box) - DNA fragment wrapped around a histone complex comprising 8 core histone proteins. Nucleosomal (core) histones have their tails exposed which allows for their modifications which in turn influences chromatin organization. As shown in the chromatin accessibility box, chromatin remodelers can modify chromatin organization from accessible to a transcription factor and other DNA-binding proteins to closed chromatin. Figure created with BioRender.com

of approximately 147 base pairs of DNA wrapped around an octamer of histone proteins, nucleosomes provide the foundation for higher-order chromatin organization (Fig. 1.10) [93]. These octamers consist of two copies of each core histone protein: H2A, H2B, H3, and H4 [94]. Protruding from the nucleosomal histones are tail domains subject to various post-translational modifications influencing chromatin accessibility and gene expression.

1.4.2 Regulation of chromatin accessibility

Chromatin organization is dictated by various multiple factors including DNA methylation and histone modifications. Nucleosome positioning, a highly dynamic process, dramatically influences the accessibility of target DNA sequences [95] (Fig. 1.10). Chromatin remodelers, such as the SWI/SNF complex discussed in detail later, can reposition nucleosomes making chromatin accessible to DNA-binding proteins even in visibly condensed regions [95].

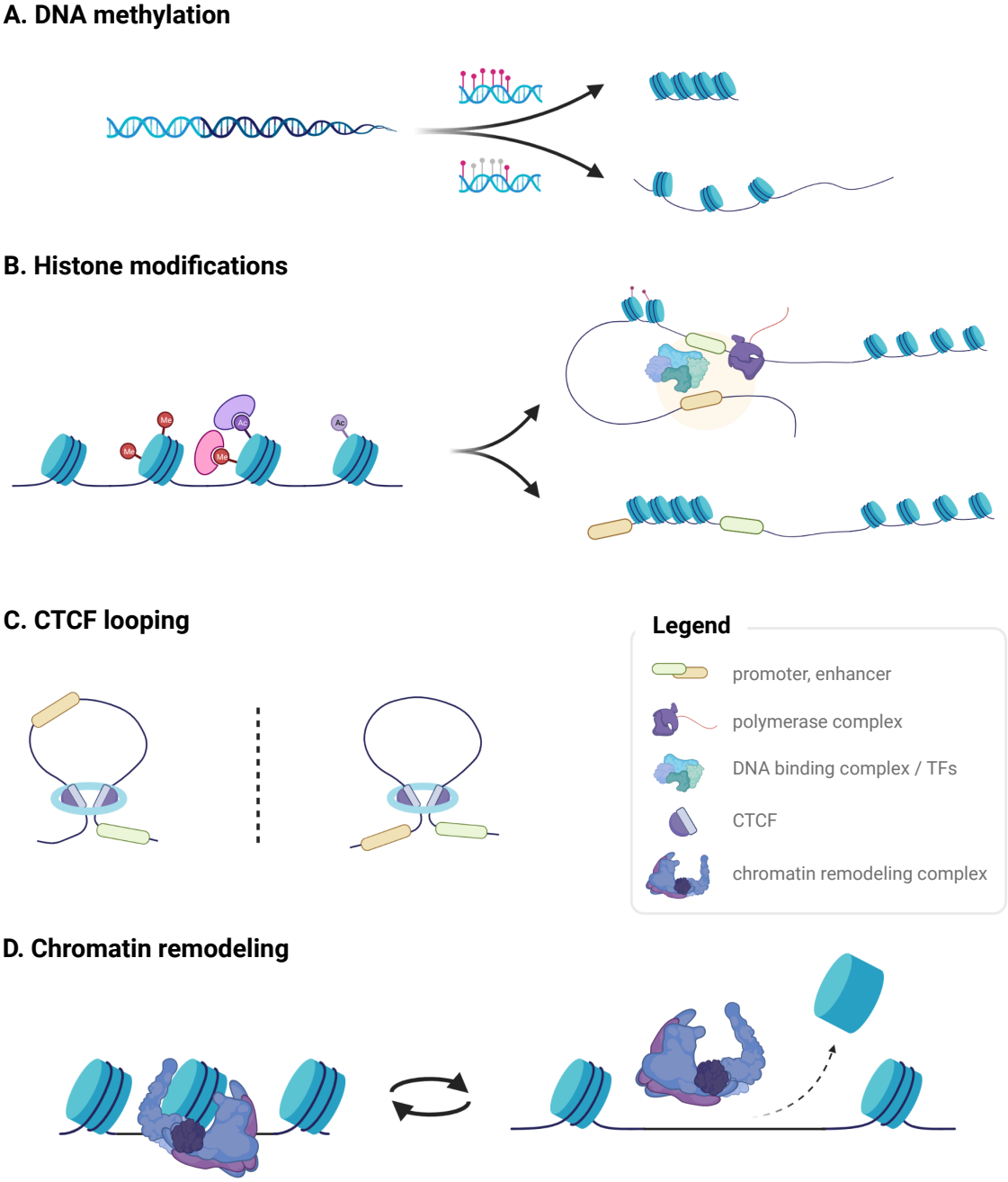


Figure 1.11: **Regulation of chromatin organization.** **A** DNA methylation levels impact chromatin organization, with higher methylation levels leading to a more closed chromatin structure. **B** Histone modifications influence chromatin organization either by directly modulating histone-DNA interactions or by indirectly recruiting effector proteins (readers) that, in turn, recruit chromatin remodelers. Accessible chromatin permits the binding of regulatory elements, initiating gene expression. **C** CTCF shapes the 3D chromatin structure by forming chromatin loops, which can either bring regulatory elements together to facilitate the binding of regulatory complexes and transcription initiation or separate them, thereby reducing gene expression under their control. **D** Chromatin remodeling complexes can reposition nucleosomes, enhancing the accessibility of genomic regions to DNA-binding proteins. Figure created with BioRender.com

1.4.2.1 DNA methylation

DNA methylation, the addition of a methyl group to the fifth carbon of cytosine residues, is a key epigenetic modification in regulating chromatin accessibility. Methylation at gene promoters and other regulatory regions often leads to transcriptional repression by influencing closed chromatin organization (Fig. 1.11A) [96]. The establishment, maintenance, and removal of DNA methylation are mediated by specialized enzymes, such as DNA methyltransferases (DNMT1, DNMT3A, and DNMT3B) and demethylases, which are responsible for the dynamic nature of this epigenetic mark [97].

1.4.2.2 Histone modifications

Histone modifications, such as acetylation, methylation, phosphorylation, and ubiquitination, represent a diverse array of chemical alterations to histone proteins that have the potential to modulate chromatin structure and impact gene regulation [98, 99]. These modifications can influence the interaction between histones and DNA or serve as recognition sites for chromatin-remodeling complexes and other regulatory factors [100]. The complex and combinatorial nature of histone modifications, often referred to as the "histone code," provides a mechanism for intricate control of chromatin accessibility and gene expression (Fig. 1.11) [101].

Several well-characterized histone lysine methylations are predominantly found on histones H3 and H4, including histone H3 lysine 4 (H3K4), lysine 9 (H3K9), lysine 27 (H3K27), lysine 36 (H3K36), lysine 79 (H3K79), and histone H4 lysine 20 (H4K20) [102]. Each of these lysine residues can exist in four distinct methylation states—unmodified, monomethylated (Kme1), dimethylated (Kme2), and trimethylated lysine (Kme3)—or alternatively, they may be present in mono-acetylated or other acetylated forms [98]. The functional significance of these post-translational modifications can vary based on their abundance and relative impact on chromatin structure and activity. For instance, the presence of H3K4me3 at the transcription start site signifies active transcription [103]. H3K4me1 typically denotes active enhancer sites, while H3K27me3 indicates a repressed region regulated by Polycomb complexes. The co-occurrence of these marks suggests a

repressed enhancer, as described in [103].

Writers, readers, and erasers Histone modifications are dynamically regulated by a diverse set of proteins, which can be categorized into three main functional groups: writers, readers, and erasers [102]. Writers are enzymes responsible for adding specific chemical modifications to histone proteins, such as histone acetyltransferases (HATs) that catalyze the acetylation of lysine residues [104] and histone methyltransferases (HMTs) that add methyl groups to lysine or arginine residues [105]. Readers are proteins that recognize and bind to specific histone modifications, facilitating the recruitment of additional regulatory factors or chromatin remodeling complexes that modulate chromatin structure and gene expression [98]. Bromodomains, for example, specifically recognize and bind to acetylated lysine residues on histone tails [106]. Finally, erasers are enzymes responsible for the removal of histone modifications, such as histone deacetylases (HDACs) that remove acetyl groups from lysine residues [107] and histone demethylases (HDMs) that remove methyl groups from lysine or arginine residues [108]. Together, these writer, reader, and eraser proteins orchestrate the dynamic regulation of histone modifications, thereby controlling chromatin structure, accessibility, and gene expression in response to diverse cellular signals and environmental cues.

KMT2 family of writers The histone-lysine N-methyltransferase 2 (KMT2) family, also known as mixed-lineage leukemia (MLL) family, is a group of enzymes that play a crucial role in modulating chromatin structure and gene expression by methylating histone H3 on lysine 4 (H3K4) [109]. The KMT2 family comprises six members, each exhibiting distinct substrate specificities and functional roles in various cellular processes, including development, differentiation, and cellular response to environmental cues [110]. The methylation of H3K4 by KMT2 enzymes results in an accessible chromatin conformation, promoting the binding of transcription factors and other regulatory proteins, thereby facilitating gene transcription. The KMT2 family members are involved in diverse cellular processes such as hematopoiesis, neuronal development, and stem cell maintenance [111]. Dysregulation of KMT2 activity has been implicated in various human diseases, includ-

ing developmental disorders and cancers [111]. For instance, SETD1B, also known as KMT2G, has been linked to several developmental issues, including neurodevelopmental delay [112].

COMPASS complex KMT2 family members form COMPASS-like complexes. The COMPASS (Complex of Proteins Associated with Set1) complex is a highly conserved multi-subunit histone methyltransferase complex, initially discovered in yeast, that plays a pivotal role in regulating gene expression and chromatin organization [113]. The enzymatic activity of the COMPASS complex is modulated by its various accessory proteins, which are essential for stabilizing the complex, enhancing its methyltransferase activity, and regulating its substrate specificity [114].

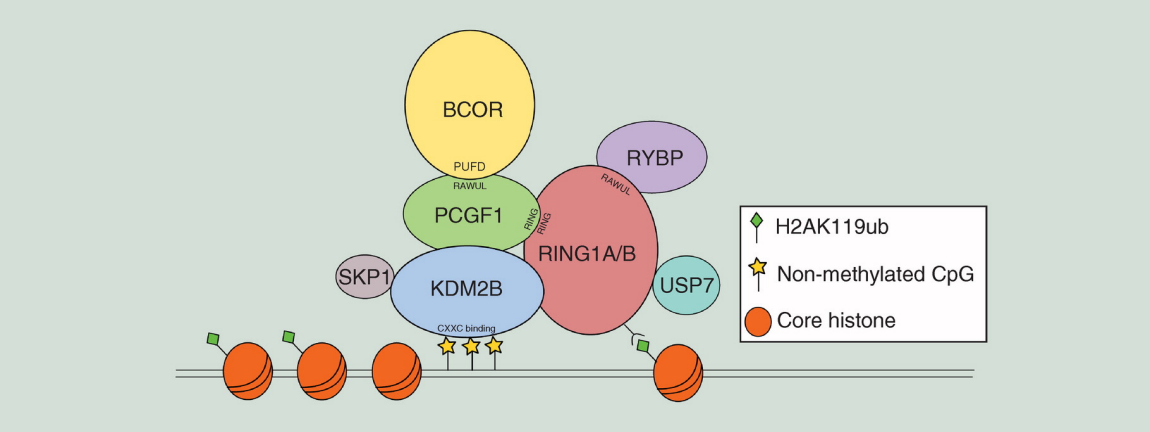


Figure 1.12: **Schematic of ncPRC1.1 model.** The function of the complex is ubiquitination of histone H2A at Lys119 (H2AK119ub). The core complex consists of the catalytic enzyme RING1A/B, which forms a dimer with PCGF1. BCOR connects to PCGF1, while RYBP interacts with RING1A/B. Chromatin recruitment is facilitated by KDM2B, which recognizes nonmethylated CpG islands through its CXXC-binding domain. SKP1, with KDM2B, and USP7 function as a deubiquitinating enzyme. Figure from [115].

Polycomb repressive complexes 1 Polycomb repressive complex 1 (PRC1) is a multiprotein complex that plays a crucial role in gene silencing and epigenetic regulation in developmental and physiological processes [116] PRC1 is one of the key members of the Polycomb group (PcG) protein family, which is involved in maintaining the transcriptional repression of target genes, particularly those associated with developmental

regulation and cell fate determination [117]. PRC1 functions primarily through its ability to ubiquitinate histone H2A at lysine 119 (H2AK119ub1), a modification associated with a repressive chromatin state [118, 119]. This modification contributes to the compaction of chromatin and restricts the access of the transcription machinery to target genes, thereby maintaining their transcriptional repression. PRC1 exists in canonical and six non-canonical configurations [115]. Interestingly, a non-canonical BCOR-PRC1.1 complex (shown in Fig. 1.12) represses differentiation programs in human embryonic stem cells [120]. PRC1 and other PcG complexes play an essential role in the epigenetic regulation of gene expression during development and cellular differentiation.

1.4.2.3 ATP-dependent nucleosome-remodelling

Nucleosome remodeling complexes, including switch/sucrose non-fermentable (SWI/SNF) and nucleosome remodeling and histone deacetylation (NuRD), play essential roles in the regulation of chromatin structure and accessibility, thereby modulating gene expression [121].

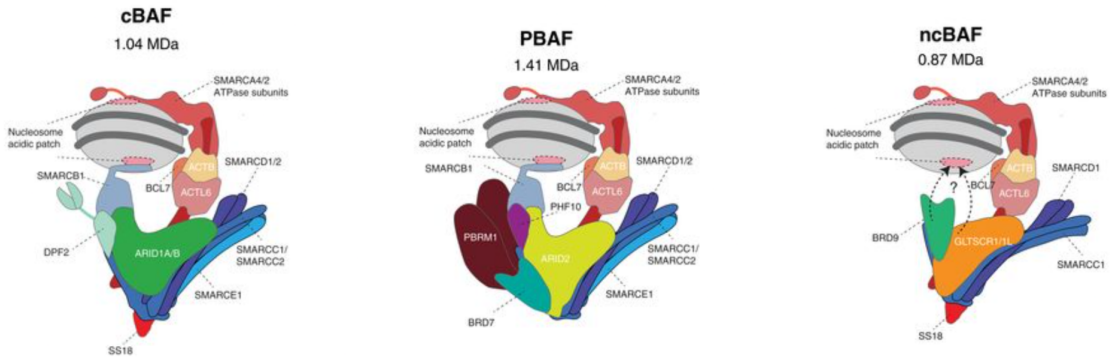


Figure 1.13: **SWI/SNF complex.** All SWI/SNF complex configurations contain an ATPase, either the SMARCA4 or SMARCA2 protein, known as BRG1 and BRM, respectively. These complexes also share common subunits such as SMARCC1, SMARCC2, SMARCD1, ACTL6, ACTB, and BCL7. **cBAF** features either ARID1A or ARID1B as its central component, with SS18 and DPF2 proteins exclusive to this configuration. In **PBAF**, ARID2 is the central component, accompanied by PBRM1, PHF10, and BRD7, which are exclusively present in this configuration. The **ncBAF** complex has GLTSCR1 or GLTSCR1L as its central component, and BRD9 is uniquely found in this configuration. Figure from [122].

SWI/SNF complex SWI/SNF complexes are ATP-dependent chromatin remodelers that use the energy derived from ATP hydrolysis to slide or evict nucleosomes, resulting in increased accessibility of DNA to transcription factors and other regulatory proteins [123]. These complexes are composed of several subunits, including the central ATPase subunit, either BRG1 or BRM, which is responsible for driving the nucleosome remodeling activity [124]. SWI/SNF complexes are diversified into three assemblies, canonical BAF (cBAF), polybromo-associated BAF (PBAF), and noncanonical BAF (ncBAF), differing in constituting members (Fig. 1.13) [125, 126]. All three complexes contain the core subunits, including SMARCC1 (also called BAF155), SMARCC2 (BAF170), SMARCD1 (BAF60) and either of the ATPases SMARCA4 (BRG1) or SMARCA2 (BRM). Still, they also contain numerous variable subunits that give each complex a distinct identity. The canonical complex is the most well-studied variant and is involved in a broad range of biological processes, including transcriptional activation, repression, and maintenance of cell identity [127]. cBAF comprises shared core subunits and accessory subunits such as ARID1A or ARID1B (BAF250), ACTL6A or ACTL6B, BCL7A, BCL7B or BCL7C and SS18, among others (Fig. 1.13) [122]. PBAF is characterized by ARID2 as its central component with PBRM1 and BRD7 exclusively found in this confirmation (Fig. 1.13). PBAF complexes play roles in transcriptional regulation, cell cycle control, and DNA damage response [127]. PBRM1, also known as BAF180, enables PBAF complexes to specifically recognize and bind to acetylated histones through bromodomains, which can influence target gene selection and transcriptional regulation [128]. ncBAF is characterized by the absence of some key cBAF subunits, such as ARID1A/B and SMARCD1/2/3, and the presence of unique subunits, including BRD9 and GLTSCR1 (Fig. 1.13) [125]. Each configuration of SWI/SNF complex shows distinct patterns of occupancy with cBAF found at enhancer sites, PBAF at active promoters, and ncBAF at CTCF sites [126].

NuRD complex Nucleosome Remodeling and Deacetylase (NuRD), also known as Mi-2, complex combines chromatin remodeling and histone deacetylase activities to regulate chromatin structure and gene expression [130]. The NuRD complex contains an ATPase subunit, such as CHD3 or CHD4, which mediates chromatin remodeling, and histone

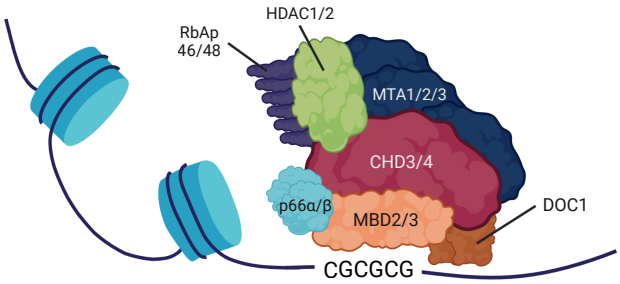


Figure 1.14: **Schematin of the NuRD complex.** NuRD complex comprises of one copy of either CHD3 or CHD4, one copy of MBD2 or MBD3, three of MTA1, MTA2 or MTA3, a dimer of DOC1, six RbAp46/48 and two of p66α/β. Reproduced from [129].

deacetylase subunits, such as HDAC1 or HDAC2, which remove acetyl groups from histone tails, promoting chromatin compaction and transcriptional repression (Fig. 1.14) [131]. NuRD complexes are crucial for the proper functioning of several developmental processes, including embryogenesis, cellular differentiation, and organogenesis [131]. Mutations in the *CHD3* and *CHD4* genes, which encode components of the NuRD complex, have been associated with Snijders Blok-Campeau syndrome, a neurodevelopmental disorder characterized by intellectual disability, speech impairments, and craniofacial anomalies [132].

1.5 Chromatin organization in cancer

Given the crucial role that chromatin organization plays in the regulation of gene expression and cellular identity, it is perhaps not surprising that its dysregulation is associated with the development and progression of cancer. In cancer cells, alterations in chromatin structure can result in abnormal gene expression, leading to the activation of oncogenes and the silencing of tumor suppressor genes [133, 134]. Furthermore, changes in chromatin organization can also contribute to genomic instability, promoting the acquisition of additional mutations that drive cancer progression. As a result, targeting the dysregulation of chromatin organization has emerged as a promising therapeutic strategy for various types of cancer.

1.5.1 DNA methylation

As discussed above, DNA methylation plays a crucial role in regulating gene expression. In cancer, aberrant DNA methylation patterns, including global hypomethylation and site-specific hypermethylation, have been implicated in the development and progression of the disease [135]. These alterations can result in the silencing of tumor suppressor genes, activation of oncogenes, and genome instability, ultimately contributing to tumorigenesis.

One example of cancer-associated hypermethylation is the promoter methylation of the *MLH1* gene in endometrial cancer. *MLH1* is a key component of the DNA mismatch repair (MMR) system, which maintains genomic integrity by correcting base-pair mismatches that arise during DNA replication [136]. Hypermethylation of the *MLH1* promoter in endometrial cancer leads to the silencing of MLH1 expression, impaired MMR activity, and the accumulation of mutations (and microsatellite instability), which can drive cancer development and progression [137, 138]. As mentioned earlier, MMR-deficient EC constitutes a distinct subtype of EC within the molecular classification.

1.5.2 SWI/SNF complex

Mutations and other dysregulation of various SWI/SNF subunits are highly recurrent in cancer, and result in altered chromatin structure and gene expression profiles that can promote tumor development and progression [127].

The most frequently mutated SWI/SNF subunit in cancer is the tumor suppressor *ARID1A*, which is recurrently mutated in multiple malignancies, including ovarian [61], endometrial [13], and gastric cancers [139]. Indeed, *ARID1A* is one of the most frequently mutated cancer genes with alterations in 45.2% of EC and clear-cell ovarian cancers, 18.7% of gastric, 9.4% of colorectal [140]. The importance of SWI/SNF complex function in cancer suppression is further illustrated by the fact that multiple other SWI/SNF subunits are frequently mutated in various cancers. For example, *PBRM1* is frequently mutated in renal cell carcinoma, *SMARCA4* in small-cell carcinoma of the ovary, and *SMARCB1* in the rhabdoid tumor, to name a few [141]. Overall, almost 25% of all cancers harbor a mutation in at least one member of SWI/SNF complex [141].

Given the widespread involvement of SWI/SNF dysregulation in cancer, this complex has emerged as an attractive therapeutic target. Efforts are underway to develop small molecule inhibitors and other therapeutic strategies targeting specific SWI/SNF subunits or their associated regulatory pathways, intending to restore normal chromatin structure and gene expression profiles to inhibit tumor growth and progression [142].

1.5.3 NuRD complex

While less common that SWI/SNF alterations, dysregulation of the nucleosome remodeling and deacetylase (NuRD) complex has been demonstrated in several cancer type. The NuRD complex, a multifunctional chromatin remodeling complex, plays a critical role in controlling gene expression, DNA repair, and cell differentiation. In ovarian cancer, high expression of *CHD4*, a core component of the NuRD complex, was associated with poor prognosis and *CHD4* was found to contribute to metastasis [143]. In endometrial cancer, mutations in *CHD4* were found to promote endometrial cancer stemness [144].

1.5.4 Polycomb repressive complex

Polycomb repressive complexes (PRCs) play an important role in cancer by aberrent regulation of gene expression through chromatin modification and remodeling. PRCs are involved in various biological processes, including cell differentiation, proliferation, and stem cell maintenance. In cancer, dysregulation of PRC function can lead to the aberrant silencing of tumor suppressor genes and activation of oncogenes, thus promoting tumor growth and progression [145]. One notable component of the PRC1 family is BCL6 corepressor (BCOR), which has been implicated in several cancers, including acute myeloid leukemia and medulloblastoma [146, 147]. Alterations in BCOR expression or function can contribute to PRC dysregulation, leading to the disruption of normal cellular processes and the promotion of oncogenic pathways.

1.5.5 Dysregulation of chromatin in cancer initiation and progression

Alterations in chromatin modifiers have been demonstrated to contribute to cancer progression and metastasis. For instance, in breast cancer, the NuRD complex, recruited by RUNX2, has been shown to promote cancer progression and bone metastasis [148]. PRC2 and KMT2D-COMPASS were identified as regulators of epithelial-mesenchymal plasticity, contributing to metastasis [149].

Several EC drivers, including *PIK3CA*, *PIK3R1*, and *KRAS* were found to be mutated in normal endometrial tissue [33, 37]. PTEN loss, detected by immunohistochemistry, was observed in 27% of normal endometrium samples, while hotspot *KRAS* and PIK3CA mutations were detected in 28% and 14% of samples, respectively [33]. No patients exhibited a loss of ARID1A [33], suggesting that perturbation of ARID1A could contribute to disease progression.

To better understand the role of chromatin modifiers and chromatin organization in endometrial cancer, it is essential to investigate the sequence of genomic alterations and the tumor's evolution. I discuss my work on understanding cancer evolution in Section 2.3.5 and Chapter 4.

1.6 Heterogeneity and cancer evolution

1.6.1 Tumor heterogeneity

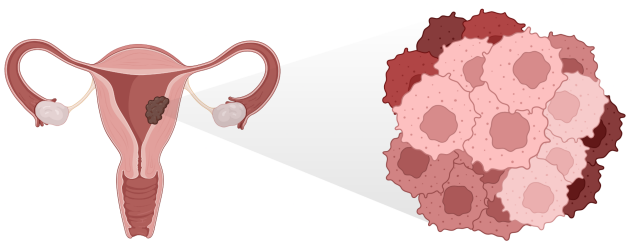


Figure 1.15: **Illustration of tumor heterogeneity.** Created with BioRender.com

Tumor heterogeneity can refer to either *inter*tumoral heterogeneity, which is the het-

erogeneity between different tumors within a cancer type, or *intra*tumoral heterogeneity, which refers to diverse subpopulations of cells within a single tumor, exhibiting distinct genetic, phenotypic, and functional traits (Fig. 1.15) [150–154]. In this work, I will use the term tumor heterogeneity to refer to intratumoral heterogeneity.

Tumor heterogeneity arises from the complex interplay of genetic and epigenetic alterations, selective pressures exerted by the tumor microenvironment, and stochastic processes that drive tumor evolution and proliferation through complex tissue admixtures. Intratumoral heterogeneity has significant clinical implications, as it contributes to therapy resistance, disease recurrence, and patient outcomes [153–155]. Moreover, tumor heterogeneity poses a challenge to personalized medicine, as a single biopsy may not capture the full range of genetic and phenotypic diversity present within a tumor [153]. Consequently, there is a growing need to develop more sophisticated methods for characterizing and understanding tumor heterogeneity to facilitate the development of effective and targeted therapeutic strategies.

1.6.2 Cancer evolution

Throughout tumor development, constituent cells acquire multiple alterations that contribute to heterogeneity, as illustrated in Fig. 1.16. Some of these will be drivers - genetic and epigenetic changes that bestow selective advantages. In contrast, passenger alterations, though more frequent, exert minimal to no influence on the tumor's growth and survival.

When a cell obtains a driver mutation that confers a selective advantage (e.g., an enhanced growth rate), it establishes a clone that is now in competition with its parent (Fig. 1.16A). Consequently, this clone can expand, potentially out-competing other clones (Fig. 1.16B). Throughout the tumor's lifespan, various compositions of subclonal populations are present. Analyzing the tumor at a single time point can only offer a picture of its current state and a limited glimpse of its dynamic evolutionary progression (Fig. 1.16C). Furthermore, the spatial location of the biopsy sampling may impact the detection of less prevalent clones in the analysis. It is important to highlight that analyzing a single time point provides only a glimpse of the evolutionary history of the tumor.

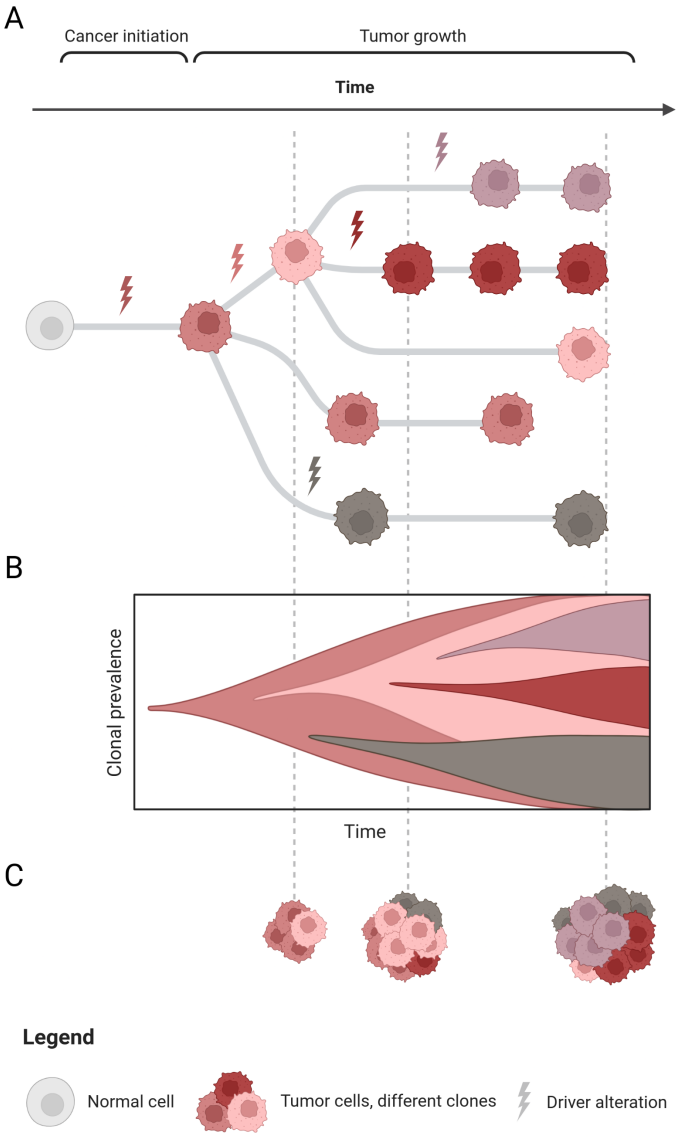


Figure 1.16: **Illustration of branched tumor evolution. A** Clones in tumors evolve by accumulating genetic and epigenetic alterations (mutations, copy number alterations, etc.). First, the initial set of alterations for transformation from a normal to a malignant cell is necessary. Then, clones further accumulate driver alterations (annotated by lightning bolts) that lead to branched lineages from their parental clones. **B** Representation of prevalence of clones in branched evolution. New clones arise with an evolutionary advantage and subsequently undergo clonal expansion, out-competing the original cells. **C** Illustration of tumor cell composition at three different time points. Each cell color indicates the cell corresponds to a distinct clone. Created with BioRender.com

1.7 Motivation of this study

Chromatin organization serves as a central regulator of gene expression and cell phenotype, with its dysregulation playing a significant role in cancer development. In endometrial cancer, the most prevalent gynecological malignancy and a leading cause of morbidity and mortality, chromatin dysregulation through genomic alteration of its modifiers is particularly common. As discussed in this Chapter, *ARID1A*—a key component of the SWI/SNF remodeling complex—is among the most frequently mutated driver genes in endometrial cancer. Other chromatin remodelers, such as *CHD4* and *BCOR*, have also been identified as drivers in endometrial cancer. Despite this, a systematic analysis of alterations in chromatin modifiers and their functional consequences in endometrial cancer has yet to be performed. Furthermore, the study of somatic mutation in normal endometrium showed that while several cancer driver genes are frequently mutated in normal tissue, the mutations in *ARID1A* are notably absent in that tissue. This distinction underscores the unique pathogenic role of *ARID1A* mutations in endometrial cancer, further highlighting the importance of investigating chromatin dysregulation in this disease.

To address this gap, this study first aims to conduct a comprehensive analysis of genomic alterations in chromatin modifiers, as detailed in Chapter 2. This analysis, leveraging the largest cohort of endometrial cancers with extensive DNA sequencing to date, aims to elucidate the spectrum and prevalence of these alterations. Understanding these genomic changes provides a foundational basis for subsequent functional studies.

Furthermore, in this thesis, I delve into the functional consequences of perturbations in chromatin modifiers identified based on the research described in Chapter 2. Employing CRISPR-Cas9 gene editing, our research investigates the impacts of these alterations in normal and malignant endometrial cells. The outcomes, including changes in transcriptomic profiles and chromatin accessibility, are thoroughly detailed in Chapter 3. This approach bridges the gap between genomic alterations and their functional implications, illuminating the pathogenic mechanisms in EC.

In Chapter 2, I explore endometrial cancer tumor heterogeneity to understand the timing of chromatin dysregulation. In Chapter 4, I discuss the challenges associated with

current methods for ordering tumor alterations and uncovering evolutionary trajectories. I also propose a deep learning solution that could aid in ordering genomic events and contribute to understanding the order of occurrence of alterations to chromatin modifiers.

In this thesis, I aim to investigate the hypothesis that chromatin modifiers play a crucial role in EC pathogenesis by affecting gene expression and cellular phenotype. I demonstrate the extent of genetic alterations in chromatin modifiers in Chapter 2, their prognostic value in Section 2.3.4, and explore their timing in Section 2.3.5. Furthermore, I explore the consequences of the loss of chromatin modifiers in Chapter 3.

In Chapter 4, I propose and investigate the framing of cancer evolution as a sequence of critical and background alterations analogous to the ordering of words in a sentence. With this effort, I hope to contribute to addressing the critical question of the timing and sequence of chromatin modifier alterations and how they influence the evolutionary trajectory and heterogeneity of EC tumors.

Chapter 2

Whole genome and chromatin landscape in endometrial cancer

Contents

2.1 Background 41
2.1.1 Molecular subtypes of Endometrial Cancer 42
2.1.2 Molecular landscape of the coding EC genome 44
2.1.3 Chromatin modifiers in Endometrial Cancer 44
2.1.4 Unanswered questions 46
2.2 Introduction 47
2.2.1 Genomics England endometrial cancer cohort 47
2.2.2 GEL EC cohort - overview 48
2.2.3 Contributions 49
2.3 Results 49
2.3.1 Molecular EC subgroups in the GEL EC cohort 49
2.3.2 Chromatin modifiers present among genes identified as drivers . 65
2.3.3 Chromatin remodeling pathways are recurrently altered in endometrial cancer 77
2.3.4 Relationship between molecular alterations and overall survival 81
2.3.5 Temporal Heterogeneity 90
2.4 Discussion 96
2.5 Methods 102
2.5.1 Genomics England data 102
2.5.2 TCGA somatic variation 109

2.1 Background

Endometrial cancer (EC) is the most common gynecological cancer in the UK [156], US [2], and countries with high Human Development Index [1, 3]. The incidence rate has

increased by >132% in the past two decades [4].

As discussed in more detail in Section 1.1.1, EC can be split into endometrioid and non-endometrioid endometrial cancer. Endometrioid endometrial cancer accounts for 80% of all EC cases. Serous endometrial carcinoma (SEC), uterine carcinosarcoma (UCS), and clear cell endometrial carcinoma (CCEC) are the most common types of non-endometrioid EC accounting for 10%, $< 5\%$ and $< 5\%$, respectively. Low-grade (grade 1 and 2) EEC corresponds to type I of the Bokhman classification, while high-grade EEC and non-endometrioid cancers correspond to type II. Although survival is generally good (5-year survival $\geq 86\%$) [5] for type I cancers, the outcome in high-risk molecular subtypes is considerably poorer.

2.1.1 Molecular subtypes of Endometrial Cancer

As discussed previously in Section 1.2.4, molecular subtyping of EC according to TCGA classification holds prognostic and predictive value. The four categories defined by TCGA [13], based on the samples molecular profiles, consist of:

- POLE subtype - ultramutated cases with polymerase epsilon (*POLE*) exonuclease domain mutations;
- MSI - hypermutated tumors with microsatellite instability (MSI) caused by deficient DNA mismatch repair;
- CN high - lowly mutated tumors with a high burden of CN alterations; and
- CN low - lowly mutated tumors with a low somatic copy number (CN) alterations burden.

Initially, the TCGA study examined 232 tumors, followed by further analysis and the release of additional data on an increased number of samples. In subsequent references to the TCGA EC cohort, I will be referring to the UCEC and UCS cohorts as described in [157]. The UCEC cohort included in the PanCancerAtlas study consists of 507 samples with 32%, 28%, 24%, and 16% of samples assigned to CN high, CN low, MSI, and POLE subtypes, respectively (Fig. 2.1). While UCS samples were not formally assigned to the molecular subtypes defined in TCGA UCEC studies, they resemble the CN high samples Fig. 2.2 [15].

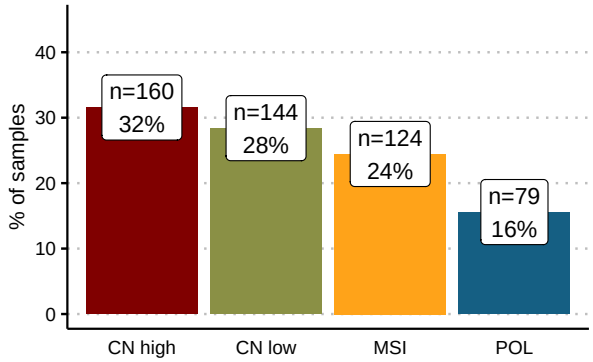


Figure 2.1: **Molecular subtypes among samples in the UCEC TCGA cohort.**

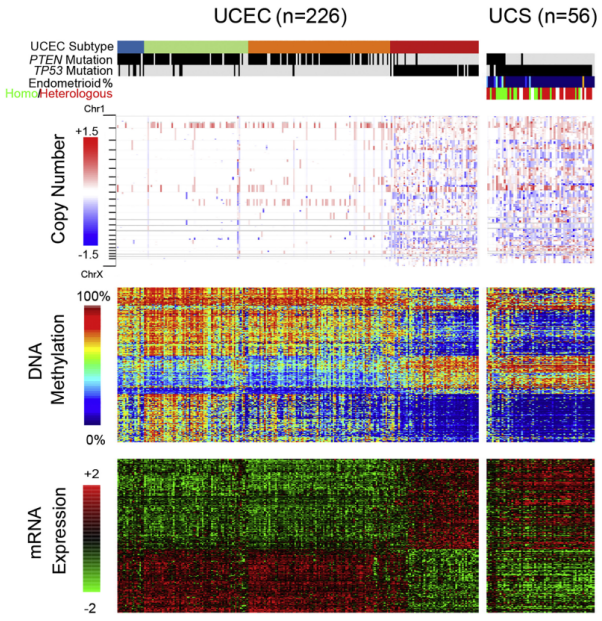


Figure 2.2: **Comparison of TCGA UCEC and UCS cohorts.** UCS samples resemble CN high from UCEC cohort. Adapted from [15].

2.1.2 Molecular landscape of the coding EC genome

TCGA described the coding genome of the EC, noting the differences in the tumor mutation burden across the molecular subtypes - with the POLE subtype characterized by an order of magnitude more mutations than the MSI (hypermutated) subtype. CN low and CN high are characterized by 100 times less mutation per Mb than POLE samples. I discuss the molecular landscape of EC in more detail in Section 1.2.5.

2.1.3 Chromatin modifiers in Endometrial Cancer

As discussed in Section 1.4, chromatin organization plays an important role in gene expression and regulation of various cell processes. It has also been implicated in many cancers (Section 1.5). SWI/SNF is one of the main chromatin remodeling complexes and mutations in *ARID1A* - a core subunit of the SWI/SNF complex are over-represented in uterine cancer (Fisher's Exact test on TCGA mutation data: endometrial cancer vs. other cancer types, p-value $< 2.2 \times 10^{-16}$). Over 40% UCEC samples containing a mutation in *ARID1A* gene - more than any other cancer type - including those with a similar or higher median mutation rate. When counting mutations across subunits of SWI/SNF, the fraction of samples with mutations in any of them in UCEC is equal to 58% (Fig. 2.3).

2.1.3.1 Chromatin modifiers in other diseases of the endometrium

Endometriosis is a disease in which endometrium-like tissue is present outside the uterus [32]. Lesions can be found in the ovaries, fallopian tubes, bladder, and other body parts. Several EC drivers - including *PIK3CA*, *PIK3R1*, *FBXW7*, *CHD4* and *KRAS* were found to be mutated in normal endometrial tissue [33, 37] as well as endometriotic tissue with the absence of cancer [32, 34, 35]. PTEN loss detected by IHC was observed in 27% of samples of normal endometrium, and hotspot *KRAS* and *PIK3CA* mutation were detected in 28% and 14% of samples, respectively. No patients exhibited loss of ARID1A [33].

As discussed in more detail in Section 1.2.2, normal endometrium, as well as endometriosis tissue, often accumulate mutations in known cancer genes. However, *ARID1A* mutations, prevalent in the endometriotic epithelium, were either absent or significantly

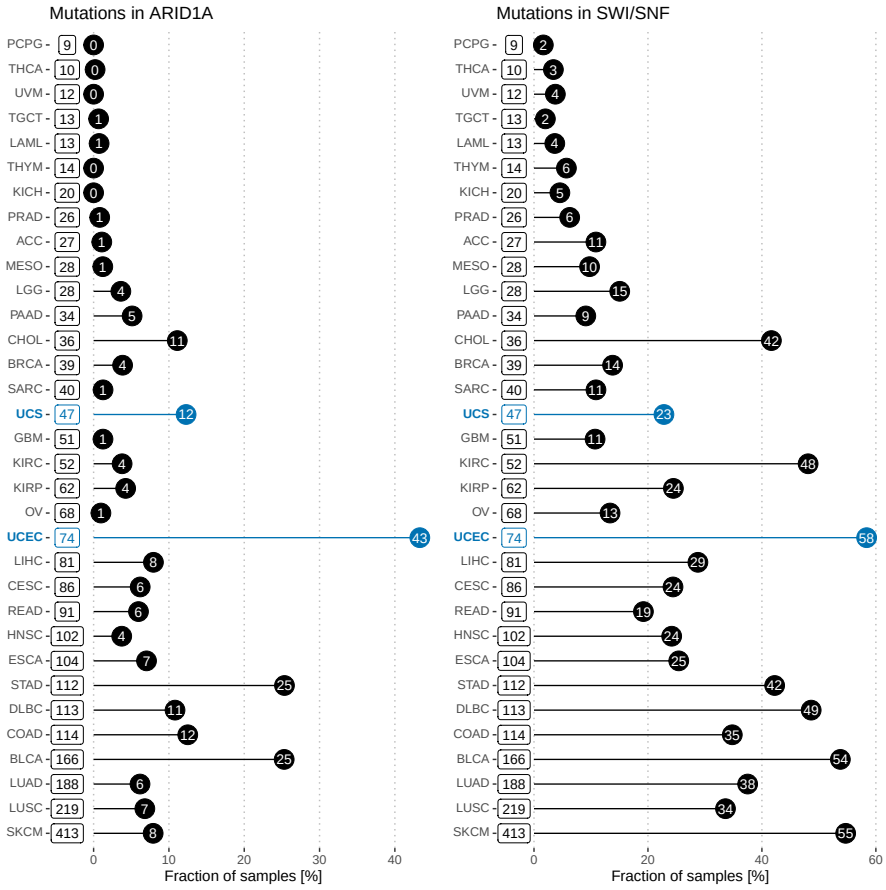


Figure 2.3: **Fraction of samples with mutations in *ARID1A* (on the left) and across SWI/SNF complex (on the right).** Mutations accessed from MC3 dataset [158], only non-silent coding mutations counted. Cancer types are ordered based on the median number of mutations across samples. TCGA EC cancer types (UCEC and UCS) are highlighted in blue.

less frequent in normal endometrium (Fig. 2.4) [34, 35, 37].

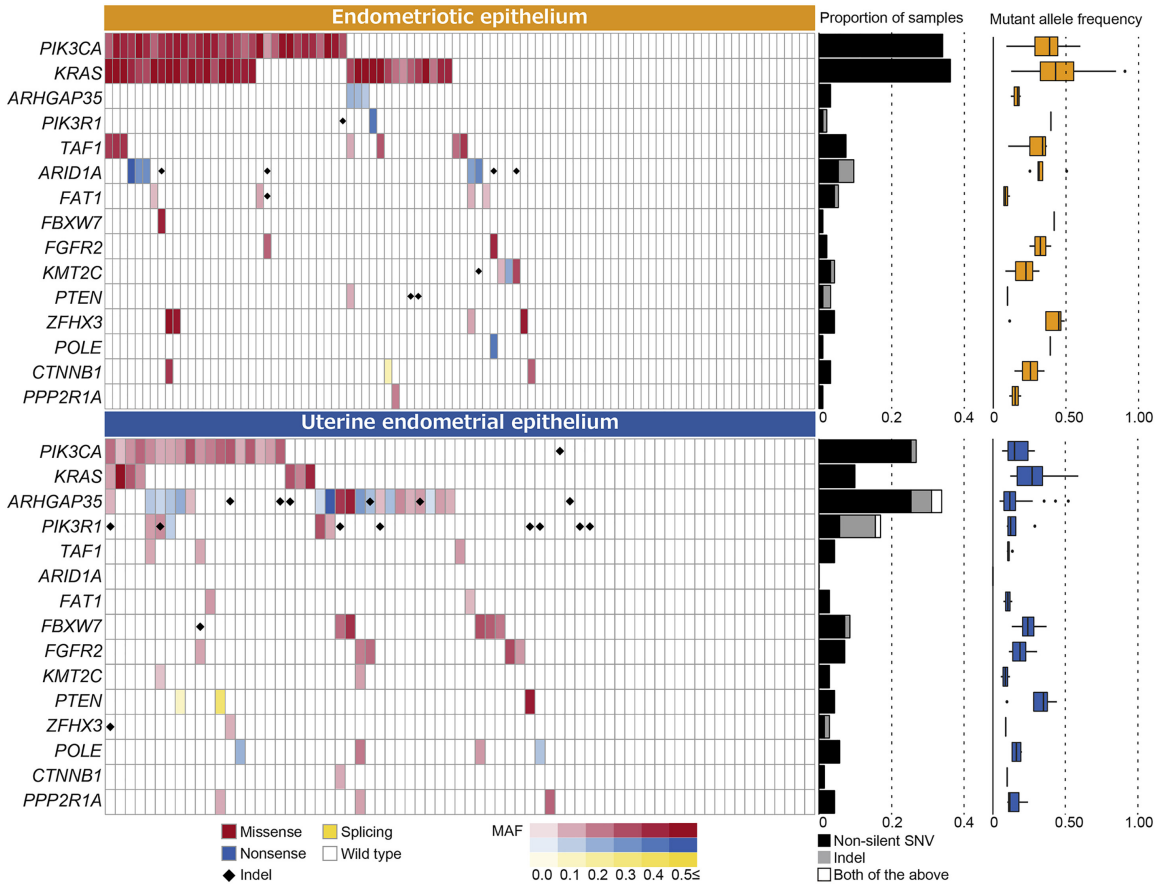


Figure 2.4: **Mutation landscape of endometriotic and uterine endometrial epithelium samples.** Missense, nonsense, splicing and indel mutations in samples endometriotic epithelium (top panel) and normal endometrium (bottom panel). Mutation allele frequency (MAF) of all but indel mutations shown in the intensity of the color. Fraction of samples with mutations in the barplot and distribution of MAFs in the boxplot shown on the right. Figure from [34].

2.1.4 Unanswered questions

TCGA generated an irreplaceable asset for the research community and provided major insights into the molecular biology of endometrial (and other) cancers. TCGA data is freely available to the research community alongside a suite of computational packages to streamline downloading and analyzing the data (for example, R TCGAbiolinks package [159]) and online applications to explore the datasets. However, several research questions and technical considerations remain ten years after the endometrial cancer landscape publication.

Mutation data available in the TCGA dataset covers only the coding sequences of the cancer genomes. The Pan-Cancer Analysis of Whole Genomes (PCAWG) study was an international collaboration of the International Cancer Genome Consortium (ICGC) and TCGA consortia set up to leverage WGS to analyze the somatic and germline variations in coding and non-coding genome, explore the mutational signatures, cancer heterogeneity and identify structural variation. The pan-cancer analyses published in 2020 provided a valuable resource in the space of pan-cancer analyses. However, only a fraction of cancer types were represented in the samples set - for example, only 50 samples from endometrial cancer (solely from the TCGA UCEC cohort) were included. Additionally, the tumor samples had a bimodal coverage distribution with modes at 38×, and 60× [160]. The above can explain the lack of discovery power in the non-coding region analyses [161].

ARID1A has been suggested as a causative gene of microsatellite instability [60], and selected other chromatin modifiers have been described and studied in the context of endometrial cancer. However, systematic analysis of alteration in chromatin modifiers and their functional consequences in EC, outside of small sample-sized, targeted studies [162], has not been performed.

2.2 Introduction

In this chapter, I describe the biggest to-date EC cohort and discuss findings obtained from the genomic analysis of this cohort. Next, I will explore the alterations in chromatin modifiers with respect to EC molecular subtypes and genomic features.

2.2.1 Genomics England endometrial cancer cohort

The Genomics England (GEL) 100,000 Genomes Project was set up in order to sequence 100,000 genomes from around 75,000 individuals with rare diseases and cancers. Participants were recruited from standard NHS care at genomic medicine centers; no preselection was carried out or quotas established.

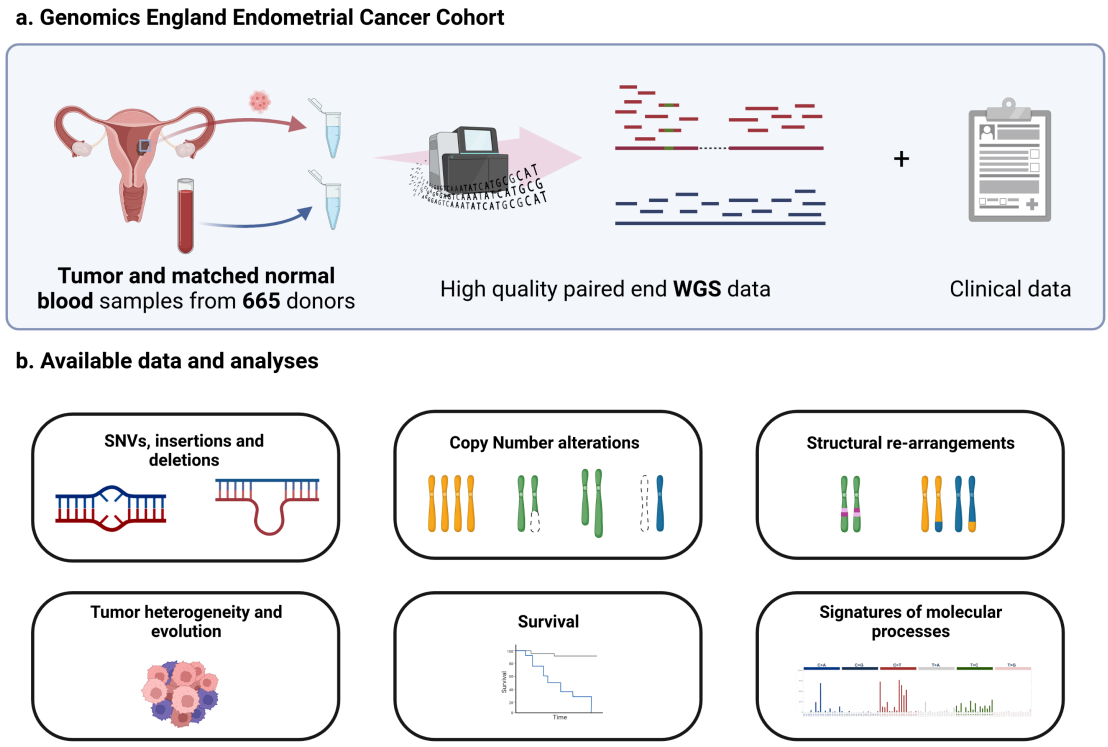


Figure 2.5: **Overview of the data available in the GEL EC cohort**. Created with BioRender.com

2.2.2 GEL EC cohort - overview

After careful exclusion of samples from non-endometrial cancers and detailed quality control (manuscript in preparation), the GEL EC cohort includes samples from 665 participants, making it the biggest cohort of endometrial cancers. Most cases were primary tumors resected by hysterectomy (657 samples, 99%), while eight samples came from metastatic tissue.

The dataset includes whole genome sequencing (WGS) at an average depth of 100x from both - tumor and matched blood (normal reference) tissue.

Leveraging WGS sequencing, we were able to describe the mutational landscape of the full EC genome, identify driver genes, observe signatures of molecular processes, identify copy number alterations, and perform heterogeneity analysis Fig. 2.5b. Analysis of whole-genome sequencing together with clinical data enabled survival analyses that showed the prognostic value of driver mutations and other molecular alterations in this real-world cohort.

2.2.3 Contributions

Analysing the GEL EC cohort is a collaborative effort undertaken by the Endometrial Cancer Genomics England Clinical Interpretation Partnership (EC GeCIP) consortium which I am part of.

Alphabetical list of EC GeCIP contributors

- Ahmed Ahmed
- Claudia Arnedo-Pac (CAP)
- James Brenton
- Tjalling Bosse
- Giulio Caravagna
- Daniel Chubb (DC)
- David N. Church (DNC)
- Alex J. Cornish (AJC)
- Emma Crosbie
- William C. H. Cross
- Richard Edmondson
- Anna Frangou (AF)
- Trevor Graham
- Andreas Gruber (AG)
- Aliah Hawari (AH)
- Jacob Househam
- Richard S. Houlston
- **Katarzyna Kedzierska (KK)**
- Ben Kinnersley
- Eszter Lakatos
- Nuria Lopez-Bigas
- Nirupa Murugaesu
- Boris Noyvert (BN)
- Alona Sosinski
- Ignacio Soriano
- Andrea Sottoriva
- Avraam Tapinos (AT)
- Ian Tomlinson
- Steve Thorn
- David Wedge
- Helen White
- Genomics England (GEL) Research Consortium

2.3 Results

2.3.1 Molecular EC subgroups in the GEL EC cohort

In the GEL EC cohort, samples were assigned to the (1) POLE subtype if they have a pathogenic mutation in the polymerase-ϵ exonuclease domain (as defined in [163]), (2) MSI status identified by microsatellite instability detection from genomic data using mSINGS [164]; microsatellite stable (MSS) samples were assigned to (3) CN high and (4) CN low subtypes based on the fraction of aberrant genome (more details on assignments to

Analysis	Contributors
Sequencing, mapping and variant calling	GEL Research Consortium
Data QC	AJC, BK, GC, AG
Subtypying	AJC, AG, DNC
Driver identification	CAP, BK, DNC Visualisation: **KK**
Structural variant calls	AJC
Copy number analysis	Copy number calls: AF Significant copy number alterations: BK Processing, visualisation: **KK**
Molecular signatures	Mutations: AG Copy number: AH Structural variants: DC, AT
Clinical data	AJC, BN
Alterations in chromatin modifiers	**KK**
Survival analysis	**KK**
Heterogeneity analysis	**KK**

Table 2.1: **Contributions to analyses relevant to the content included in the thesis.**

molecular subtypes available in Section 2.5.1.2).

While CN high, with 32% of samples, was the most abundant subtype in TCGA cohort Fig. 2.1 in the GEL EC cohort, 22% of samples are assigned to CN high molecular subgroup. CN low is the most abundant subtype with 41%, followed by MSI - 30% and with 6% of samples associated with POLE subtype Fig. 2.6. The differences in the molecular subtypes distributions can be attributed to differences in study recruitment. TCGA followed a targetted accrual (20% of samples with Grades 1 and 2, 20% with Grade 3, and 10% with serous cancer type), while patients in the GEL cohort were not following predefined quotas.

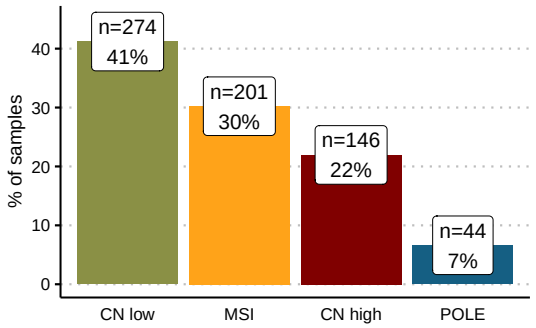


Figure 2.6: **Molecular subtypes among samples in the GEL EC cohort.**

2.3.1.1 Mutations across molecular subgroups

	EC GEL		TCGA UCEC [13]
	whole genome	coding genome	coding genome
CN low	1.5	1.8	2.3
CN high	2.1	2.0	2.9
MSI	29.3	24.0	18.0
POLE	270.6	181.5	232.0

Table 2.2: **Tumor mutation burden across molecular subtypes.** Mean tumor mutation burden (TMB, number of mutations per Mb) across molecular subtypes in EC GEL cohort and TCGA UCEC cohort shown.

As described in Section 1.2.5, TCGA characterized tumor mutation burden variation across molecular subtypes in the coding genome of 248 endometrial cancers. Here, I describe the mutational landscape across the whole genome of 665 endometrial cancers (Fig. 2.7 and Table 2.2).

In analyzing the genomic mutation profiles, CN low samples were observed to have the lower mutational burden, indicated by a median count of 3,941 SNVs and 533 indels (Fig. 2.7 and Table 2.2). This baseline mutational load contrasts with the CN high samples, which exhibit a marginally elevated mutational count, while the ultramutated POLE samples present a dramatically different profile, with a median increased by factors of 127 and 187 in SNVs and indels, respectively, compared to CN high and CN low samples (Fig. 2.7 and Table 2.2). In this mutation spectrum, hypermutated MSI samples represent an intermediate category, with a median of 91,254 mutations (Fig. 2.7 and Table 2.2), and notably, they exhibit the highest indel proportion, constituting 59% of their total mutations.

Similarly to coding and non-coding genome size proportions, on average, 1% of all mutations are found in coding regions. (Fig. 2.7)

2.3.1.2 Copy number alterations

Copy Number high subtype was identified a subtype with increased frequency of the copy number alterations. In GEL cohort CN high were separated from CN low samples based on the value of the Percent of genome aberrant (PGA) - threshold for assignment

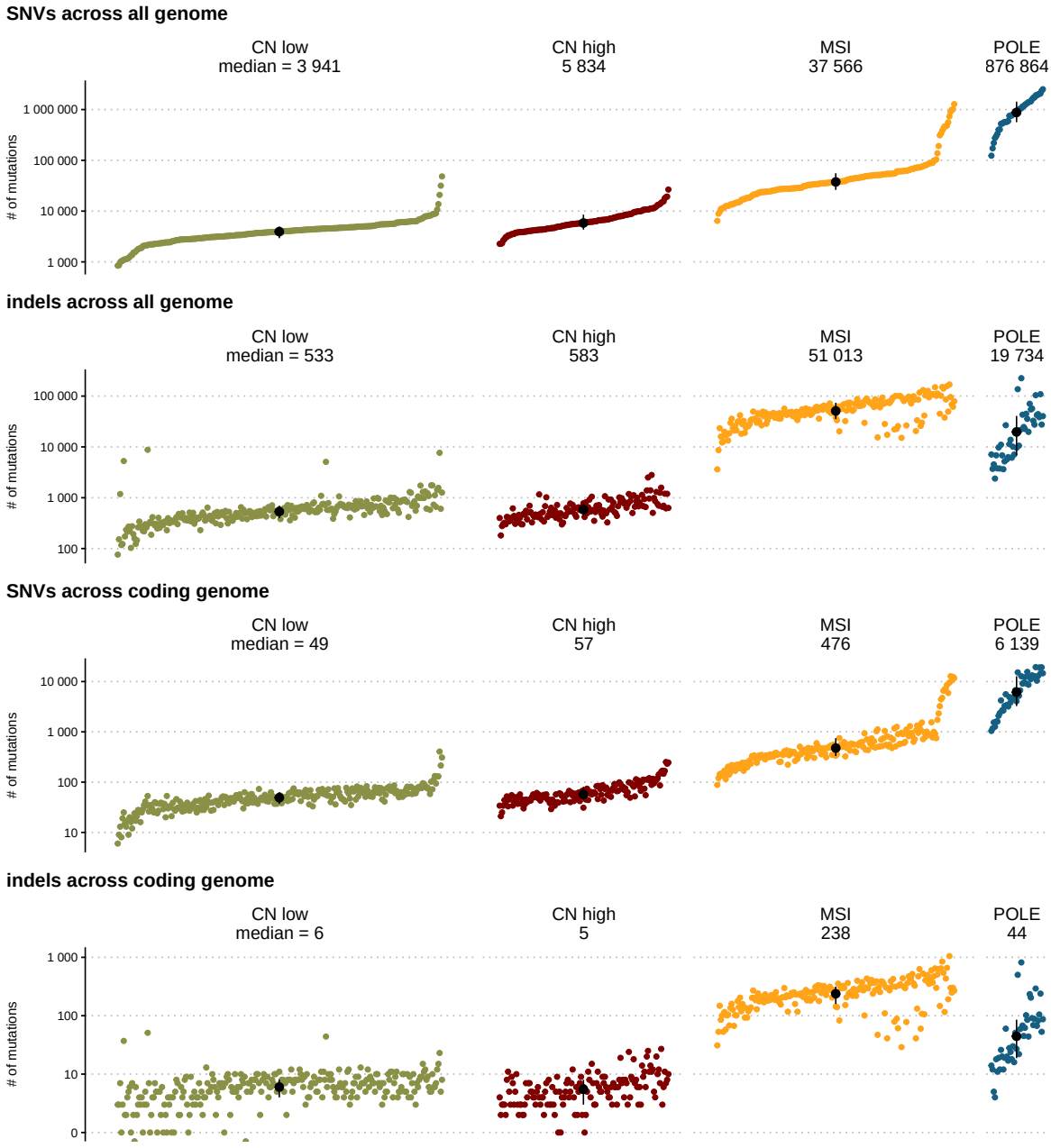


Figure 2.7: **POLE samples accumulate 100 times more variants than CN low and CN high samples.** Number of SNVs and short insertions and deletions across samples in all and only coding regions shown. Samples are sorted according to SNVs across the whole genome, ordering preserved through all mutation subsets. Although there is a correlation between the number of SNVs and indels, the relationship is not perfect as visible by a lower degree of order in subsequent panels. I discuss the relationship between the number of SNVs and indels in Section 2.3.4.2.

to either subtypes was estimated after identifying the minimum value of a smoothed binomial distribution. Median PGA for CN low and CN high samples is equal to 9.49% and 79.02%, respectively. MSI and POLE samples have less copy number alterations with median PGAs equal to 6.74% nd 1.37%, respectively. Fig. 2.8

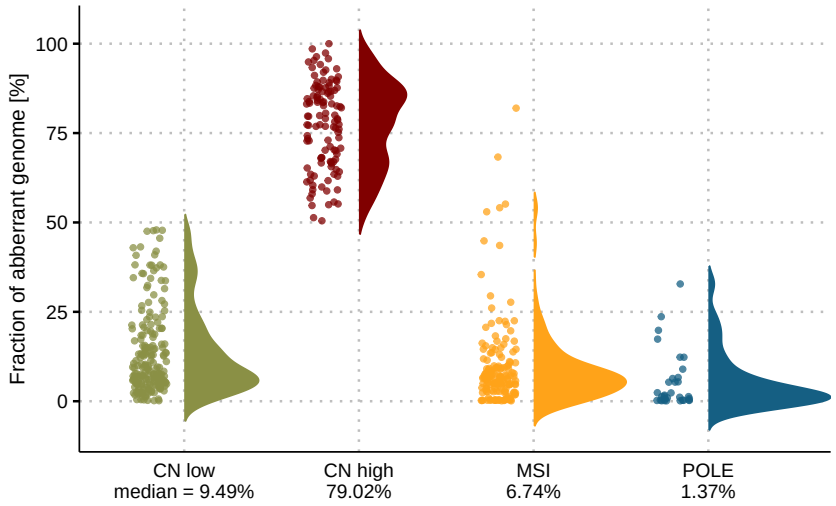


Figure 2.8: **Burden of CNAs assessed by the percentage of genome aberrant (PGA) according to endometrial cancer molecular subtype.** - a fraction of the genome affected by copy number alterations. Samples were assigned to CN high and CN low subtypes based on PGA, with the threshold chosen as the minimum of the smoothed binomial distribution.

2.3.1.3 Signatures of molecular processes

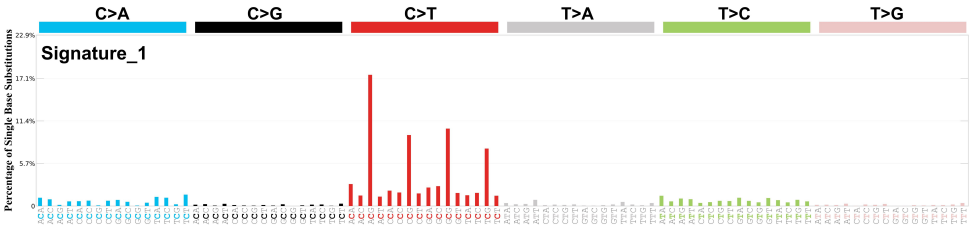
Somatic alterations - mutations, copy number changes, and structural variants - are caused by exogenous and endogenous mutational processes. Each mutational process generates a characteristic mutational signature. Individual cancer genomes incorporate multiple mutational signatures, as the accumulated alterations result from multiple mutational processes.

Compared to prior WES, the WGS available for the GEL EC cohort provided substantially greater power to identify mutational signatures, owing to the greater breadth of sequence and, thus, mutation data.

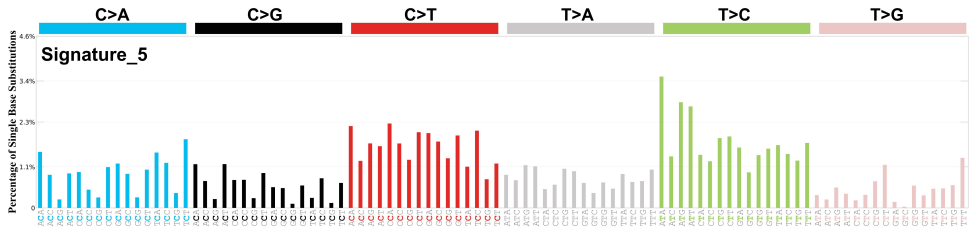
A small subset of UCEC samples (n=51) was included in the PCAWG study investigating the mutational signatures. Signatures of single base substitutions 1 (SBS1) and 5 (SBS5) Fig. 2.9 were found across all cancers, including UCEC. Both signatures correlate

with the age and division rates of tissues. Other signatures found in the EC samples include SBS10a and SBS10b associated with *POLE* mutation, SBS6, SBS15, SBS26, and SBS44 associated with defective DNA mismatch repair. [165]

We extracted the signatures from the genomes across the subtypes (Fig. 2.10). As expected, we observe the expected SBS1 and SBS5 signatures across all subtypes. *POLE* mutation signatures, SBS10a, SBS10b, and SBS14, are present in the POLE subtype. In agreement with previous findings, we also observe signatures of defective DNA mismatch repair SBS15, SBS26 and SBS44, as well as SBS21. APOBEC activity signatures, SBS2 and SBS13 are present in CN high and fewer CN low and MSI samples.



(a) **Signature SBS1** - deamination of 5-methylcytosine to thymine which generates G:T mismatches in double-stranded DNA. Clock-like signature.



(b) **Signature SBS5** - signature associated with unknown etiology. Clock-like signature.

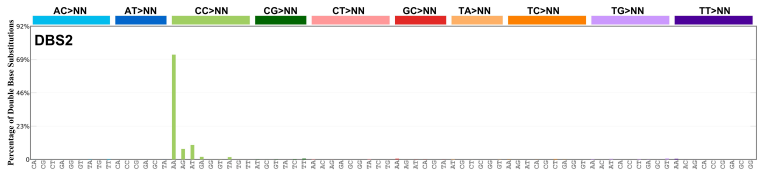
Figure 2.9: **Profiles of SBS signatures present across all cancer types.** Profiles downloaded from COSMIC Mutational Signatures (v3.3 - June 2022) website [166], [165].

Among doublet base substitutions (DBS) identified in UCEC samples in the PCAWG study was the DBS2 Fig. 2.11a, a signature associated with tobacco smoking, and DBS4, DBS6 and DBS9 - signatures of unknown etiology [165].

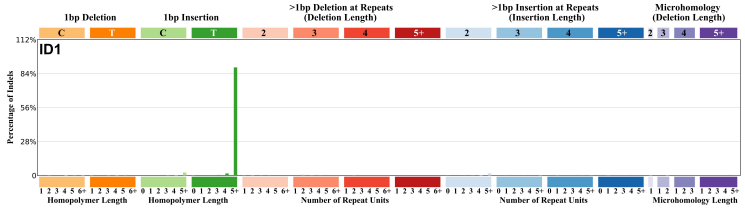
As the PCAWG study, we observe DBS2 - signature associated with Tobacco smoking present among the samples, predominately in CN low and CN high samples. We also observe *POLE* mutation signature DBS3 and defective DNA mismatch repair (DBS10) in POLE subtype with residual activities across other subtypes (Fig. 2.10).



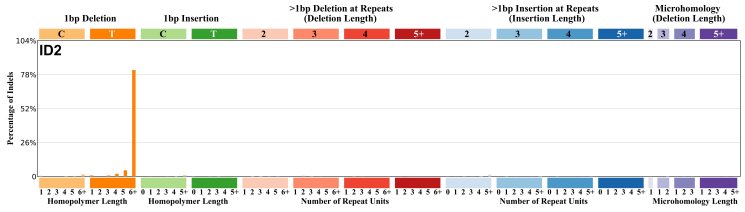
Figure 2.10: **Mutational signatures across EC subtypes. SBS** single base substitutions signature, **DBS** double base substitutions signatures, **ID** indels signatures, **CN** copy number signatures, **SV** structural variation signatures.



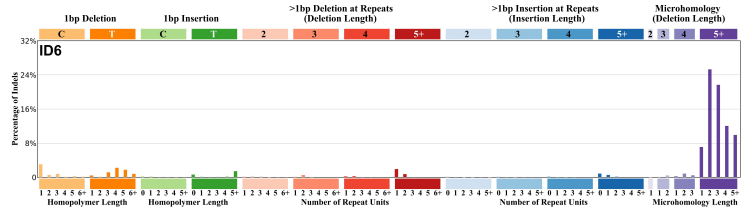
(a) **Signature DBS2** - associated with exposure to tobacco smoking.



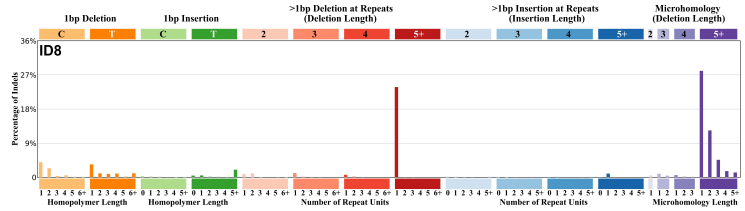
(b) **Signature ID1** - signature of slippage during DNA replication of the replicated strand. Predominately 1bp insertions of T in long ($\geq$ 5) thymine mononucleotide repeats.



(c) **Signature ID2** - signature of slippage during DNA replication of the template DNA strand. Predominately 1bp deletions of T in long ($\geq$ 5) thymine mononucleotide repeats.



(d) **Signature ID6** - repair of DNA double strand breaks by non-homologous DNA end-joining mechanisms. Characterized predominantly by $\geq$ 5-bp deletions.



(e) **Signature ID8** - repair of DNA double strand breaks by non-homologous DNA end-joining mechanisms. Characterized predominantly by $\geq$ 5-bp deletions.

Figure 2.11: **Profiles of DBS and ID signatures present in UCEC samples analysed in PCAWG study.** Profiles downloaded from COSMIC Mutational Signatures (v3.3 - June 2022) website [166], [165]

From small insertions and deletions (ID) signatures associated with replication slippage and defective DNA mismatch repair (ID1 Fig. 2.11b, and ID2 Fig. 2.11c) and double-strand repair by non-homologous end-joining (ID6 Fig. 2.11d, and ID8 Fig. 2.11e) were identified in the UCEC samples [165].

We were able to recapitulate the presence of signatures ID1 and ID2 - associated with replication slippage and ID6 and ID8 - associated with double-strand break repair. Interestingly, signatures ID1 and ID2 are the only ones identified in POLE and MSI subtypes. Both subtypes accumulate more insertions and deletions due to the defective DNA mismatch repair, which dominates the signature decomposition (Fig. 2.10).

Copy Number (CN) signatures identified in UCEC and UCS include signatures of diploidy (CN1) and tetraploidy (CN2), chromothripsis before whole genome duplication (WGD)(CN6), diploid chromosomal instability (CIN) (CN9) and homologous recombination deficiency (HRD) (CN17). Focal loss of heterozygosity (LOH) and double WGD (CN11) were identified in UCS only [167].

Across the four subtypes, we observe CN1 dominating all but CN high subtypes, where CN2, a signature for tetraploidy, is associated with copy number alterations in 62% of samples. Instead of CN11, we identify CN15, which is associated with chromosomal LOH and double WGD.

At the time of writing, structural variant (SV) signature analysis is still underway. Here, I included an approximation of the SV signature describing the SV calls. As anticipated, SVs were enriched in CN high subtype. A detailed description of SV is beyond the scope of this work.

Using the outputs of the mutational signatures, I performed unsupervised hierarchical clustering based on the fraction of alterations assigned to each signature (Fig. 2.12, distance based on Kendall rank correlation as described in Section 2.5.1.3). This agnostic clustering identified 5 clusters that mostly agreed with the molecular subtypes. Clusters 1 and 2 contain most of the CN high samples, and cluster 3 contains mostly CN low samples. Cluster 4 contains exclusively POLE samples and one MSI sample. Cluster 5 includes almost all MSI samples and 10% of CN low samples.

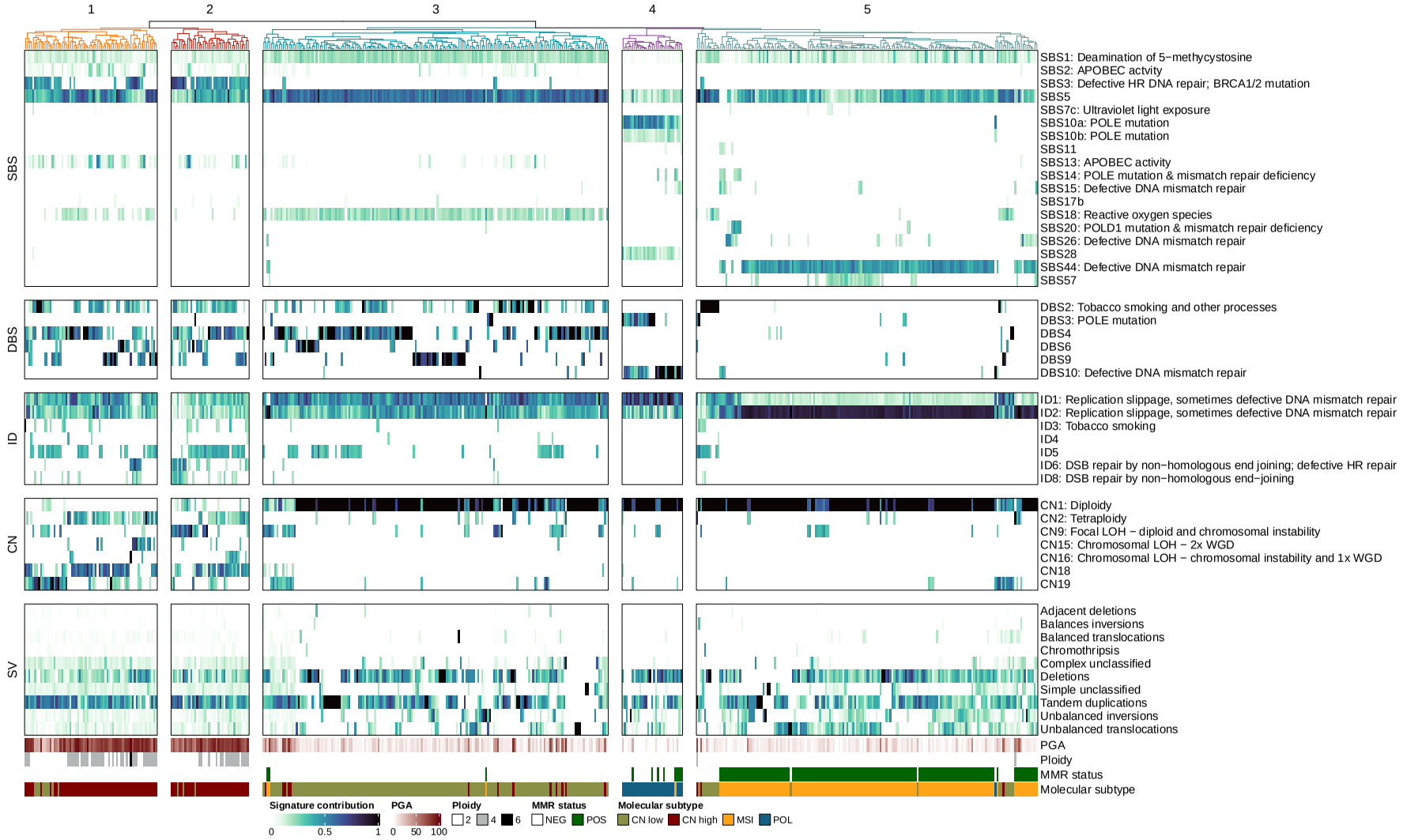
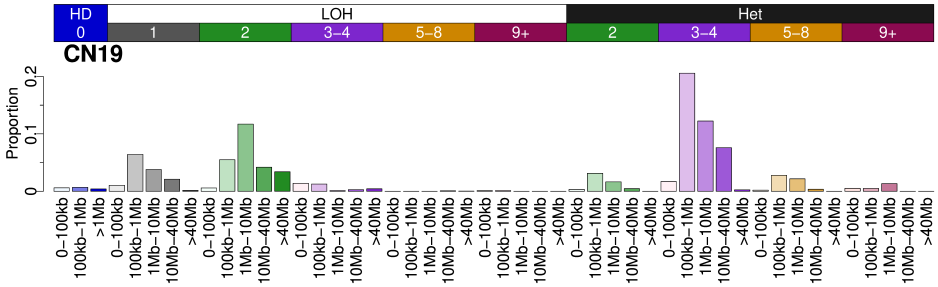
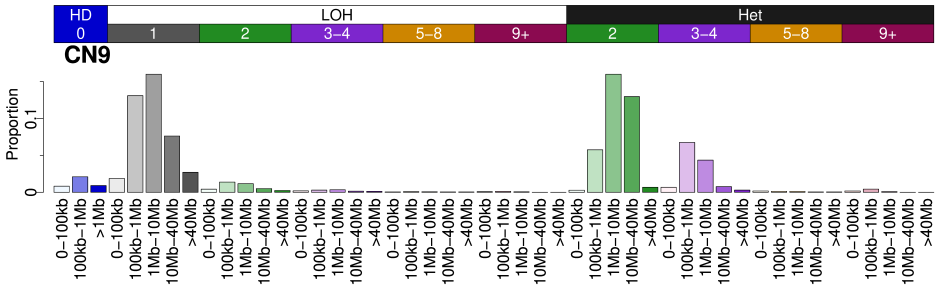


Figure 2.12: **Overview of molecular signatures in EC.** Fraction of genomic alterations assigned to single base substitutions (SBS), doublet base substitutions (DBS), small insertions and deletions (ID), copy number (CN), and structural variation (SV) signatures shown. Samples clustered on Kendall rank correlation distance across all shown signatures.

Cluster 1 is characterized by higher activity of signature CN19 (Fig. 2.13a) and cluster 2 by higher activity of CN9 (Fig. 2.13b). The latter is associated with chromosomal instability on a diploid background. CN19 is a signature with unknown etiology often found in samples with whole genome duplication [167]. The two signatures share some similarities, which can suggest that the two record the same processes at different stages - CN9 can be a representation of CN19 after more time has passed and more amplified regions were lost. However, a thorough investigation, including timing the copy number alterations and intersecting them with the assignments to signatures, would be required to verify this and understand the difference between the two processes.



(a) **Signature CN19** - signature with unknown etiology associated with high presence of genomic scars of homologous recombination deficiency [168], with the tandem duplicator phenotype [169], with the pyrgo rearrangement phenomenon (see Hadi et al., 2020) and with samples that are once-genome-doubled.

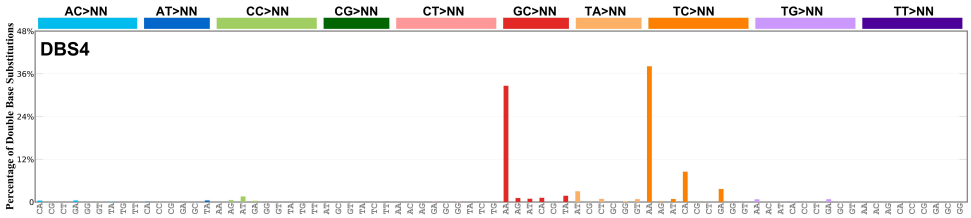


(b) **Signature CN9** - signature associated with chromosomal instability on diploid background.

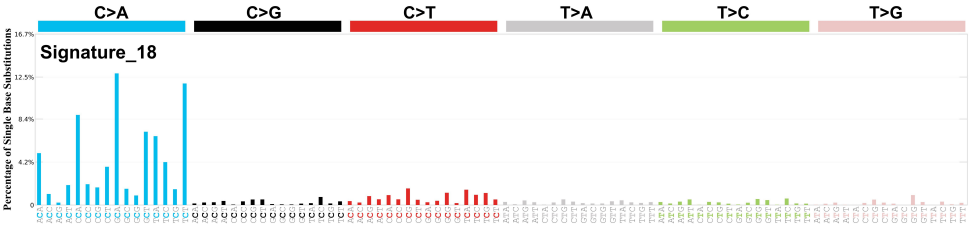
Figure 2.13: **Profiles of CN signatures present in clusters 1 and 2 of clustering based on frequencies of alterations assigned to each signature.** Profiles downloaded from COSMIC Mutational Signatures (v3.3 - June 2022) website [166], [165].

Cluster 3, predominantly comprised of CN low samples, can be characterized by mutations associated with clock-like signatures: SBS1, SBS5, and SBS44; and signature SBS18 associated with possible damage by reactive oxygen species (Fig. 2.14).

Aside from the high presence of SBS10a, SBS10b, and DBS3 (*POLE* mutation signa-

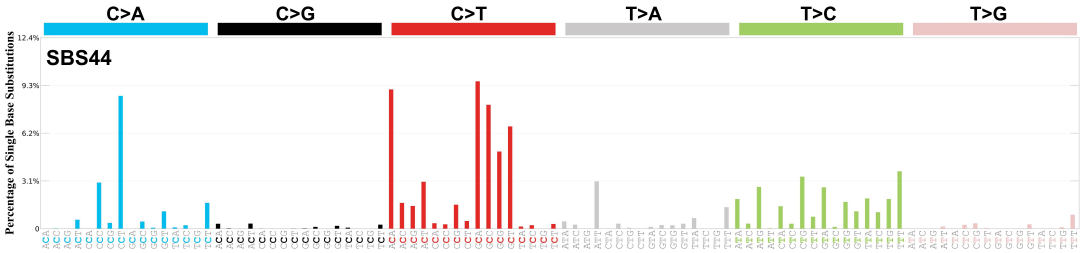


(a) **Signature DB4** - signature associated with unknown etiology. Clock-like signature.

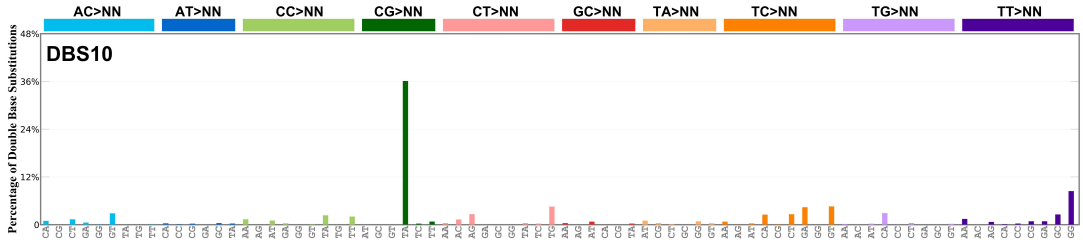


(b) **Signature SBS18** - signature associated with damage by reactive oxygen species.

Figure 2.14: **Profiles of SBS signatures present in CN low samples.** Profiles downloaded from COSMIC Mutational Signatures (v3.3 - June 2022) website [166], [165]



(a) **Signature SBS44** - associated with defective DNA mismatch repair. In EC, this signature is associated with late replication time. Regions with the H3K27ac hi stone mark (present at enhancer marks) are depleted for mutations associated with this signature. [170]



(b) **Signature DBS10** - associated with defective DNA mismatch repair. When observed, present in high numbers. Usually present alongside signatures ID1 and ID2.

Figure 2.15: **Profiles of selected mutational signatures.** Profiles downloaded from COSMIC Mutational Signatures (v3.3 - June 2022) website [166], [165]

ture), samples with POLE subtype show high activities of DBS10 and ID1 signatures. In contrast, MSI samples have shown high activities of signatures SBS44 and ID2 (Fig. 2.15). All four signatures are associated with defective DNA mismatch repair and slippage during DNA replication, highlighting both similarities and differences in the molecular processes in the two subtypes. Additionally, some of the MSI tumors show the presence of SBS20 - a signature of *POLD1* mutation in combination with loss of DNA mismatch repair.

