Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

Siao-Han Wong, M.Sc.

born in: ChiaYi, Taiwan

Oral examination: 6th July, 2021

Tumor Evolution and Heterogeneity in High-Grade Serous Ovarian Cancer

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. med. Frederik Marmé

SUMMARY

Ovarian cancer (OC) is a heterogeneous disease and can be delineated into five major histological subtypes. In 2018, the disease caused 184,799 deaths worldwide, where the majority of them were due to high-grade serous carcinoma (HGSC). HGSC accounts for 70% of OCs and have a common late-stage diagnosis. Patients usually show initial favorable response to chemotherapy but later on subject to disease relapse and development of acquired resistance.

Understanding the disease biology is important for its early detection and effective treatment. The pathogenesis of HGSC had remained obscure until the identification of its tubal origin. Studies on precursor lesions had gained insights into very early events; however, the carcinogenesis process remained largely underexplored. Meanwhile, molecular subtyping became more important due to the success of targeted therapy using poly(ADP-ribose) polymerase (PARP) inhibitors. However, current clinical trials use different assays and a consensus approach for patient stratification is lacking.

In this thesis, whole genome sequencing (WGS) was used to profile tumor-normal sample pairs from ovarian cancer patients, and a subset of them had samples collected from different anatomical sites. This multi-sample cohort (HIPO59) is suitable for addressing questions about molecular stratification and tumor pathogenesis.

Unifying contemporary DNA-based classifications, more evidences were provided here and consolidated a concept of HGSC dichotomy. Our data suggested that the previously proposed genomic subgroups in HGSC (H-HRD and H-FBI) are characterized by different extent and onset timing of homologous recombination repair (HRR) defect as well as CCNE1 pathway activation. Specifically, HRR deficiency is a common feature and acquired early in H-HRD cases, whereas H-FBI tumors often have CCNE1 pathway activation as an early event. Mechanistic details supporting this observation were revealed by several layers of evidences. Among them, the subgroups also showed differences in surrogate biomarkers for PARP inhibitor response. The HGSC dichotomy reflects meaningful biology of the disease, provides a new perspective of interpreting known biomarkers, and holds the potential for better describing the patient subset that are more likely to benefit from PARP inhibitor treatment.

To get insight into tumor evolution in real-world time, the evolutionary trajectory was computationally reconstructed for each tumor. This highlights an early bifurcation of carcinogenesis pathways in the HGSC dichotomy, despite a common scenario of very early TP53 mutation and an eventual chromosomal instability (CIN) phenotype seen in both subgroups. Furthermore, the reconstructed sample phylogeny trees inform about potential early and late driving events for tumors from all individuals. This can facilitate personalized oncology by providing clinical implications in cascade testing, therapeutic planning and disease monitoring.

Besides, our data raised additional questions worthwhile further investigations. First of all, heredity was found as an influential factor in all histotypes. Contributing factors other than BRCA mutations underlined the link between hereditary DNA repair-deficiency syndromes and ovarian cancer predisposition. Secondly, a variable degree of intra-patient heterogeneity (IPH) was observed in pre-treatment samples of OC and the IPH can potentially stratify patients into distinct prognostic groups. As the method summarizing IPH here supports a wide range of high-throughput profiling platforms, developing a standardized assay suitable for a larger cohort can help evaluate its clinical utility.

Overall, these findings corroborate the concept of HGSC dichotomy by providing mechanistic underpinnings from the aspect of tumor evolution. This integrative viewpoint allows for re-interpreting contemporary knowledge about HGSC, and will help scientists formulate questions about subtype-specific pathogenesis and vulnerabilities. Furthermore, both the dichotomy and IPH status allow for molecular stratification with potential clinical implications. Altogether, the findings in this thesis provide novel opportunities for discovering solid understanding about the biology of HGSC, as well as facilitating personalized oncology in ovarian cancer treatment.

ZUSAMMENFASSUNG

Das Ovarialkarzinom (OC) ist eine heterogene Erkrankung und lässt sich in fünf große histologische Subtypen einteilen. Im Jahr 2018 führte die Erkrankung weltweit zu 184.799 Todesfällen, von denen die meisten auf ein hochgradiges seröses Karzinom (HGSC) zurückzuführen waren. HGSC machen 70% der OCs aus und werden häufig im Spätstadium diagnostiziert. Die Patienten sprechen in der Regel zunächst gut auf Chemotherapie an, erleiden aber später einen Krankheitsrückfall und entwickeln eine Therapieresistenz.

Die Biologie hinter OC zu verstehen ist wichtig für die Früherkennung und eine effektive Behandlung. Die Pathogenese des HGSC war bis zur Entdeckung ihres tubulären Ursprungs ungeklärt. Studien an Vorläuferläsionen lieferten zwar Einblicke in sehr frühe Mutationsereignisse, der Prozess der Karzinogenese blieb jedoch weitgehend unerforscht. Inzwischen hat die molekulare Subtypisierung durch den Erfolg der zielgerichteten Therapie mit Poly(ADP-Ribose)-Polymerase (PARP)-Inhibitoren an Bedeutung gewonnen. In aktuellen klinischen Studien werden jedoch unterschiedliche Assays verwendet und es fehlt ein einheitlicher Ansatz zur Patientenstratifizierung.

In dieser Arbeit wurde Gesamt-Genom-Sequenzierung (WGS) verwendet, um Proben von Ovarialkarzinom-Patientinnen zu analysieren. Bei einem Teil der Patientinnen wurden Proben von verschiedenen Stellen im Tumor entnommen. Diese Multi-Proben-Kohorte (HIPO59) ist gut dafür geeignet, die molekulare Stratifizierung der Patientinnen zu verbessern und Einblicke in die Tumorpathogenese zu gewinnen.

Durch die Vereinheitlichung aktueller DNA-basierter Klassifizierungen wurden hier weitere Beweise für die HGSC-Dichotomie geliefert und das Konzept somit gefestigt. Unsere Daten deuten darauf hin, dass die zuvor vorgeschlagenen genomischen Untergruppen bei HGSC (H-HRD und H-FBI) durch ein unterschiedliches Ausmaß und einen unterschiedlichen Zeitpunkt des Einsetzens eines Defekts in der homologen Rekombinationsreparatur (HRR) sowie der Aktivierung des CCNE1-Signalwegs gekennzeichnet sind. Insbesondere ist der HRR-Defekt ein häufiges Merkmal und wird in H-HRD-Fällen früh erworben, während bei H-FBI-Tumoren die CCNE1-Signalweg-Aktivierung häufig als frühes Ereignis auftritt. Mechanistische Details, die diese Beobachtung unterstützen, wurden durch mehrere Anhaltspunkte herausgearbeitet. So zeigten die Untergruppen auch Unterschiede bei den Surrogat-Biomarkern für das Ansprechen auf PARP-Inhibitoren. Die HGSC-Dichotomie ist sehr aussagekräftig für die Entstehung der Erkrankung, bietet eine neue Perspektive für die Interpretation bekannter Biomarker und gibt Einsichten in Patientgruppen mit wahrscheinlich gutem Ansprechen auf eine PARP-Inhibitor-Behandlung.

Um einen Einblick in die Tumorevolution zu erhalten, wurde die Evolutionskurve für jeden Tumor rechnerisch rekonstruiert. Dies zeigt eine frühe Bifurkation der Karzinogenese in der HGSC-Dichotomie, trotz gemeinsamer früher TP53-Mutationen und

späterer chromosomaler Instabilität (CIN) in beiden Untergruppen. Darüber hinaus liefert der rekonstruierte Proben-Phylogenie-Baum Informationen über potenzielle weitere frühe und späte treibende Ereignisse für alle Tumoren. Dies kann die personalisierte Onkologie unterstützen, indem es klinische Implikationen für Kaskadentests, Therapieplanung und Krankheitsüberwachung liefert.

Außerdem warfen die Daten zusätzliche Fragen auf, die weitere Untersuchungen erfordern. Zunächst einmal wurde genetische Vererbung als einflussreicher Faktor in allen Histotypen gefunden. Neben BRCA-Mutationen verdeutlichten auch andere Einflussfaktoren den Zusammenhang zwischen erblichen DNA-Reparaturdefekten und einer Prädisposition für Eierstockkrebs. Zweitens wurde ein variabler Grad an Intra-Patienten-Heterogenität (IPH) in unbehandelten OC-Proben beobachtet, was möglicherweise der Stratifizierung von Patientinnen in verschiedene prognostische Gruppen dienen kann. Da die hier zusammengefasste IPH-Methode eine breite Palette von Hochdurchsatz-Analyseverfahren unterstützt, kann die Entwicklung eines standardisierten Assays für eine größere Kohorte helfen, den klinischen Nutzen von IPH zu bewerten.

Insgesamt bestätigen diese Ergebnisse das Konzept der HGSC-Dichotomie, indem sie mechanistische Einblicke aus der Perspektive der Tumorevolution liefern. Diese integrative Sichtweise ermöglicht eine Neuinterpretation des gegenwärtigen Wissens über HGSC und wird Wissenschaftlern dabei helfen, Fragen zur subtypspezifischen Pathogenese und Vulnerabilität zu formulieren. Darüber hinaus ermöglichen sowohl die Dichotomie als auch der IPH-Status eine molekulare Stratifizierung mit potenziellen klinischen Implikationen. Alles in allem bieten die Ergebnisse dieser Arbeit neue Möglichkeiten, um ein solides Verständnis über die Biologie von HGSC zu erlangen sowie die personalisierte Onkologie bei der Behandlung von Eierstockkrebs zu verbessern.

CONTENTS

III RESULTS

LIST OF FIGURES

LIST OF TABLES

Part I

BACKGROUND

1 TUMOR DEVELOPMENT

1.1 HALLMARKS OF CANCER

Tumor follows a multistep development, through which it acquires biological capabilities to become malignant and achieve dissemination. Hallmark capabilities of tumor was proposed by Hanahan & Weinberg in 2000[1] and updated in 2011[2]. The current eight hallmark of cancer serve as the organizing principle of tumor and are briefly summarized below.

Sustaining proliferative signaling

Mitogenic signaling is usually induced by growth factors acting on cells through ligand-receptor binding, and facilitated by different intracellular transducers regulating cell cycle, cell growth, cell survival and energy metabolism. Cancer cells sustain proliferative signaling in several ways, including autocrine proliferative stimulation, stimulating surrounding normal cells which in turn provide them with growth factor supply, increasing sensitivity to ligand, altering molecular components to achieve constitutive activation, or disruption of negative-feedback mechanisms and counteracting responses like senescence and apoptosis.

Evading growth suppressors

RB and TP53 are two canonical suppressors that limit cell growth and proliferation. RB integrates extracellular and intracellular signals and decides whether or not to proceed through cell cycle, its absence therefore permits persistent cell proliferation. TP53, on the other hand, integrates intracellular signals about genomic damage and cell stresses and decides either to halt the progression to cell cycle, or to trigger apoptosis in case of overwhelming or irreparable damage. Despite their importance, functional redundancy of these genes exists in the biological network. Therefore, in animal models lacking a functional RB gene or TP53 gene, neoplasia were observed only later in life. Two other proliferation-suppressive mechanisms include contact-mediated growth inhibition and TGF-beta signaling, where the former is compromised in cancers and the latter is redirected to activate epithelial-to-mesenchymal transition (EMT) program.

Resisting cell death

Apoptosis is a process of programmed cell death. In response to extrinsic or intrinsic signals, mediating proteases are activated and initiate downstream proteolysis cascades,

which eventually lead to cell death. The intrinsic apoptotic program, in particular, is considered a natural barrier to carcinogenesis and its dysfunction is implicated in many cancers. For example, TP53 integrate internal signals of DNA breaks and chromosomal abnormalities and is key to the induction of apoptosis through the route of DNA-damage sensing. Therefore, loss of TP53 tumor suppressor function is often observed in cancer. However, this is not the only way to achieve this hallmark and multiple other apoptosis-avoiding mechanisms have been observed.

Enabling replicative immortality

Two intrinsic mechanisms are known to prevent normal cells from unlimited replication. The first mechanism, senescence, is when cells enter a nonproliferative while viable state. On the other hand, the second mechanism, crisis, can lead to apoptosis. To obtain unlimited replicative potential, cells have to overcome these two barriers. In tumor cells, the mechanism to obtain this immortality is mainly through activating its telomere protection mechanism, which can be achieved through telomerase re-activation or alternative lengthening mechanisms.

Inducing angiogenesis

Angiogenesis is neovasculature supporting tumor cells with required nutrients and oxygen supply. In normal processes, new vasculature is developed during embryogenesis, or in adult stage it can be transiently activated upon wound healing or female reproductive cycling. In contrast, this process is constantly activated in tumor cells by disrupting angiogenic regulators, with two known examples being the induction of VEGF-A and inhibition of TSP-1. More intriguingly, early acquisition of this hallmark has been widely observed.

Activating invasion and metastasis

Outgrowth of tumor cells requires ability to invade adjacent tissues and become a malignancy, as well as to disseminate and enable distant metastases. Along the process they acquired morphological changes, and molecularly it is best described by the loss of E-cadherin. It is a key molecule for maintaining cell-to-cell contacts and its tumor suppressor role is supported by its common disruption in cancer. With dedicated studies into the invasion-metastasis cascade process, more positive and negative regulators are found to be involved in gaining the capability for invasion and metastasis.

Reprogramming of energy metabolism

Due to the drastic change from the normal state, tumor cells have to adjust their energy metabolisms accordingly. A phenomenon termed "aerobic glycolysis" has been observed, where cancer cells can reprogram their energy metabolism and become largely

relying on glycolysis even in anaerobic condition. Concordant molecular changes observed include upregulating glucose transporter like GLUT1 which increases glucose import into the cell. This change also has been associated with oncogene (e.g. RAS, MYC) activation, tumor suppressor gene (e.g. TP53) inactivation and hypoxic conditions in tumors. Notably, within the tumor, scientists had observed two symbiotic subpopulations where one of them had this metabolic switch and provide lactate as the energy source of the other subpopulation. Nonetheless, this observation has not been generalized yet and there are still many unresolved issues surrounding this emerging hallmark of cancer.

Evading immune destruction

The theory of immune surveillance proposed that the immune system has been constantly monitoring cells in the body and eradicate the formation and progression of tumor cells. In this sense, acquiring capability to avoid this detection system would be another hallmark for cancer cells to survive. There has been increasing evidences from genetically engineered mice and clinical epidemiology suggesting the role of immune system as a barrier to tumor development. Nonetheless, the immunoevasion as a emerging hallmark still remain to be firmly established.

1.2 DNA DAMAGE REPAIR DEFECT AND CANCER

It emerged from the studies of hereditary cancer predisposition syndromes that DNA damage response (DDR) might be playing a role in cancer development. The rare syndromes are mostly due to highly penetrant variants and 5-10% of all cancers are developed in patients with such syndromes[3]. Despite the small hereditary cancer burden, the advanced understanding of their genetic basis had shed light on the development of sporadic tumors.

The following two sections describe how the link between DDR-deficiency and carcinogenesis got recognized, and how often this defect is acquired in common sporadic cancers.

1.2.1 *Cancer predisposition syndromes*

Cancer predisposition syndrome describes the situation when a person is born with specific genetic allele that increases the carrier's risk of developing cancer. These genetic alleles are usually rare in the population and associated with compromised function in key genes. As cancer is considered an age-related disease where it develops by gradually accrual of key genetic changes along lifetime, it may therefore develop earlier in individuals born with such risk alleles.

Seminal observation on the association between DDR and cancer predisposition can be traced back to a study in 1969 conducted on a rare genetic disorder[4]. *Xeroderma pigmentosum* (*XP*) is a autosomal recessive disease, where patients are conferred a 1,000-fold increase in the risk of developing skin cancer[5]. Causal mutations were found in genes functioning in the nucleotide excision repair (NER) pathway, which is responsible for the repair of ultraviolet (UV) light-induced DNA damage. The dysfunction of NER pathway therefore leads to an elevated susceptibility to skin cancer.

More studies further strengthened the bond between DNA repair pathways and carcinogenesis. *Fanconi anaemia* (*FA*) happens when pathogenic variants occurred in 22 DDR-related genes and made cells incapable of repairing DNA inter-strand cross-links (ICLs)[6, 7]. On the other hand, the genetic basis of *Bloom syndrome* was found to be the loss-of-function mutations in BLM[8], a gene encoding helicase involved in homologous recombination repair (HRR) pathway[9] and thus the defect compromises DNA double-strand break (DSB) repair. Likewise, other DNA helicases are implicated in other hereditary syndromes. Mutations in WRN, for example, is the causative factor of *Werner syndrome*[10]. And when RECQL4 is affected, it lead to three other clinical syndromes, namely *Rhmund-Thomson*, *RAPADILINO* and *Baller-Gerold syndromes*[11]. More such syndromes associated with causal link to DDR genes include *Li-Fraumeni syndrome* (TP53 gene), *Hereditary Breast and Ovarian Cancer syndrome* (*HBOC syndrome*) (BRCA1 or BRCA2 genes), *Ataxia-Telanlectasia* (ATM gene), *Cowden Syndrome* (PTEN gene), *Nijmegen breakage syndrome* (NBN gene), to name but a few.

Patients with cancer predisposition syndromes can develop a wide spectrum of neoplasms and inversely, hereditary cancers are linked to different sets of rare syndromes. Colorectal cancer, for example, is associated with several high-risk syndromes

involving defects in different DNA repair machineries. *Lynch syndrome*, characterized by microsatellite instability (MSI), results from dysfunction in DNA mismatch repair genes[12–15] and has a clinical presentation of nonpolyposis. Notably, the MSI feature renders these tumors more immunogenic and therefore these patients respond well to immunotherapy. Alternatively, alterations in MYH[16], a base-excision repair (BER) gene, had been found to be the cause of *MUTYH-associated polyposis* (*MAP*). Similarly, a subset of familial breast, ovarian, prostate and pancreatic cancers may also arise from inherited defect in HRR pathway[17].

In terms of ovarian cancer, the associated genetic factors will be discussed in Section 3.2 and associated hereditary syndromes include *HBOC syndrome*, *Lynch Syndrome* and *MAP*[18].

1.2.2 *DNA damage repair deficiency in sporadic cancers*

A re-analysis of TCGA data focused on 9 DDR pathways across 33 cancer types[19]. Researchers looked at somatic disruptions in 276 DDR genes due to three alteration mechanisms, including mutation/indel, copy number loss and epigenetic silencing.

Direct repair (DR) and HRR were the most frequently altered DDR pathways across all tumors. Among the 28 associations showing enrichment of pathway gene alterations in different cancer types, NHEJ and HRR were enriched for alterations within ovarian cancer. Notably, these pathway disruptions have distinct prognostic implications in different cancer types. HRR disruption, for example, is associated with better outcome in ovarian cancer but worse outcome in several other cancers.

In terms of alteration mechanisms, there was observed disproportionality of mutation and deletion in HRR pathway, while DR pathway was mainly altered by epigenetic silencing. Interestingly, some genes are more prone to specific alteration mechanisms. For example, ALKBH3 and MGMT were affected predominantly by silencing. On the other hand, mutations accompanied by loss of heterozygosity (LOH) was observed in one third of DDR genes, including TP53, BRCA1, BRCA2, PTEN and PER1.

In summary, the authors showed that somatic disruptions in DDR gene were ubiquitous within major cancer types. Two third of the 33 cancer types showed enrichment of alterations in at least one of the 9 DDR pathways. These alterations results from three disruption mechanisms and have different implications depend on cancer types.

2 MOLECULAR FOOTPRINTS OF BIOLOGICAL PROCESSES

In human cells, DNA damage repair mechanisms have evolved into a comprehensive and redundant system to counteract constant genotoxic insult arising from intrinsic or environmental sources, this maintains normal cellular functions and ensures faithful transmission of genetic information to daughter cells. When the balance is not maintained, DNA damage left unrepaired may be passed on to the daughter cells and appear as somatic aberrations when compared to the germline genome. As some endogenous or exogenous mutational processes and repair mechanisms tend to alter or repair DNA in specific context leading to unique alteration types, somatic mutations detected in the tumor genome would exhibit specific pattern reflective of ongoing or historical biological processes.

These genomic scars can take on forms of single base substitutions, indels, rearrangements or copy number alterations. Therefore, one can theoretically describe these unique footprints using a set of features consisting of these aberrations, together with their accessory characteristics. Given the collection of genomic scars observed in numerous tumor genomes, which is a superimposition of patterns from different processes in different cancer types, computational approaches[20] were applied to discover **mutational signatures**, described by defined features, where each of them characterizes footprints of specific combination of mutational processes and/or repair mechanism defects.

The following sections provide an overview of the discovery of footprints, identified mutational signatures and their implications in underlying biological processes.

2.1 MUTATOR PHENOTYPE

In theory, if one would have known all the causes of aberrations, it is possible to computationally infer all underlying processes. However, it is oftentimes not the case. Historically, scientists had observed large numbers of mutations present in cancers. In 1974, a hypothesis was proposed[21] stating that cancers expresses a **mutator phenotype** and accumulates mutations at a rate higher than nonmalignant cells. The origins and consequences of mutator phenotype only started to get appreciated after the invent of next generation sequencing[22].

As technology allows for detecting aberrations of more forms, other mutator phenotypes such as with excess duplication events[23] and their potential origins were gradually revealed.

2.2 SINGLE BASE SUBSTITUTION SIGNATURE

From the catalogue of somatic mutations detected in 7,042 tumors, scientists were able to identify 21 Single Base Substitution Signatures (SBS Signatures)[24], each of them characterized by a combination of SBSs arising in specific localized trinucleotide contexts. Some of the signatures were found associated with age, mutagenic exposures, DNA maintenance defect or other mutational processes, while many remained of cryptic origin.

Abnormalities in DNA maintenance, for instance, are reflected by different SBS Signatures depending on which of the repair axis being disrupted. Signature 3 is suggestive of defective HRR as it was strongly associated with BRCA1 and BRCA2 mutations within two cancer types (breast and pancreatic cancer). Despite that this signature has a rather equal representation of all features, it was further found associated with indels with microhomology at the breakpoints, a feature being utilized by NHEJ mechanisms for rejoining DSBs. On the other hand, defective MMR was suggested to be responsible for four SBS Signatures, namely Signature 6, 15, 20 and 26. These signatures were associated with substantial numbers of substitutions and small indels at nucleotide repeats.

The major component of Signature 1 is CpG>TpG transitions, based on which the underlying mutational process is linked to spontaneous deamination of 5-methylcytosine to thymine at CpG sites. As Signature 1 positively correlated with age in most cancer types, it was hypothesized that these mutations have been acquired at a relatively constant rate throughout life, and that this endogenous mutational process operates in a clock-like manner. In this sense, its footprint will proportionally reflect the chronological age of an individual. In a follow-up study, Alexandrov *et al.* confirmed that both Signature 1 and Signature 5 showed clock-like properties, the authors further hypothesized that Signature 1 mutation rate can act as a molecular clock reflecting the number of mitoses a cell had experienced.

Lastly, Signature 2 and Signature 13 were attributed to abnormal activity of members of the APOBEC family of cytidine deaminases. Signature 8 was of unknown origin and showed strong transcriptional strand bias; however it was observed to be associated with absence of BRCA1 and BRCA2 function in breast cancer[26]. Signature 9 was proposed to be due to an error-prone polymerase eta, which is involved in AID-induced somatic hypermutation as well as DNA repair by translesion synthesis. More biological processes underlying other signatures can be found in the original publication.

2.3 INDEL SIGNATURES

In the most recent study, Alexandrov *et al.* identified 17 Indel Signatures (ID Signatures) based on small insertions and deletions in 4,645 whole-genome and 19,184 whole exome sequences. They also expanded the SBS Signature set to 49 SBS Signatures.

Among other clock-wise signatures, ID1 is characterized with insertions of thymine and likely generated during DNA replication of long mononucleotides due to slippage of the nascent strand.

Abnormalities in DNA maintenance can again be traced for defects in HRR or mismatch repair (MMR). ID6 and ID8 are both related to defect in HRR, and are both characterized by deletions of size >= 5-bp. They differ in the size of microhomology at the junction, where ID6 is associated with longer stretches of microhomology and ID8 is more with no or 1-bp microhomology. Both of them were suggested to be characteristic of NHEJ mechanisms. Furthermore, ID6 correlate with Signature 3, an SBS Signature reflecting defect in HRR, whereas ID8 did not show strong correlation.

Additionally, when ID1 and ID2 appear in large amount, they were usually accompanied by new SBS Signature 6, 14, 15, 20, 21, 26 and/or 44, reflecting defect in MMR. These MMR defect-associated ID Signatures would sometimes accompanied by polymerase proofreading deficiency for POLE or POLD1 (new SBS Signature 14 and 20).

2.4 REARRANGEMENT SIGNATURES

Mutational signature analysis can be extended to structural variations (SVs). It was first investigated in 2016 in a breast cancer study, where researchers extracted six Rearrangement Signatures from 560 whole-genome sequences[26]. SVs were classified into 32 features first by regional clustering property, then by types including tandem duplications (TDs), deletions (DELs), inversions (INVs), interchromosomal translocations (TRXs), lastly by the size of the rearrangement event (1-10 kb, 10-100kb, 100-1000kb, 1-10 Mb, or >10 Mb).

There were three Rearrangement Signatures associated with HRR defect, namely RS1, RS3 and RS5. RS1 represents large TD mutator phenotype and was of unknown origin. RS3 correspond to small TD mutator phenotype and is likely due to inactivation of BRCA1. RS5 is characterized by short DELs and associated with either BRCA1 or BRCA2 inactivations.

2.5 COPY NUMBER-BASED GENOMIC SIGNATURE

It had been reported that tumors with homologous recombination deficiency (HRD) would be sensitive to drugs inducing DNA cross-links such as cisplatin. These tumors might rely on error-prone repair processes for repairing DSBs and survive, resulting in various genomic abnormalities and leading to genome instability. Three types of **genomic signatures**, which measure genome-wide count of abnormal chromosomal regions, had been proposed to indicate the degree of DSB repair incompetence, and to predict sensitivity to platinum-based treatments.

Telomeric Allelic Imbalance (TAI)

Using error-prone repair processes for repairing DNA DSBs results in genomic abnormalities including high level of allelic imbalance (AI). A study looked for associations between platinum-based treatment response and different genome-wide summary measures of AI, and found that number of regions with telomeric allelic imbalance (TAI) being predictive of cisplatin sensitivity in breast cancer cell lines[28]. Number of TAI (N_tAI), defined as number of subchromosomal regions with AI extending to the telomere, were further found to predict pathologic response to preoperative treatment in TNBC, and correlate with better initial response in the TCGA ovarian cancer cohort. In summary, this study showed that TAI is a marker of platinum sensitivity and suggestive of impaired DNA repair, and proposed that N_tAI being a useful biomarker for identifying patients with wild-type BRCA1/2 but likely benefit from platinum-based therapies.

Loss of Heterozygosity (LOH)

Given that LOH is an irreversible event, scientists hypothesized that LOH-based score would provide a more stable record compared to copy number alterations. The hypothesis was examined in three ovarian cancer cohorts and a panel of ovarian, breast, colon and pancreatic cancer cell lines[29]. Scientists first characterize inactivation of BRCA1, BRCA2 and RAD51C, and then compare the LOH-based scores in samples with and without HRR deficiency.

HRR-deficient tumors were defined by samples with germline or somatic BRCA1 or BRCA2 mutations, promoter methylation or low transcript expression of BRCA1, and exhibited homozygosity at the affected gene due to LOH. Three LOH-based scores measuring LOH of different sizes were tested for association with the HRD status. Number of short LOH regions (<15Mb) was not associated with HRD in all three cohorts, whereas whole-chromosome LOH was more abundant in HR-proficient tumors in two cohorts. The intermediate sized LOH (>15Mb but less than a whole chromosome), later on referred to as the **HRD score**, was significantly more abundant in tumors deficient in BRCA1 or BRCA2 function in all three cohorts. Similar association observed in cell line panels further showed that the this biomarker is not restricted to EOCs.

Lastly, the authors also found that promoter methylation of RAD51C, as well as PTEN-deficient tumors both lead to an elevated HRD score.

Large-scale State Transitions (LST)

In search of surrogate genomic markers for BRCA1-inactivation, researchers examined the copy number profiles of 65 basal-like breast carcinomas (BLCs) using SNP arrays, and found two features strongly predictive of BRCA1 inactivation[30], including near-diploidy and higher number of large-scale state transitions (LSTs).

In the beginning, it was observed that 79% (19/24) of near-diploid tumors had BRCA1-inactivation either through germline mutation or epigenetic silencing. Subsequently, the authors profiled the frequency of copy number segment of different sizes and found that there existed two types of segments with the prominent cutoff at 3 Mb in segment size. They then defined a **state transition** of size S Mb being a chromosomal break between adjacent segments of at least S Mb. After filtering and smoothing small-scale segments (less than 3 Mb), the number of remaining large-scale state transition can split near-tetraploid BLCs into 2 stable groups, with the most significant difference observed when S=10 Mb. As a result, number of LSTs of size 10 Mb was suggested a good proxy for large-scale genome instability.

The final classifier relies on a two-step decision rule. Tumors are first segregated into near-diploid or near-tetraploid tumors and then classified into LST_High and LST_Low subgroups by ploidy-specific LST cutoffs (15 LSTs for near-diploid tumors and 20 LSTs for near-tetraploid tumors). This classifier achieved 100% sensitivity and 90% specificity (97% accuracy) for classifying BRCA1 or BRCA2 inactivated tumors.

3 OVARIAN CANCER

Ovarian neoplasms can be separated into three main types according to the probable tissue of origin. Of these, surface epithelial-stromal tumors account for 60%, followed by germ-cell tumors (30%) and sex cord stromal tumors (8%). On the other hand, based on the degree of atypia, tumors can be classified as benign, borderline (intermediate) or malignant (carcinoma). Approximately 75%-80% of the tumors are benign, whereas around 30% of the epithelial type are malignant. Therefore, the epithelial ovarian cancers (EOCs) account for the vast majority (80%-85%) of ovarian malignancies[31].

This chapter starts with an overview of ovarian cancer (OC). Section 3.1 introduces the global burden of the disease, different histological subtypes along with their occurrence and survival statistics, as well as factors associated with increasing or decreasing risk of developing the disease. The second and third sections focus on EOC, where Section 3.2 covers its heritability and genetic factors and Section 3.3 describes the current carcinogenesis model. Lastly, Chapter 4 put special emphasis on high-grade serous carcinoma (HGSC) and go through current understandings about this major subtype of EOC.

3.1 EPIDEMIOLOGY

3.1.1 *Incidence and mortality*

Ovarian cancer is the eighth most common malignancy worldwide for women in 2018, according to a WHO report on Global cancer statistics[32]. OC was newly diagnosed in 295,414 cases and caused 184,799 deaths, it also accounts for 3.3% of prevalent cancer cases in women within five years. Over the years, the incidence and mortality rate are declining in the U.S.[33, 34], potentially influenced by the change in of hormonal therapy prescriptions[35]; however, the temporal trend varies by country[36].

Incidence and mortality rate are also heterogeneous across the globe[34, 37] and some potential associating factors were proposed[38]. Such variations can be observed in different geographical regions as well as in different country development status. As summarized in Figure 3.1, OC were more frequent and caused more deaths in highly developed countries, or in European region. Specifically, the highest incidence rate was observed in the Central and Eastern Europe and affects 11.9 per 100,000 women per year, followed by in Northern Europe (9.2) and in North America (8.4), as compared to the worldwide rate of 6.6, which is translated to a 0.72% cumulative risk of developing OC before age 75 years.

In Europe, OC is the second most common and most lethal gynaecological malignancy, representing 5.2% of cancer-related morbidity. A woman's risk of developing OC before age 75 is 1 in 93, and her chance of dying of the disease before age 75 is 1 in

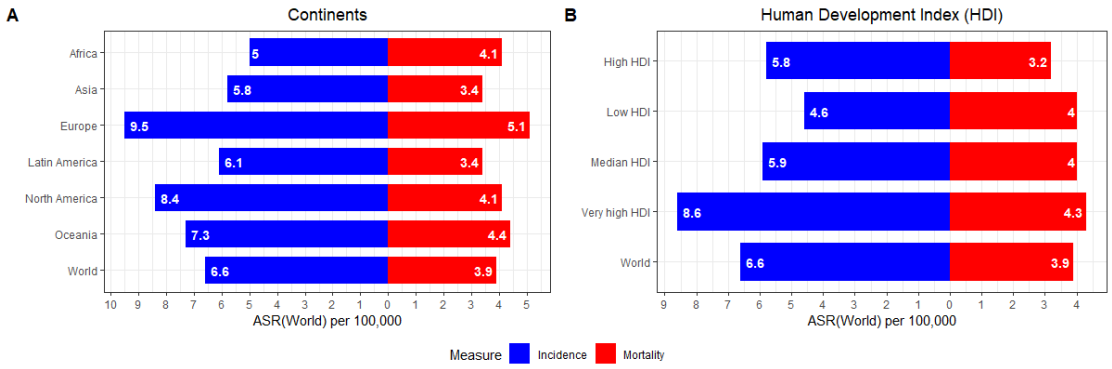


Figure 3.1: Incidence and mortality rate of ovarian cancer in 2018[32]. Age-standardised rate (ASR) stands for the number of persons affected per 100,000 women per year. The calculation was stratified by (A) geographical regions or by (B) four tiers of Human Development Index (HDI), with respective to the reference World population model.

166. However, a more precise estimation of the lifetime risk in Europe would require inclusion of a wider age range, as the life expectancy of women in Europe was 80.8 year in 2016[39].

3.1.2 *Histological classification*

Epithelial ovarian cancer can be delineated into five distinct histological subtypes, which comprise high-grade serous carcinoma (HGSC, 70%), endometrioid carcinoma (EC, 10%), clear-cell carcinoma (CCC, 5%-10%), low-grade serous carcinoma (LGSC, <5%) and mucinous carcinomas (MC, 3%)[40]. The classification is partially based on their morphologic resemblance to epitheliums of different origins. Among them, HGSC and LGSC have features resembling the epithelium of fallopian tube, and are distinguished by degree of nuclear atypia and mitosis rate[41]. Despite the communality of cell differentiation, they are different diseases. In fact, all of these subtypes vary in etiology, pathogenesis, molecular compositions, risk factors, clinical features, treatment response and prognosis. The fact that they are inherently different from each other makes ovarian cancer a group of heterogeneous malignancies instead of a single entity.

3.1.3 *Prognosis*

Ovarian cancer is more lethal than many other cancer types, as indicated by its relatively lower survival rate among other cancers[42]. Statistics from the U.S. registry (2006-2012) suggested a five-year relative survival rate of 46.2% for OC[43], meaning that the likelihood for OC patients to survive five years after diagnosis is 46.2% of that for the general population. In addition, the mortality-to-incidence ratio of OC is higher than that of all cancers combined (0.59 versus 0.55) and only behind Liver, Lung and Stomach when compared to the ten most common cancers[32].

Within OC, survival rate varies greatly based on histological subtype and stage. Except for these factors, individual's prognosis is also influenced by age, performance status, residual disease volume and BRCA status[44]. The heterogeneity in survival rate is illustrated in Figure 3.2, based on statistics of 28,118 EOC patients diagnosed between 2004 and 2014 in the U.S.[45]. In summary, the low survival rate is largely driven by late stage diagnoses, as over 60% of the OC patients were diagnosed at an advanced stage. Above all, the most common subtype HGSC accounts for more than 80% of advanced-stage EOCs and therefore represents most of the ovarian cancer mortality.

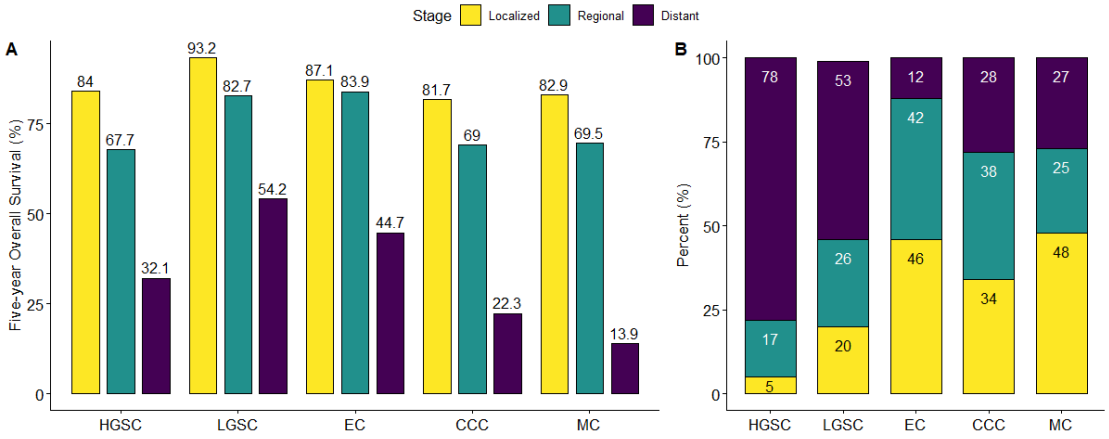


Figure 3.2: Subtype-specific statistics. (A) Overall survival stratified by subtype and disease stage. (B) Distribution of stage at diagnosis for different subtypes. In contrast to a **Localized** stage where malignancy is limited to the organ of origin, a **Regional** stage indicates that cancer has spread to nearby structures or lymph nodes, while a **Distant** stage indicates a spread to distant parts of the body, such as the liver or lungs.

Since the mid-1970s, cancer survival has increased for most common cancers in the U.S.[46] and similarly in Europe[47, 48]. Table 3.1 shows that, in contrast to the considerable survival improvements in other common cancers like prostate, colorectal and breast cancer, there has been only modest and below average increase for ovarian cancer survival over the past few decades. Therefore, the poor survival over time awaits more efforts on early-stage cancer detection and effective therapy.

3.1.4 *Risk factors and preventive factors*

Early detection of cancer would increase the chances for successful treatment. This can be made possible by organized screening programme and high-risk group identification. Despite the lack of effective screening strategy, established risk factors can help identify at-risk population so as to apply risk-reducing managements before the disease manifests.

The disease most prevalently presents in the sixth decade of life and affects predominantly perimenopausal and postmenopausal women[45]. In women with hereditary risks, the disease can occur 10 years earlier[49]. In principle, having a first-degree

Cancer Site	Survival change in the US (1975-2012)	5-year relative survival in the US (2006-2012), age-adjusted
prostate	67.8% to 99.3%	98.90%
colorectal	49.8% to 66.2%	65.10%
female breast	74.8% to 90.8%	89.70%
ovary	36% to 46.4%	46.20%
all sites	50.3% to 66.4%	66.90%

Table 3.1: Five-year relative survival and its trend in the U.S.[43, 46].

affected relative confers an increased risk of OC by three-fold[50]. The relative risk declines with both the age of the at-risk person[50] and the age of onset in their relatives[51]. On the other hand, the risk increases with the number of relatives affected[52].

In summary, older age, personal history of breast cancer, having a family history of ovarian and/or breast cancer are the main risk factors of developing OC. In line with the familial risk, some cancer predisposition syndromes carry with them increased risks of OC. These include *HBOC syndrome*, *Lynch syndrome*, *Peutz-Jegher syndrome* and some other rare syndromes[53].

In addition, hormone replacement therapy, nulliparity, and benign gynaecological conditions(endometriosis, polycystic ovarian syndrome and pelvic inflammatory disease) also lead to an elevated OC risk[54]. Other factors that confers modest risk include increased height, weight and BMI. Protective factors, on the other hand, include oral contraceptive use, increasing parity, lactation and tubal ligation[44]. Notably, different histologic types may have different risk factors[55].

3.2 HEREDITARY OVARIAN CANCER

3.2.1 *Clinical manifestation*

Hereditary ovarian cancer (HOC) is characterized by familial aggregation of cancer cases. There are three clinical manifestations[56]: (a) site-specific ovarian cancer, where only ovarian cancers are seen in excess; (b) *HBOC syndrome*, where both ovarian and breast cancers are observed; and (c) *Lynch syndrome type II*, where ovarian, colorectal or endometrial cancers aggregate in the family. Among them, the first two categories account for the majority of the hereditary cases[57]. In addition, category (a) is suggested to be a variant manifestation of category (b) as the only attributable genetic factor found was a HBOC susceptibility gene BRCA1[58, 59].

3.2.2 *Genetic factors*

HOCs account for 5% to 15% of ovarian cancers and are mainly identified by family history[60] and/or mutation carriers of high-risk susceptibility genes[61, 62]. Although family history has been a confirmed risk factor, family studies cannot exclude the contribution of nongenetic factors shared within families. Twin studies, instead, were able to distinguish heritable factor from environmental factor and suggested a significant heritability (39%) for ovarian cancer[63].

Since the first association between BRCA1 and HOCs identified in 1991[64], genetic predisposition to HOC had been, in the first two decades, attributed mainly to mutations in high-penetrance genes for *HBOC syndrome* (BRCA1, BRCA2) and *Lynch syndrome* (mainly MLH1, MSH2). Mutations in these genes are inherited in an autosomal dominant manner with incomplete penetrance regarding these rare disorders. For the past 10 years, more moderate or low penetrance genes for OC are getting appreciated. Table 3.2 listed genes conferring to an increased risk for OC and the strength of associations.

In summary, a reasonable estimate was made and suggested that the main attributable genetic factors are BRCA1 (55%) and BRCA2 (25%) in the context of the *HBOC syndrome*, followed by the *Lynch syndrome* genes (15%) and eventually additional moderate penetrance genes accounted for the rest 5%[79].

3.2.2.1 *BRCA1 and BRCA2*

It has been almost 30 years since BRCA1 and BRCA2 first shown to be linked to breast cancer[80, 81] and later on cloned[82, 83]. Both BRCA1 and BRCA2 have tumor suppressor function due to their long appreciated roles in DNA damage repair, and their loss-of-function may lead to genome instability and chromosomal rearrangements.

Following the detection of DNA damage, BRCA1 and BRCA2 are recruited to the damage sites where they share a common role of repairing DNA DSB via the HRR pathway. The process requires their coordinated interaction with other repair proteins and the nucleotide molecules using different protein domains. Localized in the nucleus, the main role of BRCA2 in HRR is to mediate the recruitment of RAD51 through a binding motif composed of BRC repeats. Other domains in BRCA2 allow its binding to PALB2 as well as single-strand DNA. On the contrary, BRCA1 has a broader role upstream to BRCA2. It is involved in three protein supercomplexes and its association with diverse binding partners enables a multi-functional role in mediating cell cycle checkpoint activation and acting in HRR, NHEJ[84], as well as in other repair pathways.

In mouse models, homozygous knockout of either gene is embryonically lethal[85, 86]. In human, germline biallelic inactivation has been observed only in BRCA2, which results in a subgroup of Fanconi anemia[87]. The inherited aberrations in heterozygous state predisposes carriers to a broad spectrum of diseases, including breast cancer, ovarian cancer, pancreatic cancer, stomach cancer, prostate cancer and, to less extent, some other cancer types[88, 89].

Germline pathogenic variants reported so far are mostly small indels or mutations that lead to protein truncation[57], e.g. frameshift indels or nonsense mutations. Their loci are dispersed across the gene bodies and hotspot mutations are uncommon. Apart

Gene	Penetrance	Frequency in unselected EOC cases (%)	Frequency in control / general population (%)	Relative risk (95% CI)	Lifetime risk (%)
BRCA1	high	9.5[62]	0.15[65]		40-41[66, 67]
BRCA2	high	5.1[62]	0.26[65]		15-18[66, 67]
RAD51C	moderate	0.41-1.1[68, 69]	0.13[70]	5.88(2.91-11.88)[69]	5.2; 8.4(80yrs)-9(80yrs)[68, 69]
RAD51D	moderate	0.35-0.6[68, 71]	0.03[70]	6.30(2.86-13.85)[71]	10(80yrs)-18.3(80yrs);12[68, 71]
BRIP1	moderate	0.92-1.36[62, 72]	0.09-0.17[70, 72]	3.41(2.12-5.54)[72]	5.8(80yrs)[72]
PALB2	moderate	0.28-0.63[62, 72]	0.09-0.13[70, 72]	2.91(1.4-6.04)[73]	3%;5%(80yrs)[73]
MSH2	moderate	0.11[74]	0.03[70]	7.97(1.1-56.6)[75]	10.4-24[76, 77]
MLH1	moderate	0.11[74]	0.05[70]	6.35(0.89-45.1)[75]	3.4-20[76, 77]
MSH6	moderate	0.26-CF(0.87)[74, 78]	0.13-CF(0.21)[70, 78]	OR=4.16 (1.95-9.47)[78]	1[77]
ATM	moderate	0.57-CF(0.98)[62, 78]	CF(0.35)-0.38[70, 78]	OR=2.85 (1.30-6.32)[78]	

Table 3.2: Ovarian Cancer Susceptibility Genes. Carrier frequency (CF) is estimated with $1.5 \times allele frequency$. Odds ratio (OR) is sometimes calculated instead of relative risk. Lifetime risk is by age 70 years unless specified.

from small-scale mutations, germline large rearrangements were also observed whereas with a lower frequency[90].

BRCA1- or BRCA2-associated ovarian cancers are in general similar to sporadic forms of the diseases, while with some specific features observed. First of all, they tend to present as high grade tumors and of serous subtype, although endometrioid and clear cell carcinoma were also observed[91, 92]. On the other hand, they are unlikely to be borderline or mucinous tumors[92]. Secondly, mutation carriers tend to be diagnosed at a younger age compared with sporadic cases, especially when mutations happened in BRCA1[91, 93]. Lastly, survival advantage was shown in carriers over noncarriers, including higher response rate to primary therapy, longer recurrence-free interval and longer overall survival[91, 93, 94].

3.2.3 *Recommendations for testing and management*

In HGSC, germline testing for BRCA1 and BRCA2 was reported to have an altogether 23% detection rate[61]. Given the high diagnostic yield, it was recommended that women affected with high-grade EOCs should be offered genetic testing and receive genetic counseling[95]. High-risk individuals are offered chemoprophylaxis and prophylactic surgeries like risk-reducing salpingo-oophorectomy[96] to reduce their ovarian cancer risk.

3.3 CARCINOGENESIS MODEL

A typical genetic model for tumor progression describes the genetic and epigenetic changes from normal tissue to benign neoplasm and eventually to carcinoma. Unlike the Fearon-Vogelstein model[97] proposed in 1990 for colorectal carcinogenesis, the carcinogenesis model for ovarian cancer was proposed late and is not yet fully clear for the time being.

In a study investigating genetic alterations in 108 sporadic serous ovarian neoplasms, Singer *et al.* [98] observed that KRAS mutations exist in around 50% of borderline tumors, non-invasive lesions and a variant of invasive carcinomas of the serous type, and that an increase in the degree of chromosomal allelic imbalance exist when comparing borderline tumors to noninvasive and invasive carcinomas. Such morphological continuum did not extend to conventional invasive serous carcinomas (mostly high-grade tumors) which are usually with wild-type KRAS and high allelic imbalance even in early-stage primary tumors. These observations suggested that serous borderline tumors and conventional serous carcinomas are unrelated and they therefore propose a preliminary concept of dualistic pathogenesis model for serous ovarian carcinomas.

Later in 2004, the authors consolidated contemporary evidences and proposed a dualistic model[99] to describe the tumorigenesis of epithelial ovarian tumors. In this model, Shih & Kurman proposed two tumor progression pathways based on morphological observations and molecular features.

Type I tumors follow a stepwise development and arise from borderline tumors. They are usually low-grade and cover a wide range of subtypes including LGSC,

MC, EC, CCC and malignant Brenner tumors. Frequent molecular changes observed in these tumors are BRAF and KRAS mutations for serous tumors, CTNNB1, PTEN mutations and MSI for endometrioid tumors and KRAS mutations for mucinous tumors. Parenthetically, this above-mentioned invasive carcinomas variant, initially named micropapillary serous carcinoma (MPSC) in 2002, was then considered synonymous with LGSC.

On the other hand, Type II tumors are those without identifiable precursor lesions and followed a seemingly de novo development. These tumors are high-grade neoplasms and include HGSC, malignant mixed mesodermal tumor (MMMT), and undifferentiated carcinoma. Frequent molecular changes observed in Type II tumors are limited, except for frequent TP53 mutations in HGSC and MMMT.

After more than a decade, this dualistic model of EOC carcinogenesis was revised in 2016[100]. Besides further subgrouping of Type I and Type II tumors based on their histotypes or molecular subtypes, the major revision lied in the change in Type II tumor pathogenesis. Originally suspected a de novo development, accumulating evidences suggest that Type II carcinomas develop from intraepithelial carcinomas in the fallopian tube and involve the ovary later on. In addition, more molecular findings were added to delineate the genetic composition of these two types. The revised carcinogenesis of type II carcinomas, especially the HGSC subtype, will be discussed in more details in Chapter 4.

4 HIGH-GRADE SEROUS CARCINOMA (HGSC)

Epithelial ovarian cancer (EOC) is a collective term for five major histotypes which are gradually recognized as different diseases[40]. This section focuses on the major histotype, the HGSC, and covers the current understanding about its clinical features, pathogenesis, molecular landscape, clinically relevant biomarkers and patient stratification.

4.1 CLINICAL FEATURES AND THERAPY

HGSC accounts for 70% of EOCs and caused most deaths since they are usually presented as a late stage and disseminated disease. Late stage presentation of HGSC has a 5-year relative survival rate of 32.1%, by contrast with 84% for early-stage disease. The difficulty in early-stage cancer detection lies in the lack of effective screening strategy as well as the lack of early and specific symptoms of the disease[101].

Despite the fact that 80% of patients seems to have favorable response to the initial platinum-based treatment, the majority of them suffered from relapse and developed resistance[102], and eventually succumb to their disease. This scenario has not substantially changed since platinum-based therapy was introduced in the late 1970s and became the standard of care for all OCs[103].

In 2014, Poly(ADP-Ribose) Polymerase (PARP) inhibitors was approved as the first histotype-specific treatment for molecularly stratified patients. Clinical trials demonstrated that this targeted therapy brought significant improvement in survival outcome of patients with BRCA-deficient tumors. Later on its efficacy was shown to extended to patients with HRD phenotype or showing chemosensitivity[104–106].

4.2 PATHOGENESIS

4.2.1 *Origination*

Ovarian carcinoma was long believed to have a mesothelial origin and arise from ovarian surface epithelium (OSE). Fathalla proposed an incessant ovulation hypothesis in 1971 and postulated that the development from OSE to epithelial neoplasms is influenced by the repetitive ovulation cycle[107]. The hypothesis posits that during cyclic rupture and repair trauma, aberration accumulated in the OSE or its cortical inclusion cysts. The accrual of DNA damage and constant exposure to follicular fluid lead to their Mullerian metaplasia from which they acquire differentiation resembling EOC subtypes and also lead to their neoplastic transformation. This hypothesis conforms with epidemiologic evidence that the number of lifetime ovulations is positively correlated with ovarian cancer risk. However, studies in the subsequent three decades

failed to find a convincing precursor of HGSCs in the OSE. These tumors were then hypothesized in the dualistic model[99] to originate from de novo development.

The discovery of (pre)neoplastic lesions in the fallopian tube had lead to a paradigm shift from mesothelial origin to tubal origin of HGSCs[100, 108]. First reported in 2001, Piek *et al.* examined fallopian tubes removed by prophylactic surgery from high-risk women and found frequent hyperplastic or dysplastic lesions in overtly normal fallopian tube epithelium (FTE). In one case carrying germline BRCA1 mutation, the dysplastic cells already exhibit p53 accumulation and lost of wild-type allele of BRCA1 gene[109]. Meanwhile, other studies performed pathologic assessment on prophylactic surgical tissues from BRCA mutation carriers and found in around 2.5% of the cases the existence of occult carcinomas in the ovaries and/or fallopian tubes[110–113]. Among them, some appear to originate in the fallopian tube. Based on these observations Piek *et al.* proposed a hypothesis in 2003 that most (hereditary) HGSCs originate from fallopian tube instead of from OSE[114].

4.2.2 *Tumor progression model*

These occult, microscopic tubal intraepithelial carcinoma (TIC) often located at the fimbriated end and are stained positive for Ki-67 and p53[115, 116]. With complete examinations on endosalpinx involvement in pelvic carcinomas, Dr. Crum's group found that TICs occurred in approximately half of ovarian carcinomas and are concomitant with pelvic carcinomas of other origins as well[117]. Mutational analysis further showed that TICs contain identical TP53 mutation with concurrent ovarian carcinomas[117–119], showing a clonal relationship between TICs and HGSCs. These observations suggest that a primary tumor arise in the fallopian tube and spread to the ovary in later stage.

Furthermore, Dr. Crum and colleagues also observed in non-neoplastic mucosa of fallopian tubes some cells with strong p53 immunostaining, designated "p53 signature", that might serve as precursor of HGSCs[118]. Their occurrence predominantly located in the fimbriae but not in cortical inclusion cysts[120]. They are featured with secretory cell type, often show evidence of DNA damage and frequently with TP53 mutations. These cells present more frequently in fallopian tubes with TICs; however the occurrence frequency in BRCA mutation carriers and normal-risk women are not significantly differnet[118, 120, 121].

To note, p53 signature only represents cells with p53 overexpression but not those with p53 null-type expression. Therefore, the precursor definition was expanded to early serous proliferations (ESPs) representing aberrant p53-expressing cells with different levels of atypia, spanning a spectrum from morphologically normal cells to proliferative lesions. Recently ESPs were shown to serve as an alternative precursor to HGSCs. In a study of 32 HGSCs cases without STIC, ESPs can be found in 40% of the cases and 75% of ESPs have clonal relationship with concurrent HGSCs[122]. With these findings Dr. Crum and colleagues proposed a precursor escape hypothesis as an alternative to explain the onset of HGSCs without co-existing STICs[123].

Not entirely excluding a mesothelial origin, the theory of tubal origin has received a widespread acceptance[124] and suggests that many of the HGSCs arise from distal

fimbriated end of the fallopian tubes. Multiple precursor types in the fimbria, including STICs and ESPs, can lead to HGSCs in the ovary. Although the exact carcinogenic sequence of how these precursors evolve remains to be further elucidated, compelling evidences collectively suggest a step-wise development of HGSC where p53 signature being the earliest lesion and developed into ESP or STIC, which ultimately transformed into pelvic HGSC[122].

4.3 MOLECULAR COMPOSITIONS

The different histotypes of EOC have distinct molecular characteristics. In low-grade serous carcinoma (LGSC), activating mutations of the RAS-MAPK pathway were frequently observed, such as mutations in KRAS (19%-55%) and BRAF (0%-33%)[125]. In the mucinous carcinoma (MC), mutation frequency of KRAS was 40%-50% and HER2 amplification was found in ~19% of the patients[126]. For tumors classified as low-grade endometrioid carcinoma (EC), mutations were frequently observed in CTNNB1 (53%), PIK3CA (40%), KRAS (33%), ARID1A (30%), PTEN (17%), PPP2R1A (17%) and PTEN (17%) genes[127]. While in clear-cell carcinoma (CCC), oncogenic mutations are found in ARID1A (50%), PIK3CA (39%), PP2R1A (15%), KRAS (14%) and PTEN (5%) genes[128]. Unlike these other subtypes that are more associated with recurrent mutations and have relatively lower SCNA burden, HGSCs are characterized by few driver mutations and widespread copy number alterations[129].

There has been extensive interest characterizing the molecular landscape of HGSC and exploring the association between molecular events and clinically relevant questions, such as diagnosis, prognosis and treatment resistance. In this section the molecular findings in major large-scale studies are reviewed.

4.3.1 *Landmark genomic studies*

4.3.1.1 *TCGA-OV*

TCGA consortium was the first to conduct a large-scale study aiming at revealing pathophysiology and clinically relevant abnormalities in HGSC genomes. In the TCGA-OV cohort published in 2011[130], 489 pre-treatment samples were comprehensively profiled on five omic platforms, including whole exome, whole transcriptome, miRNA, methylation and genome-wide copy number, in order to uncover genetic regulation at different levels.

They reported ubiquitous TP53 somatic mutations in 96% of the tumors and additional eight recurrently mutated genes at frequencies around 2%-6%, including BRCA1, BRCA2, NF1, RB1, CDK12, FAT3, CSMD3 and GABRA6. BRCA1/2 aberrations can result from germline or somatic genetic changes or epigenetic mechanism, collectively they affect 30% of the cohort.

On the contrary, there were 113 recurrent focal gains or losses across the cohort. The most common and highly amplified regions are those containing MYC, MECOM and CCNE1, which were found in 34%, 27% and 23% of the cohort, separately. Notable

focal deletions were found in regions where PTEN, RB1 and NF1 are located, while each with a frequency less than 15%.

Molecular subtypes were uncovered based on either mRNA, miRNA or methylation profiles. Besides, the authors also derived transcriptional signature predictive of survival and further identified BRCA1/2 and CCNE1 as prognostic biomarkers. In terms of pathway activation, it was proposed that HRD exhibits in ~50% of the cohort and that NOTCH and FOXM1 signalling being involved in pathophysiology of HGSC.

4.3.1.2 *ICGC-AU-OV*

As part of the ICGC consortium, the Australian group led by Dr. David Bowtell published the AOCS cohort in 2015[131]. This is the first large-scale study to characterize HGSC using whole genome sequencing (WGS) platform, as well as to include post-treatment samples to gain insight into acquired resistance mechanisms.

The cohort comprised 114 tumors from 92 patients. In line with observations in TCGA, there were prevalent TP53 mutations and additional driver genes at frequencies between 3%-6%. Moreover, researchers found that inactivation frequencies of some tumor suppressors can be increased when other regulation mechanisms were considered. For example, inactivation through gene breakage events was observed in RB1 (17.5%), NF1 (20%) as well as RAD51B and PTEN.

The authors further tested for events associated with treatment response by dividing patients into three response groups, namely resistant, refractory and chemosensitive groups. Consistent with previous findings, they showed that cases with germline or somatic BRCA1/2 mutations had favourable response, and that CCNE1 amplification is common in primary resistant and refractory diseases.

In light of this, they were able to stratify patients into three groups based on the molecular events. HRR-deficient patients, defined as those harboring HRD-related gene aberrations, had better overall survival and account for half of the cohort. On the other hand, patients with CCNE1 amplification, or patients with neither events, had similarly worse overall survival.

Lastly, this study shed light on different mechanisms leading to acquired resistance. When comparing primary and resistant samples from the same patient, they observed five molecular events associated with resistance. These include reversion of germline BRCA1 or BRCA2 mutations, loss of BRCA1 promoter methylation, desmoplasia leading to alteration in molecular subtype, and recurrent promoter fusion of the drug efflux pump MDR1 with expression up-regulation.

4.3.1.3 *OV133*

There had not been a systematic overview of histotype-specific genomic landscape until Dr. Sohrab Shah's group published the OV133 cohort (see Section 6.1.4) in 2017[132], where they showed that stratifying patients by genomic features may recapitulate major histotypes. The 20 genomic features include mutational signatures and quantitative measures of genetic alterations, with more details described in Section 4.5.2 and Section 8.5.2.

The ability of these features to surrogate diverse DNA repair deficiencies potentiate this stratification model to reveal the etiology underneath each subgroup. In EC, MMR signature showed that MSI exhibited in 28% of the cases. In CCOC, 26% of the cases were grouped according to APOBEC deamination signature, and 40% to age-related signature. Lastly, HGSC was divided into two subgroups - H-HRD and H-FBI, where the former being featured with HRD signature and the latter being enriched in foldback inversions signifying breakage-fusion-break process.

4.3.2 *DNA damage repair defect*

There is a long research history characterizing ovarian cancers with inherited HRR defects, mainly due to the well appreciated ovarian cancer susceptibility genes BRCA1 and BRCA2. Moreover, other rare inherited mutations in DDR genes were found with moderate penetrance. Their prevalence, penetrance and accountability with respect to ovarian cancer are discussed in Table 3.2.

In sporadic OCs, disruption in BRCA1 and BRCA2 can arise from other mechanisms like somatic aberration and promoter hypermethylation[133]. Overall, around 30%-40% of samples had BRCA1/2 loss from either germline or somatic mutation or methylation events[130, 131]. It is noteworthy that these different mechanisms occurred in a mutually exclusive manner among patients. Furthermore, disruption in other DDR genes were also observed in somatic setting.

In the TCGA study, researchers further looked at somatic alterations potentially inactivating the HRR axis, including EMSY amplification, PTEN focal deletion, RAD51C hypermethylation, as well as mutations in EMSY, PTEN, ATM, ATR and Fanconi anemia genes. Overall, researchers found approximately 50% of the cohort exhibiting genomic alterations that might lead to HRD.

In the ICGC study, the group also found 51% of the cohort harboring HRD. These include BRCA1/2 inactivation via the three abovementioned mechanisms, PTEN deletion, somatic mutations in RAD51C as well as germline truncating variants in BRIP1, CHEK2 and RAD51C.

Taken together, around half of the HGSC tumors are featured with defects in HRR system. Among these, BRCA1/2 inactivation account for 63% and 43%, respectively, in TCGA and ICGC study. In fact, other subtypes of EOC were reported to be associated with distinct DDR defects[132].

4.3.2.1 *TP53 Pathway*

The tumor suppressor gene TP53 is the most frequently altered gene in human cancers. It encodes p53 protein, a multifunctional transcription factor that, in response to diverse cellular stresses, controls cell cycle progression, induces apoptosis, senescence, DNA repair or diverts metabolism. It was first noticed in 1991 that overexpression of p53 protein exist in half of EOCs[134]. On the other hand, null mutations in TP53 leading to complete absence of p53 protein were also found abundant in OCs[135]. These two immunophenotype robustly reflect the existence of TP53 mutations[136].

Using sequencing approach, researchers were able to comprehensively address its prevalence and concluded that an invariable presence of p53 inactivation exists in HGSCs[137], either in the form of TP53 mutation or copy number gains in MDM2 or MDM4. In a follow-up study of TCGA, pathologists revisited the 4% of the cohort originally reported as lack of TP53 mutations and found that these cases had likely been misdiagnosed, thereby confirming TP53 mutations being present in virtually all HGSC cases in TCGA cohort[138].

4.3.2.2 *RB1 Pathway*

RB1 codes for Rb, the first described tumor suppressor protein, and its allele loss in OC was noticed in 1991[139]. Although RB1 itself is not a frequent target of mutation events in OC, Rb pathway inactivation was found to be a frequent event. This results in, among others, loss of cell cycle control at G1/S transition. Abnormalities in the Rb pathway include abnormal protein expression of p16, CDK4, cyclin D1 and phosphorylated Rb, altogether they were observed in 60.9% of EOCs[140]. Similarly, in the TCGA study this pathway was altered in 67% of the cases and mostly due to copy number change or transcriptional deregulation. Specifically, these include mRNA expression down-regulated in CDKN2A (30%) and up-regulated in CCND2 (15%), amplification of CCNE1 (20%) and CCND1 (4%), as well as deletion (8%) or mutation (3%) in RB1[130].

4.3.2.3 *RAS-PI3K Pathway*

PI3K pathway integrates upstream signals from growth factors or tyrosine kinase receptors and control survival or metabolic processes through different downstream signaling. Depend on its target, AKT can regulate apoptosis through p53, promote cell cycle progression through CCND1, or influence angiogenesis through mTOR, to name but a few[141]. In the TCGA study, aberrations in RAS/PI3K pathway members collectively affect 45% of the cohort and are predominantly through copy number changes, namely amplification of PIK3CA(17%), AKT1(3%), AKT2(6%) and KRAS(11%), as well as deletion of PTEN(7%) and NF1(8%). Mutations in these genes were also observed while with a frequency less than 1%, except for NF1 (4%)[130].

4.4 GENOMIC FOOTPRINTS AND BIOMARKERS

There has been extensive interest in identifying prognostic biomarkers in ovarian cancer[142]. The scope ranged from local changes in specific genes at the DNA, epigenetic, mRNA, and immunohistochemistry-based protein expression level, to global changes like genomic footprints and transcriptomic signatures. Some of these associations had been reproduced in different cohorts, such as local changes in BRCA1, BRCA2 and CCNE1 genes, or global changes in genomic scars (see Chapter 2) and transcriptomic molecular subtypes[143]. Nonetheless, only very limited biomarkers are readily utilized in the clinic. Beyond these, with a better resolution of DNA changes provided by WGS technique, scientists set out to explore more genomic footprints and discover more

genomic phenotypes. This section gives an overview of known DNA-based biomarkers, as well as novel genomic phenotypes observed in ovarian cancer.

4.4.1 *Known biomarkers*

4.4.1.1 *BRCA1 and BRCA2*

Inactivation of BRCA1 or BRCA2 has been a long established biomarker for better prognosis. Their loss of function lead to DNA repair defect and predict sensitivity to drugs exploiting cellular toxicity of DNA ICL, such as platinum-based chemotherapy.

The survival advantage for germline BRCA genes mutation carriers (see Section 3.2.2.1) was noted as early as in 1996[94]. Later studies confirmed this observation and further found that BRCA2 carrier status implicates even better outcome[144–146]. Furthermore, TCGA study showed that the survival depends on inactivation mechanisms, where cases with germline or somatic mutations in BRCA1/2 had better overall survival as compared to BRCA1/2 wild-type cases; however, cases with epigenetically silenced BRCA1 had survival similar to wild-type tumors[130].

In general, BRCA1/2 mutation status has been used for identifying high-risk mutation carriers, predicting better prognosis. In addition, it also serves as a predictive biomarker for better response to targeted therapy using PARP inhibitors[147].

4.4.1.2 *CCNE1*

The prognostic implications in CCNE1 was first identified in a cohort of 139 EOC patients treated on GOG Protocol 111[148]. The study showed that cyclin E, encoded by CCNE1, was an independent poor prognostic marker for overall survival and associated with CCNE1 amplification.

Later studies confirmed that CCNE1 amplification is consistently associated with primary chemoresistance[131, 149], refractory disease[131], disease-free survival[150] as well as overall survival[130, 150]. However, the prognostic effect of cyclin E expression was not always observed[150–152].

In a recent study, the Bowtell laboratory compared high cyclin E expression samples with and without CCNE1 amplification, and found that they have different pathological and biological characteristics[153]. They also found that the latter cases had better outcomes than the former ones, which possibly explains the discrepancy observed in previous studies. In summary, although not yet being used in clinical trials, CCNE1 amplification has been robustly shown a biomarker for worse prognosis.

4.4.1.3 *Homologous Recombination Deficiency (HRD)*

Based on the concept of synthetic lethality, the key determinant of sensitivity to PARP inhibitor treatment is HRD. The majority of PARP inhibitor clinical trials in ovarian cancer select patients based on the presence of germline BRCA mutations and/or prior platinum-sensitivity. Evidence from these trials suggests that some patients without

germline BRCA mutations can also benefit from the treatment[104–106] and that the therapeutic effect is more widely applicable to HRD phenotype.

Given HRD being the pivotal therapeutic target for PARP inhibitor, different surrogate markers had been proposed for this phenotype. They were developed on the basis of somatic mutations, genomic scars, transcriptional profiles, protein expression or functional assays[154]. Among them, genomic scar assays providing measures of genome instability are already commercially available, including "myChoice HRD" from Myriad Genetics and "FoundationFocus CDxBRCA LOH assay" from Foundation Medicine. The former gives a **HRD score** based on the combination of TAI, LOH and LST (see Section 2.5) and classifies tumors with score >= 42 as HRD-positive. The latter is based on the idea that fraction of genome with LOH (FLOH) is associated with chemoresistance[155] and the assay uses a genomic fraction of >= 16% as a threshold for HRD-positivity. Both assays are in use in clinical trials[105, 156] as companion diagnostic tests.

4.4.1.4 *Emerging prognostic biomarkers*

In the OV133 cohort (see Section 4.3.1.3), Wang *et al.* showed that there was prognosis difference between the two genomic subgroups of HGSC - H-FBI and H-HRD (see Section 4.5.2). They further found that foldback inversion (FBI) events are enriched in H-FBI tumors. More importantly, FBIs accompanied with high-level amplifications, abbreviated to FBI-HLAMPs, showed prognostic value transcending both BRCA1/BRCA2 mutation and known transcriptomic molecular subtypes. It serves as a biomarker for worse prognosis in the OV133 cohort, and also validated in TCGA-OV and ICGC-AU-OV cohort.

4.4.2 *Tandem Duplicator Phenotype (TDP)*

After the phenomena of excess duplication events observed in breast cancer in 2009 [23], a similar mutator phenotype received attention in HGSC[102]. Ng *et al.* from Dr. James Brenton's group performed WGS on four cell lines derived from two HGSC cases, one with and one without HRD phenotype, and observed distinct patterns of structural variants between them. The HRD case, harboring a germline BRCA2 mutation, showed small deletions (~ 12kb) and interchromosomal translocations. On the other hand, the non-HRD case, with confirmed MMR competency, had excess TDs featured in insertions (~350kb) frequently associated with copy-number gain. Shortly afterwards, this phenotype was confirmed in another WGS study. With a total of 8 HGSC cases, Dr. David Bowtell's group observed prevalent small deletions (~3.2kb) in 3 germline BRCA1/2 mutated cases, as well as excess TDs (~410kb) in 4 other cases[157].

Based on SNP array-derived copy number profile, Ng *et al.* further defined a TD-like feature that enabled them to assess the abundance of TDs from SNP array. It was estimated that 12.8% of the TCGA-OV cohort showed **Tandem Duplicator Phenotype** (TDP) and that the TDP subgroup is mutually exclusive with the subgroup harboring BRCA1/2 somatic mutations.

4.4.3 *TD-plus phenotype*

In 2016, Popova *et al.*[158] found that the highly altered genomes in TCGA-OV cohort are best characterized by interstitial gains of 2-7 Mb in size, and that this pattern is associated with CDK12 inactivation. They defined it as **CDK12 TD-plus phenotype** and reported the frequencies of its occurrence to be 3%-4% in TCGA-OV cohort, 1.7% in TCGA prostate cancer cohort and 4.2% in in-house OC cohort. Additional WGS data revealed overwhelming TDs being the source of the frequent mega-sized gains feature.

The authors then hypothesized that frequent TDs are associated with two different tandem duplicator phenotypes: one with TDs smaller than 1 Mb in size resembling the TD-like feature as previously described[102], and the other being CDK12 TD-plus phenotype with TDs of up to 10 Mb.

Although inactivation of CDK12 was reported to be associated with HRD[159] and render cells with hypersensitivity to DNA-damaging agents and to PARP inhibitors[160], the authors did not observe an elevated HRD score in these CDK12-inactivated cases, nor did they find favorable overall survival among them. This phenotype is mutually exclusive with germline/somatic BRCA1/2 mutation and with BRCA1 promoter methylation.

In summary, the CDK12 TD-plus phenotype serves as a unique genomic footprint of CDK12-inactivated tumors, with the biologial mechanism responsible for formation of these TDs remained unclear.

4.5 PATIENT STRATIFICATION

Cancer is a heterogeneous disease and there exhibits inter-patient heterogeneity. Grouping patients into more homogeneous subpopulations can provide more personalized outcome prediction and treatment planning. Ovarian cancer is recognized as a nonspecific term for histologically distinct diseases involving the ovary[40, 103]. In the major histotype HGSC, an active area of research is the molecular subtype discovery based on omics data, and in particular with transcriptome. TCGA-OV study comprehensively uncovered the subgroup structure of HGSC and derived de novo subtypes from different layers of molecular data. These include four mRNA subtypes, three miRNA subtypes and four methylation subtypes. The mRNA subtypes largely overlap with previously proposed transcriptomic molecular subtypes[143]; however association with clinical outcome was not reproduced[130].

For expression subtypes, the main limitations for their clinical utility have been the needs for a classification system robustly yielding clinically relevant subtypes and a way of unambiguously assigning single patients to subtype[161]. Recent studies partially addressed these needs by deploying expression-based subtypes with clinical grade classification assays[162, 163]. Therefore, they are getting used by targeted clinical trials for molecular subtype stratification. Nonetheless, some of the expression subtypes can be influenced by the microenvironment composition[164, 165], and individual specimen can express multiple subtype signatures[166, 167]. For the same patient, established molecular subtype can change spatially between anatomic locations[162, 165] and also

temporally along the treatment course[168]. Therefore, how transcriptomic subtypes can inform treatment decision still awaits more characterization of this dynamic behavior.

DNA-based stratification were initially based on biomarkers like CCNE1 or BRCA genes, as well as HRR gene inactivation. Nonetheless, the major studies evaluated HRR integrity using different procedures[130, 131] and there has not been a consensus approach in terms of which genes to include and which alteration mechanisms to look at. As WGS technique provides comprehensive information about DNA alterations, more DNA-based subtyping had been proposed in recent years. Two prominent DNA-based classification are introduced in this section, where TDP subgroup[169] possibly informs biological mechanisms and Shah-2017 subgroup[132] is with potential prognostic relevance.

4.5.1 *TDP subgroup classification*

Initially, tumors with TDP are identified mainly based on the frequency of TD-like events observed in SNP array-based copy-number profile. Dr. Edison Liu's group was the first to characterize TDPs in WGS studies at large scale.

Menghi *et al.*[170] developed a TDP scoring metric to systematically quantify this genomic configuration in a cohort of 277 WGS samples representing 11 cancer types. Based on the analyses across two platforms (both SNP array and WGS), they found that TDP tumors were enriched in TNBC, OV, UCEC and HCC. The development of TDP may require TP53 loss-of-function mutation, reduction of BRCA1 activity, and overexpression of DNA replication and cell cycle genes. They showed that TDP score served as a feature of BRCA1 loss, since it negatively correlates with BRCA1 expression in BRCA and OV. More importantly, TDP score was directly associated with enhanced sensitivity to cisplatin in TNBC both in vitro and in vivo. It was then hypothesized as a predictive marker of platinum-based drug sensitivity independent of tumor type.

As accumulating evidences suggest that rearrangements of different sizes result from different mechanisms, Menghi *et al.*[169] later on extended their methodology and looked into this qualitative feature of TDs. Specifically, to estimate major peaks within sample-wise TD span size distribution. In short, subsequent to separating tumors by TDP score as done previously[170], the new methodology classified TDP tumors into TDP subgroups based on the their major peak compositions.

Among the 2,720 WGS samples collected from TCGA and additional 30 studies, the authors identified 13.8% of the pan-cancer cohort showing TDP. Furthermore, 95% of the major peaks found across the cohort fall within 3 predefined size intervals. These different size classes of TDs correspond to distinct mechanisms of DNA instability, described as follow. TDPs with class 1 TDs are featured with TP53 loss-of-function events and BRCA1 deficiency and is comprised of tumors with very small TDs (~10kb). TDPs with class 2 TDs have the characteristic of CCNE1 pathway activation and correspond to intermediate (~200kb) size group. TDPs with class 3 TDs, mainly enriched in CDK12 disruption, are distinguished by their larger span size (~2Mb). Noteworthily, a tumor may bear features from more than one size groups when multiple major peaks were identified. The three most common combination of mixed groups are group 1/2

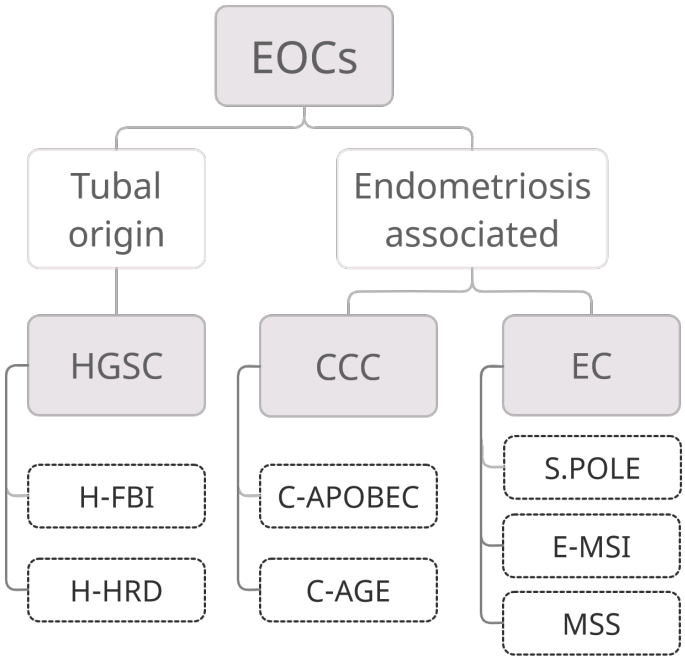


Figure 4.1: Genomic stratification recapitulate EOC histotypes. Items with grey background represent EOC and its major subtypes, while items with dashed border represent genomic subtypes. This diagram is reproduced from Figure 5 of the original publication[132].

mix, group 2/3 mix and group 1/3 mix. This method for classifying TDP provides therapeutic implications by reflecting the plausible underlying mechanisms and was recently patented.

4.5.2 *Shah-2017*

In 2017, Dr. Shah's group proposed a model capable of stratifying EOC histotypes based on genomic features[132]. As summarized in Figure 4.1, the model further divides major histotypes into seven subgroups based on different etiology described in Section 4.3.1.3.

The authors started with extracting genomic features from WGS data. The selected features are reflective of different biological processes and consist of six mutational signatures from COSMICv2 SBS Signatures, fraction of four small variant types, fraction of seven rearrangement types, three features summarizing copy number changes, and one feature reflecting degree of breakpoint homology in rearrangements. The patient stratification is then based on the unsupervised clustering of these 20 genomic features, which is described in detail in Table 4.1.

Category	Feature	Description
Signature	S.APOBEC	Signature 13 - APOBEC
Signature	S.POLE	Signature 10 - POLE
Signature	S.AGE	Signature 1 - AGE
Signature	S.BC	Signature 8
Signature	S.MMR	Signature 6 - MMR
Signature	S.HRD	Signature 3 - HRD
Variant	Nonsynonymous	Proportion of non-synonymous coding mutations
Variant	Splicesite	Proportion of splice site mutations
Variant	Stop.Lost/Gained	Proportion of stop lost or stop gained mutations
Variant	Frameshift	Proportion of frameshifting indels
SV	Foldback.Inversion	Proportion of foldback inversions
SV	Inversion	Proportion of inversions
SV	Tandem.Duplication	Proportion of tandem duplications
SV	Deletion Rearrangement	Proportion of deletions
SV	Balanced Rearrangement	Proportion of balanced rearrangements
SV	Unbalanced Rearrangement	Proportion of unbalanced rearrangements
SV Property	Homology>=5bp	Proportion of rearrangements with microhomology of >= 5bp
CNA	CN.Amplification	Proportion of genome showing copy number high-level amplification (copy number > ploidy + 2)
CNA	CN.Loss	Proportion of genome showing copy number loss (homozygous deletion or deletion LOH)
CNA	CN.LOH	Proportion of genome harbouring dominant LOH events

Table 4.1: Description of genomic features used for patient stratification. This table is reproduced from Figure S3 of the original publication[132].

5 AIMS

Ovarian cancer is usually diagnosed at late stage, together with little changes in the standard of care, survival rate of this disease had not improved much for few decades[103]. Understanding the biology of the disease is the key to its prevention, early detection, diagnosis and effective targeted treatment. To unlock the mystery of the disease and to identify biomarkers for treatment response, large-scale genomic studies were conducted on the major and most lethal histotype HGSC. The first aim of the thesis is to provide **an overview of ovarian cancer at the molecular level**, which can help to identify key pathways in the disease and understand how cancer cells achieved the hallmarks of cancer. To provide this overview, two analyses were performed:

1. characterize ovarian cancer and contrast it with other major cancer types;
2. perform cohort analysis on in-house data, and compare the result to the findings from public cohorts.

Hereditary factor is known to have significant contribution to HGSCs, where a subset of patients harbor germline inactivation in BRCA1 or BRCA2. These inherited defects render the tumors HRD phenotype and serve as a robust biomarker for treatment response. In particular, confirming carrier status of the patients can trigger cascade genetic testing and potentially prevent cancer incidence in their families. Current knowledge about HGSC suggests that half of the patients harbor somatic gene inactivation in BRCA genes or other genes in the HRR pathway. ICGC-AU-OV study further found a better prognosis in these patients. As such, there has been a tremendous interest in identifying patients with HRR gene disruption in germline or somatic settings. Therefore, the second aim of the thesis is to **identify tumors harboring inherited or acquired variants that potentially lead to DNA damage response (DDR) defect** without gene selection bias. Two analyses addressing this question are:

1. identify clinically relevant germline variants in DDR genes with the help of human geneticist;
2. identify variants most likely to induce gene inactivation in these genes and present them in a germline and somatic mutational landscape.

Recent advances in targeted therapy using PARP inhibitors opened up opportunities for further extending patient survival in those showing chemosensitivity or with HRD phenotype. The development of molecularly targeted clinical trials fostered the demand for patient stratification; however, different assays are adopted in clinical trials and there has not been a consensus in identifying patients likely respond to PARP inhibitors. In this sense, the next aim of the thesis is to **evaluate HRD phenotype in tumors**, so as to identify patients probably having chemosensitivity or potentially responding to

PARP inhibitors. In addition to the abovementioned landscape that identified potential causes of DDR defect, the following analyses further investigate this issue from different perspectives:

1. profile surrogate markers used in PARP inhibitor clinical trials, such as LOH score and HRD score;
2. incorporate two prominent DNA-based classifications, where Shah-2017 classification[132] serves as a promising biomarkers for chemosensitivity, and TDP subgroup[169] possibly reflects BRCA genes inactivation;
3. calculate activities of mutational processes reflective of HRD based on its DNA footprints in different alterations, including mutations, indels and structural variations.