Indeed 6 samples carry a mutation in *POLD1* on top of their MSI status. Samples with *POLD1* or *POLE* mutation on top of their MSI status have on average 7.5 (Wilcoxon test, p-value $= 3.8 \times 10^{-5}$) and 6.9 (p-value $= 1.4 \times 10^{-5}$) times more of mutations, respectively, than samples with MMR deficiency, but without mutation in POL genes. Samples that acquire MMR deficiency after mutation in *POLE* accumulate on average 2 times more (p-value $= 6.8 \times 10^{-4}$) mutations than POLE samples without MSI status (Fig. 2.16).

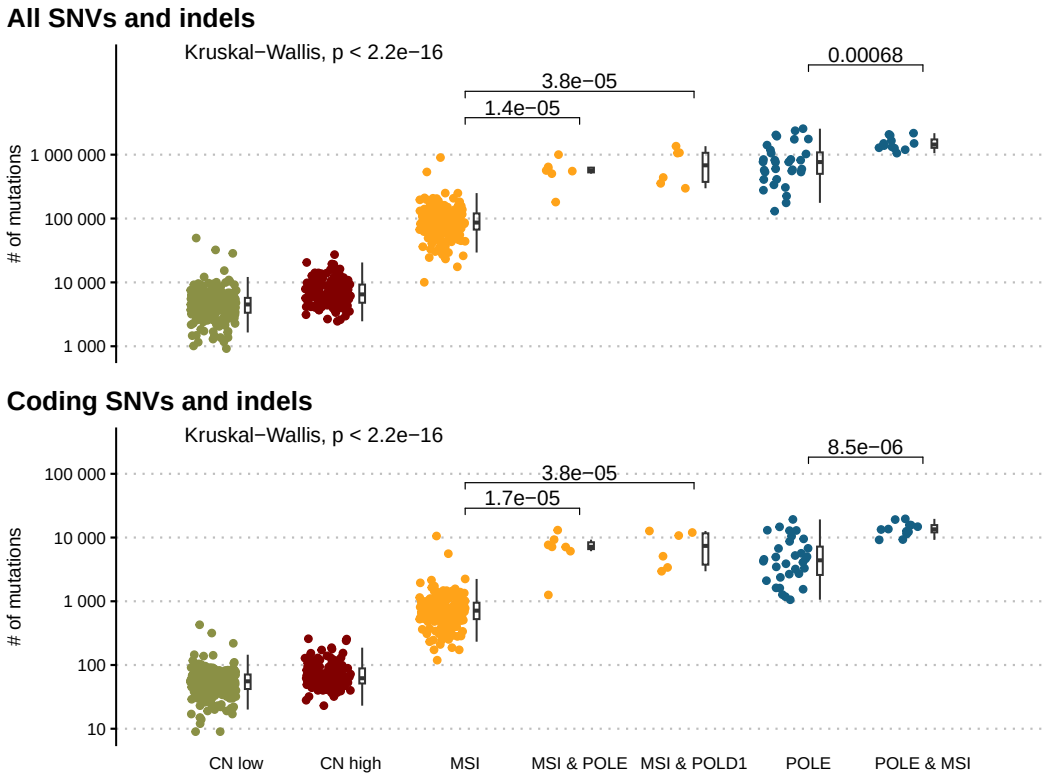


Figure 2.16: **Number of mutations with respect to molecular subtype, *POLE* and *POLD1* mutations.** p-values reported for pairwise comparisons were obtained from the Wilcoxon signed-rank test.

2.3.1.4 Coding driver mutations

The number of mutations varies between samples and is dependent on molecular subtype (Fig. 2.7), with a median number of mutations in coding regions varying from 55 in CN low samples to 5,657 POLE samples. To account for variability in tumor mutation burden and the different mutational processes across molecular subtypes, driver mutations analysis was run in all samples together and separately per molecular subtype.

With our analyses, we identified 107 drivers across all cohorts. We were able to replicate previous findings identifying drivers like *PTEN*, *ARID1A*, *PIK3CA*, *PIK3R1* and *TP53* in all subtypes. Some drivers are unique to a particular molecular subtype - the most prominent example of pathogenic polymerase-ϵ mutations exclusive to samples in the POLE subtype. We were able to recapitulate 70 known EC drivers and identify a further 22 known pancancer drivers as candidate drivers new to EC. TCGA study identifies *FBXW7* and *PPP2R1A* mutations as unique to CN high subtype [13]. In our analyses, *FBXW7* is identified as a driver in CN low, CN high, and MSI subtypes, with similar mutation spectra across subtypes (Fig. 2.18). In Section 2.3.4.3, I discuss survival modeling that identified the link between the mutation in *FBXW7* and increased hazard ratio. It is possible that the behavior of *FBXW7* resembles that of *CCNE1* amplification in ovarian cancer [171, 172].

Several genes, for example, *FGFR1*, *SF3B1*, *TAF1*, *ZBTB7B* and *ZNRF3*, were only identified in the cohort of all samples. All the genes but *TAF1* belong to the long tail of the drivers - a smaller fraction of mutated samples. These uncommon drivers indicate that while common alterations drive a substantial fraction of oncogenesis, there is also an appreciable contribution of rare variants in EC. More detailed studies investigating the long tail of drivers would contribute to a better understanding of the biology of the EC. An interesting group of drivers identified in our analyses are chromatin modifiers. Those include modifiers known to EC, such as *ARID1A*, *ARID1B*, and *ARID5B*, but also those unknown previously to EC - like *SETD1B*. I discuss chromatin modifier drivers in more detail in Section 2.3.2.

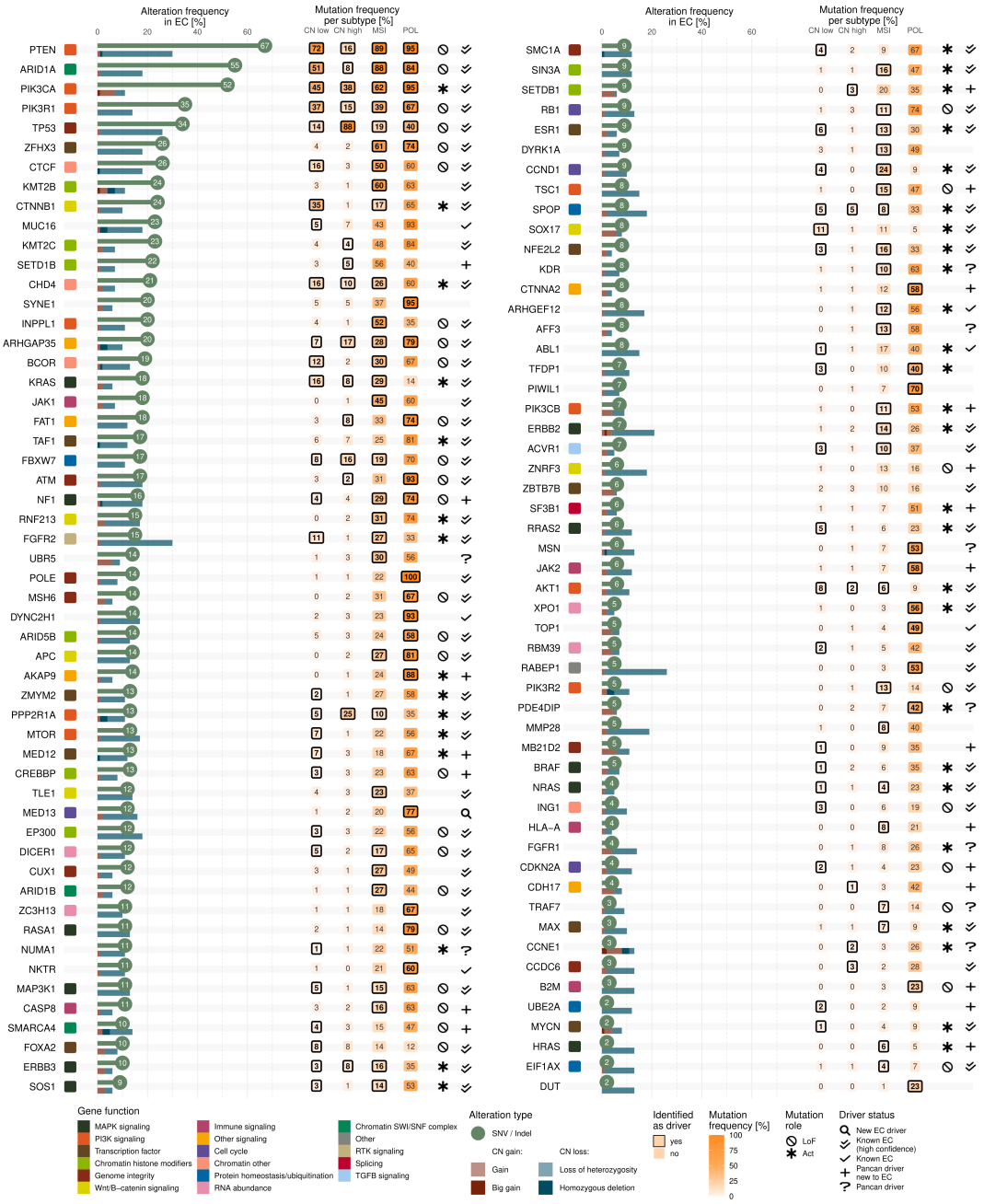


Figure 2.17: **Drivers in EC.** 107 mutational coding driver genes were identified across the whole cohort and within molecular subgroups using IntOGen [82]. Driver genes are ordered based on their mutation count. Mutations (SNVs and indels) are shown as green lollipops with copy number alterations shown as barplot. Frequency of mutations per cohort sown in a rectangle with background intensity corresponding to a frequency of alterations. If a driver was identified in a given cohort, the mutation frequency is highlighted in bold with a border around the rectangle placed in the column corresponding to the subtype. Curated gene functions annotated with color as shown in legend with a rectangle placed next to the gene name. Mutations are annotated as loss of function (LoF) and activating missense mutations (Act). Driver status - expressed as new to EC, known to EC, known Pancancer driver new to EC annotated in the last column as indicated by the legend.

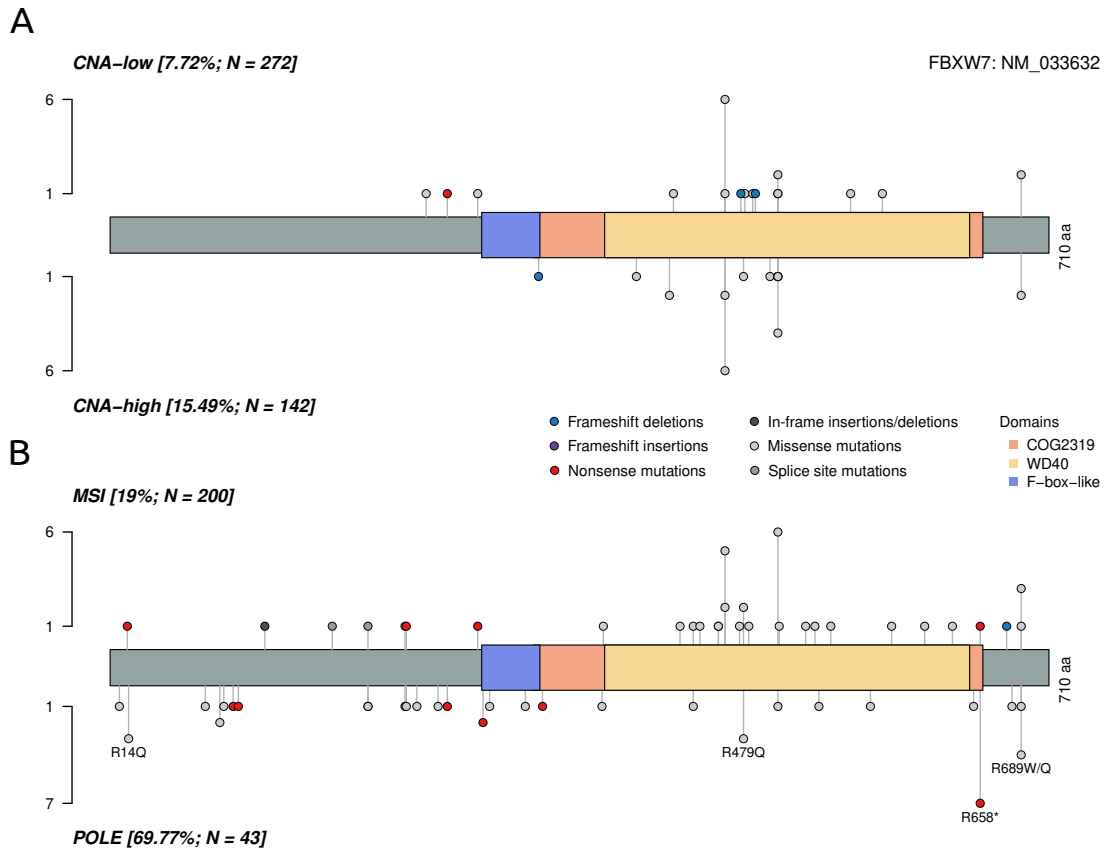


Figure 2.18: **Comparison of mutations in *FBXW7* between molecular subtypes.** **A** CN low (upper panel) and CN high (lower panel) subtypes. **B** MSI (upper panel) and POLE (lower panel). Mutations found in more than 5% of samples annotated.

2.3.1.5 Driver copy number alterations

TCGA used the copy number alteration to separate the samples into their subtypes (as discussed in Section 1.2.4). In the CN high subtype, they found recurrent focal amplifications of the oncogenes *MYC* (8q24.12), *ERBB2* (17q12), *CCNE1* (19q12), and *SOX17* (8q11.23) [13]. GEL dataset includes copy number calls performed on deep WGS. In contrast, the TCGA dataset included the SNP arrays, which are restrictive to the regions covered by the array. WGS allows for the detection of new copy number alterations with lower frequency. We used GISTIC [173] to identify significantly recurrent gains and losses in the whole endometrial cancer cohort and within molecular subgroups as defined above. We were able to recapitulate 17 out of 22 significant amplifications and 24 out of 30 significant deletions identified in the TCGA study (Fig. 2.19). For example, we observe amplification of the 19q12 region, found recurrently amplified across the cohort as well as separately in CN high and MSI samples. The 19q12 locus contains the gene *CCNE1* whose copy number is found to be a biomarker for a response of combination WEE1-ATR inhibition therapy in aggressive endometrial and ovarian cancers [174]. While we don't identify locus 8q24.12 as significantly amplified, we observe amplification of 8q24.21 locus also containing the known oncogene *MYC*. Interestingly, we also observe the amplification as significant in CN low and MSI subtype, alongside CN high. In addition, we were able to identify 20 and 23 new amplifications and deletions, respectively. For example, we identify significant deletion at 15q11.2 containing *SIN3A* in MSI and POLE subtypes. *SIN3A*, encoding a transcriptional regulatory protein, was also identified as a driver in the MSI subtype. Its suppression by miRNA has recently been found to promote metastasis in breast cancer [175].

2.3.2 Chromatin modifiers present among genes identified as drivers

As noted above, a class of genes that were commonly identified as drivers included those responsible for chromatin organization – the major focus of my research. I, therefore, undertook a systematic analysis of these.

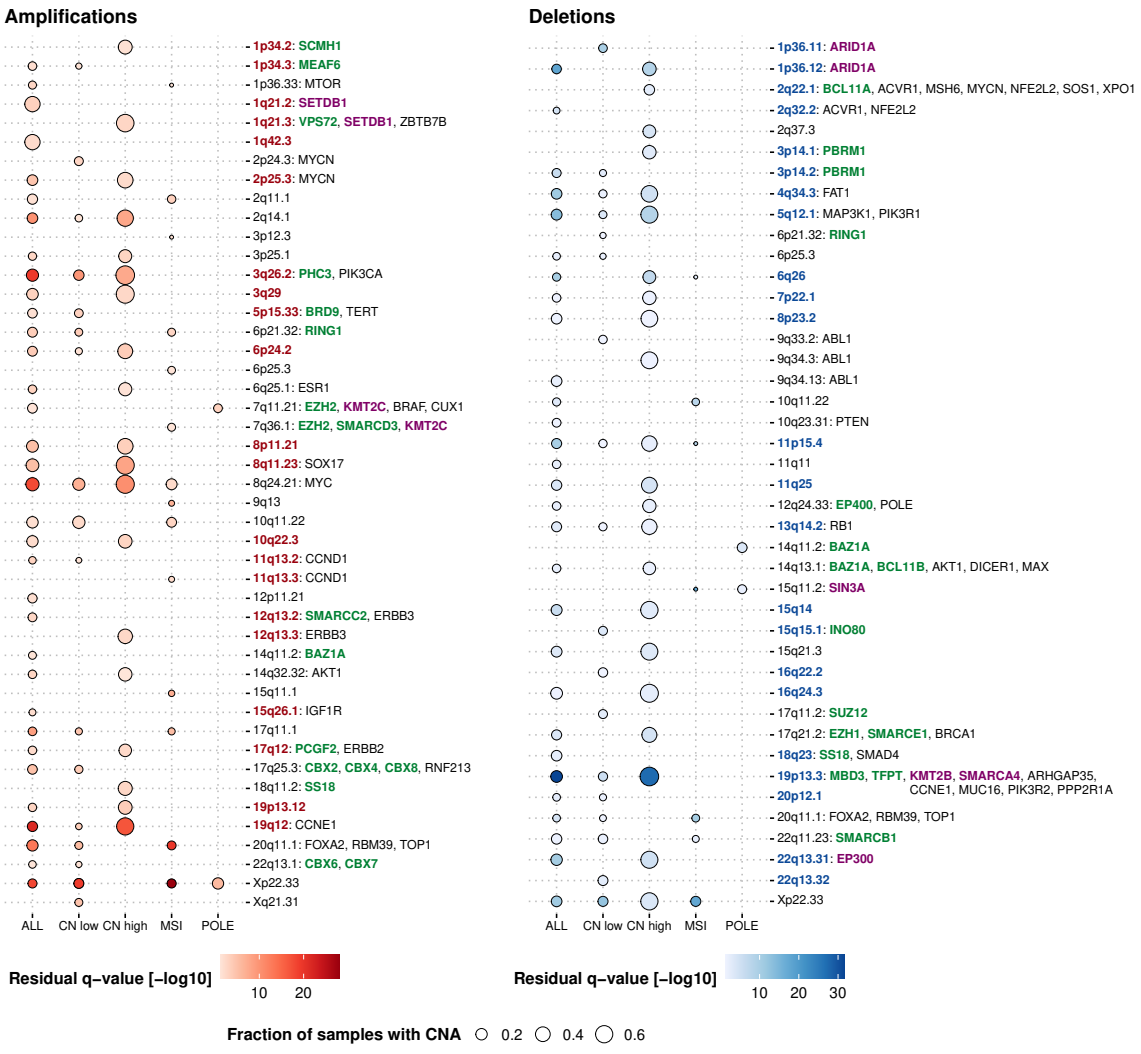


Figure 2.19: **Significant recurrent copy number alterations across molecular subtypes in EC.** Alterations found in TCGA data highlighted in red (amplifications) and blue (deletions). Selected genes in amplified loci annotated - chromatin modifiers found among drivers highlighted in purple, chromatin modifiers highlighted in green.

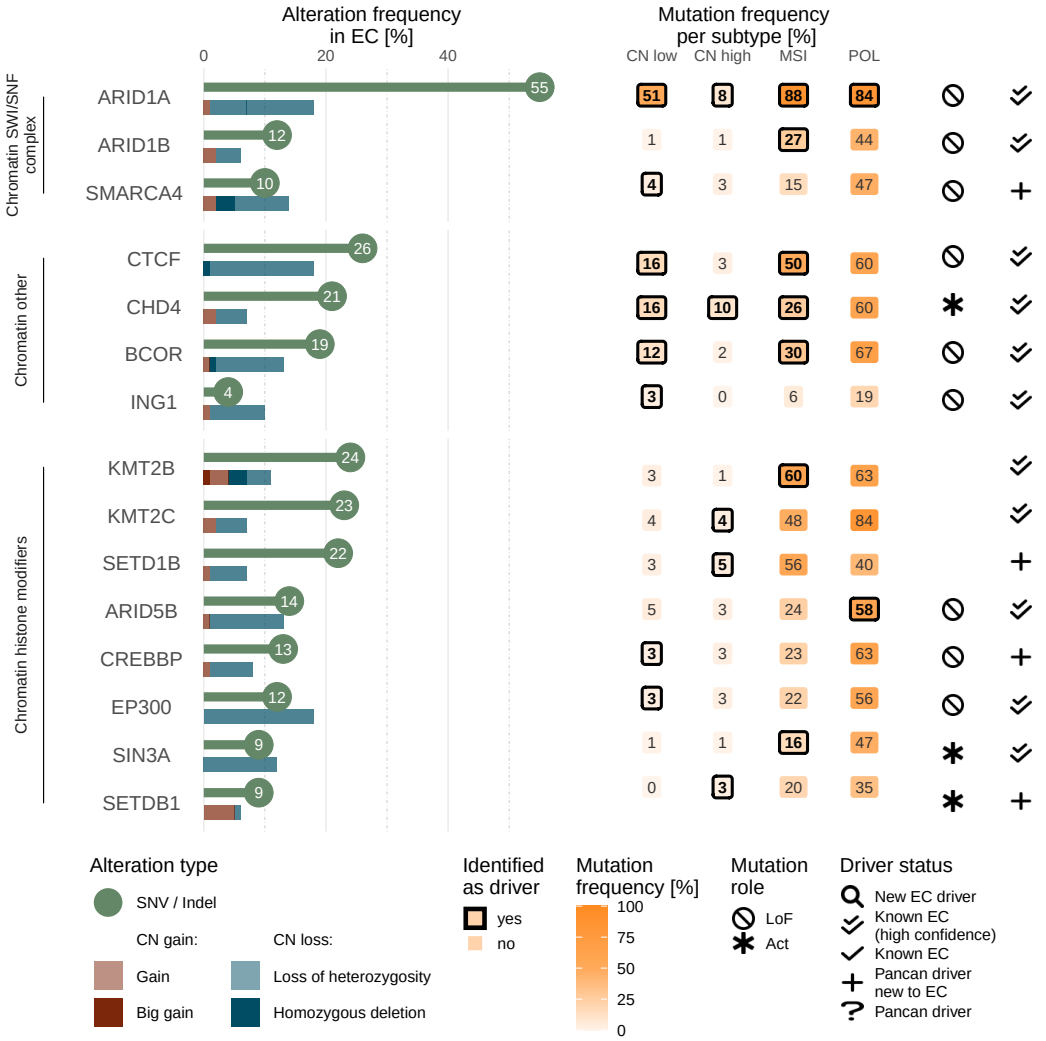


Figure 2.20: **Chromatin modifiers identified among drivers in GEL EC cohort.**

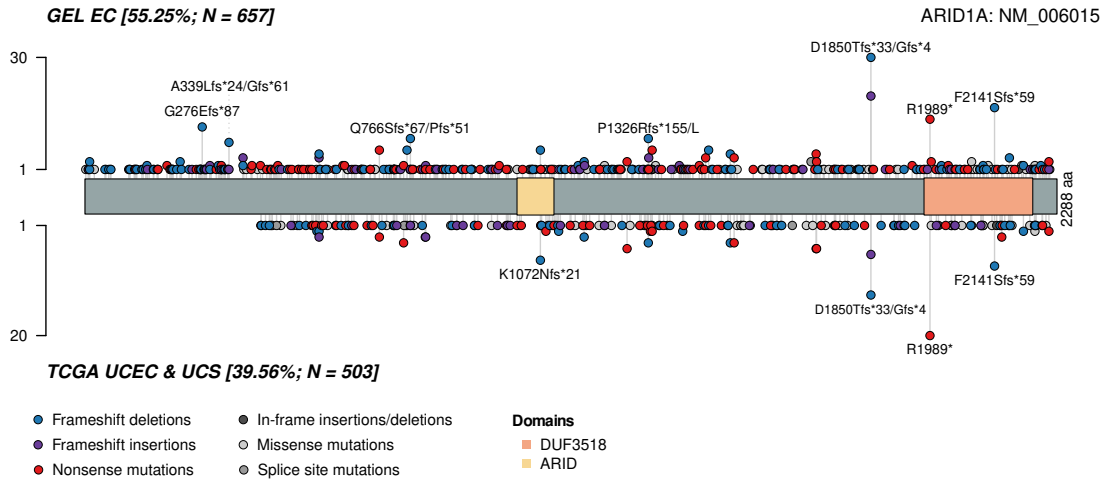


Figure 2.21: **Comparison of mutations in *ARID1A* between GEL and TCGA endometrial cancer cohorts.** Lollipop plot of mutations in *ARID1A* gene. Mutations present in more than 1% of samples are annotated with their respective change.

Out of 107 drivers identified in the GEL EC cohort, 15 are chromatin modifiers as shown in Fig. 2.20. Six of chromatin modifiers, *ARID1A*, *CTCF*, *KMT2B*, *KMT2C*, *SETD1B* and *CHD4* were found among the highly mutated drivers (top 15, mutation frequency $> 20\%$). In addition to confirming previously identified drivers, we also identified novel chromatin drivers *SMARCA4*, *SETD1B*, *CREBBP* and *SETDB1*. I discuss them in more detail below.

2.3.2.1 *ARID1A*

As described earlier, mutations in *ARID1A* are overrepresented in EC compared to other cancer types. *ARID1A* is also one of the highly mutated genes in endometrial cancer with 55% of all samples harboring a mutation (Fig. 2.21). Comparing the mutations across our and TCGA cohorts allowed the identification of potential hotspot mutations in the first 400 nucleotides unavailable in the TCGA cohort. The differences in coverage stem from the fact that data, in contrast to sample generation and processing pipelines in TCGA, GEL data does not include capturing only coding regions. *ARID1A* is characterized by truncating mutations - with mutation profiles differing between molecular subtypes (see Fig. 2.22). MSI samples accumulate frameshift mutations, while POLE samples harbor nonsense mutations with two hotspot positions - one of which is located in the DUF3518 (also known as BAF250_C) domain. More than half of CN low samples harbor mutations

in *ARID1A* - a mix of frameshifts and nonsense mutations. While only around 8.5% CN high samples carry a mutation in *ARID1A* loci containing the gene are recurrently deleted in CN low and CN high samples (Fig. 2.19). *ARID1A* is affected by different mutational processes in different molecular subtypes, and its alterations are present across most EC samples. To further understand the role *ARID1A* plays in endometrial cancer, we conducted functional studies I describe in detail in Chapter 3.

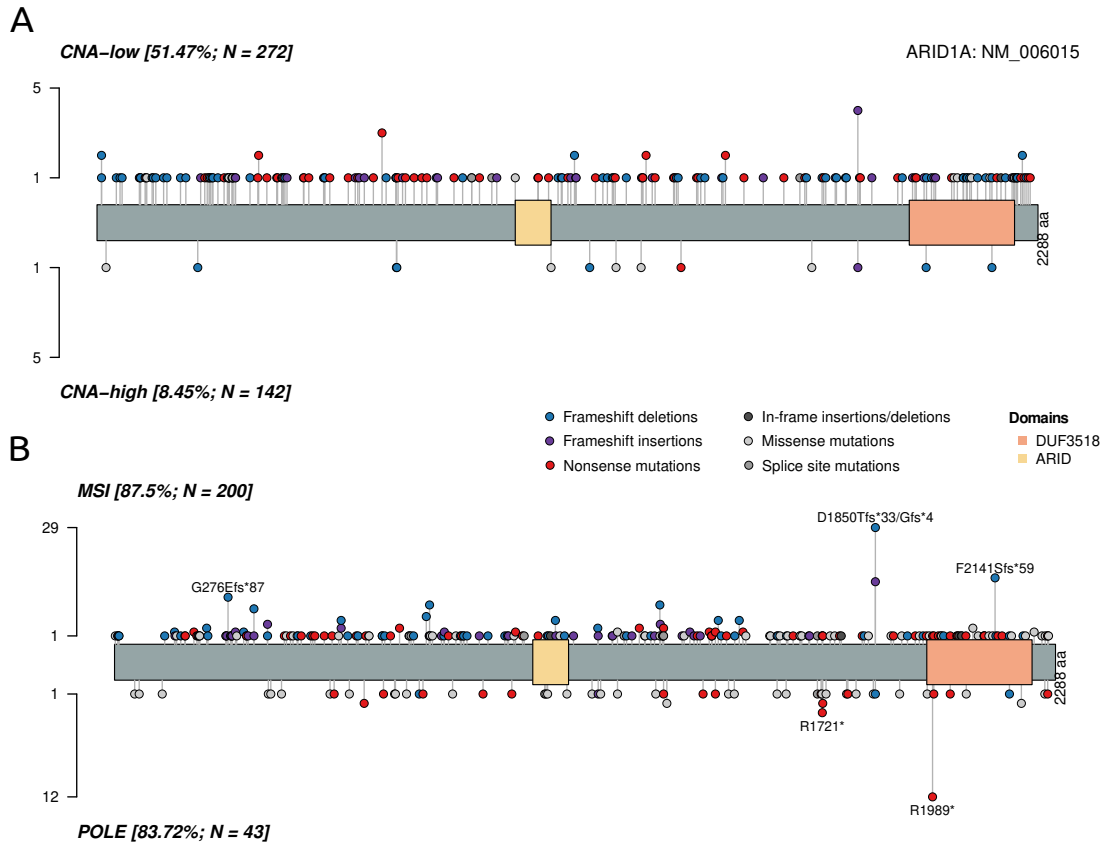


Figure 2.22: **Differences in mutations in *ARID1A* between molecular subtypes.** **A** CN low (upper panel) and CN high (lower panel) subtypes. **B** MSI (upper panel) and POLE (lower panel). Mutations found in more than 5% of samples annotated.

ARID1B *ARID1B*, *ARID1A* paralog, is also mutated in endometrial cancer Fig. 2.23. 27% and 44% of MSI and POLE samples, respectively, harbor mutations in *ARID1B*, most of them frameshifts. Less than 1.5% of CN low and CN high samples carry mutations in the gene. In MSI subtype, 49 out of the 53 samples with mutations in *ARID1B* also harbor mutations in *ARID1A*. In the POLE subtype, 17 out of 19 samples carrying a mutation in *ARID1B* harbor mutations in both genes. To confirm the co-occurrence,

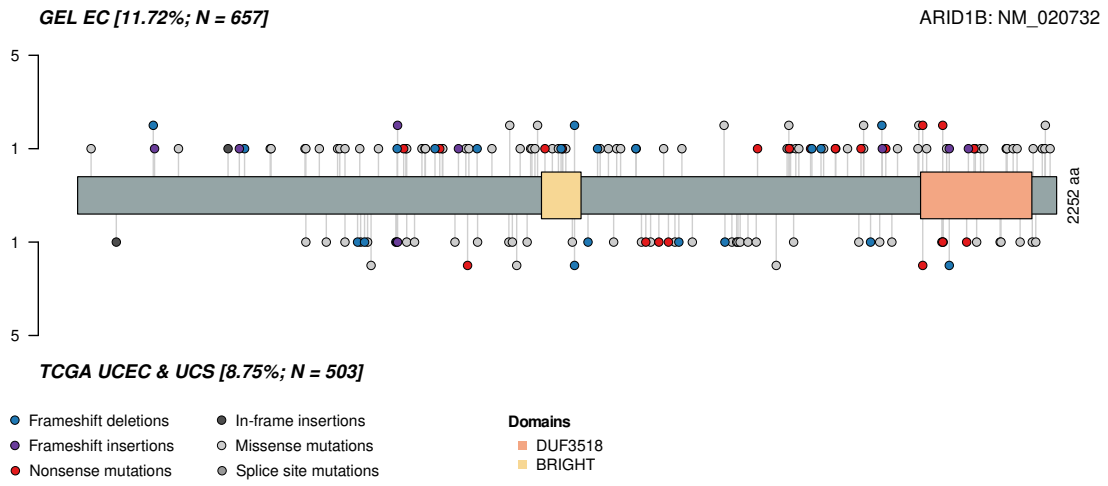


Figure 2.23: **Comparison of mutations in *ARID1B* between GEL and TCGA endometrial cancer cohorts.** Lollipop plot with mutations present in more than 1% of samples are annotated with their respective change.

a statistical method accounting for the high mutation burden of the hypermutated and ultramutated samples will be required (at the moment, the tools allowing for such analysis are unavailable in the GEL research environment).

2.3.2.2 *BCOR*

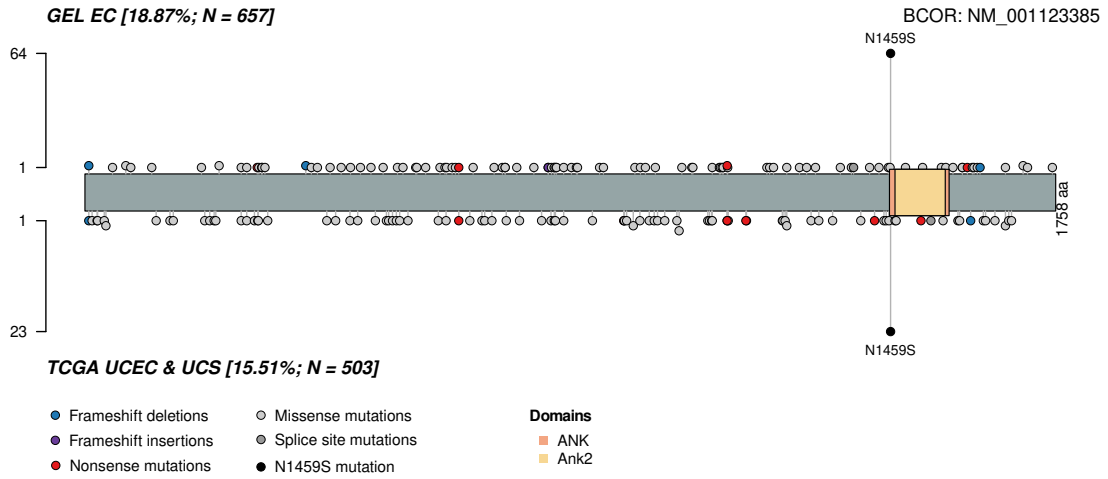


Figure 2.24: **N1459S *BCOR* mutation is located just at the site of Ankerin domain.** Lollipop plot of mutations in *BCOR*. Top - GEL EC cohort, bottom - TCGA EC cohort including UCEC and UCS samples shown. Mutations found in more than 5% of samples annotated.

BCOR, a transcriptional repressor and member of a non-canonical Polycomb repressive complex 1.1, is altered in various solid and hematologic malignancies, including acute myeloid leukemia (Section 1.4.2.2). Most alterations found in other cancer types include

internal tandem duplications, gene fusions, and loss of function mutations. However, the mutations in endometrial cancer are nearly all missense substitutions (Fig. 2.24, Fig. 2.25) at a single hotspot codon 1459S, which is near-unique to EC - 27 of 32 N1459S mutations in TCGA cohorts are found in UCEC (N1459S and other mutations are overrepresented in TCGA EC cohort, Table 2.3, χ^2 test, p-value$= 4.998 \times 10^{-4}$). N1459S mutation is located just at the site of the Ankyrin domain, the function of which remains unknown. However, Ankyrin repeat domains in other proteins have several functions, including mediation of DNA binding [176].

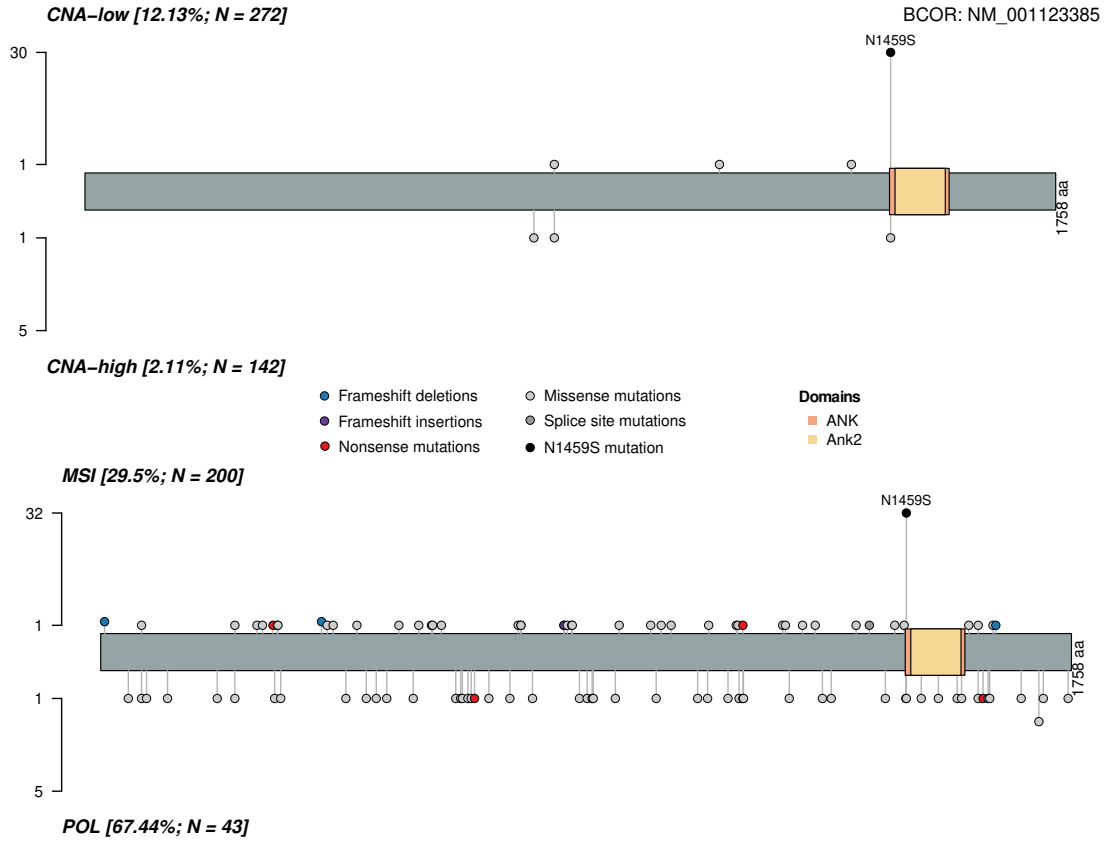


Figure 2.25: ***BCOR* N1459S mutations are frequent in samples from CNA low and MSI molecular subtypes.** Comparison of mutations in *BCOR* between molecular subtypes. **A** CN low (upper panel) and CN high (lower panel) subtypes. **B** MSI (upper panel) and POLE (lower panel). Mutations found in more than 5% of samples annotated.

In GEL EC cohort (Fig. 2.24), out of 124 samples with mutated *BCOR* 64 samples carry N1459S mutation. Overall 18.9% samples in GEL EC cohort have any non-silent mutation and 9.7% carry N1459S mutation. Although the hotspot mutation was reported in the supplementary of the TCGA UCEC landscape study [13], the biologic significance of the *BCOR* N1459S mutation is unknown (last reviewed in 2017, [162]). I report on our

		Mutations in *BCOR*		
		no mutation	other	N1459S
TCGA cohort	EC	425 (84.49%)	55 (10.93%)	23 (4.57%)
	others	8309 (97.52%)	206 (2.42%)	5 (0.06%)

Table 2.3: **Summary of *BCOR* mutations in TCGA cohorts.** Number of samples with no mutation, N1459S or other non-silent mutations reported. Percentages of samples in a given cohort reported in brackets.

efforts to understand the biological significance of the point mutant in a functional study in the Chapter 3.

2.3.2.3 *CTCF*

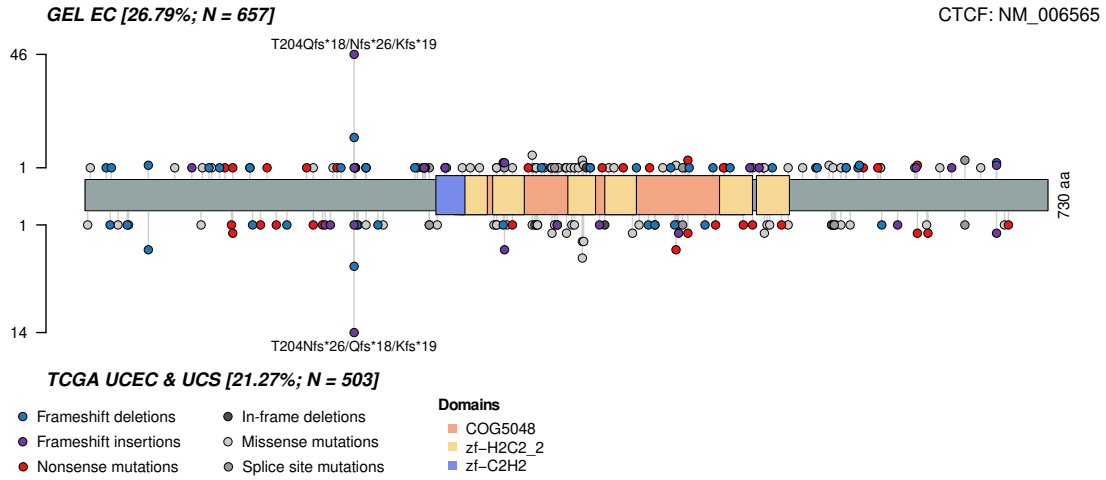


Figure 2.26: **Lollipop plot of mutations in *CTCF*.** Top - GEL EC cohort, bottom - TCGA EC cohort including UCEC and UCS samples shown. Mutations found in more than 5% of samples annotated.

CTCF, an 11-zinc finger DNA-binding protein, is a crucial architectural protein that orchestrates chromatin organization and influences gene expression by functioning as a transcriptional insulator and regulator [177, 178]. It mediates long-range chromatin interactions by forming DNA loops, thereby modulating enhancer-promoter communication and partitioning the genome into topologically associating domains (TADs) [179, 180]. CTCF truncating mutations often result in nonsense-mediated decay and commonly occur in the N-terminal portion of the gene. Hotspot frameshift alterations at CTCF codons 202 and 204 result in truncated CTCF proteins lacking zinc finger domains [181]. Truncating CTCF mutations have been identified in non-small cell lung cancer, uterine cancer

and breast cancer [182, 183]. CTCF truncating mutations have been detected in different protein domains and have been suggested to be inactivating as measured by altered binding specificities [184].

Over a quarter of mutations in *CTCF* are found at a position 204 in the amino acid sequence. Those mutations are frameshift deletions and insertions located before the functional domains, suggesting the inactivation of a tumor suppressor. *CTCF* was identified as a driver in CN low and MSI samples in which the frameshift at 204th amino acid is present (Fig. 2.20). *CTCF* is mutated in 50% of MSI samples with 27.5% carrying the p.T204 frameshift mutation.

2.3.2.4 Chromatin modifiers drivers new to EC

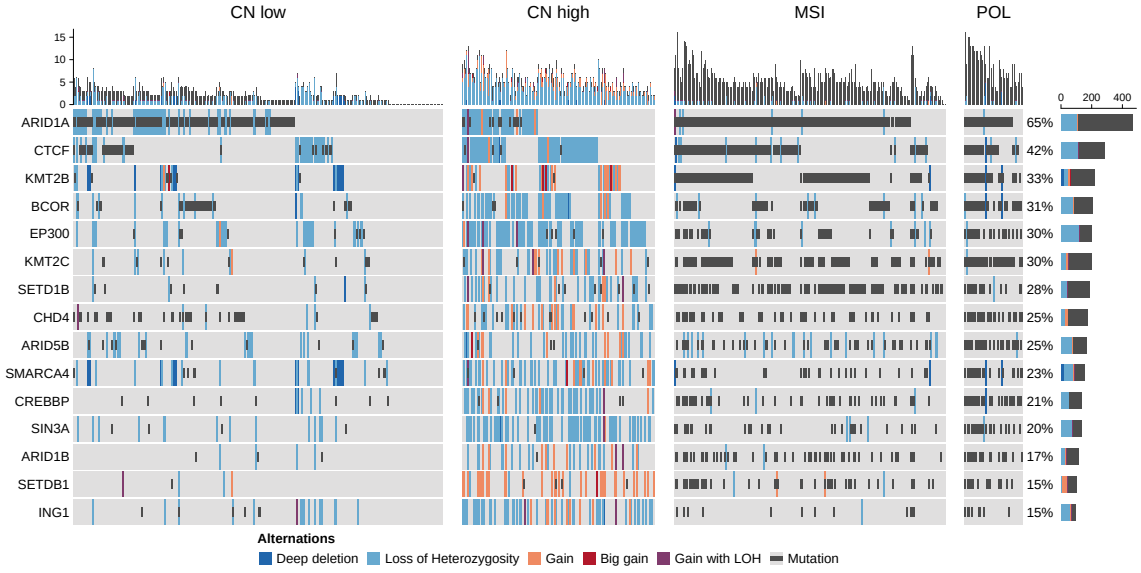


Figure 2.27: **Oncoprint of the alterations in chromatin modifiers identified as drivers.** Alterations in genes identified as drivers and annotated as chromatin modifiers presented. Copy number alterations annotated as blue, red, and purple rectangles as specified in the legend. Mutations annotated with a smaller great rectangle. Per sample sum of alterations presented in a bar in the upper panel, and per gene summary – in the bar on the right.

Out of the 15 chromatin modifiers identified with driver analysis, we found four: *SMARCA4*, *SETD1B*, *CREBBP* and *SETDB1* that are new to EC. Some were found in regions recurrently amplified (*SETDB1*) or deleted (*SMARCA4*). I present mutations and copy number alterations of chromatin modifiers identified as drivers in Fig. 2.27.

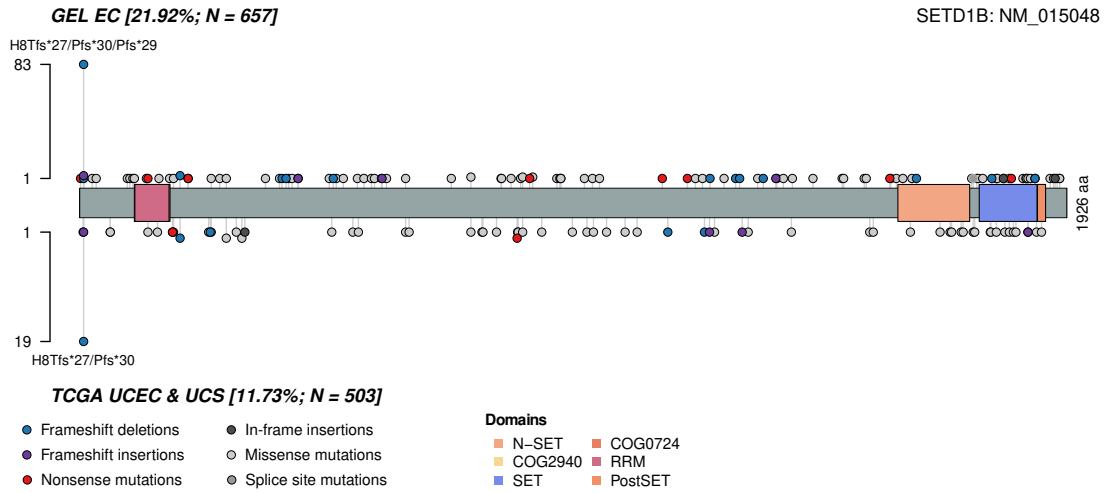


Figure 2.28: **Lollipop plot of mutations in *SETD1B***. Top - GEL EC cohort, bottom - TCGA EC cohort including UCEC and UCS samples shown. Mutations found in more than 5% of samples annotated.

SETD1B *SETD1B*, belongs to the KMT2 (MLL) family of methyltransferases (described in more detail in Section 1.4.2.2) and was identified as a putative driver in COSMIC Cancer Gene Census [185], and in several cohorts (but not EC) by IntOgen [82]. Frameshifts at the 8th codon account for almost half (83/196) of the mutations in *SETD1B* across all subtypes Fig. 2.28. In the MSI subtype, those mutations account for 55% of all *SETD1B* mutations. Interestingly, while *SETD1B* was not identified as a driver in the MSI subtype (potentially due to challenges the frameshift mutations pose to computational methods), it has been identified as a driver across all samples and in CN high subtype. Further examination of the functional consequences of loss of SETD1B and its role are needed to confirm the finding and understand the role of *SETD1B* as a driver in EC. In the next chapter, I will discuss the results of functional studies we conducted to investigate the role of *SETD1B* and other chromatin modifiers in EC.

SETDB1 *SETDB1* encodes the H3K9 methyltransferase SETDB1, and its amplification was recently identified as a mediator of immune escape [186]. We identified *SETD1B* as a mutational driver in CN high subtype (Fig. 2.20) and found the loci 1q21.2 and 1q21.3 recurrently amplified (Fig. 2.19) Missense variants consist of the majority of coding mutations found *SETD1B* (Fig. 2.29).

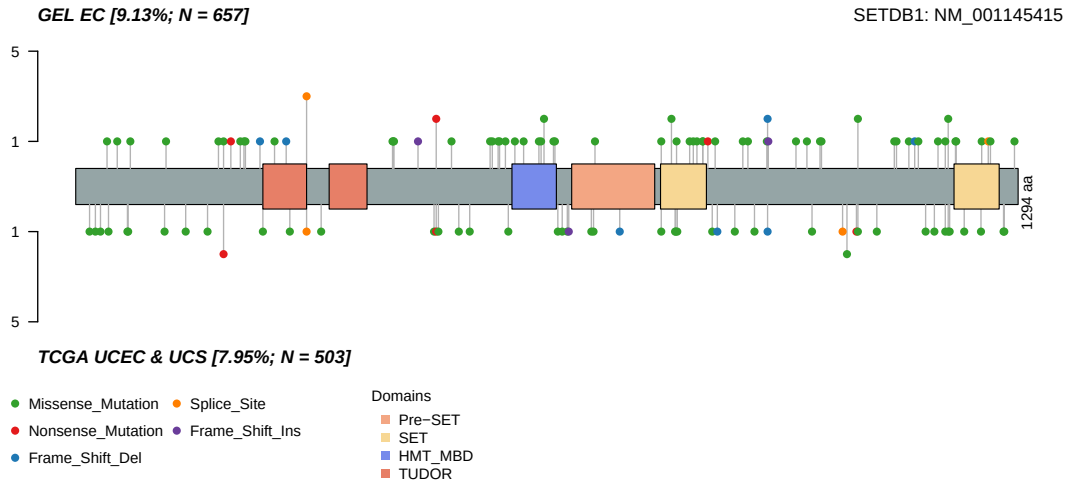


Figure 2.29: **Comparison of mutations in *SETDB1* between molecular subtypes in GEL EC cohort.** Lollipop plot with mutations present in more than 1% of samples are annotated with their respective change.

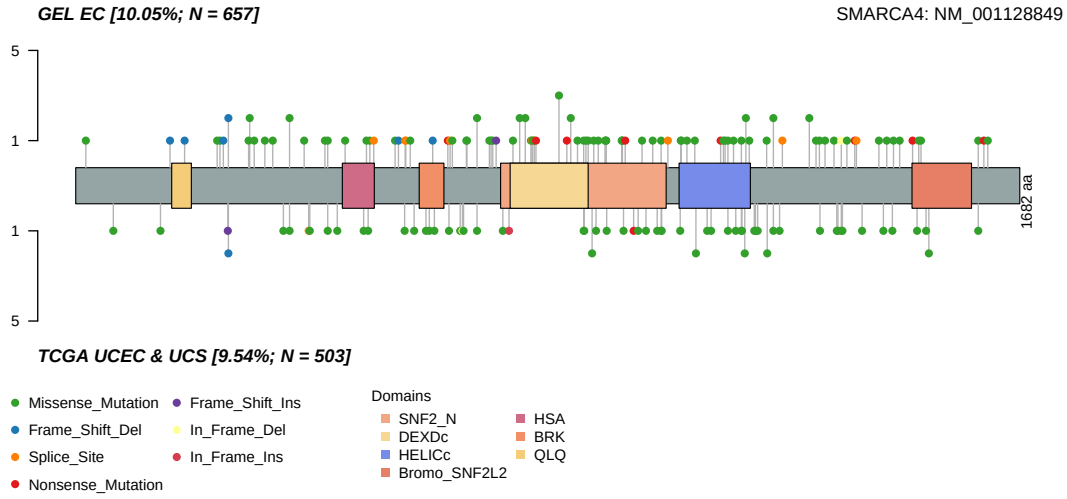


Figure 2.30: **Comparison of mutations in *SMARCA4* between molecular subtypes in GEL EC cohort.** Lollipop plot with mutations present in more than 1% of samples are annotated with their respective change.

SMARCA4 Alongside ARID1A and ARID1B, SMARCA4 is a member of the SWI/SNF remodeling complex (I discussed the complex in more detail in Section 1.4.2.3). Unlike ARID1A/B, which are only present in the canonical BAF form of the SWI/SNF complex, SMARCA4 is found in all three forms [122]. *SMARCA4* is mutated in 10% of EC samples and identified as a driver in CN low cohort. Missense variants consist of the majority of coding mutations found *SMARCA4* (Fig. 2.30). Additionally, 19p13.3 region, containing *SMARCA4*, is recurrently deleted in CN low and CN high samples (Fig. 2.19). Additionally, 19 samples (15 of CN low subtype and two each from MSI and POLE subtypes) do not carry any copy of the gene due to a homodeletion.

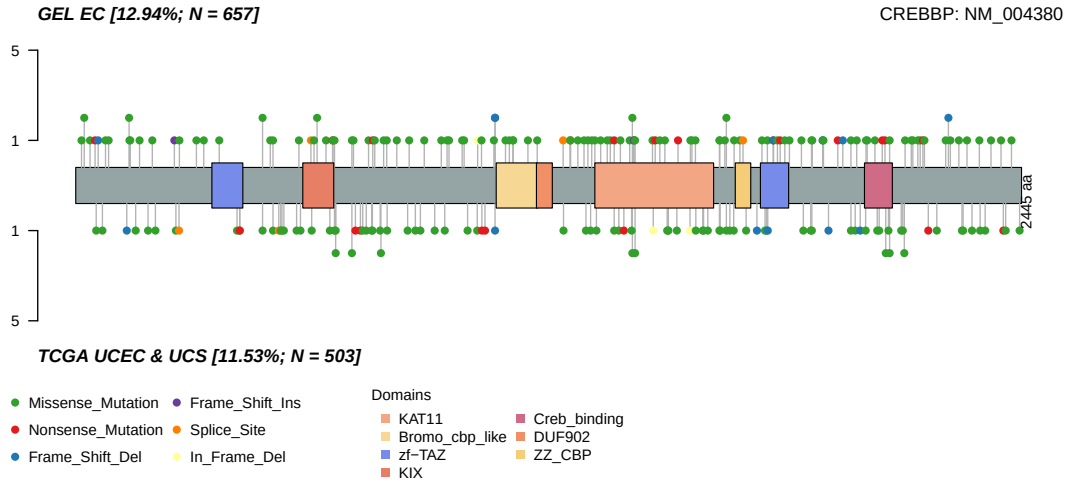


Figure 2.31: **Comparison of mutations in *CREBBP* between molecular subtypes in GEL EC cohort.** Lollipop plot with mutations present in more than 1% of samples are annotated with their respective change.

CREBBP* and *EP300 p300, encoded by *EP300* and CBP, encoded by *CREBBP*, often reffered to as one entity - CBP/p300, function as histone acetyltransferases and transcriptional co-factors. Recent studies show that inhibition of their bromodomain negatively affects the proliferation of androgen receptor (AR)-positive breast cancer cell lines [187] and decreases AR- and C-MYC–regulated gene expression in castration-resistant prostate cancer [188]. We find *EP300* and *CREBBP* altered in 30% and 21% of samples, respectively (Fig. 2.27). Both genes, based on their mutations, are identified as drivers in CN low subtype with *EP300* also identified in the cohort of all samples (Fig. 2.20). Mis-

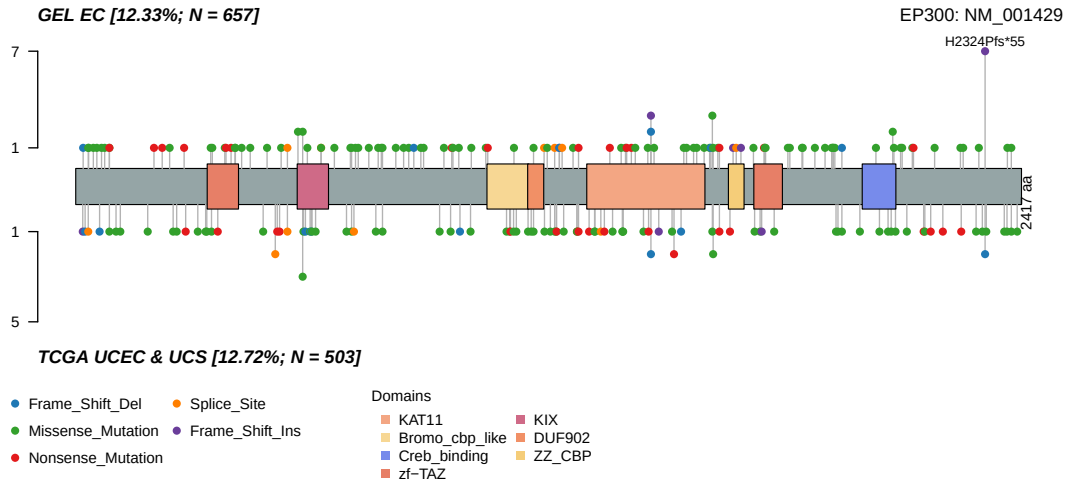


Figure 2.32: **Comparison of mutations in *EP300* between molecular subtypes in GEL EC cohort.** Lollipop plot with mutations present in more than 1% of samples are annotated with their respective change.

sense variants consist of the majority of the mutations in genes coding CBP/p300, with *EP300*, additionally accumulating franeshift insertions at the 2,324 codon (see lollipop for *CREBBP* Fig. 2.31 and *EP300* Fig. 2.32).

To understand the role of CBP/p300 in EC, we investigated the consequences of loss of *EP300* in an isogenic cell line model which I describe in Chapter 3.

2.3.3 Chromatin remodeling pathways are recurrently altered in endometrial cancer

So far, I focused on separate genes identified as drivers. However, proteins perform their function in relation to other proteins, forming complexes and interacting in small and large pathways. Often, inactivating different parts of the pathway can have a similar effect on the whole network.

2.3.3.1 SWI/SNF complex

SWI/SNF complex, an ATP-dependent chromatin remodeling complex, is one of the most frequently targeted complexes in cancer. In the earlier Section 2.3.2, I discussed several genes that encode proteins belonging to SWI/SNF complex - *ARID1A*, *ARID1B* and



Figure 2.33: **Oncoprint of the alterations in members of the SWI/SNF complex.** Selected subunits annotated - ARID1A/B subunits exclusive to cBAF marked with * and highlighted in green; equivalent subunit in PBAF - ARID2 marked with # and highlighted in yellow. PBRM1, an accessory subunit of PBAF marked with # and highlighted in brown. ATP subunits, SMARCA4/2 highlighted in red, other accessory subunits annotated in blues matching the colors in the SWI/SNF illustrations. Illustrations of SWI/SNF cBAF and PBAF variants from [122].

SMARCA4. In total, there are 29 different proteins that, in different configurations, comprise three variations of SWI/SNF complex [122]. To various extents, all 29 of the genes are altered across EC samples (Fig. 2.33). Already discussed *ARID1A*, with 65% of samples altered, is the most commonly altered gene from the SWI/SNF complex. It is affected by different molecular processes in different EC subtypes - mutations in hyper and ultramutated subtypes and copy number losses in CN high. Interestingly, *ARID1A* is also frequently mutated in CN low subtype, and those mutations are often accompanied by loss of heterozygosity (LOH), leaving the cancer cells with only a mutated copy of the gene.

ARID1B, a paralog of *ARID1A*, is altered in 17% of samples. Evidence on co-occurrences of mutations in *ARID1A* and *ARID1B* are conflicting, with a study in ovarian cancer cells confers synthetic lethality of double-knockdowns [189], while another study in the liver and other tissues reports that dual ARID1A/ARID1B loss leads to rapid carcinogenesis [190]. Mutations in *ARID1B* co-occur with *ARID1A*; however, most of the *ARID1B* mutations are found in MSI and POLE samples. Due to the high mutation burden of these samples, interpretation of co-occurrence proves difficult. I investigate the clonality of those mutations in Section 2.3.5. Exploring the relative timing of the mutations (which I explore in Chapter 4) can help better understand the relationship of those alterations.

Alterations in SWI/SNF are found in subunits of two variants of SWI/SNF - cBAF and PBAF. *ARID1A* and *ARID1B* discussed already, are exclusive to cBAF variant. The second most altered member of SWI/SNF - *SMARCE1* is an accessory unit found in both variants. Interestingly, its loss is also present in samples without alterations in *ARID1A* in CN low and CN high subtypes. *SMARCA4*, the third most altered member of SWI/SNF in EC, is an ATP-dependent subunit also found in both variants. 23% of EC samples acquire at least one alteration of *SMARCA4*, and 14 (7.3%) of CN low samples carry a complete loss of *SMARCA4*.

PBRM1, an accessory unit of the PBAF variant of SWI/SNF complex, is a known tumor suppressor in clear cell renal cell carcinoma [191]. In EC, alterations in *PBRM1*

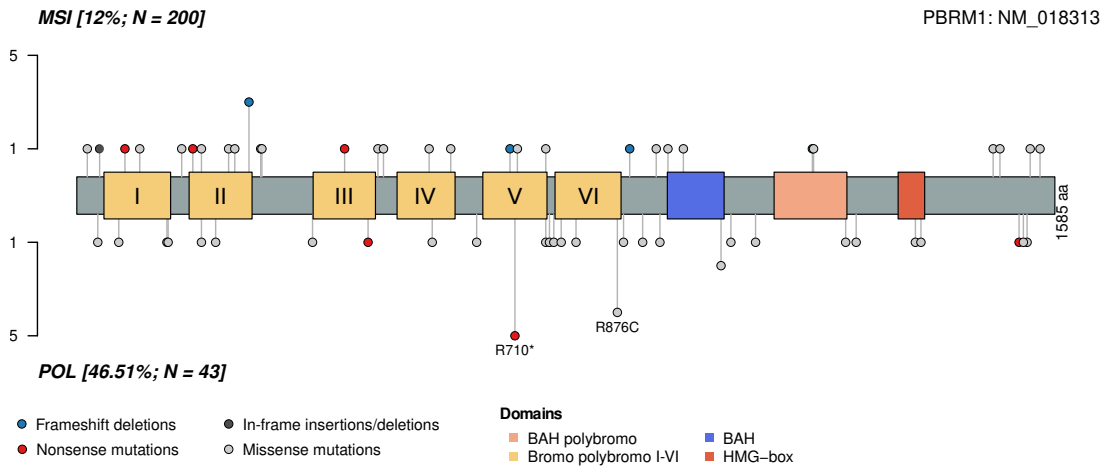


Figure 2.34: **Comparison of mutations in *PBRM1* between molecular subtypes.** MSI (upper panel) and POLE (lower panel). Mutations found in more than 5% of samples annotated.

are found in 19% of samples with 8% of samples, mostly pf MSI and POLE subtype, carrying a mutation, including 5 samples with a nonsense mutation at the R710* hotspot (Fig. 2.34). *ARID2*, an equivalent subunit to ARID1s in PBAF, is altered in 12% of samples, with no apparent pattern of mutual exclusivity or co-occurrence.

2.3.3.2 NuRD

The Nucleosome Remodeling Deacetylase (NuRD) complex regulates gene expression through nucleosome repositioning and histone deacetylation-mediated alteration of chromatin organization. NuRD is known to assemble with AR and mTor to create an mTOR-AR-NuRD complex that modulates AR-dependent gene expression in prostate cancer [192]. In EC, we find that 41% of samples carry an alteration in *CHD3*, chromodomain helicase DNA-binding an enzymatic subunit of NuRD (Fig. 2.35). Most samples (89.57%) in CN high subtype experience loss of heterozygosity of the region containing *CHD3* gene. However, it is important to note that all but one of those events are linked to loss of *TP53* as the two genes are in close proximity (both genes are located at the 17p13.1 band, *CHD3* hg38 genomic position is chr17:7888789-7912755 and *TP53* - chr17:7668421-7687490). However, in MSI subtype *CHD3* is mutated in 40% of samples with 31 participants carrying a frameshift mutation at 599 protein position. *CHD4*, a paralog of *CHD3*, is altered in 28% of samples, and *CHD5*, another paralog that forms a non-canonical NuRD complex,

is altered in 28% of samples. Overall, NuRD complex is altered in half of CN low samples with either a loss of *CHD3*, *CHD5*, complete loss of *MBD3* or a missense mutation in *CHD4*.

2.3.3.3 PRC1.1

As discussed in Section 1.4.2.2, Polycomb Repressive Complexes (PRCs) are a group of multi-protein complexes that regulate gene expression by establishing and maintaining repressive chromatin states [119]. PRC1.1, a non-canonical subtype of PRC1, is characterized by the presence of the RYBP or YAF2 proteins and has been shown to play a crucial role in early embryonic development and stem cell fate determination [145]. BCOR, a crucial subunit of PRC1.1, is mutated in 19% of EC samples, and its N1459S hotspot mutation is overrepresented in EC samples compared with other cancer types (Section 2.3.2.2).

Out of 17 genes coding members of PRC1.1, five are recurrently altered in EC (Fig. 2.36). *KDM2B*, DNA binding subunit of PRC1.1 is altered in 14% of samples - CN high samples are characterized by loss of heterozygosity and MSI and POLE - by missense mutations. *RYBP* is altered in 11% of samples. The majority of CN high samples are affected by loss of heterozygosity of at least one of the *BCOR*, *KDM2B* and *RYBP* genes. Other members of PRC1.1 altered in EC are *RNF2* and *PCGF1*, altered in 9% and 7% of samples, respectively. Most of the alterations in those genes are split between gains in samples of CN high subtype and missense mutations in POLE subtype.

2.3.4 Relationship between molecular alterations and overall survival

The whole genome sequences in GEL are accompanied by clinical data, allowing us to perform survival analyses using overall survival (recurrence data is unavailable). The median follow-up time registered in the dataset was 2.9 and 3.04 years for the median time to the event and median time to censoring, respectively. Given the 5-year survival rate for Endometrial Cancer reaching above 75% irrespective of stage in England [156],

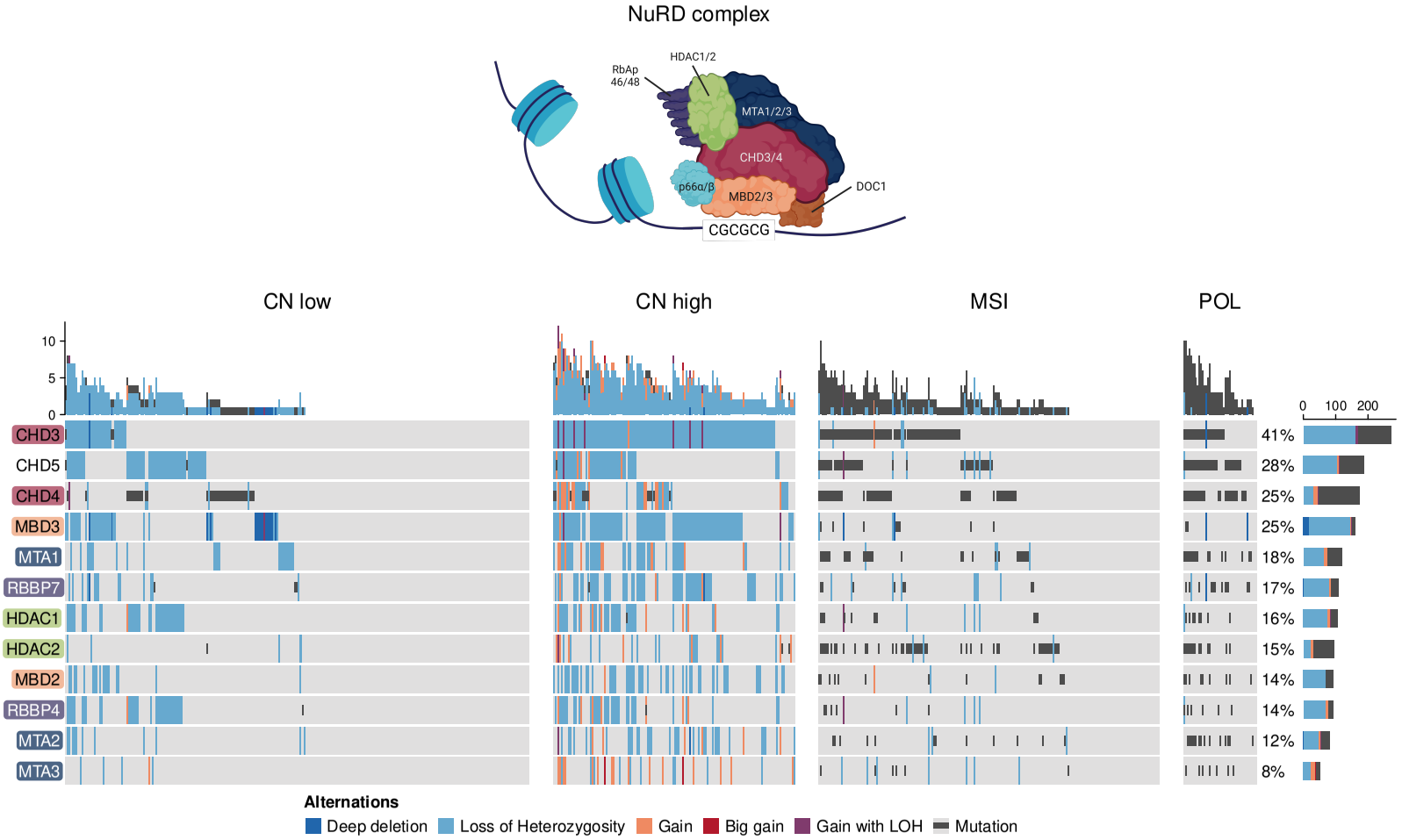


Figure 2.35: **Oncoprint of the alterations in members of the NuRD complex.** Genes highlighted with colors corresponding to the position of the proteins in the schematic. CHD5 forms a non-canonical NuRD complex and is a paralog of CHD3/4. Schematic of NuRD complex reproduced using BioRender.com from [129].

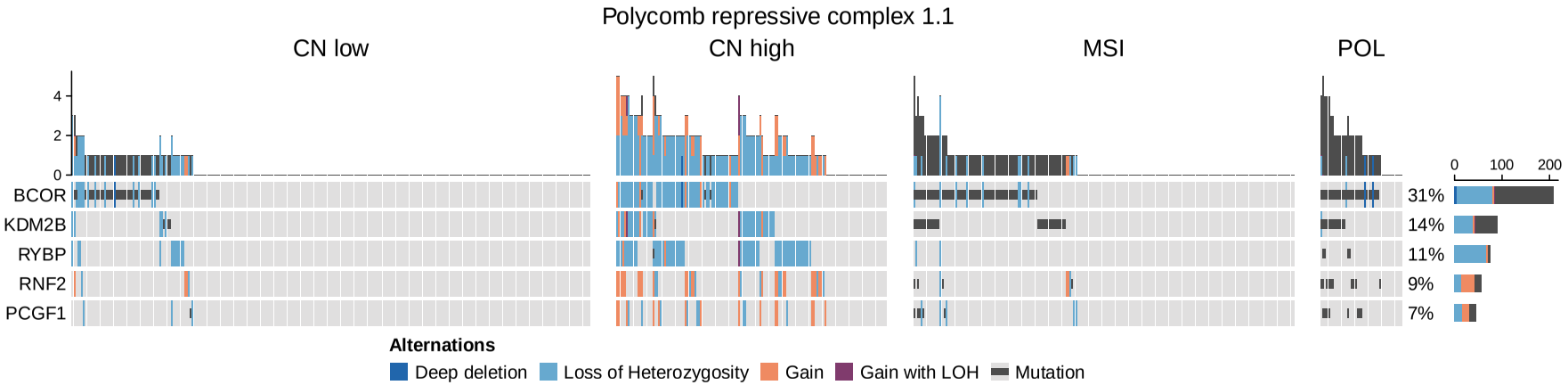


Figure 2.36: **Oncoprint of the alterations in member of the PRC1.1 complex.**

the relatively short follow-up time allows to observe only relatively large increases or decreases in hazard ratios (power for HR=0.74 calculated at 0.45 and for HR=1.5 at 0.74, Section 2.5.1.4).

2.3.4.1 CNA-high subtype is characterized by lowest survival rate

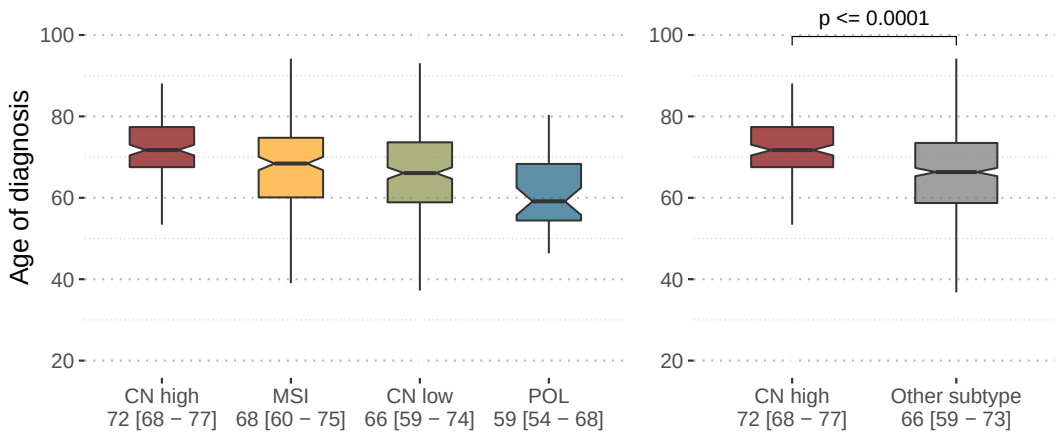
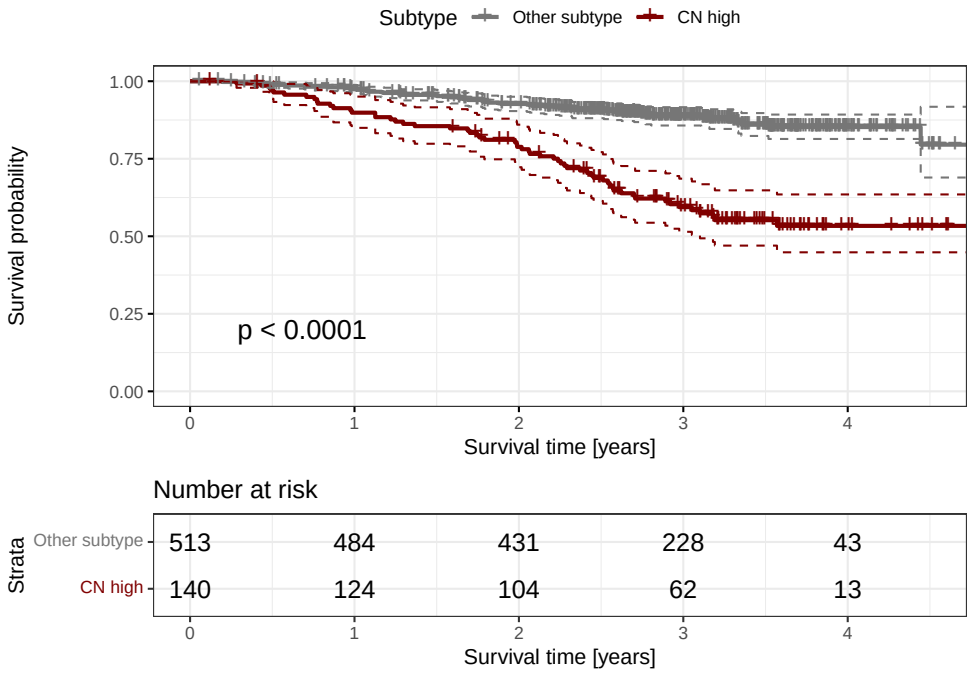
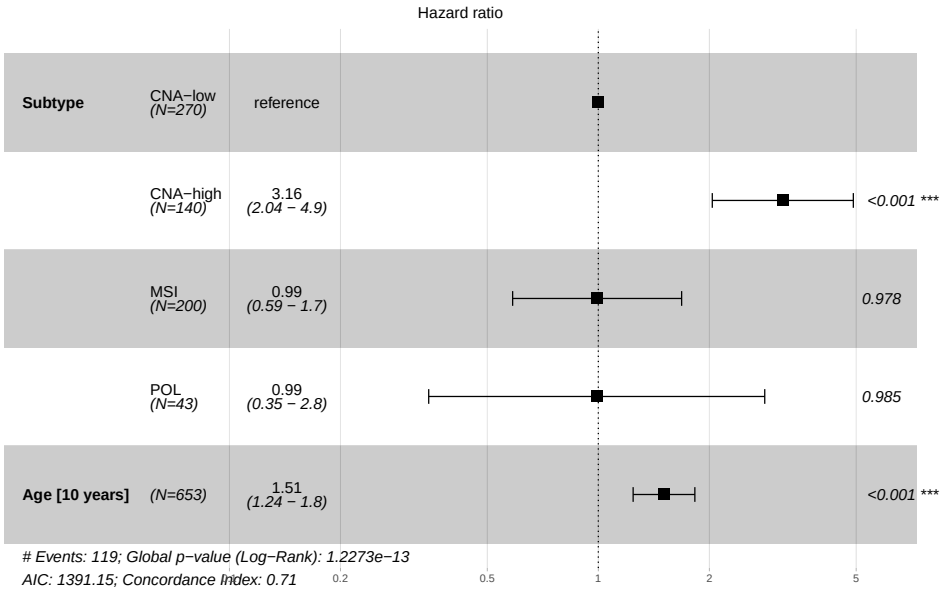


Figure 2.37: **Patients with CN high EC subtype are older at the time of diagnosis than in other subtypes.** Median age of diagnosis for each group is written under the subtype label with interquartile range (IQR) in brackets. P-value reported for Kruskal-Wallis test of medians.

CN high EC are diagnosed in older patients than other subtypes - CN high cases are diagnosed at 72 years while other subtypes at 66 years of age (Mann-Whitney U test, $p\text{-}value = 3.9 \times 10^{-8}$, Fig. 2.37). In agreement with previous reports [13], I was able to show the CN high subtype is associated with lower survival in comparison to other subtypes as shown in the Kaplan-Meier curve in Fig. 2.38a. I further confirm this by fitting a Cox Proportional-Hazards model with the age of diagnosis as a covariate to investigate the effect of CN high subtype on survival while correcting for the age of diagnosis (Fig. 2.38b). Survival adjusted for age of diagnosis is lower for patients with CN high EC - at 5.12 years post-diagnosis modeled survival for other subtypes equals 76.4% while for CN high subtype modeled survival stands at 44.3%. (Fig. 2.39)



(a) **Kaplan-Meier survival curve for EC patients with respect to molecular subtype.** Other than CN high subtypes grouped together as there was no signigicant difference in survival between CN low, MSI and POLE subtypes.



(b) **Forest plot for overall survival of EC patients. Hazard ratios estimated with Cox Proportional-Hazards model.** Out of the 4 molecular subtypes, only CN high subtype results in increased hazard ratio. We observe that risk of death increases with age, i.e. every 10 years increases the chance of death by 50%.

Figure 2.38: **CN high subtype has a lower survival rate than other EC subtypes.**

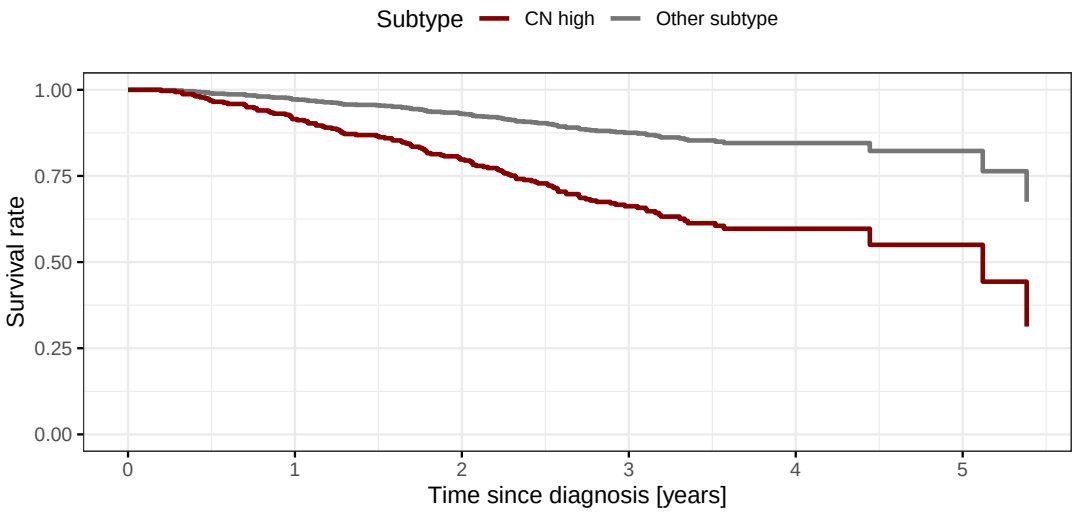


Figure 2.39: **Survival curve adjusted for age of diagnosis.** Age-adjusted 5-year survival rate for other than CN high subtypes is above 75% while for CN high it remains at slightly above 50%.

2.3.4.2 Mutation burden and the fraction of aberrant genome inform survival prediction

Endometrial cancer subtypes are characterized by the pronounced differences in the prevalence of various types of alterations - SNVs, indels, and copy number alterations, as discussed earlier. I investigated the predictive value of the variables in the context of modeling survival. A number of single nucleotide variants (SNVs) are correlated with insertions and deletions with the size of correlation dependent on the molecular subtype (Fig. 2.40). In POLE samples, the number of SNVs is the most correlated with insertions and deletions (R=0.77, p-value= 2.1×10^{-9}), and least correlated in CN low samples (R=0.41, p-value= 2.1×10^{-12}).

I investigated the tumor mutation burden (TMB) and fraction of genome in copy number altered state (PGA) (Fig. 2.41). I fitted univariate models and multivariate models adjusted for age, age and molecular subtype. TMB expressed as SNVs per Mb was associated with reduced risk of death (HR=0.69-0.86, p-value=0.007-0.025), while TMB expressed as insertions and deletions lost prognostic value when adjusted for age and subtype. At the same time, an increase by 10% in PGA increases the risk of death by 21% (Cox proportional hazard model adjusted for age and subtype, p-value = 0.005). This

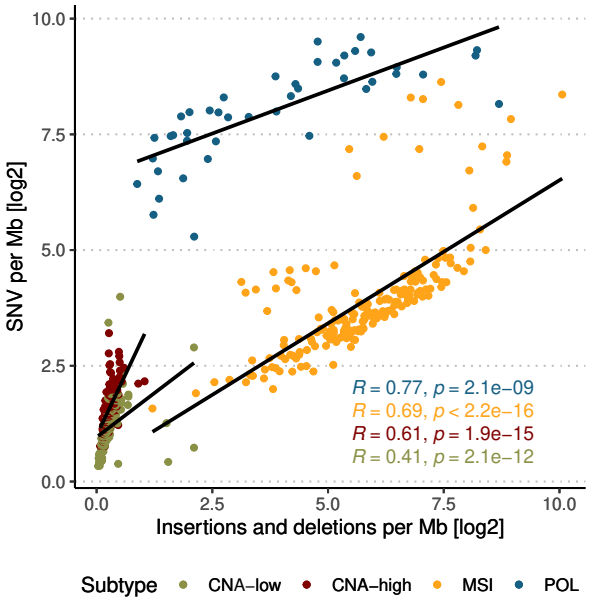


Figure 2.40: **Number of SNV mutations is positively correlated with number of insertions and deletions across subtypes.**

analysis represents the first demonstration that the burden of SNV mutations and copy number alterations are prognostic in endometrial cancer, independently of the molecular subtype.