On the other hand, the overall limited improvement in disease management is fundamentally due to unclear etiology and carcinogenesis model. Understanding the disease biology from the perspective of tumor evolution therefore holds the promise of providing new insights in advancing clinical care. In other cancer types, multi-sample design helps to identify truncal, thereby early, molecular events shared between related samples, which cannot be revealed by single sample. In this regard, another aim of the thesis is to **study intra-patient heterogeneity (IPH) with the multi-sample cohort, as well as to identify truncal events in the samples**. Relevant analyses in this regard includes:

1. reconstruct sample phylogeny trees and identify truncal and branch molecular events;
2. quantify intra-patient heterogeneity (IPH) and explore its prognostic implication.

In terms of carcinogenesis model, an important breakthrough was the identification of precursor lesions in the fallopian tube. As the tubal origin theory began to receive wider acceptance in recent years, pathological studies started to unveil early events in these lesions. Despite this, the common late diagnosis precludes the acquisition of early-stage tumors and the carcinogenesis process remained poorly characterized. Therefore the last aim of the thesis is to **provides a glimpse of tumor evolution in real-world time**. This will require the following analyses:

1. use variant timing technique to stratify small variants into tumor epochs;
2. combine multi-sample design and variant timing technique to refine tumor epochs and provide tumor evolutionary trajectory with finer time resolution;
3. calculate real-world time estimates for three major events in the sample phylogeny tree, including whole genome duplication (WGD), most recent common ancestor (MRCA) of the patient (`MRCA-PID`) and of each sample (`MRCA-SAMPLE`);
4. temporal dissect the abovementioned mutational processes activities along the evolutionary trajectory.

Part II

MATERIALS AND METHODS

6 MATERIALS

6.1 PUBLIC DATA SETS

6.1.1 *TCGA pan-cancer cohort*

The Cancer Genome Atlas (TCGA) program used multiple omics platforms to characterize tumors from a broad spectrum of cancer types. The data was processed through standardized bioinformatics pipelines and the results are shared with the public as well as visualized on web platforms with the effort of dedicated Genome Data Analysis Centers (GDACs). In particular, the BROAD GDAC Firehose provide Level 3 data, which includes aggregated, normalized, and segmented data.

The TCGA Pan-Cancer cohort consists of the 12 major cancer types published in 2013[171]. Level 3 data (run date 20160128) are retrieved from Firehose and serves as the basis of downstream analyses involved in Section 9.1. Specifically, mutation and indel callsets are derived from whole exome sequencing and used for mutation and indel burden statistics, copy number (CN) profile derived from SNP arrays are used for chromosomal instability (CIN) score calculation. The sample size of each disease cohort varies by assay type, and Table 6.1 provides an overview of these cohorts.

Study Name	Abbreviation	Mutation	Indel	CN
Bladder Cancer	BLCA	130	129	411
Breast Cancer	BRCA	992	939	1096
Colon Cancer	COAD	154	140	453
Glioblastoma	GBM	290	277	590
Head and Neck Cancer	HNSC	279	274	524
Kidney Clear Cell Carcinoma	KIRC	428	378	529
Acute Myeloid Leukemia	LAML	192	150	197
Lung Adenocarcinoma	LUAD	230	228	518
Lung Squamous Cell Carcinoma	LUSC	177	163	501
Ovarian Cancer	OV	316	260	601
Rectal Cancer	READ	69	55	166
Endometrioid Cancer	UCEC	248	247	541

Table 6.1: Overview of TCGA Pan-Cancer cohort.

In addition, cohort analysis results were retrieved for visualization. Level 3 data includes GISTIC[172] result, where the recurrence of copy number changes are calculated

and recurrent genomic regions, or peaks, are defined here as those with q-value <= 0.25. On the other hand, information of significantly mutated genes (SMGs) derived from MutSigCV[173] analysis is based on a previous publication from TCGA network[174]. SMGs are defined here as significantly mutated genes with q-value <= 0.1 in the MutSigCV result.

6.1.2 *TCGA-OV*

TCGA-OV[130] is by far the largest cohort that profiled exonic mutations and indels in 316 ovarian cancer genomes, all with HGSC histotype. This callset is retrieved from Firehose as described above and contains 20,170 somatic variants. Germline frequencies of BRCA1 and BRCA2 were further retrieved from the literature description. In the original report, cohort analyses were done to identify recurrently mutated genes and recurrent copy number alterations (CNAs). These results are obtained from supplementary data of the publication and are used in Section 9.2.1.1 and Section 9.2.3 for side-by-side comparison with HIPO59 results.

In its supplementary table 2.3b, MutSig result suggested 9 significantly mutated genes (FDR<0.15), namely TP53, BRCA1, BRCA2, NF1, RB1, CDK12, FAT3, CSMD3 and GABRA6. On the other hand, GISTIC (version 2.0) identified recurrent CNAs in terms of broad events (supplementary table 5.1) and focal events (supplementary table 5.2).

The recurrent focal events, including 63 amplifications and 50 deletions, are provided with their genomic locations in hg18 coordinate. In addition, putative targets are proposed in 43 of these focal events. These 43 driver genes define a gene set **Prior_Target_TCGA** that will be used when nominating putative targets in HIPO59 GISTIC result in Section 8.4.2.

6.1.3 *ICGC-AU-OV*

The International Cancer Genome Consortium (ICGC) coordinates large-scale cancer genome studies across the globe where TCGA also takes part in. It covers more than 50 different cancer types and aims for at least 500 samples per cancer type. Next to the TCGA-OV study, the second largest ovarian cancer study is the Australian cohort in ICGC, denoted here as ICGC-AU-OV[131]. In this cohort, 114 samples from 92 HGSC patients were profiled with WGS platform.

Somatic mutations and indels are retrieved from ICGC Data Portal[175] release 28. Among them, only variants in non-redundant primary tumors from 82 patients are considered in this study. This yields a callset containing 549,757 somatic variants. Germline frequencies of BRCA1 and BRCA2 were further retrieved from the literature description.

In the original report, IntoGen[176] analysis was done to identify recurrently mutated genes and the result suggested that TP53, RABGGTB, RB1, NF1, BRCA1, BRCA2, DNAH1 and FLNA played driver roles (FDR<0.05).

6.1.4 *OV133*

The OV133[132] consists of 133 ovarian tumors of various histotypes, including 59 HGSC, 35 CCC, 29 EC and 10 adult granulosa cell tumors. It is the third largest HGSC cohort and was also profiled with WGS platform. Throughout the thesis, OV133 will be referred to as the subset of 59 HGSC samples instead of the entire cohort, if not specified.

Although the callset of OV133 is not publicly available, the aberration status in 17 genes of interest can be retrieved from supplementary table 5 of the original report. These include somatic functional variants in BRCA1, BRCA2, RB1, POLE, RPL22, PIK3R1, PPP2R1A, KMT2B, NF1, FOXL2, CTNNB1, KRAS, PER3, PTEN, ARID1A, PIK3CA and TP53, germline functional variants in BRCA1 and BRCA2, as well as methylation status of BRCA1.

According to the literature, functional variants include mutations and indels with functional consequence falling into 4 SnpEff[177] categories, namely non-synonymous coding mutations (NON_SYNONYMOUS_CODING), stop-gained/loss mutations (STOP_GAINED, STOP_LOST), splice-site mutations (SPLICE_SITE_ACCEPTOR, SPLICE_SITE_DONOR) and frameshifts (FRAME_SHIFT).

6.2 IN-HOUSE DATA - HIPO59

This dataset comes from a research project HIPO59 under DKFZ Heidelberg Center for Personalized Oncology (DKFZ-HIPO) program. The project aims to characterize the spatial and temporal tumor heterogeneity of ovarian cancer, especially the HGSC histotype. To investigate spatial heterogeneity, multiple specimens were sampled in tumor tissues removed from different anatomical sites before treatment. High-throughput assays were then used to profile these specimens at different omic layers, including genomic information from WGS, transcriptome from RNA sequencing and methylome from methylation arrays.

The sample statistics of HIPO59 cohort is shown in Figure 6.1. In total, there are 74 samples from 42 OC patients. Among them, 55 samples were from 33 HGSC patients. The spatial heterogeneity of OC can be addressed with a subset of the cohort comprising 16 patients having multiple samples.

Most of the samples are profiled with three platforms, except for the one taken from interval debulking surgery of patient H059-ASG5U9 and was without transcriptome profiling.

The data was profiled in 2018 and the clinical information and survival status were collected and followed-up by medical doctors until January 2020. As a summary, Figure 6.2 shows the survival status of the cohort until the last follow-up.

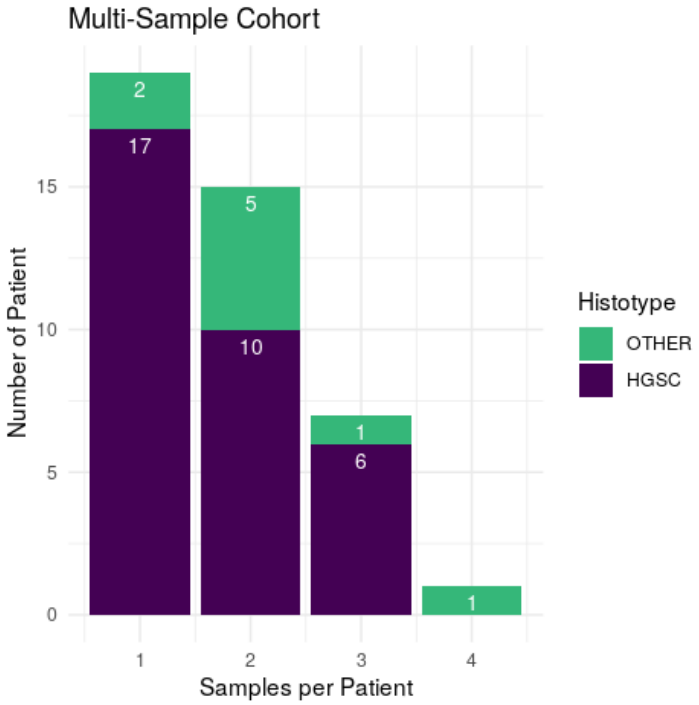


Figure 6.1: Sample statistics in HIPO59 cohort.

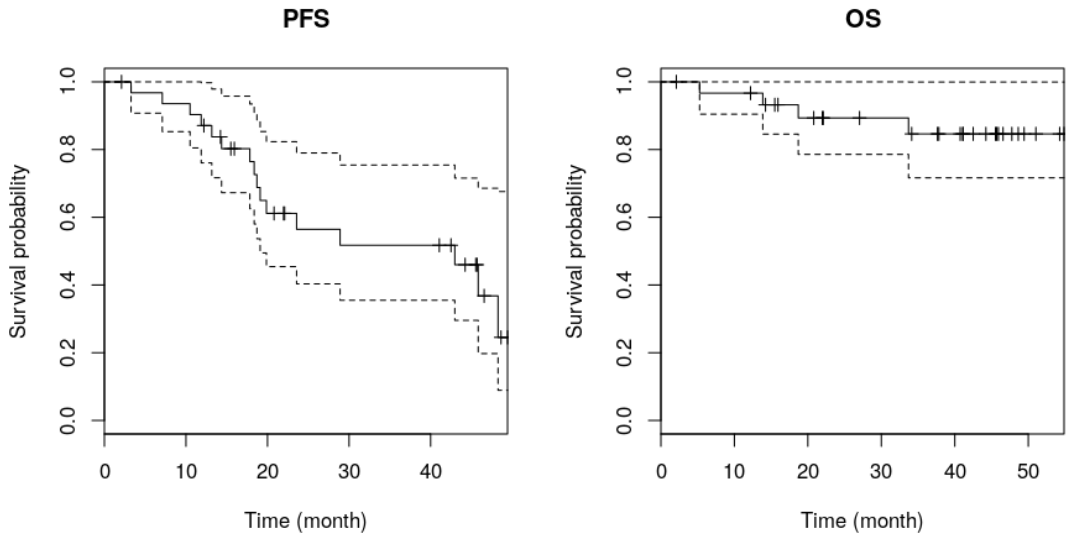


Figure 6.2: Survival status of patients in HIPO59 cohort.

6.3 GENES OF INTEREST

Cancer Gene Census (CGC)

There has been many oncogenes and tumor suppressor genes for which alterations in either germline or somatic settings are causally implicated in cancer. The COSMIC curation team collects published evidence and catalogues these genes into a high-confidence list of genes, the Cancer Gene Census (CGC). The CGC release 90[178] documented 723 cancer-associated genes. The 717 of them having genomic coordinate information were included in the gene list **Prior_Target_CGC**. Among them, 243 genes had documented role as oncogene (OG), 243 genes played as tumor suppressor gene (TSG), and another 72 genes serve as either OG and TSG depending on context.

Familial cancer genes

The **FamilialCancerGenes** is a list of 152 familial cancer related genes comes along with the Clinical Workflow (see Section 7.2).

DNA damage response genes

As described in Section 1.2.2, a previous study looked at alterations in DDR pathways within TCGA cohorts[19]. The authors defined 276 genes that play roles in coordinating responses to DNA damage. These genes are involved in nine major repair pathways as shown in Table 6.2 and comprise the gene list **PanCanDDR**. Within each pathway, a subset of genes were specifically listed. They represent those observed in HIPO59 cohort and qualified for further pathogenicity review (see Section 8.6.1) in this pathway.

Repair pathway	Number	Genes qualified for geneticist review
Direct Repair	4	-
Fanconi Anemia	41	BARD1, BLM, BRCA1, BRCA2, BRIP1, ERCC4, FANCA, FANCB, FANCC, FANCE, FANCG, FANCM, PALB2, RAD51C, SLX4
Homology-dependent recombination	88	BARD1, BLM, BRCA1, BRCA2, BRIP1, FANCM, NBN, PALB2, POLD1, RAD50, RAD51C, RAD51D, RECQL4, SLX4, WRN
Mismatch Repair	24	MLH3, PMS1, PMS2, POLD1
Non-homologous End Joining	23	FAM175A, NBN, RAD50
Nucleotide Excision Repair	51	ERCC2, ERCC4, MMS19, POLD1, POLE
Translesion Synthesis	20	-
Nucleotide pools	5	-
Others	65	ATM, ATR, ATRX, PTEN, SLX4, TP53

Table 6.2: DNA Damage Response genes involved in 9 repair pathways. This table describes the constitution of **PanCanDDR**, with the size of each DDR pathway specified. The last column listed genes qualified for pathogenicity review based on criteria in Section 8.6.1.

7 DKFZ WORKFLOW

Next generation sequencing generates primary data as sequencing reads, which went through a series of standardized analyses including alignment, variant calling and a variety of annotation and filtering steps, and finally produce variant callsets of different alterations. These analyses were wrapped up as **workflows**, which takes in sequencing data as .fastq files or .bam files, and generate variant callsets as output. The in-house HIPO59 project was processed by Omics IT and Data Management Core Facility (ODCF) using a collection of DKFZ in-house workflows.

7.1 BASIC WORKFLOWS

To start with, the `AlignmentAndQC Workflow` (version 1.2.73-1) takes in .fastq file, use BWA-MEM[179] for alignment and generates .bam file.

Single-nucleotide variants and small insertion/deletionss (indels) are processed by `SNVCalling Workflow` (version: 1.2.166-1) and `IndelCalling Workflow` (version: 1.2.177), respectively. In brief, the former uses samtools[180] and bcftools[181], and the latter uses Platypus[182] for variant calling. More details about these workflows can be found in the publication[183]. To note, as built-in module in both workflows impose a basic filter based on the location and the predicted consequence of a variant, the small variant call set, if not specified, contain only functional variants that potentially changed the translated peptide sequences. Together, these two workflows generated germline functional mutations and indels, which is referred to as **Initial call set**. From the Initial call set, the somatic fraction is referred to as **SomFxn call set**, which include somatic functional mutations and indels. As described below, the `ClinicalWorkflow` then take the germline fraction as input and further exclude variants that are less likely to be harmful.

Two other workflows, as part of DKFZ in-house workflows, were used to process HIPO59 data. The first workflow `SophiaWorkflow` (version 1.2.16) uses Sophia[184] to generate SV callset. The second workflow `ACEseqWorkflow` (version 1.2.8-4) uses ACEseq to generate copy number profiles[185].

7.2 CLINICAL WORKFLOW

Most of the consequence of germline functional variants are tolerable and collectively create the diversity in human population, only a small subset of them are pathogenic and can cause disease. It is generally accepted that the more common a variant is found in the population, the less likely it can cause severe diseases. Therefore, a second filter applied here is the germline filtering module in the `ClinicalWorkflow` (version 1.1). This module consider only variants with Variant Allele Frequency greater or equal to 0.3,

and remove those found to be common in different normal control cohorts. Specifically, variants meeting following criteria were removed: (1) found in 1000 genome and with allele frequency greater than 0.1, (2) found in dbSNP database with tag "COMMON=1", (3) found in ExAC collection with AC_HOM tag greater than 3 or AC_HET tag greater or equal to 40, and (4) found in DKFZ local normal control cohort with greater than 50 ACs.

Implemented by Dr. Barbara Hutter, the `ClinicalWorkflow` starts from the germline fraction of the Initial call set, and generates results in the **GermFxn** call set. To note, an optional filter within the germline filtering module was to restrict variants to those pertinent to **FamilialCancerGenes** genes, which is not applied during the HIPO59 germline analysis.

8 SPECIFIC TASKS

8.1 TCGA PAN-CANCER GENOME INSTABILITY MEASURES

Given the TCGA Level 3 data (see Section 6.1.1), three summarizing measures were used to quantify the extent of genome instability in each individual tumor. Following the previous work of Alexandrov *et al.*[24], mutation and indel prevalence of each tumor are defined as the number of somatic variants per mega base pairs. This is estimated with the number of somatic variants detected in protein-coding regions and assuming an average of 30Mb of the exome being effectively captured and sufficiently sequenced.

The weighted genome integrity index (wGII) describes the percentage of the genome being altered in a tumor sample. It takes values between 0 and 100 and was calculated following the original concept proposed in the literature[186]. Based on the copy number profile derived from SNP array, the ploidy of a tumor is first calculated as the weighted mean of copy number across all genomic segments. Subsequently, copy number segments are classified as gain, loss or neutral according to their copy number change relative to the ploidy of the sample, where a difference greater than 0.7 copy is considered changed. The integrity of a chromosome is described by the fraction of its genomic materials being classified as altered (either gain or loss). The altered fraction is determined separately for each of the 22 autosomal chromosomes and the average fraction across 22 autosomes determines the wGII score of the tumor.

8.2 RECURRENT MUTATIONS AND INDELS

In cohort analysis utilizing mutations and indels, the aim is to find out which genes are more likely to be subject to positive selection as compared to the background mutation rate. However, the mutation frequency at different genomic location is heterogeneous instead of being constant across the genome. The regional variation is correlated with factors like sequence context, DNA replication timing and transcriptional activity. As a result, the relative frequency of mutation and indels would not approximate positive selection during tumorigenesis unless these factors were taken into considerations.

8.2.1 *Cohort analysis using MutSigCV*

MutSigCV (Mutation Significance with Covariates)[173] models the local background mutation rate by considering patient-specific rate and region-specific rate, as well as taking into account the replication timing, transcription activity, gene length and gene sequence composition. To apply this algorithm to our cohort, MutSigCV (version 1.41)

was obtained from Broad Institute Cancer Genome Analysis website[187] and applied to mutations and indels found in 33 representative samples.

The MutSigCV result from HIPO59 cohort (n=33) is then compared to the MutSig result dereived from the TCGA-OV cohort (n=316)[130] (see Section 6.1.2)

8.2.2 *Reported significantly mutated genes profiled across cohorts*

The aim of this section is to profile the frequencies of known driver genes in different ovarian cancer cohorts. Specifically, 12 recurrently mutated genes reported in the TCGA-OV[130] and ICGC-AU-OV[131] were to be profiled in four selected cohorts, namely TCGA-OV (n=316), ICGC-AU-OV (n=82), OV133 (n=59)[132] and HIPO59 (n=33). The collection of driver gene information and callsets of public cohorts are described in Section 6.1.

To start with, variants in callsets are classified int four categories listed in Table 8.1. Given multiple variants can occur in the same gene in one sample, the variant with highest priority is used to describe how this gene is mutated in the patient. The fraction of the cohort affected by variants of different categories can then be determined for each driver gene.

Category	Definition	Priority
germline functional	any nonsynonymous, stop-gained/loss, splice-site or frameshift germline variants	1
somatic functional	any nonsynonymous, stop-gained/loss, splice-site or frameshift somatic variants	2
somatic silent (exonic)	any synonymous variant	3
somatic silent	any intronic variant	4

Table 8.1: Variant classification and priority of categories.

Nonetheless, owing to the different experimental platforms and data accessibility, some variant categories were not assayed or not released in most cohorts except for HIPO59 cohort, which offered comprehensive information in all categories. Table 8.2 summarised the available information in each cohort.

8.3 GENE BREAKAGE EVENTS IN GENES OF INTEREST

Except for inactivating mutations and indels, gene function can also be impaired by copy number alteration-associated events (CNA-associated events), where an interruption of the gene body inactivates the affected allele. The majority of them are accompanied by SVs, while a handful of them are with unknown origin. With the high resolution of WGS technique, it is possible to characterize such SVs with single base pair precision and see whether they might induce gene breakage.

Category	TCGA	ICGC	HIPO59	OV1333
germline functional	BRCA1, BRCA2	BRCA1, BRCA2	O	BRCA1, BRCA2
somatic functional	O	O	O	TP53, NF1, RB1, BRCA1, BRCA2
somatic silent (exonic)	O	O	O	X
somatic silent	X	O	O	X

Table 8.2: Data accessibility in four cohorts. Symbol "O" means the data is available for all genes in the cohort; while "X" means the other way around. When accessible information is restricted to limited number of genes, gene name of the available ones are specified.

As SV does not account for all copy number change in the genome, CNA-associated events of unknown origin are also included to enable more comprehensive characterization.

8.4 RECURRENT SOMATIC COPY NUMBER ALTERATIONS

The genomic copy number profile of a cancer genome, or the macroscopic karyotype, is a result of the accumulation of somatic copy number alterations (SCNAs). To understand which of the genomic regions underwent frequent alteration across a cohort, a common approach is to evaluate the frequency and amplitude of observed copy number changes and identify regions that are altered above the background rate.

Depending on its length, SCNA can be an arm-level broad event or a focal event that is with relatively small range. It has been observed that the abundance of SCNAs of these different types are different, where broad events are more frequent and focal events occur at a frequency negatively correlated to its length[188]. Therefore, the background rate of these two types of events should be modeled separately.

8.4.1 *Cohort analysis using GISTIC*

GISTIC (Genomic Identification of Significant Targets in Cancer)[172] identifies recurrent copy number changes in a cohort using a three-step procedure. At first, it decomposes the observed copy number profile into underlying SCNA events and separates them into broad (arm-level) and focal events by length. Having SCNAs assigned to each region, the algorithm then score a region by the probability of the SCNA's occurring by chance, where the probability was estimated based on the background rates of matching SCNA types across the cohort. Lastly, within each significant region identified, a probabilistic method was used to define a peak region that has user-defined level of confidence of containing the true driver event, in our case a 90% of confidence. In the end, the program reports significant **Regions** and its subregions (**Peaks**) most likely contain true targets.

To apply this algorithm to HIPO59 cohort, GISTIC (v2.0.23) was used to analyze the copy number profiles of 33 representative samples. Parameters chosen are listed in Table 8.3.

GISTIC Parameter	HIPO59 Run Setting
do_gene_gistic	TRUE
broad_len_cutoff	0.5
conf_level	0.9
arm_peeloff	TRUE
gene_collapse_method	extreme
ziggs.max_segs_per_sample	3500

Table 8.3: GISTIC Run Parameters.

In summary, GISITC identifies genomic regions that are altered above background rate and assigns a q-value to each region. According to the size of the region, they can be either broad events (arm-level) or focal events. For each focal event, the program further identify the corresponding peak region that most likely covers the true target.

8.4.2 *Target nomination in recurrent regions*

Once the recurrent focal events are identified, the next step is to find out why they are preferentially selected during tumor development. One of the approach is to identify genes sitting in each **Regions** or **Peaks** and nominate the most likely target gene that is subject to selective constraints.

To do so, additional prior knowledge is incorporated. First of all, a set of 43 genes previously nominated as putative targets in recurrent focal CNAs in the TCGA-OV study were used. These genes, designated 'Prior_Target_TCGA' (see Section 6.1.2), are with the first priority. Secondly, known cancer-associated genes from the Cancer Gene Census[178] are set with second priority. This 'Prior_Target_CGC' gene set (see Section 6.3) is composed of 717 oncogenes and tumor suppressor genes.

In summary, each recurrent region will initially be annotated with overlapping genes. When there are more than one gene locate in the region, a minimal set of putative target genes, if any, will be prioritized based on this nomination process.

8.5 PATIENT STRATIFICATION

8.5.1 *TDP subgroup*

8.5.1.1 *Method overview*

To implement the TDP subgroup stratification described in Section 4.5.1, Menghi *et al.* adopted a three-step classification scheme for classifying tumors into TDP subgroups.

To start with, the TDP score proposed in [170] was used to identify TDP-positive tumors. This metric accounts for both TD number and their distribution across the genome, therefore able to distinguish clustered TDs and stochastically scattered TDs. Tumors with positive score would have higher propensity for TD formation and thus classified as TDP tumors. Subsequently, sample-wise TD span size distributions are profiled, with their major peaks enumerated. These peaks were then mapped to one of the five peak classes of predetermined size intervals. In the end, each sample is assigned to a TDP subgroup according to the peak class composition in its TD span size distribution.

8.5.1.2 *Implementation*

The three-step classification scheme is implemented in this study with details described below.

First of all, the original formula (8.1) was expanded by specifying how to derive Exp_i. Assuming a uniform rate of generating TD throughout the genome, the expected number of TDs found in one chromosome is dependent on its length chr_i. Let TD_{total} stands for the total number of tandem duplications found in this sample, this can be expressed as (8.2).

$$TDP\ Score = -\frac{\sum_i |Obs_i - Exp_i|}{TD_{total}} + k \tag{8.1}$$

$$Exp_i = TD_{total} \times \frac{chr_i}{genome\ length} \tag{8.2}$$

The calculation of TDP score of a given sample thus becomes (8.3), where Obs_i is the number of TDs found in chromosome i and k is a constant 0.71 defined in the original paper. Samples with TDP score greater than 0 are classified as TDP-positive tumors, while those smaller than 0 as TDP-negative tumors.

$$TDP\ Score = -\frac{\sum_i \left|Obs_i - TD_{total} \times \frac{chr_i}{genome\ length}\right|}{TD_{total}} + k \tag{8.3}$$

Secondly, the density distribution of TD span size was profiled for TDP-positive tumors, with peaks quantified by their position (span size) and abundance (density). There might be more than one peak found in the density distribution. Given the first major peak being the largest one, additional major peaks are kept if their density <= 25% of the largest peak density.

Lastly, major peaks found in each sample were mapped to five peak classes predefined based on size interval, namely **class 0 (<1.6kb)**, **class 1 (1.64-51kb, median 11kb)**, **class 2 (51-622kb, median 231kb)**, **class 3 (622kb-6.2Mb, median 1.7Mb)** and **class 4 (>6.2Mb)**. One then can eventually assign the TDP-positive tumors to one of the TDP subgroups according to its class composition.

8.5.2 *Shah-2017*

Given the 20 genomic features used for Shah-2017 subgroup discovery (see Table 4.1), the procedure described below shows how these features were derived from the HIPO59 cohort and implemented as Table 8.4.

First of all, the contribution of COSMICv2 Signatures were calculated using the R package YAPSA (version 1.0.0)[189]. The effect of mutations and indels on corresponding genes were annotated with ANNOVAR[190]. These small variants were then selected based on ANNOVAR annotations that best match the original feature descriptions.

The rearrangements were based on SV callset from Sophia[184]. A pre-processing step was done to exclude any SVs with span size smaller than 30 bp, or any deletions with span size smaller than 1000 bp. Subsequently, all SVs were classified based on their rearrangement type and span size. The re-classification follows a sequential order of foldback inversions, inversions, tandem duplications, deletions and with the rest assigned to unbalanced rearrangements. To note, two of the rearrangement features were not implemented due to insufficient information from the SV callset.

Lastly, copy number profiles underwent a pre-processing step to exclude segments shorter than 5000 bp before deriving CNA-related features.

Following these steps, 18 genomic features were summarized in 55 HGSC samples from 33 HGSC patients. The patient stratification was then a simple implementation of hierarchical clustering with samples split into two clusters. This was done with R package pheatmap (version 1.0.12)[191], using "ward.D" clustering methods and "manhattan" distance for the clustering.

8.6 GERMLINE VARIANT ANALYSIS

Among all germline variants found in a person's genome, most of them are not pathogenic and shared with other people in the population. Their pathogenicity and implication for ovarian cancer risk can only be ascertained after a manual review process carried out by human geneticists.

In order to make the manual review feasible, it is important to exclude as many non-harmful variants as possible. This reduction was facilitated by the Clinical Workflow and described in Section 7.2. Therefore, downstream analyses in this section were based on the GermFxn call set derived from the Clinical Workflow.

8.6.1 *Clinical implication of the germline variants*

The GermFxn call set consists of rare and protein-changing germline variants. However, their pathogenicity determination rely on the evaluation by human geneticists. Due to the large amount of variants found in the GermFxn call set, only a subset of manageable amount went through geneticist review process. This subset consists of variants that are implicated in familial cancer and also have a role in the DDR process, as shown in Table 6.2. Specifically, the selection is based on the intersection of two

Category	Feature	Implementation
Signature	S.APOBEC	Signature 13 (**AC13**)
Signature	S.POLE	Signature 10 (**AC10**)
Signature	S.AGE	Signature 1 (**AC1**)
Signature	S.BC	Signature 8 (**AC8**)
Signature	S.MMR	Signature 6 (**AC6**)
Signature	S.HRD	Signature 3 (**AC3**)
Variant	Nonsynonymous	**nonsynonymous SNV** (a single nucleotide change that cause an amino acid change)
Variant	Splicesite	**splicing** (variant is within 2-bp of a splicing junction)
Variant	Stop.Lost/Gained	**stopgain**, **stoploss** (a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation/elimination of stop codon at the variant site)
Variant	Frameshift	**frameshift insertion**, **frameshift deletion** (an insertion/deletion of one or more nucleotides that cause frameshift changes in protein coding sequence)
SV	Foldback.Inversion	unbalanced inversion (**INV**) with size < 30Kb
SV	Inversion	inversion (**INV**) with size < 1Mb
SV	Tandem.Duplication	duplication (**DUP**) with size < 1Mb
SV	Deletion Rearrangement	deletion (**DEL**) with size < 1Mb
SV	Balanced Rearrangement	NA
SV	Unbalanced Rearrangement	the rest of all unclassified SVs
SV Property	Homology>=5bp	NA
CNA	CN.Amplification	Total copy number (**TCN**) > ploidy + 2
CNA	CN.Loss	Homozygous deletion, hemizygous deletion
CNA	CN.LOH	One of the allele specific copy number being 0

Table 8.4: Implementation of Shah-2017 genomic stratification. This table described how the 20 genomic features in Table 4.1 were implemented in HIPO59 cohort. Text highlighted in bold are terms used in the softwares that derived the results, namely, YAPSA for signature analysis, ANNOVAR for variant annotation, Sophia for SV discovery, and ACEseq for copy number profiling.

gene lists, namely **FamilialCancerGenes** and **PanCanDDR**, plus four additional genes (ARID1A, POLH, STK11, EPCAM) and exclude TP53.

Exceptions made in the selection rule are due to the following reasons. Causative genes of rare syndromes conferring predisposition to ovarian cancer, if not DDR genes, are additionally included. They are STK11 for *Peutz-Jegher syndrome* and EPCAM for *Lynch Syndrome*. POLH is a DDR gene functioning in both HRR and TLS subcategories. Although it is not implicated in familial cancer, it was included because of a somatic variant occurred in a patient with exceptionally high tumor mutation burden. TP53 fulfills the selection criteria however no germline variants were observed, whereas its somatic variants are already known to be important for ovarian cancer and therefore excluded. Lastly, somatic ARID1A mutations are known to be a characteristic of some ovarian cancers.

Dr. med. Laura Gieldon, as a human geneticist in the Technische Universität Dresden, reviewed the germline variants based on ACMG-criteria[192]. The variants are classified into 6 categories: Pathogenic; Likely pathogenic; variant of unknown significance (VUS); Likely benign; Benign; gene of unknown significance (GUS). Usually the pathogenic and likely pathogenic germline variants are reported to the patients and predictive testing are subsequently offered to their family members. VUS are only reported in some selected cases depend on the context.

Beside germline variants, somatic variants selected by the same criteria were also sent for pathogenicity determination using the same ACMG-criteria. Of note, the criteria was not designed for somatic variants assessment and therefore the evaluation result is not meant for treatment response decision.

8.6.2 *Functional enrichment of germline variants*

Genes do not function alone, they usually cooperate with each other to carry out a specific biological process. Given a group of genes of interest, functional enrichment is a way of knowing the functional roles these gene might collectively perform. This analysis can therefore help give a better understanding of whether germline variants are preferentially perturbed in DDR pathway and if so, which DDR axis were affected.

To start with, a total of ten gene lists were prepared to be tested. These are DDR in general (**PanCanDDR**), and its nine sub-categories as previously defined in Section 6.3. Next, genes are classified into a contingency table in two ways, namely whether it harbors any variants in the GermFxn call set, and whether it participate in the pathway of interest. A one-sided Fisher's exact test was then used to test whether there is a positive association between occurrence of germline variants and the DDR or its sub-categories. Associations with p-value smaller than 0.05 were considered significant.

This analysis was done on the entire HIPO59 cohort and repeated on a subset of the cohort considering HGSC subtype alone. The resulting p-values are compared between the two scenarios to see the effect in ovarian cancer in general compared to HGSC subtype alone.

Variant Classification	ANNOVAR Annotation
deleterious	frameshift insertion, frameshift deletion, stopgain, stoploss
unknown	nonsynonymous SNV, nonframeshift deletion, nonframeshift insertion, unknown and splicing

Table 8.5: Variant classification based on ANNOVAR annotation.

8.7 GERMLINE AND SOMATIC LANDSCAPE

Sporadic tumors may share similar abnormalities with hereditary cancers. For example, loss-of-function of BRCA1 can result from germline variant, somatic aberration or promoter hypermethylation[130]. This implies that carcinogenesis can be a common consequence of targeting same function using different mechanisms.

This section aims to integrate variants identified in both germline and somatic settings and focus on genes functioning in pre-defined pathway. Specifically, GermFxn and SomFxn call set derived from the Clinical Workflow were used and the downstream analyses consider only the 275 genes in the DDR pathway.

Furthermore, as some samples may be derived from same patient, sample-wise call sets are summarized into patient-wise call sets. When taking the intersection of sample-wise SomFxn call sets, it is equivalent to finding so-called **truncal** events. Instead, when taking the union of sample-wise SomFxn call sets, both truncal and branch events remains in the patient-wise SomFxn call sets.

8.7.1 *Variant classification*

Variants in the call sets are classified into three groups (deleterious, unknown and benign) using a two-step procedure based on functional consequence and geneticist review result.

The initial classification was based on functional consequences of the variant predicted by ANNOVAR. As listed in Table 8.5, all protein-truncating variants are classified as "**deleterious**" while other protein-changing variants are assigned "**unknown**".

Subsequently, the classification was refined according to geneticist review result. Variants assessed as (likely) benign are re-classified as "**benign**" and those assigned with (likely) pathogenic are re-classified as "deleterious". Note that this re-classification only applied to the 49 germline variants and 25 somatic variants sent to human geneticist for review, as described in Section 8.6.1. As a result, only a subset of variants in the deleterious group are deemed (likely) pathogenic, with the rest being evaluated as GUS/VUS or not short-listed for evaluation.

Regarding the DDR pathway member genes, all germline protein-truncating variants are (likely) pathogenic except for three stopgain SNVs (affecting FANCM, MLH3 and RAD1) and two INDELs (locating in APEX1 and RFC2). Two additional protein-

changing variants were re-classified as deleterious, including one splicing event in FANCC and one nonframeshift deletion in RAD51C.

On the other hand, out of the 11 somatic protein-truncating variants in the pathway, only 7 are assessed as (likely) pathogenic, including 2 SNVs (affecting BRCA2 and PTEN) and 8 INDELs (locating in ATM, BRCA1, BLM, POLH and PTEN). An additional somatic missense variant locating in PTEN (NM_000314.6, c.509G>T, p.Ser170Ile) is re-classified as deleterious and would have been missed without geneticist evaluation.

8.7.2 *Germline and somatic landscape of DNA damage response (DDR) pathways*

The germline landscape is based on GermFxn call set and the Germline Variant Number is the size of this set. On the other hand, the somatic landscape is based on the patient-wise SomFxn call set containing only truncal events. Nonetheless, the Somatic Variant Number stands for the size of the patient-wise SomFxn call sets containing both truncal and branch events.

The variant call sets were collapsed to pathway-level and specifically looking at DDR pathway and its nine sub-categories described in Table 6.2. Mutual exclusivity and enrichment test are derived from one-sided Fisher's exact tests.

8.7.3 *Aberrations potentially contributing to DDR defect*

The germline landscape is based on GermFxn call set and the somatic landscape is, different from in the previous section, based on the patient-wise SomFxn call set containing both truncal and branch events. Variants were classified into deleterious, unknown and benign as described in Section 8.7.1.

Bi-allelic inactivation, where functioning (wild-type) allele is considered not retained in tumor cells, can be achieved by either LOH or multiple hits targeting different alleles. Therefore, genes harboring these two events were considered candidates for bi-allelic inactivation.

Multiple-hits occur in a gene when more than one germline or somatic variant were identified in the same sample. LOH event, on the other hand, requires two criteria that take into consideration the local allele-specific copy number estimates for the chromosomal segment where the variant resides.

The first criteria is that the variant locate at a segment exhibiting LOH, which means the minor allele copy number equals zero. The second criteria is that the variant copy number (multiplicity) should be close to major allele copy number of the segment, with the difference between them being less than one copy. With these two criteria fulfilled, the variant is likely to locate at segment with LOH and belongs to the retained allele.

8.8 GENOMIC FOOTPRINTS ANALYSIS

Instead of developing de novo signatures, this section takes known signatures and infer their activities in the HIPO59 cohort. These analyses, conducted by established

Aberration Type	Signature Source	Signature Set	Software
Mutations	COSMICv2[195]	AC1-AC30	YAPSA(v1.0.0)[189]
Indels	COSMICv3(ID)[196]	ID1-ID18	YAPSA(v1.14.0)[197]
Structural variants	Br.Rs[26]	Br.RS1-Br.RS6	SigIT(v1.0.1)[194]

Table 8.6: Signature analyses based on footprints in mutations, indels and structural variants.

softwares, follow similar principles. They start with collecting a catalog of signatures from previous studies. Each signature is a vector of a fixed number of features with their weights and can be think of as the contribution of a biological process to a pre-defined set of footprints. Next, in each sample the observed abundance of each footprint is quantified. It is achieved by first enumerating genome-wide somatic aberrations in the sample, and then classify them into different feature categories based on footprint definitions. As the observed footprint is a collective consequence of many biological processes, the observed feature profile is an aggregation of different signatures with different activity level. A mathematical decomposition can then break down the observed profile into a set of known signatures with different weights that best describe the observation. Eventually, the signature weights are interpreted as the activity level of each process in a sample.

Three sets of signatures, each looking at footprints of different aberration types, are analyzed with two softwares (either YAPSA[193] or SigIT[194]). Table 8.6 describes the signatures and the softwares used for analyzing mutations, indels and structural variants.

To note, when performing COSMICv2 signature analyses with YAPSA(v1.0.0), a two-stage approach was adopted. In the first stage, the decomposition was based on all signatures and the contribution of all signatures were obtained. This gave an idea of which signatures were more active in the sample. A threshold of 6% (0.06) contribution to the observed footprints was used to identify **active signatures** in each sample. In the second stage, the same decomposition was repeated while using only active signatures and the final contribution of these active signatures were determined. Therefore, in each sample there would be a different set of inactive signatures with contribution set to 0.

On the other hand, the two-stage approach was also used when performing COSMICv3 signature analysis with YAPSA(v1.14.0). However, this time a cohort-wise signature-specific threshold was used as suggested in the tutorial. The set of threshold used (cutoffPCAWG_ID_WGS_Pid_df, optimal cost factor 3) contains optimal cutoff derived from training data.

When there were multiple samples from the same patient, it was observed that not all samples would have identical active signature set. In this case, a signature is said to be active in the patient when it is active in at lease one of the samples.