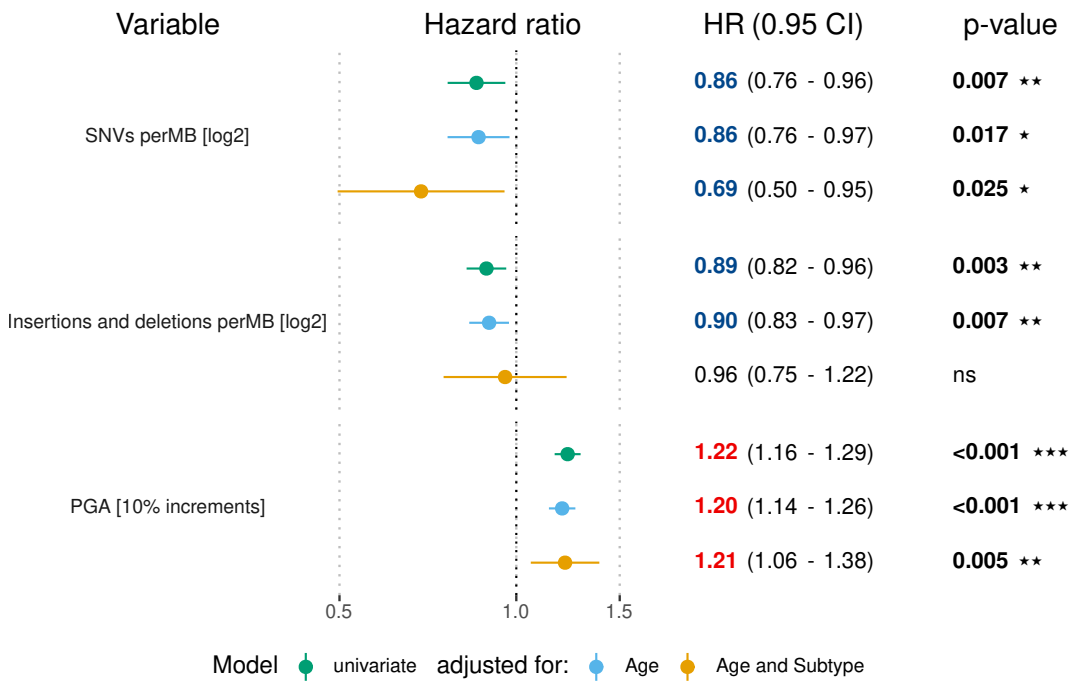


Figure 2.41: **Cox Proportional-Hazards models with respect to tumor mutation burden and fraction of altered genome (percent of altered genome, PGA).**

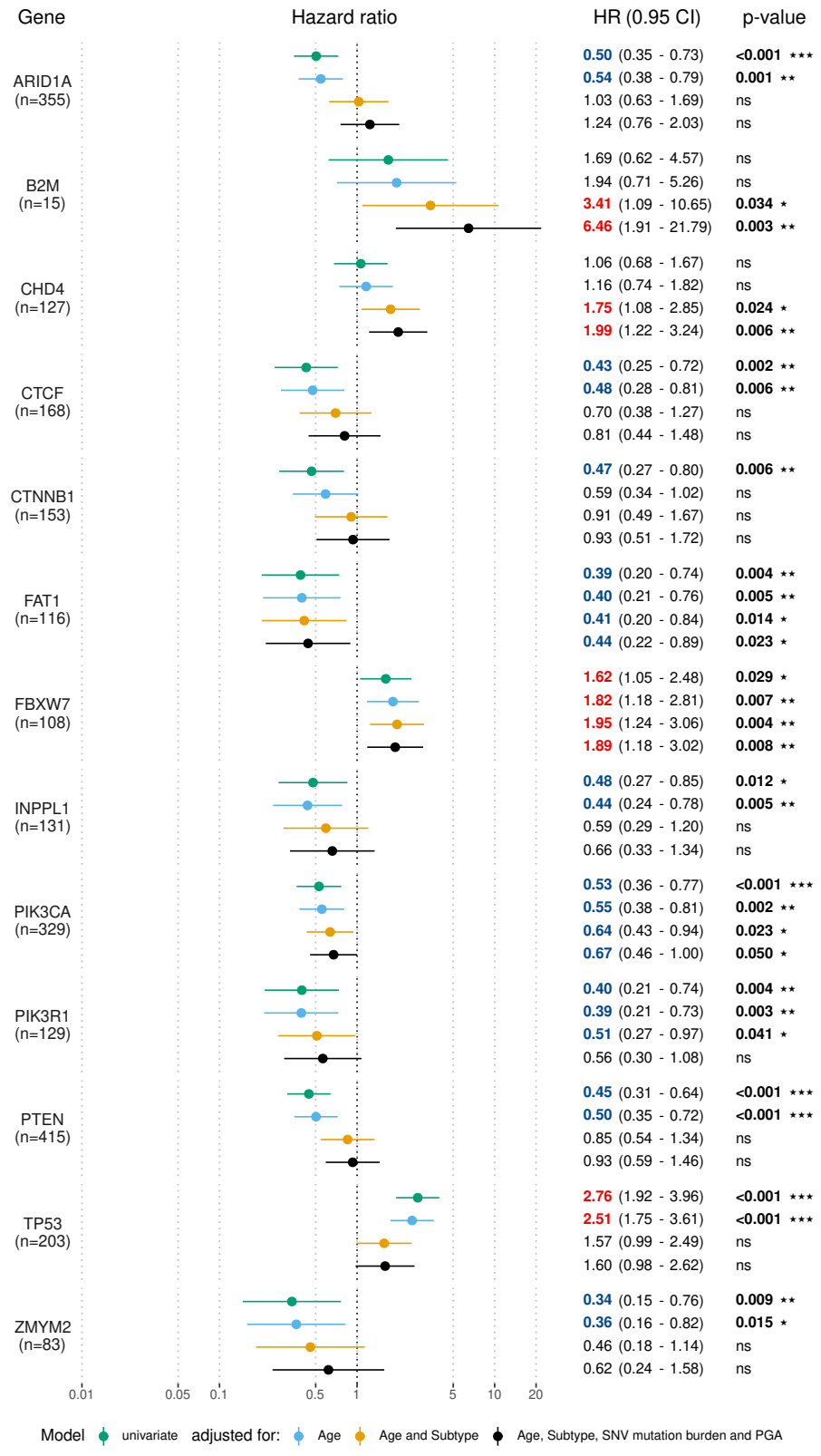


Figure 2.42: **Overall survival modeled with respect to mutations among drivers.** Survival was modeled without any co-variates (univariate), as well as adjusting for age at diagnosis, molecular subtype, and SNV mutation burden and genome alterations. Only drivers with p-value < 0.01 in any of the models shown. A full summary of fitted models is available in the supplementary materials.

2.3.4.3 Alterations in drivers are predictive of survival

I next investigated the effect of mutations in driver genes on overall survival (Fig. 2.42, full table available in the supplementary materials). I first performed a univariable analysis of all 107 coding drivers identified. I found that mutations of *ARID1A*, *PIK3R1*, *PTEN* and *CTCF* are indicative of better survival, and mutations in *TP53* are indicative of lower survival in univariate models. When adjusted for age, however, the effect is no longer present in models with a subtype as a co-variate. Overall half of the mutations in *TP53* are found in the CN high subtype, which can explain the lack of predictive value of those mutations on overall survival once the molecular subtype was included in a model. The poor outcome linked to mutations in *B2M* and *CHD4* is present in models adjusted for molecular subtype but not in those without subtype as a covariate, suggesting differences in baseline survival in each molecular subtype. Mutations in *CHD4* increase the risk of death almost twice (HR=1.99 [95% CI 1.22 - 3.24, p-value=0.006], black pointrange plot in the third row of Fig. 2.42) and mutations in *B2M* by more than six times (HR=6.46 [95% CI 1.91 - 21.79, p-value=0.003], black pointrange plot in the second row of Fig. 2.42). The mutation in *FBXW7* decreased survival with effect size on overall survival of mutations consistent across all fitted models (seventh row of Fig. 2.42, HR=1.62 [1.05-2.48], p-value=0.029 for univariate and HR=1.89 [1.18-3.02], p-value=0.008 for model adjusted for age, subtype, SNVs, and PGA).

2.3.4.4 Alterations in chromatin drivers are predictive of survival

As discussed earlier, alterations to chromatin modifiers are recurrent in EC. I sought to understand their effects on survival. I modeled the overall survival with respect to mutations in chromatin modifiers (full table available in the supplementary materials). Seven chromatin modifiers, including *ARID1A* and *CHD4*, discussed already in the Section 2.3.4.3, exhibit an effect on overall survival in at least one considered model (Fig. 2.43). Interestingly, samples with mutations in SWI/SNF subunits *PBRM1*, *SMARCA2* and *SMARCA4* exhibit decreased survival when modeled including age, subtype, TMB and PGA as covariates (black pointrange plots in rows five through seven in Fig. 2.43). Conversely,

mutations in *BCOR* and *CHD3* exhibit reduced hazard ratio only in univariate models (green pointrange plots in the second and third row in Fig. 2.43).

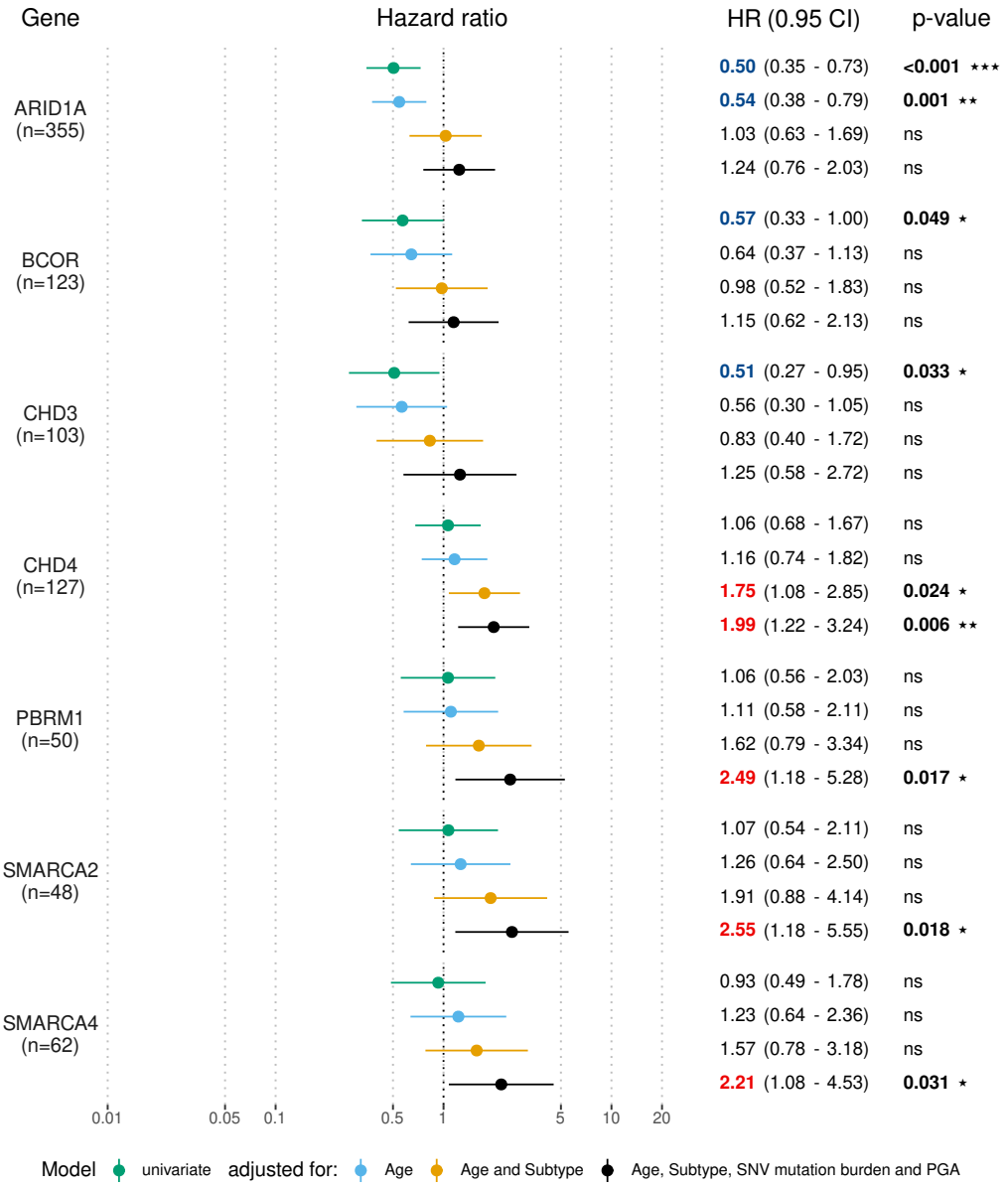


Figure 2.43: **Overall survival modeled with respect to mutations among chromatin modifiers.** Survival was modeled without any co-variates, as well as adjusting for age at diagnosis, molecular subtype and SNV mutation burden and genome alterations. Only drivers with p-value < 0.05 in any of the models shown. A full summary of fitted models is available in the supplementary materials.

2.3.5 Temporal Heterogeneity

Using the copy number profiles, sample purity estimates, and variant allele frequencies using DPClust [193] followed by extensive quality control and manual curation (described

in Section 2.5.1.5) allowed to investigate cancer heterogeneity. Most of the samples (52%) consisted of more than one clone, with twenty-three samples with four and three samples with five subclones. As shown in Fig. 2.44, 41% of CN low and 46% of CN high samples comprised of more than one subclone, while only 30% of MSI samples and 33% of POLE samples comprised a single clone.

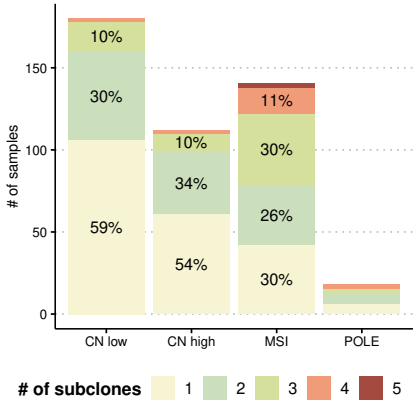


Figure 2.44: **Number of subclones per sample across molecular subtypes.** Bar plots consist of a number of samples with a given number of subclones and are annotated with the percentage of all analyzed samples in that subtype. Samples in the POLE subtype consist of 33% samples with one subclone, 33% with two subclones, and 17% with three and four, respectively.

Mutations can be categorized as clonal, shared by all cells in a sample, or subclonal, with only a subset of cancer cells possessing a particular mutation (Fig. 2.45). Clonal mutations can be further classified as early or late clonal mutations, indicating their relative order of occurrence. However, determining the timing of mutations is challenging and can only be achieved if the mutation co-occurs with a copy number alteration. For instance, a clonal mutation detected on two chromosome copies following a copy number gain suggests the mutation preceded the amplification. Conversely, if the copy number gain occurred before the mutation, it would be present on only one chromosome copy. A higher ratio of early and late to undetermined clonal mutations is observed in samples with elevated PGA and those belonging to the CN high subtype (Fig. 2.45).

I investigated the subclonal fraction of mutations in genes identified as drivers, and the distribution of subclonal and clonal mutations are visualized in Fig. 2.46. Notably, all mutations in *MYCN*, *HRAS*, and *EIF1AX* are clonal. Over 90% of mutations are clonal

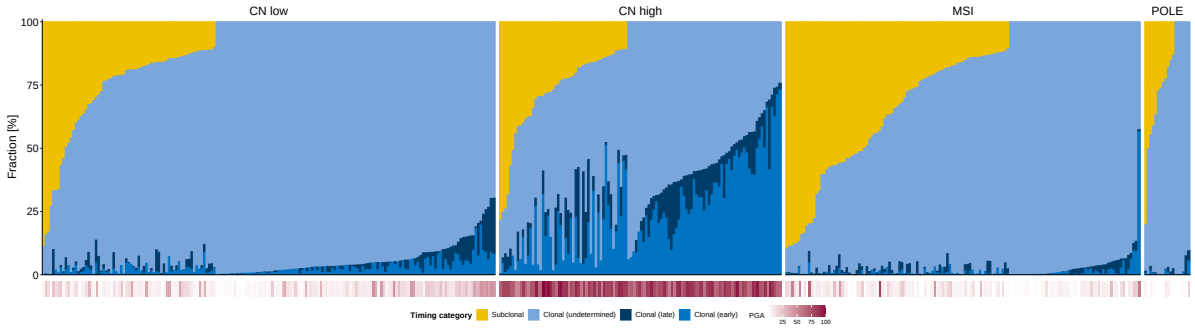


Figure 2.45: **Fractions of clonal and subclonal mutations across molecular subtypes.** Fraction of mutations annotated as subclonal and clonal - early, late, and undetermined shown. Samples are sorted according to overall subclonal and clonal contents. Percent genome aberrant (PGA) for additional context plotted.

in twenty-one drivers, with 48.1% classified as early clonal mutations in *TP53* and 30.8% in *SPOP*. Interestingly, the proportion of subclonal mutations across chromatin modifiers is higher, with over 25% of mutations categorized as subclonal in 13 out of 29 SWI/SNF complex members (Fig. 2.46A), 5 in 12 NuRD complex members (Fig. 2.46B), and 1 in 4 PRC1.1 complex members (Fig. 2.46C).

Of all coding mutations, 21.4% of missense, 28.4% of frameshift, 23.1% nonsense, and 22.2% of other mutations are subclonal Fig. 2.47A. To investigate potential differences in subclonal and clonal patterns between specific hotspot and pathogenic mutations in *POLE*, *BCOR*, and *SETD1B*, and other mutations in those genes, I examined their respective distributions. In the case of *POLE*, the majority of mutations are clonal, with 94.12% of pathogenic mutations, as defined in [163], being clonal (Fig. 2.47B). As discussed earlier, *BCOR*, a transcriptional repressor, harbors the N1459S mutation, which is almost exclusive to EC. All N1459S mutations were found to be clonal, whereas 21.62% of other missense mutations were subclonal (Fig. 2.47C). In the context of SETD1B, 12.50% of frameshift mutations and 12.82% of missense mutations were subclonal, while 17.07% of frameshift mutations detected in the 8th codon were subclonal (Fig. 2.47D).

PIK3CA, *PTEN*, and *TP53* mutations are found predominantly clonal, with a substantial fraction of the mutations classified as clonal early (Fig. 2.47E-G). The early occurrence of these mutations is consistent with their frequent presence in normal tissue and precancerous lesions [33, 34, 37].

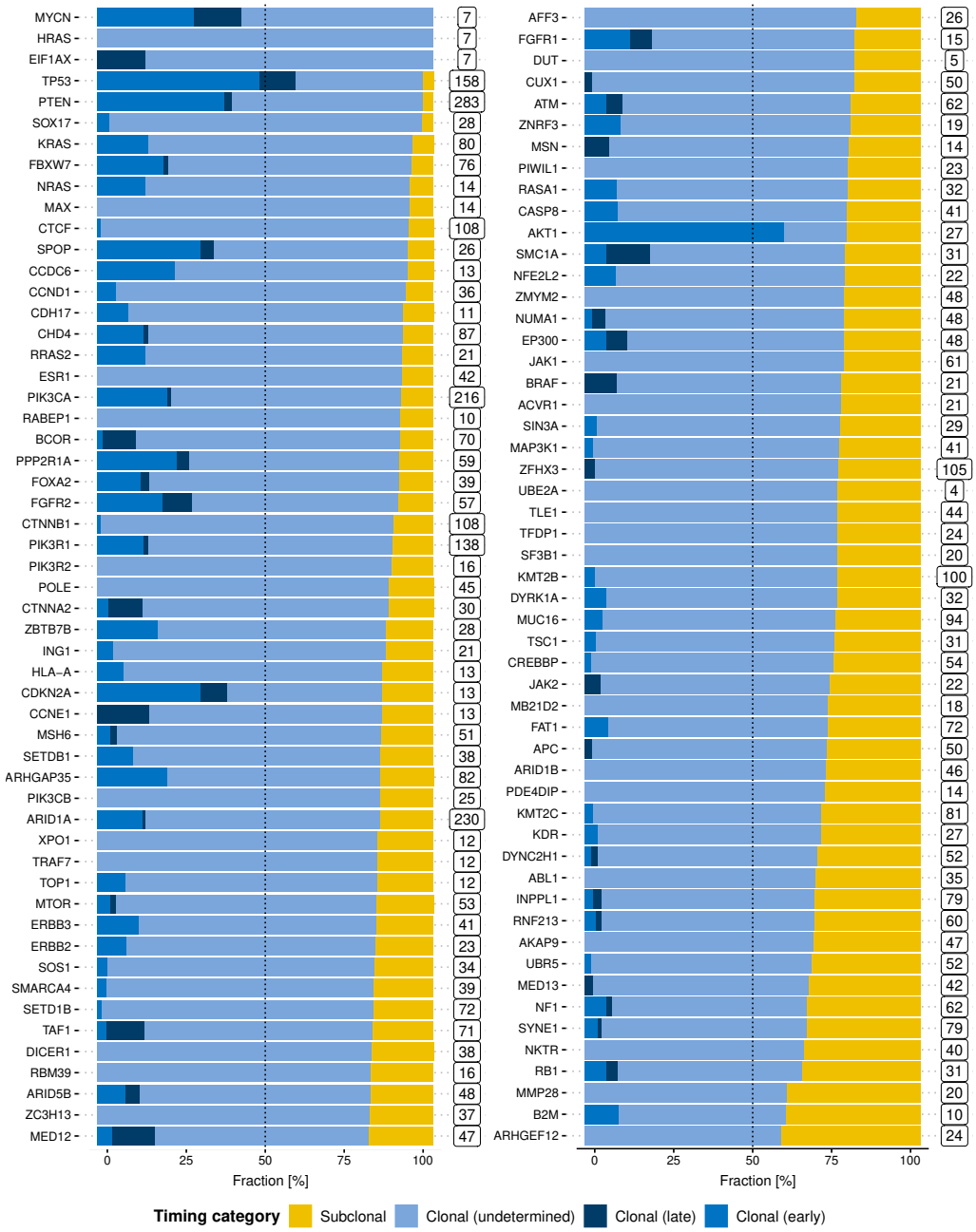


Figure 2.46: **Fraction of clonal and subclonal mutations in driver genes.** Fraction of subclonal and clonal - early, late, and undetermined shown. Genes are sorted based on their subclonal content. The number of samples with a mutation in a gene annotated with label to the right.

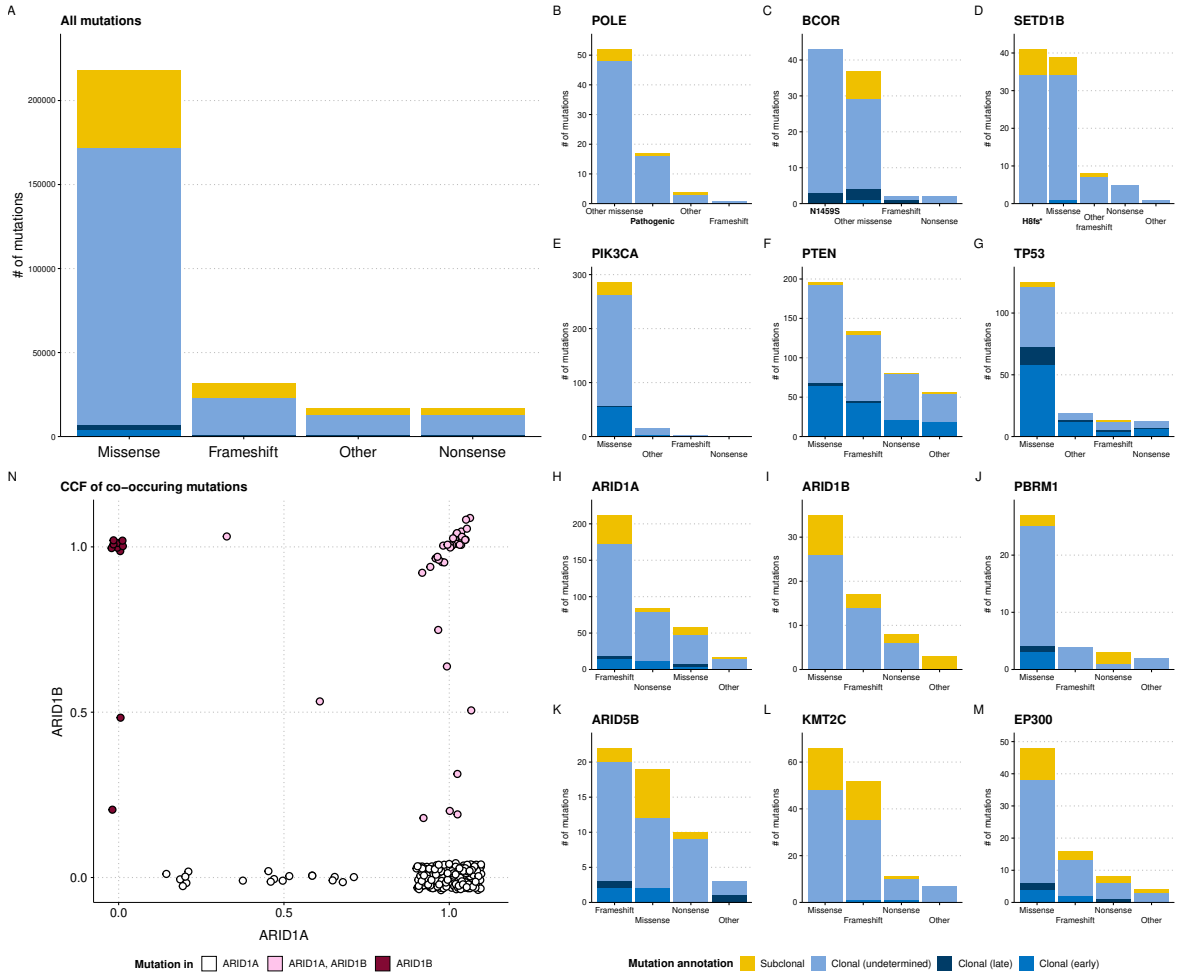


Figure 2.47: **Fraction of clonal and subclonal annotations across mutations. A** Fractions across all mutations. **B-D** Fractions shown in hotspot and other mutations. **E-G** Fractions shown in selected drivers. **H-M** Fractions shown in selected chromatin modifiers. **N** Cancer cell fractions (CCFs) of co-occurring mutations. A slight nudge was added to overlapping CCFs for visualization purposes.

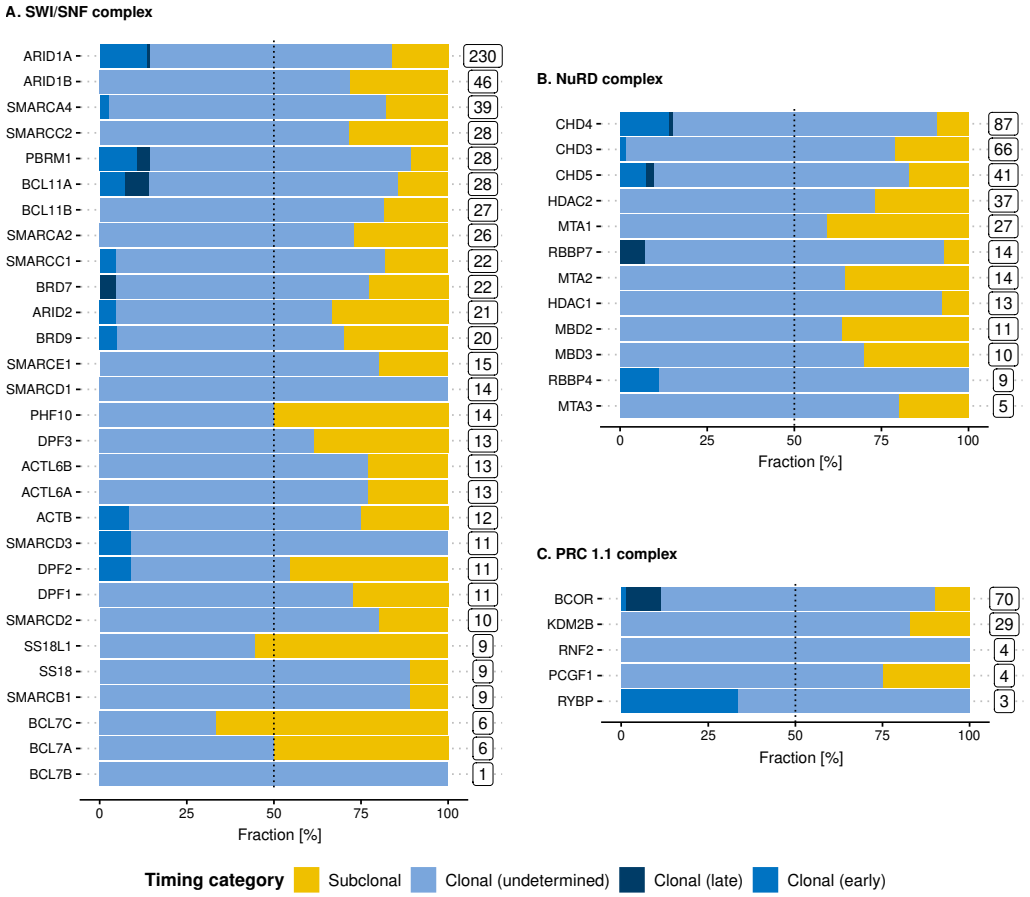


Figure 2.48: **Fraction of clonal and subclonal mutations in selected chromatin modifiers genes.** Fraction of subclonal and clonal - early, late, and undetermined shown. Genes sorted based on their subclonal content. The number of samples with a mutation in a gene annotated with a label to the right. Genes sorted in descending order of mutations. **A** Members of the SWI/SNF complex shown, **B** NuRD complex members and **C** PRC1.1 complex members.

Frameshift mutations in ARID1A and ARID1B are clonal in 81.13% and 82.35% of cases, respectively (Fig. 2.47H-I). Among 36 samples harboring ARID1A and ARID1B mutations, 27 exhibit clonal mutations in both genes. Although initial reports indicated synthetic lethality of both mutations [189], recent research suggests that dual loss can contribute to rapid carcinogenesis [190].

All frameshift mutations and 92.59% of missense mutations in *PBRM1* are clonal; however, 33.33% of nonsense mutations are subclonal Fig. 2.47J).

Subclonal fractions of missense mutations in *ARID5B* (36.84%), *KMT2C* (27.27%), and *EP300* (20.83%) are higher or similar to those across all mutations (Fig. 2.47K-M), which could suggest that these mutations occur later in tumor evolution. Nevertheless, a formal analysis of cancer evolution is necessary to uncover evolutionary trajectories and discern the sequence of acquired mutations in cancer development. I propose a deep learning approach to uncover evolutionary trajectories from whole-genome sequencing data in Chapter 4.

2.4 Discussion

In this chapter, I presented results from analyses of the largest EC cohort to date, offering a comprehensive analysis of genomic alterations across all four molecular subtypes. Our findings validate previous results from smaller cohorts and reveal novel drivers, copy number alterations, and prognostic associations. We characterized mutation and copy number alteration profiles and provided the first description of the landscape of whole genomes in EC.

In our analysis, we identified 107 candidate driver genes, including novel candidates like *FGFR1*, *SF3B1*, *ZBTB7B*, and *ZNRF3*, alongside notable copy number alterations. Particularly, we observed a significant deletion at 15q11.2, encompassing *SIN3A*, in both MSI and POLE subtypes, mirroring recent findings that implicate SIN3A suppression by miRNA in breast cancer metastasis [175]. Additionally, our study successfully replicated known drivers such as *PTEN*, *ARID1A*, *PIK3CA*, *PIK3R1*, and *TP53* across all subtypes. Notably, certain drivers appeared specific to particular molecular subtypes. Our findings

extend beyond the replication of 70 known endometrial cancer (EC) drivers, identifying an additional 22 known pancancer drivers newly associated with EC.

The TCGA study previously identified *FBXW7* and *PPP2R1A* mutations as unique to the CN high subtype [13]. However, our analysis revealed *FBXW7* as a driver gene across CN low, CN high, and MSI subtypes, exhibiting a consistent mutation spectrum (Fig. 2.18). Furthermore, I show that mutations in *FBXW7* decrease the survival EC, even when accounted for age, mutation burden, and subtype. Intriguingly, the consequence of mutation in *FBXW7* seems to mirror that of *CCNE1* amplification observed in ovarian cancer [171, 172], suggesting a possible phenocopying effect and underscoring the complexity of genetic interactions in cancer biology. More focused research would be needed to explore the causality of the *FBXW7* mutation. However, translating these findings to animal models, as illustrated in [194], underscores the complexities of *FBXW7* in tumor biology. The studied mouse models reveals that *Fbxw7* mutations alone are insufficient for endometrial tumor development, requiring additional genetic alterations such as *Pten* loss or *Trp53* mutations. This suggests a more intricate interplay of genetic factors in endometrial cancer, emphasizing the need for comprehensive models to understand the multifaceted roles of these genes in tumorigenesis. Further investigation of co-occurrence and mutual exclusivity of the driver mutations would additionally help understand their respective contributions to tumorigenesis.

Study of the impact of driver mutations is compounded by the observation that drivers commonly mutated in the GEL EC cohort, are also present in the normal endometrium and endometriotic tissue [34, 37]. Although compared to the burden of somatic mutations in normal endometrium, EC exhibits an increase in mutations ranging from 3-fold for samples classified as CN low to 651-fold for those of the POLE subtype both tissues share mutations in several genes. The study of normal endometrium identified twelve cancer driver genes (*PIK3CA*, *PIK3R1*, *ARHGAP35*, *FBXW7*, *ZFHX3*, *FOXA2*, *ERBB2*, *CHD4*, *KRAS*, *SPOP*, *PPP2R1A*, and *ERBB3*) as genes under positive selection in normal endometrial glands [37]. All of those drivers were identified in our driver analyses of the GEL EC cohort, with mutations in *PIK3CA* and *PIK3R1* found in 52% and 35% of

all samples, respectively. A study of patterns of selection in cancer and somatic tissues found that, on average, endometrial cancers carry above ten mutations under positive selection [38], suggesting an important interplay of cancer drivers. Understanding the timing of genomic alterations and mutations in cancer drivers specifically would help pinpoint the relationship between cancer drivers and disease progression. However, a robust method that would be able to leverage single tumor biopsies to reconstruct tumor evolution is needed. I describe my efforts to contribute to solving this problem in Chapter 4. Delineating the sequence of mutational events using experimental models allowing for investigating the tumors at various timepoints would further clarify the role of driver genes in cancer progression.

EC exhibits similarities with ovarian cancer. High-grade serous carcinoma (HGSC) is the most commonly diagnosed form of ovarian cancer. It is characterized by the *TP53* loss [195] and frequent copy number changes [196] similar to the CN high subtype of EC. HGSC mutation burden of 61 coding mutations [196] is close to that of CN low (49 coding mutations) and CN high (57 coding mutations). Contrary to the HGSC, which originates in the fallopian tube [197, 198], ovarian clear-cell carcinoma (OCCC), and endometroid ovarian carcinoma (EOC) can originate from endometrial tissue as a result of endometriosis [199]. Similar to the EC, OCCC is characterized by a high frequency of somatic mutations in *ARID1A* [61, 200], *PIK3CA* [201] and *PTEN* [202]. Mutations in *ARID1A*, together with the tissue of origin, are a shared factor between EC, OCCC, EOC, and endometriosis. However, while there is a robust association between endometriosis and the risk of OCCC and EOC [203], the findings are conflicting in the case of EC [32, 204]. A comprehensive study comparing the molecular profiles of ovarian cancers, endometrial cancer, and endometriosis could further increase our understanding of the diseases and aspects shared, and distinct to each of them.

Of the 107 drivers identified in EC, 15 are chromatin modifiers. Among these, mutations in *ARID1A* are particularly noteworthy, representing the second most commonly mutated driver in EC and being overrepresented in EC compared to other cancer types. Recent studies have further elucidated the role of *ARID1A* in gynecological cancers.

Restoring *ARID1A* expression in OCCC harboring mutant *ARID1A* led to a suppression of proliferation and tumor growth in mice models [205], suggesting a tumor suppressor role for *ARID1A*. Further, silencing *ARID1A* in a mouse tumor xenograft model enhanced cellular proliferation and tumorigenicity [205]. This aligns with clinical observations where the loss of *ARID1A* was associated with reduced progression-free survival in OCCC [206] and with deep myometrial invasion in EC [207]. Furthermore, as mentioned earlier, OCCC and EOC, often arising from endometriosis [61], demonstrate mutations in *ARID1A* that lead to loss of BAF250a in preneoplastic lesions, indicating that these alterations could be early events in the transformation of endometriosis into cancer. In my analysis, a significant number of EC samples exhibited mutations in both genes, with clonality observed in 27 out of 36 samples carrying mutations in both genes. This evidence challenges previous reports on the synthetic lethality of mutations in both *ARID1A* and *ARID1B* [189]. In turn, it supports an alternative hypothesis where the loss of both ARID1 genes might provide a growth advantage for the tumor [190].

Additionally, I showed that the N1459S *BCOR* mutation found in nearly 10% of EC samples is nearly unique to EC. All N1459S mutations are clonal, whereas over 20% of other missense SNVs were found to be subclonal. While this particular mutation is nearly exclusive to EC, *BCOR*'s alterations have been reported in other malignancies. Internal tandem duplications in the last exon of *BCOR* have been found in the majority of cases of clear cell sarcoma of the kidney [208, 209]. *BCOR*'s fusions with *CCNB3* have been reported in undifferentiated round cell sarcomas [210] and the *ZC3H7B-BCOR* fusion has been used as a marker of a distinct group among high-grade endometrial stromal sarcomas [211]. The role of the N1459S *BCOR* mutation in EC is not well understood, and functional studies of the N1459S *BCOR* mutation have not been performed. I describe our effort in investigating the consequences of the mutation on differentiation programs in Section 3.2.11.

I discussed the four chromatin modifiers identified as candidate EC drivers - *SMARCA4*, *SETD1B*, *CREBBP* and *SETDB1*. On top of their mutations, *SMARCA4* was also found in the region recurrently deleted. Conversely, *SETD1B* characterized by frameshift mu-

tation in the 8th codon was also found in region recurrently amplified.

Proteins often carry out their functions in coordination with other proteins, forming intricate complexes and interacting across both small and large pathways. Inactivating distinct components of a pathway can frequently result in similar effects on the entire network. In the context of endometrial cancer, the SWI/SNF pathway is commonly altered, with various members being affected by different molecular processes across distinct EC subtypes. In addition to alterations in *ARID1A* and its paralog *ARID1B*, I identified several other frequently altered pathway members. For instance, *SMARCA4* is altered in 23% of EC samples in 14 (7.3%) of CN low samples, all copies of the gene are lost. Furthermore, I found that samples with mutations in SWI/SNF subunits *PBRM1*, *SMARCA2*, and *SMARCA4* are associated with decreased survival.

In this chapter, I presented the first descriptive analysis of mutations in chromatin modifiers. It is important to recognize that these analyses are inherently limited to correlative assessments. Additional functional validation through tailored experiments is necessary to explore the consequences of these mutations. To further investigate the role of chromatin remodelers in EC, we complemented this analysis with gene editing of normal and malignant endometrial cells by CRISPR-Cas9. I describe this work in Chapter 3.

There are two major efforts to implement molecular insights into the clinical practice of EC care: Proactive Molecular Risk Classifier for Endometrial Cancer (ProMisE) [69] and Post-Operative Radiation Therapy in Endometrial Carcinoma (PORTEC) [70]. Both groups have shown added value of using molecular subgroups in clinical practice, while also reporting that the molecular subgroups should be integrated with other prognostic factors to further improve the prospective care [18, 212].

Several molecular alterations have been associated with clinical outcomes, such as mutations in *CTNNB1* [213, 214], overexpression of L1-cell adhesion molecule (L1CAM) [215–217], estrogen and progesterone receptor loss [217–220], chromosome 1q amplification [221] and other copy number alterations [222]. PORTEC4a is the first clinical trial to prospectively asses the molecular subgroups together with other molecular prognostic factors such as substantial lymphovascular space invasion, L1CAM expression, and *CTNNB1*

mutation to assign patients into the appropriate risk groups [18]. However a recent study of retrospective analyses of PORTEC3 trial cohort has shown that only estrogen receptor status has proven to hold prognostic value among the CN low, also known as no specific molecular profile (NSMP) subtype [223]. NSMP is characterized by a heterogeneity in clinical outcomes [70, 224, 225] which substantiates the need for identification of new prognostic features that would allow for better risk stratification.

I demonstrated that tumor mutation burden and PGA are prognostic of survival even when adjusted for age and molecular subtype. Additionally, I found that mutations in *FBXW7* increase the risk of death (HR=1.89 [1.18-3.02], p-value 0.008; model adjusted for age, molecular subtype, SNVs, and PGA), while mutations in *FAT1* are associated with a decreased risk (HR = 0.44 [0.22-0.89], p-value 0.023; adjusted for age, molecular subtype, SNVs burden, and PGA). In summary, this research outlines the association between cancer driver genes and patient survival in the EC cohort. It is crucial to emphasize that future studies will further scrutinize the findings, including an in-depth investigation of selected hotspots and oncogenic mutations. The results will be validated through formal follow-up analyses in other cohorts, such as the PORTEC4 study [18], further solidifying our understanding of the relationship between cancer drivers and survival in endometrial cancer. Furthermore, an investigation of censor dates, as well as more extensive modeling and the inclusion of treatment-free intervals, particularly for CN high EC cases, would increase our understanding of the survival patterns in endometrial cancer, and would further test the results of our analyses.

In this work, I explored the heterogeneity of EC tumors, providing insights into the number of subclones present in samples across molecular subtypes and the clonality of mutations across samples and within driver genes. However, to fully understand the evolution of EC, a more comprehensive analysis of cancer evolution is required to unveil evolutionary trajectories and determine the sequence of acquired mutations during cancer development. To address this, I propose a deep learning approach for deciphering evolutionary trajectories from whole-genome sequencing data, as detailed in Chapter 4.

2.5 Methods

2.5.1 Genomics England data

Samples were processed according to the GEL Sample Handling Guidance available on the GEL website. After sample collection and DNA extraction, samples were sent to Illumina Sequencing Centre at the Wellcome Genome Campus, Cambridgeshire, for library preparations and sequencing. Formalin-fixed paraffin-embedded (FFPE) samples were excluded from analyses, keeping only high quality 665 fresh frozen tumor-germline pairs (Fig. 2.49).

2.5.1.1 Sample processing

Samples were whole genome sequenced (WGS) to x100 average sequencing depth for tumors and x30 for germlines. The data was processed centrally at Genomics England via an automated pipeline described in detail in the Technical information v1.11 main document on the GEL website. Illumina's North Star pipeline (version 2.6.53.23) was used for primary WGS analysis. Reads were aligned with ISAAC (version iSAAC 03.16.02.19) [226] to human genome assembly GRCh38, the variants were called with Strelka (version 2.4.7) [227] and the germline variants were subtracted from the set.

Variants were then further filtered based on their quality. Combined maf files were then created with the vcf2maf tool [228].

2.5.1.2 Molecular subtypes and quality assessment

The samples were then extensively evaluated by histopathologists, and the sample set was manually curated, yielding 665 samples passing QC (Fig. 2.49). Eight of those are metastatic samples, and 657 are primary tumors. Endometrioid tumors comprise 71% of all primary tumors, with 14% of samples classified as serous and 7% as carcinosarcomas. The sample set also includes 23 cases of mixed endometroid and non-endometroid, 22 cases of clear cell, and 5 undifferentiated carcinomas. In my analyses, I restricted the evaluation of mutations to primary cancer samples (Fig. 2.50).

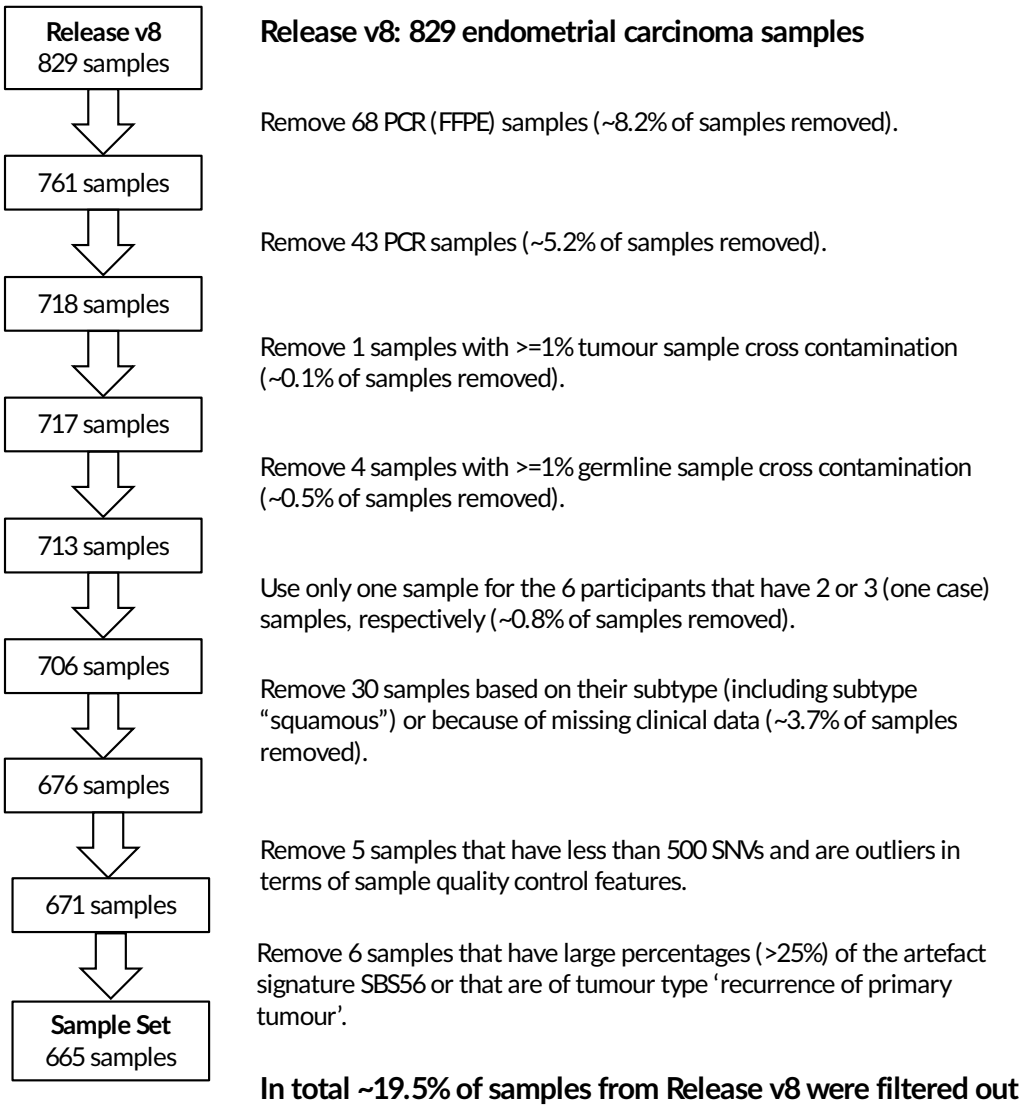


Figure 2.49: **Diagram describing exclusion criteria of samples in GEL EC cohort.** Release v8 was used for analyses with 829 endometrial carcinoma samples assembled into the GEL EC cohort. All 68 FFPE samples were excluded from the analyses (resulting in a set of n=761 samples). Only PCR-free samples were retained and samples with above 1% cross-contamination were filtered out. For tumors with multiple samples, only one sample was retained. Further, 30 samples were removed based on missing clinical data or based on their mismatched subtype. Next, samples with outlier quality control features, >20%, and recurrence tumors were filtered out, resulting in a cohort of 665 high-quality samples. Diagram created and extracted from GEL research environment by Andreas Gruber, whom I am deeply grateful for his assistance.

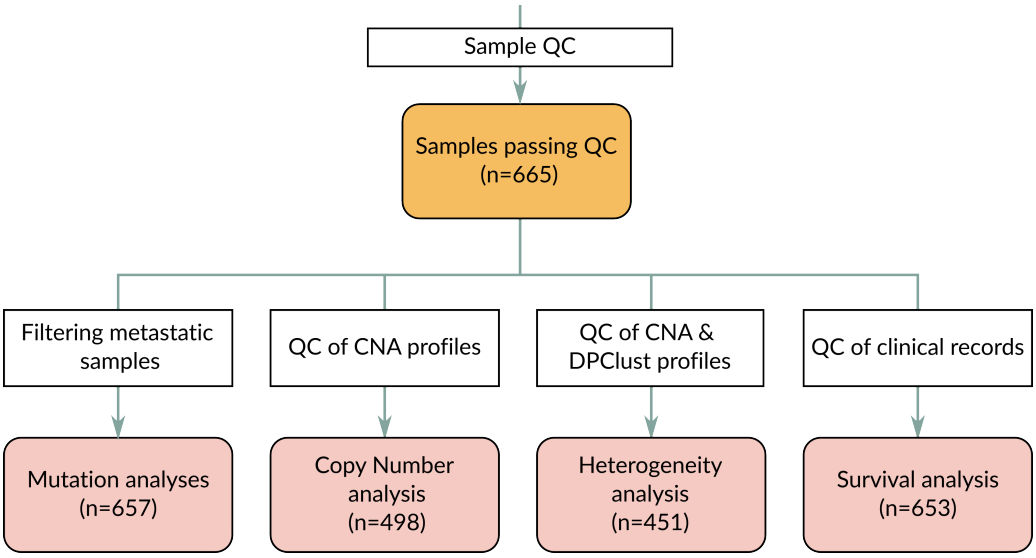


Figure 2.50: **Diagram describing exclusion and filtering of the data in my analyses.** First, the sample set was subjected to extensive quality control and evaluation of histopathology as visualized in Fig. 2.49 (resulting in n=665 samples). Further filtering of the samples was performed with respect to each of the analyses, which I describe in relevant sections of the Methods. Briefly, for mutation analyses, only primary tumors were selected (n=657). For copy number analyses, the sample set was further narrowed to those with a high-quality copy number profile (n=498). For heterogeneity and mutation timing analyses, the samples with valid copy number profiles were used for `DPClust`, of which 47 samples were excluded due to `DPClust` run fail and quality control filtering (n=451). For survival analyses, samples without complete clinical records were excluded (n=653).

Samples were assigned to molecular subtypes as follows:

- samples were assigned to the POLE subtype based on the detection of pathogenic variants described in *POLE* gene [163] (44 samples);
- remaining samples were assigned to MSI subtype based on theit MSI/MSS profiles called with mSINGS [164] (201 samples);
- remaining samples were split into CN high and CN low based on a threshold of percent genome aberrant (PGA) chosen as the minimum of the smoothed binomial distribution. 146 samples were assigned to CN high, the remaining 274 samples were assigned to CN low subtype.

2.5.1.3 Genomic analyses

Copy number analysis Somatic copy number aberrations (CNAs) were called using a framework implemented in the CleanCNA R-package as described in [229]. Briefly, genome-wide subclonal CNAs were called using Battenberg [193], and their quality was assessed. In the case of samples, a profile was not passing the quality thresholds, the purity estimate was adjusted, and CNAs were refit until the sample passed QC or the procedure was repeated four times, whichever happened first. Passing CNA profiles were produced for 498/665 samples (75%, Fig. 2.50). Sample composition of the primary tumors, including samples that pass CNAs quality control, is described in Table 2.4. CNAs for GISTIC [173] were prepared as described in [229]. GISTIC was run on the whole cohort and in each cohort of molecular subtypes separately.

	ALL	passing CNAs QC
CN low	272 (41%)	192 (39%)
CN high	142 (22%)	112 (23%)
MSI	199 (30%)	155 (32%)
POLE	44 (7%)	32 (6%)

Table 2.4: Subtype composition of primary tumors in GEL EC cohort.

Coding driver identification Protein-coding driver genes were identified using the IntOGen pipeline [82] as described in [229].

Structural variants Structural variants (SVs) were identified using a graph-based consensus approach comprising Delly [230], Lumpy [231] and Manta [232] whilst also considering support from CNAs as described in detail in [229].

Signatures Single-base-substitution (SBS), doublet-base-substitution (DBS), and insertion and deletion (indel; ID) signatures were extracted de novo and related to known COSMIC [166] signatures (v3.2) using SigProfilerExtractor [233]. SigProfilerExtractor was also used to extract copy number signatures in the tumors with profiled CNAs. SVs were grouped into footprints using the string approach proposed in [234]. SV clusters were classified as being a simple event (deletion, tandem duplication, balanced inversion, balanced translocation, unbalanced translocation, or simple unclassified) or a complex event (chromoplexy, chromothripsis, or complex unclassified). A detailed description of signatures analysis is available in [229]. To distinguish the number of genomic alterations from the information regarding their associated mutational signatures, the proportion of alterations attributed to each signature for each patient was calculated as follows:

$$s_i = \frac{n_i}{N_t},$$

where s_i represents the contribution of the i-th signature, n_i denotes the number of alterations assigned to that signature, and N_t signifies the total number of alterations across all signatures of a given type (e.g., SBS, DBS, etc.). For unsupervised clustering, the distance metric (d) was based on the Kendall rank correlation coefficient (τ), calculated using the set of signature contributions for each sample defined above. Specifically,

$$d_{m,k} = 1 - \tau(S_m, S_k),$$

where m, k are two samples, and S_m is a set of signature contributions $S_m = s_{m,i}$ for sample m.

Hazard Ratio	Power
0.25	1.00
0.50	0.97
0.75	0.35
1.25	0.23
1.50	0.60
2.00	0.97

Table 2.5: Power calculation for the survival analysis, with the event probability of 0.18%, significance level of 0.05, a sample size of 653, and using the double-sided test.

2.5.1.4 Survival analysis

Clinical data were manually curated, resulting in a comprehensive dataset comprising 653 samples (Fig. 2.50) with 119 observed events. The median observation time for all events was 2.90 years, while the median time to censoring (no death) was 3.04 years.

Power calculation Power analysis was performed with the `survivalpwr` [235] event probability equal to 18% (as calculated based on observed events), significance level of 0.05, the sample size of 653 and the double-sided test Table 2.5.

Survival curves were fitted using the Kaplan-Meier method, and the differences between groups were assessed using log-rank tests. Hazard ratios were estimated via Cox's Proportional Hazards (Cox PH) model, and the proportional hazards assumption was tested. Univariate and multivariate Cox PH models were fitted to evaluate the prognostic impact of molecular variables (tumor burden, percent genome aberrant, molecular subtype), driver mutations, and chromatin modifiers. Statistical significance was determined at a two-sided p-value of < 0.05. All survival analyses were performed in R [236] using survival package [237], and Kaplam-Meier survival curve and forest plot in Fig. 2.38 were created with survminer [238]. Other packages used for exploratory data analyses and bespoke plotting are described in more detail in Section A.2. The relationship between single nucleotide variants (SNVs) per megabase (Mb) and insertions/deletions (indels) per Mb was assessed using a test for association between paired samples. This test employed Pearson's product-moment correlation coefficients.

2.5.1.5 Heterogeneity analysis

I compared the distribution of the variant allele frequencies (VAFs) between single nucleotide variants (SNVs) and short insertions and deletions (indels). Upon visual inspection of distributions and using the Kolmogorov–Smirnov test, I decided to merge the indels with SNVs without adjusting their VAFs. I then processed merged sets of mutations using `dpclust3p` R package [239] and ran `DPClust` [193] to identify the subclones. For 20 samples, the DPClust run failed. For the remaining 478 samples, I assessed the quality of DPClust calls as follows:

- I removed all clusters with less than 10% of all mutations,
- then, I removed samples that did not have a cluster with $0.85 < \text{CCF} < 1.15$; most likely, purity estimates for those samples were overestimated (27 samples removed);
- I merged all clusters spanning $0.85 < \text{CCF} < 1.15$ range ad calculated CCF as weighted by the size of the cluster average and adjusted all CCFs of a given sample setting clonal cluster at $CCF = 1$

This produced a QC passing heterogeneity calls for 451 samples (Fig. 2.50) with 112 samples from CN low subtype, 180 CN high samples, 141 MSI samples, and 18 POLE samples. 221 samples contain at least one subclonal cluster beyond the clonal one.

Mutation copy number (m) refers to the number of chromosomes carrying the mutation, and the ploidy of sample (C) refers to the total copy number of the tumor sample. I assigned the mutations into four categories - subclonal and clonal early, late, and undetermined by adapting the approach described in [240]. Briefly,

- I assigned mutations to the clonal early set if the mutation was clonal and amplified, i.e., $\text{CCF} = 1$ and $m > 1$;
- then I annotated mutations as clonal late if the mutation was clonal, not amplified and the ploidy of the sample was higher than diploid, i.e., $\text{CCF} = 1$, $m = 1$ and $C > 2$;
- the remaining clonal mutations were annotated as clonal undetermined,
- and finally the mutations with $CCF < 1$ were annotated as subclonal.

2.5.1.6 Research environment

In light of privacy concerns, it is necessary for the analysis of GEL data to be conducted in a secure and regulated environment, specifically a research environment (RE). The GEL RE operates without internet connectivity, and all data imports and exports must be facilitated through the special portal requiring external approval. Regrettably, the infrastructure available to researchers has at times proven to be less than optimal, with frequent outages and constrained resources. All software must be requested and may take anywhere from weeks to months for approval. The documentation has improved over time, but it could still benefit from more clarity, particularly regarding export and import procedures, which can sometimes be opaque and discretionary.

In summary, the limited access to data and the constraints on computational resources have been identified as the primary challenges when working with the GEL resource. It is important to recognize that these factors can create difficulties for researchers and potentially hinder the progress of valuable findings. As a result, some planned analyses had to be postponed or reconsidered.

2.5.2 TCGA somatic variation

mc3.v0.2.8.PUBLIC.maf.gz was downloaded from National Cancer Institute Genomic Data Commons. The complex pipeline used for calling and filtering variants is described in [158]. Sample acquisition, preparation and sequencing for the UCEC and UCS cohorts is described in [15, 241].

Chapter 3

Perturbations of chromatin modifiers in endometrial cancer

Contents

3.1 Background . . . 112
3.1.1 Contributions . . . 112
3.2 Results . . . 113
3.2.1 Selection of cell lines for CRISPR screen . . . 115
3.2.2 CRISPR validation . . . 116
3.2.3 Cell line and genomic perturbations are dominant sources of variation . . . 118
3.2.4 *ARID1A* loss disrupts chromatin accessibility and gene expression in endometrium and EC models . . . 121
3.2.5 Loss of *ARID1B* reveals perturbed gene expression profiles . . . 136
3.2.6 *ARID5B* loss results in up-regulation of interferon response pathways . . . 146
3.2.7 JUND motifs are more accessible in *EP300* knockouts . . . 154
3.2.8 Loss of *KMT2C* results in up-regulation of E2F targets . . . 160
3.2.9 Loss of *SETD1B* down-regulates E2F targets . . . 164
3.2.10 Common perturbations induced by the loss of chromatin modifiers 169
3.2.11 Functional study of *BCOR* N1459S hotspot mutation . . . 175
3.3 Discussion . . . 182
3.3.1 Functional study of *BCOR* N1459S hotspot mutation . . . 187
3.4 Methods . . . 189
3.4.1 Chromatin modifiers in hEM3 and Ark2 cells . . . 189
3.4.2 *BCOR* N1459S experiment . . . 192

3.1 Background

In Section 2.3.2, I presented the results from a detailed correlative analysis of chromatin modifiers in the GEL endometrial cohort. Fifteen chromatin modifiers were identified as drivers, including those highly recurrently mutated genes such as *ARID1A* or *KMT2C*. These genes are known to play important roles in regulating chromatin structure and gene expression. Their dysregulation has been linked to the development and progression of several cancer types, including endometrial cancer [60, 242, 243]. My analysis also identified other, less well-characterized drivers, such as *BCOR*, which carries a point mutation near-unique to endometrial cancer. While this analysis provides a comprehensive overview of genetic alterations in chromatin remodellers in cancer, it does not provide insights into the mechanisms that drive tumorigenesis.

In this chapter, I recount the results of my efforts to understand the biological consequences of the disruption of chromatin modifiers in endometrial cancer. I will focus mainly on the functional analyses of cell models I designed to understand the consequences of the loss of chromatin modifiers selected based on my analyses described in Chapter 2. I will then describe our efforts to understand the consequences of the N1459S hotspot *BCOR* mutation - near-unique to EC.

3.1.1 Contributions

The work described in this chapter was carried out in collaboration with members of the Church group at the Wellcome Center for Human Genetics, University of Oxford.

Alphabetical list of past and present members of the Church group during the undertaking of work presented in this chapter:

- Reem Ali
- Matthew Brown
- David Church - principal investigator
- Yannick Comoglio (YC)
- Yun Feng
- Mark Glaire
- Luciana Gneo (LG)
- Christos Katsifis (CK)
- **Katarzyna Kedzierska (KK)**
- Laura Knight (LK)

- Charlotte Moore
- Sofia Obolenski
- Vassilena Sharlandjieva (VS)
- Tijana Tomasevic (TT)

Contributions to experimental design, data generation, and analysis are highlighted in Table 3.1. Additionally, I would like to acknowledge Matthew Brown for his helpful discussion and comments on all the work presented in this chapter. Sofia Obolenski was a Master's student interned in the group whom I advised and helped supervise during her attachment. Sofia worked on characterizing the *BCOR* N1459S mutation in the TCGA cohort, working parallel to analyses described in Section 2.3.2 that served as the basis for the functional work described later in this chapter.

	Experimental design	Experimental work	Data processing and analysis
Chromatin modifiers	**KK** YC (refining selected genes, CRISPR)	YC (CRISPR, cell culture, RNA extraction, Sanger sequencing) VS (CRISPR) CK (CRISPR) LC (cell culture, Western blot) LG (Western blot) TT, LG (sample handling) Diagenode (sample QC, library preparation, sequencing)	**KK**
N1459S BCOR	YC **KK** (advisory capacity)	OTC (CRISPR) YC (cell culture, RNA extraction) OGC (sample QC, library preparation, sequencing)	OBSGC (alignment and quantification) **KK** (quality assessment, processing, differential expression)

Table 3.1: **Contributions to the experimental design, data generation, and analysis of experiments described in this chapter. OGC** - Oxford Genomics Core, **OTC** - Oxford Transgenic Core, **OBSGC** - Oxford Bioinformatics and Statistical Genetics Core

3.2 Results

ARID1A, *ARID1B*, *ARID5B*, *KMT2C*, *EP300*, and *SETD1B* are genes encoding chromatin modifiers implicated in the development and progression of endometrial cancer. We identified all six genes as drivers in the GEL EC cohort, confirming other studies for all but *SETD1B*, which we identified as a novel candidate driver in EC. Although mutations in those genes were previously reported, studies to understand the mechanisms by which the alteration of these genes contributes to endometrial cancer pathogenesis are lacking. Addressing this is critical for understanding disease biology and developing targeted therapies for this disease. I, therefore, designed a series of experiments to investigate the consequences of the loss of those chromatin modifiers. A schematic of the experimental

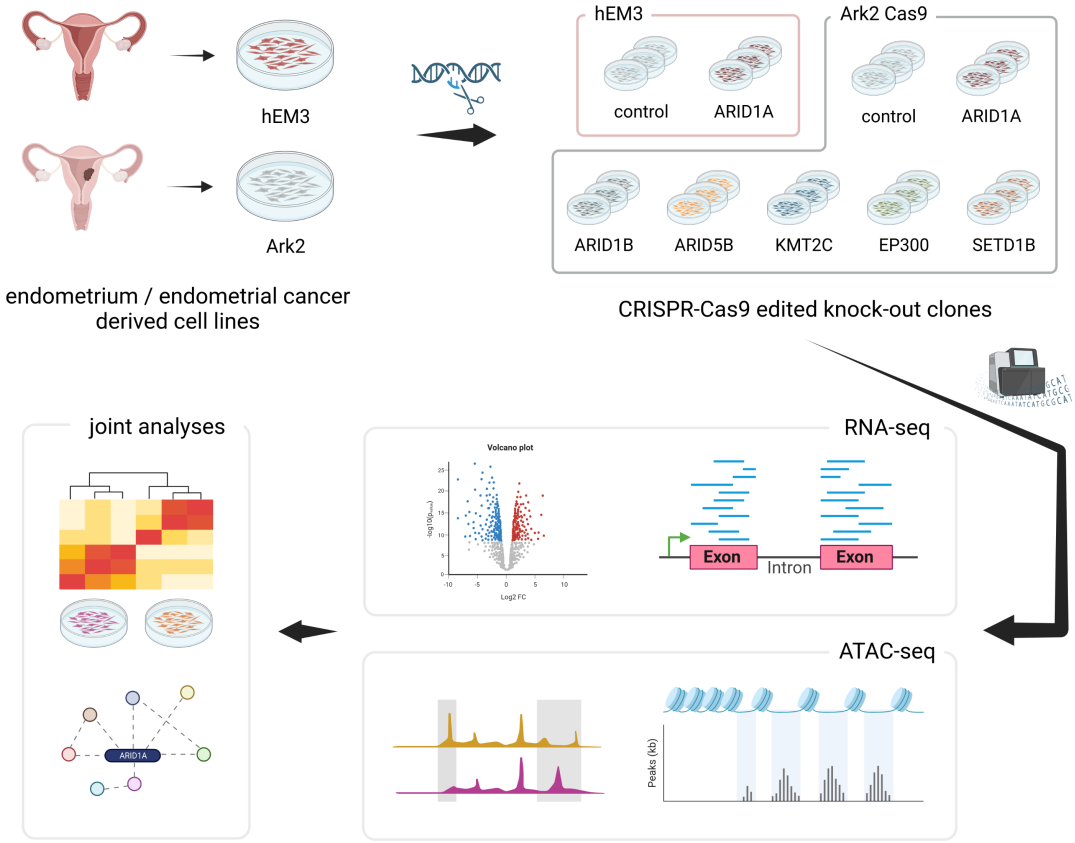


Figure 3.1: **Overview of the experiment designed to investigate the consequences of loss of chromatin modifiers.** Two cell lines were included in the experiment. **hEM3** - cell line derived from normal endometrium and **Ark2 Cas9** - cell line derived from endometrial cancer tissue further modified to stably express Cas9. CRISPR-Cas9 was used to generate knockout clones, and selected clones were grown and harvested at the confluence at three distinct passages. RNA-seq and ATAC-seq experiments were performed. Data from each experiment was analyzed separately and in conjunction. Figure created with BioRender.com.

workflow is provided as Fig. 3.1. Specifically, we acquired a cell line derived from normal human endometrium (hEM3) [244], and an hEM3 *ARID1A*^{KO} [245]. Using CRISPR-Cas9 gene editing, we[1] generated a panel of isogenic cell lines derived from the Ark2 endometrial cancer cell line [246], each with a knockout of one of the selected chromatin drivers. I discuss the steps I took to validate the CRISPR clone generation in detail in Section 3.2.2.

As illustrated in Fig. 3.1, we measured expression profiles (RNA-seq) and chromatin accessibility (ATAC-seq) from each clone and parental cell line. The former allows us to identify the differentially expressed genes and gene sets. With ATAC-seq, we identify genomic regions accessible to transcription factors and other DNA-binding proteins. From joint analyses, we can link the changes in the activities of transcription factors to changes in expression levels.

3.2.1 Selection of cell lines for CRISPR screen

To investigate the consequences of chromatin modifiers in EC, we turned to perform functional studies of their perturbations. To select the appropriate cell line, we considered the mutations among the available cell lines and the doubling time. For that, we decided to choose Ark2 cell line [246] derived from a 62-year-old African American female with endometrial serous adenocarcinoma. Ark2 cell line harbors HER2/neu-amplification [246] and does not carry any mutations in *PIK3CA*, *ARID1A* and *ARID5B* and does not carry any deleterious mutations in the remaining chromatin modifiers (based on whole exome sequencing data generated in-house).

To further our understanding of the consequences of the loss of *ARID1A*, we carried out the RNA-seq and ATAC-seq experiments on the normal endometrium cell line - hEM3 and a *ARID1A* knockout clone [244]. The hEM3 cell line is a novel human endometrial epithelial cell line developed through immortalization and clonal selection from primary human endometrium cultures. This cell line expresses protein markers characteristic of the endometrial epithelium (such as PAX8, EpCAM, cytokeratin 7/8, and ER). It does not harbor pathogenic germline mutations in genes involved in DNA mismatch repair

[1] I designed the experiments and analyzed the data. Experimental work was carried out by other lab members as described in Section 3.1.1.

or homologous repair pathways. Despite its unlimited proliferation capabilities, the cell line does not appear to be transformed, as evidenced by its lack of tumorigenicity in immunocompromised mice [244].

To observe the consequences of the *BCOR* N1459S hotspot mutation on differentiation programs, we selected mouse embryonic stem cells (mES).

3.2.2 CRISPR validation

Using the CRISPR/Cas9, we aimed to generate the knockouts (KO) of the selected chromatin modifiers. To verify the success of the KO, we subjected the clones to rigorous validation processes, including Sanger sequencing for DNA sequence confirmation and quantitative PCR (qPCR) to measure the impact on target gene expression. While we achieved a significant reduction in the expression of target genes in qPCR validation, we faced several experimental challenges.

One of the primary challenges was ensuring the complete KO of target genes. To address this, we implemented multiple validation strategies:

- Sanger sequencing and qPCR validation: All CRISPR clones were subjected to Sanger sequencing to confirm the introduction of CRISPR-induced mutations. For each clone, we sequenced the region around the target site of each guide. We experienced problems with some of the sequencing targets and received low-quality readouts. To improve the quality of Sanger sequencing, I suggested a protocol change - the use of a molecular cloning technique that can generate higher-quality sequences instead of sequencing raw PCR products. This is being implemented at the moment. While a few Sanger sequencing has to be repeated, we obtained a good-quality confirmation of successful genomic perturbation of at least one guide per clone. Subsequently, qPCR was employed to quantify the reduction in target gene expression, affirming the functional impact of the genetic modifications.

- RNA-seq quantification analysis: In my analyses, I investigated the expression levels of knockout targets (Fig. 3.2). Notably, the expression of *ARID1A* in hEM3

ARID1A^{KO} and *EP300* in Ark2 Cas9 *EP300*^{KO} was lower than in control. The expression of *ARID1A* in Ark2 Cas9 *ARID1A*^{KO}, *ARID5B* in Ark2 Cas9 *ARID5B*^{KO}, and *SETD1B* in Ark2 Cas9 *SETD1B*^{KO} was similar to that of each of the genes in the Ark2 Cas9 parental (control). This could mean that the mRNA production was unaffected, but the protein can still not be functional or not translated at all. The expression of *ARID1B* and *KMT2C* was increased, which could suggest compensatory mechanisms or incomplete knockouts, which prompted further genomic investigations.

- RNA-seq and ATAC-seq alignment analysis: I investigated the alignment of the reads from sequencing to genomic sequence to investigate the increased expression of respective genes in Ark2 ARID1B and KMT2C clones. I observed high intronic alignment in *ARID1B* (Fig. 3.3) and some intronic alignment around exon six and missing alignments in exon 5 in *KMT2C* (Fig. 3.4). Intronic alignment suggests that transcripts of a gene retained introns, which can result in the transcripts being targeted for degradation by nonsense-mediated mRNA decay (NMD). The increased levels of RNA might be a signal of increased transcription activated by a lack of protein. In the case of Ark2 *ARID1A* KO, the designed guides targeted second and third exons and resulted in 144 bp insertion in exon 2 and 1 bp insertion at the end of exon 3. As expected, there was no RNA-seq alignment to exon 2 in the Ark2 *ARID1A* KO - this does not mean that there was no expression of that exon, but more likely that the insertion was present at the RNA level and the reads did not align to the exon. This likely resulted in a dysfunctional protein or no protein product, but further protein assay verification is required.

- protein level confirmation: Recognizing that changes at the DNA and mRNA levels may not always reflect alterations in protein levels, we sought to validate the loss of target protein expression. For hEM3 *ARID1A*^{KO}, the loss of protein was confirmed through western blot analysis in the original publication [244]. The work to generate the western blot of the remaining cell lines is being carried out, facing some delays resulting from the unavailability of reliable antibodies.

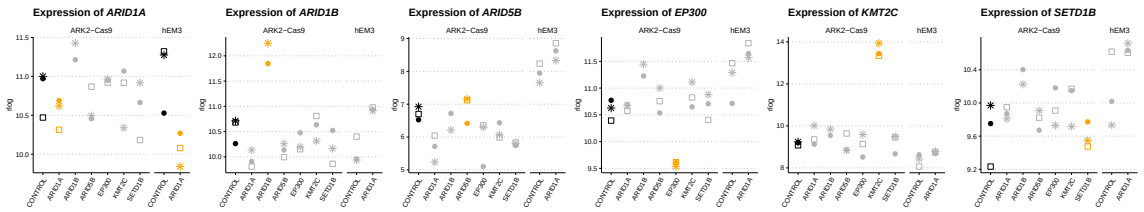


Figure 3.2: **Expression of targeted genes.** Each plot represents the expression of a given gene. Cell lines (CRISPR KO) are annotated on the x-axis. The gene targeted in a given cell line is highlighted in yellow, control in black, and other cell lines in gray.

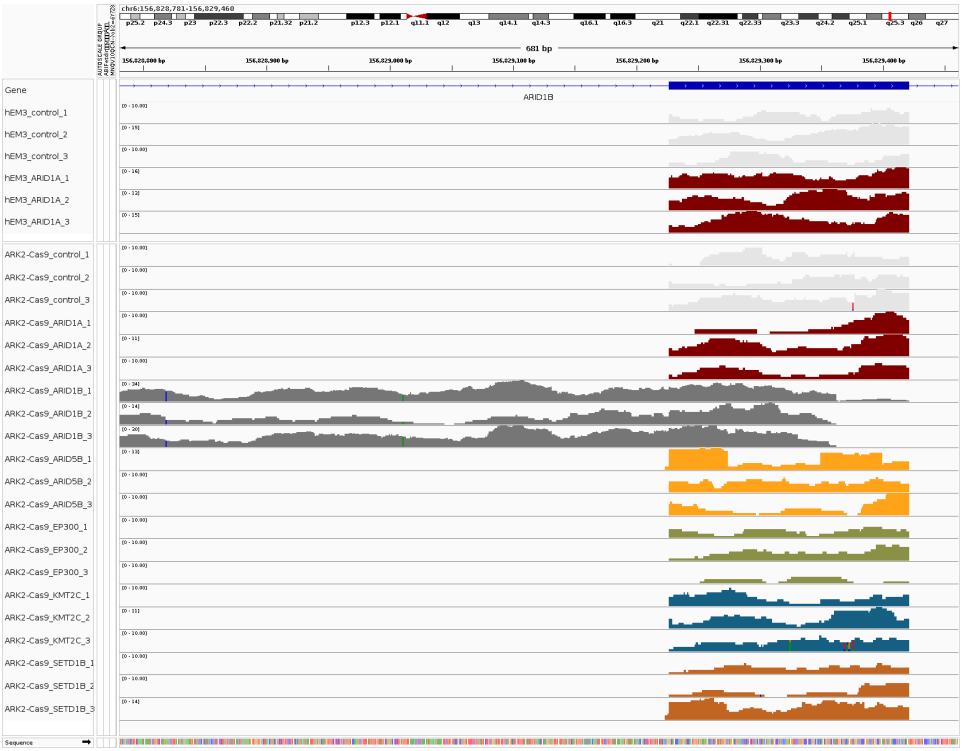
Moreover, the unsupervised analyses of expression and chromatin accessibility data, described below in Section 3.2.3, further reinforced our belief in functional knockouts. Principal component analysis (PCA) revealed distinct separation based on perturbation labels, indicating that genomic perturbations had significant effects at the cellular level, likely as a result of intended gene knockouts.

In conclusion, despite these challenges, we believe that the collective evidence from genomic sequence validation, expression analyses, unsupervised analyses, and preliminary protein-level investigations could suggest that functional KO of the targets was achieved. However, the results of the ongoing work confirming the protein loss in the cells are necessary to substantiate our findings.

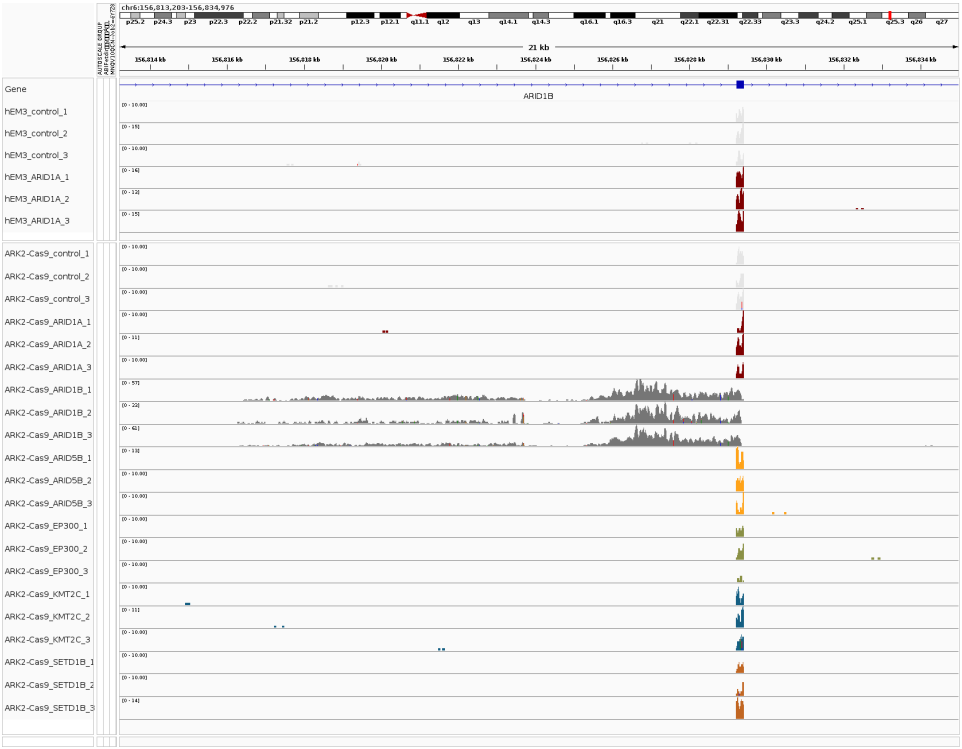
The CRISPR KOs used in the *BCOR* functional study were verified by qPCR and western blotting.

3.2.3 Cell line and genomic perturbations are dominant sources of variation

Fig. 3.5 indicates that the generated data was good quality. Principal component analysis (PCA) reveals that 88% of the variance could be attributed to the difference between the cell lines - hEM3 (cell line derived from a normal human endometrium) and Ark2 (cell line derived from malignant endometrial tissue). The separation between clones accounts for the most variance in either cell line - 46% and 48% in Ark2 Cas9 and hEM3, respectively. The median of 2,107 significantly differentially expressed genes (adjusted p-value < 0.05, and fold change, FC > 1.5 and FC < 0.67) was identified by DESeq2 (Fig. 3.6, range:

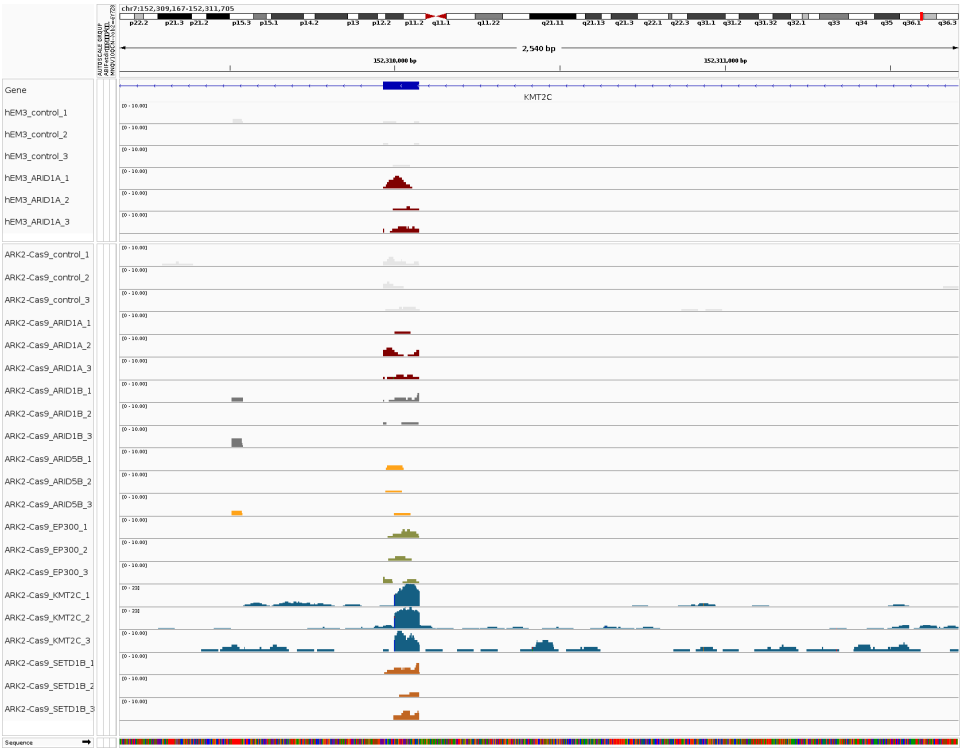


(a) 681 bp around exon 2 of *ARID1B*.

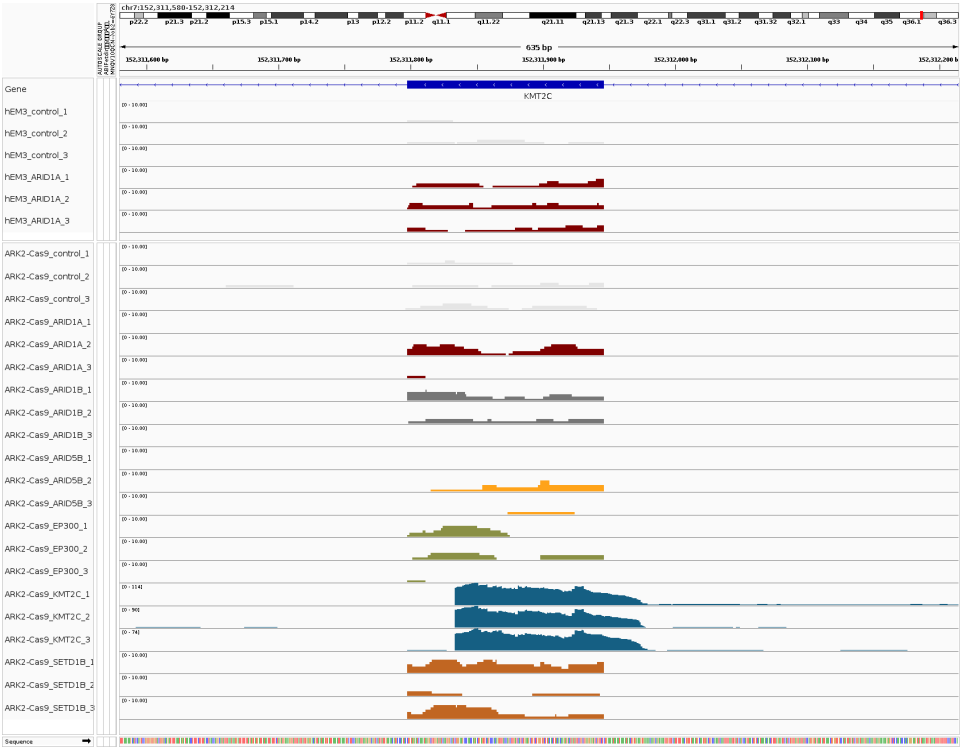


(b) 21,000 bp around exon 2 of *ARID1B*.

Figure 3.3: **Alignment of RNA-seq in the region of *ARID1B* transcript.** Ark2 control in rows 7-9 marked in light gray and Ark2 ARID1B clone in rows 13-15 marked in dark gray. Images extracted from IGV browser, hg38 used as a reference genome.



(a) 2,540 bp around exon 6 of *KMT2C*.



(b) 635 bp around exon 5 of *KMT2C*.

Figure 3.4: **Alignment of RNA-seq in the region of *KMT2C* transcript.** Ark2 control in rows 7-9 marked in light gray and Ark2 KMT2C clone in rows 22-24 marked in blue. Images extracted from IGV browser, hg38 used as a reference genome.

658−3,338). These are further discussed later.

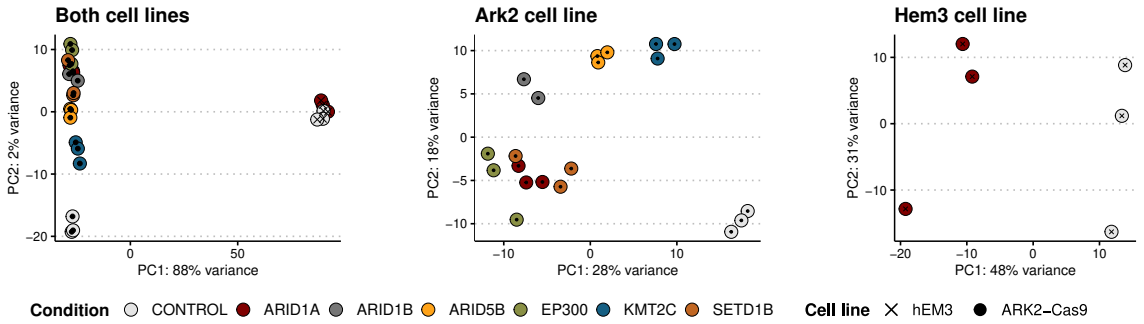


Figure 3.5: **Principal component analysis (PCA) of the RNA-seq data. Left**, the PCA across all conditions, **middle** the PCA of Ark2 cell line clones, and **right**, the PCA of the hEM3 cell line. One *ARID1B* replicate was excluded from analyses based on its distinctively different expression profile.