8.9 TUMOR HETEROGENEITY

8.9.1 *Quantifying heterogeneity between tumor samples*

Different tumors, despite derived from the same patient, can have different genomic constitutions. Their differences are oftentimes described by visualizing or enumerating shared or private mutations and indels. In this section a quantitative measure was used to summarise such differences. This measure can be further generalized to describe differences in terms of structural variants or copy number profiles.

Take small variants as an example. Somatic mutations and indels detected in two different samples are compared using a venn diagram, where each circle represent the callset from one sample. Shared variants are shown as the intersection, while private variants account for the non-overlapping fraction. Figure 8.1 visualized such comparisons for each of the 16 HGSC multiple sample sets in HIPO59. It is observed that 7 patients had majority of the variants shared between different samples. For each pair of samples, a **Jaccard Index** can be used to quantify their similarity using the concept of venn diagram. It is calculated as dividing the intersection (number of shared events) by the union (number of all events) and takes values between 0 and 1. As we are interested in the heterogeneity, a **Heterogeneity Index** is defined as $1 - \textit{Jaccard Index}$. Therefore, the smaller the Heterogeneity Index, the more similar the two samples are, a score of 0 means the two callsets are identical.

When comparing small variants, events are defined as the genomic locations of variants and events sharing the same genomic position are considered shared. When it comes to structural variants, one pair of breakpoint positions caused by one SV is considered as one event, a shared event then represent another pair of same breakpoints being found in the other sample. In terms of copy number profiles, a segment, having two alleles, is considered as two events and only alleles with the same allele-specific copy numbers between two tumors are considered shared. Eventually, each pair of sample comparison within a multiple sample set would yield three heterogeneity scores, one being heterogeneity index based on mutation and indels (`HET_MutIndel`), one being heterogeneity index based on structural variations (`HET_SV`) and another being heterogeneity index based on copy number profiles (`HET_CNA`).

8.9.2 *Stratify patients with heterogeneity score*

For each pair of related samples, three Heterogeneity Indexes (`HET_MutIndel`, `HET_SV`, `HET_CNA`) were calculated. Based on these three indexes, a k-means clustering algorithm was used to classify all comparisons into three clusters. Specifically, the 'kmeans' function R package 'stats' was used to perform the clustering, with all parameters set as default except for setting fixed seed and specifying desired cluster number to be 3. In the end, each patient is assigned a heterogeneity group based on the majority vote of the cluster membership of its within patient comparisons.

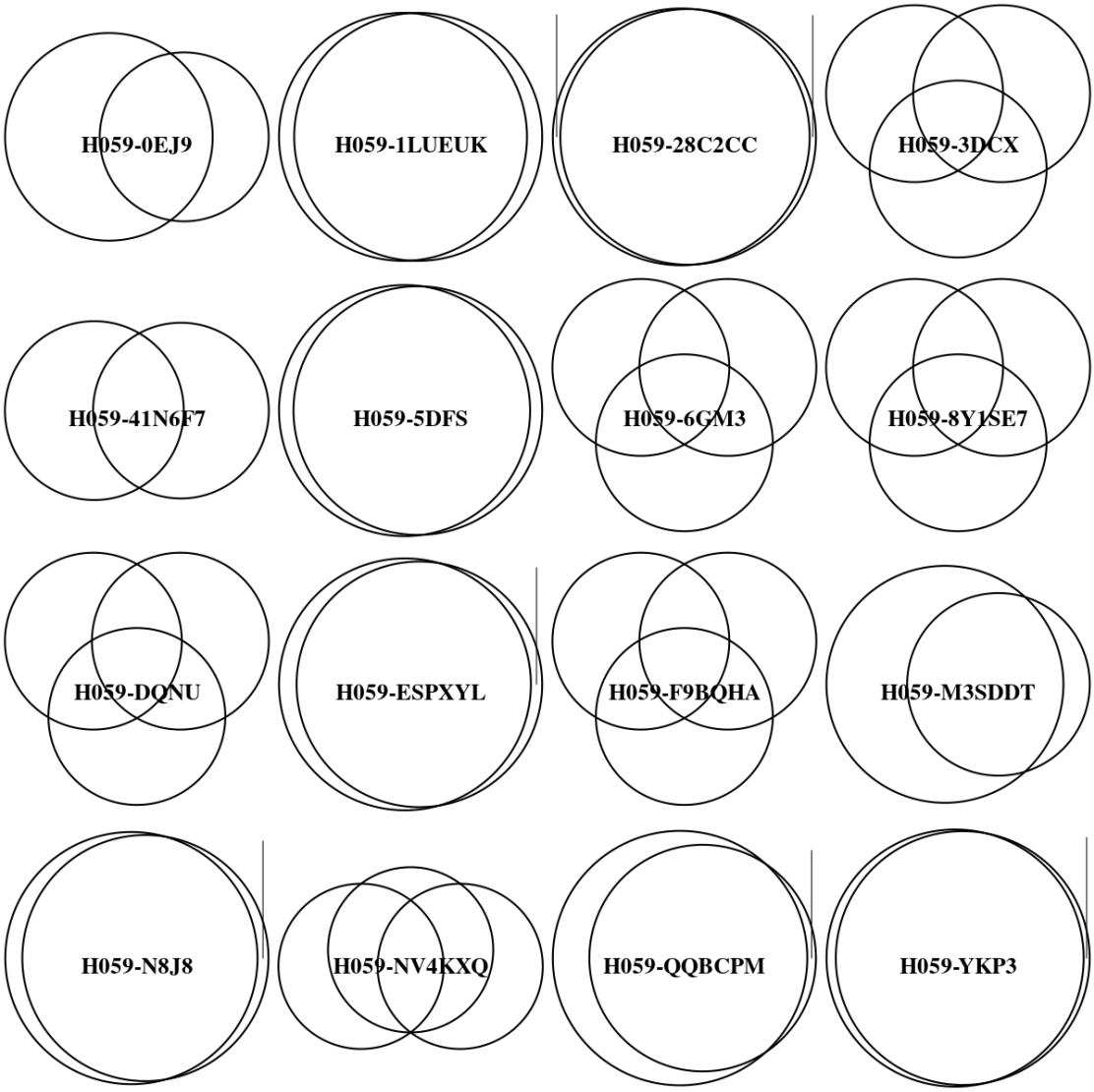


Figure 8.1: Comparison of small variant callsets detected in multiple samples of the same patient. The comparison between related samples from 16 HGSC patients are visualized with venn diagrams.

8.9.3 *Phylogenetic analysis of small variants*

Given multiple samples from an individual patient, the phylogenetic reconstruction is done based on their somatic functional mutations and indels. These small variants were converted into binary calls for the analysis. To find the most parsimonious tree, Dollo parsimony method and the "branch and bound" exhaustive search were implemented using the Dolpenny program in the PHYLIP (Phylogeny Inference Package) software version 3.6[198].

The resulting phylogenetic tree was rooted at an ancestral state 0, meaning the germline genomic sequence. Each branch length was proportional to the number of small variants acquired.

8.10 TIMING OF DRIVING EVENTS IN TUMOR EVOLUTION

Driving events were timed using the supplementary code[199] released along with the PCAWG Heterogeneity study[200], herein referred to as **Gerstung's tutorial**. It describes the process of identifying tumors with WGD, as well as determining the timing of small variants, CN gains, WGD and MRCA. Following this tutorial, MutationTime algorithm was applied to 71 samples in HIPO59 that yielded successful result. The code can be found in the Compact disc along with the thesis (**Code** Directory).

8.10.1 *Whole genome duplication (WGD) classification*

Section 7.2 and section 9 in Gerstung's tutorial described the WGD identification and the assessment of timing concordance.

By profiling the ploidy and homozygosity fraction of the genome, Figure 8.2 shows that tumors could be classified into WGD-positive and WGD-negative tumors in a two-dimensional space.

In addition, the timing of CN gains were examined to assess whether they arose from a single event. The co-occurrence of gains allowed further classification of these tumors into three patterns, namely synchronous, asynchronous and uninformative. In **synchronous** tumors, more than 75% of the timed CN gains are co-occurring, whereas tumors with lower percentage were classified as **asynchronous**. The **uninformative** cases were those having too wide (>0.5) confidence intervals of time estimates, or with only two or less chromosomes timed.

8.10.2 *Timing of small variants and copy number gains*

Gerstung *et al.* developed the **MutationTime** method for inferring whether small variants occurred early or late, as well as to estimate the timing of CN gains. The fundamental concept of this method is illustrated in Figure 1a of the original publication.

In brief, when a CN gain occurred, it doubles the mutant allele copy number of pre-existing (early) somatic variants on the duplicated region. Whereas mutations

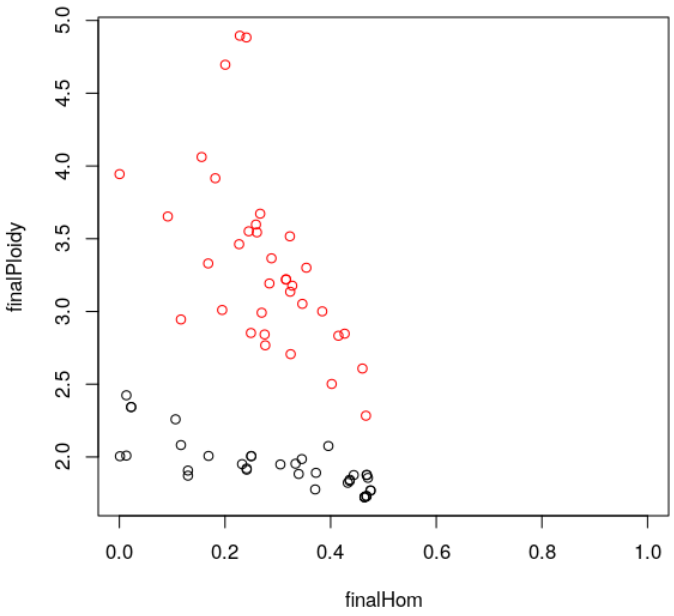


Figure 8.2: Whole genome duplication identification. In the two-dimensional space, fraction of genome with minor allele copy number equals to zero (finalHom) is profiled in the x-axis and the tumor ploidy (finalPloidy) profiled in the y-axis. Tumors were segregated into WGD-positive (red) and WGD-negative (black) tumors.

acquired late (after the CN gain), or mutations on the non-duplicated region, would remain single copy. Therefore, the method first classify variants into different categories, and then use the relative amount of duplicated mutations to non-duplicated mutations to provide the information of the timing of the CN gain.

8.10.2.1 *The concept of mutation time coordinate*

The inferred time estimate is in **mutation time** coordinate, which spans from zero mutation to the number of somatic mutations detected in the tumor. As illustrated in Figure 8.3, it is a quantification of tumor development process using the number of accumulated mutations from age zero (zygote) to the patient's age at diagnosis. As the tumor mutation burden is different between samples, mutation time is usually expressed as a percentage of all detected mutations. Therefore, when a CN gain is said to had happened at 50% in the mutation time, it means that by the time the gain happened, the number of somatic clonal mutations accumulated in the tumor amount to 50% of that observed at diagnosis.

8.10.2.2 *Timing of small variants*

To implement this idea, the authors first defined four categories of variants based on their clonality and co-amplification status, as described in Table 8.7. These categories correspond to different epochs of tumor evolution and are illustrated in mutation time coordinate in Figure 8.4, where **clonal (unspecified)** represents general clonal mutations before MRCA, **clonal (early)** stands for variants occurring before WGD,

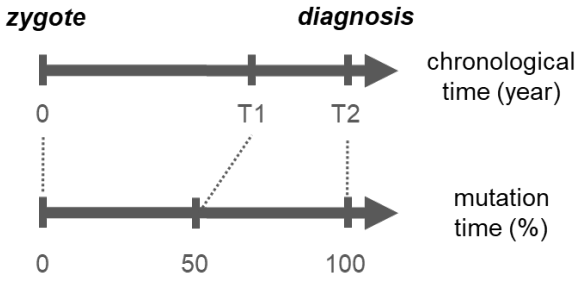


Figure 8.3: Mapping between chronological clock (upper coordinate) to molecular clock (lower coordinate). To start with, it is assumed that at diagnosis (patient age of T2) there were N somatic mutations detected in the tumor sample. At patient age 0 and mutation time 0%, the genome composition was at zygote state, where no somatic mutations were acquired yet. Patient age of T2, on the other hand, corresponds to mutation time 100%, meaning that all N mutations already developed in the tumor genome. Given that patient age of T1 corresponds to mutation time 50%, at patient age of T1 there were N/2 (50% of N) mutations accumulated in the tumor.

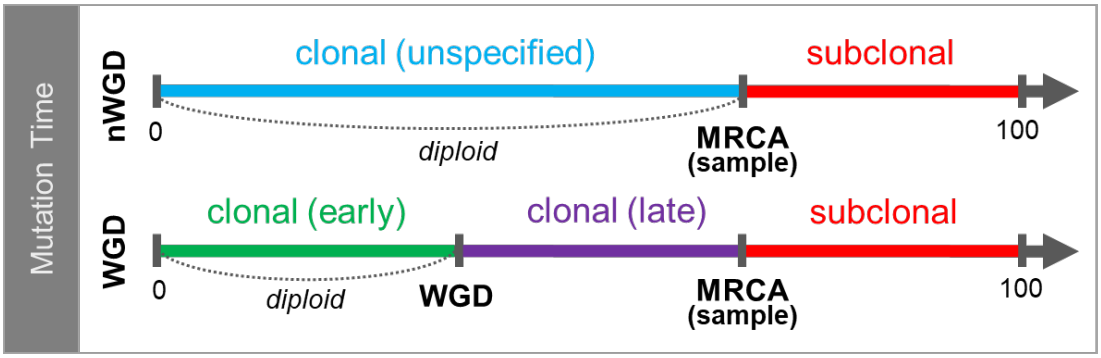


Figure 8.4: Timing of somatic variants in tumor evolution. The upper panel depicts the evolution of a WGD-negative tumor, where clonal variants cannot be further stratified. The lower panel shows the evolution of a WGD-positive tumor, where some clonal variants can be further classified into early or late depend on their relative timing to WGD occurrence.

clonal (late) are clonal variants acquired after WGD and **subclonal** variants being those appear after MRCA of the sample.

Subsequently, the timing of a CN gain was determined based on its associated clonal mutations in different categories, yielding an estimate in the mutation time coordinate.

8.10.2.3 *Timing of CN gains*

Following the fundamental idea, it is expected that when CN gain occurred early, there would be little co-amplified clonal mutations; whereas if the CN gain happened late, many duplicated mutations would be observed.

Given the categorized mutations in Table 8.7, Gerstung *et al.* calculated the time estimate of a CN gain as follows. In a duplicated genomic region, assuming that there

Categories	Definition
early clonal	two mutant copies
late clonal	one mutant copy, no retained allele
(unspecified) clonal	one mutant copy, either on duplicated or retained allele
subclonal	less than one mutant copy

Table 8.7: Somatic variants are stratified into four timing categories based on mutation copy number and local allele-specific copy number profile.

were n_1 mutations with single copy and n_2 mutations with two copies observed, the timing of the gain depends on its local allele-specific copy number configuration:

$$CN\ 2+1 : T = 3n_2/(2n_2+n_1)$$

$$CN\ 2+2 : T = 2n_2/(2n_2+n_1)$$

$$CN\ 2+0 : T = 2n_2/(2n_2+n_1)$$

in which $CN\ C_M + C_m$ refers to major (C_M) and minor (C_m) copy number of the segment.

8.10.3 *Chronological timing of major events*

Given the experimental setting of HIPO59, three events of particular interest are whole genome duplication (`WGD`), most recent common ancestor in the sample (`MRCA-SAMPLE`) and most recent common ancestor between samples (`MRCA-PID`). This section aims to identify their timing in chronological coordinate. To do so, one first has to identify their timing in mutation time and then do a mapping between the two coordinate systems.

8.10.3.1 *Mapping between two timing coordinates*

In Figure 8.3, the two coordinates are consistent in determining relative order of event timing; however there is not a linear relationship for direct transformation. This is mainly due to the time-dependent activity of different mutational processes operating in the tumor, which collectively lead to a heterogeneous mutation rate at different chronological time during tumor development.

One widely accepted approach is to construct the mutation time coordinate using only mutations generated by mutational process that is known to develop proportionally with chronological time, a so-called clock-like process. COSMICvs SBS Signature 1 is one such example[25] and it predominantly generates CpG>TpG mutations in the tumor genome. Nonetheless, despite the clock-like nature, Gerstung *et al.* showed in tutorial section 10.2 that an acceleration of CpG>TpG mutation accumulation was observed when comparing relapse samples and primary samples. Specifically, a 3.5 to 8 times acceleration was observed in 9 relapse ovarian cancer samples. In Gerstung's

tutorial, this rate acceleration is modeled by assuming that starting from mutation time **T** (%) the mutation rate increased by **A** times.

As a summary, mutation time built with CpG>TpG mutations can be mapped to chronological time when mutation rate acceleration is taken into consideration. Two parameters needed for the mapping are time of acceleration (T) and degree of acceleration (A).

8.10.3.2 *Timing of WGD and MRCA in real-world time*

Starting from WGD, a catastrophic event that happened at a single time-point and affected multiple CN gains simultaneously, its mutation time was inferred according to Gerstung's tutorial section 10.5. In short, it is a joint estimate of time-point where most CN gains co-occurred in WGD-positive tumors.

The timing of `MRCA-SAMPLE`, defined as the stage where all clonal mutations emerged, is described in Gerstung's tutorial section 10.4. The authors used the relative branch length of clonal mutations and subclonal mutations to anchor MRCA time-point in mutation time. To control for other factors affecting branch length, both clonal and subclonal branches were adjusted for their power to detect mutations as well as the subclonal phylogeny topology.

Lastly, `MRCA-PID`, defined as the time-point where all shared mutations between all samples were accumulated, can be estimated by its timing relative to `MRCA-SAMPLE`. Specifically, the relative timing can be anchored by the fraction of clonal CpG>TpG mutations shared by all samples over clonal CpG>TpG mutations detected in one sample.

After obtaining the mutation time of these major events, the mapping was done by modeling the mutation generation with two parameters T and A used in the tutorial. Parameter A assumes an acceleration of 7.5 times in the mutation rate, and parameter T profiled the starting point of acceleration across a time period T_{min} and T_{max}:

$$T_{max} = 100\%\ mutation\ time$$

$$T_{min} = max(50\%\ mutation\ time, 15\ years\ before\ diagnosis), when\ age\ is\ known$$

$$T_{min} = 80\%\ mutation\ time, when\ age\ is\ unknown$$

The final chronological time estimate of an event is taken as the median of estimates derived from all parameter combinations. To note, two patients (HO59-9BFZJ8, HO59-MRAP5C) are without age information and therefore the chronological time were not estimated. For interval debulking samples, they were assumed to be taken at the same time with other tumors as the surgery date is unknown.

8.10.4 *Re-define tumor epochs with finer time granularity*

In the PCAWG Heterogeneity study[200], four tumor epochs were defined as described in Figure 8.4 and Table 8.7. These time strata are constructed based on single samplegiven the timepoint of `WGD` or the MRCA of the sample (`MRCA-SAMPLE`). In HIPO59,

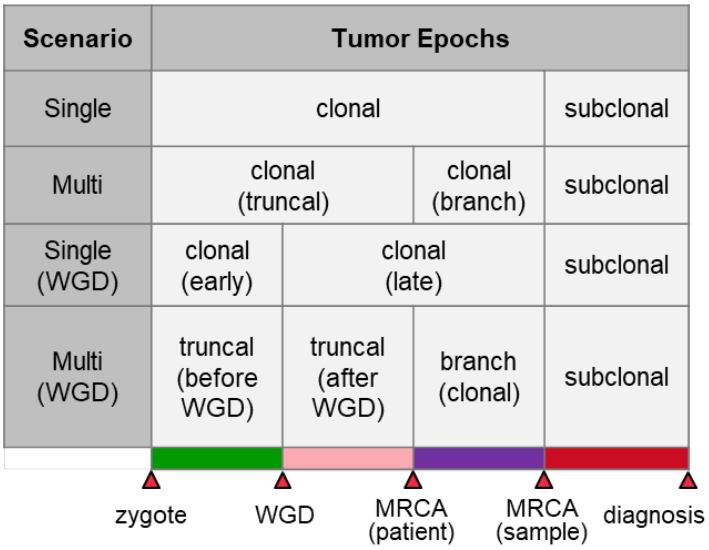


Figure 8.5: Given the multi-sample experimental design, tumor epochs for four scenarios are defined in each row. At the bottom it shows the chronological order of major events in tumor evolution stratifying the different epochs.

a multi-sample cohort, we were able to reconstruct snapshot at an additional timepoint, which is the MRCA of the patient (`MRCA-PID`). This enabled tumor epochs to be re-defined with finer time granularity. The new tumor epochs are shown in Figure 8.5, where four possible scenarios are described for samples being either single sample (Single) or multiple samples (Multi) from the same patient, and either with or without WGD. When compared to the previous definition in Figure 8.4, the "`Single`" scenario would correspond to the previous "nWGD" case, likely the "`Single(WGD)`" scenario corresponds to the previous "WGD" case. In summary, there will be two to four timepoints reconstructed computationally depending on whether multiple samples are available and whether WGD was observed.

8.10.4.1 *Timing of small variants in new tumor epochs*

According to MutationTime result, variants were originally classified into four tumor epochs shown in Figure 8.4. Whereas multi-sample cohorts provides the opportunity to reconstruct the `MRCA-PID` timepoint during tumor evolution. Therefore, small variants identified in multi-sample sets can now re-classified according to Figure 8.5.

In the "`Multi`" scenario, **clonal** variants that are observed in all samples from the same patients were re-classified as **clonal (truncal)**, while clonal variants not shared by all samples are re-classified as **clonal (branch)**.

In the "`Multi (WGD)`" scenario, **clonal [early]** variants are now renamed as **truncal (before WGD)**. For **clonal [late]** variants, those being shared among all samples from the same patient are now classified as **truncal (after WGD)**, whereas those not shared by all samples are re-classified as **branch (clonal)**. Notably, in this scenario, there can be two different models depend on whether `WGD` occurred before or after `MRCA-PID`. As will be mentioned in Section 13.3, most tumors follows the model where `WGD` happen before `MRCA-PID`, therefore the other model is not discussed in particular.

Part III

RESULTS

9 OVERVIEW OF OVARIAN CANCER

9.1 OVARIAN CANCER COMPARED WITH OTHER CANCER TYPES

Ovarian cancer (OC) has been reported to have extensive genome instability and highly rearranged genome, in contrast with its lack of recurrent mutations except for the ubiquitous TP53 aberrations. These global characteristics have not been systematically discussed in the context of other cancer types despite the extensive effort in describing this disease.

The TCGA program profiled tumors across a broad range of diseases and provides great materials for examining similarities and differences between different tumor types. In this section, three global genomic measures are compared across 12 cancer types in order to see how OC is different from other cancers. These measures, namely *mutation prevalence*, *indel prevalence* and *weighted genome integrity index (wGII)*, were derived by procedures described in Section 8.1. The abbreviation of all cancer types are used throughout this section, and their full study names can be found in Table 6.1.

9.1.1 *Low frequency of significantly mutated genes is a distinct feature of ovarian cancer*

The somatic mutation prevalence, measured by number of mutations per mega base pairs (Mbp), represents the overall rate at which mutation occurs on the genome regardless of its functional consequence. Likewise, the somatic indel prevalence can be calculated in a similar manner. Based on the mutation and indel callsets of the TCGA Pan-Cancer cohort, these rates are estimated and visualized in Figure 9.1.

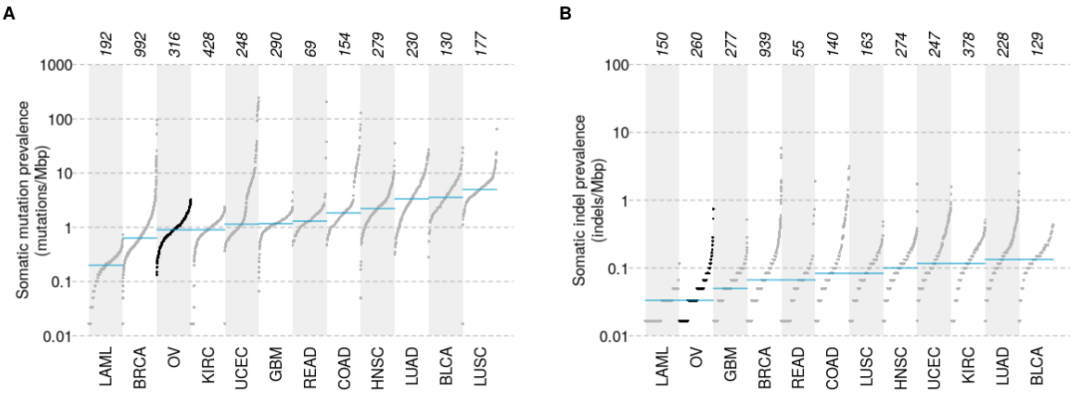


Figure 9.1: Comparison of somatic variant prevalence across TCGA Pan-Cancer cohort. The two global measures being compared are (A) mutation prevalence and (B) indel prevalence, arranged in logarithmic vertical axis. The sample size for each disease type is labeled at the top of the figure.

It appears that ovarian cancer exhibited lower mutation and indel rate when compared to the other cancer types. A consistent trend was observed in previous reports despite that the spectrum of disease type investigated were different. In the first two studies that comprehensively compared mutation burden across cancers, Lawrence *et al.*[173] and Alexandrov *et al.*[24] showed that ovarian cancer genomes were ranked in the middle among all cancer types.

Given the median to low mutability in ovarian cancer as compared to other cancer types, the next question that followed was how the mutated gene recurrence landscape look like, which to some extent reflects how selection pressure acted on the genomes.

To start with, significantly mutated genes (SMGs) in each cancer type were enumerated (see Section 6.1.1) and their frequencies in each disease were estimated based on somatic functional mutations or indels. The frequencies of the top-5 SMGs are shown in Table 9.1.

In addition to independent contribution of individual genes, their cumulative contribution would reflect how these genes collectively account for the disease of interest. By sequentially including SMGs one at a time according to decreasing frequency, their collective contribution would increase until all SMGs are included. Noteworthily, in some disease cohorts the frequency of TP53 is too large to have the influence of other SMGs reflected in the cumulative contribution. For example, TP53 aberrations occurred in 72% of the head and neck cancers (HNSC), 79.2% of the lung squamous carcinomas (LUSC) and 87.3% of the ovarian cancer (OV) cases. Therefore, TP53 was excluded when calculating cumulative contribution for all diseases.

The cumulative contribution of SMGs in TCGA Pan-Cancer Cohort is shown in Figure 9.2. In ovarian cancer, the most prevalent SMG TP53 was not considered and the second most prevalent SMG NF1 affected 4.43% of the cohort. When the third SMG RB1 was included, an additional 2.53% of the cohort was affected, followed by CDK12 (2.85%), NRAS (0.63%) and all other SMGs (0.63%).

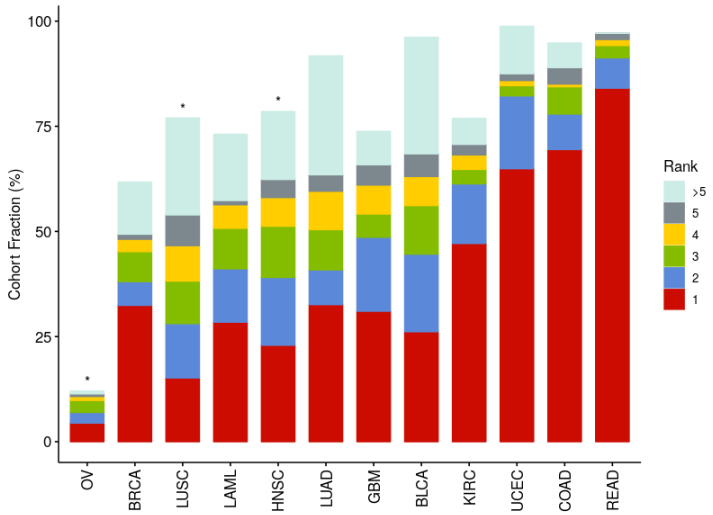


Figure 9.2: Cumulative contribution of SMGs identified by MutSigCV in the TCGA Pan-Cancer cohort. TP53 is excluded when calculating the cumulative contribution, and tumor types with TP53 prevalence >70% are marked with symbol *.

The result showed that, somatic functional variants in SMGs (excluding TP53) collectively affected 12% of the ovarian cancer cohort, as opposed to 61.7% to 98.8% in

Cohort	Number of SMGs	Frequencies of Top-5 SMGs (%)
BLCA	48	TP53 (49.2), ARID1A (26.1), KDM6A (23.9), PIK3CA (20), EP300 (16.1)
BRCA	26	PIK3CA (32.4), TP53 (30.8), CDH1 (11.2), GATA3 (9.8), MAP3K1 (7.2)
COAD	23	APC (69.5), TP53 (48), KRAS (38.3), FBXW7 (18.8), PIK3CA (16.9)
GBM	14	PTEN (31), TP53 (28.6), EGFR (26.6), PIK3R1 (11.4), PIK3CA (11)
HNSC	32	TP53 (72.4), FAT1 (22.9), CDKN2A (22.2), PIK3CA (20.8), NOTCH1 (18.6)
KIRC	12	VHL (47.1), PBRM1 (29.1), SETD2 (9.6), BAP1 (8.2), KDM5C (5.7)
LAML	71	FLT3 (28.4), NPM1 (27.4), DNMT3A (25.9), IDH2 (10.2), IDH1 (9.6)
LUAD	26	TP53 (45.2), KRAS (32.6), COL11A1 (17.4), KEAP1 (17.4), STK11 (17.4)
LUSC	23	TP53 (79.2), NFE2L2 (15.2), PIK3CA (15.2), CDKN2A (14.6), KEAP1 (12.4)
OV	8	TP53 (87.3), NF1 (4.4), CDK12 (2.8), RB1 (2.8), TOP2A (1.9)
READ	11	APC (84.1), TP53 (65.2), KRAS (55.1), FBXW7 (13), SMAD4 (11.6)
UCEC	301	PTEN (64.9), PIK3CA (53.2), ARID1A (33.5), PIK3R1 (33.5), CTNNB1 (29.8)

Table 9.1: Top five significantly mutated genes identified by MutSigCV in TCGA Pan-Cancer cohort.

other cancers. This large discrepancy can not be explained by the mutability of ovarian cancer genomes, which ranked median to low as compared to other cancers.

9.1.2 *High level of chromosomal instability (CIN) and recurrent copy number changes*

The wGII score is used to measure the CIN in a sample. As described in Section 8.1, it is derived from copy number profile and takes values between 0 to 100. The score distributions of all cancer types are shown in Figure 9.3. Ovarian cancer showed the highest level of aneuploidy and with a median score of 22.77, meaning that on average 22.77% of the genomic materials in an autosomal chromosome is subject to copy number alterations. The median wGII for other disease cohorts ranged from 0.29% to 9.7%.

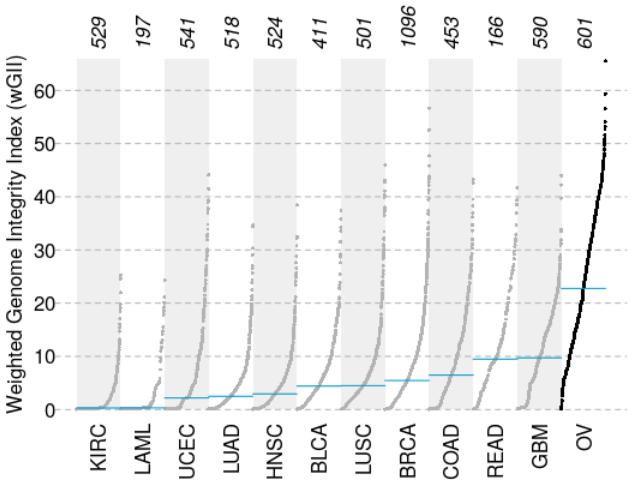


Figure 9.3: Comparison of chromosomal instability across the TCGA Pan-Cancer cohort.

Having shown the high CIN in ovarian cancer, it is interesting to see whether the copy number alterations occurred in a random manner or showed some preferences. The question can be addressed by GISTIC[172] analysis result of the 12 cancer types (see Section 6.1.1), where significantly recurrent somatic copy number alterations in each cohort were identified separately.

The frequencies of individual recurrent peak in corresponding disease cohort were enumerated in Table B.1. Furthermore, the cumulative contribution of the recurrent peaks to a disease cohort, calculated in a similar manner as did for SMGs, was visualized in Figure 9.4.

As a result, 94.8% of the ovarian samples harbored at least one recurrent peak, compared to 10.2% to 85.4% in other cancer types. The independent contribution of the top 3 recurrent peaks in ovarian cancer were 8q24.21 (49.9%), 3q26.2 (44.6%) and 19q12 (26.9%). These regions are where MYC, MECOM, CCNE1 located, separately.

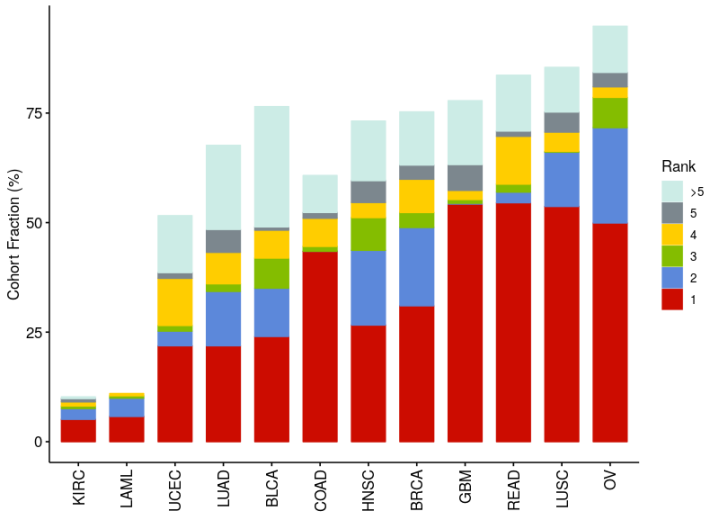


Figure 9.4: Cumulative contribution of significantly recurrent somatic copy number alterations identified by GISTIC in the TCGA Pan-Cancer cohort.

9.2 INDIVIDUAL TYPE OF SOMATIC ALTERATIONS IN HIPO59

To identify positively selected somatic events during tumorigenesis, one can perform cohort analysis that aggregate events from different patients and identify those that happen more often than expected.

Cohort analysis can be done with different alteration types. In Section 9.2.1.1, MutSigCV was used to look for genes recurrently targeted by mutations and indels, whereas Section 9.2.3 showed how GISTIC helped identify recurrent somatic copy number alterations.

As HIPO59 is a multi-sample cohort, where multiple samples are derived from the same patient, there exists intrinsic similarities between related samples. However, the different sampling number per patient would lead to their uneven influence on the cohort analysis result. To minimize this effect, only one representative sample from each HGSC patient were included when conducting MuSigCV and GISTIC analyses. The representative sample of each patient were chosen based on highest purity and are listed in Section B.2.

9.2.1 *Recurrent mutations and indels*

9.2.1.1 *TP53 is the only significantly mutated gene in HIPO59*

Among the 341,047 somatic mutations and indels found in the 33 representative samples genome-wide, MutSigCV included 152,419 events for its analysis. The result indicated that TP53 (FDR=4.19e-11) was the only gene that achieved significant recurrence threshold (FDR<0.15); while the rest of the genes had FDR being 1. The top 10 genes found by MutSigCV are shown in Table B.2.

On the other hand, the TCGA-OV study[130] reported nine SMGs using a threshold of FDR<0.15 in their MutSig analysis. Putting the two cohorts together, Figure 9.5

shows the significance level of these SMGs in TCGA-OV (n=316) as compared to that in HIPO59 (n=33) cohort, as well as the affected population size in HIPO59.

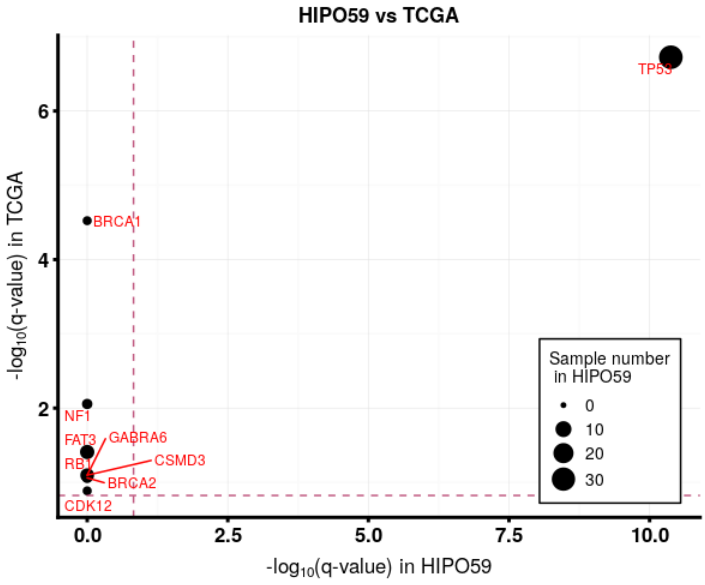


Figure 9.5: Reported significantly mutated genes and their significance in HIPO59 and in TCGA-OV. The x-axis shows the MutSigCV significance of the gene in HIPO59 cohort, while the y-axis shows the MutSig significance of it in the TCGA study. The size of each data point represents the number of patients affected in HIPO59 cohort.

The result suggests that TP53 is the only significantly mutated gene identified in HIPO59, and that some previously reported SMGs did affect a subset of this cohort while the effect is not discernible with the given cohort size.

9.2.1.2 *Reported significantly mutated genes profiled across cohorts*

In this section, mutations and indels from three public cohorts were incorporated to provide a mutational landscape of reported SMGs in HGSC. Four variant categories previously defined in Table 8.1 are used to describe gene perturbations in each patient. Due to data access restrictions, some of these categories may not be available in the public cohorts (see Table 8.2).

Figure 9.6 provides an overview of 12 reported SMGs in 4 major HGSC cohorts, where the frequencies of mutations and indels affecting each SMG were profiled. Genes are ordered by their decreasing contribution to all cohorts. Among them, CSMD3, FAT3 and GABRA6 were reported only in TCGA-OV[130] and on the other hand, DNAH1, FLNA and RABGGTB were reported only in ICGC-AU-OV[131].

In the original report, FAT3 and GABRA6 were suspected to be of lower importance since they showed very low or no expression in tumor and normal samples in the microarray data. By profiling the expression of these genes in RNAseq-based cohorts, Figure 9.7 shows that CSMD3 was also lowly expressed in both tumor (HIPO59 cohort) and normal ovary tissues (GTEx cohort[201]). Furthermore, in a pan-cancer analysis conducted by MutSigCV developers[173], CSMD3 was considered a potentially spurious cancer-associated gene due to its very long intron (>1Mb). The size of CSMD3 and other reported SMGs can be found in Figure B.1.

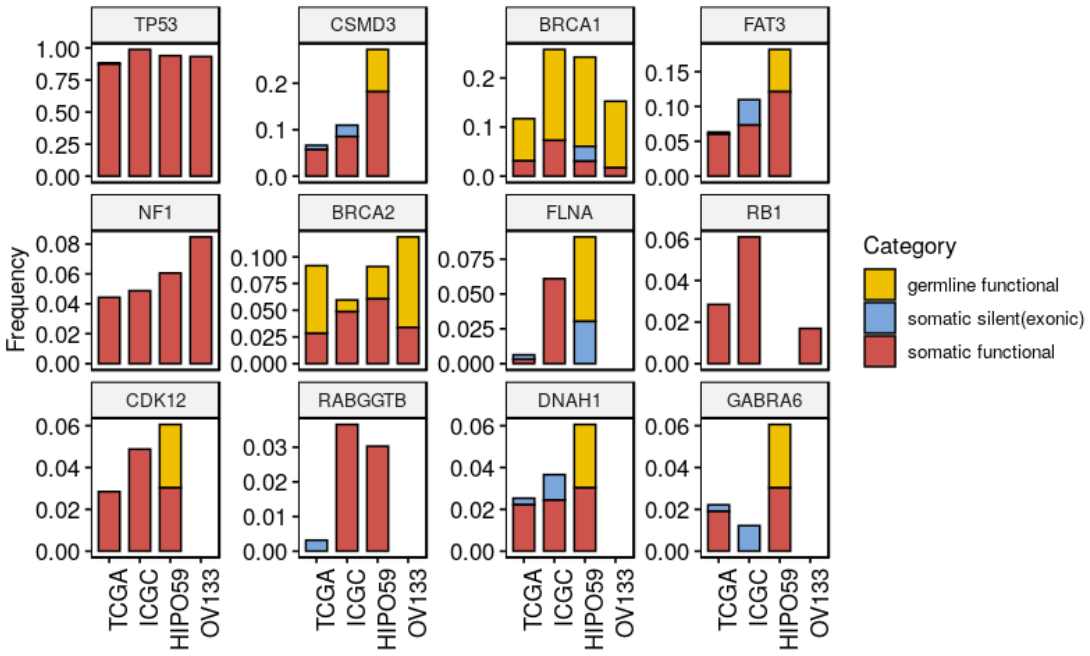


Figure 9.6: Reported significantly mutated genes and their frequency in HIPO59 and three major HGSC cohorts.

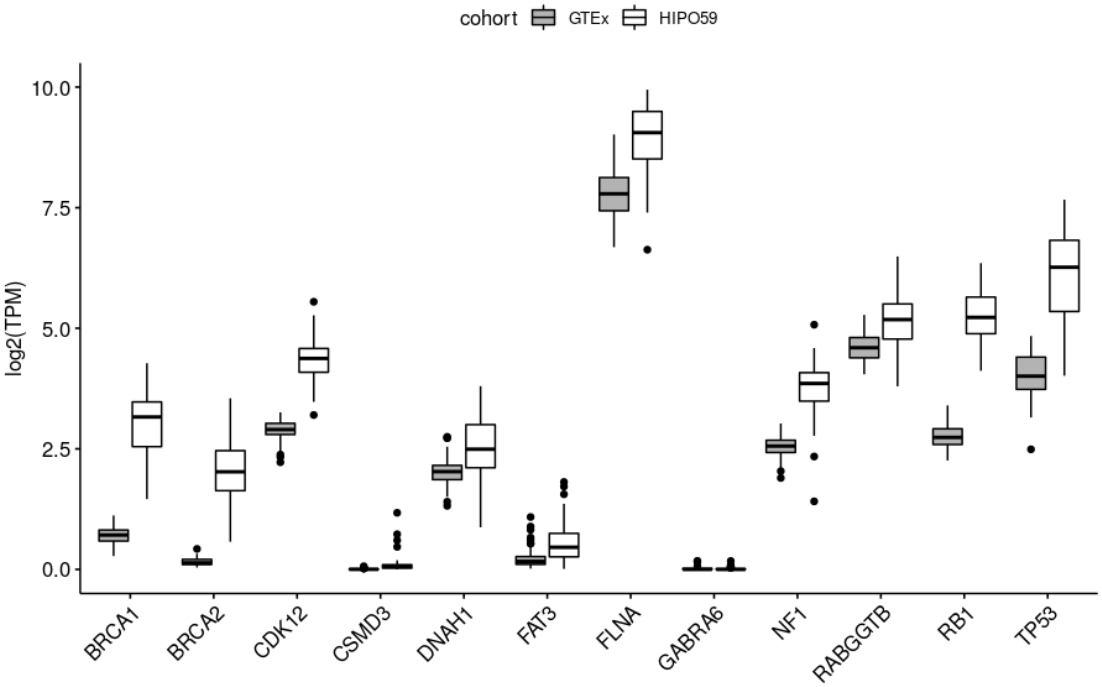


Figure 9.7: Reported significantly mutated genes and their mRNA expression in normal ovary tissues (GTEx) as well as in tumor samples (HIPO59).

9.2.2 *Gene breakage events in reported significantly mutated genes*

The WGS technique allowed researchers in the ICGC study to look into gene breakage as a result of SVs[131]. Focus on 5 SMGs, namely TP53, BRCA1, BRCA2, RB1 and NF1, they found that the inclusion of such events raised the gene inactivation frequency from 6% (mutations only) to 20% for NF1 and 17.5% for RB1. In addition, this mechanism also affected PTEN and RAD51B, where the latter plays a role in the HRR pathway.

To examine such disruption mechanism in HIPO59, SVs were profiled and the exact breakpoint positions were identified in 14 genes of interest, including PTEN, RAD51B and the 12 reported SMGs examined in Section 9.2.1.2.

9.2.2.1 *NF1 is preferably hit by gene breakage events*

A total of 424 SVs were found that involve the 14 genes of interest. Around 43% (184 out of 424) of them do not have any of its breakpoints coincide with copy number changes. The remaining 57% are named here as **CNA-associated SVs**, as when one compares the copy number state between both sides of the breakpoint position, there would be at least one of the two SV breakpoints presenting copy number differences.

In addition to CNA-associated SVs, gene breakage can also arise from any copy number changes within the gene body, in spite of the lack of associated SVs identified. Taken together, Figure 9.8 shows the number of CNA-associated events in each gene. The genes are ordered by decreasing gene length, as shown in Figure 9.9(left). The events are further stratified by how they involve the gene of interest. An "Intra-gene" event is associated with copy number change within the gene body, therefore likely breaks one of the gene alleles. On the other hand, the "Span" events are associated with copy number alterations covering the entire gene and may cause copy number dosage change of the gene without any breakage introduced.

It is expected that the larger the gene is, the more likely it is involved by these events overall, and the higher fraction of them being intra-genic events. As expected, Figure 9.8(left) shows a trend of decreasing event count in smaller genes, while NF1 and RABGGTB being more often hit. When looking at relative fraction of the two disruption types, Figure 9.8(right) further reveals that NF1 and RB1 are more often affected by gene breakage events than by gene dosage changing events.

9.2.2.2 *Gene breakage as an alternative mechanism for gene inactivation*

The next question is to estimate the percentage of the cohort harboring gene breakage events. Collected from all HGSC samples, a total of 136 SVs were found to have any of its breakpoints locate within genes of interest. These SVs were assigned to three categories considering whether it originated from a germline or somatic event, and whether it is CNA-associated. An additional fourth category describes a copy number change in the gene body without any associated SV identified. The definition and priority of the four categories are listed in Table 9.2.

Figure 9.9 shows the fraction of HGSC patients in HIPO59 harboring at least one event in one of their samples. When a patient is found with multiple events targeting

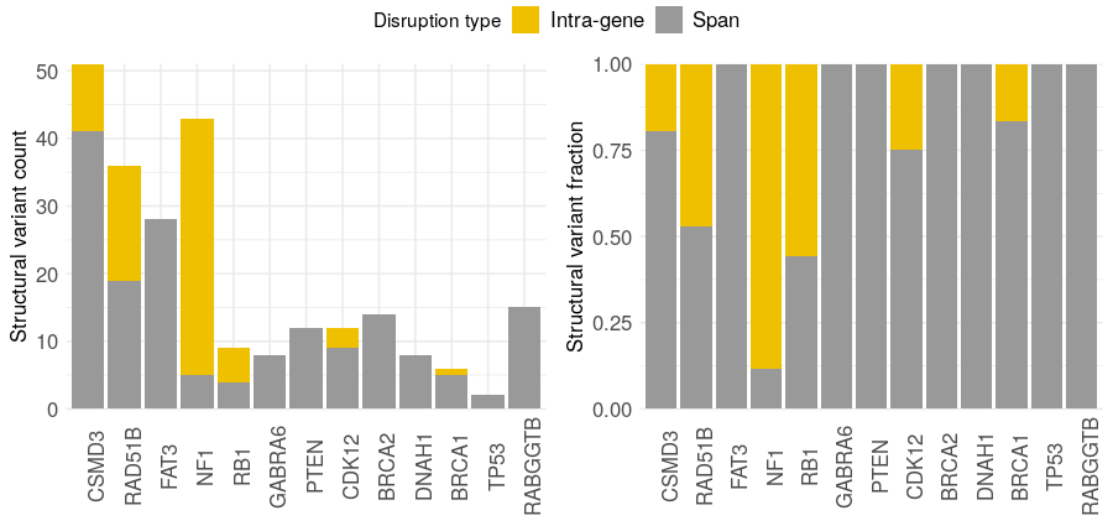


Figure 9.8: Copy number alteration-associated events involving genes of interest. An "intra-gene" event has at least one breakpoint locating within gene body, while a "span" event involves the affected gene as a whole without interrupting the gene body.

Category	Origin	Copy Number Change	Priority
somatic SV (with CNA)	somatic	SVs associated with CNA inside gene body	1
somatic CNA (no SV)	somatic	CNA within gene body, no associated SVs	2
germline SV (w/o CNA)	germline	SVs associated with CNA outside gene body SVs not associated with CNA	3
somatic SV (w/o CNA)	somatic	SVs associated with CNA outside gene body SVs not associated with CNA	4

Table 9.2: Four categories of gene breakage events and their priorities. These events can result from structural variants or copy number changes.

the same gene, the event with the highest priority was chosen. The CNA-associated events, including categories "somatic CNA (no SV)" and "somatic SV (with CNA)", indicate breakage of a gene in one of its alleles. They account for 15% of the cohort for RAD51B, 15% for NF1, 9% for CSMD3, 6% for CDK12, 6% for RB1 and 3% for BRCA1.

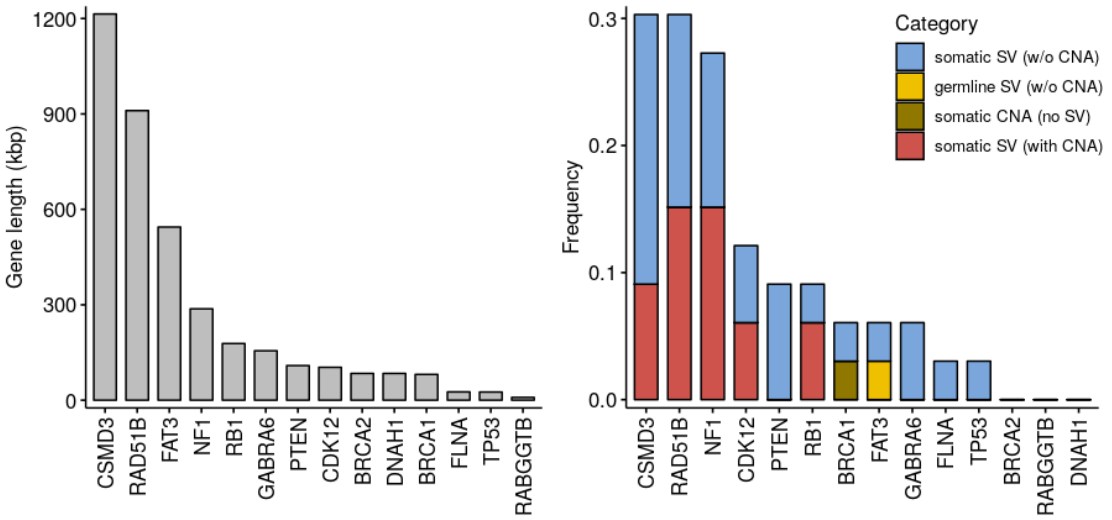


Figure 9.9: Frequency of potential gene breakage due to SVs or other CNA-associated events.

There were in total 49 non-redundant gene breakage events for 33 HGSC patients. Notably, these events are largely exclusive from the occurrence of mutations and indels. When compared to the 68 non-redundant mutation and indels for patients, only 5 co-occurrences were observed. As shown in Figure 9.10, three of them are in CSMD3, one in GABRA6 and one in RAD51B.

9.2.3 *Recurrent copy number changes*

9.2.3.1 *Prevalent chromosome arm-level loss observed in HGSC*

The GISTIC broad event analysis identified recurrent arm-level events across the HIPO59 cohort. As shown in Figure 9.11(A), 28 (61%) of 46 chromosome arms subjected to significantly recurrent alterations (q-value<0.25), and 19 (68%) of them affected more than half of the cohort. Of these, there were notably many more losses than gains, with observed 9 recurrently gained arms and 19 recurrently lost arms.

The result largely agrees with the GISTIC broad event result of TCGA-OV data shown in Figure 9.11(B), where 36 significantly recurrent alterations (q-value<0.25) were identified. The 21 events found in common are amplifications of 1q, 3q, 8q, 12p, 20q, as well as deletions of 4q, 5q, 6q, 8p, 9p, 9q, 11p, 13p, 15q, 16p, 16q, 17p, 17q, 18p, 19p and 22q.

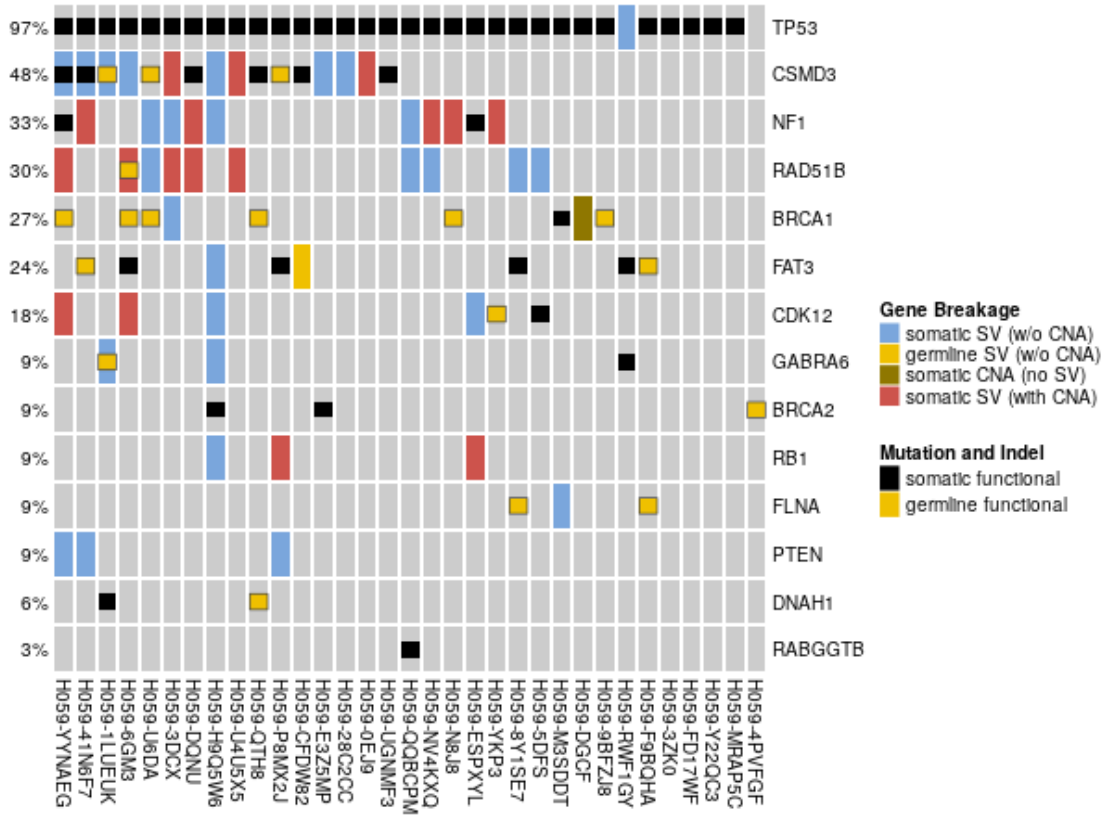


Figure 9.10: Gene inactivation frequencies estimated by integrating mutations, indels and potential gene breakage events for HGSC patients in HIPO59. Note that gene breakage events are represented by taller rectangles; while mutation and indel events are represented by shorter rectangles. Among the four types of potential gene breakage events, "somatic CNA (no SV)" and "somatic SV (with CNA)" likely induced gene breakage; while "somatic SV (w/o CNA)" and "germline SV (w/o CNA)" are candidate events whose consequences need to be further confirmed.

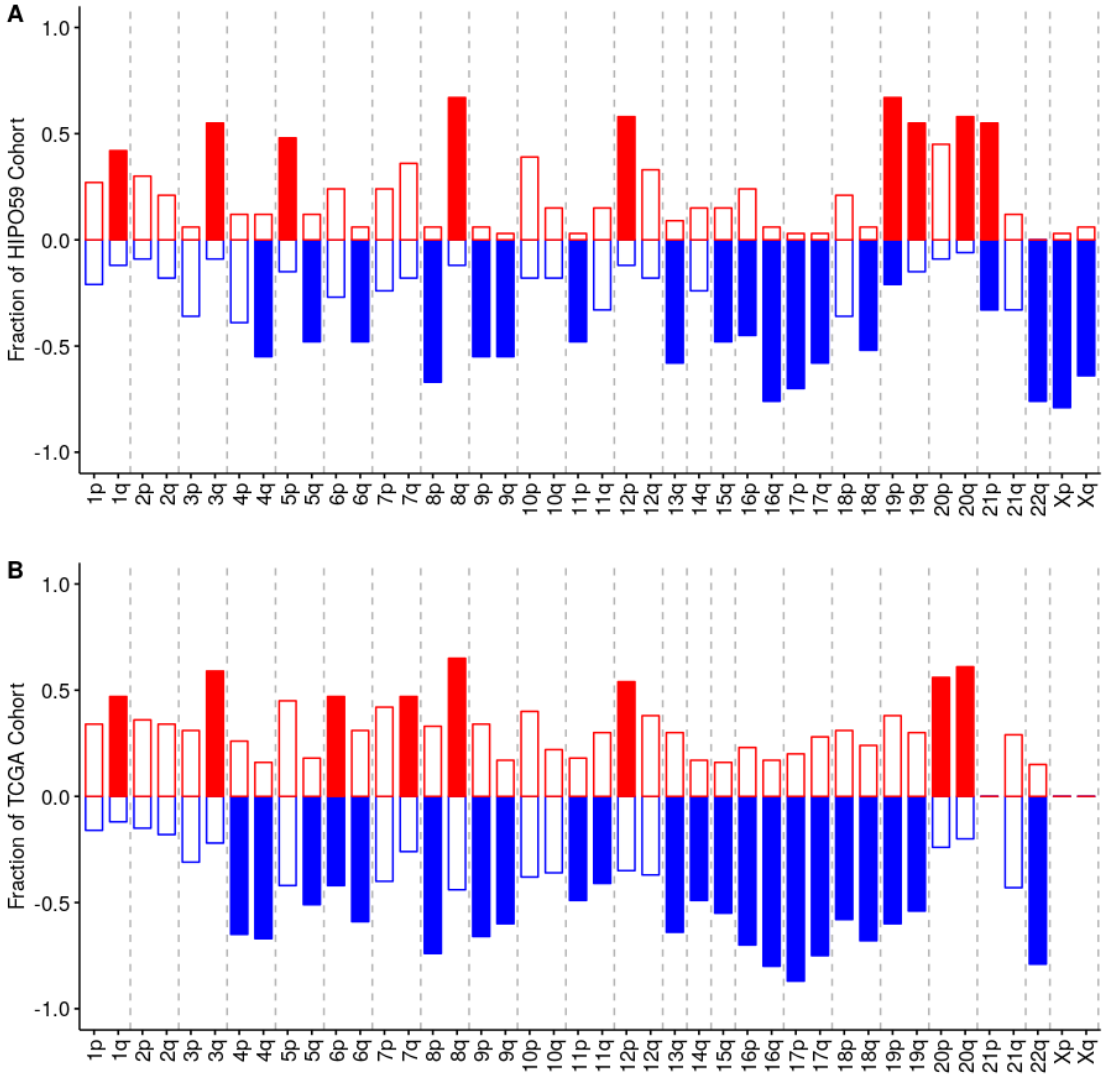


Figure 9.11: Recurrently altered chromosome arms in HIPO59 and TCGA-OV cohorts. For each chromosome arm, the frequency of it being gained or lost in the cohort is shown in the y-axis, where positive value signifies the frequency for gains and negative value for loss. Once the arm-level event is significant (q-value<0.25), the frequency bar is filled with either red(gain) or blue(loss).

9.2.3.2 *Focal copy number alterations in HIPO59*

In addition to arm-level events, GISTIC also reported recurrent focal SCNAs across the HIPO59 cohort. As shown in Figure 9.12, there were in total 21 amplifications and 32 deletions that achieved significance threshold (q-value<0.25), suggesting that alterations in these genomic regions are enriched by selective pressures.

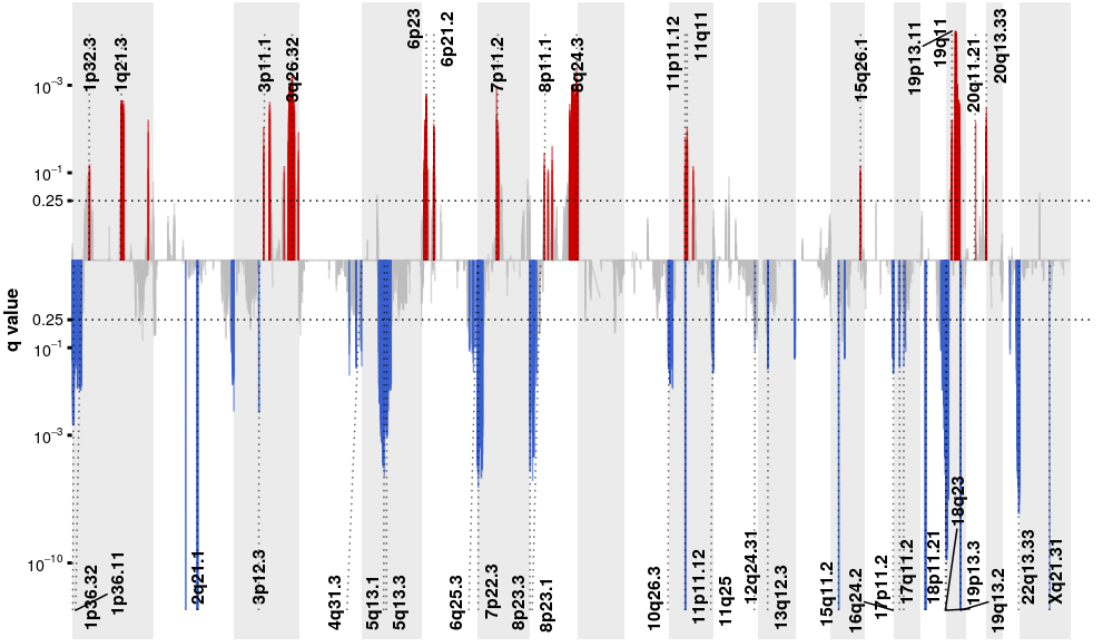


Figure 9.12: Recurrent focal SCNAs in the HIPO59 cohort. The left panel contains deletions and the right panel contains amplifications. These copy number events, each represented by a stretches of DNA, dispersed across the vertical axis from chromosome 1(top) to chromosome X(bottom). The horizontal axis profiled the significance of each event in the scale of log10(q-value), with larger peaks signifying more significant events. Peaks higher than the dashed lines denoted events with q-value less than 0.25. These significant peaks are colored with blue and red, separately, for deletions and amplifications.

For each of the 53 recurrent focal SCNAs identified, GISTIC reports its significance and the genomic coordinates of Regions and Peaks as described in Section 8.4.1. The genomic regions range in size from around 100kb to 50Mb, and cover a median of 14 genes (range from 0 to 275 genes). The number of genes overlapping each CNA is displayed in Figure 9.13.

Following the target nomination procedure in Section 8.4.2, each CNA was annotated with a minimal set of putative gene targets. In short, they are either previously suggested driver genes in the TCGA-OV study('Prior_Target_TCGA'), or being known as cancer-associated genes according to the CGC gene list('Prior_Target_CGC')[178]. Table 9.3 summarizes the 27 recurrent focal SCNAs that contain nominated driver genes. These genes are highlighted differently according to their location within the CNA.

Seven of them confirmed previous findings, including the three most prevalent gains (MYC, MECOM, CCNE1) in TCGA. Nonetheless, in HIPO59 these three genes are in

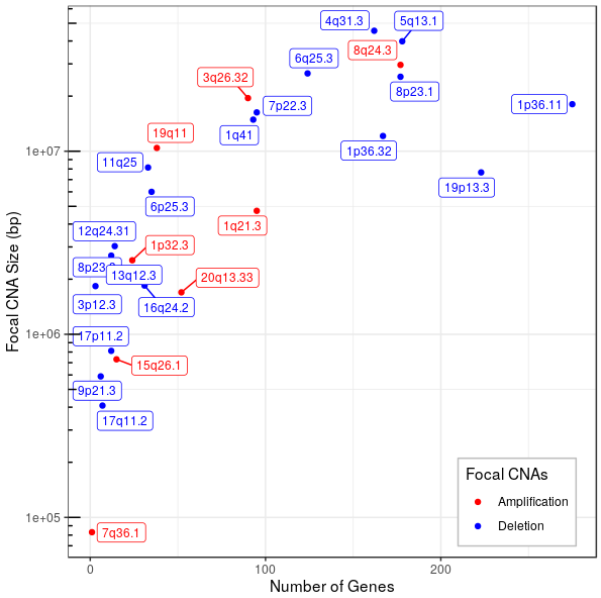


Figure 9.13: The size and affected gene number for each recurrent focal SCNA identified in HIPO59.

significant regions (Region) while outside the region most likely containing true targets (Peak), likewise observed for AURKAIP1. On the other hand, amplification of MCL1, deletion of NF1 and CDKN2A are previously proposed driving events and also sit in the subregion of Peak having the greatest amplitude and prevalence across 33 samples in HIPO59.

These driving events can affect major pathways discussed in Section 4.3. NF1 constrains Ras activity and plays as a negative regulator in RAS/PI3K pathway. In the RB1 Pathway, CCNE1 product leads to Rb inactivation and promote the progression through G1/S checkpoint. Conversely, CDKN2A product inhibit CDK4/6 products and therefore positively regulate this pathway. There are many other pathways implicated in ovarian cancer but not specifically discussed in Section 4.3. MCL1 encodes a member of the Bcl-2 family and has an anti-apoptotic function. Protein product of AURKAIP1 may function as a negative regulator of Aurora kinase A, a protein aiding cytokinesis process and often overexpressed in ovarian cancer[202].

In the 20 other recurrent focal SCNAs, some of the cancer-associated targets seems to be functionally relevant. DDR pathway, particularly genes functioning in the HRR axis (BRCA2 and WRN) were found recurrently deleted. Again, negative regulators of oncogenic pathways were found in focal deletions. In terms of RAS-PI3K pathway, the regulatory subunit of PI3K was encoded by PIK3R1, which was nominated in the 5q13.1 focal deletion. PIK3R1 not only resides in the Peak but also locates at the subregion with strongest selection signal. Regarding the RB1 pathway, CCNE1 product abundance can be controlled by ubiquitin-mediated degradation, and FBXW7 codes for one of the E3 ligases that enable this process. Lastly, components of SWI/SNF chromatin remodeling complex, including ARID1A and ARID1B, have been known linked to EOCs and were also identified in focal deletions 1p36.11 and 6q25.3, separately. Altogether, these findings show good agreement with current active areas of ovarian cancer research.

Type	Peak	Peak Size (kb)	q-value	Region Size (kb)	Genes in Region	Priority Source	Nominated Genes
Amp	19q11	25	7.2e-06	10420	38	TCGA	(CCNE1)
Amp	8q24.3	169	2.6e-04	29618	177	TCGA	(DEPTOR, MYC)
Amp	3q26.32	57	4.1e-04	19506	90	TCGA	(MECOM)
Amp	1q21.3	4275	2.8e-03	4725	95	TCGA	**MCL1**
Amp	20q13.33	10	4.7e-03	1695	52	CGC	(PTK6)
Amp	1p32.3	2537	7.2e-02	2537	24	CGC	CDKN2C, EPS15
Amp	15q26.1	707	7.5e-02	730	15	CGC	CRTC3, IDH2
Amp	7q36.1	78	1.2e-01	83	1	CGC	**KMT2C**
Del	19p13.3	520	1.7e-10	7660	223	CGC	(FSTL3, GNA11, MAP2K2, MLLT1, SH3GL1, STK11, TCF3, VAV1)
Del	7p22.3	500	8.7e-06	16293	95	CGC	(CARD11, ETV1, PMS2, RAC1)
Del	8p23.1	25510	1.8e-05	25510	177	CGC	LEPROTL1, NRG1, PCM1, WRN
Del	5q13.1	1898	2.9e-05	39849	178	CGC	**PIK3R1**, (IL6ST, MAP3K1, RAD17)
Del	8p23.3	1055	4.7e-05	2690	12	CGC	(ARHGEF10)
Del	1p36.32	3137	2.0e-03	12109	167	TCGA	(AURKAIP1)
Del	3p12.3	1833	4.4e-03	1833	3	CGC	ROBO2
Del	1p36.11	18066	2.0e-02	18066	275	CGC	ARHGEF10L, ARID1A, CASP9, ID3, MDS2, MTOR, PAX7, PRDM2, SDHB, SPEN
Del	4q31.3	45496	3.2e-02	45496	162	CGC	CASP3, DUX4, FAT1, FBXW7
Del	16q24.2	1082	3.4e-02	1846	31	CGC	(CBFA2T3)
Del	11q25	8147	3.5e-02	8147	33	CGC	FLI1, KCNJ5
Del	6q25.3	26605	3.7e-02	26605	124	CGC	ARID1B, ESR1, EZR, FGFR1OP, LATS1, MLLT4, QKI
Del	13q12.3	1939	4.3e-02	1939	12	CGC	BRCA2
Del	17p11.2	413	4.4e-02	812	12	CGC	FLCN
Del	17q11.2	408	4.8e-02	408	7	TCGA	**NF1**
Del	12q24.31	3031	9.0e-02	3031	14	CGC	NCOR2
Del	1q41	14875	1.3e-01	14875	93	CGC	ELK4, SLC45A3
Del	6p25.3	6000	2.0e-01	6000	35	CGC	IRF4
Del	9p21.3	589	2.4e-01	589	6	TCGA	**CDKN2A**

Table 9.3: Recurrent focal peaks annotated with putative target genes. Genes in bold are those within the subregion of Peak having the greatest amplitude and frequency across the cohort. Genes in bracket are those within Region but outside the Peak. When the Priority Source is "TCGA", it means the nominated genes are based on `Prior_Target_TCGA` and "CGC" stands for those from `Prior_Target_CGC`.

9.2.3.3 *Robust focal copy number alterations across cohorts*

This section compares recurrent focal SCNAs in HIPO59 with those found in the TCGA-OV study (see Section 6.1.2). As an overview, in HIPO59 there were 6 (11.3%) of 53 focal SCNAs identified in gene desert, and TCGA-OV contained 11 (9.7%) of 113 entries with no overlapping genes. In gene-containing focal SCNAs, an overlap between 24 HIPO59 entries and 27 TCGA entries was observed.

Interested in focal SCNAs harboring potential driver genes, putative target genes were nominated in TCGA entries following the same procedure in Section 8.4.2. Subsequently, the genomic regions of CNAs from HIPO59 and from TCGA were compared in order to identify overlapping **subregion** between each pair. Each subregion was annotated with the q-values in both cohorts. In the case when a CNA does not overlap with any CNAs in the other cohort, q-value was assigned 1 for the other cohort.

Figure 9.14 shows the 103 subregions containing nominated genes in either cohorts. Among them, 76 subregions were found with significant q-values only in TCGA (72) or HIPO59 (4). These private subregions appear as data points lying on the x-axis and the y-axis. Subregions private to HIPO59 usually contain the entire focal SCNA, such as 1q41 deletion, 1p32.3 amplification, 17p11.2 deletion; while sometimes partially cover the focal SCNA, like the subregion containing PCM1 in 8p23.1 deletion. In summary, 20 (74.1%) of 27 recurrent focal SCNAs that contain nominated driver genes were also found recurrent in the TCGA study.

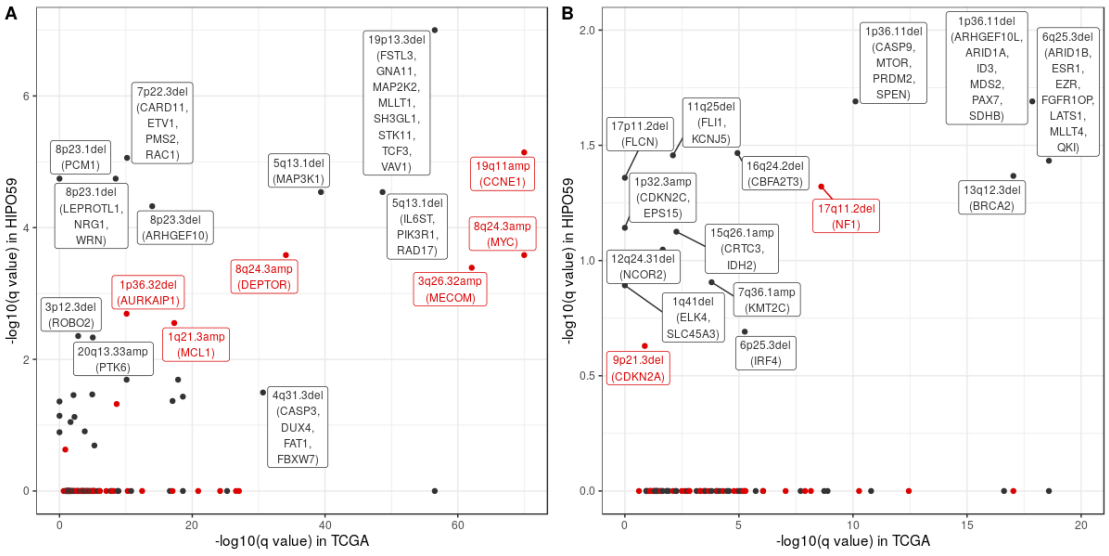


Figure 9.14: Recurrent focal SCNAs compared between HIPO59 and TCGA-OV cohort. The comparison is shown (A) as an overview or (B) zooming in on less significant regions. For each pair of CNA, x-axis and y-axis show the significance levels of its element CNAs from TCGA and HIPO59, respectively. Subregions associated with recurrent focal SCNAs in HIPO59 are labeled with corresponding name in HIPO59 as well as the putative targets in this region. The color of each subregion is based on the nomination priority. When colored in red, the putative targets are nominated based on 'Prior_Target_TCGA', while black color represents a nomination based on 'Prior_Target_CGC'.

10 PATIENT STRATIFICATION

10.1 TANDEM DUPLICATOR PHENOTYPE

This section implements the three-step classification proposed by Dr. Liu's group. It starts with calculating TDP score proposed in their publication in 2016[170], followed by assigning samples to TDP subgroup based on their publication in 2018[169].

10.1.1 *A reliable TDP score implementation*

To validate the in-house TDP score calculation described in Section 8.5.1, a re-analysis was done on a subset of the cohort used in the 2018 study. Specifically, the re-processing focus on 92 samples in the ICGC-AU-OV cohort (see Section 6.1.3), where the SV callset was downloaded from ICGC Data Portal[175] release 28. In-house scores were then compared with the original scores obtained from Table S3 of the publication[169].

First of all, Figure 10.1(A) shows that the number of TDs per sample is similar (Pearson Correlation Coefficient, R>0.99) between what derived from the SV callset and what recorded in Table S3, suggesting the consistency of the SV callset. Secondly, Figure 10.1(B) shows that in-house TDP scores had almost perfect correlation (R>0.99) with the TDP score reported in Table S3, suggesting the reliable implementation in Section 8.5.1.

Next, the TDP scores distribution in three WGS cohorts were compared. These include HIPO59 (n=53, HGSC only) and two ovarian cancer cohorts used in the 2018 study[169], namely, COSMIC_v27 (n=92, ICGC-AU-OV) and Roel_Verhaak_lab (n=49) cohort. Figure 10.2(A) profiled the in-house scores for HIPO59 and the published scores for two other cohorts, and shows that the three distributions differ in their distribution as well as modal values.

As discrepancy in BRCA1 mutation rates was also observed, the distributions might be skewed toward higher scores in different degrees. Therefore, the same scores were profiled again in Figure 10.2(B) while restricted to BRCA1 wild type samples. The HIPO59 cohort was further restricted to the representative samples of each patient (see Section B.2) to exclude bias from multi-sample setting. It shows that the BRCA1 wild type subset of three cohorts then showed distributions with more similar modal values closed to 0, the threshold for defining TDP-positive tumors.

10.1.2 *Subgroup assignment of TDP-positive tumors*

In HIPO59, 38 (72%) of 53 samples with HGSC histotype are TDP-positive tumors. They are then assigned to TDP subgroups according to Section 8.5.1. The full list of TDP assignment for each HGSC sample can be found in the Compact disc along with the

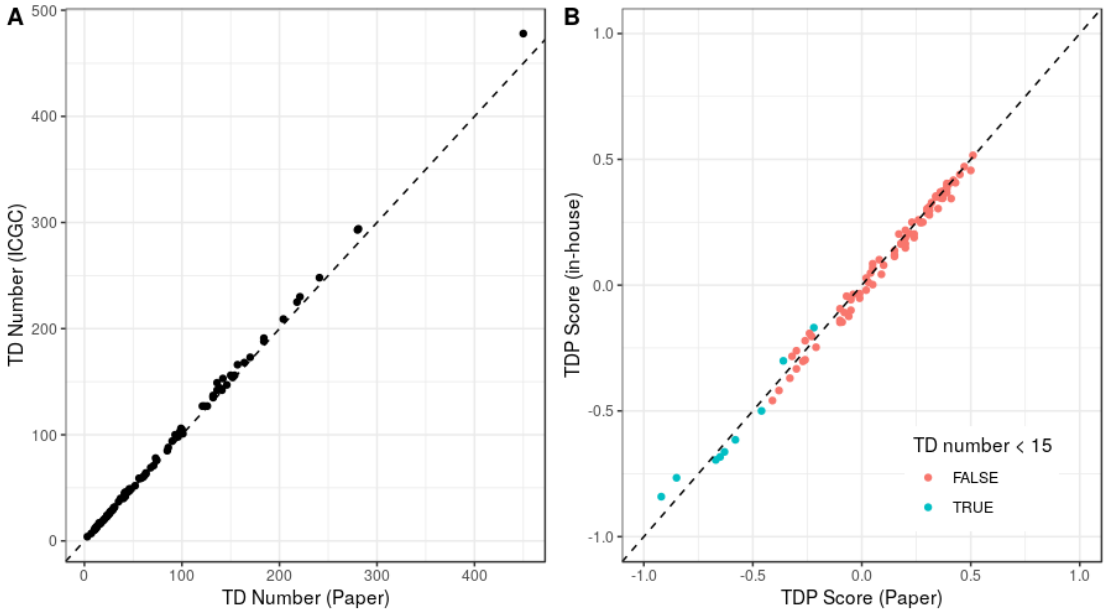


Figure 10.1: The validation was done by comparing (A) the number of TDs and (B) the TDP score between those derived from SV callset and those reported in Table S3 of the 2018 study.

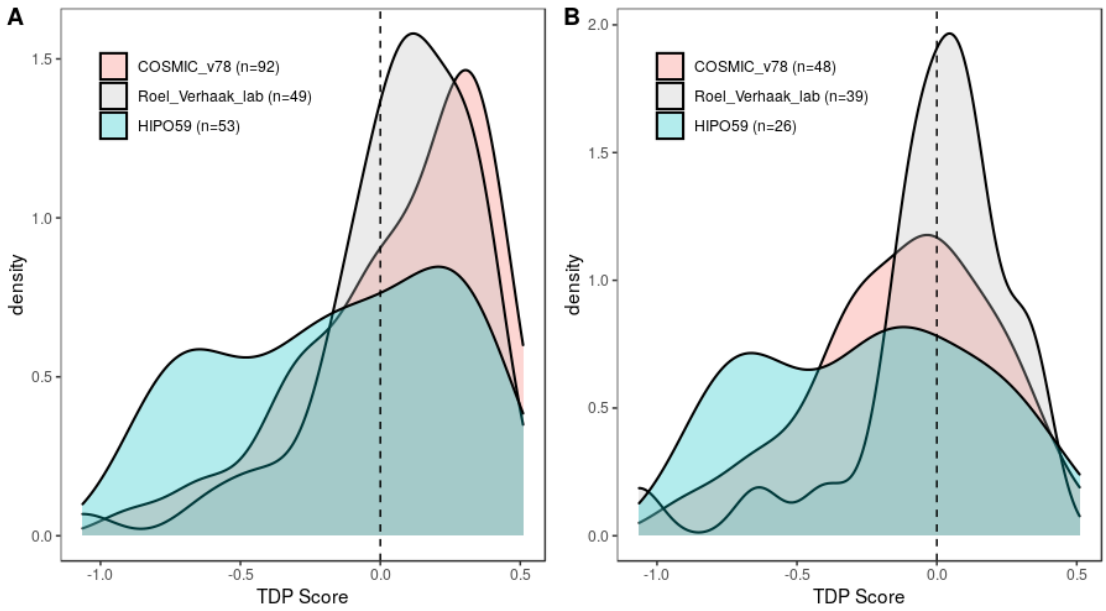


Figure 10.2: TDP score distribution among 3 cohorts. Given the published scores for public cohorts and in-house score for HIPO59 cohort, the distributions from three cohorts are shown (A) for all samples and (B) for BRCA1 wild type and representative samples. The vertical dashed line (at score=0) indicate the threshold for TDP-positive tumors classification.