Like the RNA-seq data, generated ATAC-seq data was of good quality (Fig. 3.7). PCA on accessible chromatin regions showed that 74% of the variance could be attributed to the difference between the cell lines - hEM3 and Ark2 Cas9. The separation between clones accounts for the most variance in either cell line - 65% and 41% in hEM3 and Ark2 Cas9, respectively.

3.2.4 *ARID1A* loss disrupts chromatin accessibility and gene expression in endometrium and EC models

I began the analysis of this dataset by focusing on *ARID1A*, which, as discussed previously, is a member of the SWI/SNF chromatin-remodeling complex and a tumor suppressor in several cancer types, including endometrial cancer (Section 1.4.2.3). As detailed previously, over 55% of GEL EC samples harbor a mutation in *ARID1A*, and several more samples carry the copy number alteration at the genomic region spanning *ARID1A* (Section 2.3.2.1). *ARID1A* mutations in endometrial cancer have previously been associated with aggressive disease features, such as higher tumor grade, advanced stage, and poorer prognosis [13, 247]. Loss of *ARID1A* expression in endometrial cancer has also been associated with increased tumor invasiveness and resistance to chemotherapy [60]. Moreover, *ARID1A* mutations have been suggested to be an early event in the development of endometrial cancer, as they have been detected in precancerous lesions, such as endometrial

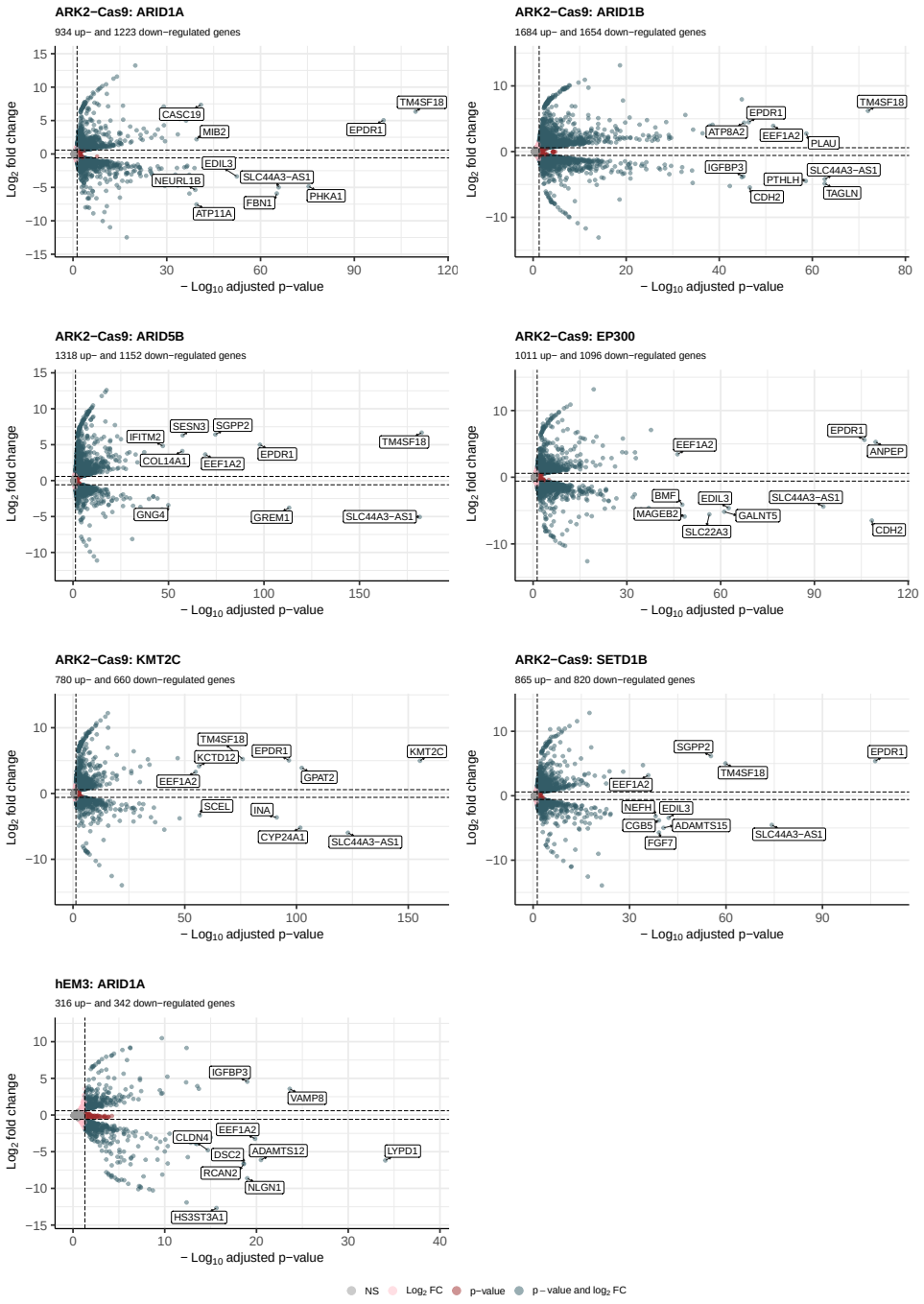


Figure 3.6: **Volcano plots of differentially expressed genes.** Top 10 genes per condition annotated. Genes passing the adjusted p-value threshold (<0.05) marked in red, fold change threshold (FC> 1.5) marked in pink, and those passing both thresholds in green.

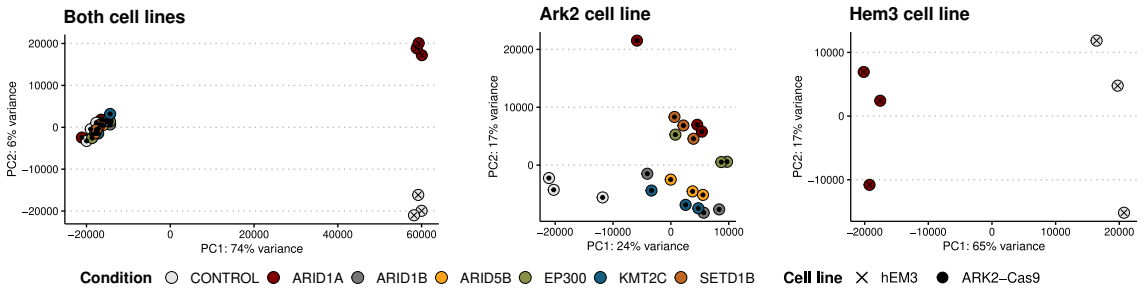


Figure 3.7: **Principal component analysis (PCA) of the ATAC-seq data. Left**, the PCA across all conditions, **middle** the PCA of Ark2 Cas9 cell line clones, and **right**, the PCA of the hEM3 cell line.

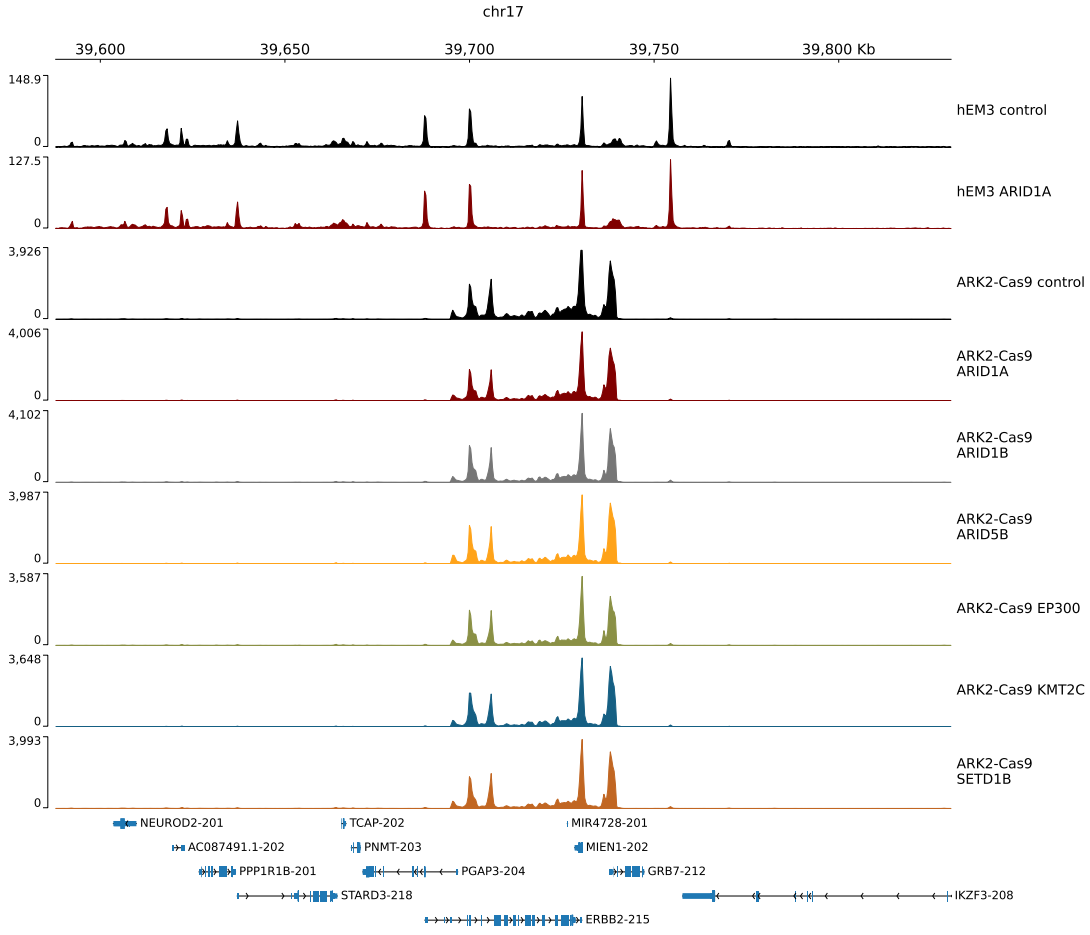


Figure 3.8: **Accessibility around *ERBB2* gene.** Regions around *ERBB2* gene are more than 20 times more accessible in EC-derived HER2 positive cell lines than in hEM3 cell lines derived from normal endometrium.

hyperplasia [247]. Therefore, *ARID1A* may serve as a potential diagnostic and prognostic marker for endometrial cancer.

3.2.4.1 Loss of *ARID1A* in normal endometrium

I used DESeq2 [248] to perform differential expression analysis of hEM3 $ARID1A^{KO}$ with respect to parental control cell line - hEM3 $ARID1A^{WT}$. This revealed 316 up-regulated genes and 342 down-regulated genes (Fig. 3.6; adjusted p-value < 0.05, FC > 1.5 and FC < 0.67). Worth of note up-regulated genes include *ERBB3*, *TGFB2* and *CREBBP*. *CREBBP* is one of the genes identified as chromatin driver and discussed in more detail in Chapter 2. Expression of a receptor tyrosine-protein kinase erbB-3 (*ERRBB3*), also identified as a driver in the GEL EC cohort, known as HER3 (human epidermal growth factor receptor 3), has been found to increase in recurrent and metastatic samples in comparison to the samples from the same patient [249]. Its binding partner, *ERBB4* (HER4), was down-regulated in hEM3 $ARID1A^{KO}$ cells.

In our study, gene set enrichment analysis (GSEA) revealed 63 significantly perturbed pathways with an adjusted p-value of less than 0.1 (Fig. 3.9A). Among these pathways, Transforming Growth Factor-β (TGF-β) signaling was notably up-regulated in the hEM3 $ARID1A^{KO}$ samples (Fig. 3.9B), and the *TGFB2* gene was found to be up-regulated as well. This observation contrasts with the findings reported in [245], wherein TGF-β signaling was up-regulated in mouse endometrium with *Arid1a* loss but *down*-regulated in hEM3 $ARID1A^{KO}$ samples.

The discrepancy between the two studies could be attributed to the differences in pathway definitions. A direct comparison of the pathway definition employed in our research and the custom pathway used by [245] was not feasible due to the latter's unavailability. Despite these differences, a comparison of the results from the differential expression analysis revealed that 201 genes were commonly dysregulated (with p-values < 0.05 and FC > 1.5 and FC < 0.67) between the two studies. Furthermore, a moderate correlation was observed between the log fold changes (Pearson's correlation $= 0.62$, p-value < 0.001). However, it is important to note that this comparison is limited to the 893 genes available

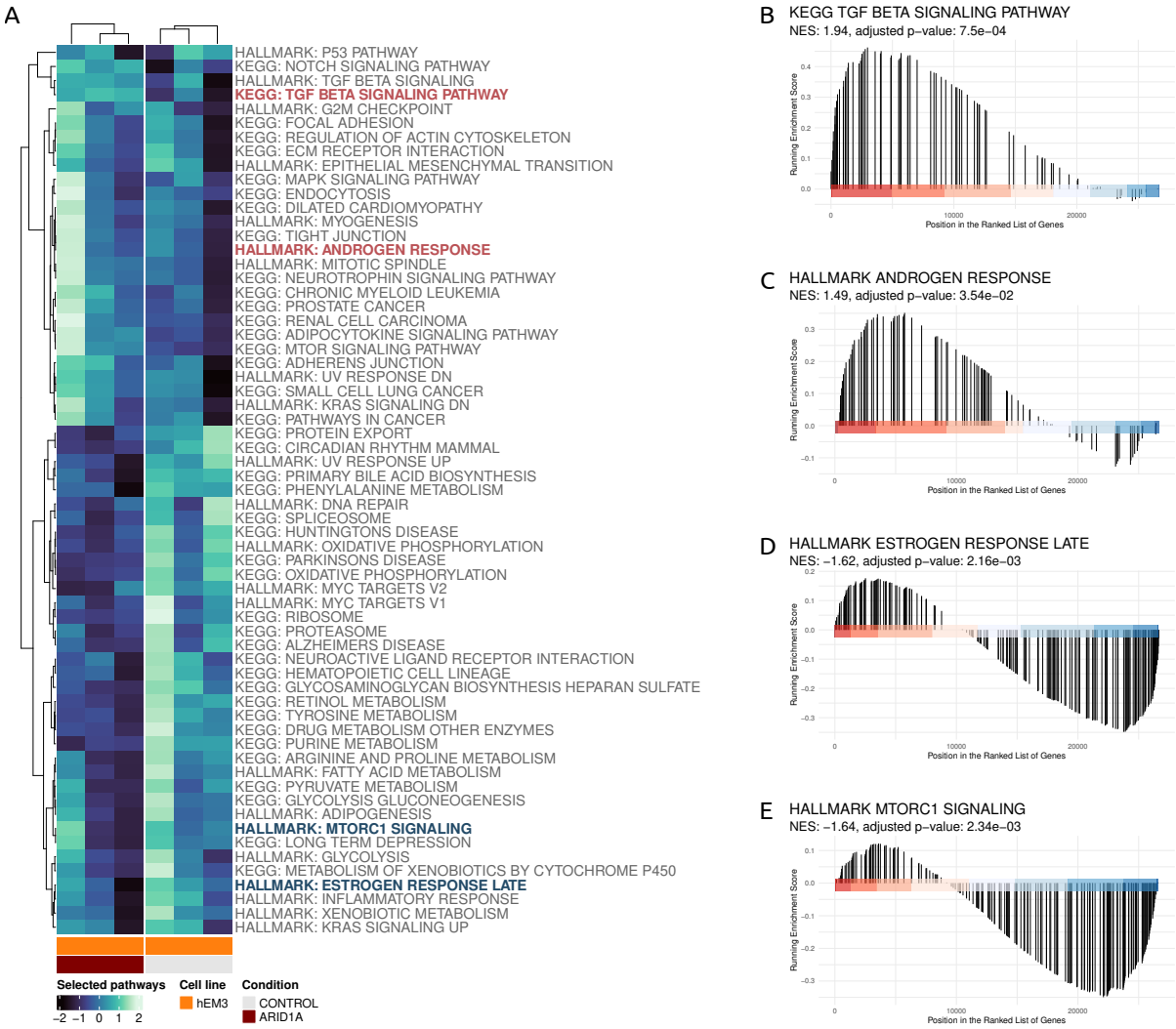


Figure 3.9: **Pathways disrupted in hEM3 *ARID1A*^{KO} cells. A** Heatmap of standardized average expression of significantly disrupted pathways (adjusted p-value < 0.1). Pathways for which GSEA plots are provided in the right panel are highlighted – up-regulated pathways marked in red and down-regulated ones in blue. For more details on clustering and visualization see Section 3.4.1.2. **B-E** Gene set enrichment plots for selected pathways. Each bar represents the gene in a pathway ordered based on differential expression score. The height of the bar plot signifies the running enrichment score at the gene's position in the list of all genes. Red to blue *heatband* shows gene rank's contribution to the enrichment score (red - enrichment in up-regulation, blue - down-regulation). NES - normalized across all permutations enrichment score, positive values signify up-regulation and negative values signify down-regulation.

in the supplementary file of [245].

TGF-β is a cytokine involved in various cellular processes such as cell growth, differentiation, and apoptosis [250]. The up-regulation of TGF-beta signaling in a cell line derived from normal endometrium with loss of *ARID1A* may suggest increased cell proliferation, tissue remodeling, or immune regulation. Additionally, it could be involved in the pathogenesis of endometriosis or endometrial hyperplasia. Several studies have reported increased levels of TGF-β in patients with endometriosis compared to controls [251–253]. Moreover, genes associated with the mitotic spindle were also found to be up-regulated, potentially supporting the increased proliferation hypothesis.

Another notable effect of *ARID1A* loss identified by GSEA was the up-regulation of androgen response (Fig. 3.9C). Androgens are synthesized in the ovary and adrenal glands and are involved in fertility regulation [254]. An imbalance of hormones, androgen specifically, is linked to disorders such as polycystic ovary syndrome (PCOS) [255] and endometriosis [256]. The up-regulation of androgen response in hEM3 $ARID1A^{KO}$ cells may indicate a mechanism contributing to endometrial hyperplasia. At the same time, the up-regulation of the androgen response was accompanied by the down-regulation of the estrogen response (Fig. 3.9D). Estrogen is a female sex hormone that plays a crucial role in regulating the menstrual cycle and maintaining the endometrial lining. *ARID1A* loss in breast cancer has previously been shown to reduce ER chromatin occupancy and target gene expression in breast cancer [257] – this result raised the possibility of similar alterations in endometrial cancer, a similar hormone-responsive malignancy. In combination, the down-regulation of estrogen response and up-regulation of androgen response strongly suggests that *ARID1A* loss leads to altered hormone sensitivity.

Other gene sets with known oncogenic activity were dysregulated in hEM3 $ARID1A^{KO}$. One example is mTORC1 signaling, down-regulated in hEM3 $ARID1A^{KO}$ (Fig. 3.9E). mTORC1, a vital player in cancer, promotes cell growth, proliferation, and survival by acting as a central hub that integrates signals from diverse pathways, including nutrient availability and growth factors. Dysregulation of mTORC1 can drive tumor development and progression through unchecked cell growth [258, 259]. Despite being hyperactivated

in a substantial number of human cancers and thus a promising target for cancer therapy, mTORC1 inhibitors have demonstrated limited success in preclinical studies. This may be attributed to the hypothesis that mTORC1 inhibition inadvertently lifts negative feedback on insulin/PI3K/Akt signaling, potentially promoting cell survival paradoxically [259].

I next explored the effects of *ARID1A* loss on chromatin accessibility using the ATAC-seq data. This allowed me to identify the perturbation of the accessibility of transcription factor binding motifs, as measured by the deviation from the expected background accessibility (Fig. 3.10). The analysis of differentially accessible motifs using chromVAR [260] revealed 69 distinct motifs with altered accessibility in hEM3 $ARID1A^{KO}$. I performed unsupervised clustering and split the motifs into more accessible and less accessible in hEM3 $ARID1A^{KO}$.

Cluster 1 comprised 55 repressed motifs belonging to Fos-related, Jun-related, and TALE-type homeodomain factors. Activator protein-1 (AP1) is a dimeric transcription factor formed through the heterodimerization of transcription factors from the Fos and Jun families. While overexpression of c-Fos and c-Jun has been demonstrated to promote cellular transformation and tumorigenesis in vitro and in vivo [261, 262], JunB and JunD have been associated with an anti-oncogenic activity. JunB expression was found to be silenced in myeloid lineage cells [263]. The role of JunD is more enigmatic, with some evidence suggesting it opposes transformation induced by Ras in mice [264].

Cluster 2 is characterized by transcription factor binding sites more accessible in hEM3 $ARID1A^{KO}$ cells. It consists of 14 activated motifs from the three-zinc finger Kruppel-related (transcription factors encoded by *KLF1*, *KLF2* and *KLF4*) and the more than three adjacent zinc fingers (*ZNF281*, *ZNF282* and *ZKSCAN1*) families.

Interestingly, motifs found more accessible in hEM3 $ARID1A^{KO}$ cells are found more often in promoter regions than those less accessible (Fig. 3.11). Motifs of KLF2, KLF1, ZNF740, ZNF281, and KLF4 were identified in proximity to the transcription start site (TSS) more often than other transcription factors (promoter region defined as a $< 3,000$ bp region around transcription start site, Fig. 3.12).

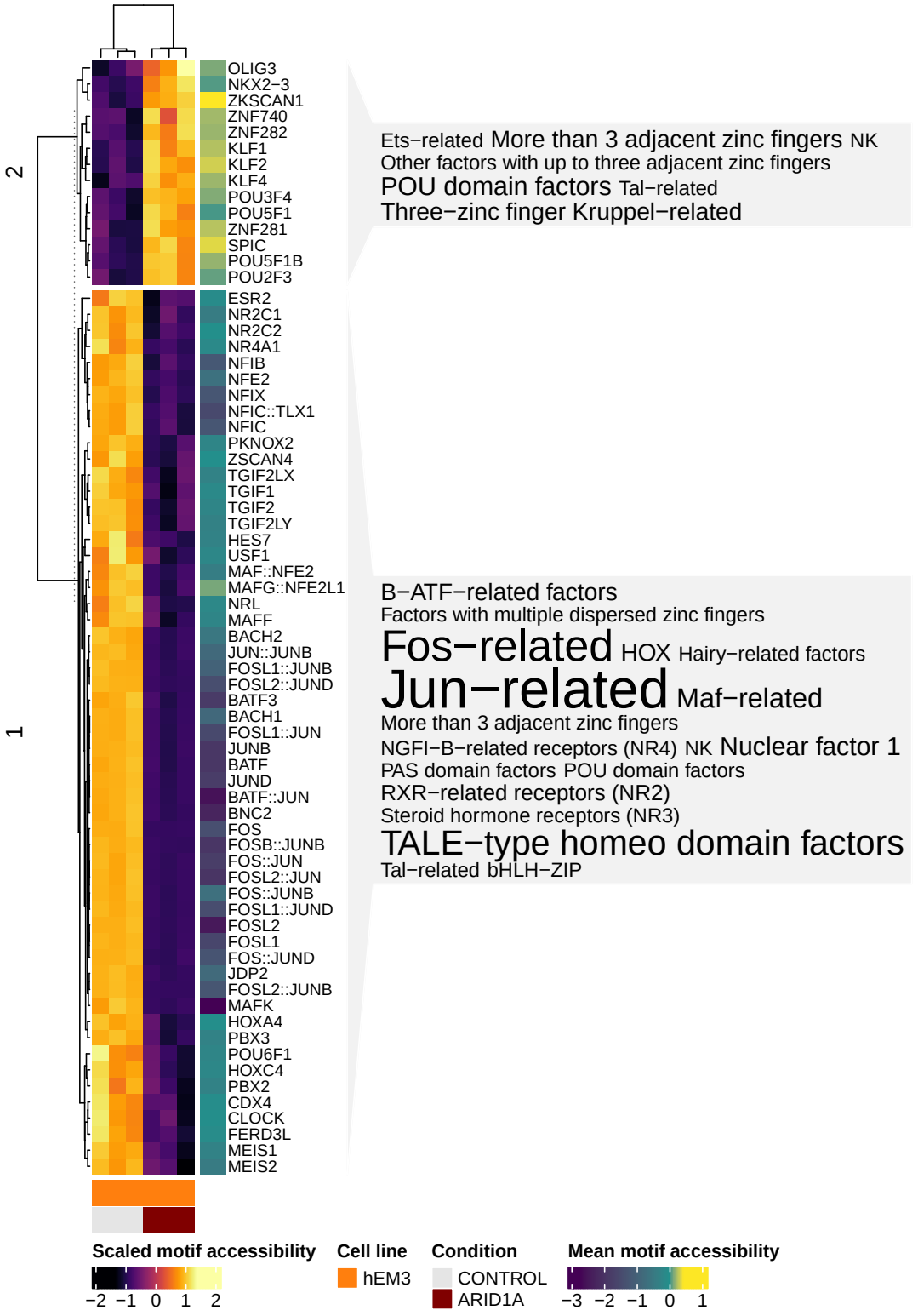


Figure 3.10: **Accessibility of motifs disrupted in hEM3 $ARID1A^{KO}$.** Heatmap of bias-corrected and normalized accessibility of motifs. Values in rows were scaled, and the row means were plotted aside to allow for the comparison of differences and the overall accessibility of motifs. The clusters were annotated with the names of motifs' families, where the font size is proportional to the number of motifs from each family present in the cluster. For more details on clustering and visualization, see Section 3.4.1.2.

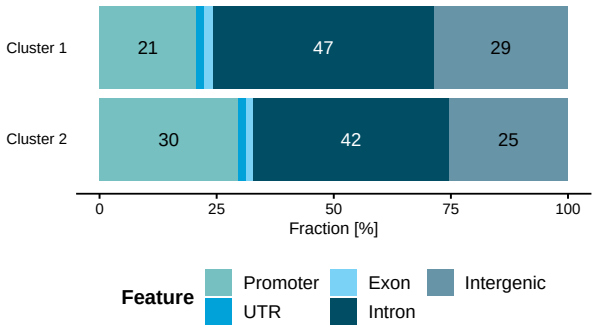


Figure 3.11: **Annotations of motifs accessibiity clusters in hEM3 *ARID1A*^{KO}.** Mean fraction of annotations assigned to the location of motifs found inaccessible sites in hEM3 *ARID1A*^{KO} cells. For per motif fractions, see Fig. 3.12.

3.2.4.2 Loss of *ARID1A* in EC derived cell line

I investigated the consequences of *ARID1A* loss in Ark2 Cas9 *ARID1A*^{KO} using CRISPR-Cas9 gene knockout technology (described in Section 3.4.1.1). Gene expression and chromatin accessibility data were generated by RNA-seq and ATAC-seq. Differential expression analysis revealed 934 up-regulated and 1,223 down-regulated genes in Ark2 Cas9 *ARID1A*^{KO} compared to Ark2 Cas9 parental cell line (FC > 1.5 and FC < 0.67, adjusted p-value < 0.05).

GSEA on Hallmark pathways in Ark2 Cas9 *ARID1A*^{KO} cells identified 15 up-regulated and 13 down-regulated pathways in knockout compared to Ark2 Cas9 parental cells (adjusted p-value < 0.05, Fig. 3.14). Up-regulated pathways included Myc targets and KRAS signaling, and down-regulated pathways included epithelial-mesenchymal transition (EMT), discussed in more detail in Section 3.2.4.3.

Differential motif accessibility results revealed three motifs more accessible in Ark2 Cas9 *ARID1A*^{KO} than in Ark2 Cas9 parental cell line (Fig. 3.13). CUX1 and CUX2 are members of a family of homeodomain transcription factors containing Cut repeat DNA-binding sequences. CUX2 was identified as oncogenes in thyroid cancer [265] and tumor suppression in glioma [266]. Similarly, CUX1 was identified as a tumor suppressor in myeloid malignancies [267] and an oncogene in prostate cancer [268]. Their role in endometrial cancer has not been described as yet. However, there is evidence that CUX1 regulates PIK3CA and promotes cell growth and metastasis via activating epithelial-

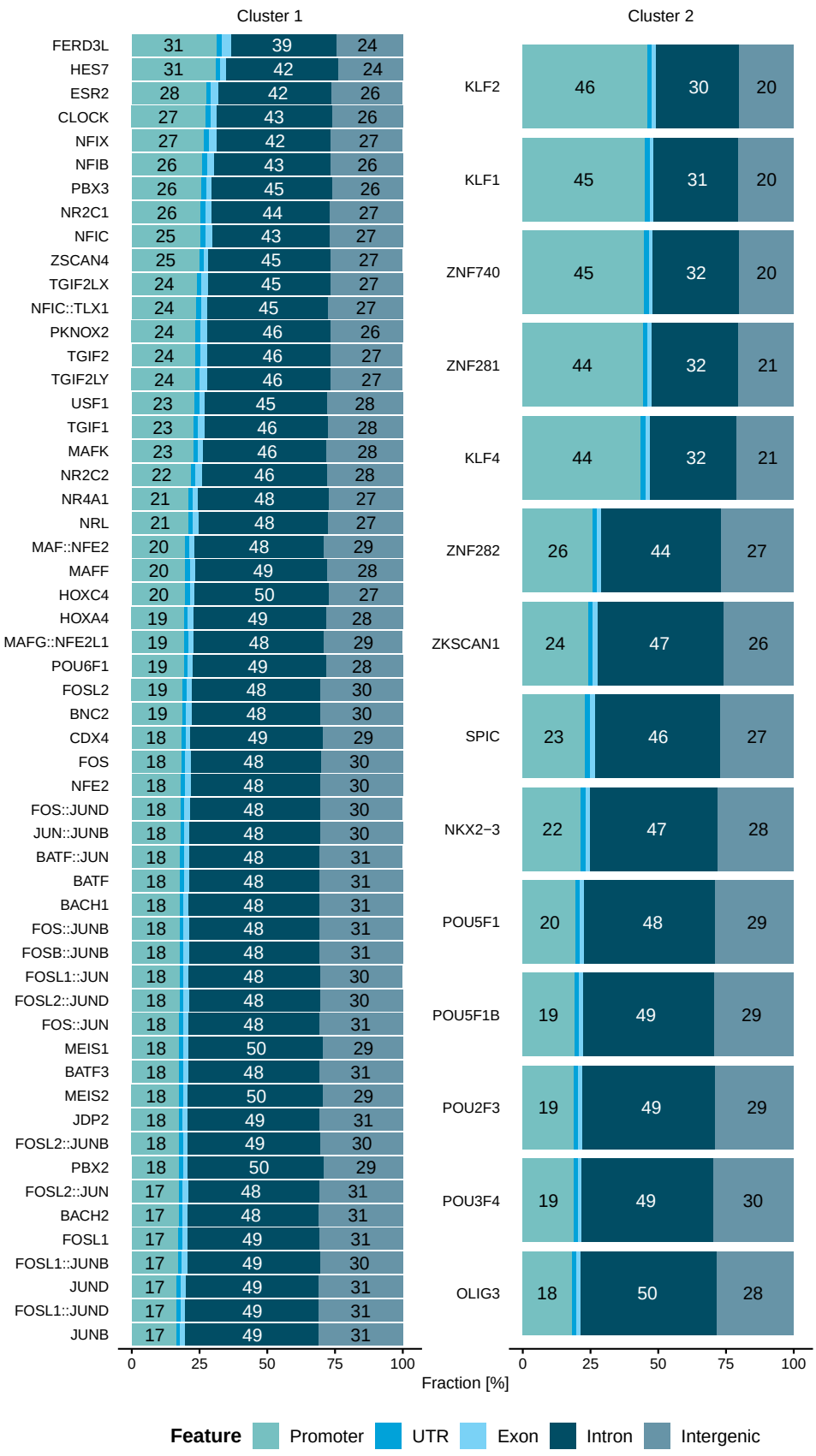


Figure 3.12: **Annotations of the motif sites in hEM3 *ARID1A*^{KO}.** Fraction of annotations assigned to the location of motifs found in accessible sites in the hEM3 *ARID1A* clone.

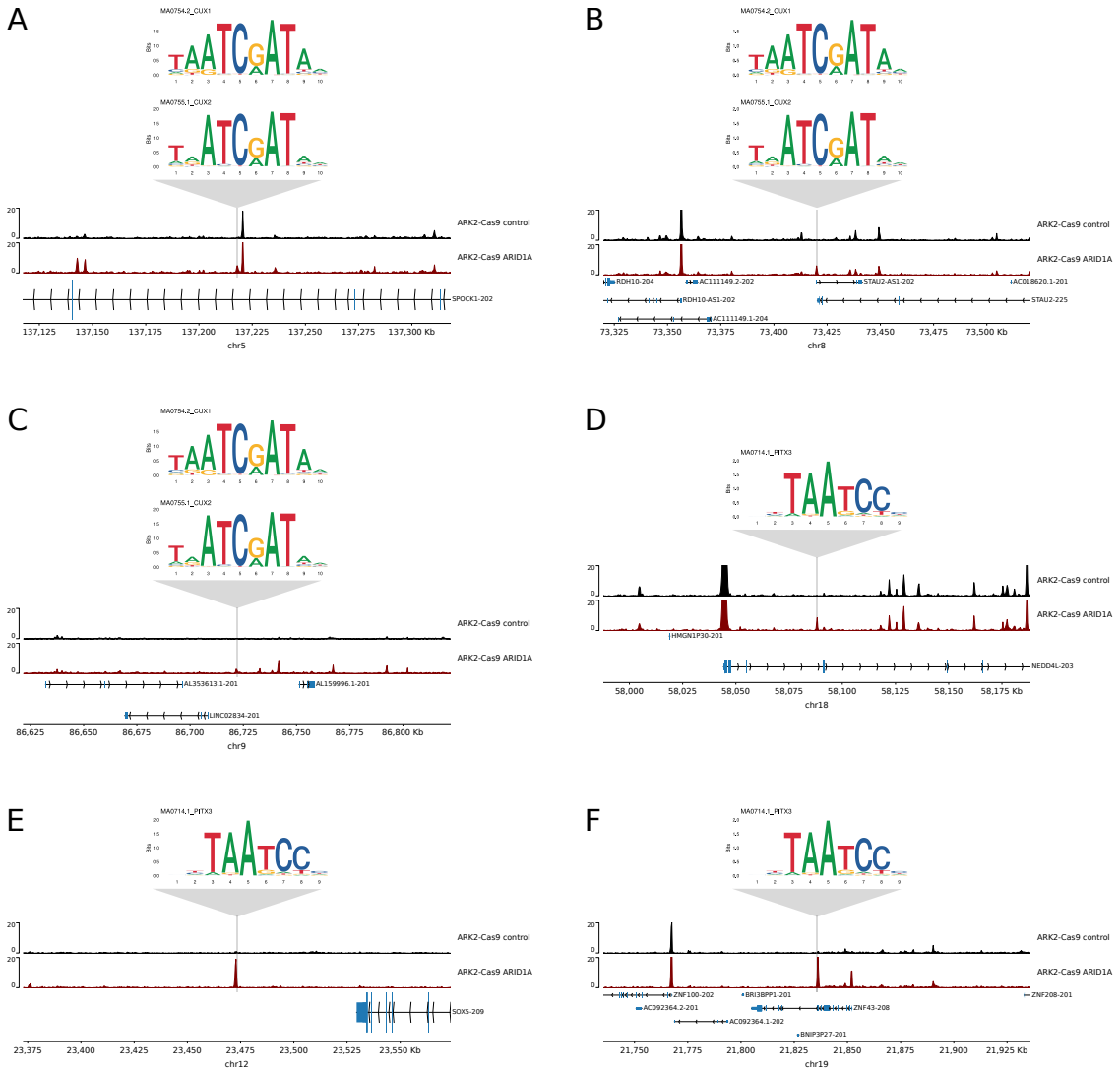


Figure 3.13: **Accessibility of regions containing the PITX3, CUX1 and CUX2 motifs.** PITX3 and CUXs motifs are found more accessible in the Ark2 Cas9 $ARID1A^{KO}$.



Figure 3.14: **Dotplot of pathways perturbed in Ark2 Cas9 *ARID1A*^{KO}.** Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the number of genes in a pathway. Key pathways discussed in the text are emphasized. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis.

mesenchymal transition (EMT) in bladder cancer [269]. PITX3 is a transcription factor of the paired like homeodomain family which was reported to regulate the development of multiple organs. *PITX3* expression was found related to tumor grade and poor prognosis in head and neck squamous cell carcinoma [270].

Regions with the CUX1 and CUX2 motifs are located in the *SPOCK1* intronic region (Fig. 3.13A), exonic region of *STAU2* (Fig. 3.13B) and intergenic regions around long non-coding RNA (lncRNA) - LINC02834 (Fig. 3.13C). Interestingly, *SPOCK1* is up-regulated in Ark2 Cas9 $ARID1A^{KO}$ ($\log_2$(FC) = 8.04, adjusted p-value = 2.21×10^{-5}). Enhancers are often found to control the expression of the genes in which intronic regions they are found [271, 272]. This can suggest that *SPOCK1* expression is regulated by CUXs transcription factor and that loss of *ARID1A* results in exposure of its motif. *SPOCK1* was found to promote EMT and invasion in gastric cancer [273] and pancreatic cancer [274]. Additionally, lncRNAs are also known to be involved in regulating cancer stem cell phenotype and EMT programs [275].

Overall, out of 78 and 83 genes up-regulated in Ark2 Cas9 $ARID1A^{KO}$ cells under the control of CUX1 and CUX2, respectively I found members of TNF-α signaling via NF-κB (*ABCA1*, *CFLAR*, *CXCL6*, *IL1A* and *SQSTM1*) and p53 pathway (*ABCC5*, *CD82*, *IL1A*, *PRKAB1* and *SLC3A2*).

PITX3 motifs are found in intergenic regions of *NEDD4L* (Fig. 3.13D) and *ZNF43* (Fig. 3.13F), as well as intergenic region around *SOX5* (Fig. 3.13E) genes. *ZNF43* and *NEDD4L* are up-regulated in Ark2 Cas9 $ARID1A^{KO}$ compared to Ark2 Cas9 parental cells ($\log_2$(FC) = 10.42, adjusted p-value = 2.5×10^{-11}, $\log_2$(FC) = 0.55, adjusted p-value = 4.95×10^{-3}, respectively).

3.2.4.3 Loss of *ARID1A* in normal endometrium and EC

Next, I investigated the commonalities and differences between the perturbation in hEM3 $ARID1A^{KO}$ and Ark2 Cas9 $ARID1A^{KO}$ cell lines (Fig. 3.15). Genes with significant perturbations can be split into four clusters (hierarchical clustering based on Pearson correlation distance). For each cluster, I performed gene set enrichment analysis (Enrichr

[276], adjusted p-value < 0.05).

Cluster 1, with expression higher in Ark2 cell lines and up-regulation in Ark2 Cas9 $ARID1A^{KO}$, is enriched for up-regulation of KRAS signaling. Cluster 2, with overall expression lower in hEM3 cell lines and down-regulation in Ark2 Cas9 $ARID1A^{KO}$, is enriched for epithelial-mesenchymal transition (EMT). EMT is also enriched in clusters 4 and 5. In cluster 4, expression is higher in hEM3 $ARID1A^{WT}$ than in other cells, and in cluster 5, overall expression is higher in the hEM3 cell line, and genes are down-regulated in the Ark2 Cas9 $ARID1A^{KO}$.

Cluster 3 is enriched for down-regulation of UV response - the expression of genes in that cluster is lower in Ark2 Cas9 $ARID1A^{KO}$ than in Ark2 Cas9 parental cells, while it is similar in control and hEM3 $ARID1A^{KO}$ cells. Cluster 6, up-regulated in hEM3 ARID1A KO, was not found to be significantly enriched for any gene set tested. *POLE* and *CREBBP*, whose expression is up-regulated in hEM3 $ARID1A^{KO}$, are among the 160 genes in that cluster. Interestingly, CBP, encoded by *CREBBP*, is a key acetyltransferase and a transcriptional cofactor. CBP, together with p300 (discussed in more detail later in this Chapter), regulates gene expression by modulating acetylation levels of histone proteins. Increased *POLE* levels could potentially suggest increased replication levels as polymerase-ϵ is crucial for replication and DNA proofreading.

Cluster 7, enriched for hypoxia and apoptosis-related genes, but also for TNF-alpha signaling via NF-κB and mTORC1 signaling, is characterized by higher expression in Ark2 Cas9 $ARID1A^{KO}$, hEM3 $ARID1A^{KO}$ and hEM3 $ARID1A^{WT}$ cells with genes are up-regulated in the Ark2 Cas9 $ARID1A^{KO}$ clone but down-regulated in the hEM3 $ARID1A^{KO}$ clone with respect to their respective parental cells.

Of 658 and 2,157 genes differentially expressed in clones with Ark2 Cas9 $ARID1A^{KO}$ and hEM3 $ARID1A^{KO}$, respectively, 61 genes are commonly up- and down-regulated across normal and malignant models. The genes can be split into four clusters to allow for data-driven interpretation, as shown in Fig. 3.16 (dendrogram was split upon visual inspection). Clusters recapitulated the up- and down-regulation and differences in baseline expression in each cell line.

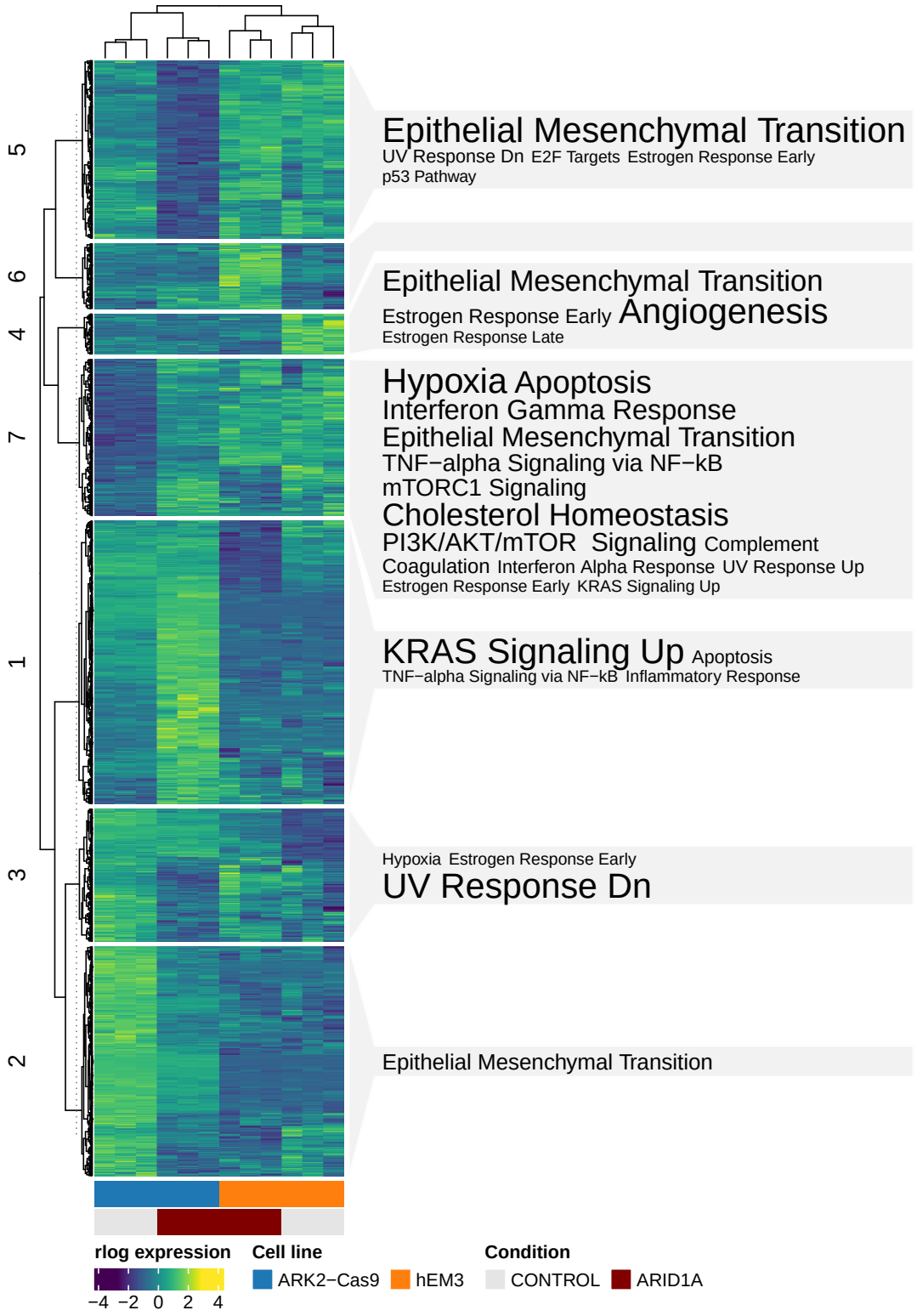


Figure 3.15: **Heatmap of the dysregulated expression across normal endometrium and EC-derived *ARID1A* knockout cell lines.** A union of genes up- (FC $>$ 1.5, p-value $<$ 0.05) and down-regulated (FC $<$ 0.67, p-value $<$ 0.05) in either of the clones shown. Clusters annotated with enriched pathways identified with Enrichr [276] (adj. p-value $<$ 0.05). The font size corresponds to the combined score - a combination of p-value and rank score as defined in [276]. For more details on enrichment analysis see Section 3.4.1.2 and for clustering and visualization, see Section 3.4.1.2.

Cluster 1, up-regulated in the event of *ARID1A* loss and with higher expression in Ark2 than hEM3, contains genes such as *ID4* and *FLG*. The role of those genes is not well described in endometrial tissue. However, expression of *ID4* is increased in colon cancer stem cells, and further knockdown of *ID4* inhibited the growth and dissemination of colon cancer cells [277]. Mutations in *FLG* have been associated with the progression of prostate cancer [278] and stomach adenocarcinomas [279].

Expression of *ADAMTS12*, member of cluster 2 - down-regulated in *ARID1A* clones, has been linked with colorectal cancer. However, the evidence seems conflicting - patients with low or *ADAMTS12* expression had a significantly poor overall survival [280], but it was also described to act as a cancer promoter [281]. Further investigation could better explain its link to *ARID1A* and the consequence of its down-regulation in endometrial cancer.

Other genes of note include *GREM1* - a member of cluster 4, which is down-regulated in *ARID1A* clones. *GREM1* overexpression was linked to inhibition of proliferation and migration in osteosarcoma [282]. Gremlin 2, a member of its family, was found to be repressed in invasive endometrial cancer [283].

Cluster 3 with higher expression in the event of *ARID1A* loss includes genes, such as *PRKCE* or *EPAS1*, whose role in endometrial cancer is currently poorly understood. However, *PRKCE* has been linked to poor prognosis in kidney renal clear cell carcinoma [284]. *EPAS1* has been found to promote non-small-call lung cancer by enhancing EMT [285].

3.2.5 Loss of *ARID1B* reveals perturbed gene expression profiles

As discussed in detail in Section 1.4.2.3, *ARID1B* is an *ARID1A* paralog and encodes a key component of the SWI/SNF complex. *ARID1B* is mutated in 11.72% of EC samples in the GEL cohort, and its loss is frequent across undifferentiated upper gynecological tract carcinomas [286]. Previous work in colorectal cancer has suggested that, similarly to ARID1A, it regulates accessibility at enhancer sites, although to a lower extent than its paralog [189]. ARID1A and ARID1B have been associated with AP1 binding sites at

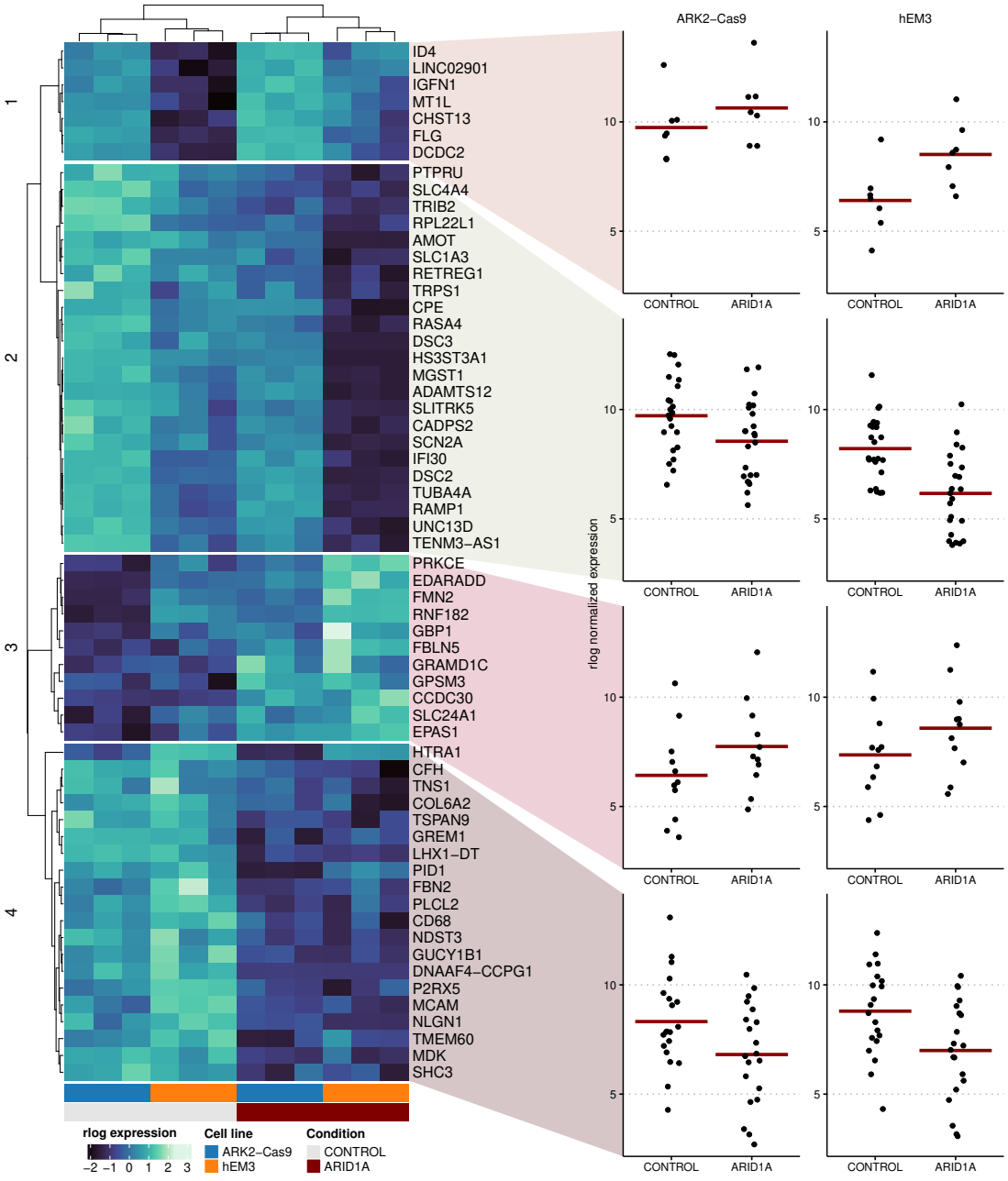


Figure 3.16: **Heatmap of expression of genes commonly disrupted across normal endometrium and EC-derived *ARID1A* knockout cell lines.** Genes commonly dysregulated in the *ARID1A* KO as defined by the differential expression analysis between a KO and respective control cell line.

enhancers in colorectal and ovarian cell lines [189]. Concurrent ARID1A and ARID1B inactivation, observed via immunohistochemistry in endometrial and ovarian dedifferentiated carcinomas, has been described [287]. However, functional analysis of *ARID1B* in endometrial cancer has not been performed previously.

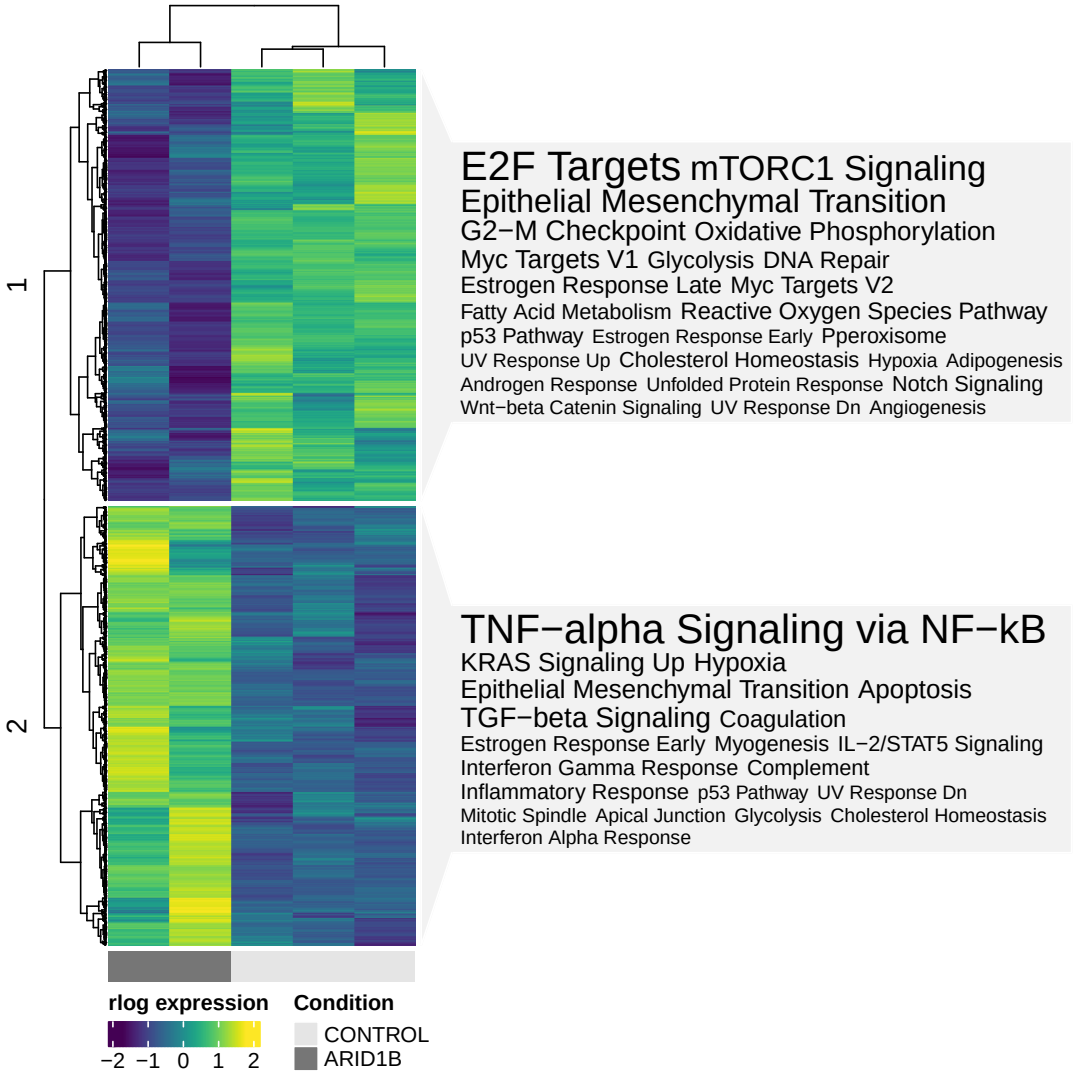


Figure 3.17: **Heatmap of differentially expressed genes in Ark2 Cas9 *ARID1B*^{KO}.** Up- and down-regulated clusters annotated with enriched pathways identified with Enrichr [276] (adj. p-value < 0.05). Standardized rlog expression plotted. The font size corresponds to the combined score - a combination of p-value and rank score as defined in [276]. For more details on enrichment analysis see Section 3.4.1.2 and for clustering and visualization, see Section 3.4.1.2.

I investigated the consequences of ARID1B loss in Ark2 Cas9 *ARID1B*^{KO} using similar CRISPR-Cas9 gene knockout technology to that employed for generating the loss of ARID1A (as described in Section 3.4.1.1). Gene expression and chromatin accessi-

bility data were generated by RNA-seq and ATAC-seq. Differential expression analysis revealed 1,623 up-regulated and 1,608 down-regulated genes (Fig. 3.17, FC $>$ 1.5 and FC $<$ 0.67, adjusted p-value $<$ 0.05). These displayed partial overlap with those differentially expressed following the loss of ARID1A, its paralogue in Ark2 cells (418 genes were commonly up-regulated, and 456 down-regulated genes were found down-regulated in both conditions). I discuss the comparison between all perturbations in more detail in Section 3.2.10. I next performed enrichment analysis to identify the dysregulated pathways in Ark2 Cas9 *ARID1B*^{KO} cells (Enrichr, adjusted p-value $<$ 0.05). Up-regulated genes were enriched for the tumor necrosis factor-alpha (TNF-α) signaling pathway through the activation of nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB). This suggests a potential role of the TNF-α/NF-κB signaling axis in response to *ARID1B* loss. Activation of NF-κB may promote the expression of anti-apoptotic and pro-proliferative genes, thereby supporting tumor cell survival and growth.

Interestingly, both down- and up-regulated genes were enriched for EMT (adjusted p-value of 2.35×10^{-10} and 3.32×10^{-7}, respectively), which suggests that *ARID1B* loss results in a more complex perturbation of the EMT pathway. For example, *SERPINE2*, a member of EMT pathway up-regulated in Ark2 Cas9 *ARID1B*^{KO} clones, was found to be induced by activated MAP kinase in intestinal epithelial cells [288] and implicated in transition between normal tissue and adenomas, as well as dysplasia and colorectal cancer [289].

GSEA on Ark2 Cas9 *ARID1B*^{KO} cells identified 20 up-regulated and 17 down-regulated pathways (Fig. 3.18, adjusted p-value $<$ 0.05). As expected from enrichment analysis, TNF-α/NF-κB signaling was identified as an up-regulated pathway (normalized enrichment score, NES = 2.20, adjusted p-value = 1×10^{-9}). Other up-regulated pathways include KRAS signaling up (NES = 2, adj. p-value = 1.4×10^{-7}), KRAS signaling up (NES = 1.43, adj. p-value = 1.1×10^{-2}), inflammation circuits (IL2 - NES = 1.6, adj. p-value = 7.2×10^{-4}, IL6 - NES = 1.64, adj. p-value = 0.004), and TGF-β signaling (Fig. 3.19, NES = 1.93, adj. p-value = 5.85×10^{-4}). Among down-regulated pathways are E2F targets (NES = -2.84, adj. p-value = 1×10^{-9}) and EMT (NES = -1.84, adj.

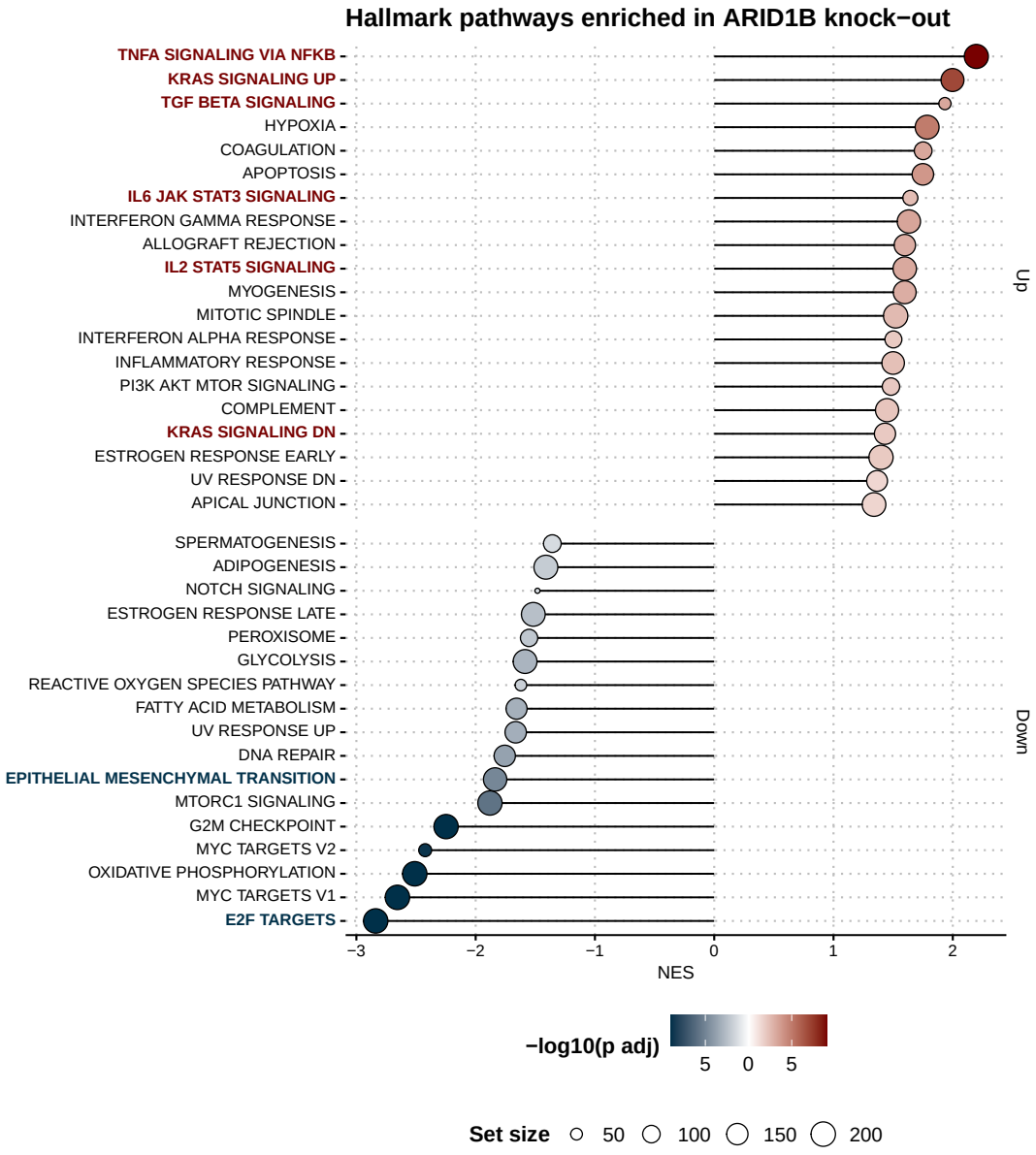


Figure 3.18: **Dotplot of enriched Hallmark pathways in Ark2 Cas9 *ARID1B^{KO}*.** Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the number of genes in a pathway. Key pathways discussed in the text are emphasized. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis.

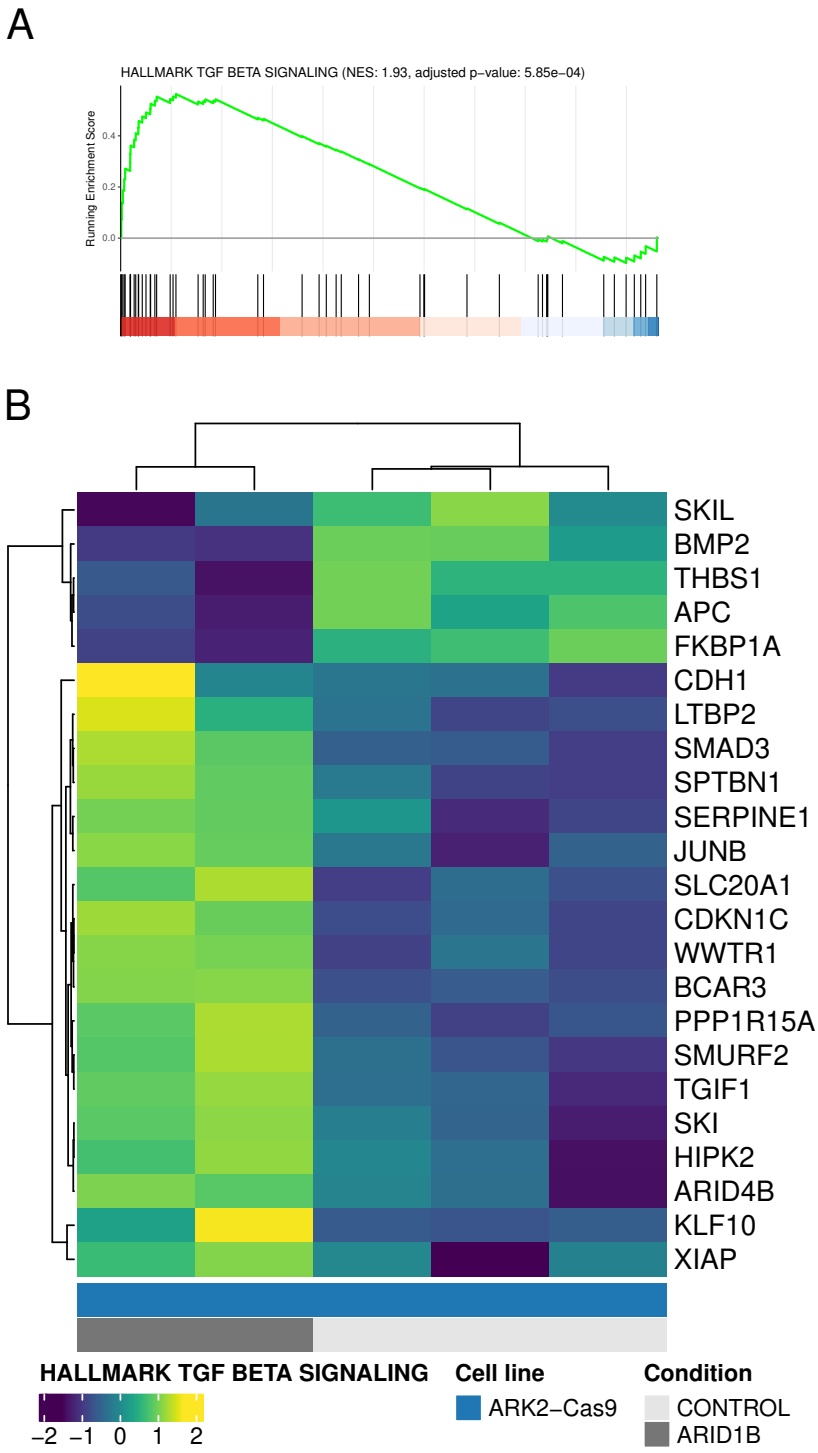


Figure 3.19: **TGF-β signaling in Ark2 Cas9 *ARID1B*^{KO}. A** GSEA plot for TGF-β signaling. Running enrichment score visualized by a green line, genes belonging to the pathway marked by black bars at the bottom, ordered by the differential expression results. **B** Heatmap of standardized rlog normalized expression of differentially expressed (FC $>$ 1.5 and FC $<$ 0.67, adj. p-value $<$ 0.05) genes belonging to TGF-β signaling pathway. For more details on enrichment analysis see Section 3.4.1.2 and for clustering and visualization, see Section 3.4.1.2.

p-value $= 1.2 \times 10^{-5}$). Those results agree with similar analyses of ARID1A/B knockouts performed in liver [190].

Interestingly, as shown in Fig. 3.19, even though the TGF-β pathway is up-regulated, the expression of its members varies. For example, genes such as *B2M* and *APC* are down-regulated, while *TGIF1*, *SMURF2* or *SERPINE1* are up-regulated. *SERPINE1* is a member of TGF-β signaling and similarly to already mentioned *SERPINE2*, its overexpression is linked to promoting malignant progression in gastric cancer [290] and is a known marker of endometriosis [291].

Differential chromatin accessibility analysis revealed 4,113 more accessible regions and 3,586 less accessible regions in Ark2 Cas9 $ARID1B^{KO}$ compared to Ark2 Cas9 parental cells. 71% of less accessible and 55% of more accessible regions are annotated in the intronic and intergenic regions (Fig. 3.20).

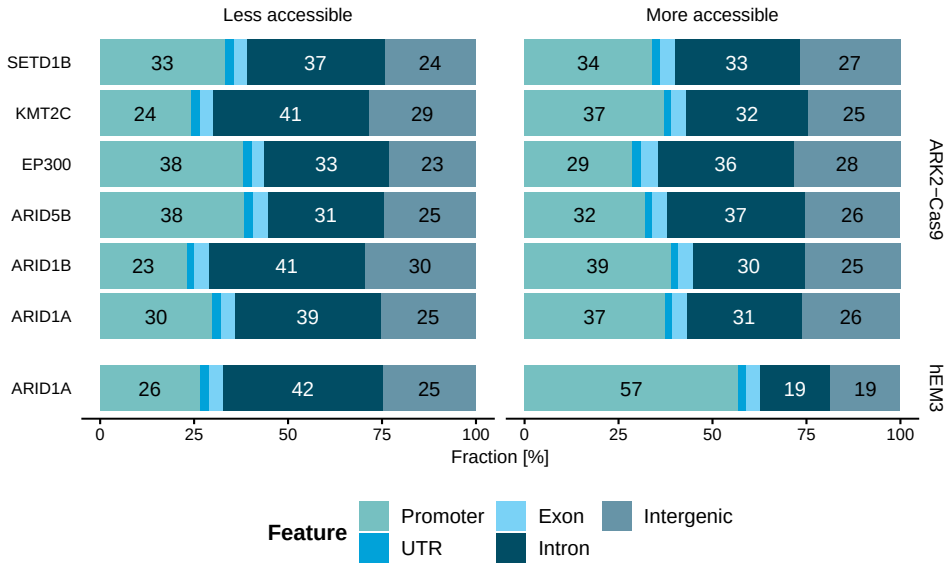


Figure 3.20: **Annotations of the differentially accessible regions.**

I annotated these regions with the closest gene, revealing 3,895 genes linked to more accessible and 3,327 genes linked to less accessible regions. When intersected with up- and down-regulated genes identified from RNA-seq analysis, I found 596 genes up-regulated and linked to more accessible regions and 395 genes down-regulated and linked to less accessible regions (Fig. 3.21B).

Next, I performed GSEA analysis on the accessibility regions linked to genes and com-

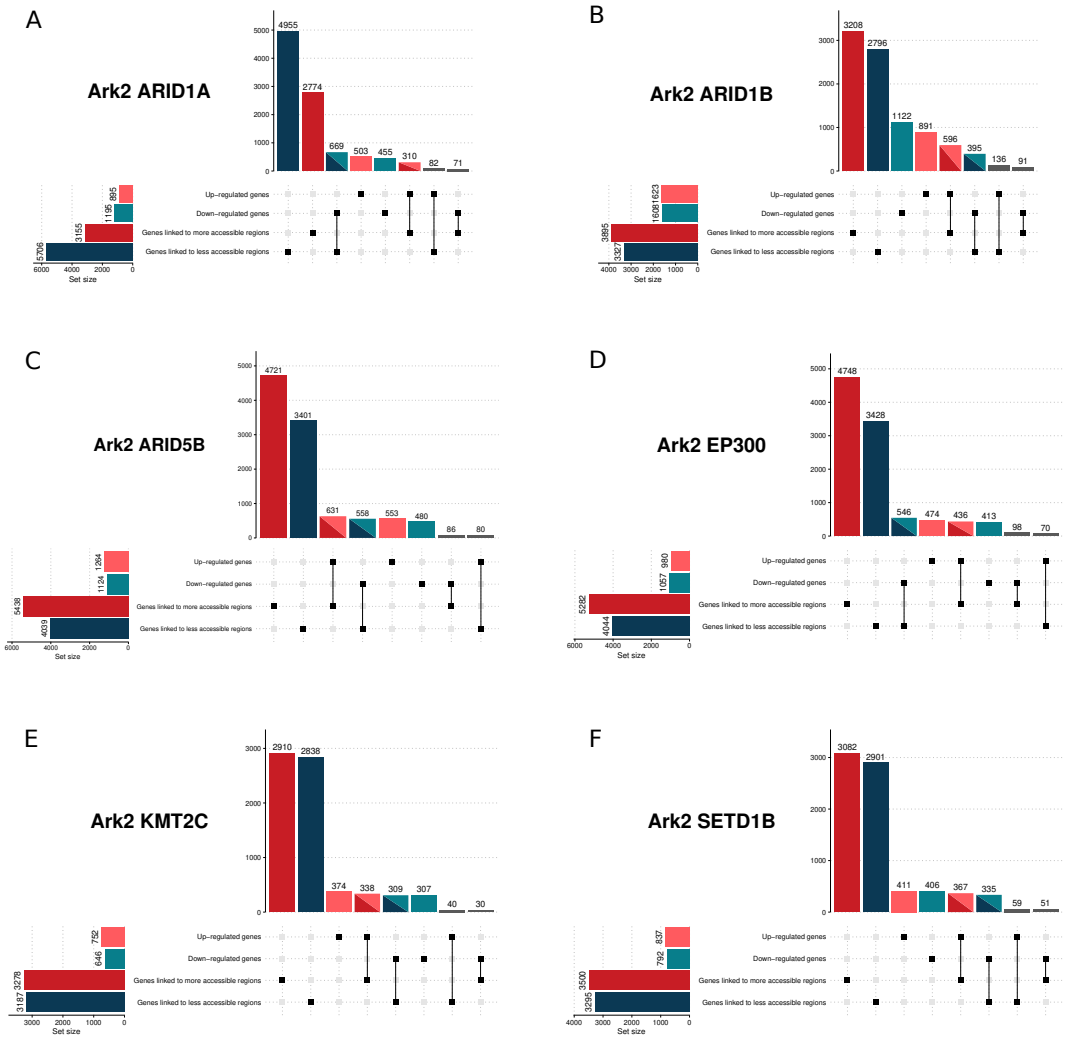


Figure 3.21: **Overlap of genes annotated to chromatin accessibility sites and dysregulated genes identified with RNA-seq analysis.** Up- and down-regulated genes identified from RNA-seq data (absolute $\log_2$ fold change $> \log_2(1.5)$, adjusted p-value < 0.05); genes linked to more and less accessible regions identified by annotating the differentially accessible (absolute $\log_2$ fold change $> \log_2(1.5)$, adjusted p-value < 0.05) regions with closest genes.

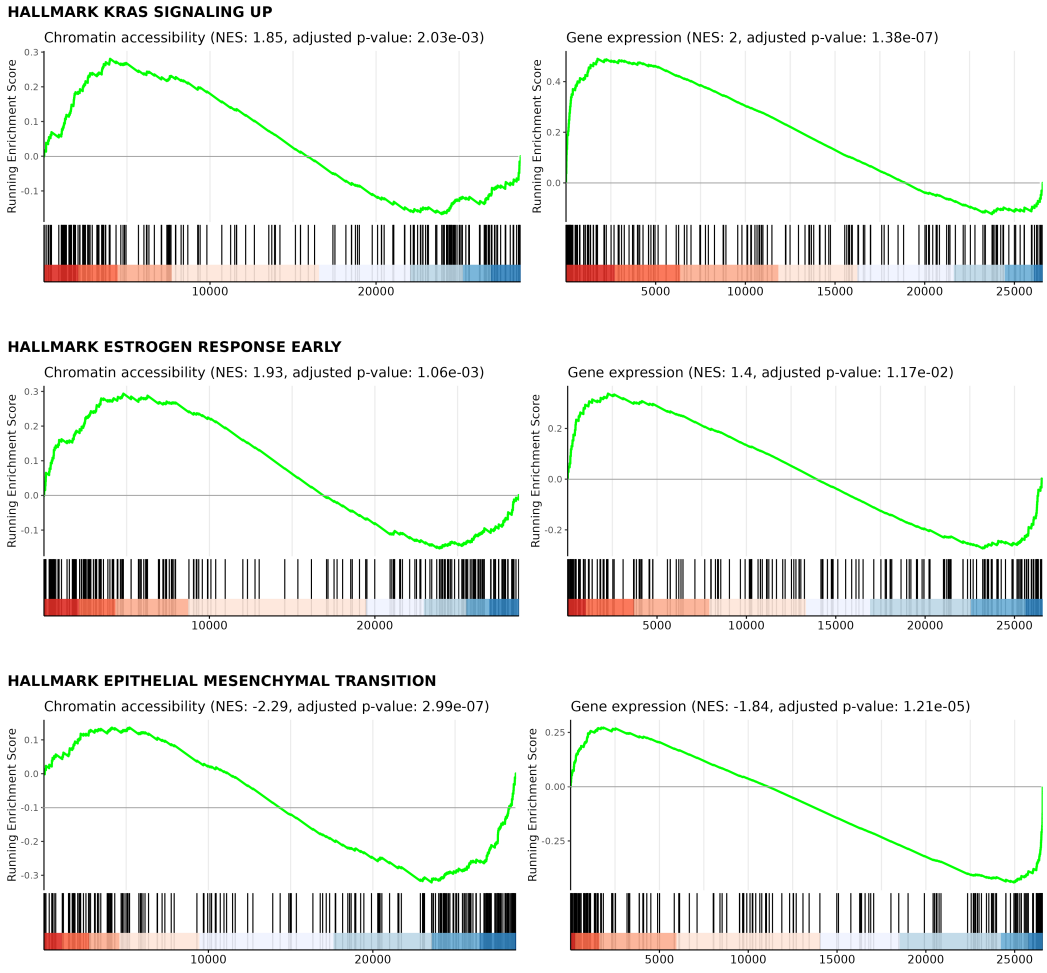


Figure 3.22: **Pathways commonly perturbed in Ark2 Cas9 *ARID1B*^{KO} at the chromatin accessibility (left) and gene expression (right) level.** GSEA plots for selected pathways. Running enrichment score visualized by a green line, genes belonging to the pathway marked by black bars at the bottom, ordered by the differential expression results.

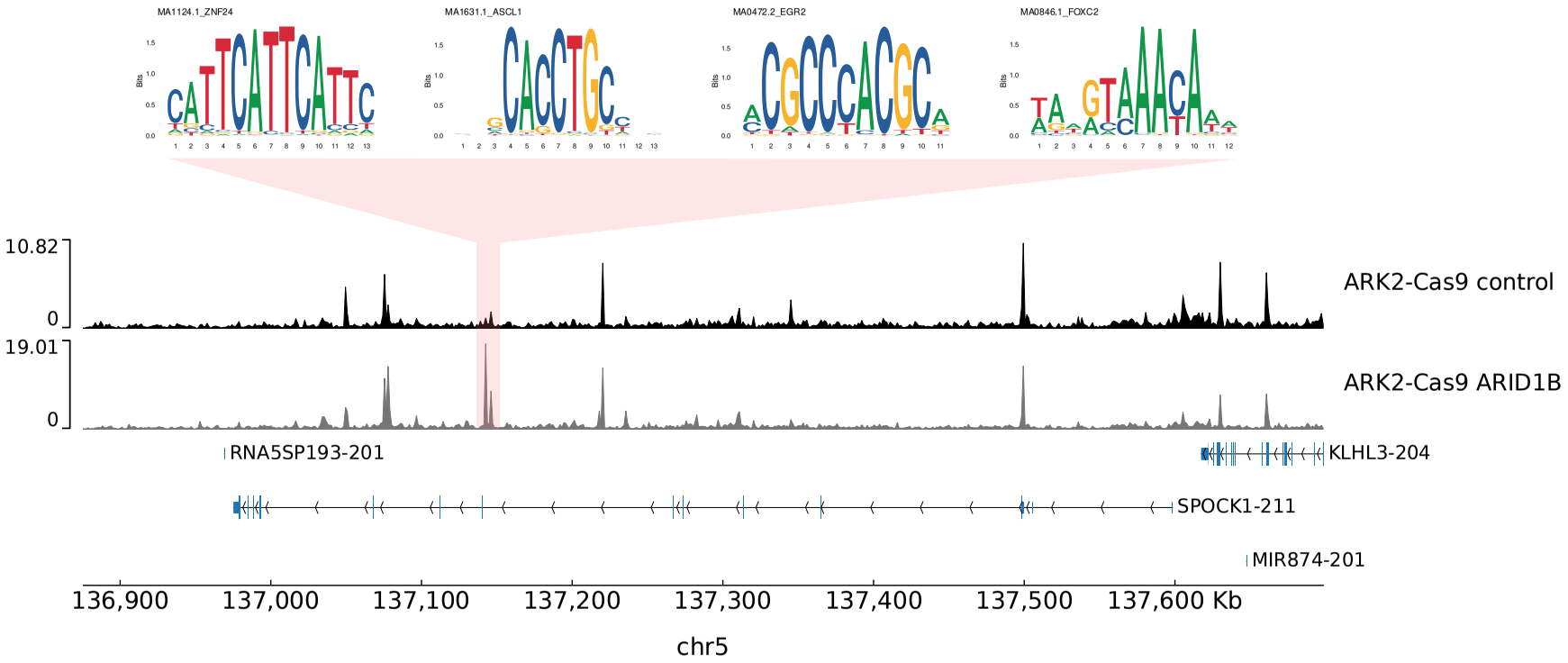


Figure 3.23: **Motifs found in the potential enhancer site near *SPOCK1* gene.**

pared the results with those identified from expression analysis (Fig. 3.22). I found KRAS signaling and estrogen response commonly up-regulated at the chromatin accessibility and expression levels. I also found EMT down-regulated in both datasets. Interestingly, the TGF-β pathway was up-regulated in gene expression but down-regulated at the accessibility level (NES $= -1.77$, adj. p-value $= 3.63 \times 10^{-2}$), possibly suggesting a more complex relationship between chromatin accessibility and expression.

Another up-regulated gene worth of note is *SPOCK1* (log2 fold change= 10.09, adj. p-value $= 1.85 \times 10^{-8}$) - a member of the EMT pathway, and also differentially expressed and accessible in Ark2 Cas9 $ARID1A^{KO}$. I found one of the more accessible chromatin regions in *SPOCK1*s intronic sequence (Fig. 3.23). This region, with increased accessibility, can serve as an enhancer site influencing the expression of the gene. I identified motif binding sites from four classes of transcription factors - ZNF24, ASCL1, EGR2, and FOXC2. The presence of those motifs, increased accessibility in Ark2 Cas9 $ARID1B^{KO}$, and up-regulation of the *SPOCK1* suggest that this is, in fact, an enhancer site that, in the event of *ARID1B* loss becomes accessible for the transcription factors to bind and regulate expression of *SPOCK1*.

3.2.6 *ARID5B* loss results in up-regulation of interferon response pathways

ARID5B is a transcriptional regulator shown to play a role in regulating cell cycle progression and apoptosis in cancer cells. Our analysis of GEL data confirmed *ARID5B* as a driver in EC - 25% of GEL EC samples harbor either a mutation, a copy number loss, or both. As with the previous chromatin modifiers, Ark2 Cas9 $ARID5B^{KO}$ cells were generated by CRISPR-Cas9 as described in Section 3.4.1.1.

There are 1,264 up-regulated (FC > 1.5, adj. p-value < 0.05) and 1,125 down-regulated (FC < 0.67, adj. p-value < 0.05) genes between Ark2 Cas9 $ARID5B^{KO}$ and Ark2 Cas9 parental cells (Fig. 3.25).

Analysis of deregulated pathways identified 18 up-regulated and 13 down-regulated pathways (GSEA on Hallmark pathways, adjusted p-value < 0.05) in Ark2 Cas9 $ARID5B^{KO}$

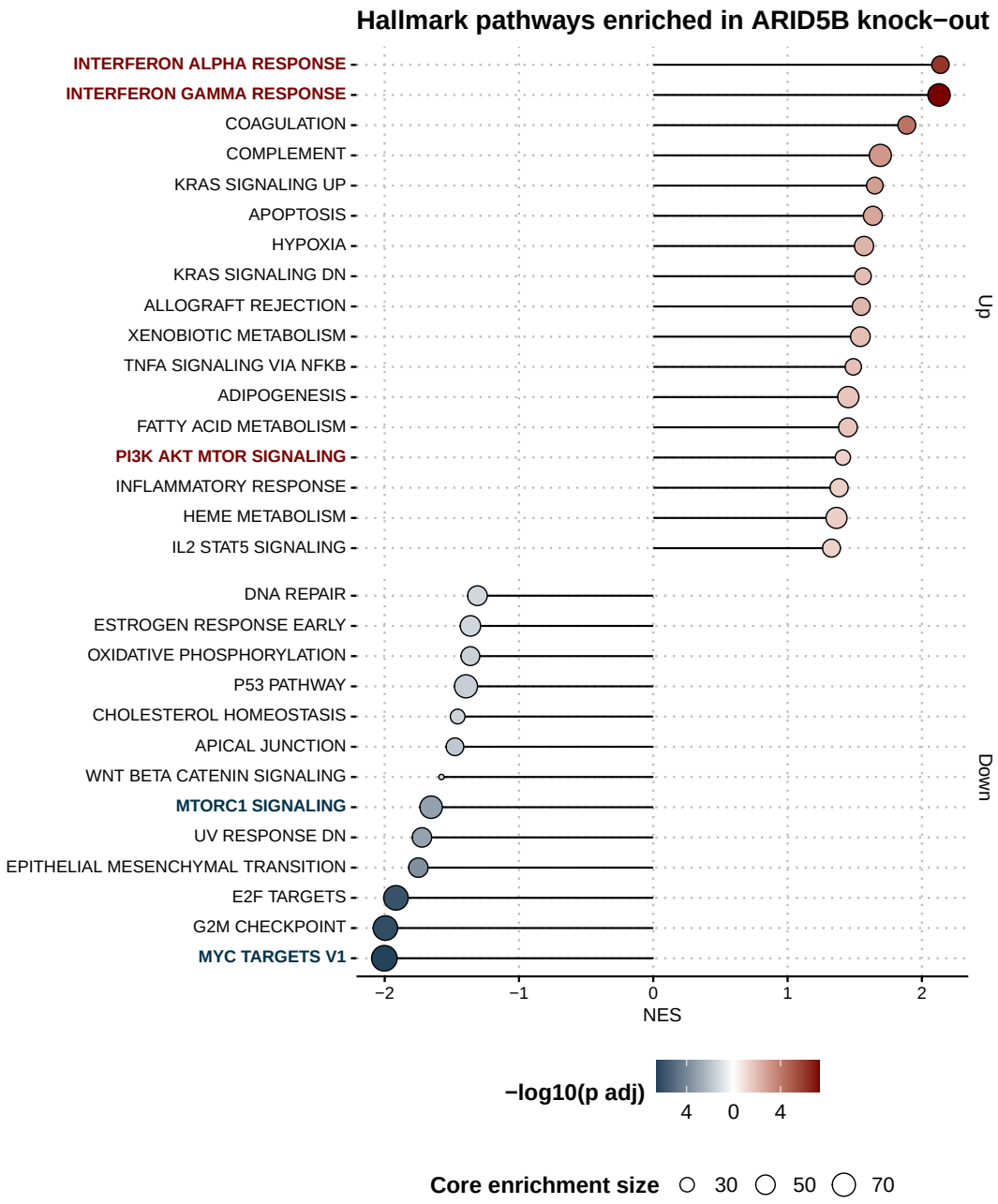


Figure 3.24: **Dotplot of perturbed pathways in Ark2 Cas9 $ARID5B^{KO}$.** Results of GSEA (adj. p-value < 0.05) on Hallmark pathways shown. Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the number of genes in a pathway. Key pathways discussed in the text are emphasized. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis.

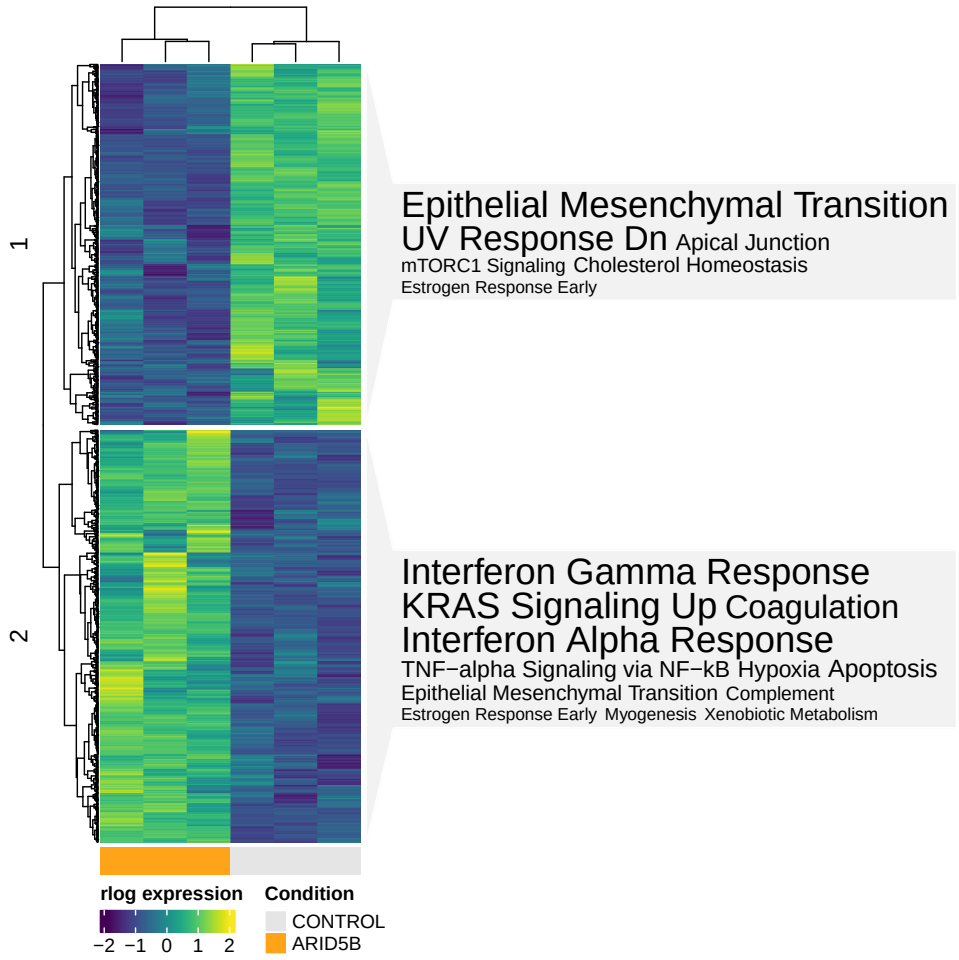


Figure 3.25: **Differentially expressed genes in Ark2 Cas9 $ARID5B^{KO}$.** Up- and down-regulated clusters annotated with enriched pathways identified with Enrichr [276] (adj. p-value<0.05). Size of the font corresponds to the combined score - combination of p-value and rank score as defined in [276].

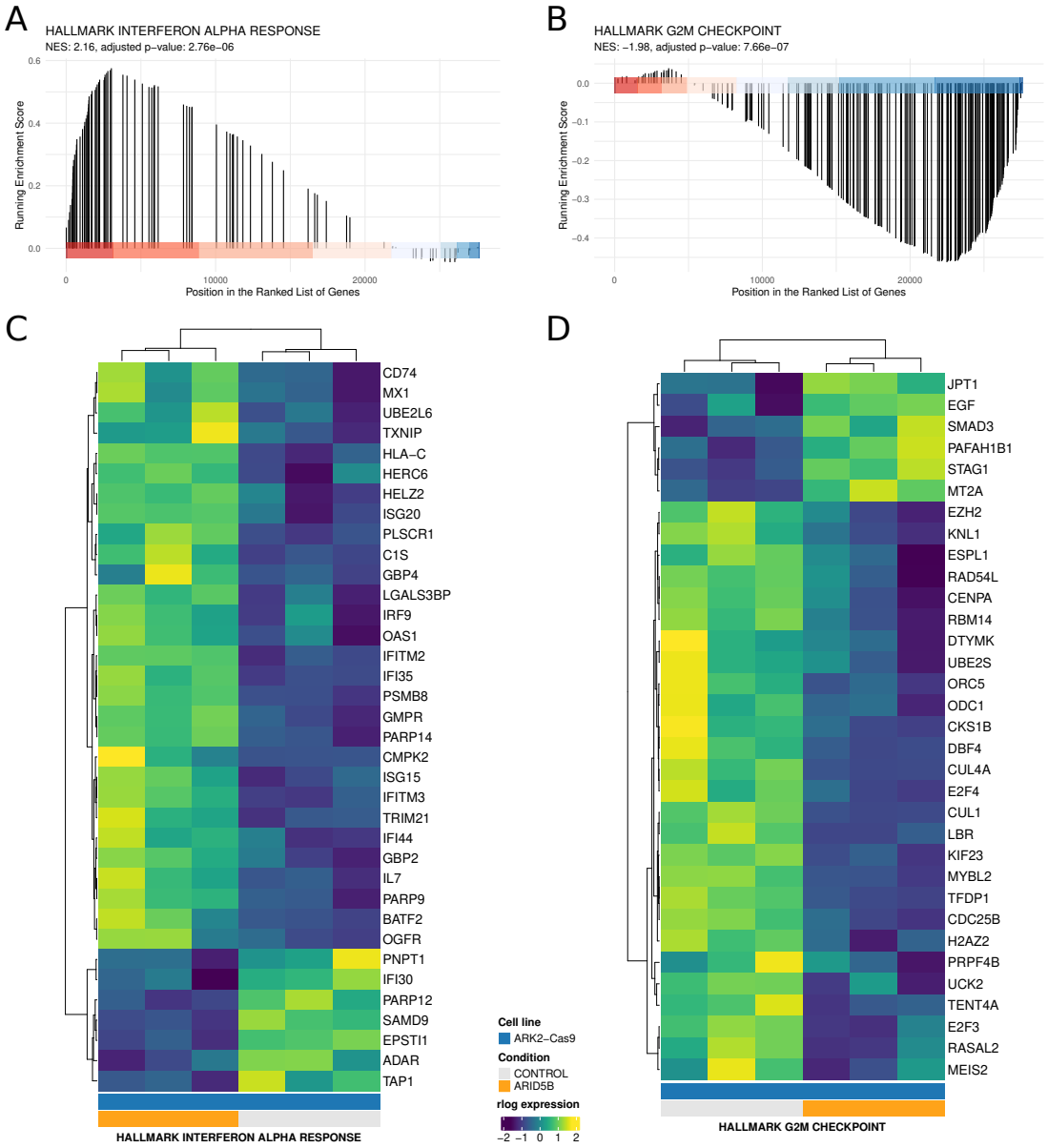


Figure 3.26: **Most dysregulated pathways in Ark2 Cas9 *ARID5B^{KO}*.** GSEA plot for up-regulated INF-α response (**A**) and down-regulated G2/M checkpoint (**B**) pathways. Heatmap of the expression of differentially expressed genes (adjusted p-value < 0.05) belonging to the INF-α response (**C**) and G2/M checkpoint (**D**) pathways.

(Fig. 3.24). Among the up-regulated pathways were interferon alpha (INF-α) response (Fig. 3.26A, NES = 2.16, adj. p-value = 2.8×10^{-6}) and interferon-gamma (INF-γ) response (NES = 2.13, adj. p-value = 2.7×10^{-8}). INF-α has been shown to regulate immune responses in various cancers. As shown in Fig. 3.26C, most of the differentially expressed genes belonging to the INF-α are up-regulated with several exceptions, such as *ADAR*. Loss of ADAR has been linked to improving cancer cells' sensitivity to immunotherapy and increased inflammation [292].

Another notable dyregulated pathway with Ark2 Cas9 $ARID5B^{KO}$ is the G2/M checkpoint, which is down-regulated (Fig. 3.26B) with 27 genes down-regulated (Fig. 3.26D, adj. p-value < 0.05) and 6 genes, such as *SMAD3*, up-regulated (Fig. 3.26D, adj. p-value < 0.05). An attenuated G2/M checkpoint can lead to an increase in cell proliferation, as cells are allowed to progress through the cell cycle more rapidly, or cycle despite the acquisition of DNA damage. Speculatively, such increased proliferation could be advantageous for cancer cells, promoting their growth and survival.

Intrestingly, two mTORC pathways are dysregulated in Ark2 Cas9 $ARID5B^{KO}$ mTORC1 signaling is up-regulated (Fig. 3.27A, NES = -1.64, adj. p-value = 4.8×10^{-4}), while PI3K/Akt/mTOR signaling is down-regulated (Fig. 3.27A, NES = 1.43, adj. p-value = 4.5×10^{-2}). The two pathways share 21 genes, which amounts to 20% of PI3K/Akt/mTOR signaling and 11% of mTORC1 signaling pathway. As shown in Fig. 3.27, five to those overlapping genes *ACTR3*, *CDKN1A*, *PDK1*, *PIK3R3* and *SQSTM1* are up-regulated, while *NFKBIB* was down-regulated. However, most of the differentially expressed genes are exclusive to either pathway.

The down-regulation of mTORC1 signaling and the up-regulation of PI3K/Akt/mTOR signaling could reflect a compensatory response by the cell to maintain cellular growth, proliferation, and survival. In this scenario, the cell could be trying to overcome the reduced mTORC1 activity by increasing the activation of the PI3K-Akt pathway, which can, in turn, activate mTORC1. Additionally, up-regulation of PI3K/Akt/mTOR signaling might reflect an attempt by the cell to overcome this negative feedback and restore the pathway's activity. However, further experimentation and analysis are required before

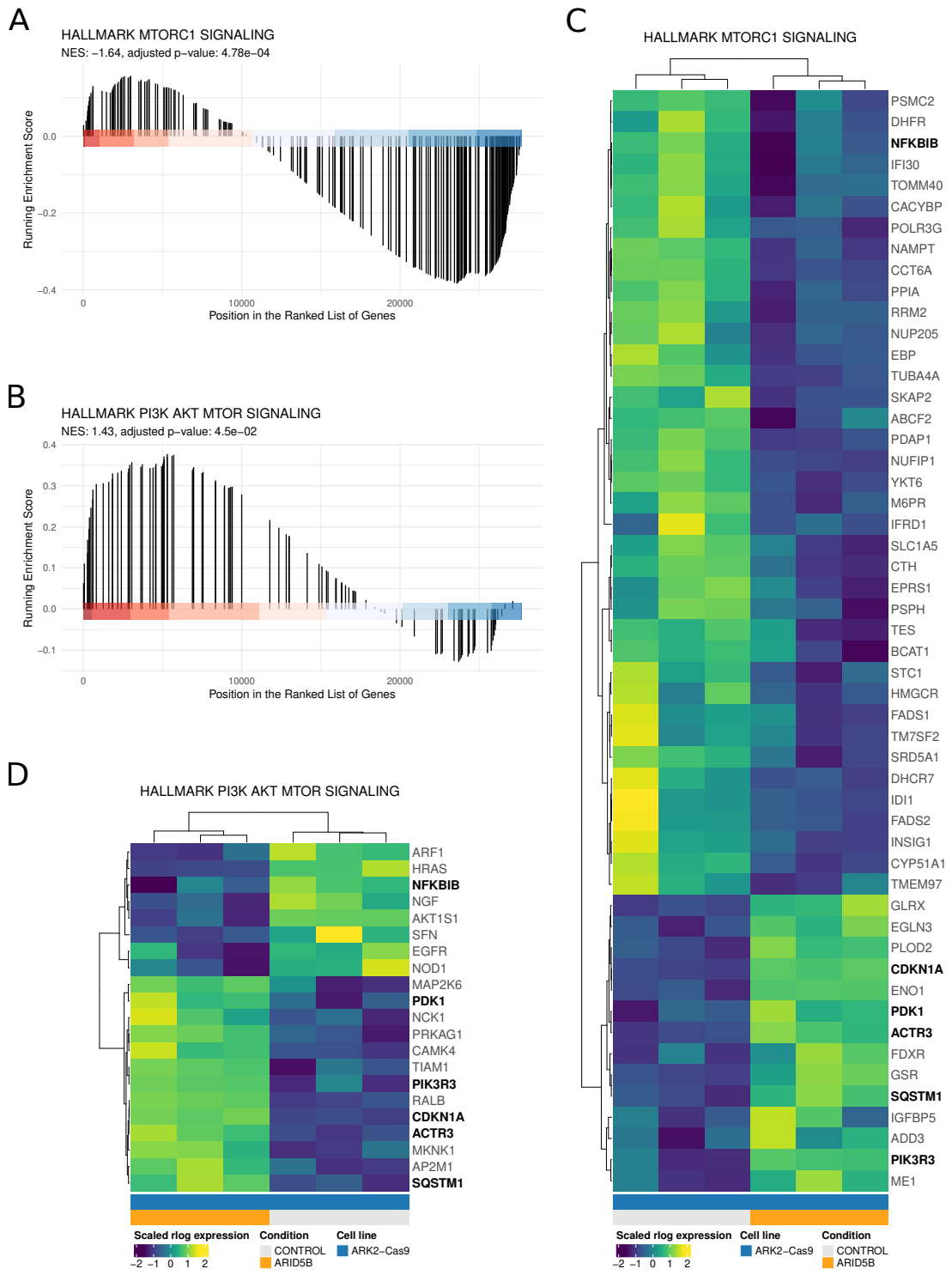
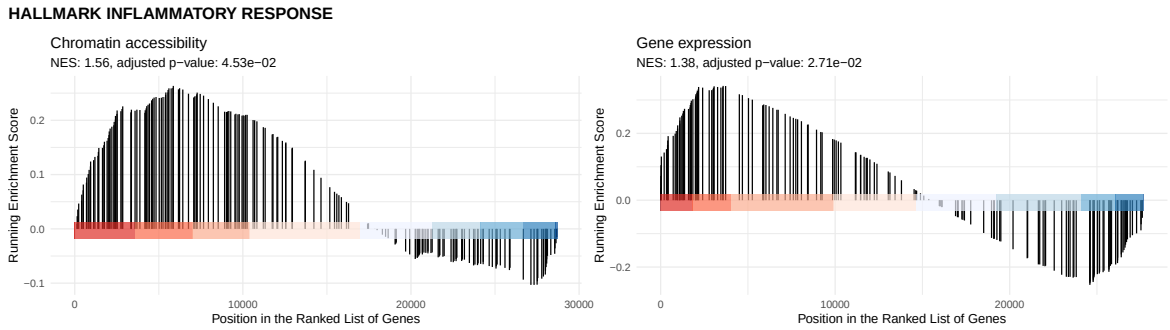
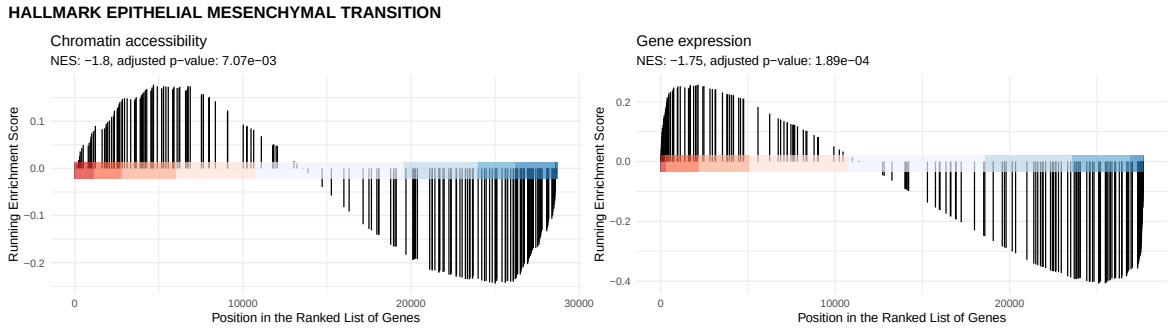


Figure 3.27: **mTORC pathways are dysregulated in Ark2 Cas9 *ARID5B^{KO}*.** GSEA plots for mTORC1 (**A**) PI3K/Akt/mTOR (**B**) signaling pathways. Heatmap of the expression of differentially expressed genes (FC $>$ 1.5 and FC $<$ 0.67, adjusted p-value $<$ 0.05) belonging to the mTORC1 (**C**) and PI3K/Akt/mTOR (**D**) signaling pathways.

firm conclusions on this dysregulation can be made.



(a) **Inflammatory response is up-regulated at the chromatin accessibility and gene expression levels.**



(b) **EMT is down-regulated at the chromatin accessibility and gene expression levels.**

Figure 3.28: **Pathways commonly perturbed in ARID5B knockout at the chromatin accessibility (left) and gene expression (right) level.**

I next analyzed ATAC-seq data from Ark2 Cas9 $ARID5B^{KO}$ cells. Differential accessibility analysis identified 8,288 more accessible regions and 11,665 less accessible regions. 56% of less accessible and 63% of more accessible regions are found in the intronic and intergenic regions (Fig. 3.20).

Annotating the dysregulated regions with the closest genes revealed 5,438 genes linked to more accessible and 4,039 genes linked to less accessible regions. When intersected with up- and down-regulated genes identified from RNA-seq analysis, I found 631 genes up-regulated and linked to more accessible regions and 558 genes down-regulated and linked to less accessible regions (Fig. 3.21C).

Two pathways are dysregulated in the same direction at both the chromatin accessibility and gene expression level (Fig. 3.28). Chromatin is more accessible around genes belonging to an inflammatory response, translating into up-regulation of the pathway at the gene expression level (Fig. 3.28a. Similarly, genes belonging to the EMT pathway are

characterized by less accessible chromatin and down-regulated expression (Fig. 3.28b).

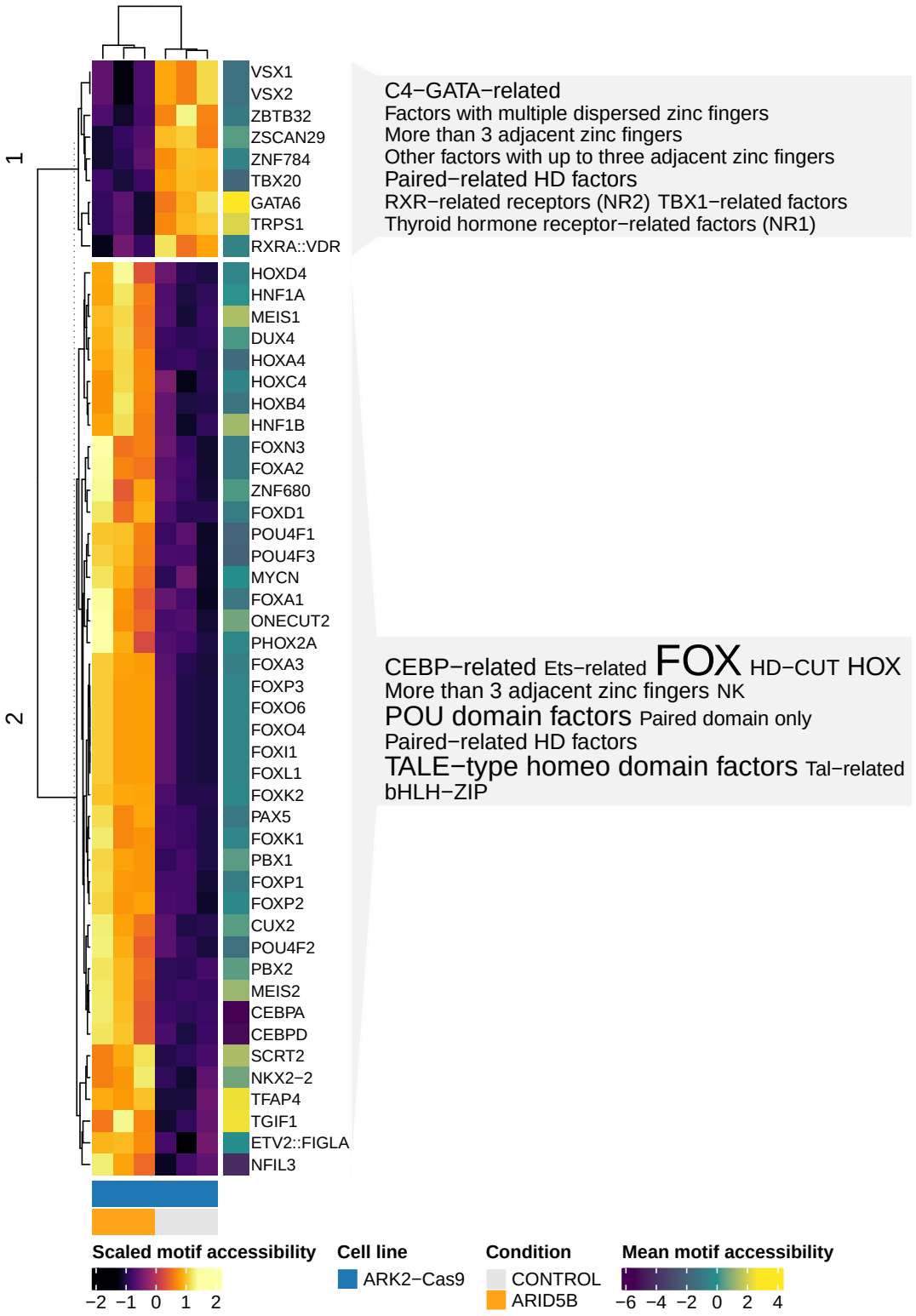


Figure 3.29: **Heatmap of disruption of motif accessibility in Ark2 Cas9 *ARID5B*^{KO}.** Heatmap of bias-corrected and normalized accessibility of motifs. Values in rows were scaled, and the row means plotted aside to allow for comparison of differences and the overall accessibility of motifs.

Next, I investigated the difference in accessibility across transcription factor binding sites. As shown in Fig. 3.29, there are two groups of motifs - those that are less accessible and those whose motifs are more accessible across the Ark2 Cas9 $ARID5B^{KO}$ genome. The former group includes motifs of GATA6, a transcription factor with high overall accessibility and less accessible regions in Ark2 Cas9 $ARID5B^{KO}$ compared to Ark2 Cas9 parental cells. The transcription factors whose motifs are more accessible in Ark2 Cas9 $ARID5B^{KO}$ include motifs of Fox family transcription factors. Two transcription factors with high accessibility and motifs more accessible in Ark2 Cas9 $ARID5B^{KO}$ are TGIF1 and TFAP4. TFAP4 is found to up-regulate cell growth in prostate cancer by positively regulating *FOXK1* [293], a transcription factor whose motif is also more accessible in Ark2 Cas9 $ARID5B^{KO}$.

3.2.7 JUND motifs are more accessible in *EP300* knockouts

p300, encoded by *EP300* gene, is a histone acetyltransferase that plays a key role in regulating gene expression in response to various cellular signals. As discussed in Section 2.3.2, *EP300* is altered in 30% of EC patients, including endometrial cancer, have been identified in several cancer types and are associated with a poorer prognosis. Loss of *EP300* has been shown to promote cancer cell proliferation, invasion, and metastasis and to contribute to the development of drug resistance [294, 295].

Ark2 Cas9 $EP300^{KO}$ cells were generated by CRISPR-Cas9 as described in Section 3.4.1.1. I performed differential expression analysis with DESeq2 [248]. Comparing Ark2 Cas9 $EP300^{KO}$ and Ark2 Cas9 parental cell lines revealed 1,011 up-regulated (FC $>$ 1.5, adjusted p-value $<$ 0.05) and 1,096 down-regulated genes (FC $<$ 0.67, adjusted p-value $<$ 0.05). I annotated the dysregulated genes with enriched pathways (Enrichr [276], adj. p-value $<$ 0.05). Up-regulated genes were enriched for INF-α and INF-γ response, while down-regulated genes were enriched for EMT. Those results were further confirmed in a gene set enrichment analysis. GSEA identified 21 up-regulated and nine down-regulated pathways (Fig. 3.31, adj. p-value $<$ 0.05). The three most up-regulated pathways include INF-α response (Fig. 3.30A, NES = 2.75, adj. p-value = 2.5×10^{-9}), INF-γ response

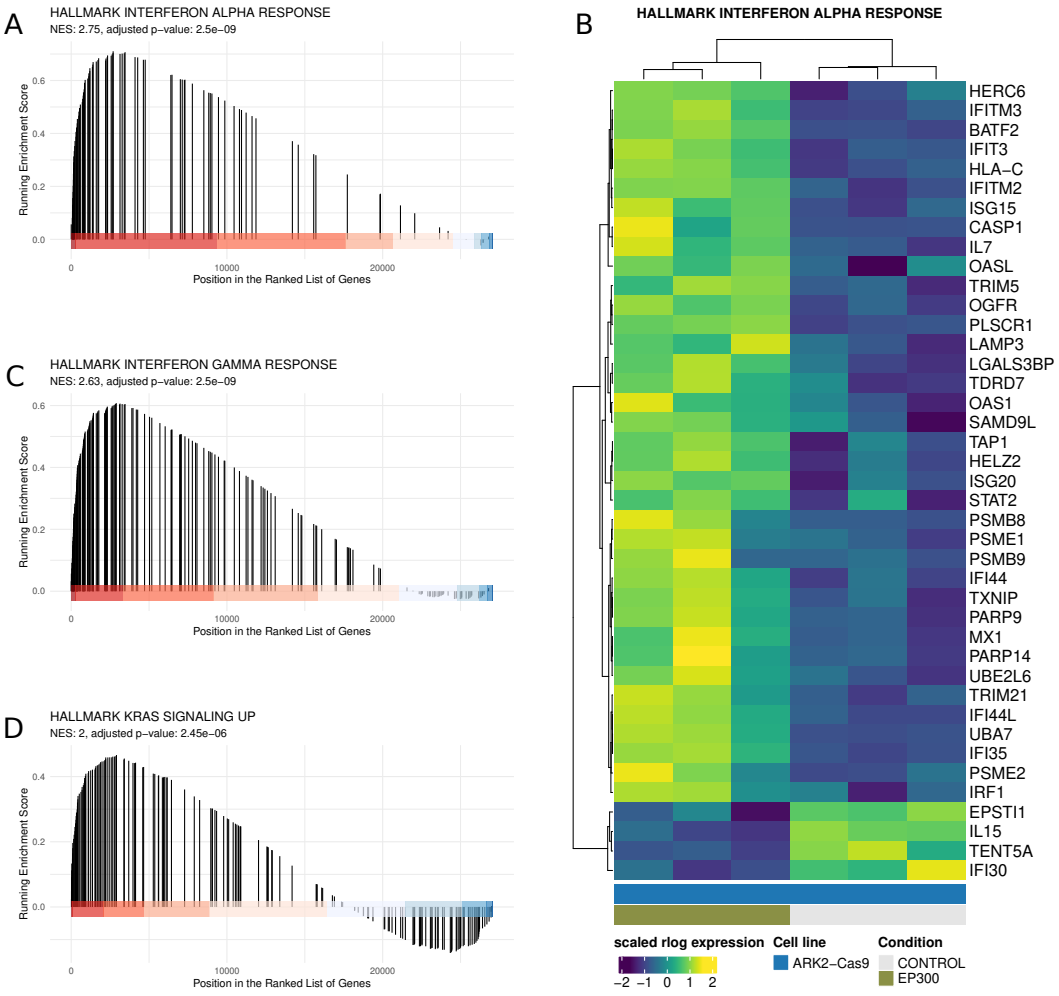


Figure 3.30: **Pathways up-regulated in Ark2 Cas9 *EP300^{KO}*. A** GSEA plot for INF-α response. **B** Heatmap of differentially expressed genes (FC $>$ 1.5 and FC $<$ 0.67, adjusted p-value $<$ 0.05), visualized using regularized logarithm-normalized expression values. For more details on enrichment analysis see Section 3.4.1.2 and for clustering and visualization, see Section 3.4.1.2.

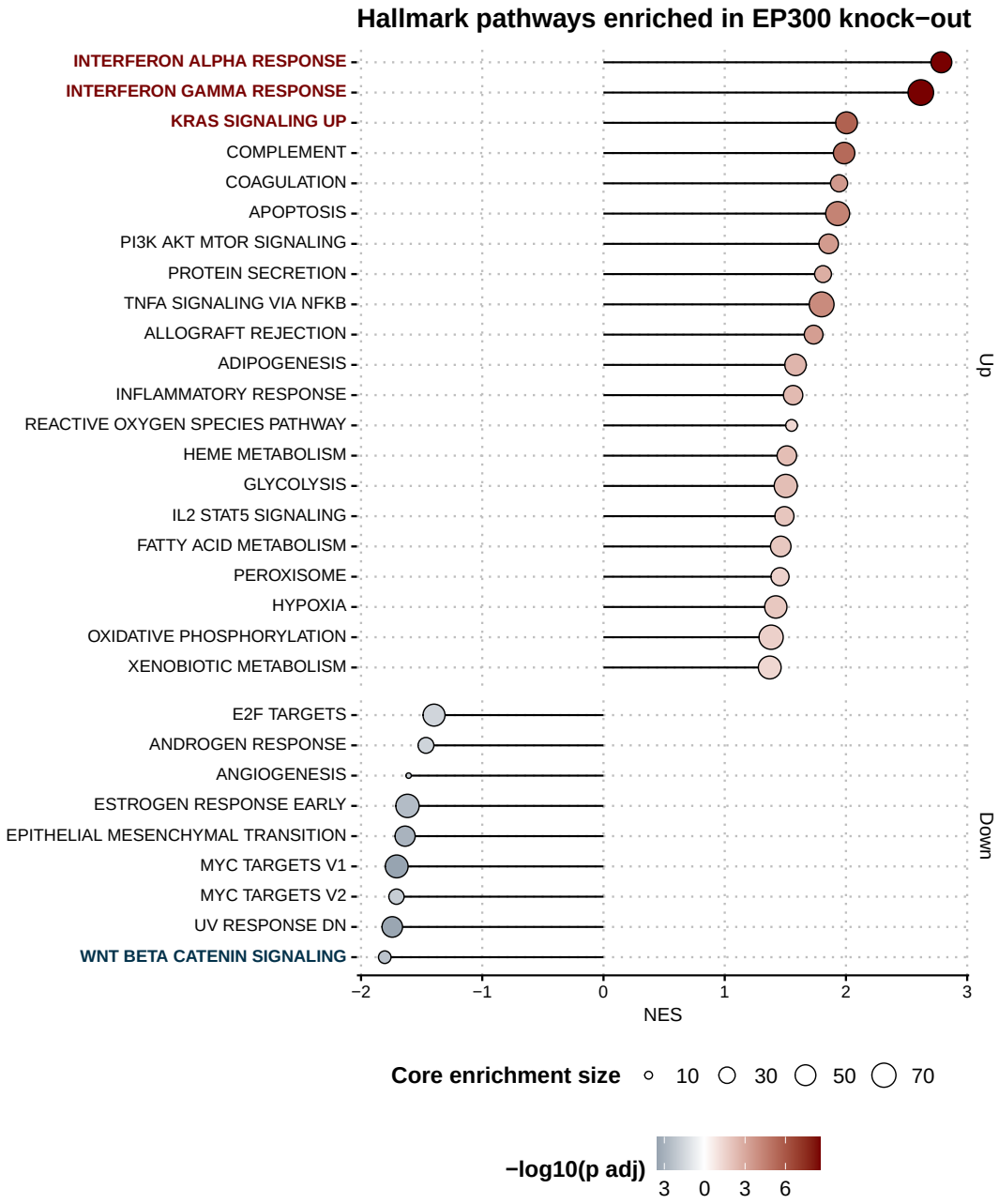


Figure 3.31: **Dotplot of pathways perturbed in Ark2 Cas9 *EP300^{KO}*.** Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the number of genes in a pathway. Key pathways discussed in the text are emphasized. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis.

(Fig. 3.30C, NES = 2.75, adj. p-value = 2.5×10^{-9}) and KRAS signaling (Fig. 3.30D, NES = 2.0, adj. p-value = 2.5×10^{-6}). Dysregulated genes belonging to INF-α response included 37 up-regulated and four down-regulated genes (Fig. 3.30B).

Activation of INF-α and INF-γ responses can have antitumor effects by inhibiting cell proliferation, promoting apoptosis, and enhancing the immune response against cancer cells [296]. On the other hand, chronic activation of INF-α and INF-γ can contribute to cancer development by promoting inflammation, angiogenesis, and immune evasion [297]. Activation of interferon responses upon p300 loss might represent a compensatory mechanism to counteract the potential tumor-promoting effects. Further experiments, such as cell viability and proliferation assays, would provide more information required to understand the consequences of up-regulation of interferon responses in the context of p300 loss.

Activation of KRAS signaling in cancer, a pathway also found up-regulated in EP300 knockout cells, is typically associated with increased tumor aggressiveness and metastasis [298, 299]. Ark2 is a HER2-positive cell line. HER2, encoded by *ERBB2*, is a known activator of the KRAS pathway [13]. This could suggest that KRAS signaling pathway activation is further heightened with the loss of *EP300*.

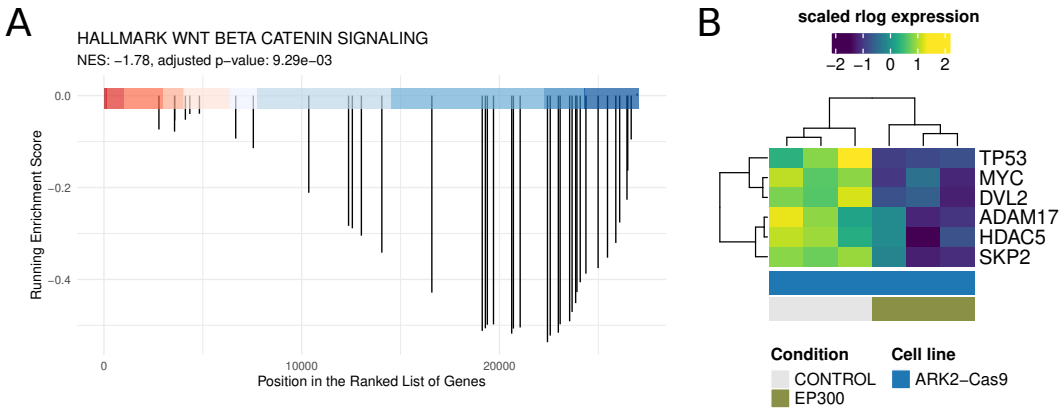


Figure 3.32: **EP300 loss results in down-regulation of Wnt/β-catenin pathway.** **A** GSEA plot of the Wnt/β-catenin pathway. **B** Heatmap of the standardized rlog normalized expression. For more details on enrichment analysis, see Section 3.4.1.2; for clustering and visualization, see Section 3.4.1.2.

As shown in Fig. 3.32, the Wnt/β-catenin pathway is down-regulated in Ark2 Cas9 $EP300^{KO}$ (Fig. 3.32A, NES = −1.78, adj. p-value = 9.3×10^{-3}). Six genes belonging to

the pathway are down-regulated (Fig. 3.32B) and they include *TP53* ($\log_2(\text{FC}) = -1.21$, adj. p-value $= 0.01$) and *MYC* ($\log_2(\text{FC}) = -1.22$, adj. p-value $= 4.9 \times 10^{-10}$).

Chromatin accessibility revealed 9,995 more and 9,183 less accessible regions in Ark2 Cas9 $EP300^{KO}$ compared to Ark2 Cas9 parental cells. As shown in Fig. 3.20, 56% of less and 63% of more accessible regions are found in intronic and intergenic regions. Only 29% of more accessible chromatin is found close to the transcription start site (within 3,000 bp). Annotating the accessible regions to the nearest genes identified 5,282 genes linked to more accessible regions and 4,044 to the less accessible regions (Fig. 3.21D).

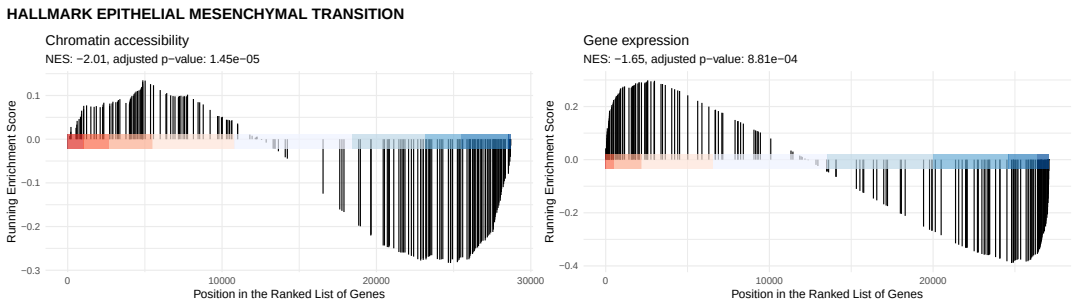


Figure 3.33: **EMT is down-regulated at the chromatin accessibility and gene expression levels in Ark2 Cas9 $EP300^{KO}$.**

As shown in Fig. 3.33, regions around the genes belonging to EMT pathway are less accessible in Ark2 Cas9 $EP300^{KO}$ cells (NES $= -2.01$, adj. p-value $= 1.5 \times 10^{-5}$), which translates to down-regulation of EMT at the expression level (NES $= -1.65$, adj. p-value $= 8.8 \times 10^{-4}$).

AP1 is a dimeric transcription factor formed by Fos and Jun family members. JUND motif is more accessible in Ark2 Cas9 $EP300^{KO}$. Although research on the specific role of JunD in EC is limited, JunD was identified as a crucial component for Ras-induced tumorigenesis in lung cancer [300]. Here, we observe the up-regulation of the KRAS signaling pathway in conjunction with the increased accessibility of JUND motifs across the EC genome. This could suggest that *EP300* facilitates KRAS pathway up-regulation by exposing JunD binding sites in EC.

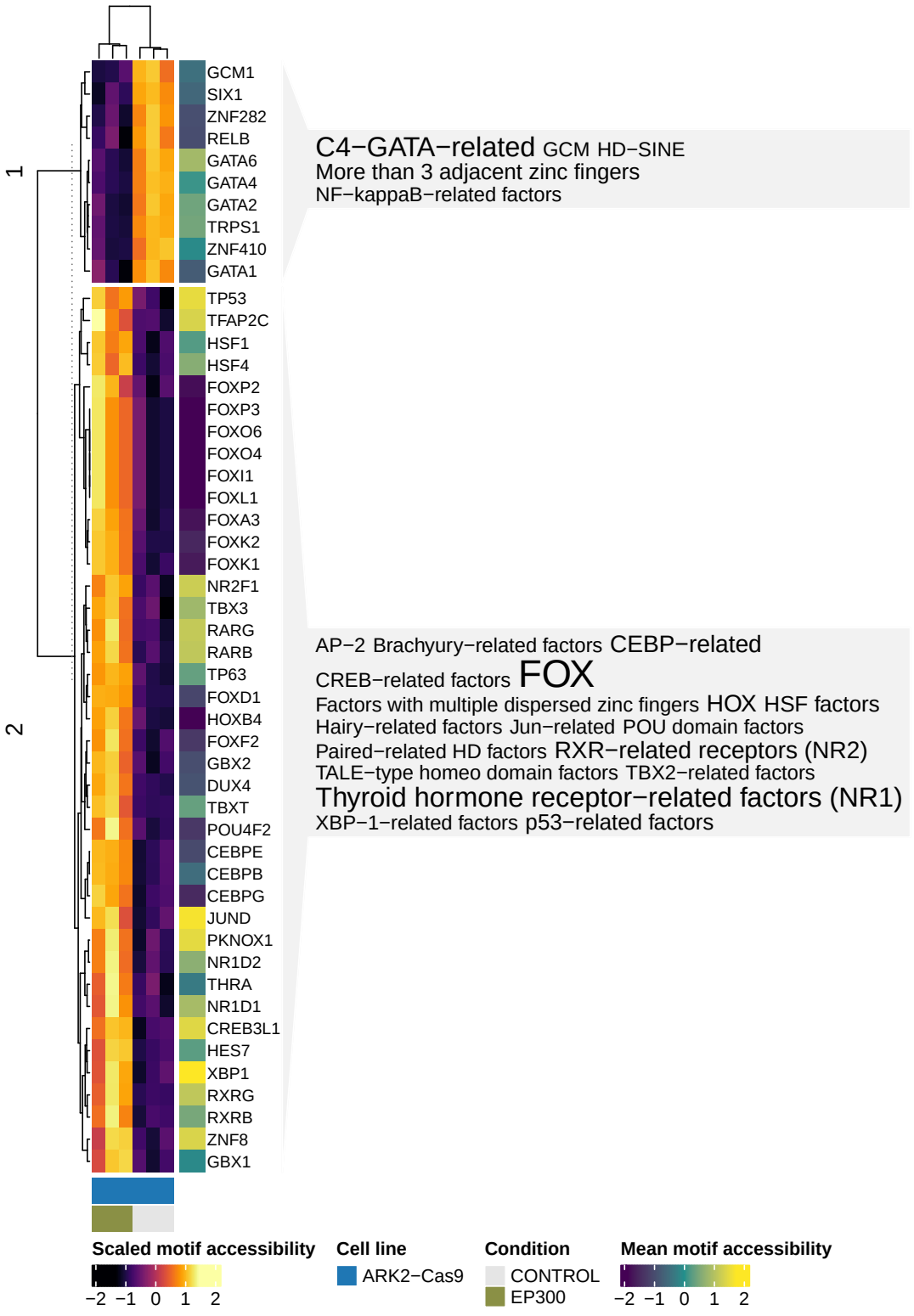


Figure 3.34: **Heatmap of disruption of motif accessibility in Ark2 Cas9 *EP300^{KO}*.** Heatmap of bias-corrected and normalized accessibility of motifs. Values in rows were standardized (scaled), and the row means were plotted aside to allow for the comparison of differences and the overall accessibility of motifs. For more details on enrichment analysis, see Section 3.4.1.2; for clustering and visualization, see Section 3.4.1.2.

3.2.8 Loss of *KMT2C* results in up-regulation of E2F targets

KMT2C encoding a histone methyltransferase is frequently mutated in various cancers and at particularly high frequency in endometrial cancer. *KM2TC* has been identified as a coding driver in CN high subtype (Fig. 2.20) and is altered in 30% of EC samples (Fig. 2.27).

Ark2 Cas9 $KM2TC^{KO}$ cells were generated by CRISPR-Cas9 as described in Section 3.4.1.1. I investigated the perturbations in expression profiles between Ark2 Cas9 $KM2TC^{KO}$ and Ark2 Cas9 parental cells by differential expression analysis. This revealed 780 up-regulated (FC $>$ 1.5, adj. p-value $<$ 0.05) and 660 down-regulated genes (FC $<$ 0.67, adj. p-value $<$ 0.05). Up-regulated genes were enriched for KRAS signaling pathway, while down-regulated genes were enriched for EMT (Enrichr [276], adj. p-value $<$ 0.05).

GSEA identified ten up-regulated and 18 down-regulated pathways (Fig. 3.35, adjusted p-value $<$ 0.05). Among them is allograft rejection (Fig. 3.36A, NES = 1.86, adj. p-value = 2.3×10^{-4}). Allograft rejection refers to the immune-mediated response against transplanted tissue or an organ from a genetically non-identical donor. The recipient's immune system recognizes the donor's tissue as foreign due to differences in major histocompatibility complex (MHC) molecules and other cell surface markers. This recognition triggers an immune response that can destroy the transplanted tissue or organ, ultimately causing the rejection of the allograft. Interpreting the up-regulation of this pathway in the context of cancer is non-trivial. There are several genes belonging to the gene set with increased expression in Ark2 Cas9 $KM2TC^{KO}$ compared to Ark2 Cas9 parental cells, such as *IFNGR2* ($\log_2$(FC) = 0.49, adj. p-value = 0.013), *HLA-A* ($\log_2$(FC) = 0.46, adj. p-value = 0.010) or *TNF* ($\log_2$(FC) = 1.36, adj. p-value = 0.007) that are closely related to inflammation. Chronic inflammation can contribute to cancer development and progression [301]. *IL11*, also up-regulated in Ark2 Cas9 $KM2TC^{KO}$ ($\log_2$ fold change = 1.28, adj. p-value = 5.4×10^{-4}), is a cytokine that plays a role in hematopoiesis, bone metabolism, and inflammation. In cancer, IL11 has been associated with tumor progression, metastasis, and therapy resistance in various cancer types, such as breast

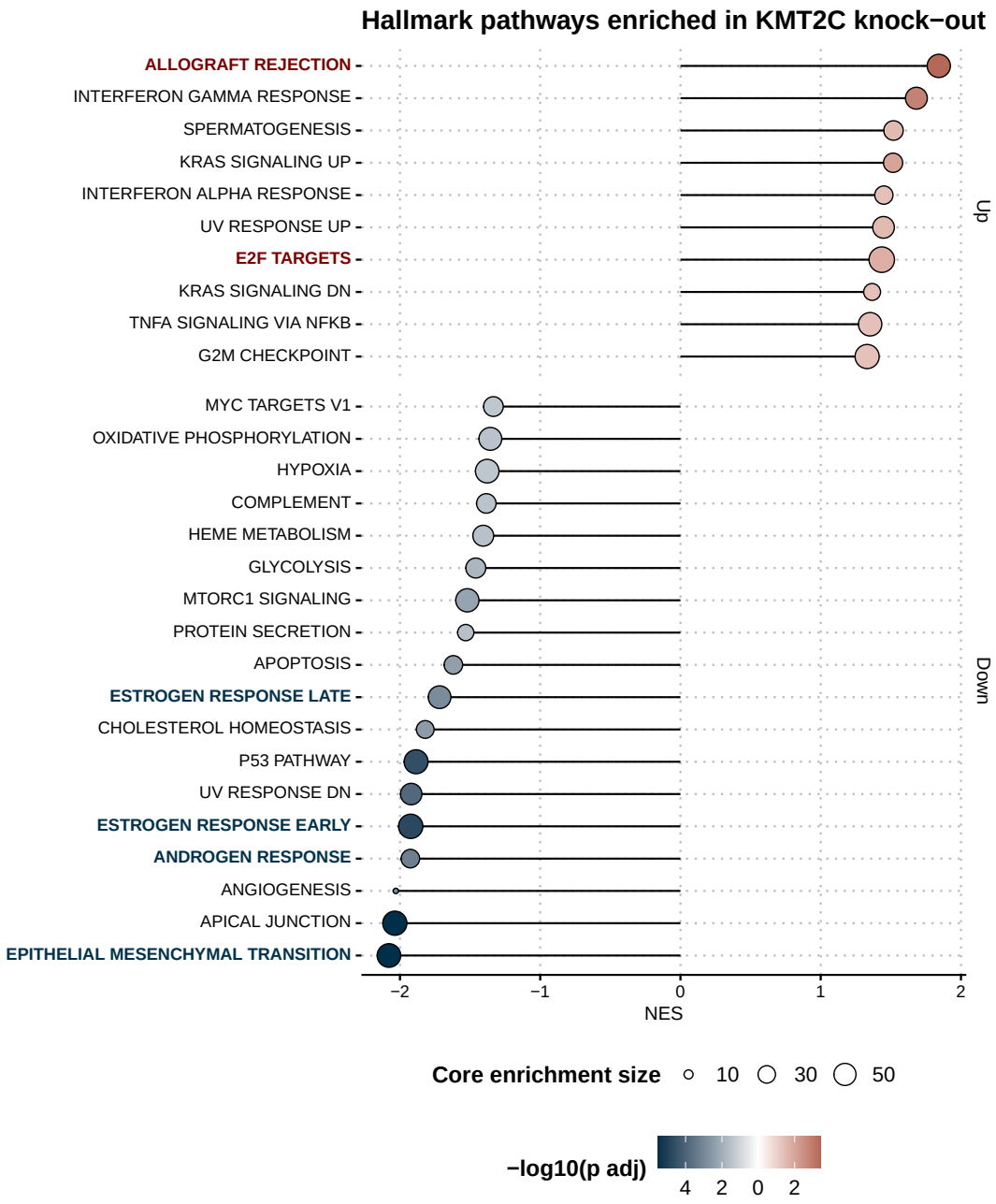


Figure 3.35: **Dotplot of pathways perturbed in Ark2 Cas9 *KM2TC^{KO}*.** Results of GSEA (adj. p-value < 0.05) on Hallmark pathways shown. Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the number of genes in a pathway. Key pathways discussed in the text are emphasized. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis.

[302], gastric [303], and colorectal [304] cancer. IL11 has been found to be up-regulated in EC [305] and has been implicated in endometrial stromal cell differentiation via activation of the JAK/STAT signaling pathway [306].

The up-regulation of E2F targets is another notable pathway identified in the KMT2C knockout (Fig. 3.36C and D, NES = 1.45, adj. p-value = 1.5×10^{-2}). The early region 2 binding factors (E2F) family has emerged as a major transcriptional regulator of cell cycle-dependent gene expression [307]. Expression of E2F targets is usually associated with increased cell division and proliferation. Overexpression of E2F targets is correlated with poor prognosis in liver [308] and pancreatic [309] cancers. Based on our findings and the available literature, it appears that the loss of KMT2C may provide a growth advantage to the tumor by enhancing cell proliferation.

Conversely, I observed down-regulation of androgen response (AR) in Ark2 Cas9 $KM2TC^{KO}$ compared to Ark2 Cas9 parental cells (Fig. 3.36E and F, NES = -1.92, adj. p-value = 2.6×10^{-4}). In EC, the significance of androgen signaling is not as well-established as that of estrogen signaling, which plays a central role in disease pathogenesis, as discussed above. The antiproliferative effect of androgen, as a result of androgens counteracting estrogen-driven pro-proliferative signaling in the endometrium, has been observed in a well-differentiated endometrial adenocarcinoma cell line [310]. A recent study on breast cancer provided a detailed explanation of how androgens can inhibit tumor growth in estrogen receptor (ER)-α-positive breast cancer cells. The study showed that the AR can displace ER and other transcriptional coactivators from estrogen response elements, resulting in the down-regulation of ER-regulated cell cycle genes and reduced breast cancer cell proliferation [311].

In the differential accessibility analysis, I identified 6,569 regions with increased accessibility and 15,572 regions with decreased accessibility with KMT2C knockout (FC < 0.67 and FC > 1.5, adj. p-value < 0.05). Among these, 70% of less accessible and 57% of more accessible regions are located within intronic and intergenic regions (Fig. 3.20).

Upon annotating the dysregulated regions with their closest genes, 3,278 genes were associated with increased accessibility, while 3,187 genes were associated with decreased

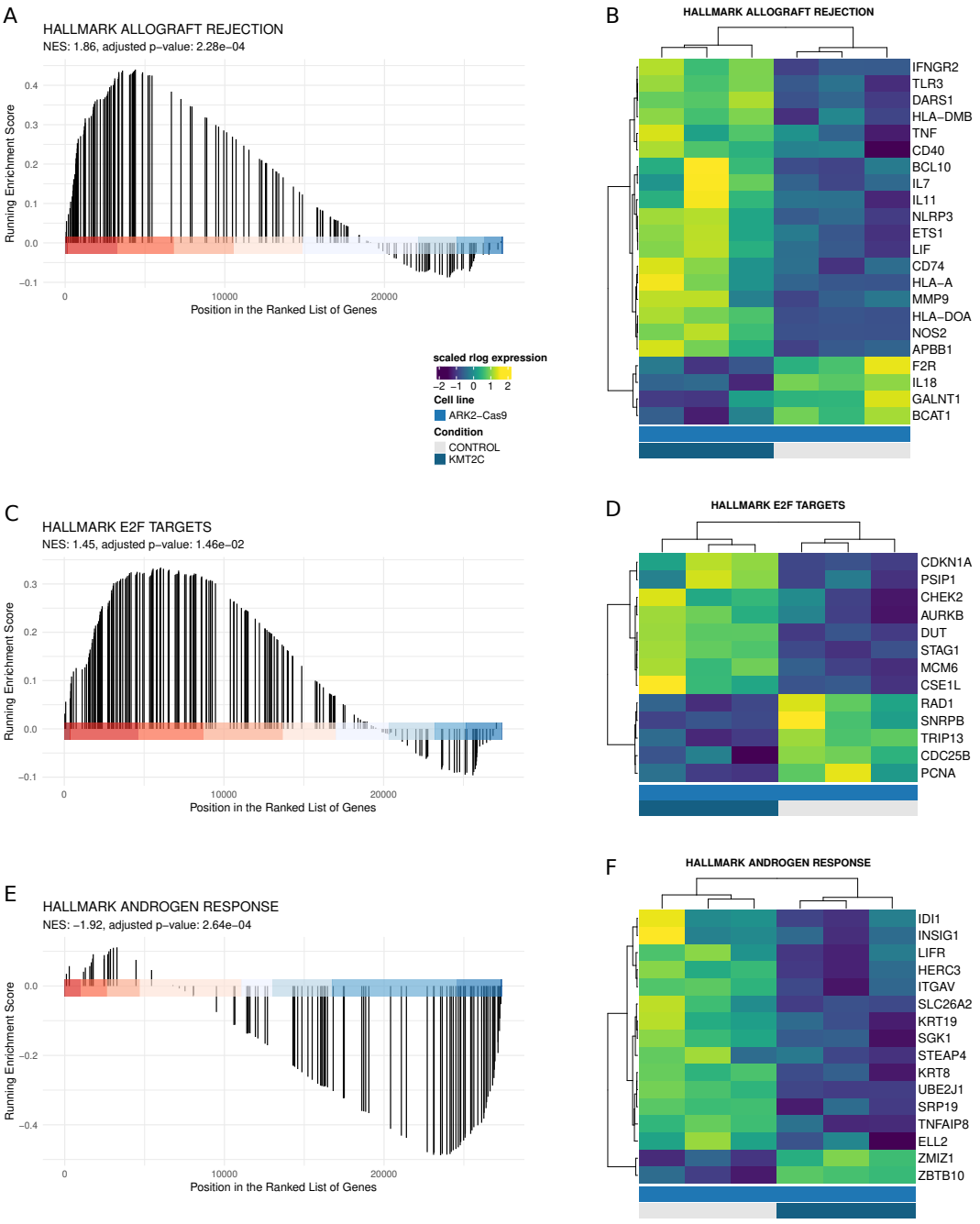


Figure 3.36: **Pathways dysregulated in Ark2 Cas9 *KM2TC^{KO}*.** On the left GSEA plots for Hallmark pathways: up-regulated allograft rejection (**A**) and E2F targets (**C**) and down-regulated androgen response (**E**). On the right heatmap of expression of dysregulated genes (FC < 0.67 and FC > 1.5, adj. p-value < 0.05) belonging to (**B**) allograft rejection, (**D**) E2F targets, and (**F**) androgen response.

accessibility. By intersecting this data with the up- and down-regulated genes identified through RNA-seq analysis, I discovered that 752 up-regulated genes corresponded to regions with increased accessibility, and 646 down-regulated genes were linked to regions with decreased accessibility (Fig. 3.21E).

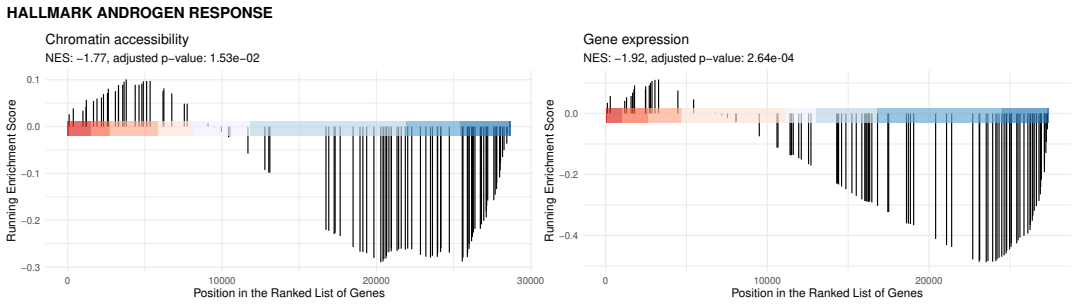


Figure 3.37: **Androgen response in Ark2 Cas9 *KM2TC^{KO}*.** Perturbation of androgen response is visible at the gene and chromatin accessibility levels. GSEA plot for chromatin accessibility on the right and gene expression on the left.

down-regulation of AR targets is also visible at the chromatin accessibility level (Fig. 3.37, NES $= -1.77$, adj. p-value $= 1.5 \times 10^{-2}$) providing further evidence of down-regulation of AR response as an effect of KMT2C loss.

3.2.9 Loss of *SETD1B* down-regulates E2F targets

SETD1B, also known as *KMT2G*, belongs to the KMT2 (MLL) family of methyltransferases. Some previous studies have suggested that SETD1B may also play a role in endometrial cancer pathogenesis, although the precise mechanisms remain to be elucidated [162]. We identified SETD1B as a recurrent coding driver in CN high subtype. We found frameshift mutation in the first codon of the gene in 12.6% of samples in the GEL EC cohort (Fig. 2.28), with 28% of all samples carrying at least a copy number alteration or a mutation (Fig. 2.27).

Ark2 Cas9 *SETD1B^{KO}* cells were generated with CRISPR-Cas9 as described in Section 3.4.1.1. Differential expression analysis of Ark2 Cas9 *SETD1B^{KO}* compared to Ark2 Cas9 parental cells identified 865 up-regulated and 820 down-regulated genes (Fig. 3.38, FC > 1.5 and FC < 0.67, adj. p-value < 0.05). Similar to the Ark2 Cas9 *KM2TC^{KO}*, up-regulated genes were enriched for the KRAS signaling pathway, while down-regulated

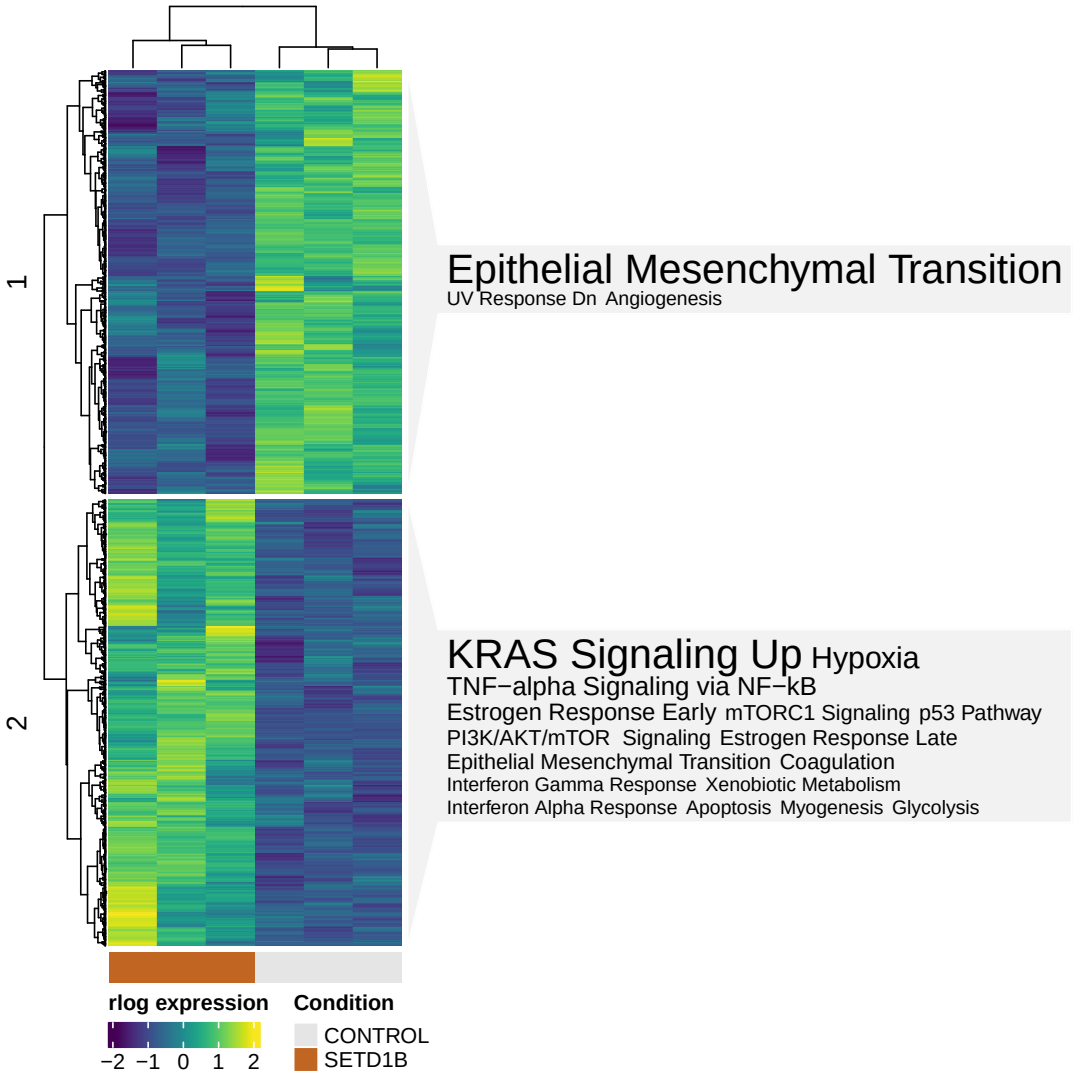


Figure 3.38: **Heatmap of expression of differentially expressed genes in Ark2 Cas9 *SETD1B^{KO}*.** The standardized rlog expression plotted. Two clusters corresponding to up-regulated (FC $>$ 1.5, adj. p-value $<$ 0.05) and down-regulated (FC $<$ 0.67, adj. p-value $<$ 0.05) genes annotated with pathways identified with enrichment analysis (adj. p-value $<$ 0.05). The font size corresponds to the combined score - a combination of p-value and rank score as defined in [276].

genes were enriched for EMT (enrichment analysis with Enrichr [276], adj. p-value < 0.05).

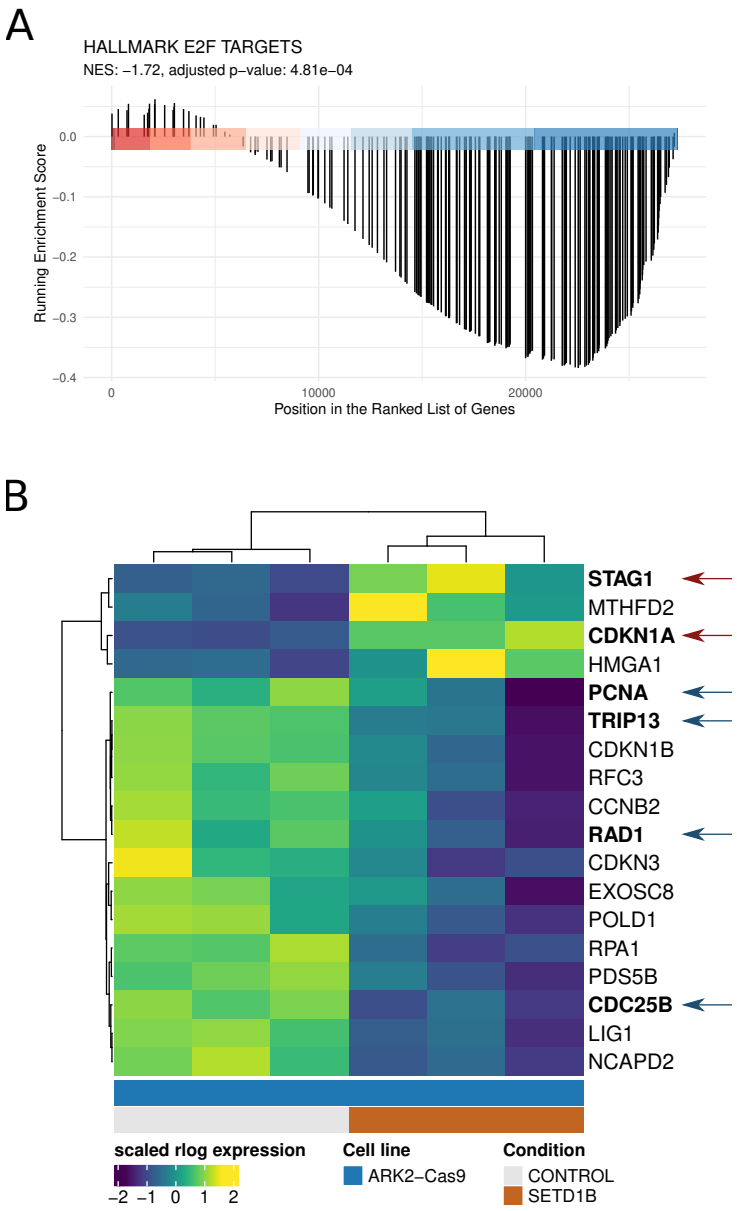


Figure 3.39: **E2F targets pathway is down-regulated in Ark2 Cas9 *SETD1B^{KO}*.** **A** GSEA plot of E2F targets gene members and **B** heatmap of differentially expressed (FC > 1.5 and FC < 0.67, adj. p-value < 0.05) genes belonging to E2F targets gene set. Genes up- and down-regulated in Ark2 Cas9 *SETD1B^{KO}* annotated with arrows – red for up-regulated and blue for down-regulated.

GSEA on Ark2 Cas9 *SETD1B^{KO}* compared to Ark2 Cas9 parental cells identified 20 up-regulated and nine down-regulated pathways (adj. p-value < 0.05). Intriguingly, I observed the down-regulation of E2F targets (Fig. 3.39A, NES $= -1.72$, adj. p-value $= 4.8 \times 10^{-4}$), which is in contrast to the direction seen in Ark2 Cas9 *KM2TC^{KO}*. Despite

the overall opposing direction of the pathway, several genes exhibited dysregulation in the same direction between the two knockouts of KMT methyltransferase family members (Fig. 3.39B). *CDKN1A* is a tumor suppressor gene that inhibits cell cycle progression by blocking cyclin-dependent kinases (CDKs). Usually, its up-regulation indicates an attempt to halt cell cycle progression and prevent uncontrolled cell division [312]. Conversely, we observed the down-regulation of *CDKN1B*, also a CDK inhibitor. In summary, the up-regulation of certain E2F target genes and the down-regulation of many others may suggest a complex interplay between altered cell cycle regulation, DNA repair, and genomic instability. This could indicate the cancer cells' attempt to balance uncontrolled proliferation while preserving genomic integrity.

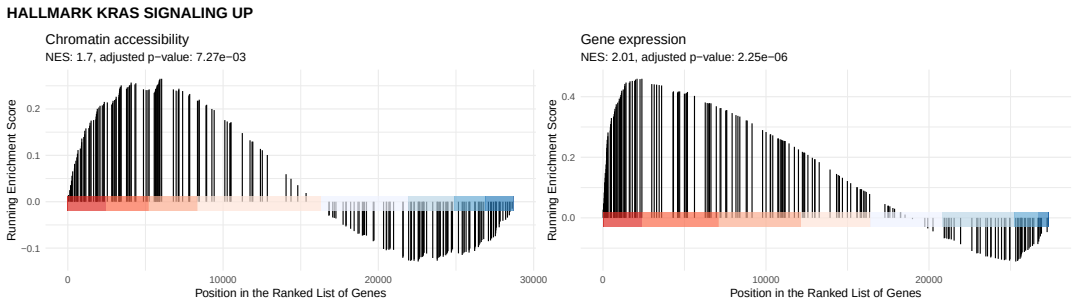


Figure 3.40: **KRAS signaling is up-regulated at the chromatin accessibility and gene expression levels.** GSEA plot for KRAS signaling across chromatin accessibility (**on the left**) and gene expression (**on the right**).

Examining chromatin accessibility by the ATAC-seq data identified 6,648 regions with increased accessibility and 6,361 regions with decreased accessibility (FC $>$ 1.5 and FC $<$ 0.67, adj. p-value $<$ 0.05). Both categories had comparable proportions of intergenic and intronic regions, with 60% of the more accessible regions and 61% of the less accessible regions falling into these categories (Fig. 3.20).

I mapped the accessible regions to the nearest genes, identifying 3,500 genes linked to more accessible and 3,295 genes linked to less accessible regions. The overlap with differentially expressed genes amounted to 837 up-regulated genes near more accessible regions and 792 down-regulated genes near less accessible regions (Fig. 3.21F).

GSEA on the differentially accessible regions revealed increased accessibility around genes belonging to KRAS signaling (NES = 1.7, adj. p-value = 7.3×10^{-3}), hypoxia (NES

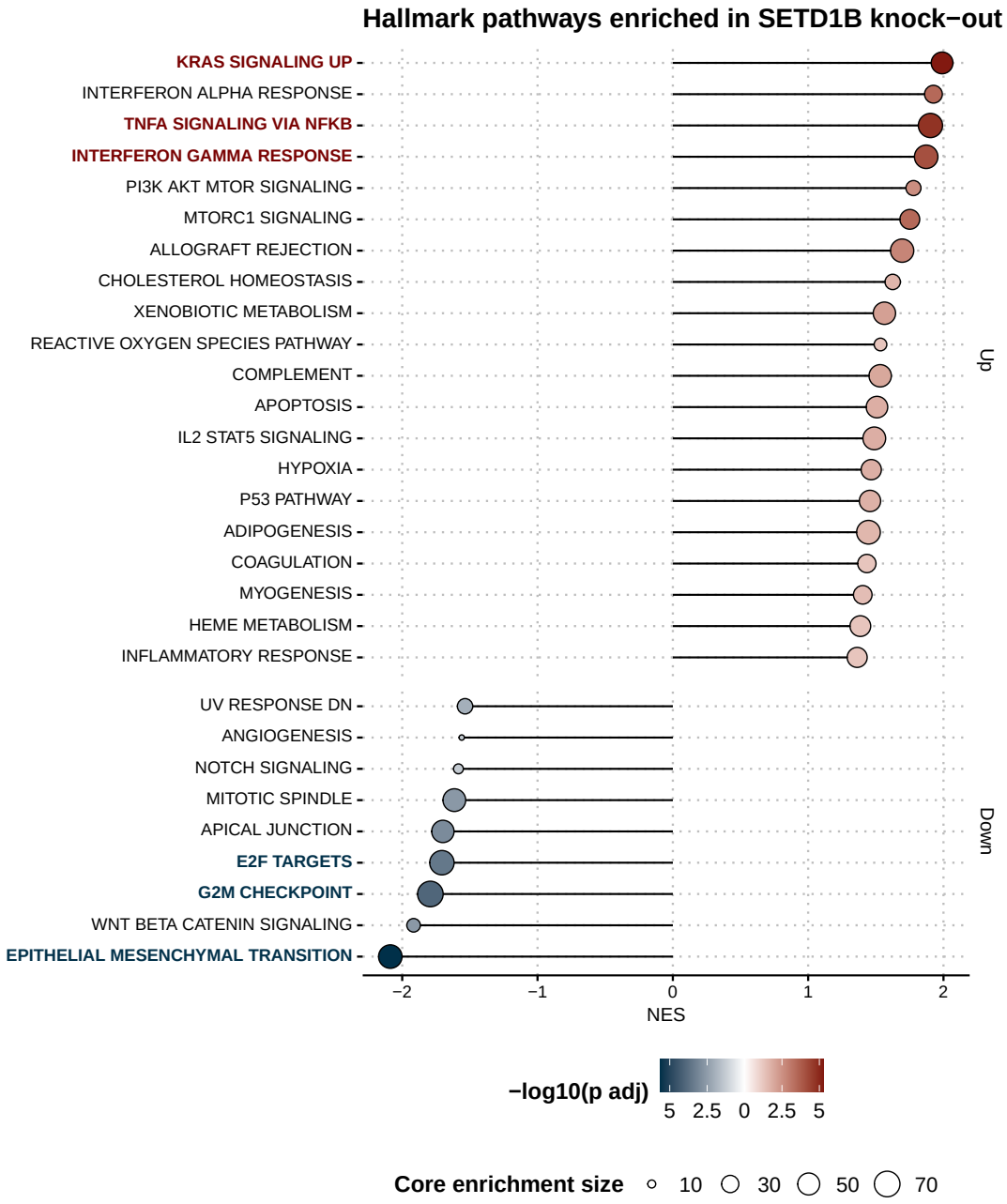


Figure 3.41: **Dotplot of pathways perturbed in Ark2 Cas9 *SETD1B^{KO}*.** Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the number of genes in a pathway. Key pathways discussed in the text are emphasized. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis

$= 1.54$, adj. p-value $= 4.2 \times 10^{-2}$), and adipogenesis pathways (NES $= 1.66$, adj. p-value $= 1.0 \times 10^{-2}$). This translated into the up-regulation of those pathways as determined by GSEA analysis on gene expression (Fig. 3.41). Similarly, we see down-regulation of EMT (NES $= -2.07$, adj. p-value $= 1.1 \times 10^{-4}$) and UV response (NES $= -1.96$, adj. p-value $= 4.1 \times 10^{-4}$) pathways and chromatin accessibility decrease around the genes from those pathways.

As discussed earlier, activation of KRAS signaling in cancer is typically associated with increased tumor aggressiveness and metastasis [298, 299]. As shown in Fig. 3.40, I observe up-regulation of KRAS signaling both at the chromatin accessibility (NES $= 1.7$, adj. p-value $= 7.3 \times 10^{-3}$) and gene expression (NES $= 2.01$, adj. p-value $= 2.3 \times 10^{-6}$) levels.

3.2.10 Common perturbations induced by the loss of chromatin modifiers

An examination of the dysregulated genes in the Ark2 EC-derived cell line, following knockouts of six different chromatin modifiers, revealed a commonality of 132 up-regulated genes and 99 down-regulated genes across all conditions (on average, differential expression analyses revealed 1,101 up-regulated and 1,099 down-regulated genes for each condition). Despite the unique dysregulation profiles stemming from the loss of individual chromatin modifiers, several pathways, including KRAS signaling, consistently exhibited up-regulation in all knockouts (Fig. 3.42). As previously discussed, the activation of KRAS signaling in cancer is generally linked to heightened tumor aggressiveness and metastasis [298, 299]. Ark2 is a HER2-positive cell line, with HER2 being encoded by *ERBB2* and recognized as a KRAS pathway activator [13]. This observation suggests that the activation of the KRAS signaling pathway may be further amplified following the loss of any of the studied chromatin modifiers.

Another frequently up-regulated pathway involves the tumor necrosis factor-alpha (TNF-α) signaling pathway (Fig. 3.42), which is mediated by the activation of nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB). This implies a potential

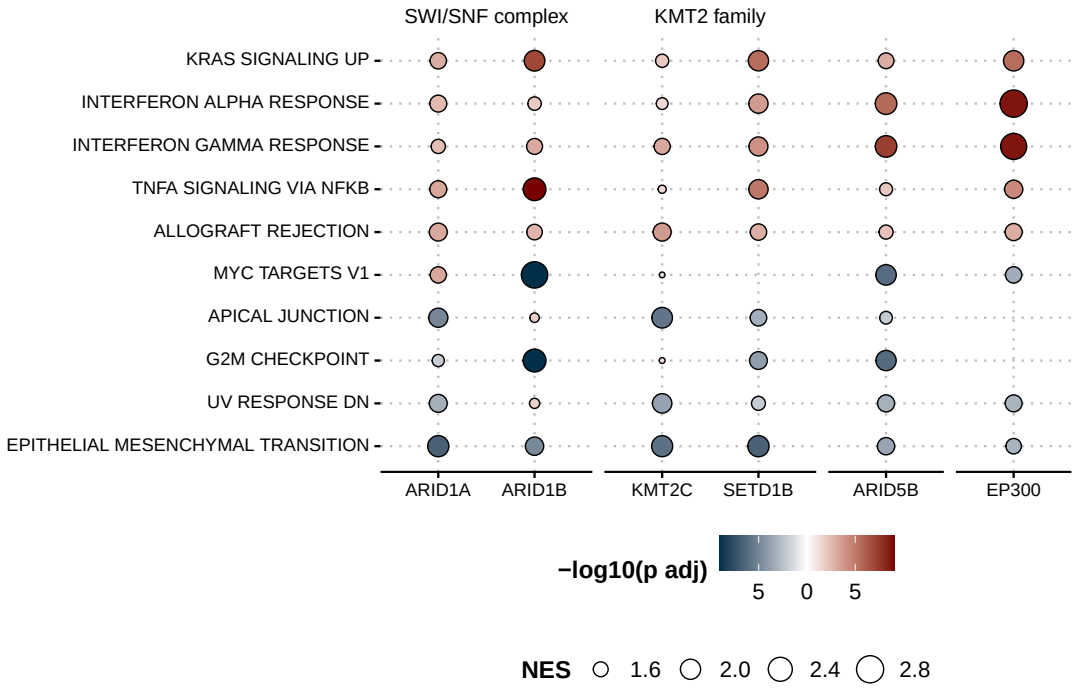


Figure 3.42: **Dotplot showing enrichment of the Hallmark pathways across clones.** Results of GSEA (adj. p-value < 0.05) on Hallmark pathways shown. Normalized Enrichment Scores (NES) for each pathway are calculated using GSEA, enabling comparison across pathways. Positive scores (in red) indicate up-regulation, while negative scores (in blue) signify down-regulation. The dot color intensity represents the adjusted p-value, while the dot size corresponds to the NES. Refer to Section 3.4.1.2 for a description of the gene set enrichment analysis.

role for the TNF-α/NF-κB signaling axis in response to perturbed chromatin organization. By activating NF-κB, anti-apoptotic and pro-proliferative genes may be expressed, thereby promoting tumor cell survival and growth.

Other frequently up-regulated pathways include interferon (INF)-α and INF-γ response (Fig. 3.42). INF-α and INF-γ response activation can exhibit antitumor effects by suppressing cell proliferation, encouraging apoptosis, and enhancing the immune response against cancer cells [296]. Conversely, sustained activation of INF-α and INF-γ may contribute to cancer development by facilitating inflammation, angiogenesis, and immune evasion [297]. The activation of interferon responses following the loss of chromatin modifiers could represent a compensatory mechanism aimed at counterbalancing the potential tumor-promoting effects.

Deciphering allograft rejection, another commonly up-regulated pathway (Fig. 3.42), within the context of a tumor proves to be more intricate. The up-regulation of allograft rejection pathways in a cancer cell line could stem from the cell line's adaptation to *in vitro* culture conditions. As a cell line is established and maintained, various genetic and epigenetic alterations may occur, potentially activating specific pathways, including those associated with immune responses. The observed up-regulation of the allograft rejection pathway might not bear direct biological relevance to the experiment itself; instead, it could be an artifact arising from experimental conditions or data analysis. For example, the pathway could encompass genes with other pertinent functions. To confirm the significance of these findings, further validation and evaluation of their biological relevance would be required.

While EMT is down-regulated across all conditions, the situation is more intricate when examining individual members of its pathway, as depicted in Fig. 3.43. Certain genes, such as *TGFBI* and *IGFBP3*, consistently display down-regulation across all conditions. Conversely, several genes demonstrate recurrent up-regulation among the investigated knockouts. For instance, *SPOCK1*, which is up-regulated in all cases except for the *SETD1B* knockout, has been previously discussed in the context of ARID1s loss and found to promote breast cancer progression [313], pancreatic cancer metastasis [314], and colon

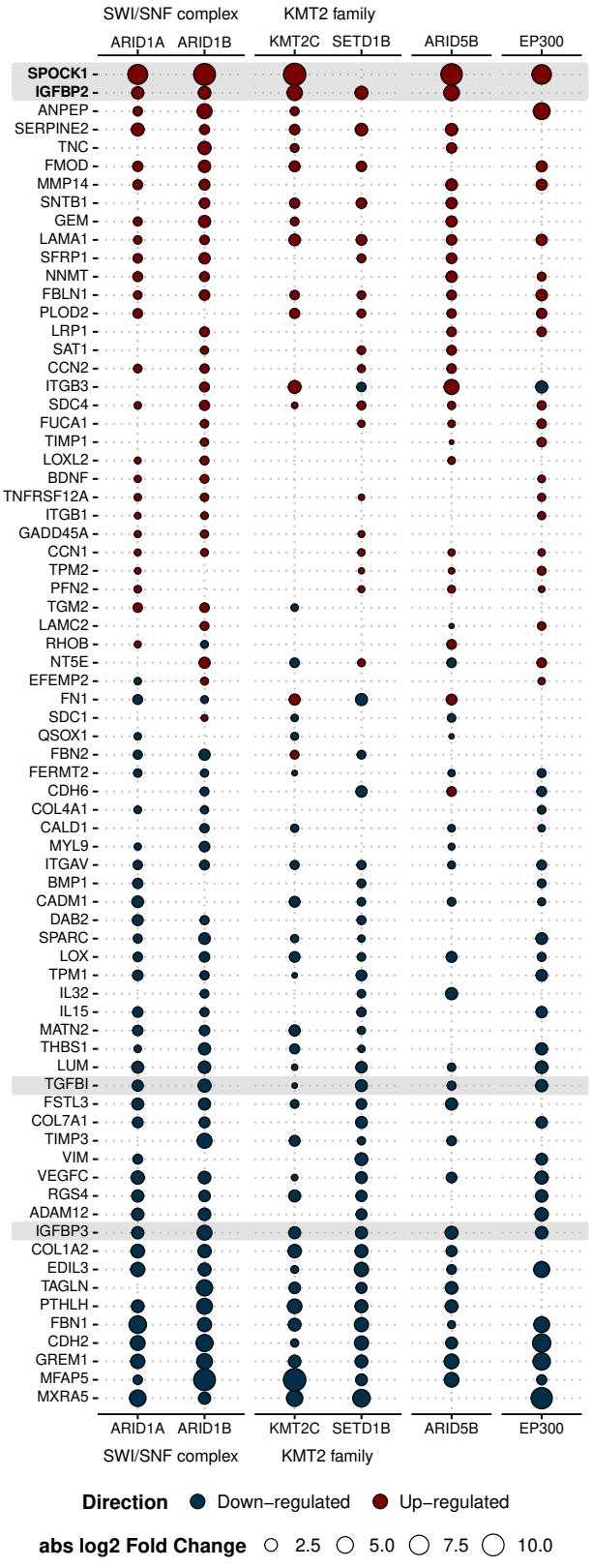


Figure 3.43: **Dysregulation of members of EMT pathway across chromatin modifiers knockouts.** Genes significantly deregulated across at least 3 conditions shown. Genes discussed in more detail in the text highlighted. Red color indicates up-regulation, while blue signifies down-regulation. The dot size corresponds to the absolute $\log_2$ fold change.

cancer proliferation and migration [315]. Another gene, *IGFBP2* (insulin-like growth factor binding protein 2), is up-regulated in all knockouts except for Ark2 Cas9 $EP300^{KO}$. Recent studies have identified the up-regulation of *IGFBP2* as a critical factor in ovarian cancer pathogenesis [316]. These findings suggest that the loss of chromatin modifiers may prompt cancer cells to adapt and exploit specific pathways to their advantage.