Patient	TDP Subgroup
H059-1LUEUK,H059-F9BQHA	Non TDP
H059-NV4KXQ	Non TDP,TDP group 1/4mix
H059-0EJ9, **H059-41N6F7**	Non TDP,TDP group 2
H059-N8J8	TDP group 1
H059-6GM3, **H059-QQBCPM**	TDP group 1,TDP group 1/2mix
H059-8Y1SE7	TDP group 1,TDP group 1/3mix
H059-3DCX, H059-5DFS, H059-ESPXYL, H059-YKP3	TDP group 2
H059-28C2CC	TDP group 2/3/4mix,TDP group 3/4mix
H059-DQNU	TDP group 2/3mix

Table 10.1: TDP subgroup heterogeneity in multiple samples from the same patient. Patients with heterogeneous subgroup assignment in corresponding samples are highlighted with bold font.

thesis (**HIPO59_Subtype_TDP.txt**). When further restricted to representative samples of each patient (see Section B.2), the subgroup assignment for HIPO59 is summarized and compared to two public cohorts in Figure 10.3.

In summary, 24 (72.7%) of 33 HGSC patients had at least one TDP-positive sample. By assuming a specific class of TDs being active in a patient when its activation found in any of the samples from the patient, both most prevalent class 1 and class 2 were active in 14 (42.4%) patients, and less frequent class 0, class 3 and class 4 were active in 1, 4, and 5 patients, separately.

To validate the correspondence of known molecular features with each class, alterations in BRCA1 and CDK12 were examined. Seven (50%) of 14 patients with active class 1 TDs had either somatic or germline mutations or indels in BRCA1 and had all their samples showing activation of class 1 TDs. This correspondence was not observed for the two patients (H059-3DCX, H059-DGCF) harboring potential gene breakage events in BRCA1. On the other hand, none of the 6 patients harboring small variants or potential gene breakage events in CDK12 showed active class 3 TDs.

Notably, when there are multiple samples available from the same patient, they don't always show the same subgroup. Table 10.1 listed HGSC patients in HIPO59 and the subgroups found in their corresponding samples. In summary, 7 (47%) of 15 HGSC patients that are with more than one sample showed heterogeneous subgroup phenotype. Among them, 3 patients are found with both TDP-positive and TDP-negative tumors. Interestingly, 5 (71%) of 7 patients have heterogeneous subgroups that differ only by activation of class 2 TDs, suggesting that class 2 TDs show greater variability in different samples from the same patient compared to other classes of TDs.

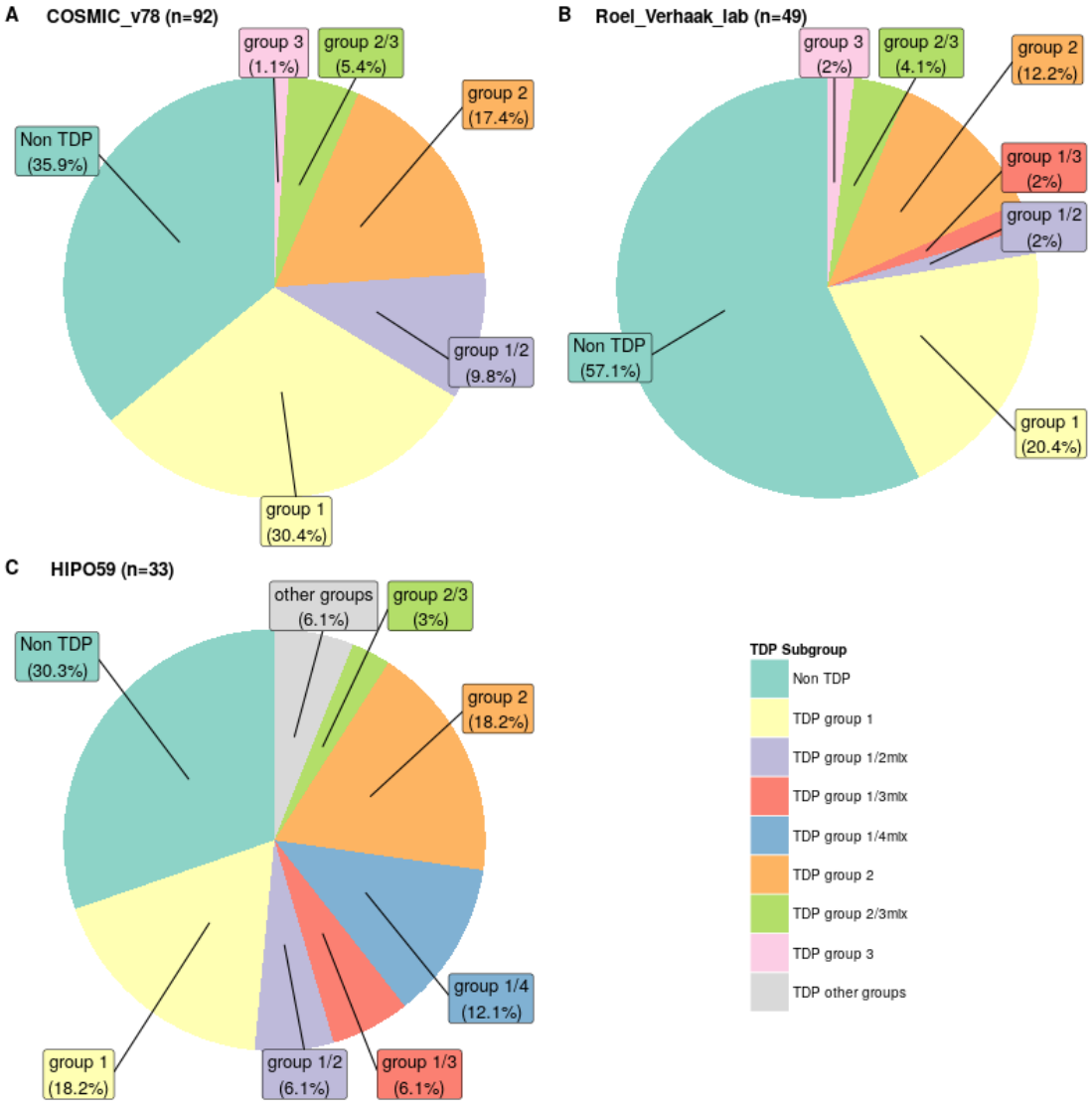


Figure 10.3: TDP subgroup composition in 3 cohorts. Given the published subgroups for public cohorts and in-house assignment for HIPO59 cohort, the subgroup composition of patients in three cohorts are shown for (A) COSMIC_v27 (ICGC), (B) Roel_Verhaak_lab and (C) HIPO59 (HGSC representative samples only). Rare subgroups like "2/3/4mix" or "0/2mix" are put together as "other groups".

10.2 SHAH-2017

Section 4.5.2 described how Wang *et al.* applied unsupervised clustering to their discovery cohort (OV133) and found two subgroups within HGSC histotype. This result is reproducible using the information in Table S5 of the original publication[132].

To see whether the same subgroup structure exists in the HIPO59 cohort, this method is implemented in Section 8.5.2. However, due to the different experimental design and processing pipeline, adjustment was made to the original method during the implementation. This section focus on validation of the adjusted method and the stratification of HIPO59 cohort.

10.2.1 *Adjusted methodology gives same conclusion in the discovery cohort OV133*

There are three major differences in the adjusted methodology. The validation is done by applying adjusted method on OV133 cohort, and see how they might affect the stratification result.

First of all, the entire OV133 cohort encompasses four different major histotypes (see Section 6.1.4), while the majority of HIPO59 cohort are HGSCs (see Section 6.2). Therefore, one cannot rule out the possibility that normalization across a heterogeneous cohort being an influential step to achieve the reported stratification. To address this question, a re-analysis was done on the OV133 cohort while restricting the clustering to only samples of HGSC histotype and originally classified as either H-HRD or H-FBI. The new stratification is shown in Figure B.2, where five samples were classified differently from the original method.

It might seem that the adjusted method yielded misclassified cases. However, survival analysis based on new cluster subgroups suggested the opposite. Figure B.3 compared the survival differences based on two methodologies, the adjusted method unexpectedly identified new subgroups with stronger prognostic value. The result showed that it is reasonable to conduct the cluster discovery based on HGSC subtype only, as in the case of HIPO59.

The second difference lies in the feature set collection. There are twenty genomic features used in the original method. Almost all of them would be possible to be extracted accordingly in the HIPO59 cohort, except for the "Microhomology between SV junctions" and "Balanced Rearrangement" features, due to the different structural variation caller used. Therefore, a re-analysis was done on the HGSC subset of OV133 cohort using only 18 genomic features. In the end there was only one sample classified differently from the original method and the prognostic value of the new subgroup was unchanged.

Lastly, although Table 8.4 implement each feature in a way best match the description in the publication, the feature extraction process is not exactly the same as the authors did to the OV133 cohort. To examine whether these features are compatible, the distributions of these features in the two cohorts were compared. A quantile-quantile plot visualized the comparison for each feature in Figure B.4. There seems to be slightly more foldback inversions and frameshift indels identified in OV133 cohort, other

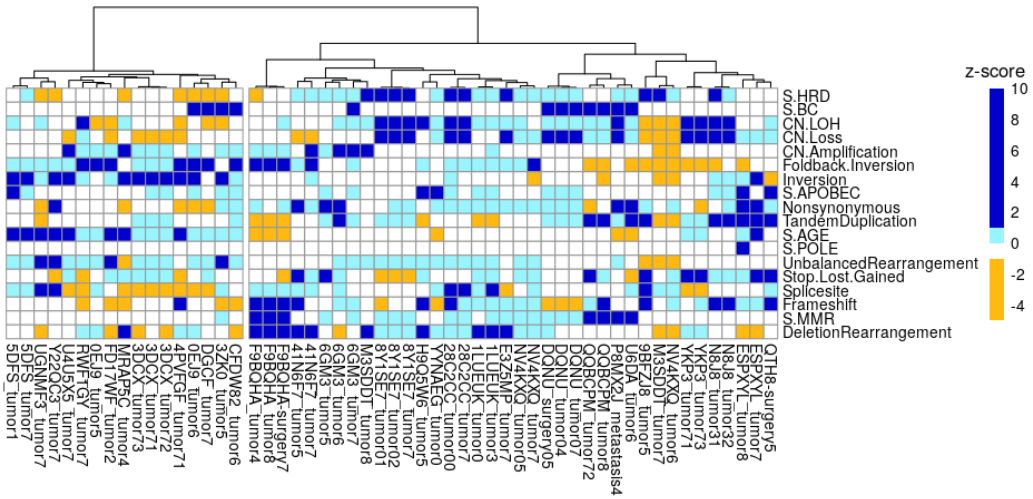


Figure 10.4: Stratify the HIPO59 cohort based on 18 genomic features.

feature distributions either look similar or show differences potentially due to different mutational processes in two cohorts.

Having confirmed that the adjusted method, hereafter referred to as **Shah18**, would not yield markedly difference in stratifying patients in OV133 cohort, this adjusted method was then applied to HIPO59 cohort.

10.2.2 *Shah18 revealed inherent subgroups in HIPO59 cohort*

Based on the 18 selected features extracted according to Section 8.5.2, two clear subgroups emerged in a hierarchical clustering of HGSC samples in HIPO59, as shown in Figure 10.4. Out of the 33 patients, 13 of them were classified as H-FBI group, while the other 20 patients in the H-HRD group. Noteworthily, unlike the heterogeneity seen in other stratification methods, all samples from the same patient were consistently classified into the same subgroup.

To warrant the generalizability of this approach to HIPO59, we would like to see whether the genomic subgroups in two independent cohorts reflect the same inherent structure of HGSC. A dimension reduction method was utilized to address this question. Based on the original 18 genomic features, principal component analysis (PCA) was used to construct new features, the principal components (PCs), where the order of new PCs represent the decreasing degree of their ability to capture the variance in the data.

Figure 10.5 shows that the first PC in both cohorts are both capable of capturing the subgroup information. On top of that, the cosine similarity between the two first PCs was 0.58, showing a high similarity in their compositions.

In summary, the results show that two inherent genomic subgroups of HGSC can be readily revealed by the Shah18 method. More importantly, different samples derived

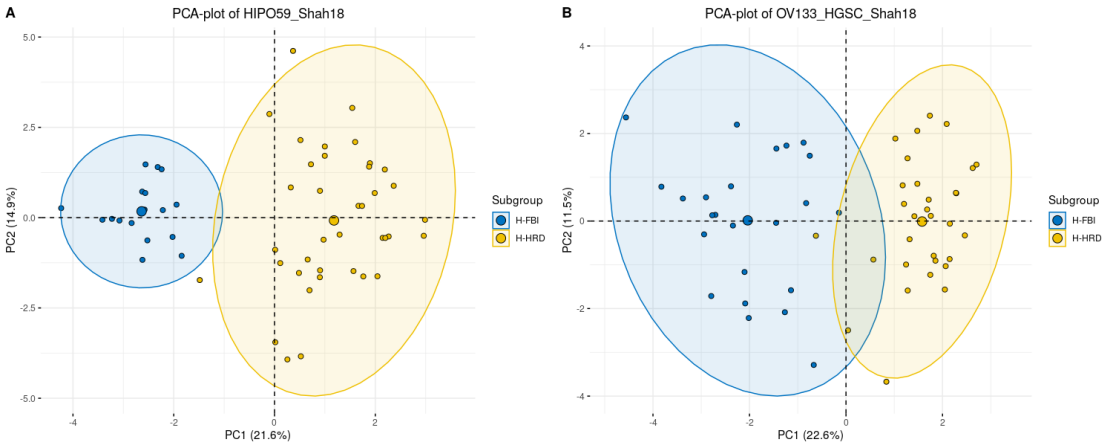


Figure 10.5: PCA of the 18 genomic features in OV133 and HIPO59 cohort, both are restricted to HGSC samples only.

from the same patient exhibit a consistent genomic subgroup. In HIPO59, 13 (39.3%) of 33 patients are with H-FBI subgroup. The prevalence is close to what was observed in the OV133 cohort, where 24 (41%) of 59 patients belongs to H-FBI subgroup.

11 DNA DAMAGE RESPONSE DEFECT IN OVARIAN CANCER

The pivotal role of BRCA cancer susceptibility genes in HGSC has been established after years of studies. Their major impact come with their function in the DNA double-strand break repair via homologous recombination repair (HRR) pathway. When subject to loss-of-function mutations in BRCA genes, tumor cells show BRCAness features and harbor homologous recombination defect (HRD) footprints in their genomes.

In TCGA study the researchers concluded that half of the cohort harbor HRD through either germline BRCA1/2 mutations, hypermethylation of BRCA1 promoter or somatic mutations in selected HRR genes[130], suggesting a genetic heterogeneity under the HRD phenotype. It is therefore important to explore whether other HRR genes are affected in our cohort and what are the consequences.

The three objectives in this chapter are, in the first place, to identify patients that carry inherited pathogenic variants in BRCA genes; secondly, to examine the integrity of DDR axis in both germline and somatic setting; and lastly, to interrogate genomic footprints of DDR defect in a subtype-specific manner.

11.1 GERMLINE PATHOGENIC VARIANTS OCCUR IN NOT ONLY BRCA GENES

It has been repeated reported that a subset of ovarian cancer patients are BRCA mutation carriers. These are high penetrance genes leading to hereditary cancers, with details described in Section 3.2.2.1. Carrier status is important not only to patient themselves but also to their family members who might share the same predisposing variants. In addition to BRCA1 and BRCA2, a handful of other genes are found to predispose their carriers, while with moderate penetrance, to ovarian cancer. Therefore, we seek help from human geneticists to determine the clinical relevance of observed germline variants.

Focused on 36 genes, a collection of 74 unique variants, including 49 germline variants and 25 somatic variants, were described in Section B.5 and sent to human geneticist for review. In principle, these variants represent a prioritized subset that arise from genes functioning in DDR pathway and simultaneously implicated in familial cancers, with the selection procedure described in Section 8.6.1. Determined by the geneticist, the pathogenicity of germline variants are shown in Figure 11.1. Among them, around 24.5% (12 out of 49) are confirmed (likely) pathogenic, 8.2% (4 out of 49) classified as (likely) benign and the rest 67.3% (33 out of 49) being either GUS or VUS.

There are eight patients harboring rare BRCA1 variants and seven of them are confirmed carriers. The only BRCA2 variant is with unknown significance. An addition of seven patients were identified carriers of other DDR genes that would have usually gone unnoticed. When stratified by histotypes, 30% (10 out of 33) of the HGSC and 44% (4 out of 9) of other subtypes carry pathogenic germline variants in DDR pathways.

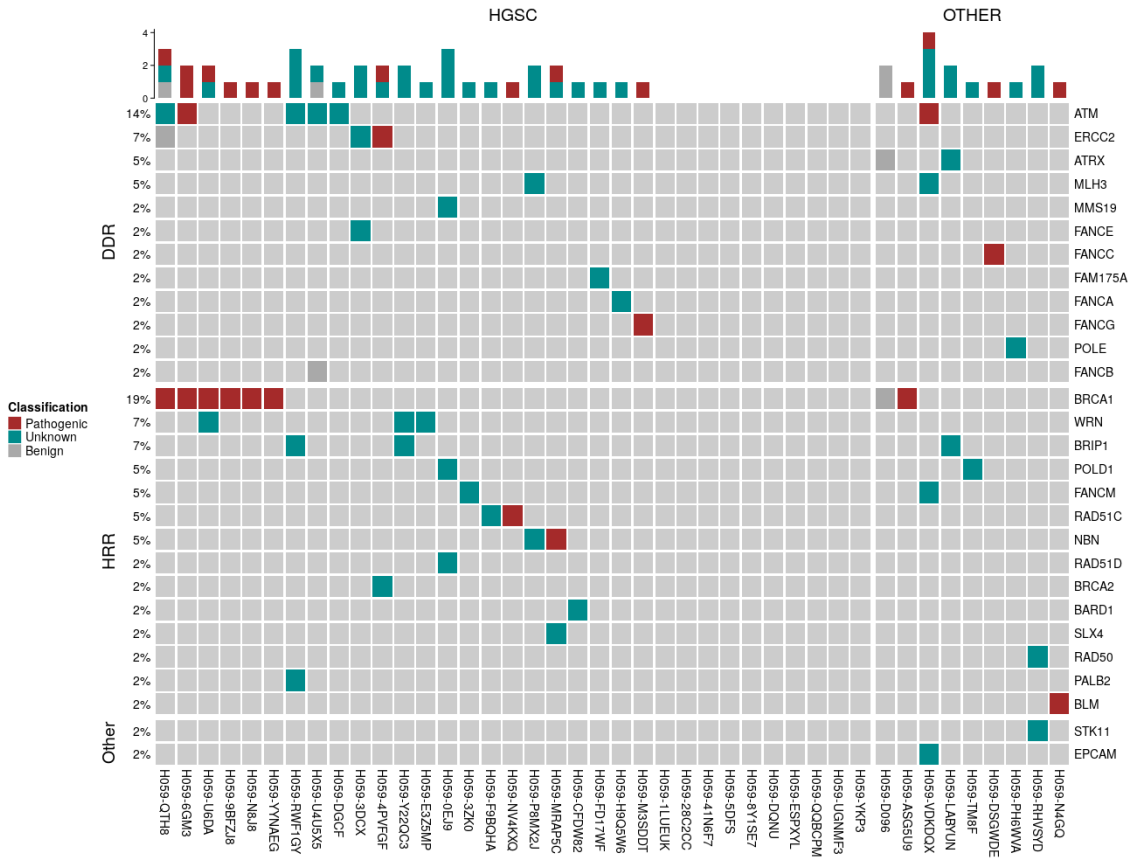


Figure 11.1: Germline variants pathogenicity. Germline variants reviewed by human geneticists and classified in accordance with the ACMG-criteria[192]. Genes harboring (likely) pathogenic variants are colored in red, and those harboring (likely) benign variants are colored in grey, while the rest are with unknown significance (VUS or GUS) and colored in green. The columns represent patients in the HIPO59 cohort and are divided into **HGSC** subtype or all **OTHER** subtypes. The rows correspond to genes and stratified by their function in **HRR** axis, other **DDR** sub-categories, or selected due to **Ohter** reasons.

These include BRCA1, RAD51C, ATM, NBN, ERCC2 and FANCG in HGSC patients and BRCA1, ATM, FANCC and BLM in patients with other histotypes. According to the geneticist's comment, BRCA1 and RAD51C are ovarian cancer predisposing genes associated with an increased risk of 60-69% and 10% respectively. On the other hand, carriers of ATM and NBN are predisposed to *HBOC syndrome* and also conferred with increased breast cancer risk of 27% and 23% respectively; whereas ERCC2 and FANCC are HBOC candidate genes. The last two genes are associated with autosomal recessive diseases, specifically they are FANCG for *Fanconi Anemia* and BLM for *Bloom Syndrome*.

The classification of germline variants can be summarised with a rule that most missense variants are classified as VUS, while the rest tend to be classified as (likely) pathogenic. The only five exceptions include four germline missense variants in BRCA1, ERCC2, FANCB, ATRX being classified as (likely) benign and one germline frameshift insertion in EPCAM classified as VUS. According to geneticist's comment, mutations in

Associated Disease	Evidence	Gene (Variant Count)	HIPO59
Ovarian Cancer, HBOC	predisposition	BRCA2, RAD51C, RAD51D, BRIP1(3)	6 VUS
HBOC	predisposition	PALB2, ATM(4), NBN	6 VUS
HBOC	candidate	RAD50, BARD1, FAM175A, ERCC2, SLX4	5 GUS
colon cancer	predisposition	POLE, POLD1(2)	3 VUS
colon cancer	candidate	MLH3(2)	2 GUS
other hereditary diseases		ATRX, CDH1, FANCA, FANCE, STK11, WRN(3)	8 VUS

Table 11.1: VUS or GUS in disease-associating genes, summarized from geneticist's comments.

the last exons of EPCAM can lead to epigenetic silencing of MSH2, which causes *Lynch Syndrome*. However, the variant observed here is intronic and therefore with unknown significance.

Many of the variants, albeit with unknown significance, occurred in additional genes implicated in ovarian cancer, colon cancer, HBOC as well as other hereditary diseases. Table 11.1 provides an overview of these VUSs. The four VUS or GUS not listed below are one from EPCAM, one from MMS19, and two from FANCM. According to geneticist's comment, MMS19 is without established association with diseases, while FANCM's association with Fanconi Anemia is under debate[203].

In summary, an overall of 33% (14 out of 42) of the HIPO59 cohort are carriers of disease-predisposing genes. Among them, 57.1% (8 out of 14) carry high or moderate penetrance genes for ovarian cancer, 28.6% (4 out of 14) carry pathogenic variants in HBOC predisposition or candidate genes, while 14.3% (2 out of 14) are carriers of other hereditary diseases.

11.2 RARE GERMLINE VARIANTS ARE ENRICHED IN DDR PATHWAYS

In the previous section, hereditary factor was shown to contribute to a significant proportion of the HIPO59 cohort. Variants involved ranged from VUS to high penetrance and the affected functions possibly span the spectrum of different DDR sub-pathways. Nonetheless, we have not yet excluded the possibility that this significant contribution is due to gene selection bias.

The GermFxn call set, comprising 9,056 germline variants in 5,796 genes, was derived from the ClinicalWorkflow as described in Section 7.2. It consists of germline variants that are rare in the population, as well as bearing a potential of changing the expression or the structure of its protein product. The number of rare variants found in each patient ranges from 177 to 331, with a median number of 209. The objective of this section is therefore to perform functional enrichment analysis and see whether these

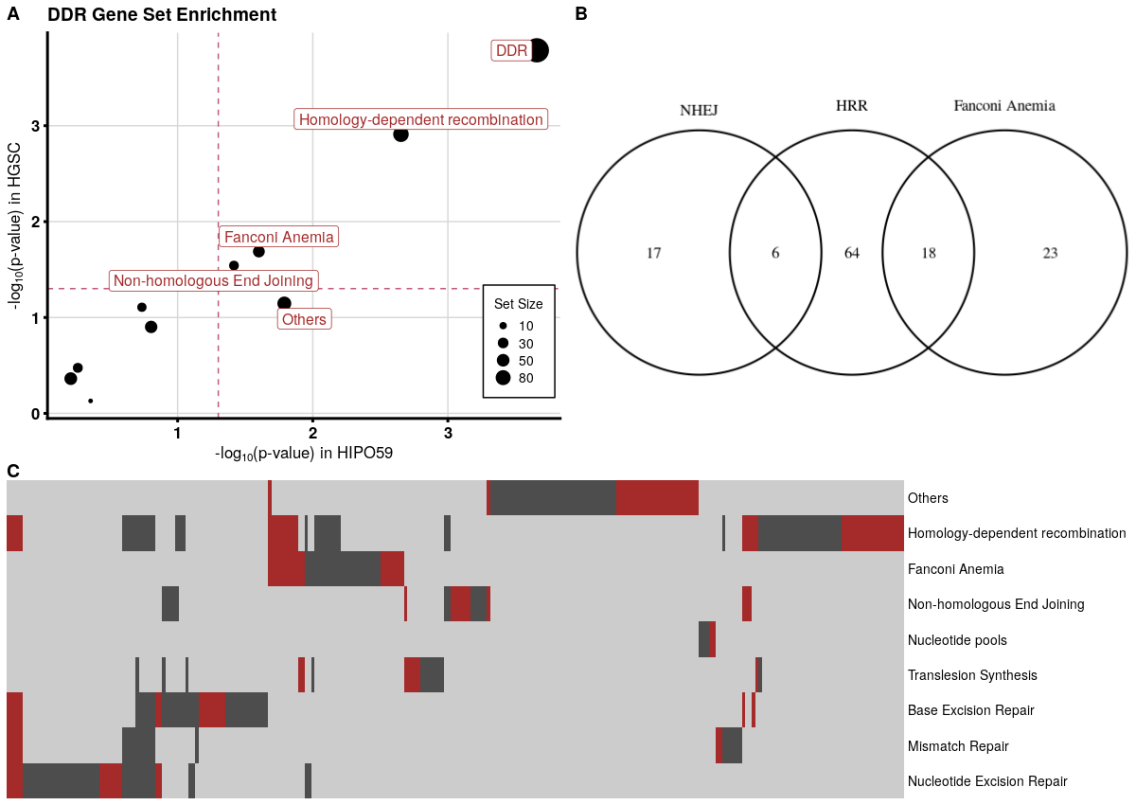


Figure 11.2: Functional enrichment analysis of germline variants. The p-value of ten gene sets in two scenarios (HIPO59 or HGSC) are shown in (A), where gene sets being significant in either scenario are labeled with its name. In (B) a Venn Diagram shows the degree of membership redundancy between significant pathways. A detailed look at the redundancy is shown in (C), where genes are arranged in column and pathways in the rows. The color encoded a gene's membership of the pathways as well as whether germline variants were observed in HIPO59. Specifically, light grey stands for non-members, dark grey stands for members of the pathway without germline variants observed, red stands for pathway members with germline variant observed.

variants preferentially arise from DDR pathway or its nine sub-categories. More details about the analysis is described in Section 8.6.2.

In this analysis, ten gene sets are tested in two scenarios: one using the entire GermFxn call set from **HIPO59** and the other using the subset found only in **HGSC** patients. The p-value of one-sided Fisher's exact test in the two scenarios were listed in Table B.3, and visualized in Figure 11.2(A). In either scenario the germline variants are significantly associated with four gene sets, including the overall DNA Damage Response pathway (DDR, p-value=0.00022) and its three subpathways, namely, Homology-dependent recombination repair (HRR, p-value=0.00223), Fanconi Anemia pathway (p-value=0.0251) and Non-homologous End Joining pathway (NHEJ, p-value=0.0383). The "Others" sub-category (p-value=0.0162) was enriched only when considering the entire cohort.

Of note, there are redundancies in the definition of these gene sets due to repair pathway cross-talks and multi-functional genes. The Venn Diagram in Figure 11.2(B) shows

the member overlap in three significant subpathways. When looking at the gene level, Figure 11.2(C) further showed that some germline variants from multi-functional genes simultaneously contribute to different sub-categories and might therefore contribute to their associations.

Using an unbiased approach, this section showed that there is indeed more germline variants arising from DNA damage response pathway, especially from its sub-categories related to DNA double-strand break repair.

11.3 GERMLINE AND SOMATIC LANDSCAPE OF DDR PATHWAYS

In addition to the aforementioned germline pathogenic variants associated with DDR defect, the objective of this section is to see how often such defect occurs in the somatic setting. Assuming driving events would already exist in tumor founder clone, only truncal somatic events are considered here, with details described in Section 8.7.2.

Figure 11.3 summarizes DDR pathway disruptions due to truncal mutations and indels in each patient. The upper panel displays, in each of the nine DDR sub-categories, number of genes harboring germline variations; while the lower panel displays, in the DDR pathway as well as in its HRR sub-category, number of genes targeted by somatic events. Of note, recurrent driver genes like TP53 and BRCA genes function in different DDR sub-categories (as described in Table 6.2), therefore, their status are specifically annotated at the top of each panel.

In germline setting, rare variants were observed in 1 to 10 DDR pathway member genes, with a median number of 4 genes. This number does not correlate with the total number of germline variants found in each patient (R=0.2, p-value=0.2). As described in the previous section, 33% (14 out of 42) of the patients harbor germline pathogenic variants and half (7 out of 14) of them locate in BRCA1. There was no evidence of a higher rare germline variant burden in patients carrying germline pathogenic variant (Welch t-test, p-value=0.79).

On the other hand, truncal somatic events occurred in 1 to 9 DDR genes, with a median of 2 genes. This number correlates with the Somatic Variant Number found in all related samples (R=0.53, p-value=0.0076). Deleterious variants, classified according to Section 8.7.1, were found in 38% (16 out of 42) of the HIPO59 cohort, where 69% (11 out of 16) of the cases were attributed to BRCA1 (1 patient), BRCA2 (2 patients) and TP53 (8 patients). There was no evidence of a higher somatic variant burden in patients carrying somatic deleterious variant (Welch t-test, p-value=0.57).

When stratified for histotype and genomic subtype, germline pathogenic or truncal somatic deleterious events in DDR pathway gens occur in 80% (16 out of 20) of the HGSC H-HRD subgroup, 46% (6 out of 13) of the HGSC H-FBI subgroup and 66.7% (6 out of 9) of the OTHER subgroup comprising all other histotypes (see Table B.4). When focusing on HGSC histotype, a tendency of enrichment in H-HRD subgroup (OR=4.43, p-value=0.051) was observed.

Mutual exclusivity was also observed in the occurrence of germline pathogenic events and truncal somatic deleterious events (OR=0.174, p-value=0.0253). Moreover, when stratified into genomic subgroups, the mutual exclusivity pattern exists only in the

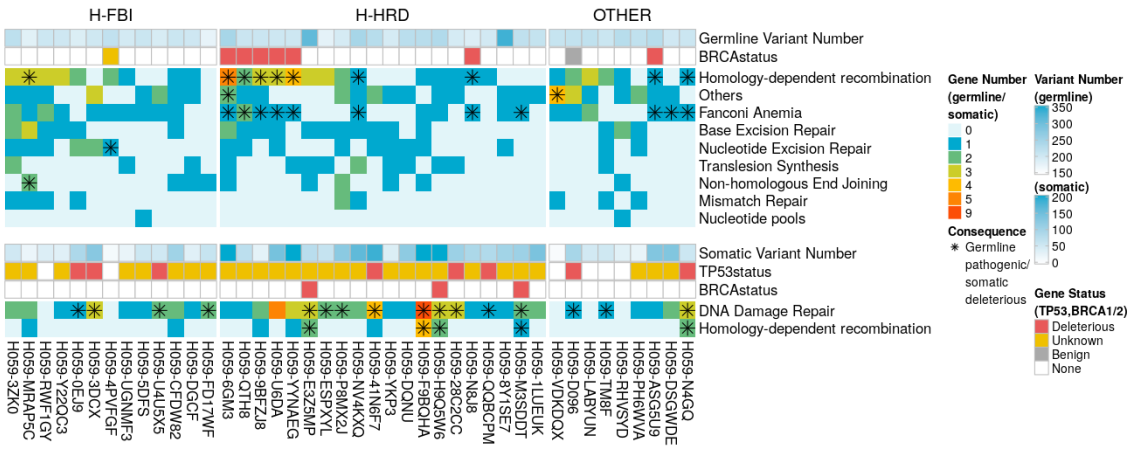


Figure 11.3: DNA damage response disruptions in germline (upper panel) and in somatic (lower panel) settings. Patients are arranged in columns and stratified into three groups, namely the two HGSC genomic subgroups (H-FBI and H-HRD) and all other ovarian cancer histotypes (OTHER). Rows represent pathways and are separated into two panels. Each panel consists of annotations (variant number, driver status) and a matrix displaying number of pathway member genes affected. Whenever a germline pathogenic variant or a truncal somatic deleterious variant exists in any of the pathway member genes, the corresponding matrix cell is labeled with symbol *. Color encoding for the matrices and annotations can be found in the left and right column of the legend, separately.

H-HRD subgroup (OR=0.083, p-value=0.0249) but not in the H-FBI subgroup (OR=0, p-value=0.462).

Collectively, the result suggested that 66.7% (28 out of 42) of the HIPO59 cohort harbor potential driving events in DDR pathway genes, either in the form of germline pathogenic variant and/or truncal somatic deleterious variant. When restricted to HGSC subtype, these events tend to enrich in H-HRD subgroup. Moreover, they occur in a mutually exclusive manner, especially in the H-HRD subgroup.

11.4 ABERRATIONS POTENTIALLY CONTRIBUTING TO DDR DEFECT

In order to provide a comprehensive landscape at the gene level, the DDR pathway is decomposed into its 276 member genes in this section. Out of the 128 member genes that contain germline or somatic variants, Figure 11.4 shows the 42 genes having either deleterious or potential bi-allelic events in any of the samples. Deleterious events, classified according to Section 8.7.1, encompass pathogenic or protein-truncating mutations and indels. Bi-allelic inactivation, on the other hand, potentially results from genes targeted with multiple-hits or LOH.

First of all, potential driving events described in the previous section (see Figure 11.3 and Table B.4) were examined for bi-allelic inactivation.

Driving events in known driver genes, including TP53, BRCA1 and BRCA2, were found in 18 patients and mostly accompanied by LOH. BRCA1 variants were oftentimes observed as germline pathogenic variants and affected 8 patients. Somatic truncal

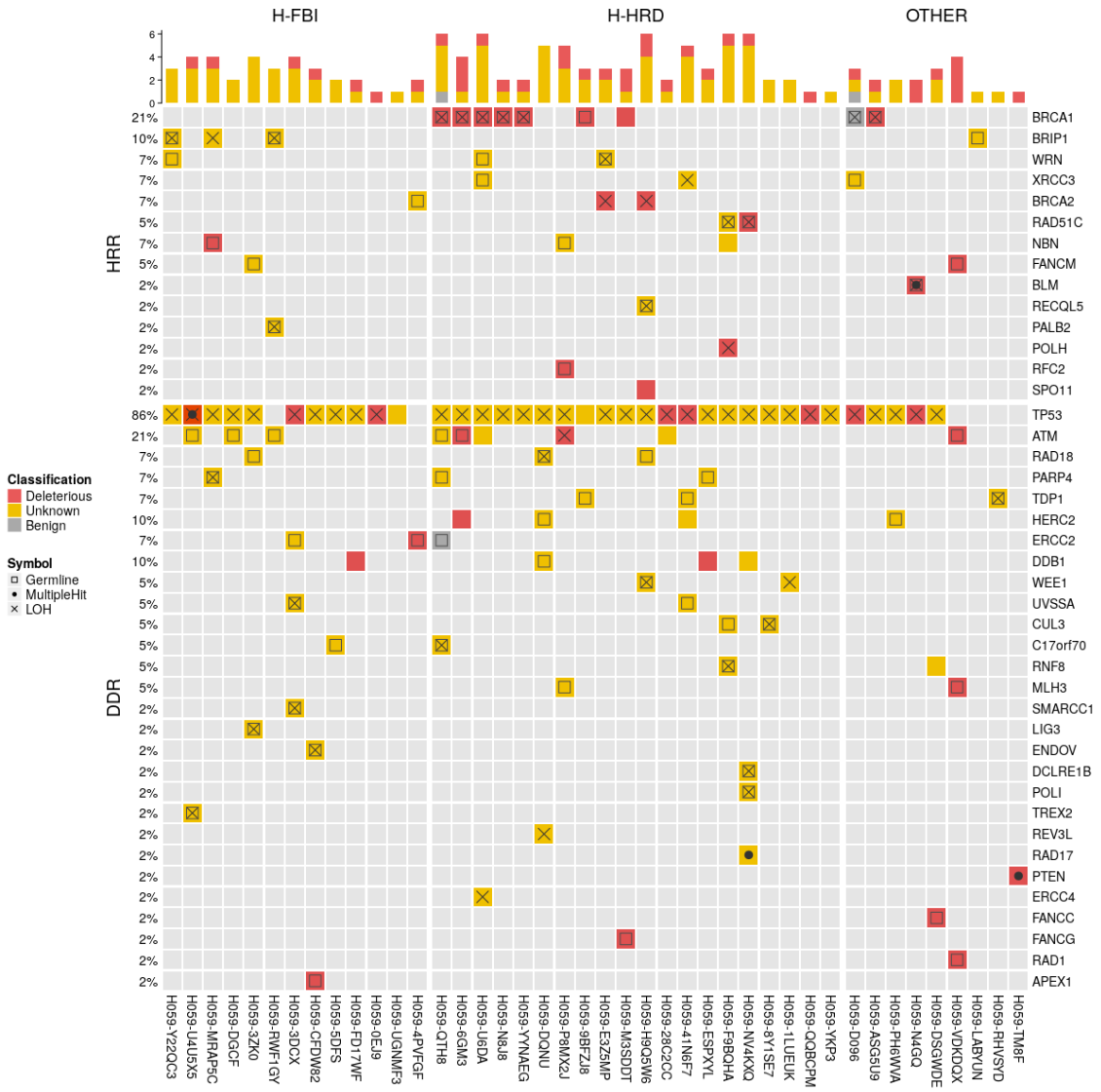


Figure 11.4: Germline-somatic landscape of DDR pathway genes. Patients are arranged in columns and stratified into three groups, while genes in row are grouped by their function into either HRR axis or other DDR pathways. Different colors represent different predicted functional consequences, where red stands for pathogenic or transcript-truncating events, yellow for protein-changing events and grey for benign variants assessed by geneticists. Different symbols stand for variants being either germline event (□), multiple variants occurring in the same gene (●) or accompanied by loss of heterozygosity(×).

deleterious variants were detected in TP53 in 8 patients and BRCA2 in 2 patients. Of these variants, only two were observed without LOH and one of the samples (H059-9BFZJ8) potentially had too low purity (13%) to discern copy number changes.

Some of these 18 patients have additional potential driving events co-existing. These include germline pathogenic variants in ATM, FANCG, BLM and truncal somatic deleterious variant in SPO11, BLM. These variants were mostly without LOH except for BLM in patient H059-TM8F. It was targeted with multiple hits, but only the somatic variant of BLM being accompanied by LOH.

Regarding patients without such events in known drivers, other potential driving events were observed. Three patients in the H-FBI subgroup had disruptions in ERCC2, DDB1 and NBN genes but without LOH. Four patients in the H-HRD subgroup harbored events in DDB1, POLH, RAD51C and ATM, where all but DDB1 lost their wild type alleles. Additional driving events were also observed in three patients with all OTHER histotypes. Among them, two patients had events in ATM and FANCC and without LOH; however ATM occurred in another sample (H059-VDKDQX, tumor6) with very low purity (20%). Notably, the third patient harbors two deleterious variants in PTEN that likely target different alleles. Overall, among patients without known driving events, potential driving events were accompanied by LOH more often in patients from the H-HRD subgroup.

Secondly, variants other than potential driving events were also profiled in Figure 11.4. Among HGSC patients, truncal somatic variants were observed in TP53 in 93.9% (31 out of 33) of the cases and only one low purity sample(H059-9BFZJ8, tumor5) showed no evidence of wild type allele lost. Other variants in known drivers include a BRCA1 germline benign variant (with LOH) and a BRCA2 germline VUS variant (without LOH).

Recurrent LOH events also occurred in genes other than known drivers. In total, there were seven additional genes that contain more than one germline or somatic variants accompanied by LOH. These are ATM(1/9, denoting the finding where out of 9 patients harboring germline or somatic event in any of related samples, one of them was accompanied with LOH), BRIP1(3/4), PARP4(1/3), RAD51C(2/2), WEE1(2/2), WRN(1/3) and XRCC3(1/3).

Similar to BRCA1 that can be disrupted in either germline or somatic setting in different patients, such pattern also apply to other disease-predisposing genes. Among genes described in Table 11.1 , this phenomena was also observed in ATM, BARD1, BLM, BRCA2, BRIP1, NBN, POLD1 and SLX4.

Collectively, potential driving events in known driver genes tend to be recurrent and biallelically inactivated. In other genes, such events were more often with LOH in patients without known driving events and belong to the H-HRD subgroup. Other variants of unknown significance were also observed with recurrent LOH or targeted in both germline and somatic setting and are therefore worth further investigation.

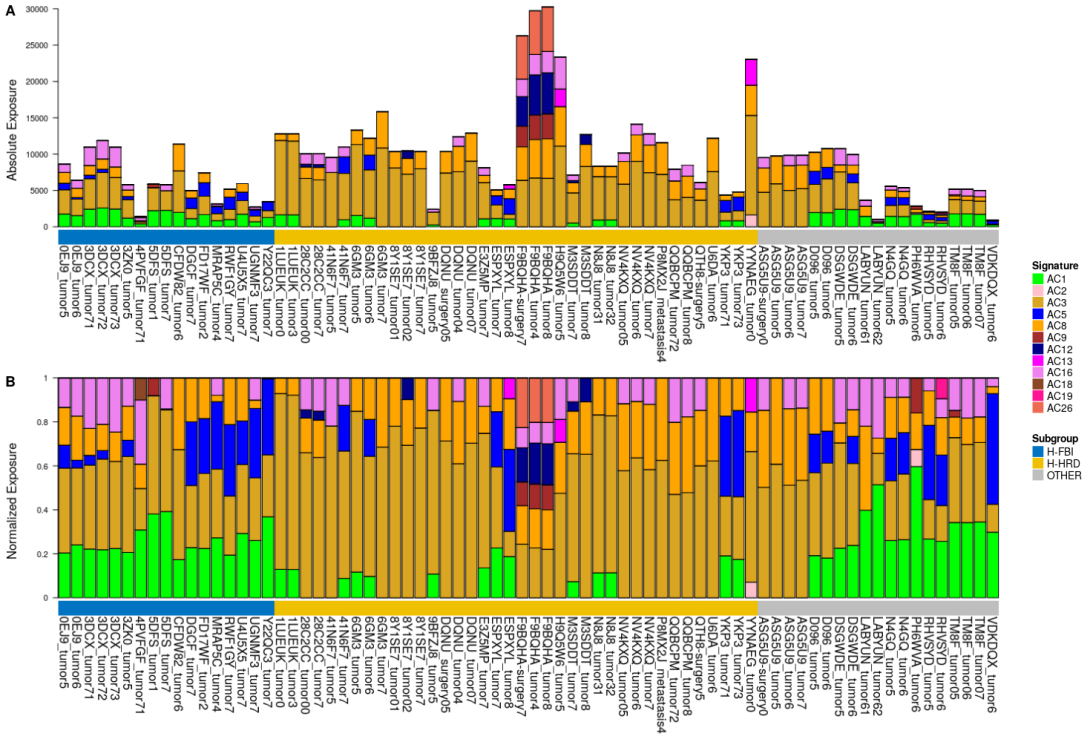


Figure 11.5: Mutational signature analysis in HIPO59, where the absolute exposure is shown in (A) and the relative contribution shown in (B). Samples are ordered by subgroups and sample name.

11.5 GENOMIC FOOTPRINTS OF DDR DEFECT

When DDR system is not functioning properly, the genetic information of parental cells can be transmitted to daughter cells with lower fidelity. Genomic footprints resulting from unrepaired or wrongly repaired DNA damage can present in various forms, ranging from small mutations, indels to different structural variants.

This section uses these downstream evidences to infer the underlying mutational processes operating in the tumor cells. In most of the subsections, the activity of all known signatures are considered, except for Section 11.5.1 and Section 11.5.2 which focus on active signatures identified in each sample.

11.5.1 *Single Base substitution signature*

The activities of 30 known mutational signatures from COSMICv2[195] are profiled in Figure 11.5. Among the 12 signatures being active in at least one sample in HIPO59 cohort, the three contributed the most to the cohort are listed in Table 11.2, and the full list can be found in Table B.5.

The three active signatures contributed the most are AC3, AC8 and AC1 (see Table 11.2). AC3, a footprint of HRD, was active in most of the tumors except for H059-PH6WVA. AC8 is a signature of unknown process and also observed in the

Signature	Patient	Average Fraction	Signature Description
AC3	41	47.4	defect DNA DSB homologous recombination repair
AC8	40	19.3	unknown
AC1	30	24	spontaneous deamination

Table 11.2: Top three active mutational signatures in HIPO59. The third column summarizes average fraction of each signature among patients with active signatures (second column).

majority of the cohort. AC1, a clock-wise signature, is active in 71% of all patients. Notably, AC1 is active in all H-FBI subgroup, but only 45% of H-HRD cohort.

There were also biologically meaningful signatures occurring in a small subset of samples. The three patients with hypermutator phenotype can be explained by either MMR defect signature (AC26), APOBEC signatures (AC2 and AC13) or Polymerase eta signature (AC9). Specifically, H059-F9BQHA had active AC9 and AC26, H059-H9Q5W6 had AC13, whereas in H059-YYNAEG, both AC2 and AC13 were observed.

On the other hand, some patients without hypermutator phenotype also harbor footprint from the aforementioned four signatures. These are H059-PH6WVA (AC2, AC9), H059-ESPXYL (AC13), H059-5DFS (AC9), and H059-TM8F (AC9).

Notably, H059-ESPXYL harbors a germline stopgain variant in BRCA2. This variant was excluded by ClinicalWorkflow due to its commonality in the population, and was also evaluated as benign by geneticist. Although its pathogenicity is under debate and association with breast cancer was reported[204], there were not much of AC3 footprint observed in this patient.

11.5.2 *Indel signature*

The activities of 18 known indel signatures from COSMICv3[196] are profiled in Figure 11.6. There were 9 active signatures identified in HIPO59 cohort, the four contributed the most to the cohort are listed in Table 11.3, and the full list can be found in Table B.5.

The four active signatures contributed the most are ID6, ID12, ID1 and ID8 (see Table 11.3), where ID6 and ID8 are again related to DSB repair defect. Among the three patients showing excess burden of mutations, only H059-F9BQHA, the one with active AC26 (MMR defect), showed also excess of indels and high activities of ID1 and ID2 signatures.

11.5.3 *Rearrangement signature*

The activities of 6 rearrangement signatures from previous breast cancer study[26] are profiled in Figure 11.7. The four signatures contributed the most to the cohort are listed in Table 11.4, and the full list can be found in Table B.5.

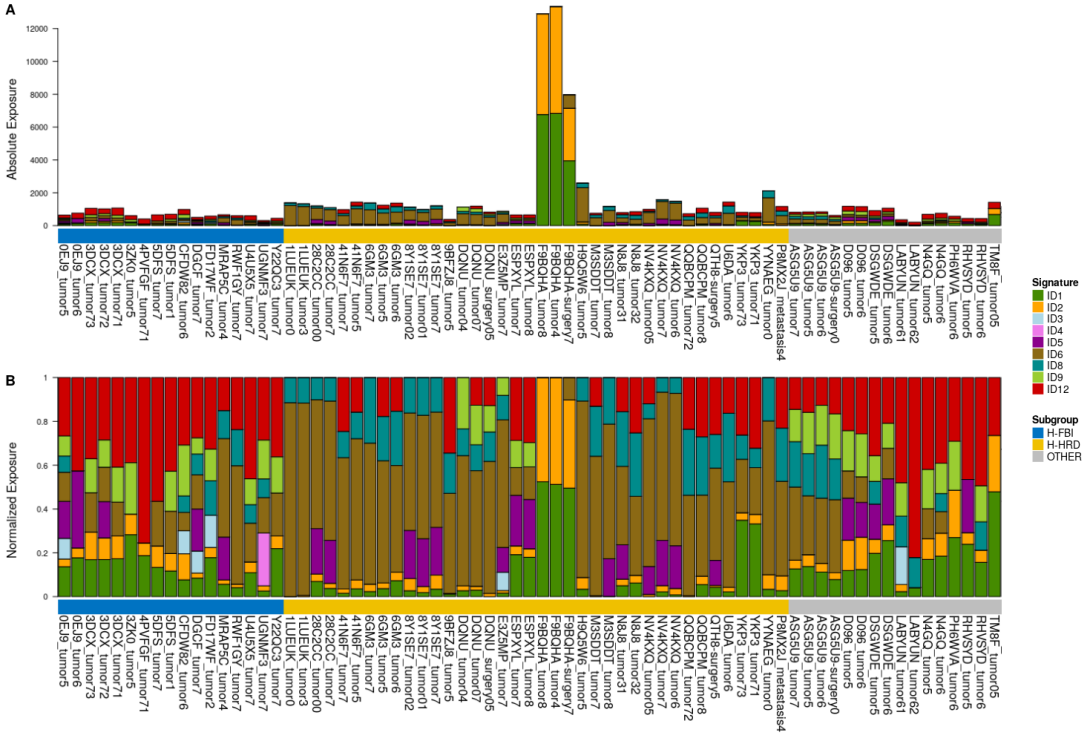


Figure 11.6: Indel signature analysis in HIPO59, where the absolute contribution is shown in (A) and the relative contribution shown in (B). Samples are ordered by subgroups and sample name.

Signature	Patient	Average Fraction	Description
ID6	35	39.98	DSB repair by NHEJ; defective HRR
ID12	35	30.6	unknown
ID1	40	14.12	Replication slippage, sometimes defective DNA MMR
ID8	31	15.7	DSB repair by NHEJ

Table 11.3: Top active indel signatures in HIPO59. The third column summarizes average fraction of each signature among patients with active signatures (second column).

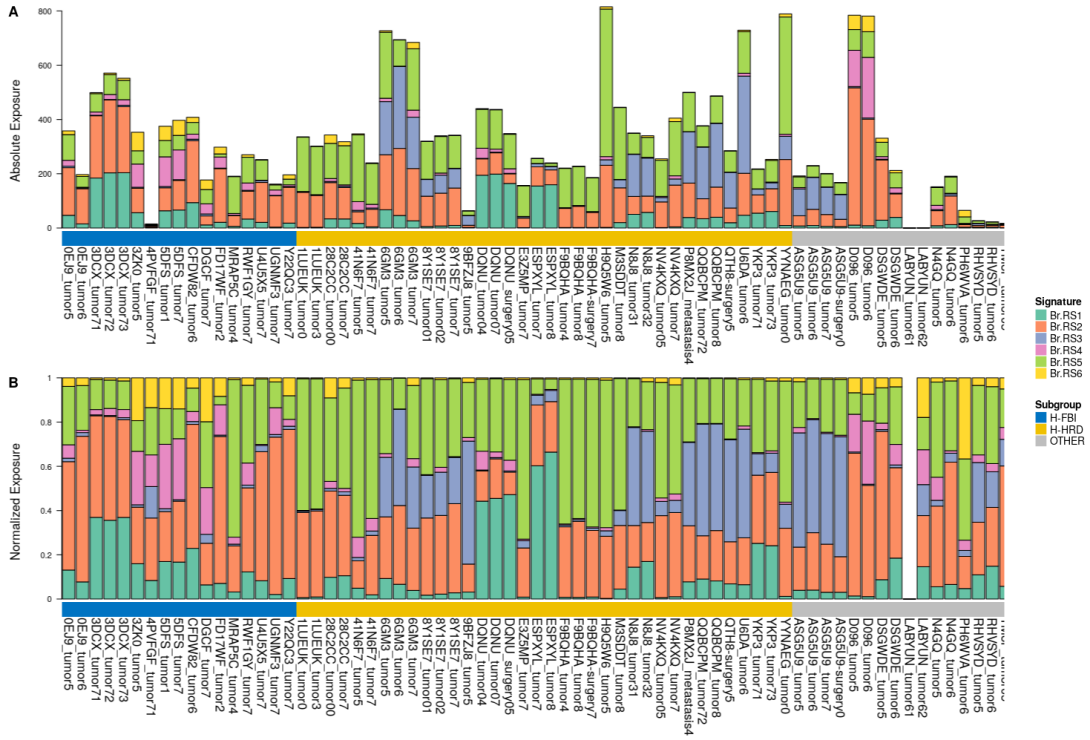


Figure 11.7: Rearrangement signature analysis in HIPO59, where the absolute contribution is shown in (A) and the relative contribution shown in (B). Samples are ordered by subgroups and sample name.

Signature	Average Fraction	Patient
Br.RS2	42	33.8
Br.RS5	41	32.2
Br.RS3	18	29.1
Br.RS1	27	18.6

Table 11.4: Top active rearrangement signatures in HIPO59. The third column summarizes average fraction of each signature among patients with active signatures (second column).

The four active signatures contributed the most to the cohort are RS2 (34%), RS5 (31%), RS3 (14%), RS1 (12%), where RS1, RS3 and RS5 are associated with defects in HR as described in the breast cancer study. In HIPO59 cohort, RS2 and RS5 were observed to be active in majority of the patients, while RS3 and RS1 were active in only a subset of the cohort.

11.5.4 *Comparing signature activities between samples and subgroups*

When comparing samples in pair-wise manner, it is shown in a later section (see Section 12.1) that samples are only similar when derived from the same patient. This holds true when comparing mutations, indels, structural variants and copy number changes. It is therefore interesting to investigate the similarity of signature activities between different samples. To perform these comparisons, the normalized exposure were used and in terms of mutational signatures, the normalized exposures from all 30 signatures (stage I result in Section 8.8) were used.

For all possible pairs of samples, the Pearson correlation coefficients were calculated between pairs of signature activities, and the distribution of pairwise correlations between related and unrelated samples are profiled in Figure 11.8. In consistent with observations in the later section (see Figure 12.3(A)), signature activity profiles usually showed high correlation between related samples. However, unlike those observed in mutation and indels (see Figure 12.1), structural variants (see Figure B.6) and copy number profiles (see Figure B.7), there were sometimes similar signature activity profiles observed between unrelated samples (see Figure B.5).

Next, activities of all signatures from three types of signature sets are compared between HGSC genomic subgroups. Figure 11.9 shows only those with significant differences using Welch t-test with bonferroni correction (see Table B.6 for test result). Interestingly, in all three types of signature sets, signatures related to DSB repair defect were all significantly higher in H-HRD subgroup, including AC3 in mutational signatures, ID6 and ID8 in indel signatures, as well as RS3 and RS5 in rearrangement signatures.

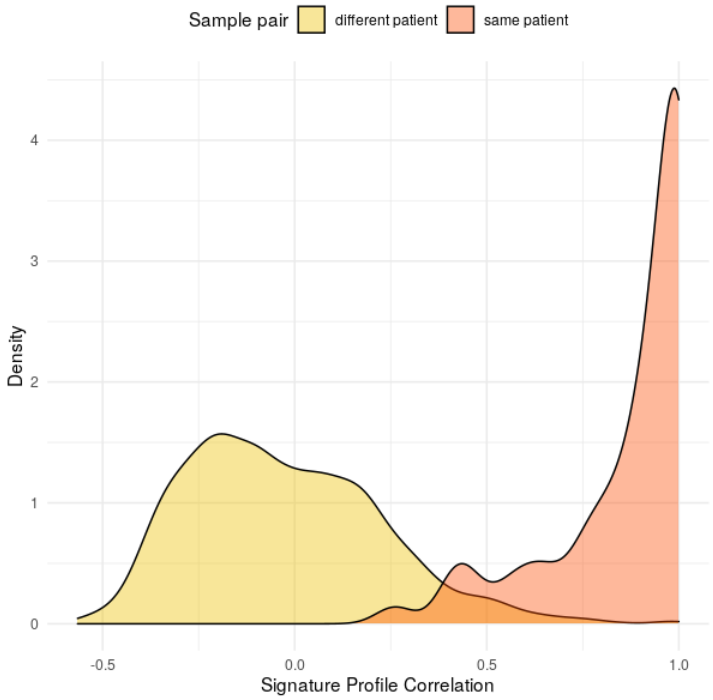


Figure 11.8: From each pair of samples, a Pearson correlation coefficient was calculated using their signature activities. The distribution of correlation scores between related samples (red) and between unrelated samples (yellow) are profiled separately.

11.5.5 *Homologous recombination deficiency score*

Another genomic footprint, **HRD score**, is summarized from copy number profiles. It has been routinely used in clinical trials to identify patients that might respond better for PARP inhibitor treatment in different cancer types. Therefore, it is of clinical relevance to know how the scores behaves in two HGSC genomic subgroups.

At first, the correlation between different component scores were compared. Figure 11.10(A) shows that LOH and LST had high correlation (R=0.69) between each other, and they both correlate well with the composite HRD score (R>0.85). On the other hand, TAI score showed low correlation with other component scores (LOH, LST) nor with the HRD score.

Next, the four scores were compared between HGSC genomic subgroups using Welch t-test. Figure 11.10(B) shows that both component scores (LOH, LST) and composite HRD score are significantly higher in H-HRD group, while LST score is not discriminative between the two groups. The test result can be found in Table B.7.

Furthermore, to quantify the ability of the four scores to discriminate two genomic subgroups, their receiver operating characteristic (ROC) curves were calculated using R package 'ROCit' and visualized in Figure 11.11. The four curves with area under the ROC curve (AUC) ordered from high to low are HRD score (0.8954), LOH (0.8799), LST (0.8783) and lastly TAI (0.5564).

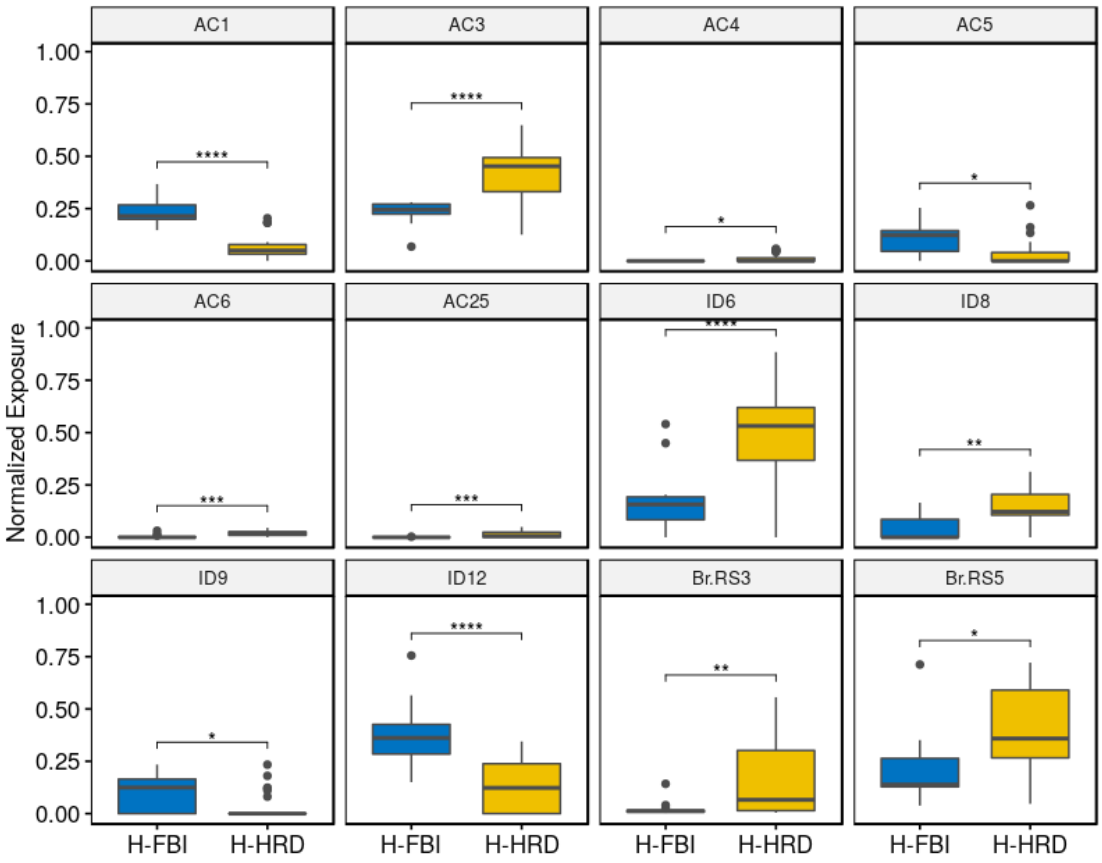


Figure 11.9: Significant differences in signature activities between HGSC genomic subgroups. Samples are divided into group 1 (17 H-FBI tumors). A Welch t-test was performed for each of all 54 signatures from three types of signature sets, and all the p-values were adjusted for bonferroni correction. The test result can be found in Table B.6 and only those with p.adj < 0.05 are shown in the figure.

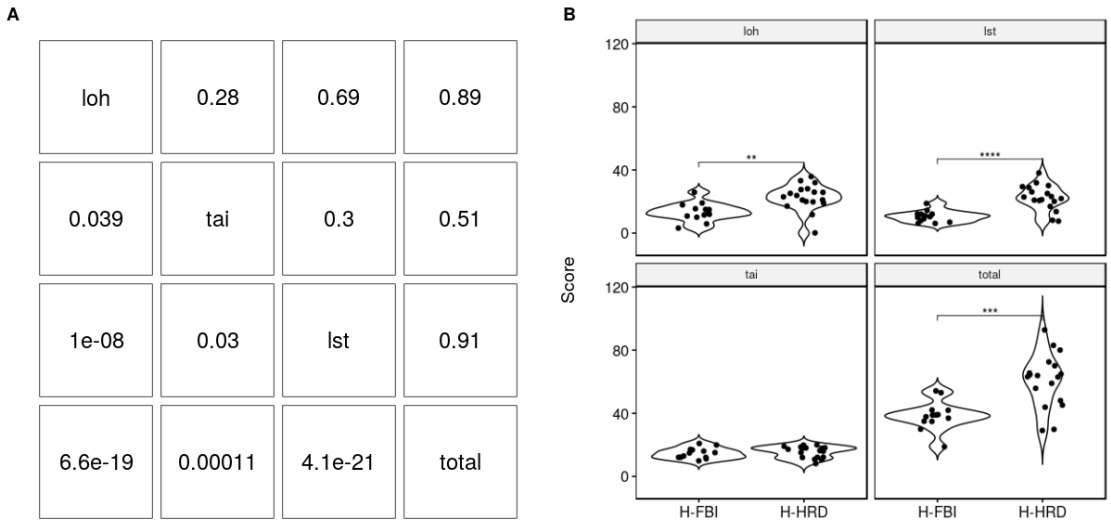


Figure 11.10: The mutual correlation between HRD score (total) and its three component scores (LOH, TAI and LST) are calculated in (A), where the upper panel shows the pair-wise Pearson correlation coefficient and the lower panel shows the p-value. These scores were also compared between two HGSC subtypes using Welch t-test. Comparisons with significant p-value (<0.05) are marked with stars.

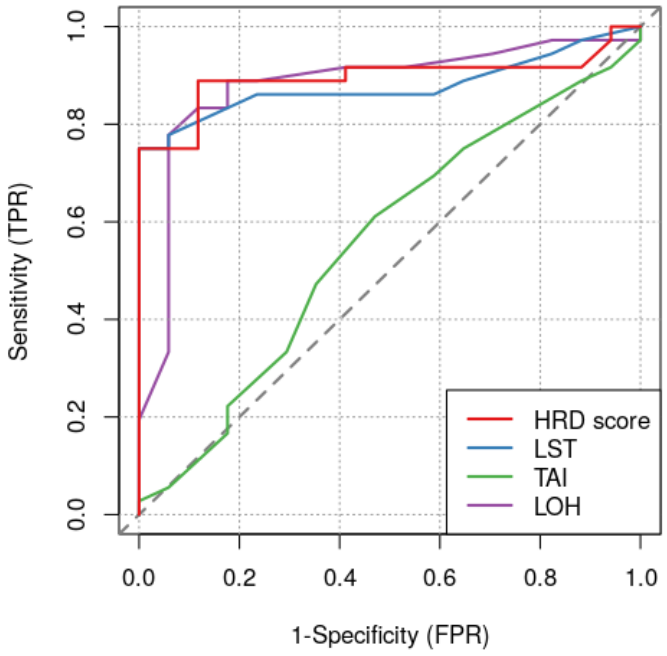


Figure 11.11: Discriminative power of HRD score and its three component scores for separating genomic subgroups.

12 TUMOR HETEROGENEITY

HIPO59 is a multi-sample cohort (see Section 6.2) and there are 23 patients with multiple sites sampled. In this section, quantitative measures are used to compare genomes between tumor pairs. **Jaccard Index**, as described in Section 8.9.1, can be used to measure their similarity. It takes values between 0 to 1, with a larger value representing a higher degree of similarity. Inversely, **Heterogeneity Index**, defined as $1 - \textit{Jaccard Index}$, tells the opposite. The following sections use these measures to look at similarities between tumors in Section 12.1 and stratify patients in Section 12.2. Given the patient stratification, tumor phylogeny for each patient are visualized in Section 12.3 and their implication in patient survival was investigated in Section 12.4.

12.1 QUANTIFYING SIMILARITY BETWEEN TUMOR SAMPLES

In HIPO59 there are 23 patients with multiple samples. A pair-wise comparison between all these samples were done based on somatic functional mutations and indels, where a Jaccard Index measures the similarity between two callsets from the sample pair. Figure 12.1(A) visualize the result of all comparisons in a heatmap. On the other hand, Figure 12.1(B) shows only the scores for related samples from the same patient.

Using the same approach, sample similarity can also be measured based on structural variants (see Figure B.6) and copy number profiles (see Figure B.7). The three quantified heterogeneity measures for each sample can be found in the Compact disc along with the thesis (**Tumor_Heterogeneity_SAMPLE.txt**).

In principle the three measures show similar tendency while with different dynamic ranges. Two pairs of comparisons are visualized in Figure 12.2 as examples. Subfigures A, C, E compares two similar samples (tumor7 and tumor5) from patient H059-DQNU, while Subfigures B, D, F contrasts two very different samples (tumor7 and tumor5) from patient H059-6GM3. For mutations and indels, their variant allele frequency (VAF) in two samples are compared. In Figure 12.2(A) most variants lie close to the diagonal, meaning their VAFs are similar. In contrast, many variants lying on the x-axis or y-axis in Figure 12.2(B) shows that they are private to either samples. Structural variants are compared in circos plot in Figure 12.2(C,D), where each SV is visualized by one line connecting two genomic locations. Copy number profiles are compared by showing the log2 value of copy number ratio between two samples. Segments lying at baseline 0 means the two samples have same allele specific copy number at this segment. Altogether it shows that the Jaccard Index reflect true heterogeneity between samples.

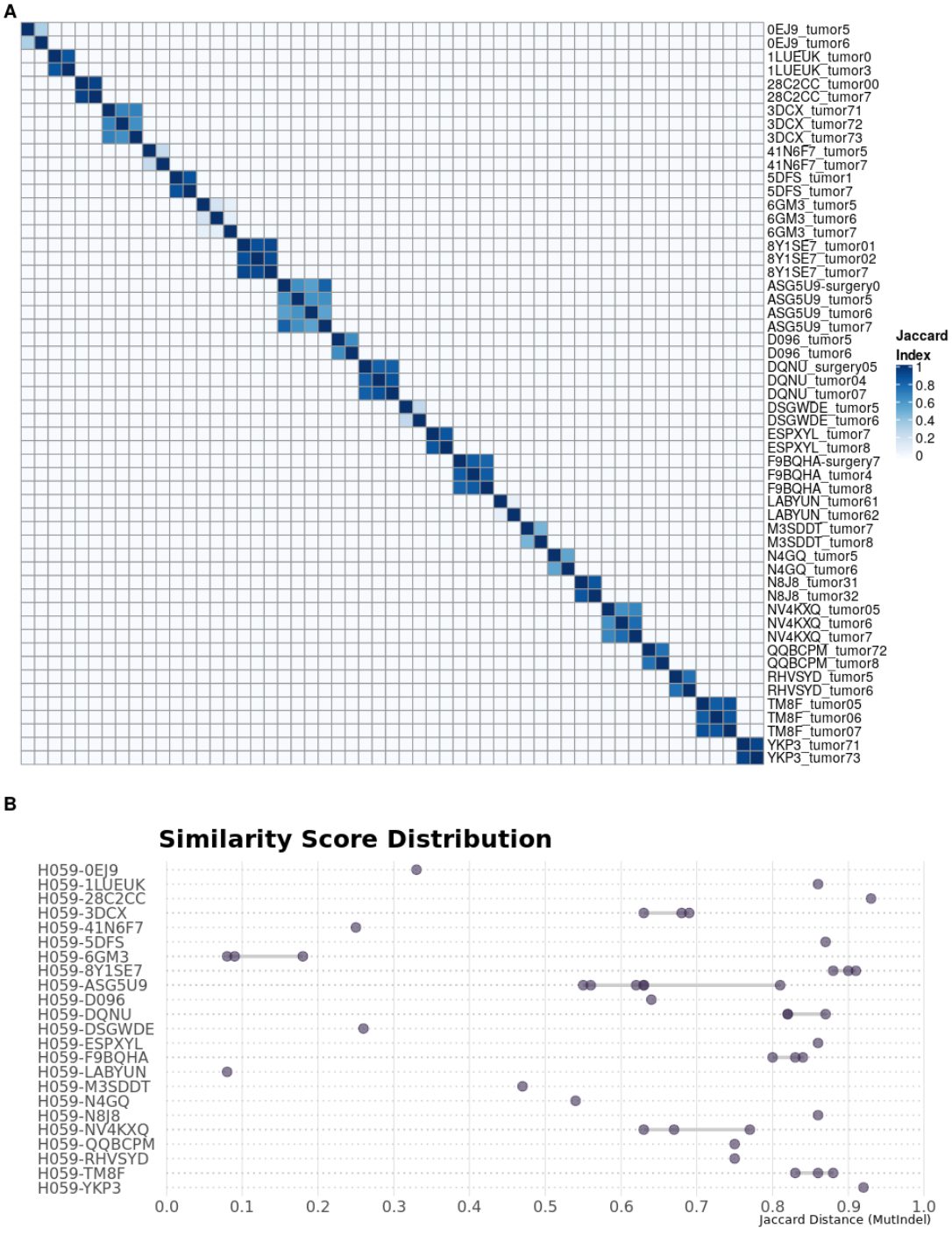


Figure 12.1: Pair-wise sample comparisons based on small variants. In (A) similarity between all sample pairs are shown, where samples arranged in columns and rows are in the same order. Each cell is one Jaccard Index of corresponding sample pair. (B) shows the degree of sample similarity in each patient, where a data point represent one score between a pair of related samples from this patient. All comparisons from the same patient are connected with a line to visualize the range of similarity score.

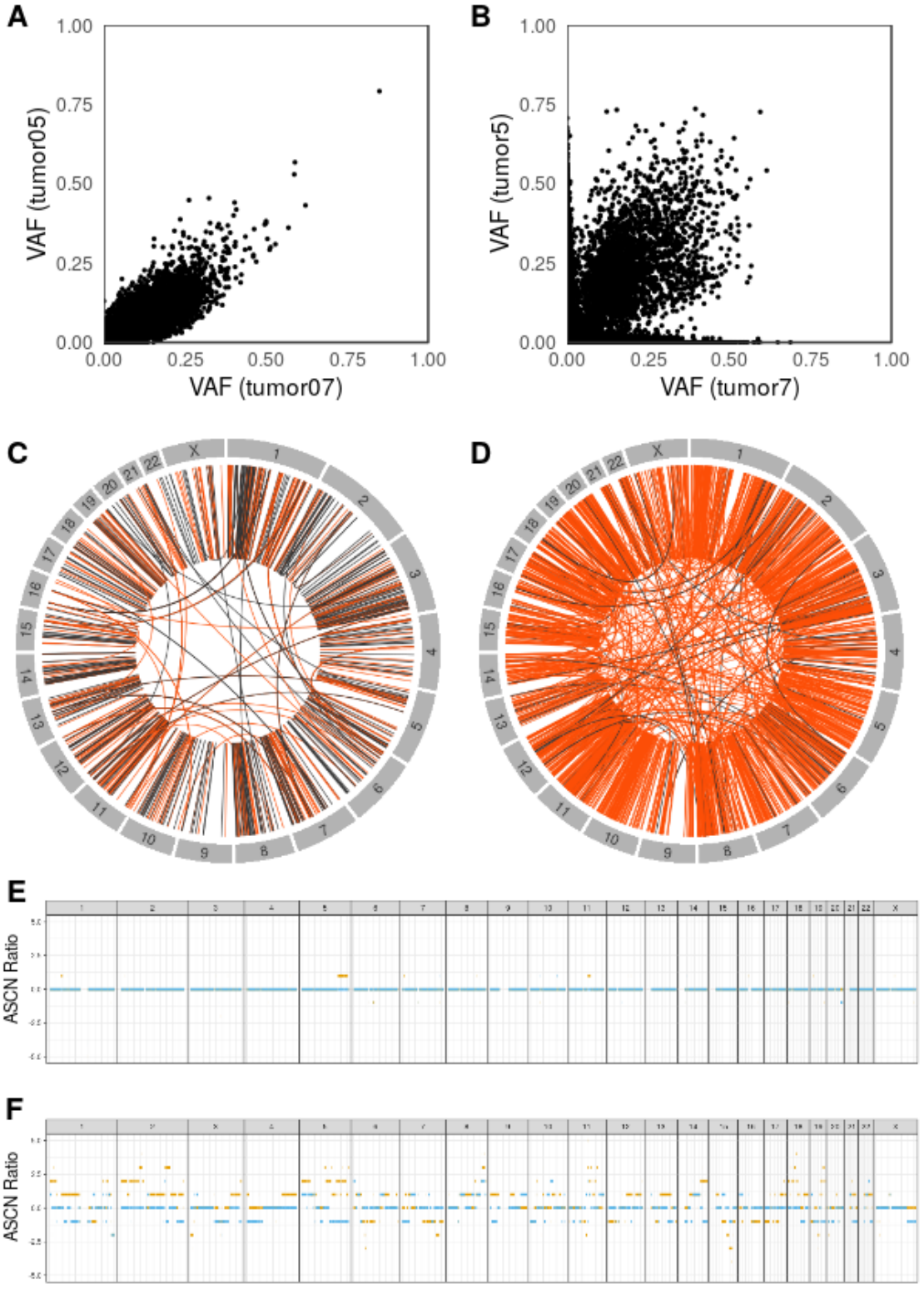


Figure 12.2: Examples showing two sample pairs with high (A,C,E) and low (B,D,F) similarity. Small variants have their VAFs compared in (A) and (B), with similar variants locating at the diagonal. Structural variants are compared in (C) and (D), where shared SV events are colored in grey and private ones colored in orange. In (E) and (F) the allele-specific copy number (ASCN) profiles are compared. The major allele and minor allele are colored in orange and blue, separately.

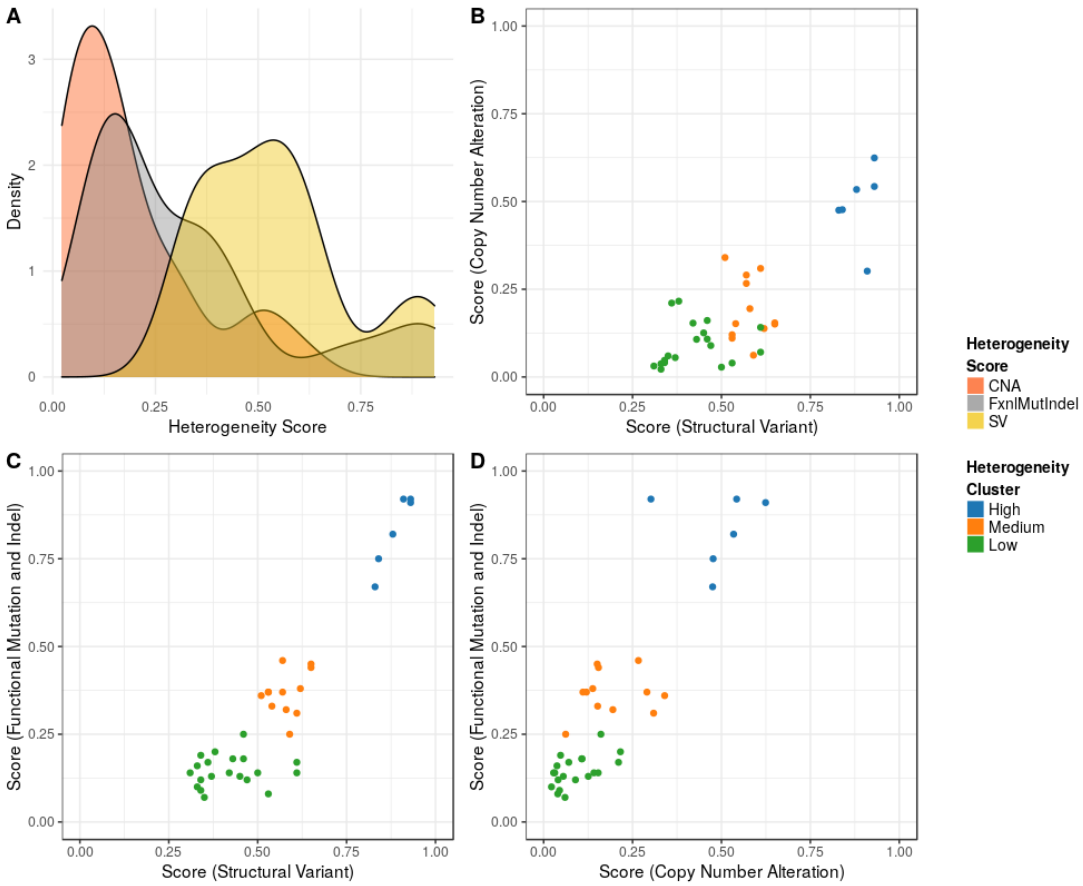


Figure 12.3: From each pair of sample comparisons, three heterogeneity indexes were independently derived from small variants, structural variants and copy number alterations. The distribution of three scores are compared in (A) and their mutual relationships profiled in (B,C,D). Based on these scores, each comparison is further clustered into high, medium and low heterogeneity groups encoded by blue, orange and green color, separately.

12.2 STRATIFY PATIENTS WITH HETEROGENEITY SCORE

Three Heterogeneity Indexes, derived from Jaccard Index, were calculated for each sample pair and their different dynamic ranges were shown in Figure 12.3(A). These within patient comparisons were each visualized as a data point in Figure 12.3(B,C,D), where the relationship between three scores were profiled. The scatter plots show that these scores were mutually correlated. After having the comparisons classified into three clusters according to Section 8.9.2, the cluster membership was further encoded by different colors in the scatter plots.

The three clusters are named High, Medium and Low heterogeneity groups. Table 12.1 shows the number of sample pairs assigned to each group, as well as the final group assignment of each patient. This result can also be found in the Compact disc along with the thesis (**Tumor_Heterogeneity_PID.txt**). In the end, there are 4 patients in the

group with high heterogeneity, 6 in the Medium group and 11 in the Low group. Notably, 4 (67%) of 6 patients in the Medium group are with non-HGSC histotype.

In addition, for the 7 patients having more than two samples, most of the pair-wise comparisons were consistently assigned to the same heterogeneity group. The only exception was found in H059-ASG5U9, where her 3 pre-treatment samples showed medium heterogeneity, while the interval debulking sample were more similar to one of the pre-treatment sample.

12.3 PHYLOGENETIC TREE OF TUMOR SAMPLES

A phylogenetic tree embeds the evolutionary relationships among different samples from the same patient. Constructed from small variants, Figure 12.4 displays such trees for 15 HGSC patients underwent heterogeneity stratification, with heterogeneity group encoded by different colors in the subplot titles.

At the top of a tree shows a green root node representing the germline genome of the patient. Each tree starts with the germline genome and end with leaf nodes representing different tumors sampled for sequencing. While traversing the tree from the root node (germline) to a leaf node (tumor), the altitude difference encodes the number of small variants accumulated during tumor transformation. Notably, in every tree there is one orange node named **Most Recent Common Ancestor (MRCA)**, which is a proposed ancestor of observed tumors. It is assumed that the MRCA once existed at some time point during tumor development and gave rise to different tumors.

As the altitude from germline node to MRCA node encodes the shared small variants between all samples, this distance would reflect the heterogeneity status between samples. As expected, patients in High heterogeneity group have MRCA node closer to the root node, implying an earlier divergence of different tumors; whereas patients in Low heterogeneity group likely had their MRCA emerged later in the tumor phylogeny. This suggests a correspondence between the heterogeneity grouping and the phylogeny topology.

12.4 POTENTIAL CLINICAL IMPLICATION OF HETEROGENEITY STATUS

To investigate the prognosis value of the heterogeneity grouping, the survival status was compared between HGSC patients in Low heterogeneity group (n=10) versus patients in High or Medium heterogeneity groups (n=5). Figure 12.5 suggests a trend that patients with lower heterogeneity between samples tend to have unfavorable outcomes.