Subsequently, I examined the transcription factors across chromatin modifiers. Utilizing the Kruskal-Wallis nonparametric test (through the ChromVar package [260]), I identified differentially accessible transcription factor binding sites (TFBS) across the chromatin modifier knockouts. I visualized the accessibility of these motif binding sites in Fig. 3.44. After visually inspecting the dendrogram, I divided the motifs into five clusters.

Cluster 1, which comprises motifs from the C4-Gata-related family, exhibits decreased accessibility of motifs across chromatin modifier knockouts, with the most significant decrease observed in the context of KMT2C loss. Cluster 2 reveals increased accessibility of CEBP-related TFBS in Ark2 Cas9 $SETD1B^{KO}$, Ark2 Cas9 $ARID1A^{KO}$, and Ark2 Cas9 $EP300^{KO}$, particularly. The heightened accessibility in Ark2 Cas9 $EP300^{KO}$ could imply increased activation of CBP, a transcriptional coactivator closely related to p300, as a compensatory mechanism for the loss of EP300. Interestingly, this cluster also encompasses ESR1 TFBS, which displays increased accessibility in Ark2 Cas9 $SETD1B^{KO}$, Ark2 Cas9 $ARID1A^{KO}$, and Ark2 Cas9 $EP300^{KO}$ loss, but decreased accessibility in Ark2 Cas9 $KM2TC^{KO}$, Ark2 Cas9 $ARID5B^{KO}$, and Ark2 Cas9 $ARID1B^{KO}$. The increased accessibility of ESR1 motifs does not seem to correlate with the dysregulation of estrogen or androgen response pathways.

Cluster 3 is distinguished by the enhanced accessibility of FOX TFBS in all knockouts except for Ark2 Cas9 $KM2TC^{KO}$. In the case of Ark2 Cas9 $ARID1A^{KO}$ and Ark2 Cas9 $ARID5B^{KO}$ knockouts, CUXs TFBS display increased accessibility. I previously elaborated on the role of CUXs transcription factors in relation to the presence of CUXs TFBS near up-regulated genes in the Ark2 Cas9 $ARID1A^{KO}$. The loss of *ARID5B* appears to similarly expose additional CUX TFBS, thereby augmenting the regulation of gene expression via CUX transcription factors.

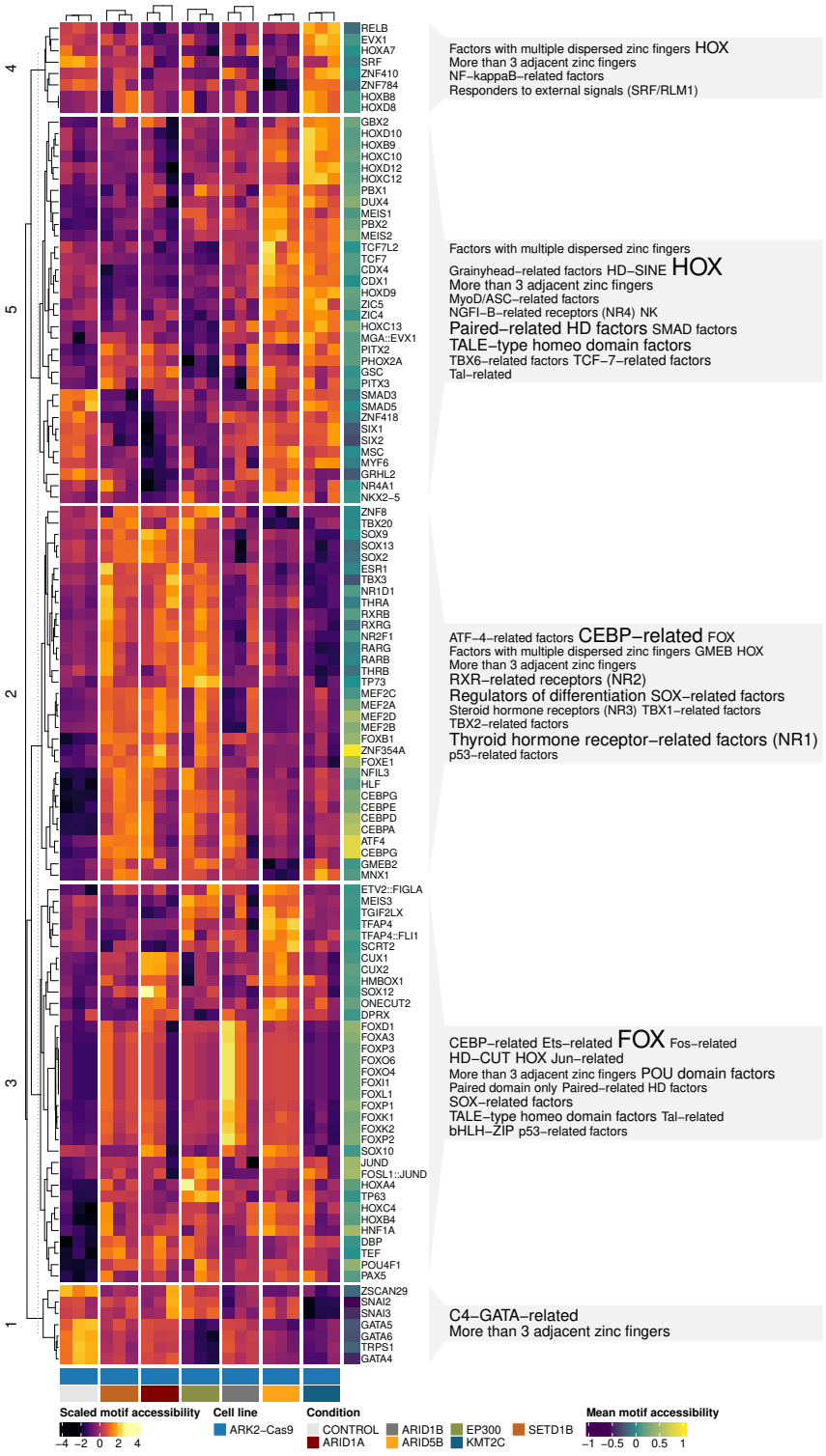


Figure 3.44: **Heatmap of motif accessibility scores across Ark2 Cas9 clones.** Heatmap of bias-corrected and normalized accessibility of motifs. Values in rows were standardized (scaled), and the row means were plotted aside to allow for the comparison of differences and the overall accessibility of motifs. For more details on enrichment analysis, see Section 3.4.1.2; for clustering and visualization, see Section 3.4.1.2.

TFBS of Hox family, notably HOXB8, HOXD8, and HOXA7, are found in cluster 4. This cluster is characterized by lower accessibility across conditions, except for Ark2 Cas9 $KM2TC^{KO}$, where accessibility of sites of transcription factors in this cluster is increased. Although the role of those transcription factors in EC remains undefined, HOXB8 was found to enhance the proliferation and metastasis of colorectal cancer [317], while HOXA7 was found to promote metastasis of KRAS mutant colorectal cancer [318].

Cluster 5 consists of more HOX family TFBS with increased accessibility in Ark2 Cas9 $ARID5B^{KO}$ and Ark2 Cas9 $KM2TC^{KO}$ compared to Ark2 Cas9 parental cells. Interestingly, SMAD3 and SMAD5 TFBS, also members of cluster 5, are found more accessible in Ark2 Cas9 $ARID5B^{KO}$ and Ark2 Cas9 $KM2TC^{KO}$ but show decreased accessibility across other knockouts compared to the Ark2 Cas9 parental cells. *SMAD3*, but not *SMAD5*, expression is up-regulated across all conditions. Notably, the inactivation of *Smad2* together with *Smad3* induced uterine cancer in mice [319].

3.2.11 Functional study of *BCOR* N1459S hotspot mutation

As discussed in Section 2.3.2.2, *BCOR* is frequently altered in various cancers. However, the *BCOR* N1459S point mutation found frequently in EC (9.7% samples in EC GEL cohort carry this mutation) is almost exclusive to endometrial cancer. Investigation of the correlates of *BCOR* N1459S in EC using gene expression and genomic data from EC TCGA cohort revealed an enrichment of estrogen and androgen response pathways among up-regulated gene sets in patients with N1459S *BCOR* mutation in comparison to those without *BCOR* alteration [320]. BCOR is a member of a non-canonical polycomb repressive complex 1 (PRC 1.1). PRC 1.1 has been found to repress differentiation programs in human embryonic stem cells [120].

There is 89.04% similarity between human BCOR and mouse Bcor protein sequences (calculated with Clustal Omega [321]). N1459S hotspot mutation in human *BCOR* corresponds to N1463S in mouse *Bcor*, with the sequence exactly the same between the species within 20 amino acids of the mutation as shown in Fig. 3.45. To observe the consequences of the mutation on differentiation programs, we designed a functional study in mouse em-

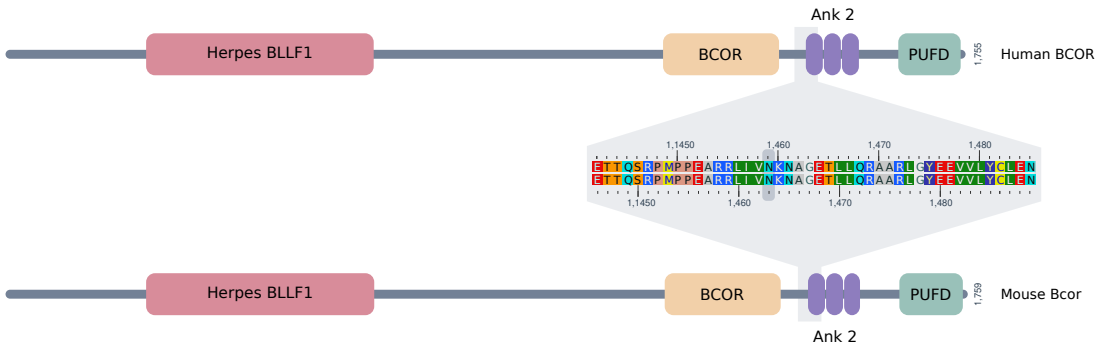


Figure 3.45: **Comparison of human BCOR and mouse Bcor protein sequences.** The region around mutation in question, N1459S in human BCOR, highlighted. Figure created with Inkscape.

bryonic stem cells (mES) where we introduced the N1463S mutation to model the N1459S human hotspot (mES $Bcor^{N1463S}$) and mES $Bcor^{KO}$ (Fig. 3.46). With this design, we obtained an isogenic cell line model of *BCOR* loss and N1459S hotspot mutation. Cells were kept at early passage under conditions optimized for stem cells, and expression profiling of the clones under basal conditions was performed by RNA-seq.

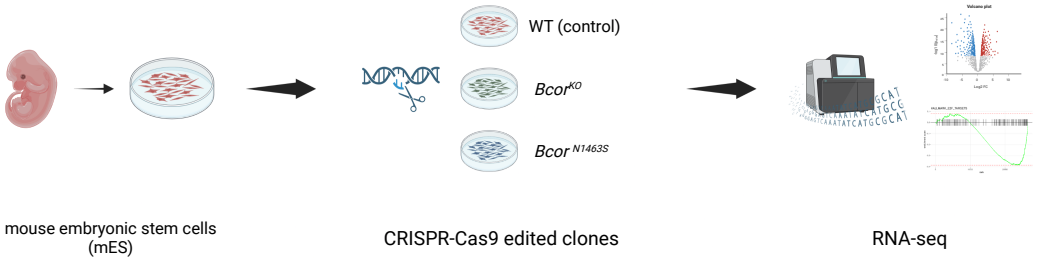


Figure 3.46: **Illustration of the functional experiment to understand the consequences of *BCOR* N1459S hotspot mutation.** N1463S *Bcor* mutation was introduced to model the N1459S *BCOR* - hotspot mutation found in endometrial cancer. Knockout (mES $Bcor^{KO}$) was achieved by 120bp deletion and insertion. Figure created with BioRender.com

Preliminary assessment of expression data revealed significant batch effect as shown in Fig. 3.47A. I processed the data, including merging the technical replicates and removing batch effect (Fig. 3.47B). Surprisingly, subsequent analysis of expression of *Bcor* and *Bcor* associated genes did not reveal obvious differences between either mES $Bcor^{N1463S}$ or mES $Bcor^{KO}$ and parental cell lines (Fig. 3.47C). This may result from the detection of *Bcor* alternative transcripts, as protein loss in knockouts had previously been confirmed by western blotting. Acute knockdown of *BCOR* expression resulted in increased expression

of *POU5F1* and *NANOG* and decreased expression of *SOX2* in human embryonic stem cells (hESC) [120]. However, in our data, there was no observable change in the expression of *Pou5f1* and *Nanog* in mES $Bcor^{KO}$ compared to the parental cell line. Expression of *Sox2* was slightly increased (FC = 1.23, p adj. = 0.015). Expression of *Foxa2* was increased in mES $Bcor^{KO}$ (FC = 2.83, p adj. = 0.040), whereas its down-regulation was observed in hESC cells. I also observed high disagreement in differential expression genes between mES $Bcor^{N1463S}$ clones (Fig. 3.47D).

I further investigated the dysregulation of pathways in mES $Bcor^{N1463S}$ compared to the parental cell line and, as with the differentially expressed genes, there was little agreement in the results between the clones (Fig. 3.50, Fig. 3.51). I found pathways related to interferon response and TNF-α signalling via NF-κB among the most up-regulated gene sets (Fig. 3.48, Fig. 3.50). Most significant down-regulated pathways included KRAS signaling (Fig. 3.49, Fig. 3.51). However, pathways and genes related to stemness were not identified among the significantly disrupted.

DNA repair has also been found up-regulated across all but one mES $Bcor^{N1463S}$ clone. Observing the increased activity of the DNA repair pathway might be an effect of the cells being incubated with MEK/ERK pathway inhibitor (PD0325901) and GSK3 inhibitor (CHIR99021) to arrest them in undifferentiated mouse embryonic stem cell state.

Overall, the results from these experiments are difficult to interpret and inconclusive. Further experiments are necessary to address the biological significance of *BCOR* N1459S mutation. To observe differences in differentiation, a time course experiment in conditions allowing for differentiation should be performed. It is also possible that the effect of N1459S mutation is dependent on cellular processes unique or enriched in endometrial tissue and that the effects of the mutation should be investigated in human endometrial cell lines or organoids, or a mouse knock-in should be performed. Additionally, further investigation should be performed to exclude the mutation's passenger character. The sequence of *BCOR* is located on the X chromosome, which in females undergoes a process referred to as X chromosome inactivation (XCI) [322]. XCI is a process to repress the expression of genes from one copy of the X chromosome in order to maintain the sex-

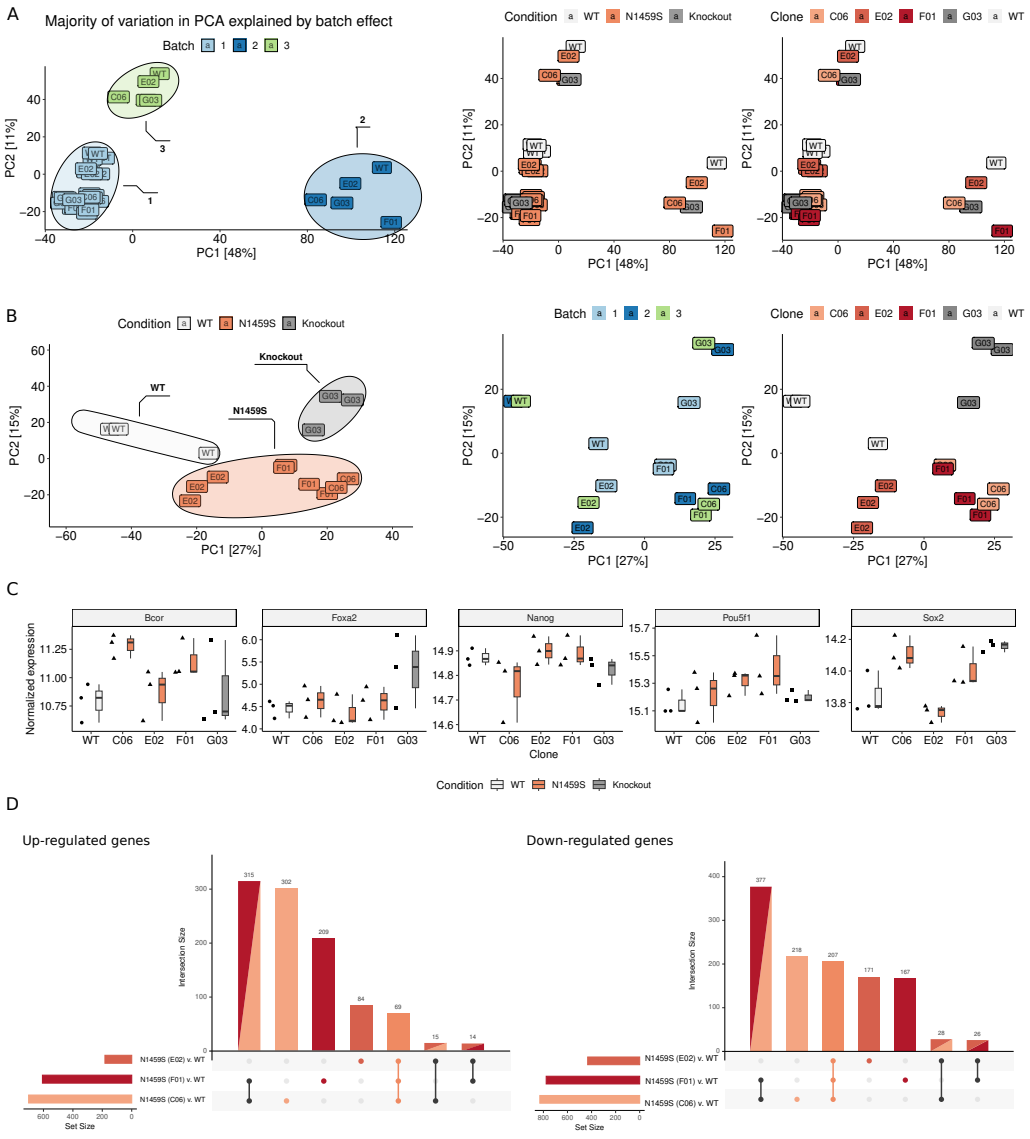


Figure 3.47: **Quality assessment of the RNA-seq data from the isogenic mES *Bcor*^{N1463S} and mES *Bcor*^{KO} cell lines.** WT denotes mES parental cell line, N1459S denotes the mES *Bcor*^{N1463S} and knockout – mES *Bcor*^{KO}. **A** Principal component analysis (PCA) of normalized counts; the plots show separation across PC1 and PC2 based on batch effect. **B** PCA of normalized counts with batch effect removed and technical replicates merged; the batch effect is no longer present among top principal components. **C** Normalized expression of selected genes. Although *Bcor* expression remained at the same level as in the parental cell line, the knockout was confirmed with Western blot. **D** Overlap of genes identified as deferentially expressed (FC $>$ 1.5 and FC $<$ 0.67, adjusted p-value $<$ 0.05) across mES *Bcor*^{N1463S} (N1459S) clones.

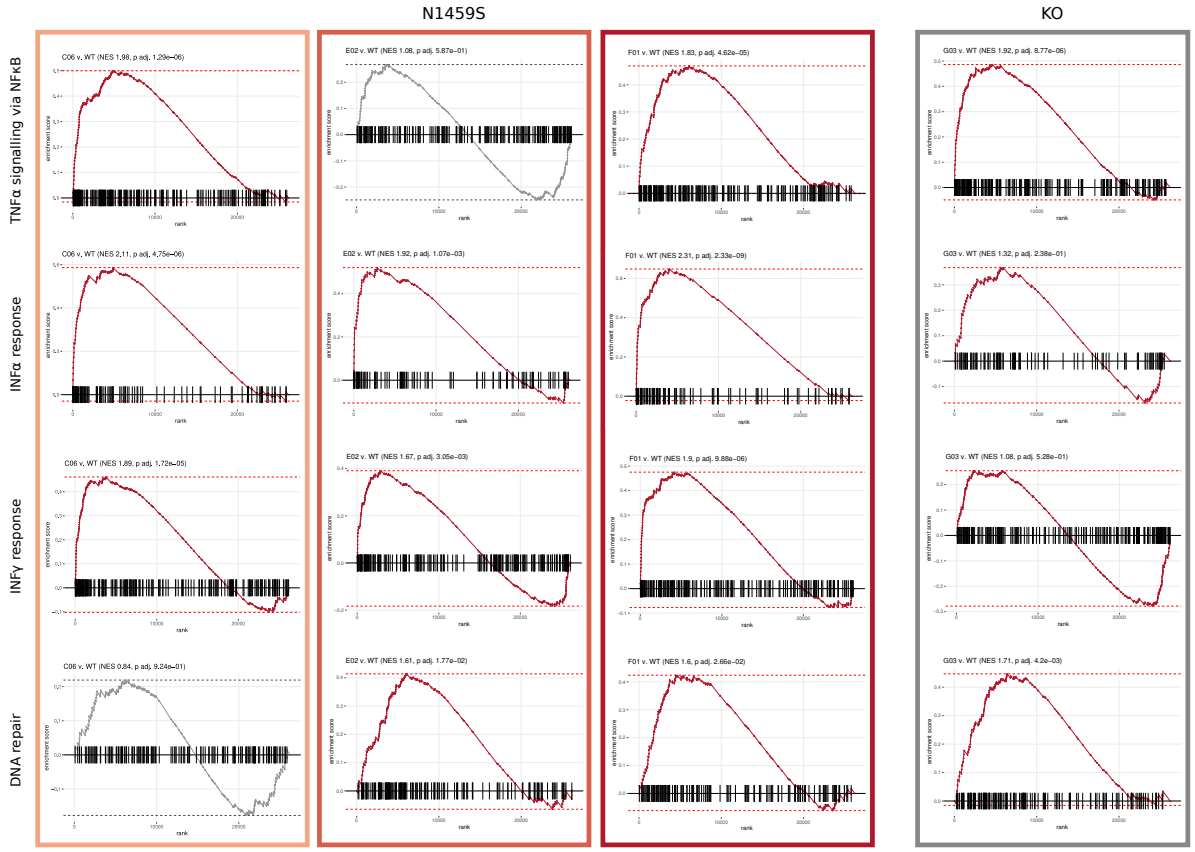


Figure 3.48: **Gene sets up-regulated in mES *Bcor*^{N1463S} (N1459S) clones.** GSEA plots for selected pathways across *Bcor* clones. Plots with non-significant signal plotted in gray. The color scheme for clones retained from Fig. 3.47.

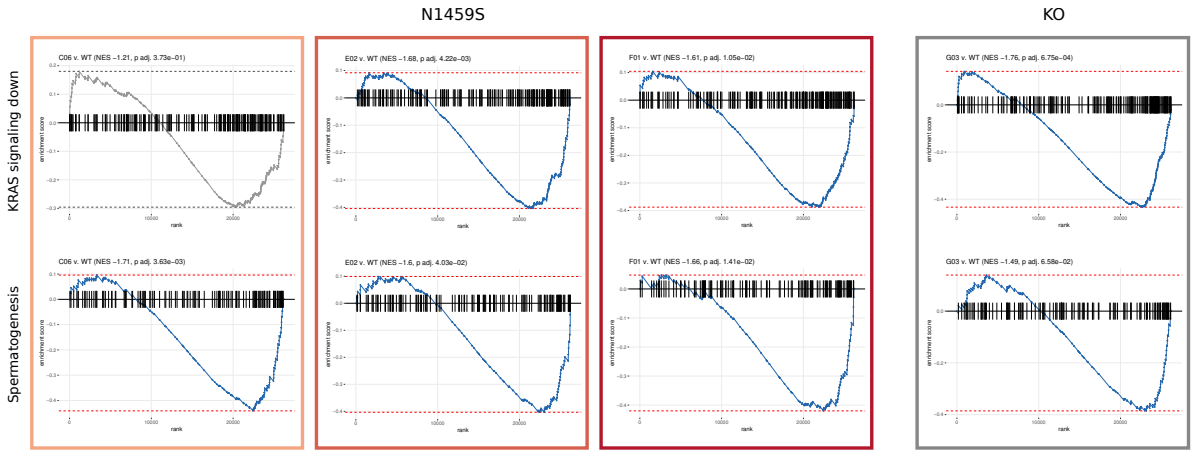


Figure 3.49: **Gene sets down-regulated in mES *Bcor*^{N1463S} (N1459S) clones.** GSEA plots for selected pathways across clones. Plots with non-significant signal plotted in gray. The color scheme for clones retained from Fig. 3.47.

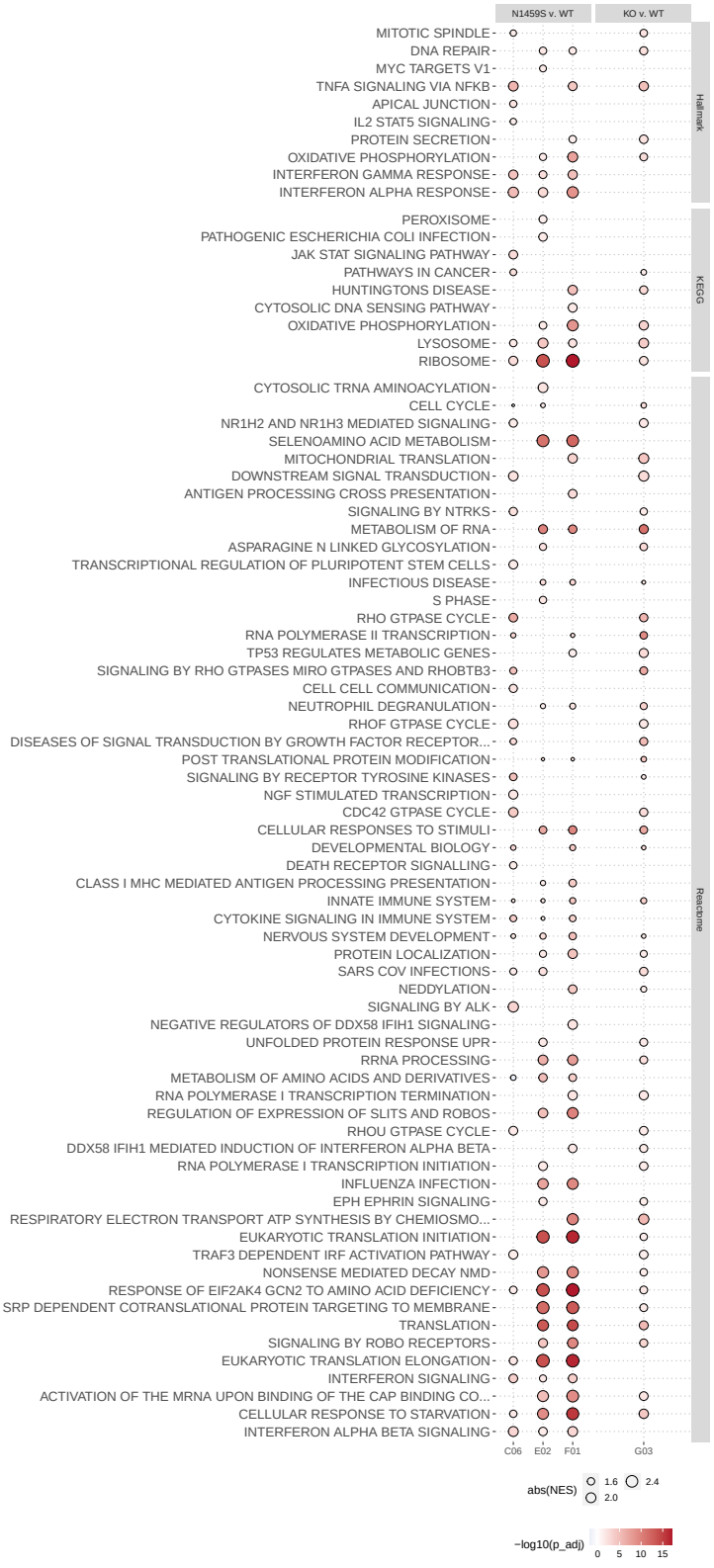


Figure 3.50: **Gene sets up-regulated in N1459S clones.** Results of gene set enrichment analysis (GSEA). Gene sets are significantly dysregulated in at least one clone shown.



Figure 3.51: **Gene sets down-regulated in N1459S clones.** Results of gene set enrichment analysis (GSEA). Gene sets significantly dysregulated in at least one clone shown.

chromosome dosage compensation. Interestingly, non-canonical PRC1 has been found to interact with X-linked X-inactive-specific transcript (Xist) - a long non-coding RNA that initiates XCI [323]. I was able to confirm the N1459S mutation in one sample from RNA-seq available in EC TCGA dataset. However, not enough data was available to perform statistical analysis. A further experiment with access to genotyped patient samples (such as those in the EC GEL dataset) would allow verifying the expression of a mutated copy.

3.3 Discussion

In this chapter, I report the results of functional work to understand the consequences of the loss of chromatin modifiers in endometrial cancer. Motivated and informed by my descriptive analyses of alterations of chromatin modifiers in the GEL EC cohort, I aimed to understand the role of those chromatin modifiers in endometrial cancer cell biology. Although *ARID1A*, and to a lesser extent its paralog *ARID1B*, have been studied in endometrial cancer models, little if anything is known about the consequences of mutations in *ARID5B*, *EP300*, *KMT2C*, *SETD1B*. To the best of my knowledge, this study was the first joint attempt to understand the implications of the loss of each of the six chromatin modifiers. We also looked at the consequences of the loss of *ARID1A* to the normal endometrium in an isogenic cell line model derived from hEM3 - normal endometrium derived cell line. We also attempted to understand the implications of the near-exclusive to EC *BCOR* hotspot mutation.

With our findings, we were able to reproduce several findings previously reported in the literature. For example, p300 and CBP are closely related transcriptional coactivators responsible for regulating gene expression via altering acetylation levels of histone proteins. Wilson and Reske et al. [324] find that *ARID1A* mutations promote p300-dependent endometrial invasion, while we observe the up-regulation of CBP protein in the loss of *ARID1A* in the hEM3 cell line.

For the first time, we describe the gene expression and chromatin accessibility profiles as an effect of chromatin modifiers loss. The investigation of chromatin modifier knock-outs revealed dysregulation of critical pathways such as KRAS signaling, TNF-α/NF-κB

signaling, interferon responses, and the down-regulation of EMT. Our findings underscore the consequences that loss of chromatin modifiers has tot the chromatin accessibility and gene expression profiles, as well as the potential compensatory mechanisms cancer cells employ in response to the loss of chromatin modifiers.

The results presented in this study highlight significant effects of *ARID1A* loss on hormonal regulation and transcription factor accessibility in normal endometrial cell lines. Notably, *ARID1A* loss led to the up-regulation of androgen response and down-regulation of estrogen response in the normal endometrium, implicating a potential mechanism for endometrial hyperplasia. The altered hormonal sensitivity may contribute to disorders like PCOS and endometriosis, conditions closely linked to hormone imbalances [254–256]. Similarly, the down-regulation of estrogen response mirrors findings in breast cancer [257].

Differential motif accessibility analyses uncovered increased accessibility for CUX1, CUX2, and PITX3 motifs in *ARID1A*-deficient EC-derived cells, pointing towards altered transcriptional regulation. The roles of CUX1 and CUX2 as oncogenes or tumor suppressors vary across cancer types, indicating a complex interplay with cancer pathology [265–268]. The presence of these motifs in regulatory regions associated with *SPOCK1* and lncRNA suggests a potential mechanism by which *ARID1A* loss might influence gene expression patterns critical for cancer progression, including EMT and invasion [271–275].

The perturbation of the EMT pathway following *ARID1B* loss in EC-derived cell lines reveals a nuanced modulation of gene expression. Up-regulated and down-regulated genes within this pathway, including the notable up-regulation of *SERPINE2*, highlight the complex role of *ARID1B* in EMT regulation. Induction of *SERPINE2*, known to be stimulated by the activated MAP kinase pathway in various contexts [288], has been implicated in the progression from normal tissue to adenomas and further to colorectal cancer [289].

Moreover, the differential expression within the TGF-β pathway, particularly the up-regulation of *TGIF1*, *SMURF2*, and *SERPINE1*, alongside the down-regulation of *B2M* and *APC*, indicates a selective modulation of this pathway by *ARID1B* loss. Overexpression of *SERPINE1*, akin to *SERPINE2*, underscores its role in promoting malignant

progression, similar to that in gastric cancer [290] and as a marker in endometriosis [291].

Loss of *EP300* leads to decreased EMT pathway gene accessibility and expression, alongside an increase in KRAS pathway activity. This dual impact highlights *EP300* as a critical regulator of tumorigenic pathways in EC, with the enhanced KRAS signaling via increased JunD motif accessibility suggesting a potential link between *EP300* loss and oncogenic signaling. Although research on the specific role of JunD in EC is limited, JunD was identified as a crucial component for Ras-induced tumorigenesis in lung cancer [300].

The *KMT2C* KO experiment revealed an up-regulation of E2F target genes. Overexpression of E2F targets is associated with poor prognosis in liver [308] and pancreatic [309] cancers. This finding suggests that *KMT2C* loss might confer a proliferative advantage to EC cells. Furthermore, loss of *KMT2C* also results in the down-regulation of AR. The antiproliferative effect of androgen, as a result of androgens counteracting estrogen-driven pro-proliferative signaling in the endometrium, has been observed in a well-differentiated endometrial adenocarcinoma cell line [310]. A recent study on breast cancer showed that the AR can displace ER and other transcriptional coactivators from estrogen response elements, resulting in the down-regulation of ER-regulated cell cycle genes and reduced breast cancer cell proliferation [311].

The impact of *SETD1B* loss on cell cycle regulation contrasts with the effects observed following *KMT2C* loss, notably in the modulation of E2F targets. The down-regulation of E2F targets following *SETD1B* loss suggests a distinct regulatory influence compared to the up-regulation seen with *KMT2C* knockout. This divergence highlights the complex role of chromatin modifiers in cell proliferation and cycle control. Interestingly, despite the overall contrasting impact on E2F targets, specific genes such as *CDKN1A* exhibit up-regulation. Up-regulation *CDKN1A* has been linked to an attempt to arrest cell cycle progression and limit uncontrolled division [312]. Meanwhile, the down-regulation of *CDKN1B* further complicates the landscape of cell cycle regulation and suggests an intricate balance between promoting and inhibiting cell proliferation.

The consistent up-regulation of KRAS signaling across various chromatin modifier

knockouts underscores the critical role of this pathway in cancer progression. Such uniform activation suggests that the KRAS pathway is a pivotal element in the malignancy of EC-derived cell lines. Activation of KRAS signaling is generally linked to tumor aggressiveness and metastasis [298, 299]. Additionally, HER2 positivity in the Ark2 cell line, with *ERBB2* as a known activator of KRAS signaling [13], may amplify these effects, reinforcing the significance of KRAS pathway activation in the context of chromatin modifier loss on the HER2 positive background.

The nuanced modulation of the EMT pathway, with certain genes consistently down-regulated across knockouts, suggests a broad yet selective impact of chromatin modifier loss on cancer progression. *SPOCK1* has been found up-regulated in all conditions except Ark2 Cas9 $SETD1B^{KO}$. *SPOCK1* has been associated with cancer progression in breast [313], pancreatic [314], and colon [315] cancers. Similarly, *IGFBP2* has been up-regulated in all conditions except in Ark2 Cas9 $EP300^{KO}$. *IGFB2* have been implicated in ovarian cancer pathogenesis [316].

We have generated a valuable resource that can be extended beyond the analyses conducted in this thesis. For instance, utilizing the chromatin accessibility profiles obtained in this experiment, one could investigate differential transcription factor (TF) footprinting. TF footprinting is a computational analysis that detects the occupancy of binding sites by transcription factors. Through this analysis, we could further examine the changes in transcription factor recruitment to chromatin upon the loss of a chromatin modifier.

In the current analyses, I associated accessibility regions with the nearest gene. A more biologically informed approach, incorporating information about topologically associating domains (TADs), could refine the annotation and elucidate the relationship between accessibility and expression regulation. For differentially accessible chromatin and TFBS, I relied on the curated regions defined by the TCGA study across various tissues [325]. Enhancing the resolution of the analyses could involve extending the regions to the consensus of peaks found across all conditions in this experiment.

While the study sheds light on the effects of perturbations in chromatin modifiers, it has been limited to one cancer-derived cell line, thereby restricting the understanding

of the broader molecular implications. To gain deeper insights and validate the findings would require increasing the representation of endometrial-cancer-derived cell lines used for the experiment. Performing similar experiments in EC-derived cell lines with diverse backgrounds could help confirm the directionality of changes resulting from the loss of chromatin modifiers. Moreover, following up the study with DNA-binding experiments, for example, CUT & RUN [326] or CUT & Tag [327] could help understand the dynamics of the interaction of the chromatin modifiers with the DNA sequence.

Another limitation of this research is the reliance on CRISPR experiments in established cell lines, which may not fully recapitulate the complex *in vivo* tumor microenvironment or the genetic diversity found in primary tumors. Using primary cells for CRISPR experiments could potentially offer more physiologically relevant insights but also introduce challenges related to variability and accessibility. Alternatively, the use of endometrial organoids could be explored. However, initial tries in the Church group have proven challenging as, in contrast to color organoids, endometrial organoids require many more mice, making this technically challenging.

Additional experiments, such as proliferation assays, would provide invaluable context to the omics assay results. Furthermore, a rescue phenotype experiment would offer invaluable validation of the phenotype observed in our results.

Furthermore, to comprehensively understand the effects of common mutations in chromatin modifiers and beyond, it would be beneficial to conduct high-throughput single-cell CRISPR screens, like Perturb-ATAC [328], sci-ATAC [329], or Spear-ATAC [330] followed by in-depth experiments. Those assays can accomplish this by using cells that express Cas9, a library of gRNAs, and simultaneous readout of chromatin accessibility and gRNA detection at the single-cell level. These experiments can investigate chromatin accessibility perturbations resulting from knockdowns of specific genes, including common mutations like prevalent hotspots and gain-of-function mutations. Further, specialized investigations of selected mutations can further our understanding of their impact on chromatin modifiers and EC.

Mutations in *ARID1A* are shared between endometrial cancer and endometriosis while

largely absent from normal endometrium [34, 35]. In our study, we identified additional potential links between the dysregulation of chromatin modifiers in EC and its consequences with the markers of endometriosis - such as overexpression of *SERPINE1* following the loss of *ARID1B*. However, more research is needed to verify the similarities between the role of chromatin modifiers in both diseases and the potential role of those modifiers in the transition from endometriosis to ovarian cancer. A functional study of dysregulation of chromatin modifiers in the endometriotic, ovarian cancer, and endometrial cancer cell lines would help further understand the role of chromatin modifiers in disease progression.

Another avenue worth exploring is investigating the combination of mutations. As demonstrated in research, chromatin modifier mutations rarely occur in isolation. A targeted experiment focusing on a double-knockout of *ARID1A* and *ARID1B* can serve as an excellent starting point.

3.3.1 Functional study of *BCOR* N1459S hotspot mutation

The experiment investigating the N1459S hotspot mutation in *BCOR* yielded inconclusive results. This could be attributed to the fact that the mES cells used in the research appeared to recover from the perturbations to the chromatin. In another study of BCOR-PRC1.1 complex, authors of [120] performed a similar experiment where they investigated the consequences of the inactivation of *BCOR* gene in human embryonic stem cells (hES). The authors of the study employed small hairpin RNA (shRNA) and demonstrated that the shBCOR cells were visibly dissimilar from the control cells. The shBCOR cells spent a longer duration in the G1 phase, while the S phase was reduced. The study also showed that the depletion of BCOR prompted endodermal and mesodermal differentiation of hES. It is feasible that the effects of BCOR loss may not be observable after a delay in measurement. To tackle this, an alternate system for introduction of the the *Bcor* point mutation, such as a tetracycline-inducible system, could be explored. However, this may not capture the intricacies of gene dosage. To detect differences in differentiation, a time course experiment under differentiation-permitting conditions could help understand the role of *BCOR* and its mutation. Additionally, the use of chromatin accessibility or DNA-

protein binding techniques, such as ChIP-seq or CUT & RUN [326], could illuminate the implications of the mutation. It is also conceivable that the effect of the N1459S mutation depends on cellular processes that are unique or abundant in endometrial tissue. To investigate this, the mutation's effects should be studied in human endometrial cell lines, organoids, or a mouse knock-in should be performed.

3.4 Methods

3.4.1 Chromatin modifiers in hEM3 and Ark2 cells

3.4.1.1 Experimental work

hEM3 cell lines were acquired from the authors [244]. Ark2 Cas9 $ARID1A^{KO}$ cell line generated with CRISPR Cas9 was described in [246].

CRISPR-Cas9 First, Ark2 Cas9 stably expressing cell line was established. Next, for each of the chromatin modifiers, two guide sequences were designed according to the following rules: (1) guide target not in the first exon, (2) not in the last exons, (3) in an exon present in all isoforms, (4) each guide in separate exon and (5) with an IDT an on-target score > 40 and an off-target score > 60. Several clones per condition were isolated, and the best clone per condition was chosen based on Sanger sequencing and qPCR results. Three separate passages were used as biological replicates. RNA for RNA-seq library preparation was extracted with a Qiagen RNeasy kit. Cell pellets were frozen for the ATAC-seq experiment. I describe the steps I took to validate the CRIPSR-Cas9 experiments in Section 3.2.2.

RNA-seq The RNA-seq experiment has been conducted by RNA-seq services (Diagenode Cat# G02030000). RNAs have been firstly quantified using Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Q32852) and secondly checked for integrity using RNA 6000 Pico Kit (5067-1513, Agilent) on a 2100 Bioanalyzer system (Agilent). Library preparation has been done with 15ng of input RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB# E7490) followed by NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB# E7760) and NEBNext Multiplex Oligos for Illumina Index Primers (NEB# E6440). The generated DNA libraries were purified using Agencourt AMPure XP (Beckman Coulter). Purified libraries were quantified, and their size was assessed with QIAxcel (Qiagen). The libraries were pooled and sequenced on the NovaSeq (paired-end, 50 bp).

ATAC-seq The ATAC-seq experiment has been conducted byDiagenode ATAC-seq Profiling service (Diagenode Cat# G02060000). **Nuclei isolation optimization** Cell viability and cell counting of the test samples were performed via microscope (AE2000; Matic). Cells were centrifuged at 500g before proceeding with the nuclei isolation optimization, testing two different conditions for the Lysis buffer 1 composition(0,1% and 0,4% of Tween20/Igepal) using the reagents of the ATAC-seq Kit from Diagenode (Diagenode Cat# C01080002). **Transposition** Cell viability and cell counting were performed manually. 100.000 cells were centrifuged at 500g before proceeding with the nuclei isolation and transposition reaction using the ATAC-seq Kit from Diagenode (Diagenode Cat# C01080002). According to the nuclei isolation optimization results, the Tween20/Igepal concentration in Lysis buffer 1 (0,1% originally) has been increased to 0,4%. 50.000 isolated nuclei were lysed and transposed for 30 minutesat $37°C$ using the loaded Tagmentase (Diagenode Cat# C01070012).Transposed DNA was then purified on Diapure columns (Diagenode Cat# C03040001). **Library preparation** The Diagenode ATAC-seq Profiling service has conducted the ATAC-seq library preparation (Diagenode Cat# G02060000). Libraries were prepared from purified transposed DNA using NEBNext High-Fidelity PCR MasterMix (NEB, M0541) and Illumina indexing primers. The number of library amplification cycles was determined by qPCR analysis using NEBNext High-Fidelity PCR MasterMix (NEB, M0541) on LightCycler96 System (Roche). After their amplification, the libraries were size selected and purified using Agencourt AMPure XP (Beckman Coulter) and quantified using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Q32854). Finally, their fragment size was analyzed by High Sensitivity NGS Fragment Analysis Kit (DNF-474) on a Fragment Analyzer (Advanced Analytical Technologies, Inc.) The libraries were pooled and sequenced on the NovaSeq (paired-end, 50 bp).

3.4.1.2 Computational analyses

RNA-seq Raw reads were processed using custom Snakemake pipeline [331]. Briefly, the quality of the reads was assessed with fastqc [332], adaptor sequences were trimmed

with cutadapt [333] using trim_galore [334]. Expression was quantified with salmon [335], and processed reads were then aligned to hg38 using STAR aligner [336]. Quality metrics were assembled and assessed with multiqc [337]. Upon investigation of unsupervised clustering and PCA of the biological replicates, one outlier Ark2 Cas9 $ARID5B^{KO}$ replicate was removed from the analyses. Differential expression analysis was performed with DESeq2 [248]. Genes were annotated as up-regulated with FC > 1.5 and adjusted p-values < 0.05, and down-regulated with FC < 0.67 adjusted p-values < 0.05. P-values were adjusted using Benjamini-Hochberg false discovery rate correction procedures as described in [248].

ATAC-seq ATAC-seq reads were processed with a custom pipeline [338] written in Snakemake [339] available on GitHub: github.com/kzkedzierska/atac. Briefly, the quality of the reads was assessed with fastqc [332], adaptor sequences were trimmed with cutadapt [333] using trim_galore [334]. Processed reads were then aligned to hg38 using bwa aligner [340]. The alignments were then filtered and processed with picard [341] and samtools [342]. Peaks were called with macs2 [343] and hmmratac [344]. Quality metrics were assembled and assessed with multiqc [337]. Differential accessibility and motif analysis was performed with ChrAccR [325, 345] and ChromVar [260]. DESeq2 [248] was used to identify differentially accessible regions and annotated as more accessible with FC > 1.5 and adjusted p-values < 0.05, and less accessible with FC < 0.67 adjusted p-values < 0.05.

Joint analyses Genomic alignment was visualized with IGV [346] and pygenomtracks [347]. Exploratory data analysis, visualization, and in-depth data analysis were performed in R [236] with a suite of packages described and listed in more detail in Appendix A.

Enrichment analysis Gene set enrichment analysis in gene clusters was performed with Enrichr [276] using the enrichR package [348]. Gene set enrichment analysis (GSEA) [349], as implemented in clusterProfiler [350] and fgsea [351], on differentially expressed genes, or genes linked to differentially accessible regions, was performed. Genes were ranked using Wald statistic calculated with DESeq2. Enrichment scores are calculated

as described in [349]. Briefly, a distribution of ranks between the pathway genes and the full gene list is compared using Kolmogorov–Smirnov-like statistic. A higher score symbolizes the enrichment of genes at one end of the list (top or bottom), while a lower score describes similar distributions between the two lists, i.e., no enrichment. Scores were normalized to account for the size of the pathway and allow for comparison of the scores across pathways. See [351] for details on normalization and p-value estimation. Hallmark [349, 352] or KEGG [353, 354] gene sets were used if not specified differently.

Clustering and visualizations For visualizations and clustering, a regularized logarithm (rlog) normalization is performed as implemented in the DESeq2 package to allow for comparison between conditions and genes. Unless specified differently, for visualizations, values were standardized (scaled) across features (genes, accessibility regions, motifs). Each feature was centered and scaled by subtracting the mean and dividing by the standard deviation. Standardization was performed to avoid the dominance of the highly expressed genes in the visualization. The clustering of features was performed on the distance d calculated as follows $d = 1 - \rho_{X,Y}$, where ρ denotes a Pearson correlation between features X and Y. Samples were clustered based on Euclidean distance. If not specified differently, a complete linkage was used for clustering.

Heatmaps were created with ComplexHeatmap [355]. GSEA plots were created with custom functions relying on gseaplot2 function from GuangchuangYu/enrichplot package [350]. Other visualizations were created with custom code based on ggplot2 and its extensions (a list of packages commonly used in data analyses in this thesis is available in Appendix A).

3.4.2 *BCOR* N1459S experiment

3.4.2.1 Experimental work

CRISPR The murine Bcor gene was annotated against locus NC_000086 on NCBI reference sequence and used to design guide RNAs (gRNAs) for gene editing by CRISPR-Cas9. gRNAs and repair templates to introduce the murine equivalent of the human

N1459S mutation were designed by the transgenic core facility at the Wellcome Centre of Human Genetics (sequences provided in Table S1). Gene editing of murine embryonic stem cells was done by the core facility using standard methods; successful editing of targeted clones was performed by sequencing extracted DNA. Immunoblotting confirmed the loss of Bcor protein in mES cells with mES $Bcor^{KO}$.

RNA-seq Parental mES cells and targeted Bcor clones for RNAseq analysis were thawed and were passaged for maximum of 2-3 passages under standard conditions on the fibroblast feeder layer prior to RNA extraction. RNA was extracted and purified with the RNeasy Mini kit (Qiagen) and Turbo DNA-Free kit (Ambion; mouse cells only). The Oxford Core Genomics Facility conducted sequencing. Sequencing libraries were generated using TruSeq technology (Illumina) for the Illumina HiSeq Genome Analyzer. Libraries were sequenced to generate 40 mln fragments.

3.4.2.2 Computational analyses

RNA-seq The raw reads were processed and aligned according to the Oxford Bioinformatics and Statistical Genetics Core internal pipeline. Briefly, the adapters were trimmed using skewer [356], and the reads were aligned to mouse genome GRCm38 using hisat2 aligner [357]. Read counts were generated with featureCounts [358]. The counts were then imported using tximport package [359] and differential expression analysis was conducted using DESeq2 [248]. For visualizations counts were normalized using rlog procedure available through DESeq2. The batch effect removal was conducted using the linear model as implemented in limma package [360].

Chapter 4

Heterogeneity and evolution in cancer

Contents

4.1 Background 195
4.1.1 Relevance of cancer evolution and tumor heterogeneity 197
4.2 Methodology for ordering genomic events 197
4.2.1 Methodologies for reconstructing evolutionary trees 198
4.2.2 Problem setup 201
4.2.3 Existing approaches for reconstructing evolutionary trajectories from single-sample bulk sequencing 201
4.2.4 Deep Learning approach 203
4.3 Discussion and conclusions 225

4.1 Background

The complexity of cancer stems, in part, from its capacity for adaptive evolution, rendering it a formidable challenge in understanding and treatment. Recent advances in sequencing technologies have facilitated exploration and a more profound understanding of the genomic heterogeneity within and between tumors, illuminating the mechanisms underlying their initiation and progression. Understanding cancer evolution and tumor heterogeneity is paramount for advancing cancer research, understanding cancer progression, and devising targeted and effective therapeutic strategies [361–363]. Nonetheless, considerable knowledge gaps remain despite the progress made thus far.

In Chapter 2, I presented the analysis of mutation subclonality in the GEL EC cohort. Nonetheless, a comprehensive investigation of cancer evolution is required to delineate the

evolutionary trajectories underlying endometrial carcinoma.

In this chapter, I delve into the complexities involved in elucidating cancer evolutionary trajectories and the challenges associated with this pursuit. I propose a novel approach to tackle these challenges by employing Recurrent Neural Networks (RNNs) to learn and model the patterns within evolutionary sequences. This innovative method aims to accurately predict subsequent driver events, offering valuable insights into the dynamic and intricate process of cancer evolution.

As early as the 1970s, Nowell [150] and others proposed that mutations in cancer are accumulated through selection for mutations that promote cell proliferation. With the advancements in sequencing technology, new methods to investigate the trajectories of cancer evolution have emerged.

One of the seminal works in the field was the study conducted by the Evolution and Heterogeneity Working Group of the PCAWG consortium[1] [240]. The study was based largely on a single sample of primary, untreated tumors and normal tissue samples for reference, with only a small proportion representing metastases or recurrences. In their analysis, authors used statistical modeling to infer the clonal frequency (i.e., fraction of cancer cells carrying a particular mutation) and mutation copy number (i.e., reflecting the number of chromosomes carrying a given mutation), which they used as a proxy for ordering genomic alterations. This allowed them to classify the mutations into clonal (early, late, or undetermined) or subclonal mutations. This approach was further complemented by *in silico* simulations to trace the evolutionary paths of individual tumors. The authors were able to identify several driver mutations that preceded the diagnosis by years and sometimes decades, providing an opportunity for early detection, as well as proving the significance of understanding the tumor evolutionary trajectories. However, having access to single samples of the primary tumors and due to the constraints of the statistical modeling, the research inherently permits only a partial reconstruction of cancer's evolutionary trajectories, highlighting the complexities and limitations in fully delineating the relative timing of mutational events.

[1]PCAWG - The Pan-Cancer Analysis of Whole Genomes was a study of whole genomes from the International Cancer Genome Consortium and TCGA.

4.1.1 Relevance of cancer evolution and tumor heterogeneity

Improving our understanding of tumor evolution is essential for developing more effective therapeutic strategies. For instance, a study involving 30 samples from four renal cell carcinoma patients revealed that over 60% of all somatic mutations were not identifiable across every tumor region [153]. Another study of 104 triple-negative breast cancers found highly variable subclonal frequencies [364]. These indicate that although tumor subclonal composition can be reconstructed from standard bulk sequencing approaches, only single-cell sequencing or deep sequencing of multi-region samples can secure the full coverage of subclones present in the tumor.

By acknowledging the heterogeneous nature of tumors, more comprehensive and personalized treatment plans can be designed to target all subclones. An example of this approach can be seen in the treatment of chronic lymphocytic leukemia (CLL), where identifying multiple subclones and their evolutionary trajectories informed the development of combination therapies to target both the dominant and minor subclones, thereby increasing treatment efficacy [365]. Evolution-aware therapeutic strategies like this may ultimately lead to the development of modified preventive therapies that target *all* subclones, thereby mitigating the risk of treatment resistance and improving patient outcomes.

In summary, the study of cancer evolution and tumor heterogeneity has far-reaching implications for cancer research and treatment, and its consideration is vital for developing effective and durable therapeutic strategies.

4.2 Methodology for ordering genomic events

In this work, I focus on branched tumor evolution, as described earlier in Section 1.6.2. Tumor composition, and therefore the inferred evolutionary trees, can exhibit significant differences at various time points in tumor evolution, as illustrated in Fig. 4.1A. To reconstruct cancer evolutionary trees, one needs to know which mutations are present in which clones and what their prevalence is across cells - a cancer cell fraction (CCF) (as

shown in Fig. 4.1B).

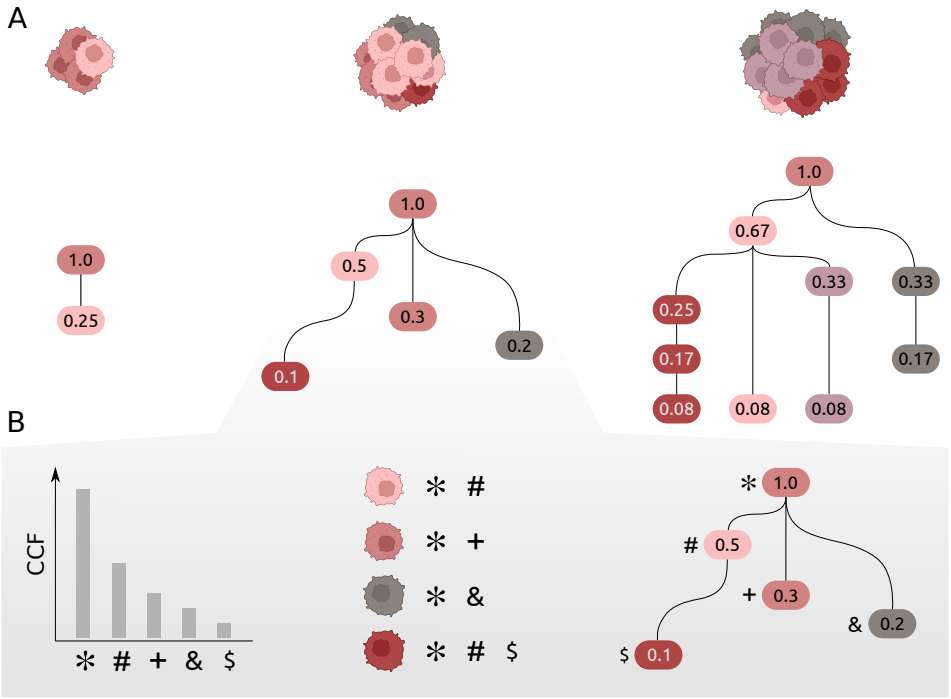


Figure 4.1: **Evolutionary trees. A** Evolutionary trees for the three scenarios as presented in Fig. 1.16 and discussed in Section 1.6.2. Numbers in nodes represent cancer cell fraction (CCF). **B** We can reconstruct a phylogenic tree based on the CCFs and information on which alterations (marked with symbols) co-occur in subclones.

4.2.1 Methodologies for reconstructing evolutionary trees

Current approaches in deciphering evolutionary trajectories include reconstructing tumor phylogenetic trees [366]. By ordering the alterations found in tumor samples, phylogenetic trees of cancer evolution can be reconstructed. Unfortunately, due to challenges spanning large quantities of mutations, including both passenger and driver events, copy number alterations affecting variant allele frequencies (VAF), and VAF being a coarse function of sequencing depth [367], those inferences usually result in multiple equally probable versions of phylogenetic trees. To narrow down the search space approaches to leverage multiple tumor samples [368–370] or multi-region samples are applied [371–374]. However, the differing evolutionary patterns across various samples further complicate the former approaches. The multi-region approach, on the other hand, requires acquiring samples spanning several regions of the tumor, which is costly and more challenging logistically.

Methods that leverage the multi-region samples are also more complex and often require manual curation [239, 366]. The advent of single-cell technologies has allowed for a more direct investigation of cancer progression. Rather than inferring the cell populations from bulk studies, a more straightforward approach to reconstructing the evolution trajectories from the alterations at single-cell resolution could offer an improvement. However, single-cell WGS (scWGS) is still limited by allelic dropouts, non-uniform coverage, PCR errors, and recombinations [375]. Alternative approaches, such as targeted sequencing or single-cell transcriptomics, present their own limitations.

Despite its challenges, given the available data, inferring the evolutionary trees directly from single-sample bulk samples is necessary. An interesting approach to the problem involves pairing the investigation of single-sample bulk sequencing combined with allele phasing. Haplotypes represent combinations of alleles at multiple loci on a single chromosome, inherited together as a unit from a single parental chromosome. Phasing is the process of determining the specific chromosomal origin of alleles, enabling the identification of haplotypes. It can be done directly from long-read sequencing if the mutations are found on the same fragment or from short reads sequencing using statistical methods, such as hidden Markov models in Beagle [376] or Bayesian-based models as used in Octopus, a haplotype-based variant caller [377].

Phasing enables the assignment of mutations to specific chromosome copies, which subsequently allows for determining the relationship between mutations and identifying distinct subclones. For instance, if two mutations, A and B, are found on the same chromosome copy with CCFs of 0.6 and 0.3, respectively, we can infer that the subclone with mutation A is parental to the clone with mutations A and B. This knowledge is crucial for accurately reconstructing the tumor's evolutionary tree.

Nevertheless, frequently, analyses are limited to bulk single-sample short-read sequencing without phasing information. Bulk sequencing technology effectively generates high-quality sequencing data from samples procured for diagnostic and evaluative purposes, amassing a substantial repository of invaluable data points. However, reconstructing the trees with only CCFs presents significant challenges. For instance, the CCF sequence

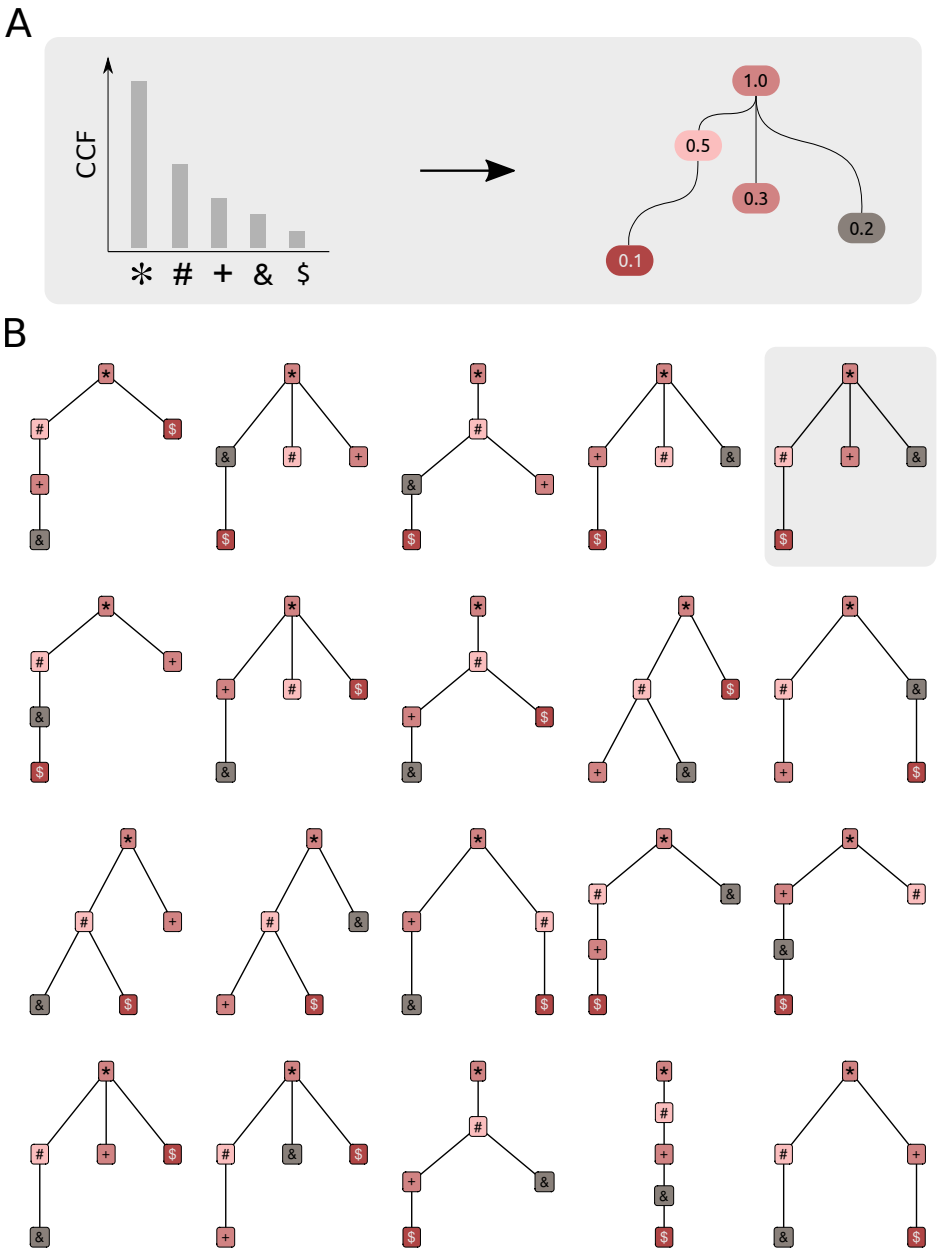


Figure 4.2: **Inferring the tree from CCFs. A** The aim is to reconstruct the tree from the CCFs of the mutations. **B** Twenty different evolutionary trees can be reconstructed from the sequence of CCFs in A. The actual tree is highlighted in grey.

demonstrated in Fig. 4.1B can correspond to as many as 20 distinct plausible evolutionary trees, as shown in Fig. 4.2. Crucially a single sample is unlikely to encompass all the clones in the tumor, so it might not accurately represent the complete evolutionary tree. This limitation underscores the challenge of reconstructing tumor evolutionary trees based solely on single-sample bulk sequencing data without phasing information.

4.2.2 Problem setup

Inference of individual trees can provide deep insights into how tumors evolve, for instance, the identification of polyclonal seeding in [378]. Most methods used in the field rely on jointly analyzing multiple samples from each cancer. At the same time, most available datasets consist of single biopsies of a given tumor. Therefore, there is a clear need for methods that can be applied to cross-sectional data from many patients rather than multi-region or longitudinal data from a few samples. Alternative approaches have been proposed that do not aim to reconstruct the accurate evolutionary history of individual tumors but instead aim to infer common evolutionary trajectories shared between tumors.

Comparing evolutionary behavior between tumors is not straightforward. As shown in Fig. 4.1A, the compositions of clones and the CCFs for each alteration differ at various time points, reflecting the dynamic nature of tumor evolution. This variability makes it challenging to compare the CCFs of alterations between samples directly. Instead, a more viable approach involves ordering the alterations based on their CCFs and focusing on comparing these orderings. One approach to examining the evolutionary tree within a given tumor is to consider each branch as a distinct evolutionary trajectory specific to a particular clone, which can be compared across multiple samples.

4.2.3 Existing approaches for reconstructing evolutionary trajectories from single-sample bulk sequencing

Estimating evolutionary trajectories continues to be a complex challenge, as only a limited number of published methods have moved beyond curated approaches.

One statistical approach, proposed by authors in [379], employs a Bradley-Terry [380]

model used for paired comparison data, which allows the estimation of the strength or preference of items or individuals by analyzing the outcomes of pairwise contests. For each pair of selected copy number alterations, the authors compared the number of times a given alteration occurred before the other. Assembling all pairwise comparisons together allowed the reconstruction of a relative ordering of those CNAs.

A recent method, described in [381], builds on this and utilizes the Plackett-Luce model [382, 383] to generate a probability distribution over relative orderings. More details about the method can be found in the supplementary materials (see Section B.2.1). Briefly, the authors use a Monte Carlo approach to estimate plausible trees from patient-level CCF distribution. They generate sampled orderings by sampling plausible trees and then creating an ordering based on each tree to capture the inherent variability and uncertainty in the data. The Plackett-Luce model is employed to infer the order in which events occur by generating a probability distribution over relative orderings. A Plackett-Luce mixture model [384] is used to determine whether subpopulations in the data exhibit different orderings. Maximum likelihood estimation and the Bayesian Information Criterion are applied to determine appropriate parameter values and the number of mixture components, respectively.

The curated approach offers the advantage of incorporating biological knowledge about specific cancer types and the consequences of alterations. This method provides a more tailored analysis that takes into account the nuances of the biological system. However, this approach is time-consuming and not easily reproducible, as it relies on expert input, which can be subjective and vary between experts.

On the other hand, the two statistical approaches are computationally intensive, providing a more systematic and objective analysis. These methods can be automated and scaled up to analyze large datasets, making them more reproducible. However, they still require a researcher's intervention to narrow down and identify the driver events before the modeling stage. In the case of the mixture model, an additional challenge is specifying the underlying number of mixture components, which can impact the accuracy of the analysis and may require further investigation or validation.

A data-driven approach would provide further insights into the complex nature of cancer evolution. By leveraging machine learning, a new solution could uncover the underlying evolutionary trajectories without the need to pre-select alterations. This unbiased approach would allow for the identification of previously unknown driver events and relationships among alterations, contributing to a more comprehensive understanding of the disease. Additionally, a data-driven approach would be highly adaptable and could be easily applied to multiple cancer types.

4.2.4 Deep Learning approach

4.2.4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) [385–388] represent a class of artificial neural networks specifically designed to model sequential data by incorporating memory elements that allow the network to maintain and utilize information from previous time steps (Fig. 4.3A). Unlike traditional feedforward neural networks, RNNs possess the unique capability of retaining state information through loops, enabling the network to capture temporal dependencies within a given sequence. This property makes RNNs particularly well-suited for tasks involving time series data, natural language processing, and speech recognition.

The basic architecture of an RNN, shown in Fig. 4.3A, consists of input (annotated as x_t, where t is a time point in a time series), hidden state (h_t), and output (y_t) with the hidden layer containing recurrent connections. When processing a sequence, the input at each time step (x_t) is combined with the hidden state from the previous time step (h_{t-1}) to produce an updated hidden state (h_t), which in turn is used to generate an output (y_t). Earlier inputs, x_{t-1}, can influence the output y_t at time t and later via the network's hidden states. This process is repeated for each element in the input sequence, allowing the RNN to model sequential relationships of varying lengths. The weights of the connections within the network are adjusted during training through a process called backpropagation through time (BPTT), an extension of the standard backpropagation algorithm that accounts for the temporal aspect of RNNs.

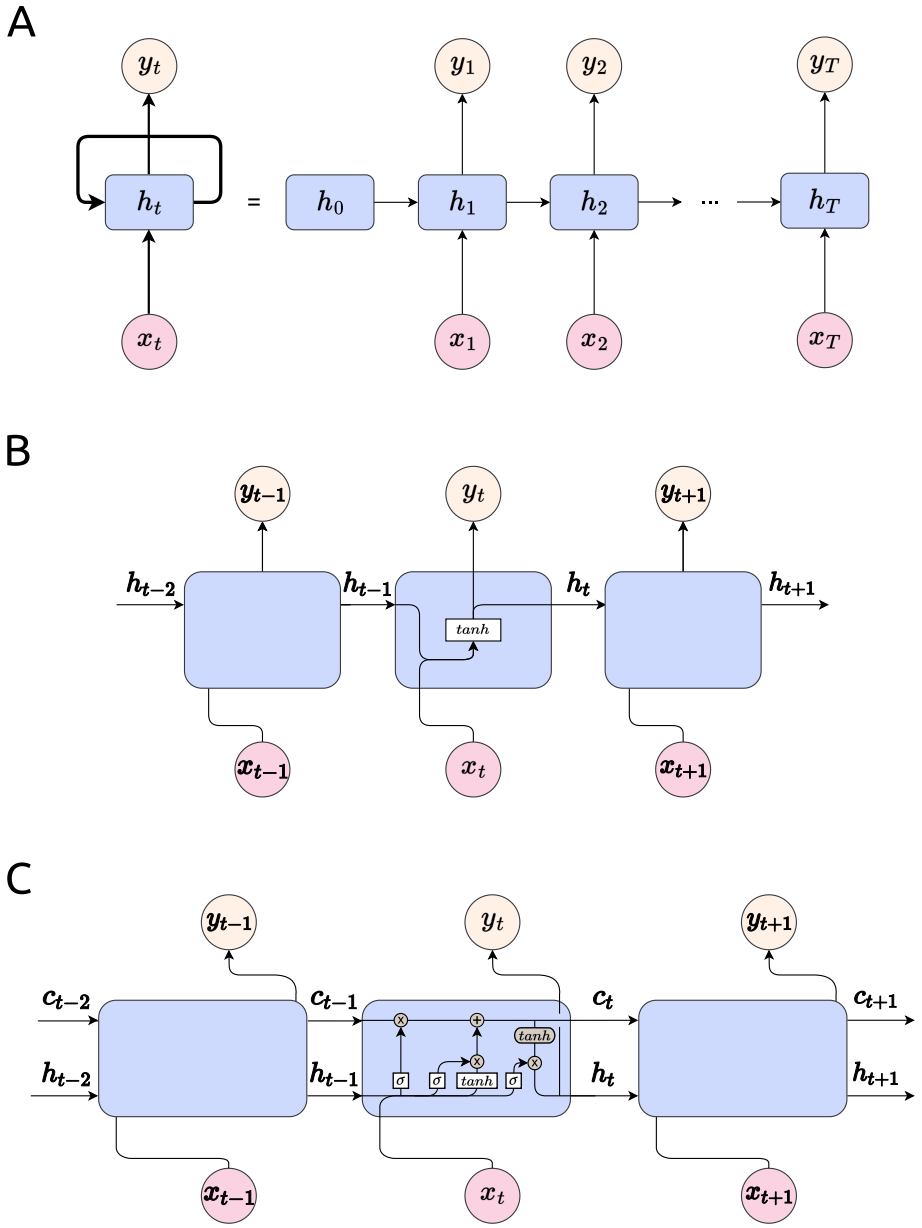


Figure 4.3: **Illustration of RNN and LSTM models. A** General schema of RNN models. **B** Architecture of vanilla RNN. **C** Architecture of LSTM. Figure created with draw.io and Inkscape.

Typically, the hidden state is defined by

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b_h),$$

and the output is defined as a function of the hidden state

$$y_t = W_y h_t + b_y.$$

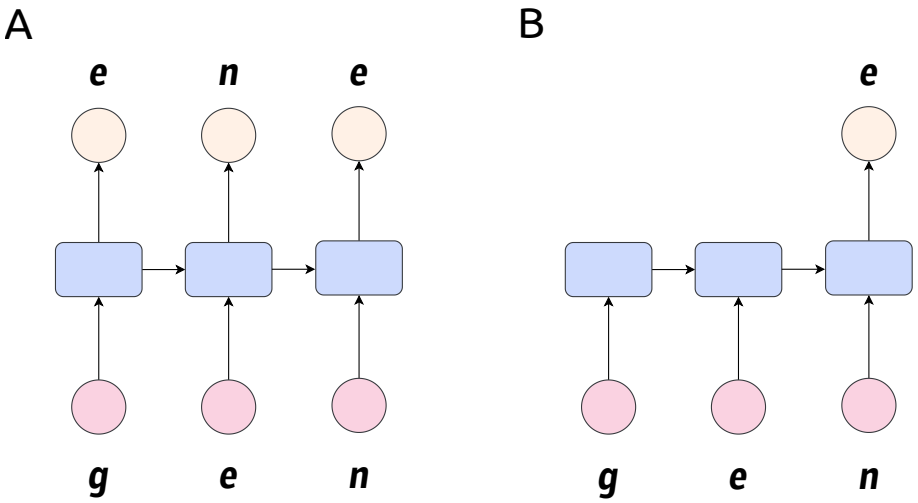


Figure 4.4: **Selected types of RNNs. A Many-to-many** The model generates the prediction for each time point. For example, for the word *gene*, with the input of *gen*, the model generates the prediction of *ene*. **B Many-to-one** In this model, we are only interested in the last letter as a prediction; hence for *gen* input, the model will return only *e*. Figure created with draw.io and Inkscape.

There are various types of RNNs that reflect how the network processes the inputs and outputs. In this work, I will focus on two types of RNNs - many-to-many (Fig. 4.4A) and many-to-one (Fig. 4.4B) networks. In the many-to-many model, as presented in Fig. 4.4A, the model takes a sequence as an input and outputs a shifted sequence. We illustrate how this works using a task of predicting letters in a world *gene*. As input, we provide the model with the first three letters in a sequence - *g*, *e*, and *n*. For each position in the sequence (t), the model predicts the $t + 1$ position, i.e., for $t = 1$ letter *g*, the model predicts *e*, for $t = 2$ letter *e* model predicts *n* etc. However, if we are more interested in predicting missing characters, we could choose a simpler model - many-to-one. As shown in Fig. 4.4B, the many-to-one model in predicting the next letter would only output one

letter at the final time point, $t = 3$. In this case, for input *gen*, the model would output a prediction *e*.

4.2.4.2 Long Short-Term Memory

Although powerful in their application to sequential data, RNNs can be prone to issues such as vanishing or exploding gradients during training, which can impede learning long-range dependencies. The vanishing gradient problem occurs when gradients become too small as they are propagated backward through time, causing the network to lose the ability to learn from distant time steps. Conversely, exploding gradients arise when the gradients become excessively large, leading to unstable and oscillatory training behavior. These issues can result in suboptimal performance and hinder the network's ability to capture complex temporal patterns.

Advanced RNN architectures, such as Long Short-Term Memory (LSTM) networks [389], have been developed to address these challenges. These architectures incorporate specialized gating mechanisms that modulate the flow of information and gradients through the network, thereby mitigating the issues associated with vanishing and exploding gradients. LSTMs replace the hidden states in vanilla RNN (Fig. 4.4B) with a memory cell involving an input gate, an output gate, and a forget gate to control the flow of information into, within, and out of memory cells, enabling the network to remember and forget information over a longer time scales selectively. As shown in Fig. 4.3C, information from previous time steps is separated into a cell state c_t and hidden state h_t.

4.2.4.3 Applications of RNNs and LSTMs

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have found widespread use across various domains due to their ability to model sequential data effectively. Some of the most frequent applications include language modeling, where RNNs and LSTMs learn the structure of the text and predict the next word or character in a sequence [390]; machine translation, where they have been employed to translate text between different languages by capturing the context and structure of the input [391]; and

speech recognition, where they process and transcribe spoken language by learning the complex temporal dependencies in acoustic signals [392, 393]. Additionally, these neural networks have been utilized in sentiment analysis for understanding the sentiment of text documents [394], named entity recognition to identify and classify entities in a text, and time series prediction, where they excel at learning patterns and dependencies in time-varying data. Furthermore, RNNs and LSTMs have been applied in creative domains such as music and video generation. They generate new compositions or classify videos by learning temporal structures and patterns [395, 396].

Applications in computational biology RNNs have emerged as a powerful tool in computational biology, offering a range of applications due to their ability to model sequential data effectively. PubMed search of applications of RNNs and LSTMs results in 1,427 and 3,789 hits, respectively[2]. They have been particularly valuable in analyzing biological sequences, such as DNA, RNA, and protein sequences, where they can capture the intricate patterns and relationships within the data (Alipanahi et al., 2015). By leveraging the sequential nature of protein sequences, RNNs have been demonstrated to accurately predict the secondary structure elements, such as alpha-helices, beta-sheets, and random coils, which are critical for understanding protein function and stability [397]. Additionally, RNNs have been employed in the study of RNA sequences, where they have shown great potential in predicting RNA secondary structure, folding, and the function [398].

Another area where RNNs have significantly impacted is the analysis of gene expression data. RNNs can be applied to model gene regulatory networks, allowing for the identification of complex regulatory patterns and interactions between genes. This can aid researchers in understanding the mechanisms governing gene expression and regulation, which is vital for both fundamental biological research and translational applications, such as drug discovery and personalized medicine [399].

[2]For applications of RNNs, the term "RNN" was used, for LSTMs - "LSTM". PubMed searched on 7/04/2023 with a publication date within 5 years filter applied

4.2.4.4 Employing RNNs to unravel cancer evolutionary trajectories

The primary motivation for developing a Deep Learning approach is to create a solution capable of learning the underlying evolutionary trajectories in cases where the number of trajectories is unknown. In contrast to the discussed statistical models, a Deep Learning RNN-based approach serves as a generative model that not only distinguishes the underlying evolutionary trajectories but also learns conditional distributions $x_{t+1}|(x_1, ..., x_t)$. This enables the model better to capture the complex dynamics and dependencies within the data, offering a more flexible and powerful solution for understanding tumor evolution.

Learning the sequence of driver events in the evolutionary trajectory can be compared to learning a sentence where alterations are comparable to words. For example, two driver events can be separated by a sequence of passenger alterations (from the evolutionary trajectory perspective). Similarly, two connected words can be separated by a sequence of other words. For example, in a sentence:

Yesterday, seeing all the rain, I was grateful I brought my umbrella with me.

yesterday is connected to words *is* and *bring* by informing their past tense form, and *rain* provides a context for the most likely item - *umbrella*. RNNs were designed to tackle the task of learning and "remembering" the long context of the sequence.

In this work, I opted to examine both a vanilla RNN, as a baseline, and an LSTM model under two configurations: a single-layer many-to-one architecture (Fig. 4.5A) and a two-layer architecture, wherein the first layer employs a many-to-many architecture and the second layer utilizes a many-to-one setup (Fig. 4.5B). For the two-layer model, both layers are of the same type, either consisting of two vanilla[3] RNNs or two LSTMs.

Self attention Attention, introduced in [391, 400] and popularized in [401], a key component in many modern deep learning architectures, allows a neural network to focus selectively on different parts of an input sequence, effectively capturing contextual information and long-range dependencies. In self-attention, each element in the sequence is

[3]vanilla is a term used to describe a simple model architecture, usually used as a baseline for more complex setups.

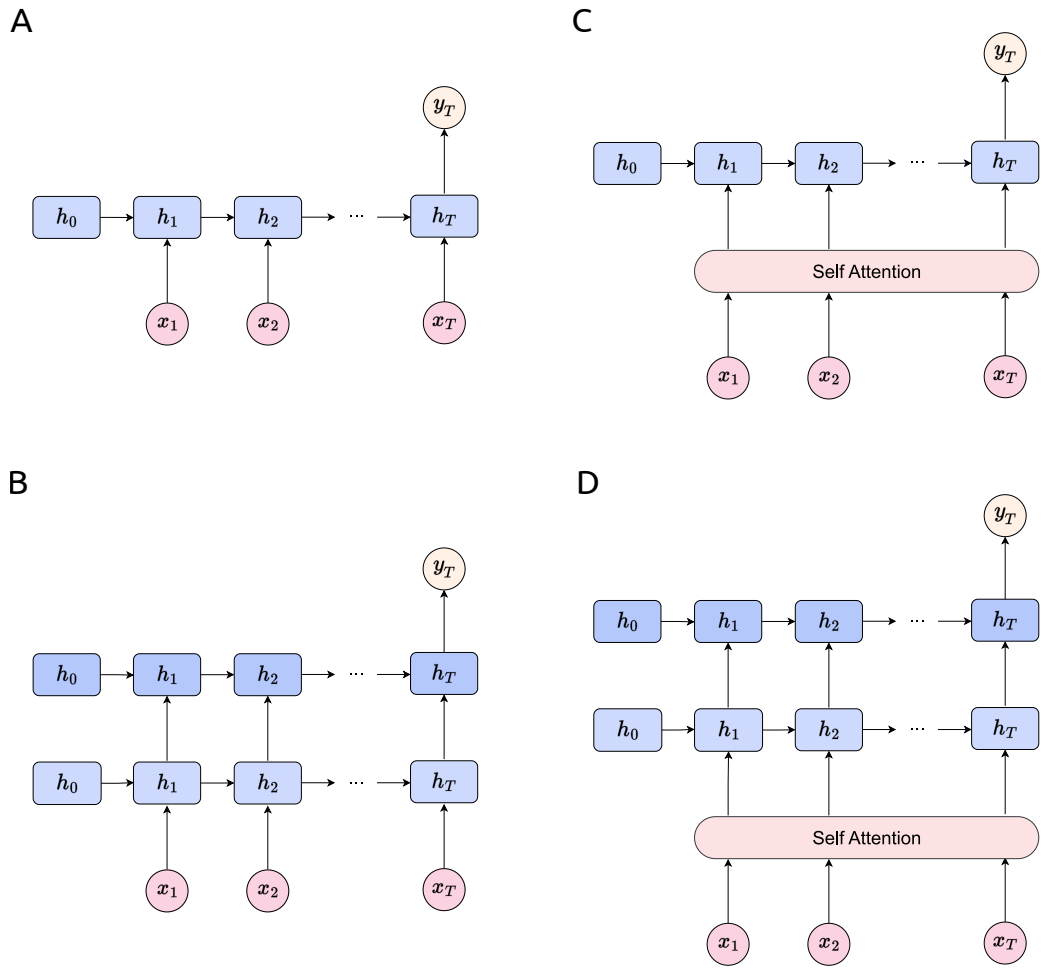


Figure 4.5: **Illustration of the selected models' architecture.** **A** Single-layer model with many-to-one architecture. **B** Two RNN/LSTM layers model with many-to-many output in the first and many-to-one architecture of the second layer. **C** Single-layer RNN/LSTM model with many-to-one architecture and self-attention. **D** Two RNN/LSTM layers model with many-to-many output in the first and many-to-one architecture of the second layer with self-attention layer. Created with draw.io and Inkscape.

assigned a weighted representation based on its relationships with other elements. This is achieved by computing a series of dot products between the elements' embeddings, followed by a softmax operation to produce the attention weights. The final output is a linear combination of the original embeddings, weighted by the attention scores. By dynamically enabling the model to weigh the importance of each element in the sequence, self-attention has significantly improved the performance of deep learning models in tasks such as machine translation, natural language understanding, and sequence-to-sequence prediction.

As mentioned earlier, learning the sequence of driver events in the evolutionary trajectory of a cancer can be compared to learning a sentence where alterations are comparable to words. This metaphor holds especially true if we consider that not all words carry the same information content in predicting the next words. The primary function of attention is to enable the model to focus dynamically on relevant parts of the input sequence while processing information rather than relying on fixed-length representations. Attention learns to weigh the importance of different elements in a sequence based on the task. For example, in a sentence *Ada is performing a high-throughput sequencing experiment* words *is* and *a*, although grammatically correct, do not consist of much information in the context of predicting the next words in the sequence. On the other hand, *high-throughput* often precedes *sequencing*, offering much more information pertaining to the consecutive words. The attention layer can learn which words provide the necessary context and therefore to which of them to pay attention.

In another example, a question-answering task, attention can help the model focus on the most relevant parts of the passage to find the correct answer. For example, attention applied to a question *Could you please help me describe the process of photosynthesis?* would allow the model to focus on the underlined parts.

Similarly, in the context of cancer evolution, passengers, often found in random orders, do not inform the evolutionary trajectory, while drivers guide the path of cancer evolution. Adding a self-attention layer enables the model to learn which alterations to focus on.

In order to explore the role of attention mechanisms in cancer evolution, I evaluated

modified versions of the models discussed in the previous section by incorporating a self-attention layer (Fig. 4.5C-D).

4.2.4.5 Synthetic data

As explained in Section Section 4.2.4.1, there are a number of options for constructing a deep learning sequence model. We will therefore create a number of models and then assess and quantify how well each model has captured the underlying evolutionary trajectory. As evolutionary trees in real data are largely unknown, I created synthetic data sets that mimic the intricate evolutionary trajectories found in real data and used these to evaluate the models. These synthetic data sets consist of varying lengths of ordered alterations, including signal - representing ordered driver alterations - and background - representing random passenger alterations.

The three evolutionary trajectories, as visualized in Fig. 4.6, include:

1. **short driver sequence** consists of two driver sequences with 4 shared driver alterations in a different order;

2. **long driver sequence** consists of two driver sequences with 3 shared driver alterations in the same order, 7 shared driver alterations in a different order, and 5 driver alterations separate to each sequence;

3. **early diverging sequence** consists of two driver sequences with 3 shared driver alterations in the same order, 7 driver alterations unique to each sequence, and 1 shared driver at different positions in the orderings.

The ordering created based on cancer cell fraction estimates from sequencing data consists of driver (signal) and passenger (background) alterations. To mimic this phenomenon, I sampled the sequence of driver alterations and randomly inserted the passenger alterations into the sequence. For example, in creating a data set based on the short driver sequence, I created an ordering of a length of 10 alterations as follows:

1. choose the underlying driver sequence at random;

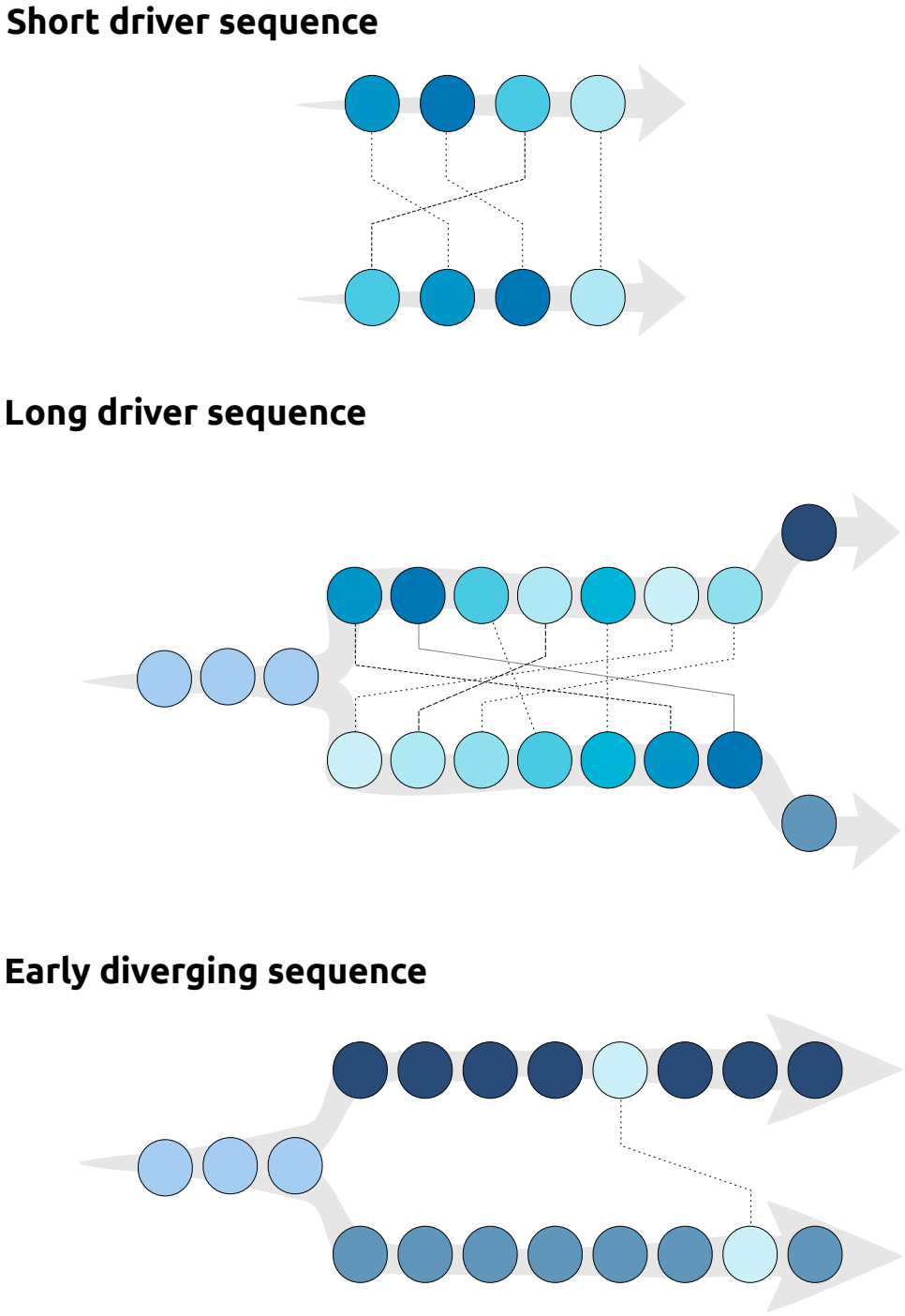


Figure 4.6: **Driver sequences used in simulations.** Each circle denotes a separate driver alteration. Arrows indicate the direction of ordering (i.e., progression of time). **Top panel** Short driver sequence consists of four common drivers in different orders. **Middle panel** Long driver sequence consists of three common drivers (light blue) early in the evolutionary trajectory, seven common drivers (various shades of blue) in different orders in the middle, and five separate driver alterations (dark shades of blue) further down the evolutionary trajectory. **Bottom panel** Early diverging sequence consists of three common drivers (light blue) early in the evolutionary trajectory, seven separate drivers (dark shades of blue), and one common driver at different time points.

Inputs	Prediction target
x_1	$y = x_2$
x_1, x_2	$y = x_3$
x_1, x_2, x_3	$y = x_4$
x_1, x_2, x_3, x_4	$y = x_5$
x_1, x_2, x_3, x_4, x_5	$y = x_6$

Table 4.1: **Example inputs and predictions for a many-to-one model.** For the sampled first sequence from Fig. 4.7 at $n = 4$ X = ["gene O", "loh 14q13", "gene M", "gene G"] and Y = "loh 9p19".

2. drivers are assigned to random positions in the sequence while preserving their order;

3. the six passengers are chosen randomly from a set of (i) four passengers shared between two separate sequences (shown in beige in Fig. 4.7) and (ii) four passengers unique to the sequence of passengers (marked in green and red in Fig. 4.7).

I repeated those steps creating a total of 100 sequences. For each of those sequences, I created 9 pairs of the input sequence (X) of length n ($n \in 1, 2, ..., 9$) and output (Y) with $n + 1$ element as shown in Table 4.1.

Similarly, I created the following three data sets by applying the following modifications:

- **long driver sequence data set** is a set of 30 alterations long sequences created by sampling driver sequence from long driver sequence and a set of (i) 15 shared passenger alterations and (ii) 15 exclusive passengers each. This sequence type contains a 1:1 ratio of the driver to background ratio.

- **long driver sequence (with more background) data set** is a set of 30 alterations long sequences created by sampling driver sequence from long driver sequence and sets of (i) (i) 30 shared passenger alterations and (ii) 30 exclusive passengers each. This sequence contains a 1:2 ratio of the driver to background ratio.

- **early diverging sequence data set** consist of sequences of 30 alterations and incorporates the two driver sequences as shown in Fig. 4.6 with sets of (i) 15 shared passenger alterations and (ii) 15 exclusive passengers each. This sequence type contains a 1:1 ratio of the driver to background ratio.

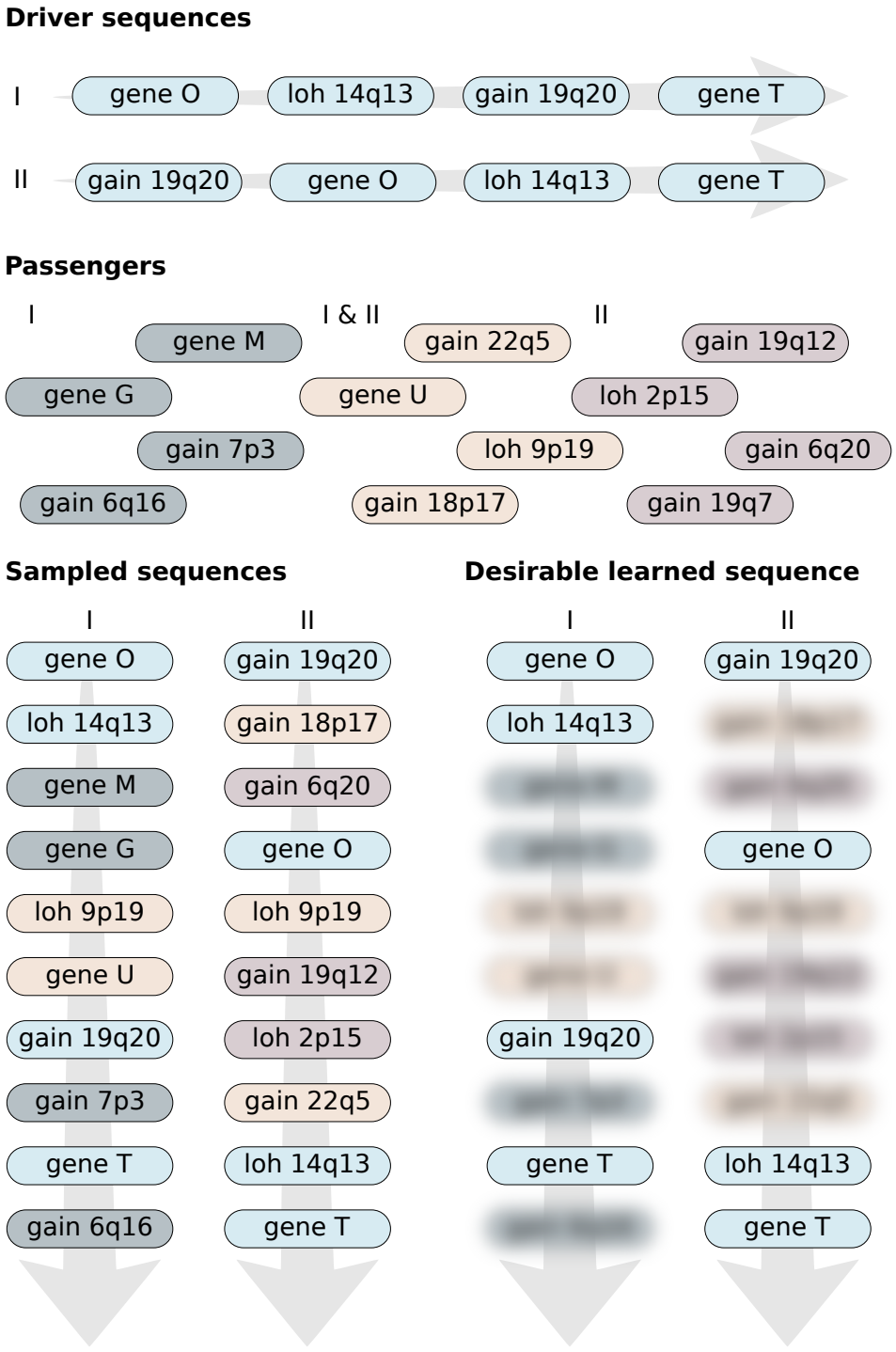


Figure 4.7: **Visualization of the synthetic data generation based on short driver sequences.** The alteration labels, added to illustrate the hypothetical evolutionary trajectory, are generated randomly to resemble copy number alterations and mutations. In this setup, there are two underlying driver sequences. The driver alterations are shared between the sequences with three alterations in a different order. The set of passengers exclusive to the first sequence shaded in green, those exclusive to the second one marked in red, and the shared ones in yellow. As shown in the example sampled sequences, the ordering of drivers is preserved, while passengers are sampled from the set of exclusive and shared sets and placed at random. The model will aim to learn the driver sequences while learning to ignore the passenger alterations.

Generated sequences were further split 80:20 into the train-test split. The models were trained on the train sequences and evaluated on the test sequences. Accuracy during training was calculated with respect to predicting the next event in the sequence.

4.2.4.6 Model evaluation

Following the similarities between the task of predicting the next event in evolutionary trajectory and the next word in a sequence I decided to investigate the application of the models often used in the Natural Language Processing (NLP) field in predicting the sequence of events. I explored the use of vanilla RNN and LSTM in the following configurations: a) one layer (Fig. 4.5A), b) two layers (Fig. 4.5B), c) one layer with attention (Fig. 4.5C) and d) two layers with attention (Fig. 4.5D).

In all model architectures, the implementation includes an output layer - a dense layer with softmax activation, defined as follows

$$f(s)_i = \frac{e^{s_i}}{\sum_{j=1}^{C} e^{s_j}},$$

where element $f(s)_i$ denotes the probability for an occurence of i-th alteration, s is the score for all alterations and C is the total number of all alterations under consideration. The output can be interpreted as the probability distribution over the alterations, with the highest score attributed to the most likely next event, i.e. the model's prediction denoted by $\hat{y}$

$$\hat{y} = \arg\max_{i=1,\ldots,C} f(s)_i,$$

The model is trained using the cross-entropy loss function (CE), which is defined as

$$\text{CE} = -\sum_{i=1}^{C} o_i \log(f(s)_i),$$

where o_i is the ground truth – a one-hot vector with 1 marking the correct event and 0 for all others. For example, in a set of four alterations: A, B, C, and D, an alteration B would be encoded $[0, 1, 0, 0]$).

We can rewrite the CE as follows:

$$\mathrm{CE} = -\log\left(\frac{e^{s_y}}{\sum_{j=1}^{C} e_{s_j}}\right),$$

where s_y is the network's score for the true alteration. The cross-entropy loss effectively measures the discrepancy between the model's predicted probability distribution and the ground truth, driving the model to make more accurate predictions of the next event in the sequence.

This loss allows for training the model, but since the aim is that the model will learn the underlying *driver* evolutionary sequences I measured the accuracy of predicting the next *driver* alteration to evaluate models and choose the one that performs best. I define the accuracy of predicting the next *driver* alteration as follows

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(\hat{y}_n = \tilde{y}_n),$$

given

$$\mathbb{I}(\hat{y}_n = \tilde{y}_n) = \begin{cases} 1 & \text{if } \hat{y}_n = \tilde{y}_n \\ 0 & \text{otherwise} \end{cases},$$

where N is the number of sequences and $\tilde{y}_n$ is the next *driver* alteration[4].

In assessing the accuracy of predicting the next driver alteration, I compared the performance of RNN and LSTM models across four simulated datasets (see Fig. 4.8A, Table S4). The single-layer RNN model achieved an average accuracy of 24%, with the best performance observed for short driver sequences (accuracy of 0.36) and the lowest for long driver sequences with increased background (accuracy of 0.10). In contrast, the single-layer LSTM model demonstrated superior performance, achieving an average accuracy of 0.57, with better performance on short driver sequences (accuracy of 0.69) and lower performance on long driver sequences with increased background (accuracy of 0.40).

Introducing a second layer significantly increased the number of learned parameters.

[4]Not to be mistaken with y_n – the next alteration, the ground truth.

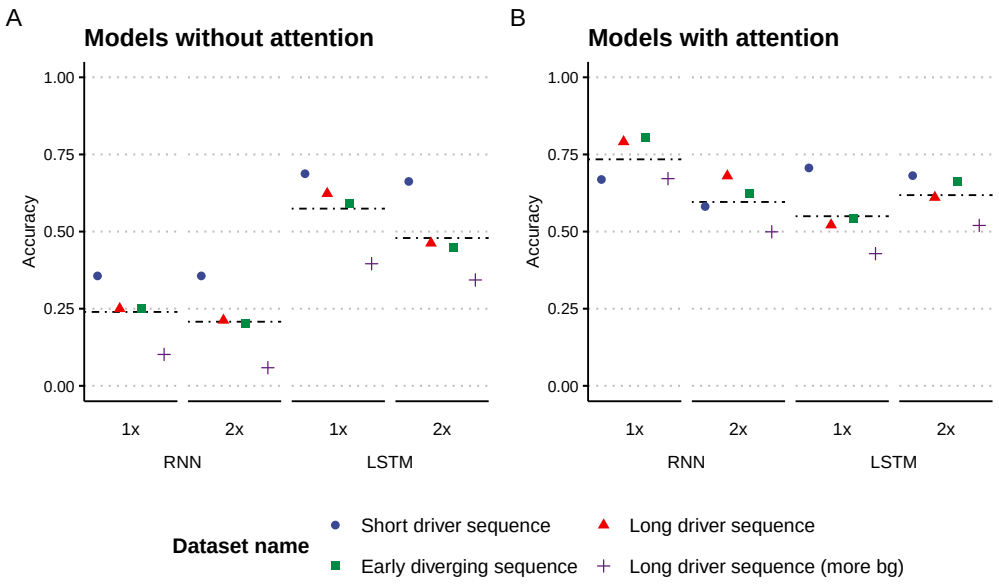


Figure 4.8: **Accuracy of models in predicting the next driver event.** Accuracy was measured on 4 different data sets, with mean accuracy plotted as a dotted horizontal line. **A** Models without attention layer. **B** Models with attention layer.

For instance, in the short driver dataset, the number of parameters in the RNN model increased from 4,166 to 9,216, while for the LSTM model, it grew from 14,216 to 34,416. However, the performance of these two-layer models was inferior to that of their single-layer counterparts (Fig. 4.8A).

Incorporating an attention layer substantially improved the performance of the single-layer RNN model, yielding an average accuracy of 0.73 (see Fig. 4.8B, Table S5). This model exhibited higher accuracy for the long driver (0.79) and early diverging sequences (0.81) compared to short (0.67) and long driver with a lower signal-to-background ratio (0.67) sequences. Also, the two-layer RNN model with attention demonstrated enhanced accuracy relative to the two-layer RNN model without attention, however, it underperformed compared to the single-layer RNN model with attention (Fig. 4.8B). Interestingly, the addition of attention decreased the accuracy of the single-layer LSTM model, while the opposite effect was observed for the two-layer model.

Overall, the highest accuracy was achieved by the single-layer RNN model with attention.

4.2.4.7 Intepreting self-attention

Despite the inherent black-box nature of neural networks, particularly RNNs, the proposed approach offers a distinct advantage in terms of interpretability. By incorporating a self-attention layer, researchers can scrutinize and visualize the aspects of the data that the model is focusing on. This, in turn, enables the development of an intuitive understanding surrounding the model's application and underlying mechanisms.

In a separate experiment, I trained the LSTM model on two datasets, one with short driver sequences and the other with long driver sequences. After training, I extracted the attention layer weights and visualized them to analyze the model's focus during prediction.

For the short driver sequence dataset, attention weights are displayed in Fig. 4.9A, while the number of occurrences at a given time point is illustrated in Fig. 4.9B. The attention layer assigns greater importance to driver alterations that appear earlier in the sequence. This is evident as the model largely ignores the last driver alteration in the sequence since it does not contribute to predicting subsequent events. Shared passenger alterations, which occur more frequently in sequences, receive relatively more attention. The alteration in gene G, found exclusively in sequence I (as shown in Fig. 4.7), is deemed more important due to its early appearance in the sequence, serving as a distinguishing event.

For the long driver sequence dataset, passenger alteration positions appear random, whereas the driver alteration positions exhibit a discernible organization (Fig. 4.10B). The attention layer assigns the highest importance to driver alterations (Fig. 4.10A). Alterations that are more frequently encountered in training sequences receive greater attention. Consistent with the short driver sequence dataset, the model learns not to focus on the last occurring drivers, as they do not inform subsequent alterations.

Overall, as expected, attention scores are indeed higher for drivers than for passengers (Fig. 4.11). The difference is more pronounced for a short driver sequence (Fig. 4.11A) than for a longer sequence (Fig. 4.11B).

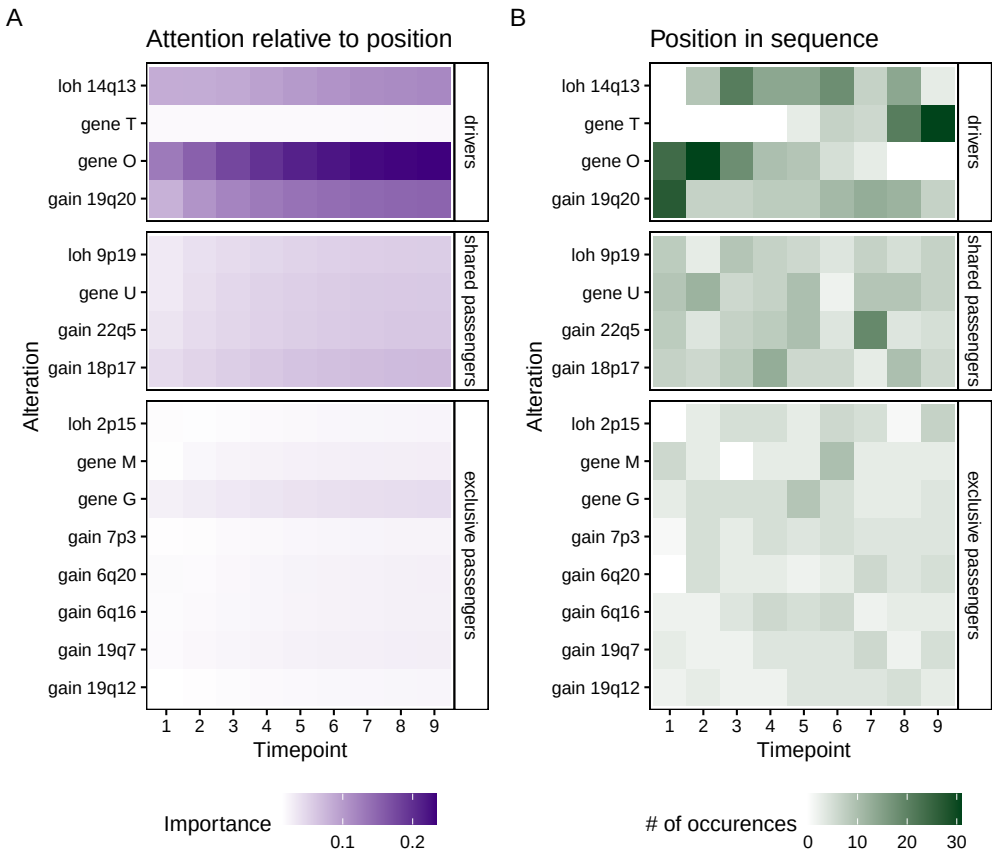


Figure 4.9: **Visualization of the self-attention layer in the short driver sequence model.** **A** Visualization of attention weights. **B** Occurrence of the alterations in the sequence.

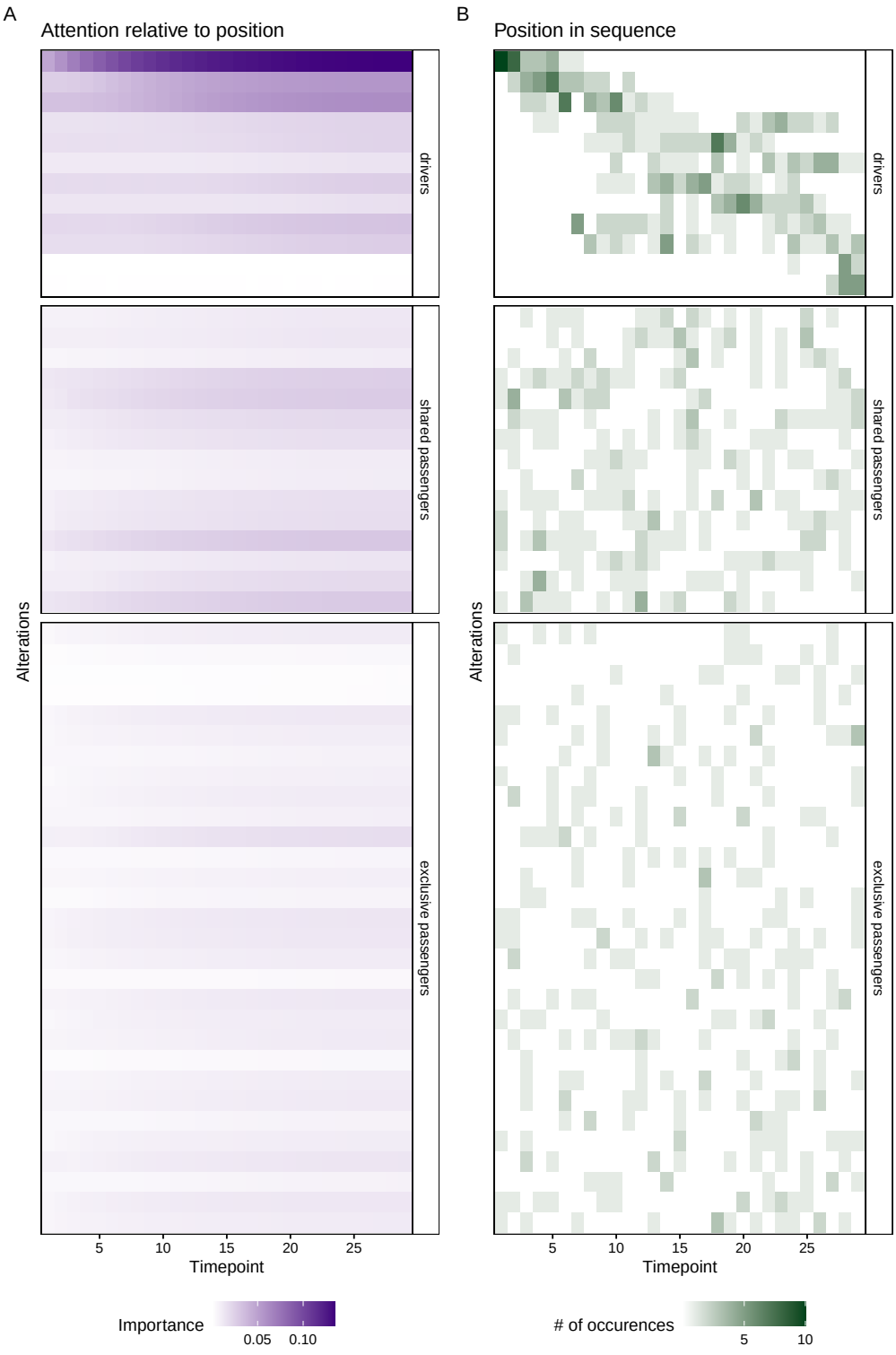


Figure 4.10: **Visualization of the self-attention layer in the long driver sequence model.** **A** Visualization of attention weights. **B** Occurrence of the alterations in the sequence.

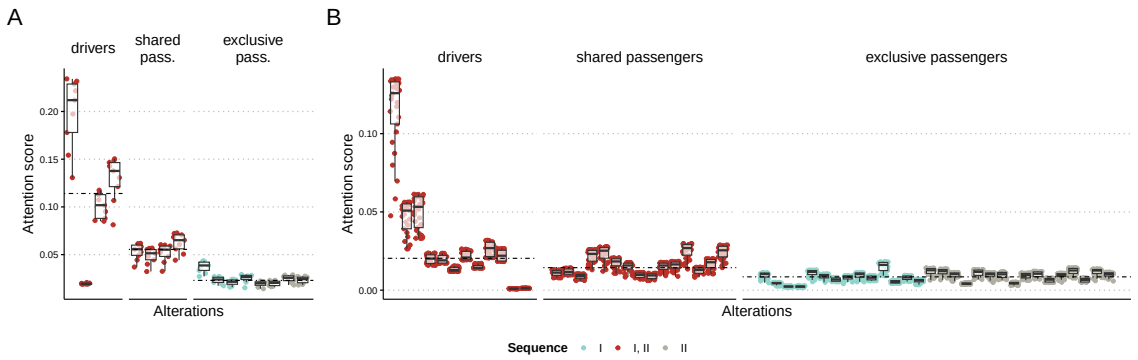


Figure 4.11: **Attention score per alteration types.** Scores from the same experiments visualized in Fig. 4.9 and Fig. 4.10.

4.2.4.8 Regularization

A primary challenge when working with neural networks is the potential for overfitting the data. In the context of RNN models for predicting the evolutionary sequence of genomic alterations, overfitting could result in the model learning the exact sequence of the training dataset, including the position of passenger alterations, which would hinder its generalization capabilities. To address this issue, I explored various regularization techniques:

- L1 penalty (Lasso regularization) - It adds the absolute values of the weights multiplied by a regularization parameter to the loss function, encouraging sparsity in the model by driving some weights to zero.

$$L_1(\mathbf{w}) = \lambda \sum_{i=1}^{n} |w_i|,$$

  where $L_1(\mathbf{w})$ denotes the L1 penalty term, λ denotes penalty coefficient and w_i denotes model weights.

- L2 penalty (Ridge regularization) - It adds the squares of the weights multiplied by a regularization parameter to the loss function, discouraging large weights and reducing the complexity of the model.

$$L_2(\mathbf{w}) = \lambda \sum_{i=1}^{n} (w_i)^2,$$

where $L_2(\mathbf{w})$ denotes the L2 penalty term, λ denotes penalty coefficient and w_i denotes model weights.

- L1L2 penalty (Elastic Net regularization) - It combines both L1 and L2 penalties, promoting a balance between sparsity and complexity control.

$$L_{1L2}(\mathbf{w}) = \lambda_1 \sum_{i=1}^{n} |w_i| + \lambda_2 \sum_{i=1}^{n} w_i^2,$$

where $L_{12}(\mathbf{w})$ denotes the L1L2 penalty term, λ_1, λ_2 denotes penalty coefficients and w_i denotes model weights.

Additionally, I employed dropout [402], a technique that randomly deactivates a proportion of neurons during training, preventing over-reliance on any single neuron and promoting robustness.

To evaluate the effectiveness of various regularization techniques in mitigating overfitting and promoting generalizability, I applied them to two distinct model architectures: one consisting of an attention layer and a single LSTM layer and another featuring an attention layer with two consecutive LSTM layers. The regularization techniques were applied to different components of the models, including the weights (kernel), biases (bias), and layer outputs (activity), as shown in Fig. 4.12 (Table S6).

In the single LSTM layer model, the most notable improvement in accuracy was observed when Elastic Net regularization was applied to the layer output for the short sequence dataset (increasing accuracy from 0.71 to 0.81, Fig. 4.12, Table S6). However, the L1 (Lasso) activity penalty yielded a more substantial improvement for long driver and early diverging driver sequences (an improvement from 0.52 to 0.70 and 0.63 to 0.78, respectively, Fig. 4.12, Table S6). For the two-layer model, the highest accuracy across all datasets was achieved when L2 (Ridge) penalty was applied to the layer output improving the accuracy most significantly in long driver sequence (from 0.56 to 0.66, Fig. 4.12, Table S6). This suggests that the choice of regularization technique and its application may vary depending on the model architecture and the nature of the input sequences.

Nonetheless, apart from the enhanced performance on the short driver sequence, the

overall accuracy remained lower than that of a single-layer RNN network incorporating a self-attention layer. Additional experiments are warranted to identify the optimal model, encompassing parameter searches for each method and evaluating regularization techniques in alternative architectures.

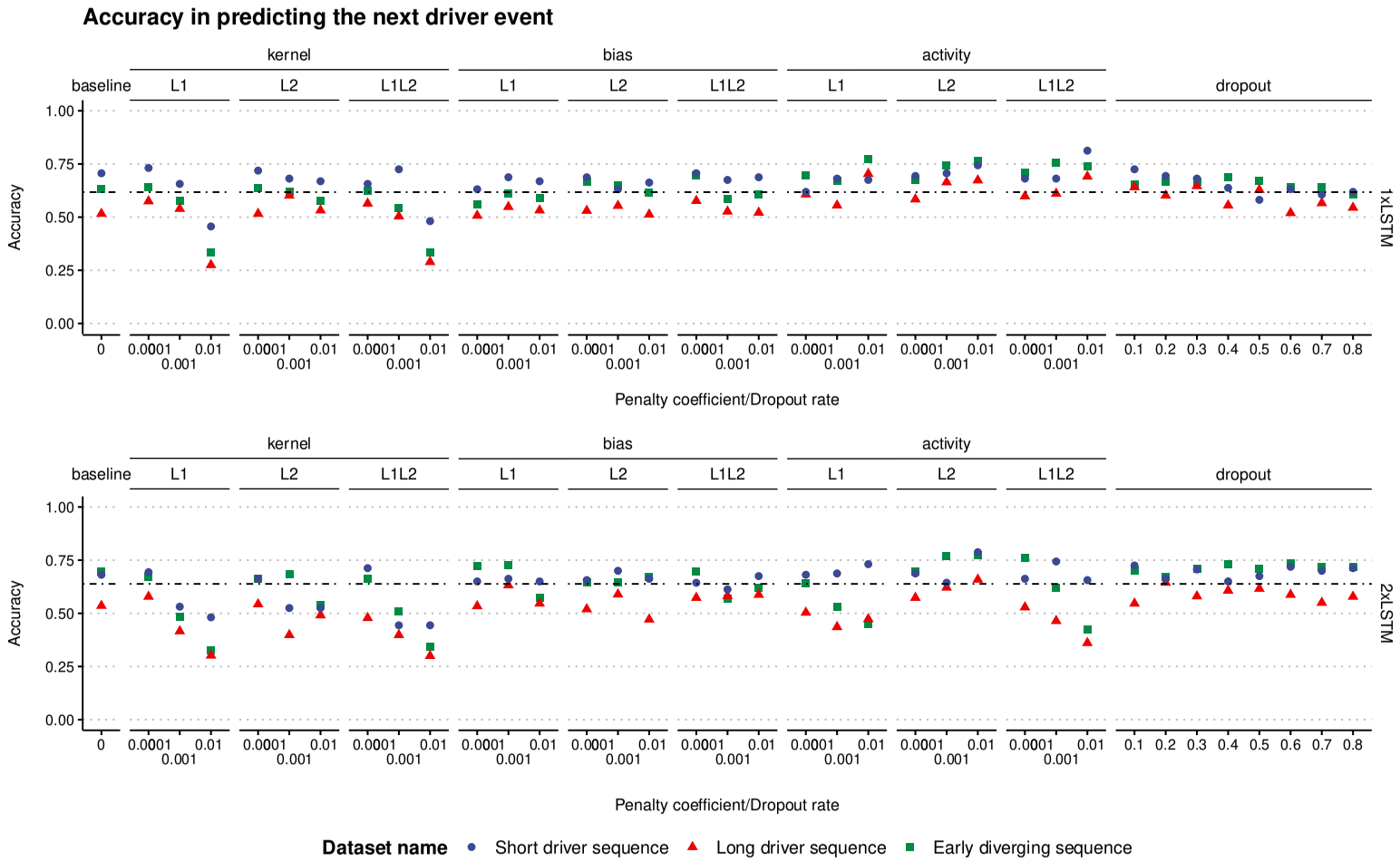


Figure 4.12: **Accuracy of models in predicting the next driver event with respect to regularization methods.** A dashed horizontal line shows a mean accuracy for a baseline model - trained without any regularization method.

4.3 Discussion and conclusions

Reconstruction of evolutionary trees and trajectories is challenging, as current methods often necessitate manual curation and rely on numerous assumptions, as discussed in Section 4.2.1. In this research, preliminary results from an ML-based approach employing Recurrent Neural Networks (RNNs) with an attention layer are presented as a proof of concept. This method draws an analogy between the roles of words in sentences and mutations in the history of cancer, distinguishing between driver mutations, which shape the evolutionary trajectory, and evolutionarily neutral passenger mutations. The motivation behind the method is to create a more general-purpose method that could be applied without having to model the various assumptions directly and that can directly leverage the bulk sequencing of a single tumor sample, allowing it to be used retrospectively with the already available data. Despite the promising nature of these initial findings, the study faces several limitations that warrant further investigation.

To develop a more refined and robust methodology, it is imperative to address the current reliance on synthetic datasets, which lack the complexity and noise characteristic of real-world data. This, however, requires access to a well-established ground truth.

Currently, there is limited published work studying the evolution of endometrial cancer, which makes it harder to leverage the high-quality genomic data from GEL EC cohort. At the same time, this underscores the need for further research. One of the few studies of EC evolution was the study of endometrial carcinoma and abdominopelvic metastasis samples [403]. In their research, authors analyzed biopsies of primary tumors, primary tumors matched with metastases, metastases with unmatched primary tumors, and precursor lesions from 52 individuals. Using WES and single nucleotide polymorphism (SNP) array profiling, they were able to describe the mutational landscape of the coding genome alongside the copy number analyses of the samples. Next, by creating a customized strategy to create incidence matrices, authors reconstructed the phylogenetic trees using hierarchical clustering of the distance between the primary tumors and metastasis samples. Careful consideration of the matched samples allowed them to identify the mutations shared between the biopsies and those that were unique for each of them.

Based on the phylogenetic analyses, authors concluded that the metastases were likely to have originated from a common ancestral subclone that was not identified in the primary tumor biopsy [256]. Interestingly, in their study, authors report mutations in *ARID1A* as subclonal, whereas the PCAWG study [240] as well as in the GEL EC cohort study, found the *ARID1A* mutations to be largely clonal. Further investigation into the timing of *ARID1A* and other mutations in larger cohorts and using higher fidelity methods alongside validation using experimental assays is needed to understand the timing of those alterations.

Conversely, there is more mature and established research into colorectal [404–406], prostate [407, 408] and renal [409] cancer evolution. For example, a recent study into prostate cancer evolution revealed prostate cancer disease types that arise as a result of divergent trajectories of a stochastic evolutionary process in which specific genetic alterations can tip the balance toward convergence to either route [407]. In their work, authors use an assembly of ML and statistical methods to investigate the genomic landscape of the tumors thoroughly, describe divergent evolution and identify evotypes of prostate cancer. Applying the RNN method, which could potentially streamline the research in the future, could prove beneficial and would serve as a validation for the approach. All the mentioned studies can be used for more informed generation of synthetic data, as well as validation of the method in real-life scenarios. After the method has been validated using real data with known ground truth, such as prostate cancer, it can be applied to the high-quality data of the GEL EC cohort.

Aside from validation, the proposed method is not free from its inherent limitations. Firstly, the model relies on a separate method to estimate the CCF and currently does not assume any degree of noise in those measurements. Modification of the model's architecture, which would either allow the use of the VAF estimates directly or incorporate the uncertainties from the CCF estimation step, would improve the model. Furthermore, the proposed method can only be as good as the available data. Extensive batch effects, as well as imperfections in sequencing, low coverage, sequence biases, and other technical issues, can make the model inherently biased. Strategies that would explore the extent

of the bias would enhance the explainability of the model, as well as provide users with clearer information about the model's limitations. Moreover, the current design of the model assumes that it needs to be trained for each of its use cases, requiring access to a Graphics Processing Unit (GPU), which makes it less accessible to researchers. Furthermore, the model could potentially wrongly identify the passengers that frequently appear as a consequence of the mutational process as a result of driver mutation. Those instances would be difficult to distinguish and will require further modifications to the model, for example, by including inductive bias (i.e., domain knowledge regarding protein function or mutational consequences) that could additionally improve the interpretability and explainability of the method.

To further improve the model architecture, a deeper exploration of attention-based models is warranted. We saw that incorporating an attention layer enhances model performance by learning the distinction between drivers and passengers. Comparing the current models with simpler attention-based models or more sophisticated architectures, such as transformers, could be advantageous. Transformer models are cutting-edge machine learning architectures that utilize self-attention mechanisms to process and generate sequences of data, demonstrating exceptional capabilities in natural language understanding and generation (for example, that seen in ChatGPT [410]). Owing to their capabilities in capturing complex dependencies within textual data, transformers have achieved state-of-the-art performance across various natural language processing tasks. Previously adapted to genomic data, for instance, in predicting gene expression from histone modifications [411], transformer models hold potential for further improvement in predicting the next driver event and identifying evolutionary trajectories, thus paving the way for the development of an efficacious model. However, a key limitation of these models lies in their original design for processing complex, lengthy texts and reliance on vast amounts of data. Consequently, they may struggle to learn sequences in a cancer evolution context due to the limited available data. This challenge could potentially be addressed through transfer learning and sophisticated batch correction methods, which enable leveraging data from multiple experiments and possibly different cancer types. Nonetheless, such an

approach would necessitate further in-depth investigation.

While the method presented in this thesis appears promising and serves as a solid foundation, further exploration and validations are necessary before the model can be used to generate insight into the evolution of endometrial cancer.

Chapter 5

Conclusions and future work

5.1 Molecular landscape of endometrial cancer

This thesis recounts a comprehensive analysis of the largest endometrial cancer (EC) cohort to date, examining genomic alterations across all four molecular subtypes. The study validated previous findings, revealed novel drivers, copy number alterations, and prognostic associations, and provided the first description of whole-genome landscapes in EC. Mutational processes diverge among subtypes, often converging in their alterations, such as the prevalence of nonsense mutations in POLE samples, indel mutations in MSI samples, and copy number losses in CN high samples, jointly affecting shared driver genes between subtypes. We observe an increase in the number of mutations in a subset of samples harboring a mutation in *POLD1* on top of their MSI signature and in those with *POLE* mutation. The results of the coding drivers analysis identified 107 driver genes, including new candidates. The significant copy number analysis revealed, among others, a significant recurrent deletion in MSI and POLE samples at 15p11.2 containing *SIN3A*, a transcriptional regulatory protein whose suppression was implicated in promoting metastasis in breast cancer [175].

Clear cell endometrial carcinoma, not included in the TCGA study and represented by 3% of samples in the GEL EC cohort, remains less studied than other EC cases. Over 80% of CCEC patients are characterized by p53 abnormality or by having no specific molecular profile (NSMP) [412]. The molecular classification of CCEC, particularly of

the NSMP group, is vital for determining prognosis and guiding therapy. More research in this area is needed to address the need for further understanding of the disease and to better adjust therapeutic strategies.

A natural progression of the study on the GEL EC cohort would include exploring additional data modalities. Incorporating transcriptomic data alongside extensive genomic analyses could help elucidate the consequences of specific genomic backgrounds on gene expression profiles. Furthermore, integrating the genomic data with methylation profiles or, ideally, chromatin accessibility data would offer a more in-depth understanding of the regulatory mechanisms underpinning endometrial cancer. This multi-modal approach could provide a more comprehensive picture of the biology of the disease, ultimately informing future research and therapeutic strategies.

Building upon the foundation laid by initiatives such as TCGA, I have shown the benefits of employing WGS at high coverage in the GEL EC cohort. For example, the use of WGS has proven beneficial in identifying regions of the exome not covered by the TCGA WES assay, exemplified by the inclusion of the beginning of the *ARID1A* sequence region. Taking into account the technological advancements, the multi-omic approach to the GEL EC cohort has a true potential of identifying new insights.

Furthermore, the increased representation of the CN low subtype, sometimes referred to as samples with no specific molecular profile, in the GEL EC cohort makes it a great candidate cohort for further investigation. This subtype could potentially be defined by an interplay of genomic and other features that could contribute to a deeper understanding of the disease and refine therapeutic approaches.

Moreover, recent publications, exemplified by [413], highlight the benefits of jointly modeling multiomic features to refine molecular subtyping and enhance clinical relevance. Similarly, a recent study of response to neoadjuvant chemotherapy in high-grade serous ovarian cancer [414] has shown that integrating molecular data with radiomics can successfully predict chemotherapy response, suggesting potential pathways for improving therapeutic outcomes through data integration. These advancements underline the value of a comprehensive, multi-modal approach in elucidating the complex biology of endome-

trial cancer and tailoring more effective treatment strategies.

Through informal discussions, I have become aware of plans to conduct similar multi-omic integration projects in the Genomics England cohort. However, those plans have yet to be formalized and funded. I hope that this project will be funded and the study will be made possible as this would benefit our understanding of the disease and help inform research in the space of modern therapies.

In the thesis introduction, I've discussed the research into heritable genetic conditions. These conditions target genes that are integral to the pathogenesis of endometrial cancer, underscoring the necessity of integrating the study of germline variations with analyses of somatic mutations in cancer tissues. Such a comprehensive approach could enhance our understanding of the molecular landscape of endometrial cancer, particularly in elucidating the origins of its molecular subtypes, including the interplay between MSI and POLE mutations reported in our research.

5.2 Mutations in drivers are predictive of survival

This study demonstrated the prognostic value of tumor mutation burden (TMB) and the fraction of the aberrant genome (percent genome aberrant, PGA) for survival outcomes in endometrial cancer, even after adjusting for age and molecular subtype. I also investigated the relationship between several driver genes and patient survival, finding that, for instance, mutations in *FBXW7* were associated with decreased overall survival across all examined models (i.e., univariate and adjusted for age, molecular subtype, and TMB). More complex relationships were observed for *CHD4*, *B2M*, and *TP53*, while mutations in *PIK3CA* and *PIK3R1* appeared to increase survival.

The findings from the survival analyses will undergo further validation in other cohorts, such as the PORTEC study [18], which will augment our comprehension of the relationship between cancer drivers and survival in endometrial cancer. Due to the limited power of the GEL EC cohort data, a multivariate model could not be fitted. Future research should investigate the application of a multivariate model combined with regularization, which could assist in identifying the minimal set of variables necessary for accurate survival

prediction. Moreover, an analysis informed by the types of alterations - for instance, focusing on oncogenic alterations - could further emphasize the prognostic value of these alterations.

The utility of translating these findings into clinical practice lies in their potential to improve patient outcomes. Hopefully, by enabling more accurate risk stratification, healthcare providers could potentially offer personalized treatment plans that are directly informed by the molecular characteristics of a patient's tumor. This approach, in line with an adaptation of precision medicine in oncology, has the potential to improve survival rates and quality of life for patients with EC.

Expanding our datasets to include larger sample sizes encompassing germline variations discussed in the previous section, alongside detailed clinical and molecular data, would enable more nuanced survival analyses. This integration promises not only to identify additional predisposing variants but also to correlate specific variants with distinct molecular subtypes of EC. Moreover, aligning these genetic insights with clinical outcomes could facilitate more precise survival predictions, potentially uncovering differential response rates to treatment based on genetic predisposition. Such advancements could significantly contribute to personalized medicine, offering targeted therapeutic strategies and improving patient outcomes in endometrial cancer.

5.3 Chromatin modifiers are frequently altered in endometrial cancer

Motivated by earlier findings of frequent mutations in selected chromatin modifiers, I conducted a systematic investigation of genetic defects in chromatin modifiers in EC. Chromatin modifiers are frequently altered across all four molecular subtypes. Coding drivers analysis revealed 15 chromatin modifiers identified as drivers, including four newly identified as candidate EC drivers - *SMARCA4*, *SETD1B*, *CREBBP* and *SETDB1*.

Alterations in key chromatin modifiers are prevalent. For example, EC samples harbor a p.T204 hotspot frameshift mutations in *CTCF*. Alterations in chromatin modifiers differ

in nature between the molecular subtypes - as per their underlying molecular processes, MSI samples are characterized by frameshift mutations, while CN high samples by losses and gains. For instance, alterations in *ARID1A* underline the differences in underlying molecular processes between subtypes - mutations in *ARID1A* in MSI samples are dominated by the frameshift indels, in POLE samples by hotspot nonsense mutations and losses in CN high samples. Interestingly, CN low samples include nonsense and frameshift mutations, as well as losses of hterezogyosity. This serves as an example of convergence in altering the drivers shared between molecular subtypes.

Alterations in chromatin modifiers are not exclusive to a single member of the complex but rather present across different subunits. In the context of endometrial cancer, the SWI/SNF pathway is commonly altered. In addition to alterations in *ARID1A* and its paralog *ARID1B*, I identified several other frequently altered pathway members. For instance, *SMARCA4* is altered in 23% of EC samples and in 14 (7.3%) of CN low samples all copies of the gene are lost. Furthermore, I found that samples with mutations in SWI/SNF subunits *PBRM1*, *SMARCA2*, and *SMARCA4* are associated with decreased survival. Alterations are also abundant in other chromatin remodeling complexes, such as NuRD and PRC1.1, and chromatin modifiers, such as p300/CBP. The correlative analysis of alterations in chromatin modifiers provides further evidence of the significance of chromatin organization and its regulation in endometrial cancer.

5.4 Endometrial cancer hetergoneity

In my research within the GEL EC cohort, I also investigated the heterogeneity of EC tumors, which offered valuable insights into the clonality of mutations. The analysis revealed increased fractions of clonal mutations in some cases, such as BCOR N1459S, but not in others. I highlighted the heterogeneity across various driver genes and chromatin modifiers.

Future research should prioritize understanding the heterogeneity across all alterations, encompassing copy number changes. A more comprehensive analysis, including the ordering of alterations, is necessary to unravel the evolution of EC. In this thesis, I proposed

a deep learning approach for deciphering evolutionary trajectories from whole-genome sequencing data as a potential solution.

5.5 *ARID1A* loss in normal endometrium cell lines causes hormone response imbalance

A functional study of *ARID1A* loss in the hEM3 cell line, a model for normal endometrium, showed an activation of the TGF-β pathway and genes associated with the mitotic spindle, possibly suggesting an increased proliferation. Moreover, the experiment revealed an up-regulation of the androgen response and a down-regulation of the estrogen response, suggesting dysregulation of hormone response pathways in the context of *ARID1A* loss.

Further work should focus on confirming those findings. A good place to start would include a proliferation assay to test the increased proliferation hypothesis. It would also be beneficial to conduct a rescue phenotype study and investigate the expression profiles and chromatin accessibility. Examining androgen (AR) and estrogen receptor (ER) binding on chromatin could help elucidate the implications of hormone imbalance. Exploration of transcription factor footprinting, with a focus on AR and ER, is also recommended.

5.6 Functional consequences of perturbations of chromatin modifiers in EC

I presented the results of functional work to understand the consequences of chromatin modifier loss in endometrial cancer. Guided by our descriptive analyses of chromatin modifier alterations in the GEL EC cohort, we aimed to understand the role of these chromatin modifiers in endometrial cancer cell biology. While *ARID1A* and its paralog *ARID1B* have been studied in endometrial cancer models, little is known about the consequences of mutations in *ARID5B*, *EP300*, *KMT2C*, and *SETD1B*. This study was the first attempt to understand the implications of the loss of all six chromatin modifiers.

The analysis across the chromatin modifiers knockouts revealed dysregulation of criti-

cal pathways such as KRAS signaling, TNF-α/NF-κB signaling, interferon responses, and down-regulation of epithelial-mesenchymal transition (EMT). These findings highlight the consequences of chromatin modifier loss on chromatin accessibility and gene expression profiles, as well as the potential compensatory mechanisms cancer cells employ in response to the loss of chromatin modifiers.

CUXs motifs were found more accessible in Ark2 Cas9 *ARID1A*^{KO}. Although the role of CUXs transcription factors has not been described as yet, there is evidence that CUX1 regulates PIK3CA and promotes cell growth and metastasis via activating EMT in bladder cancer [269].

Although EMT was found down-regulated across all investigated chromatin modifiers in EC-derived cell lines, several genes in the pathway were up-regulated, *SPOCK1* being the most notable example. Expression of *SPOCK1* was up-regulated in all conditions except for *SETD1B* loss. Although its role in endometrial cancer remains to be described, *SPOCK1* was found to promote breast cancer progression [313], pancreatic cancer metastasis [314], and colon cancer proliferation and migration [315]. I identified a potential enhancer site accessible in the Ark2 Cas9 *ARID1B*^{KO} but not in the parental cell line, while in Ark2 Cas9 *ARID1A*^{KO} cells, a CUXs motif was found in the intronic region of the gene sequence, accessible in the knockout but not in Ark2 Cas9 parental cells.

The loss of *ARID5B* was associated with the up-regulation of interferon (INF) response. Both INF-α and INF-γ response pathways were up-regulated in Ark2 Cas9 *ARID5B*^{KO}. Interestingly, despite the up-regulation of these pathways, certain genes, such as *ADAR*, were down-regulated. Loss of ADAR has been implicated in enhancing cancer cells' sensitivity to immunotherapy and promoting increased inflammation [292].

Remarkably, two mTOR pathways were found to be dysregulated in *ARID5B* deficient cells: mTORC1 signaling was up-regulated, while PI3K/Akt/mTOR signaling was down-regulated. The down-regulation of mTORC1 signaling and the up-regulation of PI3K/Akt/mTOR signaling could represent a compensatory response by the cell to preserve cellular growth, proliferation, and survival. In this context, the cell might attempt to counteract the reduced mTORC1 activity by increasing the activation of the PI3K-

Akt pathway, which can subsequently activate mTOR. Additionally, the up-regulation of PI3K/Akt/mTOR signaling could reflect the cell's effort to overcome negative feedback and restore the pathway's activity. Nonetheless, further experimentation and analysis are necessary to draw definitive conclusions on this dysregulation.

AP1 is a dimeric transcription factor composed of Fos and Jun family members. In the case of *EP300* loss, the JUND motif exhibits increased accessibility. Although research on JunD's specific role in EC is limited, JunD has been identified as a critical component in Ras-induced tumorigenesis in lung cancer [300]. The observed up-regulation of the KRAS signaling pathway, along with the increased accessibility of JUND motifs in *EP300* loss, may suggest that *EP300* contributes to the up-regulation of the KRAS pathway by making JunD binding sites more accessible in EC.

Interestingly, the loss of *KMT2C*, a member of the KMT2 (MLL) family of methyltransferases, leads to the up-regulation of E2F target genes. In contrast, the loss of *SETD1B*, also a member of the KMT2 family, results in the down-regulation of E2F target genes. Despite the opposing directions of dysregulation, several genes are commonly dysregulated in both *KMT2C* and *SETD1B* deficient cells. The concurrent up-regulation of specific E2F target genes and down-regulation of numerous others may imply a complex interaction among altered cell cycle regulation, DNA repair, and genomic instability. This observation could suggest that cancer cells are attempting to strike a balance between uncontrolled proliferation and the maintenance of genomic integrity.

To further investigate the consequences of common mutations in chromatin modifiers and beyond, high-throughput assays followed by more in-depth experiments are recommended. This could be achieved using high-throughput single-cell CRISPR screens, such as Perturb-ATAC [328], sci-ATAC [329], or Spear-ATAC [330]. All three methods involve cells stably expressing Cas9, a library of gRNAs and rely on the simultaneous readout of chromatin accessibility and gRNA detection at the single-cell level. Such experiments could facilitate the investigation of chromatin accessibility perturbations as a result of knockdowns of selected genes. An ideal experiment would include a modification of the method, enabling the investigation of not only knockdowns but also the replication of

common mutations, such as prevalent hotspots, gain-of-function mutations, and others. Subsequently, more specialized investigations of selected mutations would be warranted to further our understanding of their impact on chromatin modifiers and endometrial cancer.

An additional avenue worth exploring involves investigating the combination of mutations. As demonstrated in this research, chromatin modifier mutations rarely occur in isolation. Although a targeted experiment focusing on a double-knockout of *ARID1A* and *ARID1B* could serve as an excellent starting point, further identification of targets should be preceded by an in-depth analysis of mutation co-occurrence and mutual exclusivity. It is worth noting that such analysis needs to employ a method capable of incorporating the hypermutated and ultramutated context of the MSI and POLE EC subtypes. By examining the interplay between these co-occurring mutations, researchers may better understand their impact on endometrial cancer and chromatin modifier function.

5.7 *BCOR* N1459S hotspot mutation is near-exclusive to EC

Almost 1 in 10 samples (9.7%) of the GEL EC cohort carry N14559S BCOR hotspot mutation. In Pancancer TCGA cohort, the N1459S mutation is near unique to endometrial cancer cohort (χ^2 test, p-value $= 4.998 \times 10^{-4}$).

The functional study in the mouse embryonic fibroblast cells with mutation generated with CRISPR-Cas9 yielded inconclusive results. This could be due to the fact that the mES cells used in the study appeared to have recovered from the perturbations to the chromatin. The study of the acute knockdown of *BCOR* in hES showed that the depletion of *BCOR* prompted endodermal and mesodermal differentiation of hES.

To address the discrepancy fo the results between our and the reported study, an alternative system for introducing the *BCOR* point mutation, such as a tetracycline-inducible system, could be explored. A time course experiment under differentiation-permitting conditions could be considered to detect differences in differentiation. Additionally, using chromatin accessibility or DNA-protein binding techniques could provide insight into the

implications of the mutation. Furthermore, it is also possible that the effect of the mutation depends on cellular processes that are unique or abundant in endometrial tissue. To investigate this, the mutation's effects should be studied in human endometrial cell lines and organoids, or a mouse knock-in should be performed.

5.8 RNN-based approach to uncover cancer evolutionary trajectories

I presented an idea to build an RNN-based approach to uncover cancer evolutionary trajectory leveraging the solutions designed for large texts and sequential data.

Although the preliminary findings show potential, additional research is necessary to establish a more sophisticated and reliable method. Despite being a suitable starting point, the synthetic datasets currently employed are significantly shorter and less noisy than real-world data. It is recommended to generate synthetic data that more accurately reflect actual conditions, utilizing well-documented evolutionary trajectories such as those observed in colorectal [415] and prostate [416] cancers.

A broader examination of attention-based models is advised. By incorporating an attention layer, the model's performance is enhanced by learning to distinguish between driver and passenger mutations. Evaluating the existing models against simpler attention-based models or more advanced architectures, like transformers, could provide valuable information. Transformers are cutting-edge machine learning architectures that use self-attention mechanisms to process and generate data sequences, demonstrating remarkable natural language understanding and generation capabilities. Due to their ability to capture complex dependencies within textual data, transformers have achieved top performance across various natural language processing tasks. Previously adapted for genomic data, such as predicting gene expression from histone modifications, transformer models can improve predictions of subsequent driver events and the identification of evolutionary trajectories, ultimately resulting in a highly effective model. However, one major limitation of these models is their original design for managing intricate, lengthy texts

and dependence on vast amounts of data. Consequently, they may struggle to learn sequences in a cancer evolution context due to the limited available data. This challenge could potentially be addressed through transfer learning and advanced batch correction techniques, enabling the use of data from multiple experiments and possibly different cancer types. Nonetheless, adopting such an approach would necessitate further in-depth investigation.

5.9 Final remarks

In summary, this thesis offers insights into the role of chromatin modifiers in endometrial cancer. It presents the first comprehensive descriptive analysis of alterations in chromatin modifiers and, for the first time, associates mutations in several driver genes with changes in survival.

Additionally, through utilizing transcriptomic and chromatin accessibility data in tailored isogenic cell line experiments, this work provides an unprecedented exploration into the effects of common mutations in chromatin modifiers on endometrial cancer cells. Collectively, these findings contribute to an enhanced understanding of chromatin modifier dysregulation in endometrial cancer.

Lastly, in this work, I introduce the concept of employing a Recurrent Neural Network (RNN) based approach for uncovering evolutionary trajectories in cancer. The described methodology serves as a solid foundation for future research.

Appendices

Appendix A

Software and packages

A.1 Software

The analyses were performed using the software listed in Table A.1. Custom pipelines were made available on GitHub. The notebooks with code and data needed to produce figures and analysis will be released together with the final version of the plots upon completion of the project.

Software name	Citation
R	[236]
Python	[417]
Bash	[418]
Snakemake	[339]
BWA	[340]
Picard	[341]
Trim Galore!	[334]
Mutect2	[419, 420]
FastQC	[332]
Salmon	[335]
STAR	[336]
vcf2maf	[228]
Strelka	[227]
ISAAC	[226]
Cutadapt	[333]
MultiQC	[337]
DPClust	[193, 239]
Battenberg	[193]
pygenometracks	[347]

Table A.1: List of Software and Programming Languages used in the analyses.

A.2 R packages

Most of the exploratory data analyses and plotting have been carried out in R [236] and the invaluable packages are listed in Table A.2.

package name	citation
ATACseqQC	[421]
BiocManager	[422]
ChIPseeker	[423]
clusterProfiler	[424]
ComplexHeatmap	[355]
DESeq2	[248]
glue	[425]
ggthemes	[426]
ggrepel	[427]
ggsci	[428]
ggtext	[429]
gtsummary	[430]
gmodels	[431]
Gviz	[432]
knitr	[433–435]
limma	[360]
lubridate	[436]
maftools	[437]
patchwork	[438]
pbmcapply	[439]
RColorBrewer	[440]
survival	[237]
survminer	[238]
TCGAbiolinks	[159]
tidyverse	[441]
TxDb.Hsapiens.UCSC.hg38.knownGene	[442]
viridis	[443]

Table A.2: List of R packages used in the analyses. Packages commonly used in exploratory data analysis and creating visualizations are highlighted in bold.

A.3 Python modules

module name	citation
keras	[444]
numpy	[445]
pandas	[446]
scipy	[447]
tensorflow	[448]

Table A.3: List of Python packages used while working on this thesis.

Appendix B

Supplementary Information

Contents

B.1 Chapter 3 248
B.1.1 *BCOR* N1459S hotspot mutation experiment 248
B.1.2 Chromatin complexes definitiions 250
B.2 Chapter 4 253
B.2.1 Statistical approach 253
B.2.2 Deep Learning approach 255

B.1 Chapter 3

B.1.1 *BCOR* N1459S hotspot mutation experiment

Sequences of the guide RNAs and primers used to generate mES $Bcor^{KO}$ and mES $Bcor^{N1463S}$ are presented in Table S1. All sequences are annotated in reference to the sequence accession number NC_000086, created on 16/07/2019 and last modified on 20/03/2020.

Primer name	Primer sequence	Length	Annotation
Y-062-R-mBCOR-geno	gggggtgttaaagggtcaca	20-mer	Binds at: 10,413 <- 10,432
Y-064-R-mBCOR-geno-2	tctgtcctcttgtcctggcc	20-mer	Binds at: 11,249 <- 11,268
Y-061-F-mBCOR-geno	acaaagagccaacagccaga	20-mer	Binds at: 9941 ->9960
Primer 3	atgaacagcgagagggtcc	19-mer	Binds at: 24,811 <- 24,829
Test 1	GGGGACAACTTTGTATACAAAAGTTGTaactcagcgggttacgtgggtgacc	52-mer	Binds at: 19,679 <- 19,704
Y-001-F-BCoR_Exon4	GTATATAGCACTGAAGCCATTTGGG	25-mer	Binds at: 22,173 ->22,197
Y-005-F-BCor-N1459S Mut	cggagacttattgtcaGtaagaacgctggcg	31-mer	Binds at: 10,158 <- 10,188
Y-006-R-BCor-N1459S Mut	cgccagcgttcttaCtgacaataagtctccg	31-mer	Binds at: 10,158 ->10,188
Y-007-F-GW-BCoR	GGGGACAAGTTTGTACAAAAAAGCAGGCTTAGAAGGAGATAGAACCATGctctcagcaacccccctgtatgg	72-mer	Binds at: 24,846 <- 24,872
Y-009-F-BCoR	atgctctcagcaaccccc	18-mer	Binds at: 26,269 ->26,279; 102,881 <- 102,895
Y-012-F-mBCoR-N1459S Mut	ctcggagacttattgtcaGtaagaatgctggcga	34-mer	Binds at: 10,157 <- 10,190
Y-013-F-mBCoR-seq1	cctgccagtgacaaacagagcc	22-mer	Binds at: 22,088 <- 22,109
Y-015-F-mBCoR-seq3	agtctgtttcctgggcaccttgc	23-mer	Binds at: 20,527 <- 20,549
Y-016-F-mBCoR-seq4	agtgagggccctgctgtgact	21-mer	Binds at: 19,866 <- 19,886
Y-019-F-mBCoR-seq7	gatacctatagcgatttgtttga	23-mer	Binds at: 2239 <- 2261
Y-022-F-hBCor-Seq4	aacccaagaggatcccaaaccattt	25-mer	Binds at: 19,833 <- 19,857
Y-023-R-hBCoR-Seq4	aaatggtttgggatcctcttgggtt	25-mer	Binds at: 19,833 ->19,857
Y-023bis-R-hBCoR-Seq4	cccacacaaaatggtttgggatcctcttgggtt	33-mer	Binds at: 19,825 ->19,857
Y-025-R-hBCoR-Seq3	agtggcttatctgtgacgtctttgg	25-mer	Binds at: 21,298 ->21,322
Y-031-R-qP-BCoR-202-C	cacgaccagccttccag	17-mer	Binds at: 63,512 <- 63,522
Y-032-R-qP-BCoR-202-D	gacctgaatgcctcatttggag	22-mer	Binds at: 20,674 ->20,695
Y-033-R-qP-BCoR-A	atgaacagcgagagggtcc	19-mer	Binds at: 24,811 <- 24,829
Y-035-F-qP-BCoR-C	ctcagtgctgccaggcca	18-mer	Binds at: 22,408 ->22,425; 39,640 <- 39,649
Y-054-F-mBCOR-A	gagcacgtctgagagcagaag	21-mer	Binds at: 1148 ->1168
Y-055-R-mBCOR-A	ccccaggaccagaagaccc	19-mer	Binds at: 2275 <- 2293
Y-056-F-mBCOR-B	ggggcttcaaagagatcatg	20-mer	Binds at: 12,474 ->12,493
Y-057-R-mBCOR-B	gactgggaaaggttgaaagg	20-mer	Binds at: 18,304 <- 18,323
Y-063-F-mBCOR-geno-2	gtgcagagaacagcagttgc	20-mer	Binds at: 9258 ->9277

Table S1: **Sequences of the primers used in the BCOR experiment.** Sequences of primers used to generate and confirmed mES $Bcor^{KO}$ and mES $Bcor^{N1463S}$.

B.1.2 Chromatin complexes definitiions

The following tables list the proteins/genes comprising the SWI/SNF complex (Table S2) and NuRD complex (Table S3). Both tables are available in the *.tsv* format in the Supplement archive distributed together with this document.

HUGO Symbol	Protein name	Subunit category	Complex	Exclusivness
SMARCA2	BRM	Helicase/ATPase	found in only mSWI/SNF	SMARCA4
SMARCA4	BRG1	Helicase/ATPase		SMARCA2
SMARCB1	INI1	Core subunit		
SMARCC1	BAF155	Core subunit		
SMARCC2	BAF170	Core subunit		
SMARCD1	BAF60A	Accessory subunit		SMARCD2, SMARCD3
SMARCD2	BAF60B	Accessory subunit		SMARCD1. SMARCD3
SMARCD3	BAF60C	Accessory subunit		SMARCD1, SMARCD2
SMARCE1	BAF57	Accessory subunit		
ACTL6A	BAF53A	Accessory subunit		ACTL6B
ACTL6B	BAF53B	Accessory subunit		ACTL6A
ARID1A	BAF250A	Accessory subunit	found in only mSWI/SNF	ARID1B
ARID1B	BAF250B	Accessory subunit	found in only mSWI/SNF	ARID1A
PBRM1	BAF180	Accessory subunit	found in only non-canonical SWI/SNF	
ARID2	BAF200	Accessory subunit	found in only non-canonical SWI/SNF	
BRD7	BRD7	Accessory subunit	found in only non-canonical SWI/SNF	
ACTB	ACTB	Accessory subunit		
PHF10	BAF45A	Accessory subunit		BAF45B, BAF45C, BAF45D
DPF1	BAF45B	Accessory subunit		BAF45A, BAF45C, BAF45D
DPF3	BAF45C	Accessory subunit		BAF45A, BAF45B, BAF45D
DPF2	BAF45D	Accessory subunit		BAF45A, BAF45B, BAF45C

Table S2: **SWI/SNF complex subunits.** If subunit is exclusive for a given complex it is noted in the Complex column. Main subunits listed, mutually exclusive subunits noted in the Exclusiveness column. Adapted from [449]. mSWI/SNF - mammalian SWI/SNF = canonical SWI/SNF

HUGO Symbol	Full name	Previous symbols	Aliases	Chromosome
CHD3	chromodomain helicase DNA binding protein 3		Mi-2a,ZFH,Mi2-ALPHA	17p13.1
CHD4	chromodomain helicase DNA binding protein 4		Mi-2b,Mi2-BETA	12p13.31
CHD5	chromodomain helicase DNA binding protein 5			1p36.31
HDAC1	histone deacetylase 1	RPD3L1	HD1,GON-10,KDAC1	1p35.2-p35.1
HDAC2	histone deacetylase 2		RPD3,YAF1,KDAC2	6q21
MBD2	methyl-CpG binding domain protein 2			18q21.2
MBD3	methyl-CpG binding domain protein 3			19p13.3
MTA1	metastasis associated 1			14q32.33
MTA2	metastasis associated 1 family member 2	MTA1L1	MTA1-L1	11q12.3
MTA3	metastasis associated 1 family member 3		KIAA1266	2p21
RBBP4	RB binding protein 4, chromatin remodeling factor		RbAp48,NURF55,lin-53	1p35.1
RBBP7	RB binding protein 7, chromatin remodeling factor		RbAp46	Xp22.2

Table S3: **NuRD complex.** Gene list from HUGO Gene Nomenclature Committee (HGNC) website [450], accessed on 3/09/2020.

B.2 Chapter 4

B.2.1 Statistical approach

Estimating evolutionary trajectories remains a complex challenge, as only a limited number of published methods have moved beyond human-curated approaches. A most recent method, described in [381], employs a Monte Carlo approach to estimate plausible trees from patient-level CCF distribution and employs the Plackett-Luce model [382, 383] to generate a probability distribution over the relative orderings.

The method employs Monte Carlo simulations to generate a sampled ordering. This process involves two steps: first, a single plausible tree is sampled from the data, and then, based on this tree, alterations are subsampled to create an ordering corresponding to a selected branch. This approach effectively captures the inherent variability and uncertainty in the underlying data. Furthermore, the authors used a Plackett-Luce mixture model [384] to determine whether subpopulations in the data exhibit different orderings.

B.2.1.1 The Plackett-Luce model

Given a set of alterations ordered according to their CCF, we would like to infer the order in which these events generally occur. To do this, the authors of [381] use a Plackett-Luce model, formulated as a *ranking* method, that returns a value quantifying the ranking preference. However, the authors use an *ordering* formally defined as the inverse of the ranking preference mapping [384]. Essentially, instead of a set of rankings, they use a set of elements sorted by their ranks. It is worth noting the Plackett-Luce model does not return any temporal information outside the expected order of events. We have a set of N alterations:

$$A = \{a_1, a_2, \ldots, a_N\},$$

then we can apply Luce's choice axiom [382, 451]. The axiom states that the probability of selecting one event over another from a set of events is independent of the presence or absence of the other events in the set. We can therefore write the probability

of observing event i as

$$P(c_i|A) = \frac{\alpha_i}{\sum_j \alpha_j},$$

where $\{\alpha_i\}$ are the coefficients that quantify the relative probability of observing the i^{th} event. To reflect the ordering aspect of our application, we refer to this value as the *proclivity*. Plackett [383] used this formalism to construct a generative model in which all N events are randomly sampled from A without replacement (i.e., a *permutation*). If we let Λ correspond to a permutation of the set A such that $\lambda_k \in A$ and $\lambda_1 \prec \lambda_2 \prec \ldots \prec \lambda_N$, then we write the probability density of a single ordering as

$$P(\Lambda) = \prod_k^N \frac{\alpha_{\lambda_k}}{\sum_{j \in \Lambda^{(k)}} \alpha_j},$$

where α_{λ_k} is the proclivity associated with event λ_k, and $\Lambda^{(k)} = \{\lambda_k, \lambda_{k+1}, \ldots, \lambda_N\}$ is the set of possible events after $k-1$ events have occurred.

B.2.1.2 The Plackett-Luce mixture model

To account for the possible multiple trajectories, the authors use PLMIX - a mixture Plackett-Luce model as described in [384]. PLMIX assumes that the population consists of a finite number, G, of distinct subpopulations. By reinstating Luce's choice axiom, the selection of each CNA (copy number alteration) can be considered independent, conditional on the mixture component. The probability of observing the ordering Λ_s for the s^{th} sample is expressed as

$$P(\Lambda_s) = \sum_g^G \omega_g P_g(\Lambda_s),$$

where ω_g denotes the weight parameters that quantify the probability that sample s belongs to subgroup g. Maximum likelihood estimation via an EM algorithm is employed to determine the appropriate parameter values [384]. Furthermore, the Bayesian Information Criterion (BIC) estimation is used to select the number of mixture components, with the formula

$$BIC = N \log(M) - 2\ell(\Theta_{ML}),$$

where Θ_{ML} represents the parameter set maximizing the log-likelihood $\ell(\cdot)$, N is the number of parameters, and M is the number of samples.

B.2.2 Deep Learning approach

Model	Short driver sequence	Long driver sequence	Early diverging driver sequence	Long driver sequence (more bg)
1xRNN	0.36	0.25	0.25	0.10
2xRNN	0.36	0.21	0.20	0.06
1xLSTM	**0.69**	**0.62**	**0.59**	**0.40**
2xLSTM	0.66	0.46	0.45	0.34

Table S4: **Accuracy of predicting the next driver event - models without attention.** The best accuracy for a given data set is marked in bold.

Model	Short driver sequence	Long driver sequence	Early diverging driver sequence	Long driver sequence (more bg)
1xRNN	0.67	**0.79**	**0.81**	**0.67**
2xRNN	0.58	0.68	0.62	0.50
1xLSTM	**0.71**	0.52	0.54	0.43
2xLSTM	0.68	0.61	0.66	0.52

Table S5: **Accuracy of predicting the next driver event - models with attention.** The best accuracy for a given data set is marked in bold.

			1xLSTM			2xLSTM		
	Penalty	Coefficient	SD	LD	ED	SD	LD	ED
baseline			0.71	0.52	0.63	0.68	0.54	0.70
	L1	0.0001	0.73	0.57	0.64	0.69	0.58	0.67
	L1	0.001	0.66	0.54	0.58	0.53	0.42	0.48
	L1	0.01	0.46	0.28	0.33	0.48	0.30	0.33
	L2	0.0001	0.72	0.52	0.64	0.66	0.54	0.66
kernel	L2	0.001	0.68	0.60	0.62	0.53	0.40	0.69
	L2	0.01	0.67	0.53	0.58	0.53	0.49	0.54
	L1L2	0.0001	0.66	0.56	0.62	0.71	0.48	0.66
	L1L2	0.001	0.72	0.50	0.54	0.44	0.40	0.51
	L1L2	0.01	0.48	0.29	0.33	0.44	0.30	0.34
	L1	0.0001	0.63	0.51	0.56	0.65	0.53	0.72
	L1	0.001	0.69	0.55	0.61	0.66	0.63	0.72
	L1	0.01	0.67	0.53	0.59	0.65	0.55	0.57
	L2	0.0001	0.69	0.53	0.67	0.66	0.52	0.64
bias	L2	0.001	0.63	0.55	0.65	0.70	0.59	0.65
	L2	0.01	0.66	0.51	0.61	0.66	0.47	0.67
	L1L2	0.0001	0.71	0.58	0.70	0.64	0.57	0.70
	L1L2	0.001	0.68	0.53	0.59	0.61	0.58	0.57
	L1L2	0.01	0.69	0.52	0.61	0.68	0.59	0.62
	L1	0.0001	0.62	0.61	0.70	0.68	0.50	0.64
	L1	0.001	0.68	0.56	0.67	0.69	0.44	0.53
	L1	0.01	0.68	**0.70**	**0.78**	0.73	0.47	0.45
	L2	0.0001	0.69	0.58	0.68	0.69	0.57	0.69
activity	L2	0.001	0.71	0.66	0.74	0.64	0.62	**0.77**
	L2	0.01	0.74	0.67	0.77	**0.79**	**0.66**	**0.77**
	L1L2	0.0001	0.68	0.60	0.71	0.66	0.53	0.76
	L1L2	0.001	0.68	0.61	0.76	0.74	0.46	0.62
	L1L2	0.01	**0.81**	0.69	0.74	0.66	0.36	0.42
		0.1	0.72	0.64	0.65	0.72	0.55	0.70
		0.2	0.69	0.60	0.67	0.66	0.64	0.67
		0.3	0.68	0.65	0.67	0.71	0.58	0.71
		0.4	0.64	0.56	0.69	0.65	0.61	0.73
dropout		0.5	0.58	0.63	0.67	0.68	0.62	0.71
		0.6	0.63	0.52	0.64	0.72	0.59	0.74
		0.7	0.61	0.57	0.64	0.70	0.55	0.72
		0.8	0.62	0.54	0.61	0.71	0.58	0.72

Table S6: **Accuracy of models in predicting the next driver event with respect to regularization methods.** The highest values in each column are shown in bold.

Bibliography

[1] J. Ferlay, M. Colombet, I. Soerjomataram et al. Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, 149(4):778–789, 2021. doi: 10.1002/ijc.33588.

[2] R. L. Siegel, K. D. Miller, N. S. Wagle et al. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1):17–48, 2023. doi:10.3322/caac.21763.

[3] H. Sung, J. Ferlay, R. L. Siegel et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi:10.3322/caac.21660.

[4] B. Gu, X. Shang, M. Yan et al. Variations in incidence and mortality rates of endometrial cancer at the global, regional, and national levels, 1990–2019. *Gynecologic Oncology*, 161(2):573–580, 2021. doi:10.1016/j.ygyno.2021.01.036.

[5] J. V. Bokhman. Two pathogenetic types of endometrial carcinoma. *Gynecologic Oncology*, 15(1):10–17, 1983. doi:10.1016/0090-8258(83)90111-7.

[6] J. S. Berek, X. Matias-Guiu, C. Creutzberg et al. FIGO staging of endometrial cancer: 2023. *International Journal of Gynecology & Obstetrics*, 162(2):383–394, 2023. doi:10.1002/ijgo.14923.

[7] E. J. Crosbie, S. J. Kitson, J. N. McAlpine et al. Endometrial cancer. *The Lancet*, 399(10333):1412–1428, 2022. doi:10.1016/S0140-6736(22)00323-3.

[8] K. H. Lu and R. R. Broaddus. Endometrial Cancer. *New England Journal of Medicine*, 383(21):2053–2064, 2020. doi:10.1056/NEJMra1514010.

[9] M. Koskas, F. Amant, M. R. Mirza et al. Cancer of the corpus uteri: 2021 update. *International Journal of Gynecology & Obstetrics*, 155(S1):45–60, 2021. doi:10.1002/ijgo.13866.

[10] N. Concin, X. Matias-Guiu, I. Vergote et al. ESGO/ESTRO/ESP guidelines for the management of patients with endometrial carcinoma. *International Journal of Gynecologic Cancer*, 31(1), 2021. doi:10.1136/ijgc-2020-002230.

[11] G. Bogani, I. Ray-Coquard, N. Concin et al. Uterine serous carcinoma. *Gynecologic Oncology*, 162(1):226–234, 2021. doi:10.1016/j.ygyno.2021.04.029.

[12] G. Bogani, I. Ray-Coquard, N. Concin et al. Clear cell carcinoma of the endometrium. *Gynecologic Oncology*, 164(3):658–666, 2022. doi:10.1016/j.ygyno.2022.01.012.

[13] D. A. Levine. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013. doi:10.1038/nature12113.

[14] S. R. Kim, B. T. Cloutier, S. Leung et al. Molecular subtypes of clear cell carcinoma of the endometrium: Opportunities for prognostic and predictive stratification. *Gynecologic Oncology*, 158(1):3–11, 2020. doi:10.1016/j.ygyno.2020.04.043.

[15] A. D. Cherniack, H. Shen, V. Walter et al. Integrated Molecular Characterization of Uterine Carcinosarcoma. *Cancer Cell*, 31(3):411–423, 2017. doi:10.1016/j.ccell.2017.02.010.

[16] R. A. de Jong, H. W. Nijman, T. F. Wijbrandi et al. Molecular markers and clinical behavior of uterine carcinosarcomas: focus on the epithelial tumor component. *Modern Pathology*, 24(10):1368–1379, 2011. doi:10.1038/modpathol.2011.88.

[17] B. G. Wortman, T. Bosse, R. A. Nout et al. Molecular-integrated risk profile to determine adjuvant radiotherapy in endometrial cancer: Evaluation of the pilot phase of the PORTEC-4a trial. *Gynecologic Oncology*, 151(1):69–75, 2018. doi:10.1016/j.ygyno.2018.07.020.

[18] A. S. V. M. v. d. Heerik, N. Horeweg, R. A. Nout et al. PORTEC-4a: international randomized trial of molecular profile-based adjuvant treatment for women with high-intermediate risk endometrial cancer. *International Journal of Gynecologic Cancer*, ijgc, 2020. doi:10.1136/ijgc-2020-001929.

[19] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799, 2004. doi:10.1038/nm1087.

[20] A. Wong and J. Ngeow. Hereditary Syndromes Manifesting as Endometrial Carcinoma: How Can Pathological Features Aid Risk Assessment? *BioMed Research International*, 2015:e219012, 2015. doi:10.1155/2015/219012.

[21] J. Wijnen, W. d. Leeuw, H. Vasen et al. Familial endometrial cancer in female carriers of MSH6 germline mutations. *Nature Genetics*, 23(2):142–144, 1999. doi:10.1038/13773.

[22] K. H. Lu, J. O. Schorge, K. J. Rodabaugh et al. Prospective Determination of Prevalence of Lynch Syndrome in Young Women With Endometrial Cancer. *Journal of Clinical Oncology*, 25(33):5158–5164, 2007. doi:10.1200/JCO.2007.10.8597.

[23] S. Gordhandas, R. M. Kahn, C. Gamble et al. Clinicopathologic features of endometrial cancer with mismatch repair deficiency. 2020. doi:10.3332/ecancer.2020.1061. URL http://ecancer.org/en/journal/article/1061-clinicopathologic-features-of-endometrial-cancer-with-mismatch-repair-defic

[24] C. Palles, J.-B. Cazier, K. M. Howarth et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, 45(2):136–144, 2013. doi:10.1038/ng.2503.

[25] D. N. Church, S. E. Briggs, C. Palles et al. DNA polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. *Human Molecular Genetics*, 22(14):2820–2828, 2013. doi:10.1093/hmg/ddt131.

[26] D. N. Church, E. Stelloo, R. A. Nout et al. Prognostic Significance of POLE Proofreading Mutations in Endometrial Cancer. *JNCI: Journal of the National Cancer Institute*, 107(1):dju402, 2015. doi:10.1093/jnci/dju402.

[27] T. M. Starink, J. P. W. v. d. Veen, F. Arwert et al. The Cowden syndrome: a clinical and genetic study in 21 patients. *Clinical Genetics*, 29(3):222–233, 1986. doi:10.1111/j.1399-0004.1986.tb00816.x.

[28] J. A. Hobert and C. Eng. PTEN hamartoma tumor syndrome: An overview. *Genetics in Medicine*, 11(10):687–694, 2009. doi:10.1097/GIM.0b013e3181ac9aea.

[29] K. Yizhak, F. Aguet, J. Kim et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, 364(6444):eaaw0726, 2019. doi:10.1126/science.aaw0726.

[30] I. Martincorena and P. J. Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015. doi:10.1126/science.aab4082.

[31] I. Martincorena, J. C. Fowler, A. Wabik et al. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, 2018. doi:10.1126/science.aau3879.

[32] K. T. Zondervan, C. M. Becker and S. A. Missmer. Endometriosis. *New England Journal of Medicine*, 382(13):1244–1256, 2020. doi:10.1056/NEJMra1810764.

[33] V. Lac, T. M. Nazeran, B. Tessier-Cloutier et al. Oncogenic mutations in histologically normal endometrium: the new normal? *The Journal of Pathology*, 249(2):173–181, 2019. doi:10.1002/path.5314.

[34] K. Suda, H. Nakaoka, K. Yoshihara et al. Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Reports*, 24(7):1777–1789, 2018. doi:10.1016/j.celrep.2018.07.037.

[35] M. S. Anglesio, N. Papadopoulos, A. Ayhan et al. Cancer-Associated Mutations in Endometriosis without Cancer. *New England Journal of Medicine*, 376(19):1835–1848, 2017. doi:10.1056/NEJMoa1614814.

[36] V. Lac, L. Verhoef, R. Aguirre-Hernandez et al. Iatrogenic endometriosis harbors somatic cancer-driver mutations. *Human Reproduction*, 34(1):69–78, 2019. doi:10.1093/humrep/dey332.

[37] L. Moore, D. Leongamornlert, T. H. H. Coorens et al. The mutational landscape of normal human endometrial epithelium. *Nature*, 580(7805):640–646, 2020. doi:10.1038/s41586-020-2214-z.

[38] I. Martincorena, K. M. Raine, M. Gerstung et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.e21, 2017. doi:10.1016/j.cell.2017.09.042.

[39] M. L. Rudd, J. C. Price, S. Fogoros et al. A Unique Spectrum of Somatic PIK3CA (p110α) Mutations Within Primary Endometrial Carcinomas. *Clinical Cancer Research*, 17(6):1331–1340, 2011. doi:10.1158/1078-0432.CCR-10-0540.

[40] L. W. Cheung, B. T. Hennessy, J. Li et al. High Frequency of PIK3R1 and PIK3R2 Mutations in Endometrial Cancer Elucidates a Novel Mechanism for Regulation of PTEN Protein Stability. *Cancer Discovery*, 1(2):170–185, 2011. doi:10.1158/2159-8290.CD-11-0039.

[41] M. K. McConechy, J. Ding, M. C. Cheang et al. Use of mutation profiles to refine the classification of endometrial carcinomas. *The Journal of Pathology*, 228(1):20–30, 2012. doi:10.1002/path.4056.

[42] M. K. McConechy, J. Ding, J. Senz et al. Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles. *Modern Pathology*, 27(1):128–134, 2014. doi:10.1038/modpathol.2013.107.

[43] M. K. McConechy, L. N. Hoang, M. H. Chui et al. In-depth molecular profiling of the biphasic components of uterine carcinosarcomas. *The Journal of Pathology: Clinical Research*, 1(3):173–185, 2015. doi:10.1002/cjp2.18.

[44] S. Zhao, S. Bellone, S. Lopez et al. Mutational landscape of uterine and ovarian carcinosarcomas implicates histone genes in epithelial–mesenchymal transition. *Proceedings of the National Academy of Sciences*, 113(43):12238–12243, 2016. doi:10.1073/pnas.1614120113.

[45] H.-J. An, S. Logani, C. Isacson et al. Molecular characterization of uterine clear cell carcinoma. *Modern Pathology*, 17(5):530–537, 2004. doi:10.1038/modpathol.3800057.

[46] D. F. DeLair, K. A. Burke, P. Selenica et al. The genetic landscape of endometrial clear cell carcinomas. *The Journal of Pathology*, 243(2):230–241, 2017. doi:10.1002/path.4947.

[47] M. Le Gallo, M. L. Rudd, M. E. Urick et al. Somatic mutation profiles of clear cell endometrial tumors revealed by whole exome and targeted gene sequencing. *Cancer*, 123(17):3261–3268, 2017. doi:10.1002/cncr.30745.

[48] L. N. Hoang, M. K. McConechy, B. Meng et al. Targeted mutation analysis of endometrial clear cell carcinoma. *Histopathology*, 66(5):664–674, 2015. doi:10.1111/his.12581.

[49] S. A. Byron, M. Gartside, M. A. Powell et al. FGFR2 point mutations in 466 endometrioid endometrial tumors: relationship with MSI, KRAS, PIK3CA, CTNNB1 mutations and clinicopathological features. *PloS One*, 7(2):e30801, 2012. doi:10.1371/journal.pone.0030801.

[50] M. Le Gallo, M. L. Rudd, M. E. Urick et al. The FOXA2 transcription factor is frequently somatically mutated in uterine carcinosarcomas and carcinomas. *Cancer*, 124(1):65–73, 2018. doi:10.1002/cncr.30971.

[51] M. Le Gallo, A. J. O'Hara, M. L. Rudd et al. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nature Genetics*, 44(12):1310–1315, 2012. doi:10.1038/ng.2455.

[52] E. Kuhn, R.-C. Wu, B. Guan et al. Identification of Molecular Pathway Aberrations in Uterine Serous Carcinoma by Genome-wide Analyses. *JNCI: Journal of the National Cancer Institute*, 104(19):1503–1513, 2012. doi:10.1093/jnci/djs345.

[53] S. Zhao, M. Choi, J. D. Overton et al. Landscape of somatic single-nucleotide and copy-number mutations in uterine serous carcinoma. *Proceedings of the National Academy of Sciences*, 110(8):2916–2921, 2013. doi:10.1073/pnas.1222577110.

[54] M. E. Urick, M. L. Rudd, A. K. Godwin et al. PIK3R1 (p85α) Is Somatically Mutated at High Frequency in Primary Endometrial Cancer. *Cancer Research*, 71(12):4061–4067, 2011. doi:10.1158/0008-5472.CAN-11-0549.

[55] B. Meng, L. N. Hoang, J. B. McIntyre et al. POLE exonuclease domain mutation predicts long progression-free survival in grade 3 endometrioid carcinoma of the endometrium. *Gynecologic Oncology*, 134(1):15–19, 2014. doi:10.1016/j.ygyno.2014.05.006.

[56] B. G. Wortman, C. L. Creutzberg, H. Putter et al. Ten-year results of the PORTEC-2 trial for high-intermediate risk endometrial carcinoma: improving patient selection for adjuvant therapy. *British Journal of Cancer*, 119(9):1067–1074, 2018. doi:10.1038/s41416-018-0310-8.

[57] M. E. Urick and D. W. Bell. Clinical actionability of molecular targets in endometrial cancer. *Nature Reviews Cancer*, 19(9):510–521, 2019. doi:10.1038/s41568-019-0177-x.

[58] G. L. Mutter, N. M. Monte, D. Neuberg et al. Emergence, Involution, and Progression to Carcinoma of Mutant Clones in Normal Endometrial Tissues. *Cancer Research*, 74(10):2796–2802, 2014. doi:10.1158/0008-5472.CAN-14-0108.

[59] G. L. Mutter, M.-C. Lin, J. T. Fitzgerald et al. Altered PTEN Expression as a Diagnostic Marker for the Earliest Endometrial Precancers. *JNCI: Journal of the National Cancer Institute*, 92(11):924–930, 2000. doi:10.1093/jnci/92.11.924.

[60] T. Bosse, N. T. ter Haar, L. M. Seeber et al. Loss of ARID1A expression and its relationship with PI3K-Akt pathway alterations, TP53 and microsatellite instability in endometrial cancer. *Modern Pathology*, 26(11):1525–1535, 2013. doi:10.1038/modpathol.2013.96.

[61] K. C. Wiegand, S. P. Shah, O. M. Al-Agha et al. ARID1A Mutations in Endometriosis-Associated Ovarian Carcinomas. *New England Journal of Medicine*, 363(16):1532–1543, 2010. doi:10.1056/NEJMoa1008433.

[62] T. H. Kim, J.-Y. Yoo, Z. Wang et al. ARID1A Is Essential for Endometrial Function during Early Pregnancy. *PLOS Genetics*, 11(9):e1005537, 2015. doi:10.1371/journal.pgen.1005537.

[63] X. Wang, S. Khatri, R. Broaddus et al. Deletion of Arid1a in Reproductive Tract Mesenchymal Cells Reduces Fertility in Female Mice1. *Biology of Reproduction*, 94(4):93, 1–14, 2016. doi:10.1095/biolreprod.115.133637.

[64] T.-L. Mao, L. Ardighieri, A. Ayhan et al. Loss of ARID1A Expression Correlates With Stages of Tumor Progression in Uterine Endometrioid Carcinoma. *The American Journal of Surgical Pathology*, 37(9):1342, 2013. doi:10.1097/PAS.0b013e3182889dc3.

[65] M. E. Sherman, M. E. Bur and R. J. Kurman. p53 in endometrial cancer and its putative precursors: Evidence for diverse pathways of tumorigenesis. *Human Pathology*, 26(11):1268–1274, 1995. doi:10.1016/0046-8177(95)90204-X.

[66] S. F. Lax, B. Kendall, H. Tashiro et al. The frequency of p53, k-ras mutations, and microsatellite instability differs in uterine endometrioid and serous carcinoma. *Cancer*, 88(4):814–824, 2000. doi:10.1002/(SICI)1097-0142(20000215)88:4<814::AID-CNCR12>3.0.CO;2-U.

[67] J.-W. Jeong, H. S. Lee, H. L. Franco et al. β-catenin mediates glandular formation and dysregulation of β-catenin induces hyperplasia formation in the murine uterus. *Oncogene*, 28(1):31–40, 2009. doi:10.1038/onc.2008.363.

[68] M. R. Moroney, E. Woodruff, L. Qamar et al. Inhibiting Wnt/beta-catenin in CTNNB1-mutated endometrial cancer. *Molecular Carcinogenesis*, 60(8):511–523, 2021. doi:10.1002/mc.23308.

[69] A. Talhouk, M. K. McConechy, S. Leung et al. A clinically applicable molecular-based classification for endometrial cancers. *British Journal of Cancer*, 113(2):299–310, 2015. doi:10.1038/bjc.2015.190.

[70] E. Stelloo, R. A. Nout, E. M. Osse et al. Improved Risk Assessment by Integrating Molecular and Clinicopathological Factors in Early-stage Endometrial Cancer—Combined Analysis of the PORTEC Cohorts. *Clinical Cancer Research*, 22(16):4215–4224, 2016. doi:10.1158/1078-0432.CCR-15-2878.

[71] A. Talhouk, M. K. McConechy, S. Leung et al. Confirmation of ProMisE: A simple, genomics-based clinical classifier for endometrial cancer. *Cancer*, 123(5):802–813, 2017. doi:10.1002/cncr.30496.

[72] L. Vermij, V. Smit, R. Nout et al. Incorporation of molecular characteristics into endometrial cancer management. *Histopathology*, 76(1):52–63, 2020. doi:10.1111/his.14015.

[73] A. Jamieson, A. Lum, S. Leung et al. O005/#638 ProMisE-2: a one step DNA-based endometrial cancer molecular classifier. *International Journal of Gynecologic Cancer*, 32(Suppl 3), 2022. doi:10.1136/ijgc-2022-igcs.7.

[74] F. Barra, G. Evangelisti, L. Ferro Desideri et al. Investigational PI3K/AKT/mTOR inhibitors in development for endometrial cancer. *Expert Opinion on Investigational Drugs*, 28(2):131–142, 2019. doi:10.1080/13543784.2018.1558202.

[75] B. M. Slomovitz, Y. Jiang, M. S. Yates et al. Phase II Study of Everolimus and Letrozole in Patients With Recurrent Endometrial Carcinoma. *Journal of Clinical Oncology*, 33(8):930–936, 2015. doi:10.1200/JCO.2014.58.3401.

[76] P. T. Soliman, S. N. Westin, D. A. Iglesias et al. Everolimus, Letrozole, and Metformin in Women with Advanced or Recurrent Endometrioid Endometrial Cancer: A Multi-Center, Single Arm, Phase II Study. *Clinical Cancer Research*, 26(3):581–587, 2020. doi:10.1158/1078-0432.CCR-19-0471.

[77] P. A. Ott, Y.-J. Bang, D. Berton-Rigaud et al. Safety and Antitumor Activity of Pembrolizumab in Advanced Programmed Death Ligand 1–Positive Endometrial Cancer: Results From the KEYNOTE-028 Study. *Journal of Clinical Oncology*, 35(22):2535–2541, 2017. doi:10.1200/JCO.2017.72.5952.

[78] A. Oaknin, A. V. Tinker, L. Gilbert et al. Clinical Activity and Safety of the Anti–Programmed Death 1 Monoclonal Antibody Dostarlimab for Patients With Recurrent or Advanced Mismatch Repair–Deficient Endometrial Cancer: A Nonrandomized Phase 1 Clinical Trial. *JAMA Oncology*, 6(11):1766–1772, 2020. doi:10.1001/jamaoncol.2020.4515.

[79] V. Makker, M. H. Taylor, C. Aghajanian et al. Lenvatinib Plus Pembrolizumab in Patients With Advanced Endometrial Cancer. *Journal of Clinical Oncology*, 38(26):2981–2992, 2020. doi:10.1200/JCO.19.02627.

[80] K. Nakamura, E. Aimono, S. Tanishima et al. Olaparib Monotherapy for BRIP1-Mutated High-Grade Serous Endometrial Cancer. *JCO Precision Oncology*, (4):283–290, 2020. doi:10.1200/PO.19.00368.

[81] D. Hanahan. Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, 12(1):31–46, 2022. doi:10.1158/2159-8290.CD-21-1059.

[82] F. Martínez-Jiménez, F. Muiños, I. Sentís et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020. doi:10.1038/s41568-020-0290-x.

[83] N. D. Dees, Q. Zhang, C. Kandoth et al. MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, 22(8):1589–1598, 2012. doi:10.1101/gr.134635.111.

[84] M. T. Chang, S. Asthana, S. P. Gao et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology*, 34(2):155–163, 2016. doi:10.1038/nbt.3391.

[85] C. Tokheim, R. Bhattacharya, N. Niknafs et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Research*, 76(13):3719–3731, 2016. doi:10.1158/0008-5472.CAN-15-3190.

[86] F. Martínez-Jiménez, F. Muiños, E. López-Arribillaga et al. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. *Nature Cancer*, 1(1):122–135, 2020. doi:10.1038/s43018-019-0001-2.

[87] L. Mularoni, R. Sabarinathan, J. Deu-Pons et al. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology*, 17(1):128, 2016. doi:10.1186/s13059-016-0994-0.

[88] F. Dietlein, D. Weghorn, A. Taylor-Weiner et al. Identification of cancer driver genes based on nucleotide context. *Nature Genetics*, 52(2):208–218, 2020. doi: 10.1038/s41588-019-0572-y.

[89] M. S. Lawrence, P. Stojanov, P. Polak et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013. doi:10.1038/nature12213.

[90] D. Weghorn and S. Sunyaev. Bayesian inference of negative and positive selection in human cancers. *Nature Genetics*, 49(12):1785–1788, 2017. doi:10.1038/ng.3987.

[91] C. Arnedo-Pac, L. Mularoni, F. Muiños et al. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics*, 35(22):4788–4790, 2019. doi:10.1093/bioinformatics/btz501.

[92] S. L. Klemm, Z. Shipony and W. J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019. doi:10.1038/s41576-018-0089-8.

[93] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, 2009. doi:10.1038/nature07667.

[94] R. D. Kornberg and J. O. Thomas. Chromatin Structure: Oligomers of the Histones. *Science*, 184(4139):865–868, 1974. doi:10.1126/science.184.4139.865.

[95] M. G. Poirier, M. Bussiek, J. Langowski et al. Spontaneous Access to DNA Target Sites in Folded Chromatin Fibers. *Journal of Molecular Biology*, 379(4):772–786, 2008. doi:10.1016/j.jmb.2008.04.025.

[96] I. Keshet, J. Lieman-Hurwitz and H. Cedar. DNA methylation affects the formation of active chromatin. *Cell*, 44(4):535–543, 1986. doi:10.1016/0092-8674(86)90263-1.

[97] A. Nishiyama and M. Nakanishi. Navigating the DNA methylation landscape of cancer. *Trends in Genetics*, 37(11):1012–1027, 2021. doi:10.1016/j.tig.2021.05.002.

[98] A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, 2011. doi:10.1038/cr.2011.22.

[99] M. Tan, H. Luo, S. Lee et al. Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification. *Cell*, 146(6):1016–1028, 2011. doi:10.1016/j.cell.2011.08.008.

[100] T. Jenuwein and C. D. Allis. Translating the Histone Code. *Science*, 293(5532):1074–1080, 2001. doi:10.1126/science.1063127.

[101] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, 2000. doi:10.1038/47412.

[102] T. Kouzarides. Chromatin Modifications and Their Function. *Cell*, 128(4):693–705, 2007. doi:10.1016/j.cell.2007.02.005.

[103] A. Kundaje, W. Meuleman, J. Ernst et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. doi:10.1038/nature14248.

[104] S. Y. Roth, J. M. Denu and C. D. Allis. Histone acetyltransferases. *Annual Review of Biochemistry*, 70:81–120, 2001. doi:10.1146/annurev.biochem.70.1.81.

[105] K. Hyun, J. Jeon, K. Park et al. Writing, erasing and reading histone lysine methylations. *Experimental & Molecular Medicine*, 49(4):e324–e324, 2017. doi:10.1038/emm.2017.11.

[106] C. Dhalluin, J. E. Carlson, L. Zeng et al. Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, 399(6735):491–496, 1999. doi:10.1038/20974.

[107] X.-J. Yang and E. Seto. The Rpd3/Hda1 family of lysine deacetylases: from bacteria and yeast to mice and men. *Nature Reviews Molecular Cell Biology*, 9(3):206–218, 2008. doi:10.1038/nrm2346.

[108] Y. Shi, F. Lan, C. Matson et al. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell*, 119(7):941–953, 2004. doi:10.1016/j.cell.2004.12.012.

[109] A. Shilatifard. Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. *Current Opinion in Cell Biology*, 20(3):341–348, 2008. doi:10.1016/j.ceb.2008.03.019.

[110] D. Hu, A. S. Garruss, X. Gao et al. The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nature Structural & Molecular Biology*, 20(9):1093–1097, 2013. doi:10.1038/nsmb.2653.

[111] B. P. Mishra, K. M. Zaffuto, E. L. Artinger et al. The Histone Methyltransferase Activity of MLL1 Is Dispensable for Hematopoiesis and Leukemogenesis. *Cell Reports*, 7(4):1239–1247, 2014. doi:10.1016/j.celrep.2014.04.015.

[112] M. J. A. Weerts, K. Lanko, F. J. Guzmán-Vega et al. Delineating the molecular and phenotypic spectrum of the SETD1B-related syndrome. *Genetics in Medicine*, 23(11):2122–2137, 2021. doi:10.1038/s41436-021-01246-2.

[113] T. Miller, N. J. Krogan, J. Dover et al. COMPASS: A complex of proteins associated with a trithorax-related SET domain protein. *Proceedings of the National Academy of Sciences*, 98(23):12902–12907, 2001. doi:10.1073/pnas.231473398.

[114] Y. Dou, T. A. Milne, A. J. Tackett et al. Physical Association and Coordinate Function of the H3 K4 Methyltransferase MLL1 and the H4 K16 Acetyltransferase MOF. *Cell*, 121(6):873–885, 2005. doi:10.1016/j.cell.2005.04.031.

[115] A. Astolfi, M. Fiore, F. Melchionda et al. BCOR involvement in cancer. *Epigenomics*, 11(7):835–855, 2019. doi:10.2217/epi-2018-0195.

[116] B. Schuettengruber, H.-M. Bourbon, L. D. Croce et al. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell*, 171(1):34–57, 2017. doi:10.1016/j.cell.2017.08.002.

[117] M. Sauvageau and G. Sauvageau. Polycomb Group Proteins: Multi-Faceted Regulators of Somatic Stem Cells and Cancer. *Cell Stem Cell*, 7(3):299–313, 2010. doi:10.1016/j.stem.2010.08.002.

[118] H. Wang, L. Wang, H. Erdjument-Bromage et al. Role of histone H2A ubiquitination in Polycomb silencing. *Nature*, 431(7010):873–878, 2004. doi:10.1038/nature02985.

[119] E. C. Chittock, S. Latwiel, T. C. Miller et al. Molecular architecture of polycomb repressive complexes. *Biochemical Society Transactions*, 45(1):193–205, 2017. doi:10.1042/BST20160173.

[120] Z. Wang, M. D. Gearhart, Y.-W. Lee et al. A Non-canonical BCOR-PRC1.1 Complex Represses Differentiation Programs in Human ESCs. *Cell Stem Cell*, 22(2):235–251.e9, 2018. doi:10.1016/j.stem.2017.12.002.

[121] C. R. Clapier and B. R. Cairns. The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry*, 78(1):273–304, 2009. doi:10.1146/annurev.biochem.77.062706.153223.

[122] N. Mashtalir, H. T. Dao, A. Sankar et al. Chromatin landscape signals differentially dictate the activities of mSWI/SNF family complexes. *Science*, 373(6552):306–315, 2021. doi:10.1126/science.abf8705.

[123] A. Saha, J. Wittmeyer and B. R. Cairns. Chromatin remodeling through directional DNA translocation from an internal nucleosomal site. *Nature Structural & Molecular Biology*, 12(9):747–755, 2005. doi:10.1038/nsmb973.

[124] M. L. Phelan, S. Sif, G. J. Narlikar et al. Reconstitution of a Core Chromatin Remodeling Complex from SWI/SNF Subunits. *Molecular Cell*, 3(2):247–253, 1999. doi:10.1016/S1097-2765(00)80315-9.

[125] N. Mashtalir, A. R. D'Avino, B. C. Michel et al. Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell*, 175(5):1272–1288.e20, 2018. doi:10.1016/j.cell.2018.09.032.

[126] B. C. Michel, A. R. D'Avino, S. H. Cassel et al. A non-canonical SWI/SNF complex is a synthetic lethal target in cancers driven by BAF complex perturbation. *Nature Cell Biology*, 20(12):1410–1420, 2018. doi:10.1038/s41556-018-0221-1.

[127] C. Kadoch and G. R. Crabtree. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Science Advances*, 1(5):e1500447, 2015. doi:10.1126/sciadv.1500447.

[128] B. Lemon, C. Inouye, D. S. King et al. Selectivity of chromatin-remodelling cofactors for ligand-activated transcription. *Nature*, 414(6866):924–928, 2001. doi:10.1038/414924a.

[129] M. P. Torchy, A. Hamiche and B. P. Klaholz. Structure and function insights into the NuRD chromatin remodeling complex. *Cellular and Molecular Life Sciences*, 72(13):2491–2507, 2015. doi:10.1007/s00018-015-1880-8.

[130] J. K. Tong, C. A. Hassig, G. R. Schnitzler et al. Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. *Nature*, 395(6705):917–921, 1998. doi:10.1038/27699.

[131] S. A. Denslow and P. A. Wade. The human Mi-2/NuRD complex and gene regulation. *Oncogene*, 26(37):5433–5438, 2007. doi:10.1038/sj.onc.1210611.

[132] L. Snijders Blok, J. Rousseau, J. Twist et al. CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. *Nature Communications*, 9(1):4619, 2018. doi:10.1038/s41467-018-06014-6.

[133] A. Kleppe, F. Albregtsen, L. Vlatkovic et al. Chromatin organisation and cancer prognosis: a pan-cancer study. *The Lancet Oncology*, 19(3):356–369, 2018. doi:10.1016/S1470-2045(17)30899-9.

[134] S. Zhao, C. D. Allis and G. G. Wang. The language of chromatin modification in human cancers. *Nature Reviews Cancer*, 21(7):413–430, 2021. doi:10.1038/s41568-021-00357-x.

[135] S. B. Baylin and P. A. Jones. A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734, 2011. doi:10.1038/nrc3130.

[136] J. Jiricny. The multifaceted mismatch-repair system. *Nature Reviews Molecular Cell Biology*, 7(5):335–346, 2006. doi:10.1038/nrm1907.

[137] M. Esteller, R. Levine, S. B. Baylin et al. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. *Oncogene*, 17(18):2413–2417, 1998. doi:10.1038/sj.onc.1202178.

[138] S. B. Simpkins, T. Bocker, E. M. Swisher et al. MLH1 Promoter Methylation and Gene Silencing is the Primary Cause of Microsatellite Instability in Sporadic Endometrial Cancers. *Human Molecular Genetics*, 8(4):661–666, 1999. doi:10.1093/hmg/8.4.661.

[139] K. Wang, J. Kan, S. T. Yuen et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nature Genetics*, 43(12):1219–1223, 2011. doi:10.1038/ng.982.

[140] C. Kadoch, D. C. Hargreaves, C. Hodges et al. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nature Genetics*, 45(6):592–601, 2013. doi:10.1038/ng.2628.

[141] P. Mittal and C. W. M. Roberts. The SWI/SNF complex in cancer — biology, biomarkers and therapy. *Nature Reviews Clinical Oncology*, 17(7):435–448, 2020. doi:10.1038/s41571-020-0357-3.

[142] K. C. Helming, X. Wang and C. W. M. Roberts. Vulnerabilities of Mutant SWI/SNF Complexes in Cancer. *Cancer Cell*, 26(3):309–317, 2014. doi:10.1016/j.ccr.2014.07.018.

[143] J. Wang, F. Zhong, J. Li et al. The epigenetic factor CHD4 contributes to metastasis by regulating the EZH2/β-catenin axis and acts as a therapeutic target in ovarian cancer. *Journal of Translational Medicine*, 21(1):38, 2023. doi:10.1186/s12967-022-03854-1.

[144] Y. Li, Q. Liu, D. J. McGrail et al. CHD4 mutations promote endometrial cancer stemness by activating TGF-beta signaling. *American Journal of Cancer Research*, 8(5):903–914, 2018.

[145] A. Piunti and A. Shilatifard. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nature Reviews Molecular Cell Biology*, 22(5):326–345, 2021. doi:10.1038/s41580-021-00341-1.

[146] S. S. Dhar, D. Zhao, T. Lin et al. MLL4 Is Required to Maintain Broad H3K4me3 Peaks and Super-Enhancers at Tumor Suppressor Genes. *Molecular Cell*, 70(5):825–841.e6, 2018. doi:10.1016/j.molcel.2018.04.028.

[147] E. Tiacci, V. Grossmann, M. P. Martelli et al. The corepressors BCOR and BCORL1: two novel players in acute myeloid leukemia. *Haematologica*, 97(1):3–5, 2012. doi:10.3324/haematol.2011.057901.

[148] X. Yin, X. Teng, T. Ma et al. RUNX2 recruits the NuRD(MTA1)/CRL4B complex to promote breast cancer progression and bone metastasis. *Cell Death & Differentiation*, 29(11):2203–2217, 2022. doi:10.1038/s41418-022-01010-2.

[149] Y. Zhang, J. L. Donaher, S. Das et al. Genome-wide CRISPR screen identifies PRC2 and KMT2D-COMPASS as regulators of distinct EMT trajectories that contribute differentially to metastasis. *Nature Cell Biology*, 24(4):554–564, 2022. doi:10.1038/s41556-022-00877-0.

[150] P. C. Nowell. The Clonal Evolution of Tumor Cell Populations. *Science*, 194(4260):23–28, 1976. doi:10.1126/science.959840.

[151] I. J. Fidler and I. R. Hart. Biological Diversity in Metastatic Neoplasms: Origins and Implications. *Science*, 217(4564):998–1003, 1982. doi:10.1126/science.7112116.

[152] G. H. Heppner. Tumor Heterogeneity. *Cancer Research*, 44(6):2259–2265, 1984.

[153] M. Gerlinger, A. J. Rowan, S. Horswell et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892, 2012. doi:10.1056/NEJMoa1113205.

[154] C. E. Meacham and S. J. Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013. doi:10.1038/nature12624.

[155] G. H. Heppner, D. L. Dexter, T. DeNucci et al. Heterogeneity in drug sensitivity among tumor cell subpopulations of a single mammary tumor. *Cancer Research*, 38(11 Pt 1):3758–3763, 1978.

[156] Cancer Research UK. Cancer Research UK - Uterine cancer incidence statistics. 2020. URL https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/incidence.

[157] K. A. Hoadley, C. Yau, T. Hinoue et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6, 2018. doi:10.1016/j.cell.2018.03.022.

[158] K. Ellrott, M. H. Bailey, G. Saksena et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems*, 6(3):271–281.e7, 2018. doi:10.1016/j.cels.2018.03.002.

[159] A. Colaprico, T. C. Silva, C. Olsen et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71–e71, 2015. doi:10.1093/nar/gkv1507.

[160] Pan-Cancer Analysis of Whole Genomes. 2020. URL https://www.nature.com/collections/afdejfafdb/.

[161] E. Rheinbay, M. M. Nielsen, F. Abascal et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578(7793):102–111, 2020. doi:10.1038/s41586-020-1965-x.

[162] P. García-Sanz, J. C. Triviño, A. Mota et al. Chromatin remodelling and DNA repair genes are frequently mutated in endometrioid endometrial carcinoma. *International Journal of Cancer*, 140(7):1551–1563, 2017. doi:10.1002/ijc.30573.

[163] A. León-Castillo, H. Britton, M. K. McConechy et al. Interpretation of somatic POLE mutations in endometrial carcinoma. *The Journal of Pathology*, 250(3):323–335, 2020. doi:10.1002/path.5372.

[164] S. J. Salipante, S. M. Scroggins, H. L. Hampel et al. Microsatellite Instability Detection by Next Generation Sequencing. *Clinical Chemistry*, 60(9):1192–1199, 2014. doi:10.1373/clinchem.2014.223677.

[165] L. B. Alexandrov, J. Kim, N. J. Haradhvala et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020. doi:10.1038/s41586-020-1943-3.

[166] D. Bee, A. Jones, Z. Sondka et al. COSMIC | Mutational Signatures. 2022. URL https://cancer.sanger.ac.uk/signatures/.

[167] C. D. Steele, A. Abbasi, S. M. A. Islam et al. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, 2022. doi:10.1038/s41586-022-04738-6.

[168] Z. Sztupinszki, M. Diossy, M. Krzystanek et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *npj Breast Cancer*, 4(1):1–4, 2018. doi:10.1038/s41523-018-0066-6.

[169] F. Menghi, F. P. Barthel, V. Yadav et al. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell*, 34(2):197–210.e5, 2018. doi:10.1016/j.ccell.2018.06.008.

[170] B. Otlu, M. Díaz-Gay, I. Vermes et al. Topography of mutational signatures in human cancer. 2022. doi:10.1101/2022.05.29.493921. URL https://www.biorxiv.org/content/10.1101/2022.05.29.493921v1.

[171] D. Gallo, J. T. F. Young, J. Fourtounis et al. CCNE1 amplification is synthetic lethal with PKMYT1 kinase inhibition. *Nature*, 604(7907):749–756, 2022. doi:10.1038/s41586-022-04638-9.

[172] C.-H. Yeh, M. Bellon and C. Nicot. FBXW7: a critical tumor suppressor of human cancers. *Molecular Cancer*, 17(1):115, 2018. doi:10.1186/s12943-018-0857-2.

[173] C. H. Mermel, S. E. Schumacher, B. Hill et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011. doi:10.1186/gb-2011-12-4-r41.

[174] H. Xu, E. George, Y. Kinose et al. CCNE1 copy number is a biomarker for response to combination WEE1-ATR inhibition in ovarian and endometrial cancer models. *Cell Reports Medicine*, 2(9), 2021. doi:10.1016/j.xcrm.2021.100394.

[175] M. L. Davenport, M. R. Davis, B. N. Davenport et al. Suppression of SIN3A by miR-183 Promotes Breast Cancer Metastasis. *Molecular Cancer Research*, 20(6):883–894, 2022. doi:10.1158/1541-7786.MCR-21-0508.

[176] K. Nakamura, G. Saredi, J. R. Becker et al. H4K20me0 recognition by BRCA1–BARD1 directs homologous recombination to sister chromatids. *Nature Cell Biology*, 21(3):311–318, 2019. doi:10.1038/s41556-019-0282-9.

[177] J. E. Phillips and V. G. Corces. CTCF: Master Weaver of the Genome. *Cell*, 137(7):1194–1211, 2009. doi:10.1016/j.cell.2009.06.001.

[178] J. Zlatanova and P. Caiafa. CTCF and its protein partners: divide and rule? *Journal of Cell Science*, 122(9):1275–1284, 2009. doi:10.1242/jcs.039990.

[179] S. S. P. Rao, M. H. Huntley, N. C. Durand et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680, 2014. doi:10.1016/j.cell.2014.11.021.

[180] G. Fudenberg, M. Imakaev, C. Lu et al. Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, 15(9):2038–2049, 2016. doi:10.1016/j.celrep.2016.04.085.

[181] I. Zighelboim, D. G. Mutch, A. Knapp et al. High Frequency Strand Slippage Mutations in CTCF in MSI-Positive Endometrial Cancers. *Human Mutation*, 35(1):63–65, 2014. doi:10.1002/humu.22463.

[182] S. Y. Hwang, M. A. Kang, C. J. Baik et al. CTCF cooperates with CtIP to drive homologous recombination repair of double-strand breaks. *Nucleic Acids Research*, 47(17):9160–9179, 2019. doi:10.1093/nar/gkz639.

[183] S. J. Aitken, X. Ibarra-Soria, E. Kentepozidou et al. CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biology*, 19(1):106, 2018. doi:10.1186/s13059-018-1484-3.

[184] C. Fang, Z. Wang, C. Han et al. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *Genome Biology*, 21(1):247, 2020. doi:10.1186/s13059-020-02152-7.

[185] Z. Sondka, S. Bamford, C. G. Cole et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018. doi:10.1038/s41568-018-0060-1.

[186] G. K. Griffin, J. Wu, A. Iracheta-Vellve et al. Epigenetic silencing by SETDB1 suppresses tumour intrinsic immunogenicity. *Nature*, 595(7866):309–314, 2021. doi:10.1038/s41586-021-03520-4.

[187] V. Garcia-Carpizo, S. Ruiz-Llorente, J. Sarmentero et al. CREBBP/EP300 Bromodomain Inhibition Affects the Proliferation of AR-Positive Breast Cancer Cell Lines. *Molecular Cancer Research*, 17(3):720–730, 2019. doi:10.1158/1541-7786.MCR-18-0719.

[188] J. Welti, A. Sharp, N. Brooks et al. Targeting the p300/CBP Axis in Lethal Prostate Cancer. *Cancer Discovery*, 11(5):1118–1137, 2021. doi:10.1158/2159-8290.CD-20-0751.

[189] T. W. R. Kelso, D. K. Porter, M. L. Amaral et al. Chromatin accessibility underlies synthetic lethality of SWI/SNF subunits in ARID1A-mutant cancers. *eLife*, 6:e30506, 2017. doi:10.7554/eLife.30506.

[190] Z. Wang, K. Chen, Y. Jia et al. Dual ARID1A/ARID1B loss leads to rapid carcinogenesis and disruptive redistribution of BAF complexes. *Nature Cancer*, 1(9):909–922, 2020. doi:10.1038/s43018-020-00109-0.

[191] C. J. Creighton, M. Morgan, P. H. Gunaratne et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013. doi:10.1038/nature12222.

[192] C. R. Dufour, C. Scholtes, M. Yan et al. The mTOR chromatin-bound interactome in prostate cancer. *Cell Reports*, 38(12):110534, 2022. doi:10.1016/j.celrep.2022.110534.

[193] S. Nik-Zainal, P. Van Loo, D. C. Wedge et al. The Life History of 21 Breast Cancers. *Cell*, 149(5):994–1007, 2012. doi:10.1016/j.cell.2012.04.023.

[194] M. Brown, A. Leon, K. Kedzierska et al. Functional analysis reveals driver cooperativity and novel mechanisms in endometrial carcinogenesis. *EMBO Molecular Medicine*, 15(10):e17094, 2023. doi:10.15252/emmm.202217094.

[195] A. A. Ahmed, D. Etemadmoghadam, J. Temple et al. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *The Journal of Pathology*, 221(1):49–56, 2010. doi:10.1002/path.2696.

[196] D. Bell, A. Berchuck, M. Birrer et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011. doi:10.1038/nature10166.

[197] Y. Lee, A. Miron, R. Drapkin et al. A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *The Journal of Pathology*, 211(1):26–35, 2007. doi:10.1002/path.2091.

[198] T. M. Bergsten, J. E. Burdette and M. Dean. Fallopian tube initiation of high grade serous ovarian cancer and ovarian metastasis: Mechanisms and therapeutic implications. *Cancer Letters*, 476:152–160, 2020. doi:10.1016/j.canlet.2020.02.017.

[199] R. J. Kurman and I.-M. Shih. The Origin and Pathogenesis of Epithelial Ovarian Cancer: A Proposed Unifying Theory. *The American Journal of Surgical Pathology*, 34(3):433, 2010. doi:10.1097/PAS.0b013e3181cf3d79.

[200] S. Jones, T.-L. Wang, I.-M. Shih et al. Frequent Mutations of Chromatin Remodeling Gene ARID1A in Ovarian Clear Cell Carcinoma. *Science*, 330(6001):228–231, 2010. doi:10.1126/science.1196333.

[201] S. Yamamoto, H. Tsuda, M. Takano et al. PIK3CA mutations and loss of ARID1A protein expression are early events in the development of cystic ovarian clear cell adenocarcinoma. *Virchows Archiv*, 460(1):77–87, 2012. doi:10.1007/s00428-011-1169-8.

[202] D. S. Tan, M. Iravani, W. G. McCluggage et al. Genomic Analysis Reveals the Molecular Heterogeneity of Ovarian Clear Cell Carcinomas. *Clinical Cancer Research*, 17(6):1521–1534, 2011. doi:10.1158/1078-0432.CCR-10-1688.

[203] C. Wang, Z. Liang, X. Liu et al. The Association between Endometriosis, Tubal Ligation, Hysterectomy and Epithelial Ovarian Cancer: Meta-Analyses. *International Journal of Environmental Research and Public Health*, 13(11):1138, 2016. doi:10.3390/ijerph13111138.

[204] M. Kvaskoff, F. Mu, K. L. Terry et al. Endometriosis: a high-risk population for major chronic diseases? *Human Reproduction Update*, 21(4):500–516, 2015. doi:10.1093/humupd/dmv013.

[205] B. Guan, T.-L. Wang and I.-M. Shih. ARID1A, a Factor That Promotes Formation of SWI/SNF-Mediated Chromatin Remodeling, Is a Tumor Suppressor in Gynecologic Cancers. *Cancer Research*, 71(21):6718–6727, 2011. doi:10.1158/0008-5472.CAN-11-1562.

[206] H. M. J. Werner, A. Berg, E. Wik et al. ARID1A loss is prevalent in endometrial hyperplasia with atypia and low-grade endometrioid carcinomas. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 26(3):428–434, 2013. doi:10.1038/modpathol.2012.174.

[207] A. Katagiri, K. Nakayama, M. T. Rahman et al. Loss of ARID1A expression is related to shorter progression-free survival and chemoresistance in ovarian clear cell carcinoma. *Modern Pathology*, 25(2):282–288, 2012. doi:10.1038/modpathol.2011.161.

[208] H. Ueno-Yokohata, H. Okita, K. Nakasato et al. Consistent in-frame internal tandem duplications of BCOR characterize clear cell sarcoma of the kidney. *Nature Genetics*, 47(8):861–863, 2015. doi:10.1038/ng.3338.

[209] A. Astolfi, F. Melchionda, D. Perotti et al. Whole transcriptome sequencing identifies BCOR internal tandem duplication as a common feature of clear cell sarcoma of the kidney. *Oncotarget*, 6(38):40934–40939, 2015. doi:10.18632/oncotarget.5882.

[210] Y.-C. Kao, A. A. Owosho, Y.-S. Sung et al. BCOR-CCNB3 Fusion Positive Sarcomas: A Clinicopathologic and Molecular Analysis of 36 Cases With Comparison to Morphologic Spectrum and Clinical Behavior of Other Round Cell Sarcomas. *The American Journal of Surgical Pathology*, 42(5):604, 2018. doi:10.1097/PAS.0000000000000965.

[211] N. Lewis, R. A. Soslow, D. F. Delair et al. ZC3H7B-BCOR high-grade endometrial stromal sarcomas: a report of 17 cases of a newly defined entity. *Modern Pathology*, 31(4):674–684, 2018. doi:10.1038/modpathol.2017.162.

[212] S. Kommoss, M. K. McConechy, F. Kommoss et al. Final validation of the ProMisE molecular classifier for endometrial carcinoma in a large population-based case series. *Annals of Oncology*, 29(5):1180–1188, 2018. doi:10.1093/annonc/mdy058.

[213] K. C. Kurnit, G. N. Kim, B. M. Fellman et al. CTNNB1 (beta-catenin) mutation identifies low grade, early stage endometrial cancer patients at increased risk of recurrence. *Modern Pathology*, 30(7):1032–1041, 2017. doi:10.1038/modpathol.2017.15.

[214] D. C. Costigan, F. Dong, M. R. Nucci et al. Clinicopathologic and Immunohistochemical Correlates of CTNNB1 Mutated Endometrial Endometrioid Carcinoma. *International Journal of Gynecological Pathology*, 39(2):119, 2020. doi:10.1097/PGP.0000000000000583.

[215] T. Bosse, R. A. Nout, E. Stelloo et al. L1 cell adhesion molecule is a strong predictor for distant recurrence and overall survival in early stage endometrial cancer: Pooled PORTEC trial results. *European Journal of Cancer*, 50(15):2602–2610, 2014. doi:10.1016/j.ejca.2014.07.014.

[216] F. K. Kommoss, A. N. Karnezis, F. Kommoss et al. L1CAM further stratifies endometrial carcinoma patients with no specific molecular risk profile. *British Journal of Cancer*, 119(4):480–486, 2018. doi:10.1038/s41416-018-0187-6.

[217] L. J. M. v. d. Putten, N. C. M. Visser, K. v. d. Vijver et al. Added Value of Estrogen Receptor, Progesterone Receptor, and L1 Cell Adhesion Molecule Expression to Histology-Based Endometrial Carcinoma Recurrence Prediction Models: An ENITEC Collaboration Study. *International Journal of Gynecologic Cancer*, 28(3), 2018. doi:10.1097/IGC.0000000000001187.

[218] V. Jongen, J. Briët, R. d. Jong et al. Expression of estrogen receptor-alpha and -beta and progesterone receptor-A and -B in a large cohort of patients with endometrioid endometrial cancer. *Gynecologic Oncology*, 112(3):537–542, 2009. doi:10.1016/j.ygyno.2008.10.032.

[219] J. Trovik, E. Wik, H. M. J. Werner et al. Hormone receptor loss in endometrial carcinoma curettage predicts lymph node metastasis and poor outcome in prospective multicentre trial. *European Journal of Cancer*, 49(16):3431–3441, 2013. doi:10.1016/j.ejca.2013.06.016.

[220] F. J. Backes, C. J. Walker, P. J. Goodfellow et al. Estrogen receptor-alpha as a predictive biomarker in endometrioid endometrial cancer. *Gynecologic Oncology*, 141(2):312–317, 2016. doi:10.1016/j.ygyno.2016.03.006.

[221] J. Depreeuw, E. Stelloo, E. M. Osse et al. Amplification of 1q32.1 Refines the Molecular Classification of Endometrial Carcinoma. *Clinical Cancer Research*, 23(23):7232–7241, 2017. doi:10.1158/1078-0432.CCR-17-0566.

[222] A. Momeni-Boroujeni, B. Nguyen, C. M. Vanderbilt et al. Genomic landscape of endometrial carcinomas of no specific molecular profile. *Modern Pathology*, 35(9):1269–1278, 2022. doi:10.1038/s41379-022-01066-y.

[223] L. Vermij, J. J. Jobsen, A. León-Castillo et al. Prognostic refinement of NSMP high-risk endometrial cancers using oestrogen receptor immunohistochemistry. *British Journal of Cancer*, 128(7):1360–1368, 2023. doi:10.1038/s41416-023-02141-0.

[224] E. Stelloo, T. Bosse, R. A. Nout et al. Refining prognosis and identifying targetable pathways for high-risk endometrial cancer; a TransPORTEC initiative. *Modern Pathology*, 28(6):836–844, 2015. doi:10.1038/modpathol.2015.43.

[225] T. Bosse, R. A. Nout, J. N. McAlpine et al. Molecular Classification of Grade 3 Endometrioid Endometrial Cancers Identifies Distinct Prognostic Subgroups. *The American Journal of Surgical Pathology*, 42(5):561, 2018. doi:10.1097/PAS.0000000000001020.

[226] C. Raczy, R. Petrovski, C. T. Saunders et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, 29(16):2041–2043, 2013. doi:10.1093/bioinformatics/btt314.

[227] S. Kim, K. Scheffler, A. L. Halpern et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15(8):591–594, 2018. doi:10.1038/s41592-018-0051-x.

[228] C. Kandoth. vcf2maf. 2020. URL https://github.com/mskcc/vcf2maf.

[229] A. J. Cornish, A. J. Gruber, B. Kinnersley et al. Whole genome sequencing of 2,023 colorectal cancers reveals mutational landscapes, new driver genes and immune interactions. 2022. doi:10.1101/2022.11.16.515599. URL https://www.biorxiv.org/content/10.1101/2022.11.16.515599v1.

[230] T. Rausch, T. Zichner, A. Schlattl et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012. doi:10.1093/bioinformatics/bts378.

[231] R. M. Layer, C. Chiang, A. R. Quinlan et al. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, 2014. doi:10.1186/gb-2014-15-6-r84.

[232] G. R. Bignell, C. D. Greenman, H. Davies et al. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–898, 2010. doi:10.1038/nature08768.

[233] S. M. A. Islam, M. Díaz-Gay, Y. Wu et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics*, 2(11):100179, 2022. doi:10.1016/j.xgen.2022.100179.

[234] C. H. Li, S. D. Prokopec, R. X. Sun et al. Sex differences in oncogenic mutational processes. *Nature Communications*, 11(1):4330, 2020. doi:10.1038/s41467-020-17359-2.

[235] L. D'Agostino McGowan. Power calculations for survival analyses. ???? URL https://lucymcgowan.github.io/survivalpwr/index.html.

[236] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

[237] T. M. Therneau, T. L. o. S.->. p. a. R. m. until 2009), A. Elizabeth et al. survival: Survival Analysis. 2023. URL https://CRAN.R-project.org/package=survival.

[238] A. Kassambara, M. Kosinski, P. Biecek et al. survminer: Drawing Survival Curves using 'ggplot2'. 2021. URL https://CRAN.R-project.org/package=survminer.

[239] S. Dentro. DPClust pre-processing. 2022. URL https://github.com/Wedge-lab/dpclust3p. Original-date: 2015-01-04T13:54:16Z.

[240] M. Gerstung, C. Jolly, I. Leshchiner et al. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, 2020. doi:10.1038/s41586-019-1907-7.

[241] A. C. Berger, A. Korkut, R. S. Kanchi et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, 33(4):690–705.e9, 2018. doi:10.1016/j.ccell.2018.03.014.

[242] C. Toumpeki, A. Liberis, I. Tsirkas et al. The Role of ARID1A in Endometrial Cancer and the Molecular Pathways Associated With Pathogenesis and Cancer Progression. *In Vivo*, 33(3):659–667, 2019. doi:10.21873/invivo.11524.

[243] R. C. Rao and Y. Dou. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nature Reviews Cancer*, 15(6):334–346, 2015. doi:10.1038/nrc3929.

[244] Y. Park, J.-G. Jung, Z.-C. Yu et al. A novel human endometrial epithelial cell line for modeling gynecological diseases and for drug screening. *Laboratory Investigation*, 101(11):1505–1512, 2021. doi:10.1038/s41374-021-00624-3.

[245] Y. S. Rahmanto, J.-G. Jung, R.-C. Wu et al. Inactivating ARID1A Tumor Suppressor Enhances TERT Transcription and Maintains Telomere Length in Cancer Cells *. *Journal of Biological Chemistry*, 291(18):9690–9699, 2016. doi:10.1074/jbc.M115.707612.

[246] K. El-Sahwi, S. Bellone, E. Cocco et al. In vitro activity of pertuzumab in combination with trastuzumab in uterine serous papillary adenocarcinoma. *British Journal of Cancer*, 102(1):134–143, 2010. doi:10.1038/sj.bjc.6605448.

[247] B. Guan, T.-L. Mao, P. K. Panuganti et al. Mutation and Loss of Expression of ARID1A in Uterine Low-grade Endometrioid Carcinoma. *The American Journal of Surgical Pathology*, 35(5):625, 2011. doi:10.1097/PAS.0b013e318212782a.

[248] M. I. Love, W. Huber and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi:10.1186/s13059-014-0550-8.

[249] Y. Kojima, K. Sudo, H. Yoshida et al. Changes in HER3 expression profiles between primary and recurrent gynecological cancers. *Cancer Cell International*, 23(1):18, 2023. doi:10.1186/s12935-022-02844-z.

[250] V. Syed. TGF-beta Signaling in Cancer. *Journal of Cellular Biochemistry*, 117(6):1279–1287, 2016. doi:10.1002/jcb.25496.

[251] B. Y. Gueuvoghlanian-Silva, P. Bellelis, D. F. Barbeiro et al. Treg and NK cells related cytokines are associated with deep rectosigmoid endometriosis and clinical symptoms related to the disease. *Journal of Reproductive Immunology*, 126:32–38, 2018. doi:10.1016/j.jri.2018.02.003.

[252] K.-K. Chang, L.-B. Liu, L.-P. Jin et al. IL-27 triggers IL-10 production in Th17 cells via a c-Maf/RORgamma/Blimp-1 signal to promote the progression of endometriosis. *Cell Death & Disease*, 8(3):e2666–e2666, 2017. doi:10.1038/cddis.2017.95.

[253] H.-J. Choi, M.-J. Park, B.-S. Kim et al. Transforming growth factor β1 enhances adhesion of endometrial cells to mesothelium by regulating integrin expression. *BMB Reports*, 50(8):429–434, 2017. doi:10.5483/BMBRep.2017.50.8.097.

[254] K. A. Walters, U. Simanainen and D. A. Gibson. Androgen action in female reproductive physiology. *Current Opinion in Endocrinology, Diabetes and Obesity*, 23(3):291, 2016. doi:10.1097/MED.0000000000000246.

[255] W. Ye, T. Xie, Y. Song et al. The role of androgen and its related signals in PCOS. *Journal of Cellular and Molecular Medicine*, 25(4):1825–1837, 2021. doi:10.1111/jcmm.16205.

[256] D. A. Gibson, I. Simitsidellis, F. Collins et al. Androgens, oestrogens and endometrium: a fine balance between perfection and pathology. *Journal of Endocrinology*, 246(3):R75–R93, 2020. doi:10.1530/JOE-20-0106.

[257] G. Xu, S. Chhangawala, E. Cocco et al. ARID1A determines luminal identity and therapeutic response in estrogen-receptor-positive breast cancer. *Nature Genetics*, 52(2):198–207, 2020. doi:10.1038/s41588-019-0554-0.

[258] M. Laplante and D. M. Sabatini. mTOR Signaling in Growth Control and Disease. *Cell*, 149(2):274–293, 2012. doi:10.1016/j.cell.2012.03.017.

[259] R. A. Saxton and D. M. Sabatini. mTOR Signaling in Growth, Metabolism, and Disease. *Cell*, 168(6):960–976, 2017. doi:10.1016/j.cell.2017.02.004.

[260] A. N. Schep, B. Wu, J. D. Buenrostro et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, 14(10):975–978, 2017. doi:10.1038/nmeth.4401.

[261] A. Grigoriadis, K. Schellander, Z. Wang et al. Osteoblasts are target cells for transformation in c-fos transgenic mice. *Journal of Cell Biology*, 122(3):685–701, 1993. doi:10.1083/jcb.122.3.685.

[262] M. R. Young, J.-J. Li, M. Rincón et al. Transgenic mice demonstrate AP-1 (activator protein-1) transactivation is required for tumor promotion. *Proceedings of the National Academy of Sciences*, 96(17):9827–9832, 1999. doi:10.1073/pnas.96.17.9827.

[263] E. Passegué, W. Jochum, M. Schorpp-Kistner et al. Chronic Myeloid Leukemia with Increased Granulocyte Progenitors in Mice Lacking JunB Expression in the Myeloid Lineage. *Cell*, 104(1):21–32, 2001. doi:10.1016/S0092-8674(01)00188-X.

[264] C. M. Pfarr, F. Mechta, G. Spyrou et al. Mouse JunD negatively regulates fibroblast growth and antagonizes transformation by ras. *Cell*, 76(4):747–760, 1994. doi: 10.1016/0092-8674(94)90513-4.

[265] Y. Sun, D. Ye, Y. Li et al. CUX2 functions as an oncogene in papillary thyroid cancer. *OncoTargets and Therapy*, 12:217–224, 2018. doi:10.2147/OTT.S185710.

[266] G. Yao, S. Le, S. Min et al. CUX2 prevents the malignant progression of gliomas by enhancing ADCY1 transcription. *Experimental Brain Research*, 240(12):3153–3165, 2022. doi:10.1007/s00221-022-06481-w.

[267] M. R. M. Jotte and M. E. McNerney. The significance of CUX1 and chromosome 7 in myeloid malignancies. *Current Opinion in Hematology*, 29(2):92, 2022. doi: 10.1097/MOH.0000000000000699.

[268] E. R. Dorris, A. O'Neill, A. Treacy et al. The transcription factor CUX1 negatively regulates invasion in castrate resistant prostate cancer. *Oncotarget*, 11(9):846–857, 2020. doi:10.18632/oncotarget.27494.

[269] Z. Wang, J. Shang, Z. Li et al. PIK3CA Is Regulated by CUX1, Promotes Cell Growth and Metastasis in Bladder Cancer via Activating Epithelial-Mesenchymal Transition. *Frontiers in Oncology*, 10, 2020.

[270] Y. Zhao, J. Zhao, M. Zhong et al. The expression and methylation of PITX genes is associated with the prognosis of head and neck squamous cell carcinoma. *Frontiers in Genetics*, 13, 2022.

[271] H. Chen, H. Li, F. Liu et al. An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Scientific Reports*, 5(1):8465, 2015. doi:10.1038/srep08465.

[272] V. S. Bachu, S. Kandoi, K. U. Park et al. An enhancer located in a Pde6c intron drives transient expression in the cone photoreceptors of developing mouse and human retinas. *Developmental Biology*, 488:131–150, 2022. doi:10.1016/j.ydbio.2022.05.012.

[273] D. Chen, H. Zhou, G. Liu et al. SPOCK1 promotes the invasion and metastasis of gastric cancer through Slug-induced epithelial-mesenchymal transition. *Journal of Cellular and Molecular Medicine*, 22(2):797–807, 2018. doi:10.1111/jcmm.13357.

[274] V. L. Veenstra, H. Damhofer, C. Waasdorp et al. Stromal SPOCK1 supports invasive pancreatic cancer growth. *Molecular Oncology*, 11(8):1050–1064, 2017. doi:10.1002/1878-0261.12073.

[275] E. M. McCabe and T. P. Rasmussen. lncRNA involvement in cancer stem cell function and epithelial-mesenchymal transitions. *Seminars in Cancer Biology*, 75:38–48, 2021. doi:10.1016/j.semcancer.2020.12.012.

[276] M. V. Kuleshov, M. R. Jones, A. D. Rouillard et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(Web Server issue):W90–W97, 2016. doi:10.1093/nar/gkw377.

[277] C.-T. Ha, C.-Y. Cheng, M.-Y. Zheng et al. ID4 predicts poor prognosis and promotes BDNF-mediated oncogenesis of colorectal cancer. *Carcinogenesis*, 42(7):951–960, 2021. doi:10.1093/carcin/bgab037.

[278] P. Zhang, Z. An, C. Sun et al. FLG Gene Mutation Up-regulates the Abnormal Tumor Immune Response and Promotes the Progression of Prostate Cancer. *Current Pharmaceutical Biotechnology*, 23(14):1658–1670, 2022.

[279] F. Yicheng, L. Xin, Y. Tian et al. Association of FLG mutation with tumor mutation load and clinical outcomes in patients with gastric cancer. *Frontiers in Genetics*, 13, 2022.

[280] D. Wang, T. Zhu, F.-B. Zhang et al. Expression of ADAMTS12 in Colorectal Cancer-Associated Stroma Prevents Cancer Development and Is a Good Prognostic Indicator of Colorectal Cancer. *Digestive Diseases and Sciences*, 56(11):3281–3287, 2011. doi:10.1007/s10620-011-1723-x.

[281] C. Li, X. Luo, B. Huang et al. ADAMTS12 acts as a cancer promoter in colorectal cancer via activating the Wnt/β-catenin signaling pathway in vitro. *Annals of Translational Medicine*, 8(6):301–301, 2020. doi:10.21037/atm.2020.02.154.

[282] Q. Gu, Y. Luo, C. Chen et al. GREM1 overexpression inhibits proliferation, migration and angiogenesis of osteosarcoma. *Experimental Cell Research*, 384(1):111619, 2019. doi:10.1016/j.yexcr.2019.111619.

[283] H. Tsubamoto, K. Sakata, R. Sakane et al. Gremlin 2 is Repressed in Invasive Endometrial Cancer and Inhibits Cell Growth In Vitro. *Anticancer Research*, 36(1):199–203, 2016.

[284] J. Wang, J. Jin, Y. Liang et al. miR-21-5p/PRKCE axis implicated in immune infiltration and poor prognosis of kidney renal clear cell carcinoma. *Frontiers in Genetics*, 13, 2022.

[285] Q. Zhen, Y. Zhang, L. Gao et al. EPAS1 promotes peritoneal carcinomatosis of non-small-cell lung cancer by enhancing mesothelial–mesenchymal transition. *Strahlentherapie und Onkologie*, 197(2):141–149, 2021. doi:10.1007/s00066-020-01665-1.

[286] E.-Y. Kang, B. Tessier-Cloutier, M. A. Duggan et al. Loss of ARID1B and SMARCB1 expression are specific for the diagnosis of dedifferentiated/undifferentiated carcinoma in tumours of the upper gynaecological tract and cervix. *Histopathology*, 79(2):160–167, 2021. doi:10.1111/his.14333.

[287] M. Coatham, X. Li, A. N. Karnezis et al. Concurrent ARID1A and ARID1B inactivation in endometrial and ovarian dedifferentiated carcinomas. *Modern Pathology*, 29(12):1586–1593, 2016. doi:10.1038/modpathol.2016.156.

[288] S. Bergeron, E. Lemieux, V. Durand et al. The serine protease inhibitor serpinE2 is a novel target of ERK signaling involved in human colorectal tumorigenesis. *Molecular Cancer*, 9(1):271, 2010. doi:10.1186/1476-4598-9-271.

[289] J. Selzer-Plon, J. Bornholdt, S. Friis et al. Expression of prostasin and its inhibitors during colorectal cancer carcinogenesis. *BMC Cancer*, 9(1):201, 2009. doi:10.1186/1471-2407-9-201.

[290] S. Chen, Y. Li, Y. Zhu et al. SERPINE1 Overexpression Promotes Malignant Progression and Poor Prognosis of Gastric Cancer. *Journal of Oncology*, 2022:e2647825, 2022. doi:10.1155/2022/2647825.

[291] F. T. Alotaibi, B. Peng, C. Klausen et al. Plasminogen activator inhibitor-1 (PAI-1) expression in endometriosis. *PLOS ONE*, 14(7):e0219064, 2019. doi:10.1371/journal.pone.0219064.

[292] J. J. Ishizuka, R. T. Manguso, C. K. Cheruiyot et al. Loss of ADAR1 in tumours overcomes resistance to immune checkpoint blockade. *Nature*, 565(7737):43–48, 2019. doi:10.1038/s41586-018-0768-9.

[293] Y. Gu, J. Jiang and C. Liang. TFAP4 promotes the growth of prostate cancer cells by upregulating FOXK1. *Experimental and Therapeutic Medicine*, 22(5):1299, 2021. doi:10.3892/etm.2021.10734.

[294] R. Dutta, B. Tiu and K. M. Sakamoto. CBP/p300 acetyltransferase activity in hematologic malignancies. *Molecular Genetics and Metabolism*, 119(1):37–43, 2016. doi:10.1016/j.ymgme.2016.06.013.

[295] Y.-H. Huang, K. Cai, P.-P. Xu et al. CREBBP/EP300 mutations promoted tumor progression in diffuse large B-cell lymphoma through altering tumor-associated macrophage polarization via FBXW7-NOTCH-CCL2/CSF1 axis. *Signal Transduction and Targeted Therapy*, 6(1):1–14, 2021. doi:10.1038/s41392-020-00437-8.

[296] E. C. Borden, G. C. Sen, G. Uze et al. Interferons at age 50: past, current and future impact on biomedicine. *Nature Reviews Drug Discovery*, 6(12):975–990, 2007. doi:10.1038/nrd2422.

[297] G. L. Beatty and Y. Paterson. IFN-γ Can Promote Tumor Evasion of the Immune System In Vivo by Down-Regulating Cellular Levels of an Endogenous Tumor Antigen1. *The Journal of Immunology*, 165(10):5502–5508, 2000. doi:10.4049/jimmunol.165.10.5502.

[298] F. J. H. Hoogwater, M. W. Nijkamp, N. Smakman et al. Oncogenic K-Ras Turns Death Receptors Into Metastasis-Promoting Receptors in Human and Mouse Colorectal Cancer Cells. *Gastroenterology*, 138(7):2357–2367, 2010. doi:10.1053/j.gastro.2010.02.046.

[299] D. J. Cha, J. L. Franklin, Y. Dou et al. KRAS-dependent sorting of miRNA to exosomes. *eLife*, 4:e07197, 2015. doi:10.7554/eLife.07197.

[300] E. J. Ruiz, L. Lan, M. E. Diefenbacher et al. JunD, not c-Jun, is the AP-1 transcription factor required for Ras-induced lung cancer. *JCI insight*, 6(13):e124985, 2021. doi:10.1172/jci.insight.124985.

[301] L. N. Kent and G. Leone. The broken cycle: E2F dysfunction in cancer. *Nature Reviews Cancer*, 19(6):326–338, 2019. doi:10.1038/s41568-019-0143-7.

[302] C. Sotiriou, M. Lacroix, L. Lespagnard et al. Interleukins-6 and -11 expression in primary breast cancer and subsequent development of bone metastases. *Cancer Letters*, 169(1):87–95, 2001. doi:10.1016/S0304-3835(01)00524-9.

[303] T. L. Putoczki, S. Thiem, A. Loving et al. Interleukin-11 Is the Dominant IL-6 Family Cytokine during Gastrointestinal Tumorigenesis and Can Be Targeted Therapeutically. *Cancer Cell*, 24(2):257–271, 2013. doi:10.1016/j.ccr.2013.06.017.

[304] C. Heichler, K. Scheibe, A. Schmied et al. STAT3 activation through IL-6/IL-11 in cancer-associated fibroblasts promotes colorectal tumour development and correlates with poor prognosis. *Gut*, 69(7):1269–1282, 2020. doi:10.1136/gutjnl-2019-319200.

[305] J. Yap, L. A. Salamonsen, T. Jobling et al. Interleukin 11 is upregulated in uterine lavage and endometrial cancer cells in women with endometrial carcinoma. *Reproductive Biology and Endocrinology*, 8(1):63, 2010. doi:10.1186/1477-7827-8-63.

[306] E. Dimitriadis, C. Stoikos, Y.-L. Tan et al. Interleukin 11 Signaling Components Signal Transducer and Activator of Transcription 3 (STAT3) and Suppressor of Cytokine Signaling 3 (SOCS3) Regulate Human Endometrial Stromal Cell Differentiation. *Endocrinology*, 147(8):3809–3817, 2006. doi:10.1210/en.2006-0264.

[307] D. K. Dimova and N. J. Dyson. The E2F transcriptional network: old acquaintances with new faces. *Oncogene*, 24(17):2810–2826, 2005. doi:10.1038/sj.onc.1208612.

[308] L. N. Kent, S. Bae, S.-Y. Tsai et al. Dosage-dependent copy number gains in *E2f1* and *E2f3* drive hepatocellular carcinoma. *The Journal of Clinical Investigation*, 127(3):830–842, 2017. doi:10.1172/JCI87583.

[309] W. Lan, B. Bian, Y. Xia et al. E2F signature is predictive for the pancreatic adenocarcinoma clinical outcome and sensitivity to E2F inhibitors, but not for the response to cytotoxic-based treatments. *Scientific Reports*, 8(1):8330, 2018. doi:10.1038/s41598-018-26613-z.

[310] L. P. Lovely, K. B. C. Appa Rao, Y. Gui et al. Characterization of androgen receptors in a well-differentiated endometrial adenocarcinoma cell line (Ishikawa). *The Journal of Steroid Biochemistry and Molecular Biology*, 74(4):235–241, 2000. doi:10.1016/S0960-0760(00)00127-8.

[311] T. E. Hickey, L. A. Selth, K. M. Chia et al. The androgen receptor is a tumor suppressor in estrogen receptor–positive breast cancer. *Nature Medicine*, 27(2):310–320, 2021. doi:10.1038/s41591-020-01168-7.

[312] J. Wade Harper, G. R. Adami, N. Wei et al. The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell*, 75(4):805–816, 1993. doi:10.1016/0092-8674(93)90499-G.

[313] M. Xu, X. Zhang, S. Zhang et al. SPOCK1/SIX1axis promotes breast cancer progression by activating AKT/mTOR signaling. *Aging*, 13(1):1032–1050, 2020. doi:10.18632/aging.202231.

[314] X. Cui, Y. Wang, W. Lan et al. SPOCK1 promotes metastasis in pancreatic cancer via NF-κB-dependent epithelial-mesenchymal transition by interacting with IκB-α. *Cellular Oncology*, 45(1):69–84, 2022. doi:10.1007/s13402-021-00652-7.

[315] P. Zhao, H.-T. Guan, Z.-J. Dai et al. Knockdown of SPOCK1 Inhibits the Proliferation and Invasion in Colorectal Cancer Cells by Suppressing the PI3K/Akt Pathway. *Oncology Research*, 24(6):437–445, 2016. doi:10.3727/096504016X14685034103554.

[316] Y. Wang, P. Huang, B. G. Wang et al. Spatial Transcriptomic Analysis of Ovarian Cancer Precursors Reveals Reactivation of IGFBP2 during Pathogenesis. *Cancer Research*, 82(24):4528–4541, 2022. doi:10.1158/0008-5472.CAN-22-1620.

[317] T. Wang, F. Lin, X. Sun et al. HOXB8 enhances the proliferation and metastasis of colorectal cancer cells by promoting EMT via STAT3 activation. *Cancer Cell International*, 19(1):3, 2019. doi:10.1186/s12935-018-0717-6.

[318] Y. Dang, J. Yu, S. Zhao et al. HOXA7 promotes the metastasis of KRAS mutant colorectal cancer by regulating myeloid-derived suppressor cells. *Cancer Cell International*, 22(1):88, 2022. doi:10.1186/s12935-022-02519-9.

[319] M. Kriseman, D. Monsivais, J. Agno et al. Uterine double-conditional inactivation of Smad2 and Smad3 in mice causes endometrial dysregulation, infertility, and uterine cancer. *Proceedings of the National Academy of Sciences*, 116(9):3873–3882, 2019. doi:10.1073/pnas.1806862116.

[320] S. Obolenski. *In Silico characterization of BCOR N1459S mutation in Endometrial Cancer*. Maastricht Univeristy, University of Oxford, 2020.

[321] F. Sievers, A. Wilm, D. Dineen et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, 2011. doi:10.1038/msb.2011.75.

[322] T. Robert Finestra and J. Gribnau. X chromosome inactivation: silencing, topology and reactivation. *Current Opinion in Cell Biology*, 46:54–61, 2017. doi:10.1016/j.ceb.2017.01.007.

[323] C. Chu, Q. C. Zhang, S. T. da Rocha et al. Systematic Discovery of Xist RNA Binding Proteins. *Cell*, 161(2):404–416, 2015. doi:10.1016/j.cell.2015.03.025.

[324] M. R. Wilson, J. J. Reske, J. Holladay et al. ARID1A Mutations Promote P300-Dependent Endometrial Invasion through Super-Enhancer Hyperacetylation. *Cell Reports*, 33(6), 2020. doi:10.1016/j.celrep.2020.108366.

[325] M. R. Corces, J. M. Granja, S. Shams et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413):eaav1898, 2018. doi:10.1126/science.aav1898.

[326] P. J. Skene and S. Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6:e21856, 2017. doi:10.7554/eLife.21856.

[327] H. S. Kaya-Okur, S. J. Wu, C. A. Codomo et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1930, 2019. doi:10.1038/s41467-019-09982-5.

[328] A. J. Rubin, K. R. Parker, A. T. Satpathy et al. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell*, 176(1):361–376.e17, 2019. doi:10.1016/j.cell.2018.11.022.

[329] N. Liscovitch-Brauer, A. Montalbano, J. Deng et al. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nature Biotechnology*, 39(10):1270–1277, 2021. doi:10.1038/s41587-021-00902-x.

[330] S. E. Pierce, J. M. Granja and W. J. Greenleaf. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nature Communications*, 12(1):2969, 2021. doi:10.1038/s41467-021-23213-w.

[331] K. Kedzierska. dchurch-lab/rna: as used in thesis. 2023. URL https://zenodo.org/record/7799704.

[332] S. Andrews, F. Krueger, A. Segonds-Pichon et al. *FastQC*. Babraham, UK, 2012. Published: Babraham Institute.

[333] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011. doi:10.14806/ej.17.1.200.

[334] F. Krueger. Trim Galore. 2018. URL https://github.com/FelixKrueger/TrimGalore.

[335] R. Patro, G. Duggal, M. I. Love et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

[336] A. Dobin, C. A. Davis, F. Schlesinger et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, 2013. doi:10.1093/bioinformatics/bts635.

[337] P. Ewels, M. Magnusson, S. Lundin et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016. doi:10.1093/bioinformatics/btw354.

[338] K. Kedzierska. kzkedzierska/atac: as used in thesis. 2023. URL https://zenodo.org/record/7799700.

[339] J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012. doi:10.1093/bioinformatics/bts480.

[340] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–595, 2010. doi:10.1093/bioinformatics/btp698.

[341] Broad Institute. *Picard Tools*. 2018. URL http://broadinstitute.github.io/picard/. Publication Title: Broad Institute, GitHub repository.

[342] H. Li, B. Handsaker, A. Wysoker et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. doi:10.1093/bioinformatics/btp352.

[343] Y. Zhang, T. Liu, C. A. Meyer et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008. doi:10.1186/gb-2008-9-9-r137.

[344] E. D. Tarbell and T. Liu. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Research*, 47(16):e91, 2019. doi:10.1093/nar/gkz533.

[345] D. Calderon, M. L. T. Nguyen, A. Mezger et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics*, 51(10):1494–1505, 2019. doi:10.1038/s41588-019-0505-9.

[346] J. Gao, B. A. Aksoy, U. Dogrusoz et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013. doi:10.1126/scisignal.2004088.

[347] L. Lopez-Delisle, L. Rabbani, J. Wolff et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics*, 37(3):422–423, 2021. doi:10.1093/bioinformatics/btaa692.

[348] W. Jawaid. enrichR: Provides an R Interface to 'Enrichr'. 2023. URL https://CRAN.R-project.org/package=enrichR.

[349] A. Subramanian, P. Tamayo, V. K. Mootha et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. doi:10.1073/pnas.0506580102.

[350] T. Wu, E. Hu, S. Xu et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021. doi:10.1016/j.xinn.2021.100141.

[351] G. Korotkevich, V. Sukhov, N. Budin et al. Fast gene set enrichment analysis. 2021. doi:10.1101/060012. URL https://www.biorxiv.org/content/10.1101/060012v3.

[352] A. Liberzon, C. Birger, H. Thorvaldsdóttir et al. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 2015. doi:10.1016/j.cels.2015.12.004.

[353] M. Kanehisa, M. Furumichi, Y. Sato et al. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592, 2023. doi:10.1093/nar/gkac963.

[354] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi:10.1093/nar/28.1.27.

[355] Z. Gu, R. Eils and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 2016.

[356] H. Jiang, R. Lei, S.-W. Ding et al. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(1):182, 2014. doi:10.1186/1471-2105-15-182.

[357] Y. Zhang, C. Park, C. Bennett et al. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Research*, 31(7):1290–1295, 2021. doi:10.1101/gr.275193.120.

[358] Y. Liao, G. K. Smyth and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–930, 2014. doi:10.1093/bioinformatics/btt656.

[359] C. Soneson, M. I. Love and M. D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. Technical Report 4:1521, F1000Research, 2016. URL https://f1000research.com/articles/4-1521. Type: article.

[360] M. E. Ritchie, B. Phipson, D. Wu et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015. doi:10.1093/nar/gkv007.

[361] J. Nangalia, F. L. Nice, D. C. Wedge et al. DNMT3A mutations occur early or late in patients with myeloproliferative neoplasms and mutation order influences phenotype. *Haematologica*, 100(11):e438–e442, 2015. doi:10.3324/haematol.2015.129510.

[362] E. Papaemmanuil, M. Gerstung, L. Malcovati et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627, 2013. doi:10.1182/blood-2013-08-518886.

[363] E. Papaemmanuil, M. Gerstung, L. Bullinger et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine*, 374(23):2209–2221, 2016. doi:10.1056/NEJMoa1516192.

[364] S. P. Shah, A. Roth, R. Goya et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, 2012. doi:10.1038/nature10933.

[365] D. A. Landau, E. Tausch, A. N. Taylor-Weiner et al. Mutations driving CLL and their evolution in progression and relapse. *Nature*, 526(7574):525–530, 2015. doi:10.1038/nature15395.

[366] R. Schwartz and A. A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017. doi:10.1038/nrg.2016.170.

[367] S. C. Dentro, D. C. Wedge and P. V. Loo. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor Perspectives in Medicine*, 7(8):a026625, 2017. doi:10.1101/cshperspect.a026625.

[368] R. Desper, F. Jiang, O.-P. Kallioniemi et al. Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data. *Journal of Computational Biology*, 6(1):37–51, 1999. doi:10.1089/cmb.1999.6.37.

[369] N. Beerenwinkel, J. Rahnenführer, M. Däumer et al. Learning Multiple Evolutionary Pathways from Cross-Sectional Data. *Journal of Computational Biology*, 12(6):584–598, 2005. doi:10.1089/cmb.2005.12.584.

[370] N. Beerenwinkel, J. Rahnenführer, R. Kaiser et al. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, 2005. doi:10.1093/bioinformatics/bti274.

[371] A. G. Deshwar, S. Vembu, C. K. Yung et al. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015. doi:10.1186/s13059-015-0602-8.

[372] G. Ha, A. Roth, J. Khattra et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, 2014. doi:10.1101/gr.180281.114.

[373] M. El-Kebir, L. Oesper, H. Acheson-Field et al. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015. doi:10.1093/bioinformatics/btv261.

[374] Y. Qiao, A. R. Quinlan, A. A. Jazaeri et al. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biology*, 15(8):443, 2014. doi:10.1186/s13059-014-0443-x.

[375] A. S. Nam, R. Chaligne and D. A. Landau. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nature Reviews Genetics*, 22(1):3–18, 2021. doi:10.1038/s41576-020-0265-5.

[376] B. L. Browning, X. Tian, Y. Zhou et al. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108(10):1880–1890, 2021. doi:10.1016/j.ajhg.2021.08.005.

[377] D. P. Cooke, D. C. Wedge and G. Lunter. A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology*, 39(7):885–892, 2021. doi:10.1038/s41587-021-00861-3.

[378] G. Gundem, P. Van Loo, B. Kremeyer et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015. doi:10.1038/nature14347.

[379] D. C. Wedge, G. Gundem, T. Mitchell et al. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nature Genetics*, 50(5):682–692, 2018. doi:10.1038/s41588-018-0086-z.

[380] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs. *Biometrika*, 39:324–335, 1952.

[381] N. Ansari-Pour, Y. Zheng, T. F. Yoshimatsu et al. Whole-genome analysis of Nigerian patients with breast cancer reveals ethnic-driven somatic evolution and distinct genomic subtypes. *Nature Communications*, 12(1):6946, 2021. doi:10.1038/s41467-021-27079-w.

[382] R. D. Luce. *Individual choice behavior*. Individual choice behavior. John Wiley, Oxford, England, 1959.

[383] R. L. Plackett. The Analysis of Permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975. doi:10.2307/2346567.

[384] C. Mollica and L. Tardella. Bayesian mixture of Plackett-Luce models for partially ranked data. *Psychometrika*, 82(2):442–458, 2017. doi:10.1007/s11336-016-9530-0.

[385] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi:10.1073/pnas.79.8.2554.

[386] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi:10.1038/323533a0.

[387] M. I. Jordan. Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986. Technical Report AD-A-173989/5/XAB; ICS-8604, California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, 1986. URL https://www.osti.gov/biblio/6910294.

[388] J. L. Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, 1990. doi:10.1207/s15516709cog1402_1.

[389] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735.

[390] I. Sutskever, J. Martens and G. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 1017–1024. Omnipress, Madison, WI, USA, 2011. ISBN 9781450306195.

[391] D. Bahdanau, K. Cho and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. 2016. doi:10.48550/arXiv.1409.0473. URL http://arxiv.org/abs/1409.0473. ArXiv:1409.0473 [cs, stat].

[392] T. Mikolov, M. Karafiát, L. Burget et al. Recurrent neural network based language model. 2010. URL https://www.semanticscholar.org/paper/Recurrent-neural-network-based-language-model-Mikolov-Karafi%C3%A1t/9819b600a828a57e1cde047bbe710d3446b30da5.

[393] M. Sundermeyer, R. Schlüter and H. Ney. LSTM Neural Networks for Language Modeling. 2012. URL https://www.semanticscholar.org/paper/LSTM-Neural-Networks-for-Language-Modeling-Sundermeyer-Schl%C3%BCter/f9a1b3850dfd837793743565a8af95973d395a4e.

[394] D. Tang, B. Qin and T. Liu. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1422–1432. Association for Computational Linguistics, Lisbon, Portugal, 2015. doi:10.18653/v1/D15-1167. URL http://aclweb.org/anthology/D15-1167.

[395] D. Eck and J. Schmidhuber. A First Look at Music Composition using LSTM Recurrent Neural Networks. Technical Report, Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 2002.

[396] J. Donahue, L. A. Hendricks, M. Rohrbach et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. 2016. doi:10.48550/arXiv.1411.4389. URL http://arxiv.org/abs/1411.4389. ArXiv:1411.4389 [cs].

[397] J. Hanson, Y. Yang, K. Paliwal et al. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, 33(5):685–692, 2017. doi:10.1093/bioinformatics/btw678.

[398] K. Sato, M. Akiyama and Y. Sakakibara. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, 12(1):941, 2021. doi:10.1038/s41467-021-21194-4.

[399] Y. Chen, Y. Li, R. Narayan et al. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 2016. doi:10.1093/bioinformatics/btw074.

[400] Y. Kim, C. Denton, L. Hoang et al. Structured Attention Networks. 2017. doi:10.48550/arXiv.1702.00887. URL http://arxiv.org/abs/1702.00887. ArXiv:1702.00887 [cs].

[401] A. Vaswani, N. Shazeer, N. Parmar et al. Attention Is All You Need. 2017. doi:10.48550/arXiv.1706.03762. URL http://arxiv.org/abs/1706.03762. ArXiv:1706.03762 [cs].

[402] N. Srivastava, G. Hinton, A. Krizhevsky et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[403] W. J. Gibson, E. A. Hoivik, M. K. Halle et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nature Genetics*, 48(8):848–855, 2016. doi:10.1038/ng.3602.

[404] B. C. Morson. Evolution of cancer of the colon and rectum. *Cancer*, 34(S3):845–849, 1974. doi:10.1002/1097-0142(197409)34:3+<845::AID-CNCR2820340710>3.0.CO;2-H.

[405] S. M. Powell, N. Zilz, Y. Beazer-Barclay et al. APC mutations occur early during colorectal tumorigenesis. *Nature*, 359(6392):235–237, 1992. doi:10.1038/359235a0.

[406] W. Cross, M. Kovac, V. Mustonen et al. The evolutionary landscape of colorectal tumorigenesis. *Nature Ecology & Evolution*, 2(10):1661–1672, 2018. doi:10.1038/s41559-018-0642-z.

[407] D. J. Woodcock, A. Sahli, R. Teslo et al. Genomic evolution shapes prostate cancer disease type. *Cell Genomics*, 4(3), 2024. doi:10.1016/j.xgen.2024.100511.

[408] S. Rao, C. Verrill, L. Cerundolo et al. Intra-prostatic tumour evolution, steps in metastatic spread and histogenomic associations revealed by integration of multi-region whole genome sequencing with histopathological features. 2023. doi:10.1101/2023.02.27.530113. URL https://www.biorxiv.org/content/10.1101/2023.02.27.530113v2. Pages: 2023.02.27.530113 Section: New Results.

[409] T. J. Mitchell, S. Turajlic, A. Rowan et al. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell*, 173(3):611–623.e17, 2018. doi:10.1016/j.cell.2018.02.020.

[410] OpenAI. GPT-4 Technical Report. 2023. doi:10.48550/arXiv.2303.08774. URL http://arxiv.org/abs/2303.08774. ArXiv:2303.08774 [cs].

[411] D. Lee, J. Yang and S. Kim. Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nature Communications*, 13(1):6678, 2022. doi:10.1038/s41467-022-34152-5.

[412] A. Travaglino, A. Raffone, A. Santoro et al. Clear cell endometrial carcinomas with mismatch repair deficiency have a favorable prognosis: A systematic review and meta-analysis. *Gynecologic Oncology*, 162(3):804–808, 2021. doi:10.1016/j.ygyno.2021.07.007.

[413] Z. Liu, Y. Zhao, P. Kong et al. Integrated multi-omics profiling yields a clinically relevant molecular classification for esophageal squamous cell carcinoma. *Cancer Cell*, 41(1):181–195.e9, 2023. doi:10.1016/j.ccell.2022.12.004.

[414] M. Crispin-Ortuzar, R. Woitek, M. A. V. Reinius et al. Integrated radiogenomics models predict response to neoadjuvant chemotherapy in high grade serous ovarian cancer. *Nature Communications*, 14(1):6756, 2023. doi:10.1038/s41467-023-41820-7.

[415] S. F. Roerink, N. Sasaki, H. Lee-Six et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, 556(7702):457–462, 2018. doi:10.1038/s41586-018-0024-3.

[416] C. Gerhauser, F. Favero, T. Risch et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer Cell*, 34(6):996–1011.e8, 2018. doi:10.1016/j.ccell.2018.10.016.

[417] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1-4414-1269-7.

[418] P. GNU. *Free Software Foundation. Bash (3.2. 48)[Unix shell program]*. 2007.

[419] A. McKenna, M. Hanna, E. Banks et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010. doi:10.1101/gr.107524.110.

[420] M. A. DePristo, E. Banks, R. Poplin et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011. doi:10.1038/ng.806.

[421] J. Ou, H. Liu, J. Yu et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*, 19(1):169, 2018. doi:10.1186/s12864-018-4559-3.

[422] M. Morgan. *BiocManager: Access the Bioconductor Project Package Repository*. 2019. URL https://CRAN.R-project.org/package=BiocManager.

[423] G. Yu, L.-G. Wang and Q.-Y. He. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383, 2015. doi:10.1093/bioinformatics/btv145.

[424] G. Yu, L.-G. Wang, Y. Han et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012. doi:10.1089/omi.2011.0118.

[425] J. Hester, J. Bryan and RStudio. glue: Interpreted String Literals. 2022. URL https://CRAN.R-project.org/package=glue.

[426] J. B. Arnold, G. Daroczi, B. Werth et al. ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. 2021. URL https://CRAN.R-project.org/package=ggthemes.

[427] K. Slowikowski. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. 2020. URL https://CRAN.R-project.org/package=ggrepel.

[428] N. Xiao. *ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'*. 2018. URL https://CRAN.R-project.org/package=ggsci.

[429] C. O. Wilke and B. M. Wiernik. ggtext: Improved Text Rendering Support for 'ggplot2'. 2022. URL https://CRAN.R-project.org/package=ggtext.

[430] D. D. Sjoberg, J. Larmarange, M. Curry et al. gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. 2023. URL https://CRAN.R-project.org/package=gtsummary.

[431] G. R. Warnes, B. Bolker, T. Lumley et al. *gmodels: Various R Programming Tools for Model Fitting*. 2018. URL https://CRAN.R-project.org/package=gmodels.

[432] F. Hahne and R. Ivanek. Statistical Genomics: Methods and Protocols. 335–351. Springer New York, New York, NY, 2016. ISBN 978-1-4939-3578-9. doi:10.1007/978-1-4939-3578-9_16. URL http://dx.doi.org/10.1007/978-1-4939-3578-9_16. Section: Visualizing Genomic Data Using Gviz and Bioconductor.

[433] Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. 2020. URL https://yihui.org/knitr/.

[434] Y. Xie. knitr: A Comprehensive Tool for Reproducible Research in R. In V. Stodden, F. Leisch and R. D. Peng (editors), *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595.

[435] Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL https://yihui.org/knitr/.

[436] V. Spinu, G. Grolemund, H. Wickham et al. lubridate: Make Dealing with Dates a Little Easier. 2023. URL https://CRAN.R-project.org/package=lubridate.

[437] A. Mayakonda, D.-C. Lin, Y. Assenov et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, 28(11):1747–1756, 2018. doi:10.1101/gr.239244.118.

[438] T. L. Pedersen. *patchwork: The Composer of Plots*. 2020. URL https://CRAN.R-project.org/package=patchwork.

[439] K. Kuang, Q. Kong and F. Napolitano. *pbmcapply: Tracking the Progress of Mc*pply with Progress Bar*. 2019. URL https://CRAN.R-project.org/package=pbmcapply.

[440] E. Neuwirth. *RColorBrewer: ColorBrewer Palettes*. 2014. URL https://CRAN.R-project.org/package=RColorBrewer.

[441] H. Wickham, M. Averick, J. Bryan et al. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi:10.21105/joss.01686.

[442] B. C. Team and B. P. Maintainer. *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*. 2019.

[443] S. Garnier. *viridis: Default Color Maps from 'matplotlib'*. 2018. URL https://CRAN.R-project.org/package=viridis.

[444] F. Chollet et al. Keras. 2015. URL https://keras.io.

[445] C. R. Harris, K. J. Millman, S. J. van der Walt et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi:10.1038/s41586-020-2649-2.

[446] T. p. d. team. pandas-dev/pandas: Pandas. 2023. URL https://zenodo.org/record/7794821.

[447] P. Virtanen, R. Gommers, T. E. Oliphant et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi:10.1038/s41592-019-0686-2.

[448] T. Developers. TensorFlow. 2023. URL https://zenodo.org/record/7764425.

[449] S. Savas and G. Skardasi. The SWI/SNF complex subunit genes: Their functions, variations, and links to risk and survival outcomes in human cancers. *Critical Reviews in Oncology/Hematology*, 123:114–131, 2018. doi:10.1016/j.critrevonc.2018.01.009.

[450] EMBL-EBI. HGNC Database, HUGO Gene Nomenclature Committee (HGNC). 2020. URL https://www.genenames.org/data/genegroup/.

[451] R. D. Luce. The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15(3):215–233, 1977. doi:10.1016/0022-2496(77)90032-3.