Patient	Histotype	High	Medium	Low	Patient Group
H059-0EJ9	HGSC	1	0	0	High
H059-41N6F7	HGSC	1	0	0	High
H059-6GM3	HGSC	3	0	0	High
H059-3DCX	HGSC	0	3	0	Medium
H059-NV4KXQ	HGSC	0	1	0	Medium
H059-1LUEUK	HGSC	0	0	1	Low
H059-28C2CC	HGSC	0	0	1	Low
H059-5DFS	HGSC	0	0	1	Low
H059-8Y1SE7	HGSC	0	0	3	Low
H059-DQNU	HGSC	0	0	3	Low
H059-ESPXYL	HGSC	0	0	1	Low
H059-F9BQHA	HGSC	0	0	3	Low
H059-N8J8	HGSC	0	0	1	Low
H059-QQBCPM	HGSC	0	0	1	Low
H059-YKP3	HGSC	0	0	1	Low
H059-LABYUN	OTHER	1	0	0	High
H059-ASG5U9	OTHER	0	5	1	Medium
H059-D096	OTHER	0	1	0	Medium
H059-N4GQ	OTHER	0	1	0	Medium
H059-RHVSYD	OTHER	0	1	0	Medium
H059-TM8F	OTHER	0	0	3	Low

Table 12.1: Heterogeneity group assignment for multi-sample patients. The three columns Cluster1, Cluster2 and Cluster3 list the number of sample-pairs classified into each cluster. Each patient is then assigned a heterogeneity group according to the voting of these sample-pair assignments, where the three clusters encode for High (Cluster 1), Medium (Cluster 2) and Low (Cluster 3) heterogeneity groups.

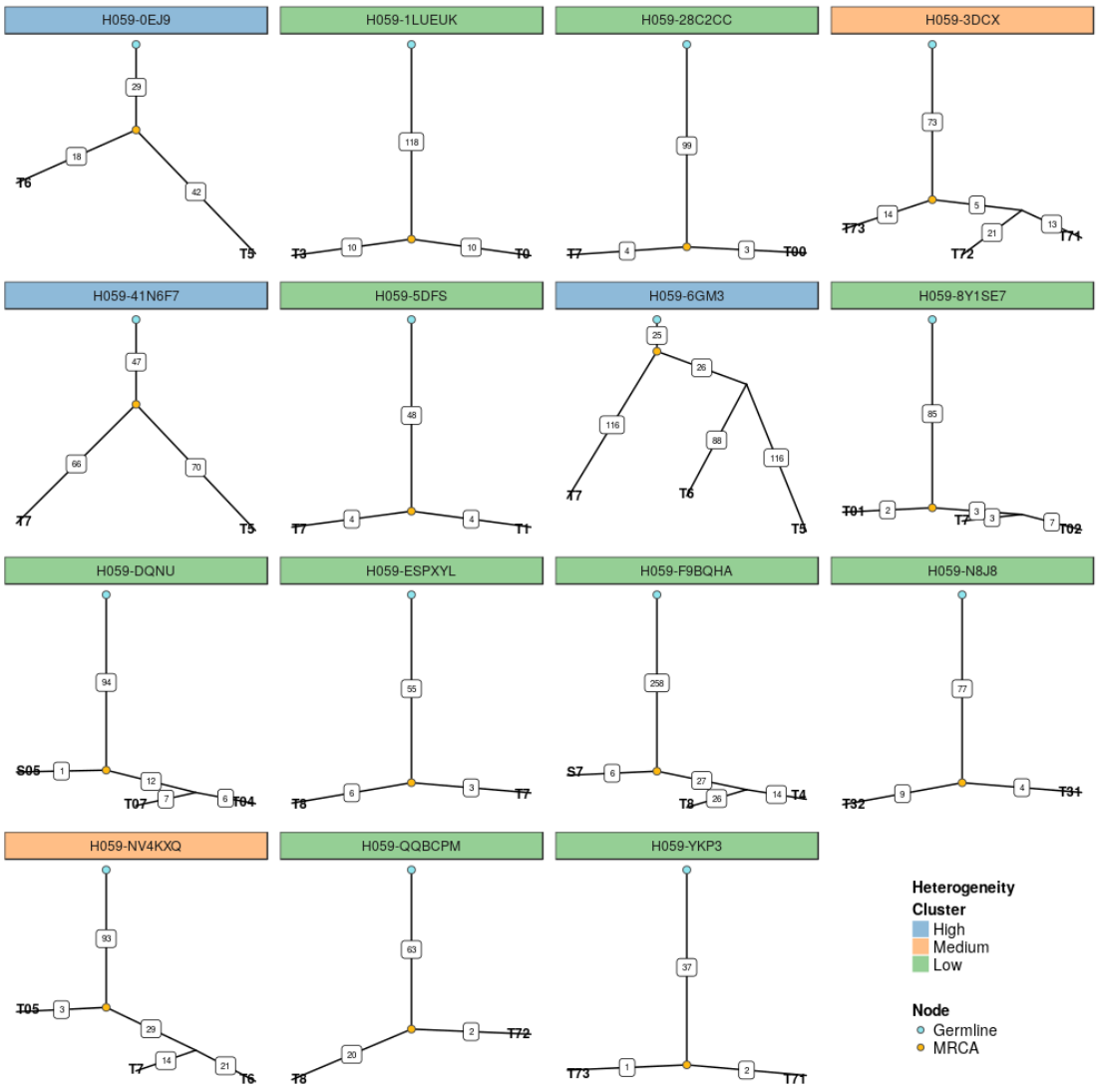


Figure 12.4: Tumor phylogenetic tree for HGSC patients. In each tree the nodes represent a specific genomic composition, where green node indicates the germline genome, orange node stands for a conceptual MRCA and leaf nodes represent observed samples. The number labeled at each branch described the number of small variant changes that are different between the two connected tree nodes. Heterogeneity group assignment of each patient is encoded by the color of patient label.

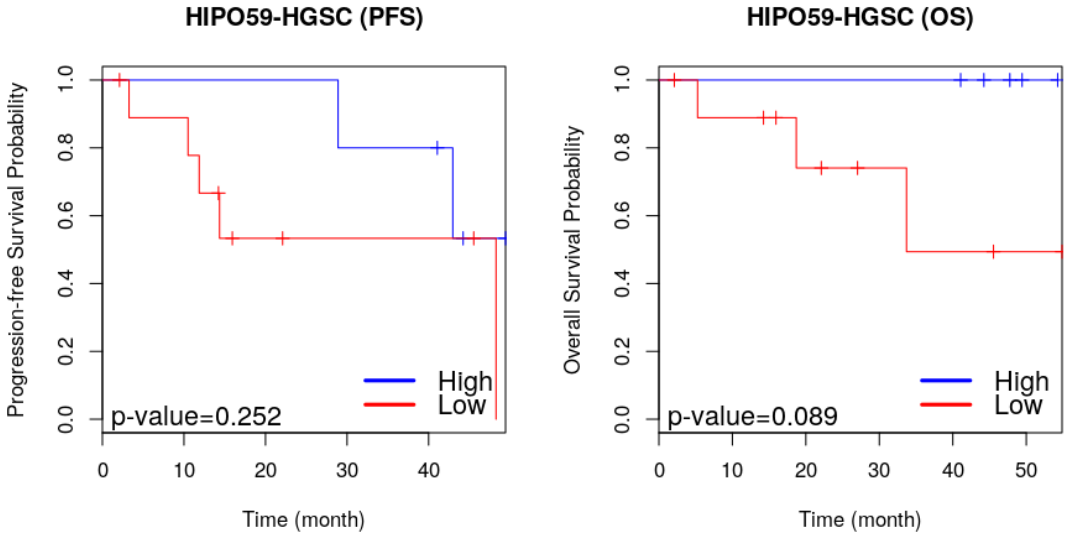


Figure 12.5: Prognostic value of heterogeneity grouping. HGSC patients are grouped in to High (high and medium heterogeneity groups, n=5) and Low (low heterogeneity group, n=10) groups. The Kaplan-Meier (K-M) plot on the left compares the progression-free survival (PFS) between two groups, where High and Low groups are encoded with blue and red colors separately. The K-M plot on the right, on the other hand, compares overall survival (OS) between the groups.

13 TUMOR EVOLUTION

Each tumor is sampled at one time point in the tumor developmental process. Given the single snapshot, scientists have been trying to computationally dissect the temporal order of observed molecular events. In this section, computational methods capable of inferring the relative time order of major molecular events were used to estimate when these events happened during tumor evolution. The multi-sample design of HIPO59 further offer finer time granularity when dissecting the process.

Starting with identifying WGD occurrence in Section 13.1, the subsequent subsections estimated the timing of small variants (Section 13.2), and the timing of WGD and MRCAs (Section 13.3). A reconstruction of tumor evolution trajectory for each patient then place these events in order in Section 13.4. Lastly, the temporal change in the activities of mutational processes were revealed in Section 13.5.

13.1 WHOLE GENOME DUPLICATION

Identifying WGD status in individual samples

The occurrence of WGD in individual tumor can be identified and classified according to Section 8.10.1. Figure 13.1 shows that WGD occurred in 35 (49%) of 71 samples in the HIPO59 cohort. When looking at the timing of copy number gains, 30 (86%) of 35 WGD-positive tumors had most of the gains occured at concordant timing. Notably, as opposed to H-HRD subgroup where WGD occurred in 36% (13 out of 36) of the cases, all 17 H-FBI tumors showed WGD.

Measures associated with WGD status

Next, the association between WGD status and copy-number associated measures were examined. For the 23 patients with multiple samples, Figure 13.2(A) shows that the median of all within patient heterogeneity index (`HET_CNA`) was associated with WGD status (with or without WGD) and independent of Histotype (HGSC or OTHER).

As to the association with CIN, Figure 13.2(B) profiles the wGII scores of 42 patients, taking the median if multiple samples were available, and showed that wGII score had a histotype-dependent association with WGD status. In non-HGSC (OTHER) histotypes, WGD is associated with higher level of CIN. However, in HGSCs the CIN level was already high in WGD-negative (nWGD) cases and therefore not distinguishable from WGD-positive cases.

Given that heterogeneity suggested a trend of better prognosis (see Figure 12.5), WGD status might also bear clinical implications as it associated with one of the heterogeneity

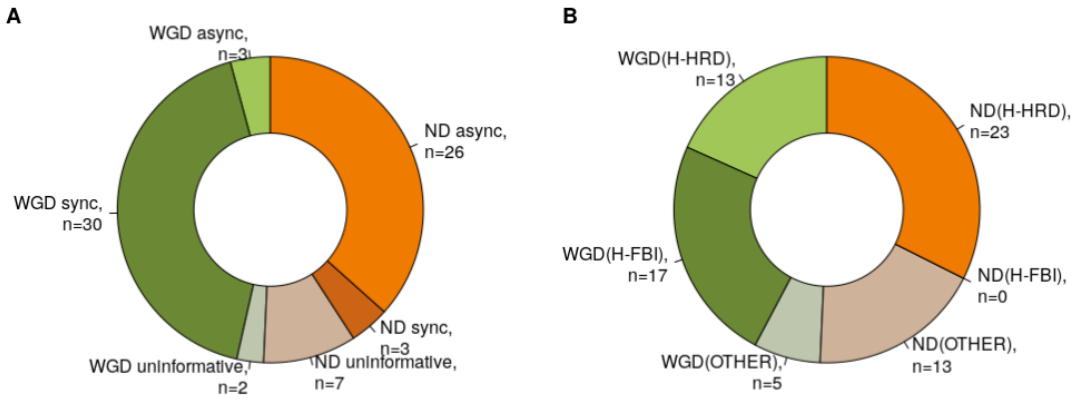
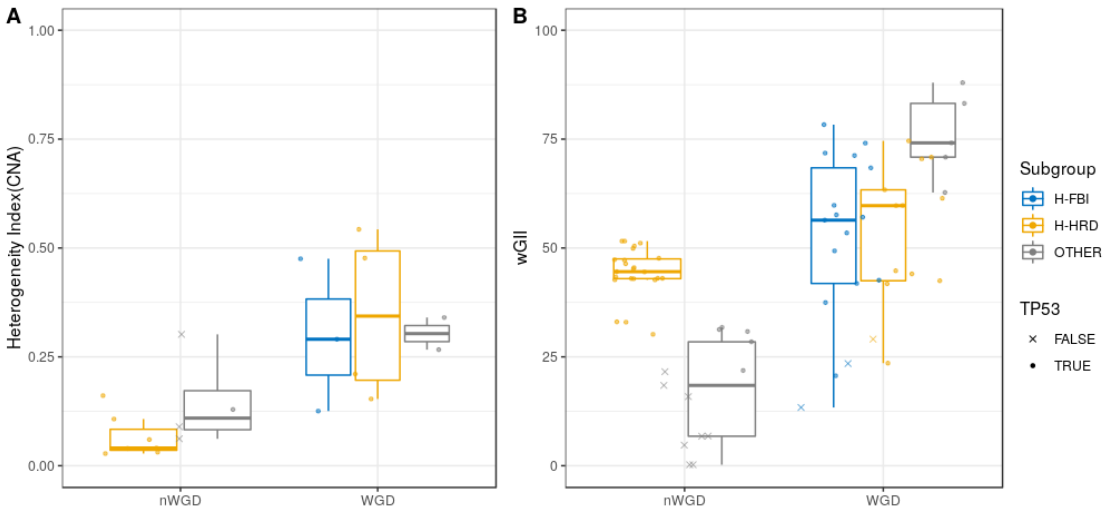


Figure 13.1: Whole genome duplication in HIPO59. WGD status stratified by timing concordance in (A) and by Shah-2017 genomic subgroups in (B). A sample can have ploidy status of either WGD (WGD) or near-diploid (ND). There were three categories of timing concordance, being synchronous (sync), asynchronous (async) and uninformative.



ANOVA (HET_CNA ~ WGD + Histotype + WGD*Histotype)

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WGD	1	0.250	0.250	17.481	7.06e-04
Histotype	1	0.007	0.007	0.475	5.00e-01
WGD:Histotype	1	0.010	0.010	0.680	4.22e-01
Residuals	16	0.229	0.014	NA	NA

ANOVA (wGII ~ WGD + Histotype + WGD*Histotype)

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WGD	1	3979.5679	3979.5679	16.657	2.22e-04
Histotype	1	510.2943	510.2943	2.136	1.52e-01
WGD:Histotype	1	4816.7481	4816.7481	20.161	6.44e-05
Residuals	38	9078.8910	238.9182	NA	NA

Figure 13.2: Association between WGD and CNA-based measures are shown for (A) heterogeneity index (HET_CNA) and (B) chromosomal instability (wGII score). In the upper panel, the scores were first stratified by WGD-positive (WGD) and WGD-negative (nWGD) and then by histotype (HGSC, OTHER), finally by HGSC genomic subgroup (H-FBI, H-HRD). Statistical tests on the associations were done with R function 'aov', with the formula and results listed in the lower panel.

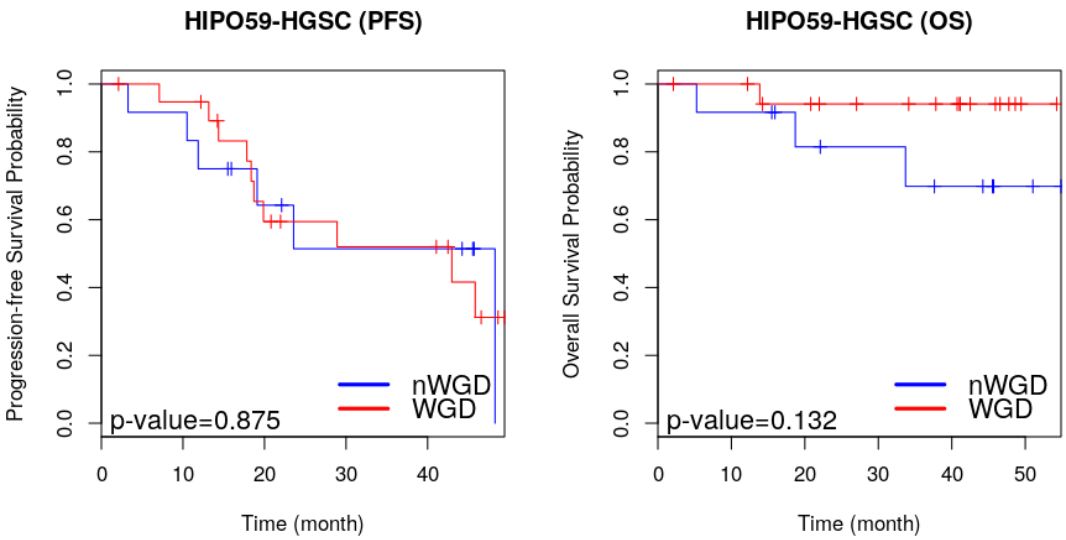


Figure 13.3: Prognostic value of WGD status. HGSC patients were grouped in to WGD (n=20) and nWGD (n=12) groups. The K-M plot on the left compares the PFS between two groups, where nWGD and WGD groups are encoded with blue and red colors separately. The K-M plot on the right, on the other hand, compares OS between the groups.

index. As expected, WGD status suggested a similar, but weaker trend for better OS in Figure 13.3.

13.2 TIMING OF SMALL VARIANTS

After excluding potential germline variants in the SomFxn call set, 615,741 mutations and 83,047 indels were analyzed with MutationTime algorithm. Given a conceptualized tumor evolution process in Figure 8.4, the algorithm categorizes small variants into four timing categories corresponding to four epochs during evolution (early clonal, late clonal, unspecified clonal and subclonal).

13.2.1 *Overview of variant timing in individual samples*

Figure 13.4 profiles the timing class composition for small variants detected in each of 71 samples in HIPO59.

When comparing the composition in mutations to that in indels, Welch t-test with bonferroni correction suggested that the three clonal categories showed similar fraction among all patients. However, there were significantly more indels classified as subclonal or NA as compared to mutations (see Figure B.8).

In general, there were no significant difference in the composition between subgroups. Instead, Figure B.8(C) used a Welch t-test with bonferroni correction and showed that in WGD-positive tumors, higher fraction of timing classes “clonal [early]” (p.adj=8.28e-7)

and “clonal [late]” (p.adj=1.09e-13) were observed when compared to WGD-negative tumors. This observation was as expected as more of the clonal variants were able to be timed in WGD-positive tumors.

13.2.2 *Timing of variants in cancer-associated genes*

Instead of aggregating variants by sample, one can also aggregate variants by gene. Following Gerstung's tutorial[199] section 4, Figure 13.5 profiled the timing class composition of functional variants in cancer-associated genes defined in Cancer Gene Census (CGC)[178] (see Section 6.3). Among the 717 genes in CGC, only 26 genes mutated in more than three samples are profiled.

The most recurrently mutated gene, TP53, were always clonal and emerged in the “clonal [early]” epoch in 61% of the time. Somatic variants in BRCA2 were classified as “clonal [NA]” for patient H059-E3Z5MP and “clonal [early]” for patient H059-H9Q5W6. The only somatic variant in BRCA1 was detected in H059-M3SDDT and classified as “clonal [NA]”.

13.2.3 *Role of DDR genes, OGs and TSGs in different tumor epochs*

To investigate whether DDR genes, TSG and OG would play role in different tumor developmental stages, timed variants were further segregated into non-overlapping gene sets. The three gene sets of interest are `DDR_PanCan` (275 DDR-related genes excluding TP53, see Section 6.3), `TSG` (200 TSGs excluding overlaps with `DDR_PanCan`, see Section 6.3) and `OG` (242 OGs excluding overlaps with `DDR_PanCan`, see Section 6.3).

Excluding TP53 variants, there were 2,025 small variants in the three gene sets being timed, and 264 of them are deleterious variants. Note that these somatic variants are collected from all HGSC patients except for H059-41N6F7, who later on found having a different order of tumor epochs. Figure 13.6 showed the temporal trend of the deleterious variant proportions for three gene sets. In the lower panel, an interesting trend was observed for H-HRD cases. Column `OG` displayed that proportion of deleterious variants in oncogenes slowly increased as tumor evolved. This trend is concordant with that of all timed variants (in column "All"). However, `DDR_PanCan` and `TSG` deleterious variants are more likely to be found in early stages in tumor evolution, which is opposite to the overall trend in column "All".

13.3 TIMING OF MAJOR EVENTS

To infer the chronological time when major events occurred, their timing in mutation time coordinate should be inferred first and then map it to chronological time coordinate as described in Section 8.10.3.1. Notably, a faithful mapping requires the mutation rate being modeled correctly.

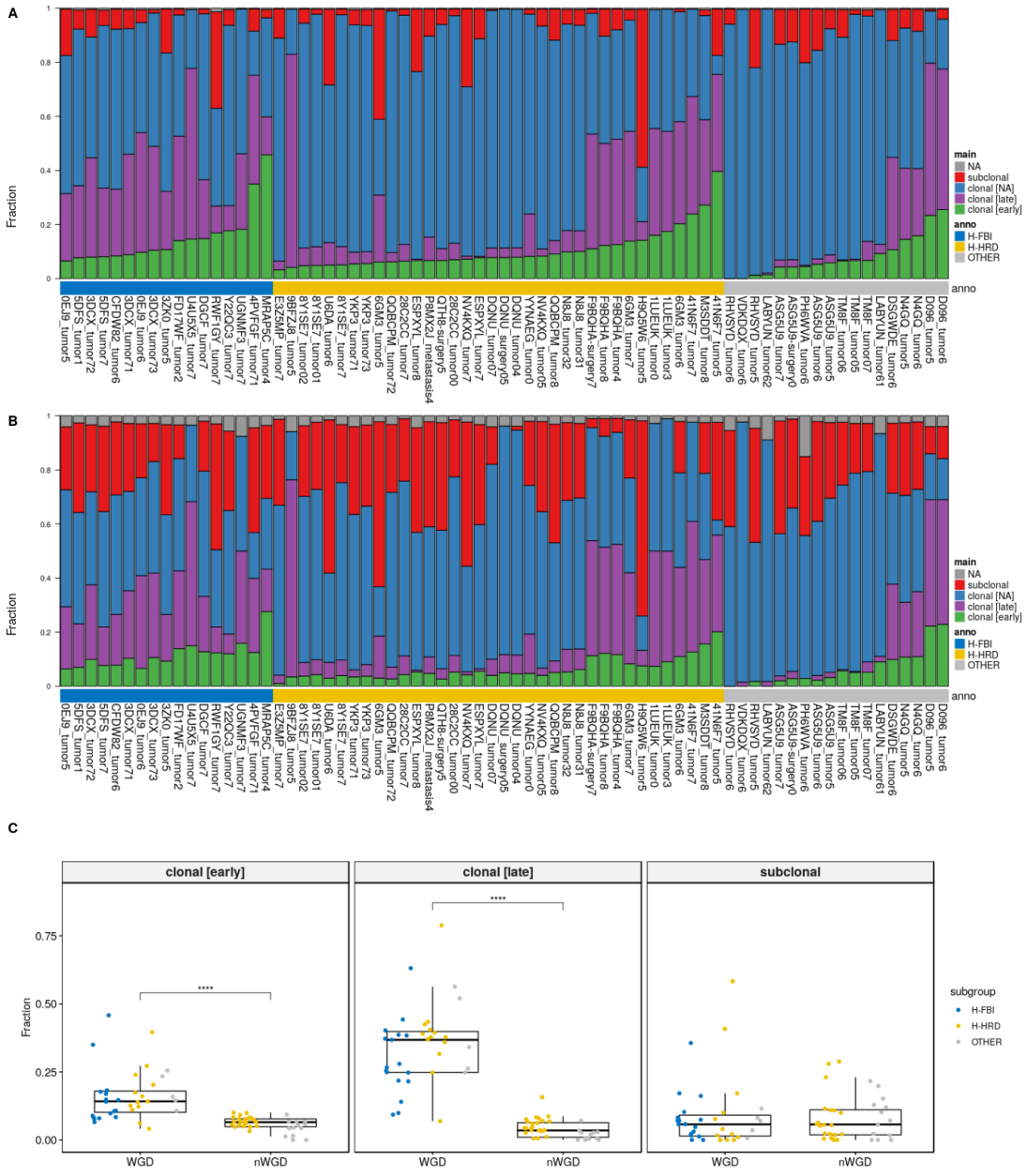


Figure 13.4: Mutations (A) and indels (B) were classified into four timing categories representing different tumor epochs. Samples are ordered by subgroups, including HGSC genomic subgroup (H-FBI, H-HRD) and histotype (OTHER). The fraction of mutations in three categories are further stratified in (C). Each boxplot compares whether WGD-positive (WGD) and WGD-negative (nWGD) tumors have different abundance in mutations of certain category, with data points colored according to subgroups.

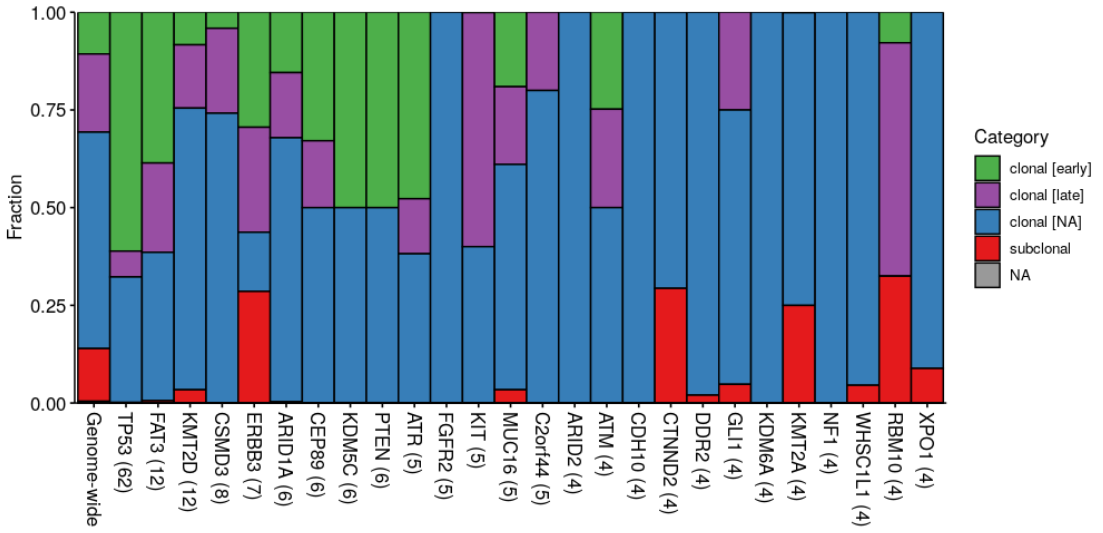


Figure 13.5: Timing of functional variants (mutations and indels) in cancer-associated genes. The x-axis is labeled with the name of the gene and the number of samples harboring functional variants in that gene.

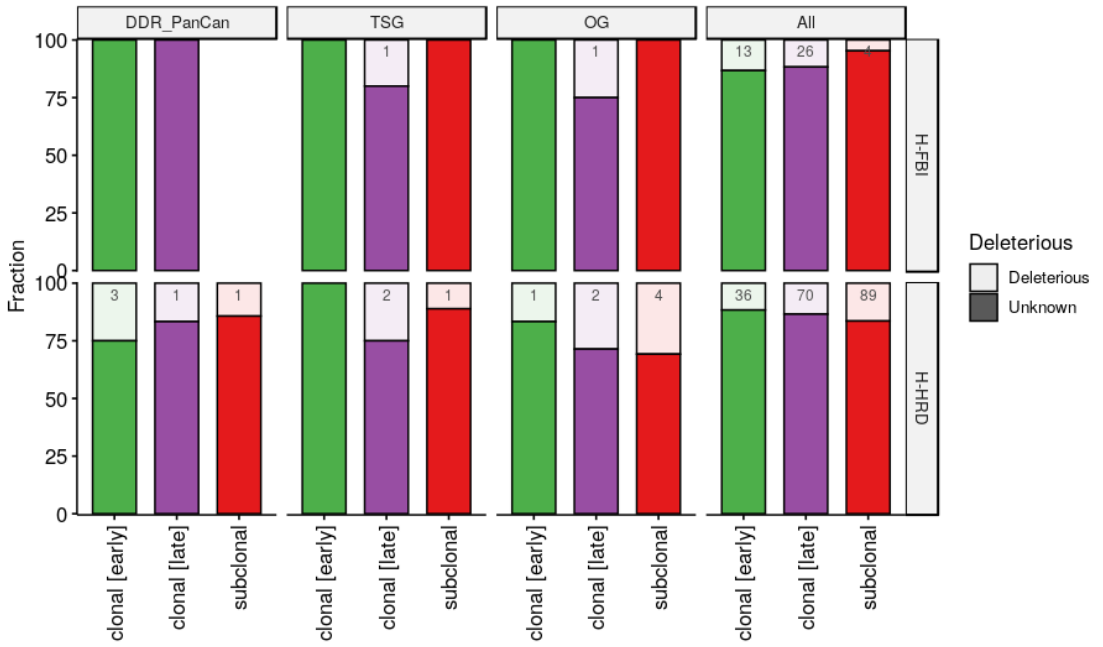


Figure 13.6: Potentially different roles of DDR-related genes, TSGs and OGs during tumorigenesis. Proportions of deleterious variants were profiled temporally and stratified into columns by gene sets (DDR-related genes, TSGs, OGs and all variants) as well as into rows by HGSC genomic subgroup (H-FBI and H-HRD). The frequency bars are colored by tumor epochs, with the lighter portion representing the fraction of deleterious variants and labeled with the number of deleterious variants.

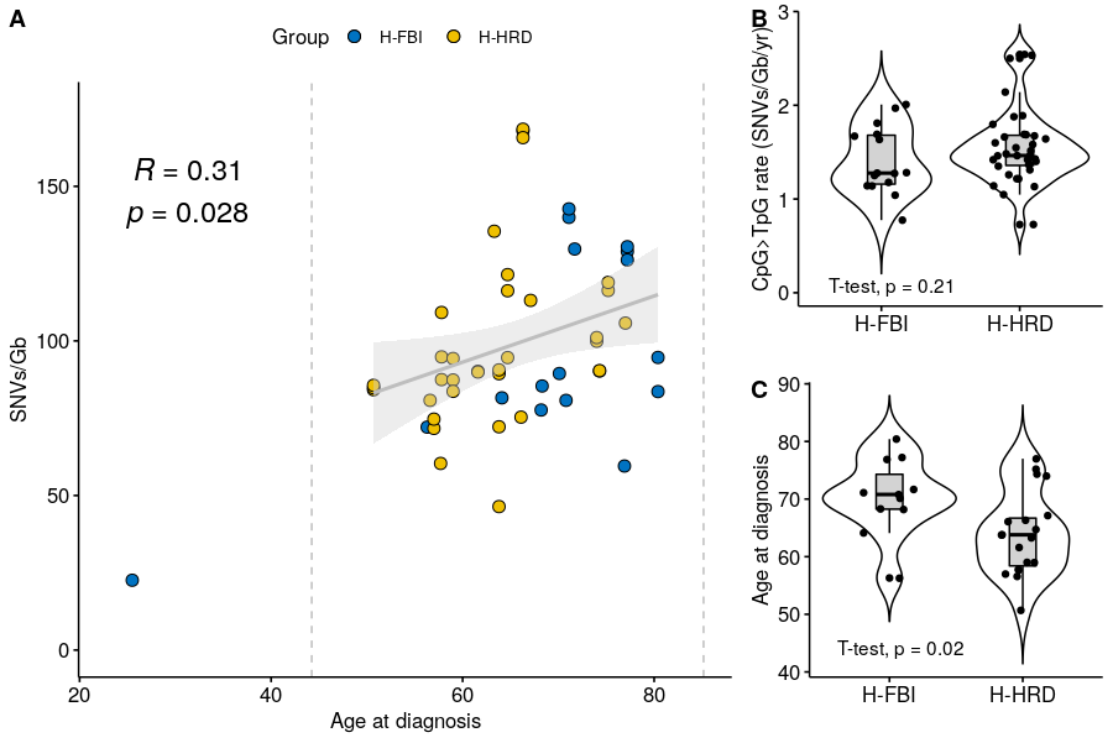


Figure 13.7: The correlation between age at diagnosis and CpG>TpG mutations, the main footprint of COSMICv2 Signature 1, is shown in (A). The CpG>TpG mutation rate (B) and age of diagnosis (C) were compared between HGSC genomic subgroups (H-FBI and H-HRD). Note that patient Ho59-4PVFGF, shown as the blue data point outside the dashed lines in (A), was identified as an age outlier and excluded in all three analyses. The outlier identification used a threshold of *median* $\pm 2 \times$ *Median Absolute Deviation*, where the lower (44 years) and upper (85 years) threshold were shown as dashed lines.

13.3.1 *Examine assumptions for molecular clock*

CpG>TpG mutation number was known to scale proportionally with chronological time[25] and considered a better molecular clock. This assumption was examined in Figure 13.7, which profiles the relationship between mutation and age in HIPO59 HGSC patients excluding one outlier being a patient (Ho59-4PVFGF) diagnosed at exceptionally young age (25 years).

In Figure 13.7(A), the CpG>TpG mutation burden showed correlation with age at diagnosis (R=0.31, p-value=0.028). A linear regression analysis was done and the fitted line shown in grey. The Pearson correlation coefficient increased to 0.45 (p-value=0.00092) when the outlier was included. Furthermore, the mutation rate remained invariable between two HGSC subgroups (see Figure 13.7(B)) and also not dependent of age (R=-0.12, p-value=0.42). Nonetheless, the CpG>TpG mutation burden did not show significant difference between two subgroups (Welch t-test p-value=0.71) despite a significantly different age distribution observed in Figure 13.7(C).

13.3.2 *Timing of WGD, MRCA-PID and MRCA-SAMPLE*

The timing of three major events were calculated according to Section 8.10.3. The mutation time and chronological time are shown in pairs in Figure 13.8 for `WGD` (A,B), `MRCA-PID` (C,D) and `MRCA-SAMPLE` (E,F). In (B,D,F), the chronological time is expressed in latency (years before diagnosis), therefore the baseline time=0 correspond to patient's age of diagnosis. The boxplot on the right gives an overview of the latency of the event in all HIPO59 samples, with the median correspond to the rectangles next to the boxplot at tick "7.5x". As the mapping between two time coordinates assumed a mutation rate acceleration of **A** times, these rectangles ticks correspond to the median latency of the cohort given different parameters of A (1x, 2.5x, 5x, 7.5x, 10x, 15x and 20x).

In ovarian cancer, `WGD` seems to happen rather early in the mutation time (A) and dates back to 2.3 to 65.4 years before diagnosis (B), with a median of 34.6 years. `MRCA-SAMPLE`, on the other hand, mostly emerge at the end of mutation time (E) and can be traced back 0 to 3.5 years before diagnosis (F). `MRCA-PID` occurred either very early or very late (C) and the latency ranged from 0.2 to 33.5 years (D). The chronological time estimates of these three major events in each sample can be found in the Compact disc along with the thesis (**HIPO59_Tumor_Evolution.txt**).

As shown in Section 12.3, the more heterogeneity observed between samples from the same patient, the earlier the divergence in the sample phylogeny tree. When interpreted in chronological time, it implies that `MRCA-PID` is traced back earlier in life. Three patients in High heterogeneity group had `MRCA-PID` estimated at 11.97, 16.13 and 31.05 years separately. One patient from Median heterogeneity group had `MRCA-PID` at 2.69 years. In Low heterogeneity group there were 10 patients with `MRCA-PID` ranged from 0.45 to 4.28 years, with a median of 1.02 years.

Furthermore, as the timing of `WGD` and `MRCA-PID` were estimated separately in each sample, one can compare the event timing between different samples from the same patient. In Figure B.9, 6 out of 7 patient had `WGD` timing estimates close to each other and likely being the same event consistently estimated. `MRCA-PID` estimation from 15 sample sets are shown in Figure B.10, where 9 sample sets had all samples overlapping each other in their 80% CIs. Two sample sets showed a deviated estimate in only the interval debulking sample, which is reasonable as the time lapse between diagnosis and interval debulking surgery was not considered in the model.

Lastly, the median latency of the three events are arranged in time sequence for each patient in Figure 13.9. Only patients with multiple samples as well as with evidence of WGD would have these events connected by segments in the figure. It is observed that the majority of the cases follow a tumor evolution model where `MRCA-PID` appear after `WGD`. The only exception happened in patient H059-41N6F7, where `WGD` might have occurred after `MRCA-PID`.

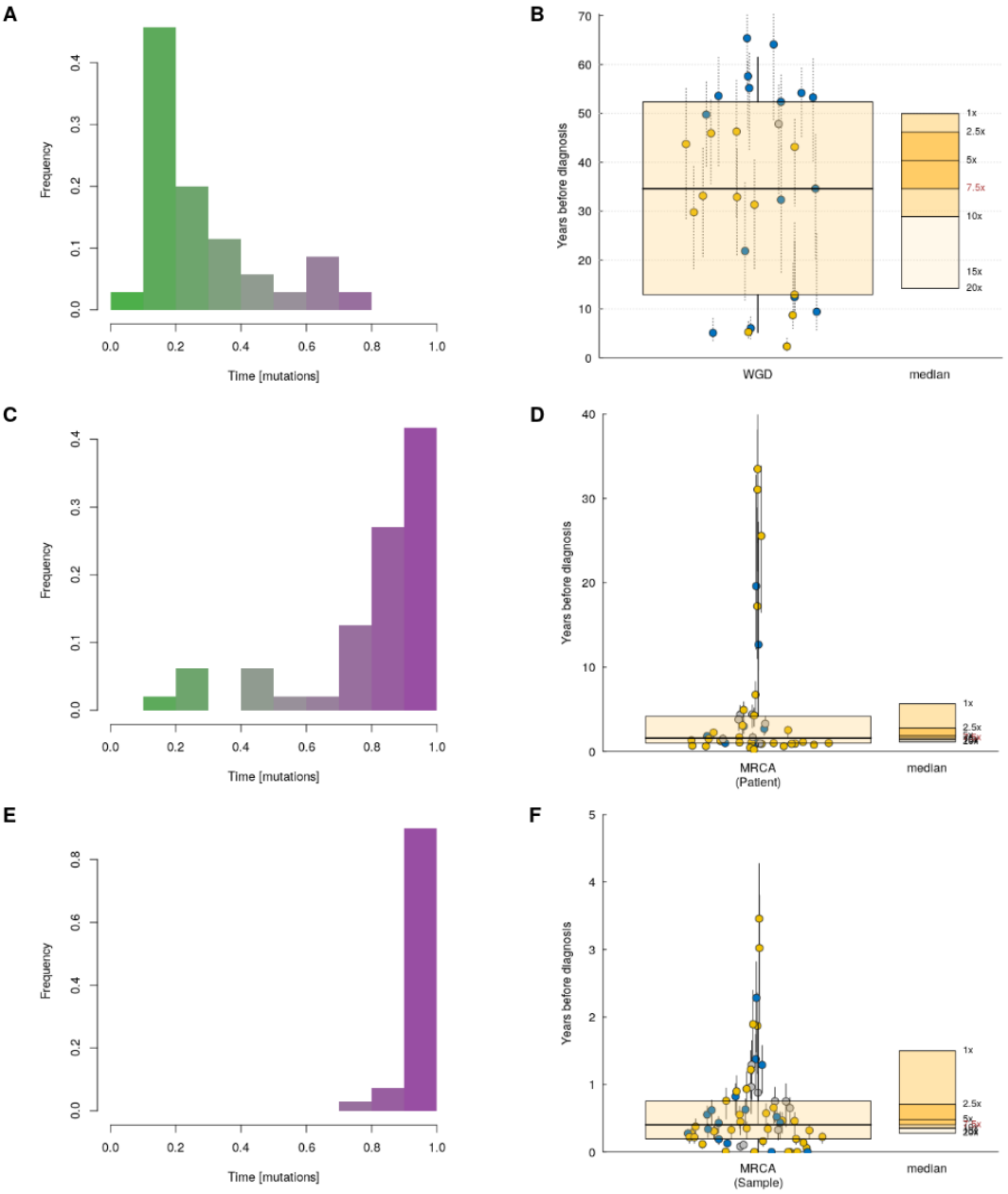


Figure 13.8: The mutation time (A,C,E) and chronological time (B,D,F) were estimated in pairs for `WGD` (A,B), `MRCA-PID` (C,D) and `MRCA-SAMPLE` (E,F). The y-axis in a mutation time plot shows the fraction of the HIPO59 cohort and the frequency bars were colored from green (early) to purple (late). In chronological plots, each data point represents the latency (with 80% CI) of the event in one sample and colored by subgroups (H-HRD: orange; H-FBI: blue; OTHER: grey). The rectangles on the right show the median time of the cohort when different mutation acceleration rate were used in the model.

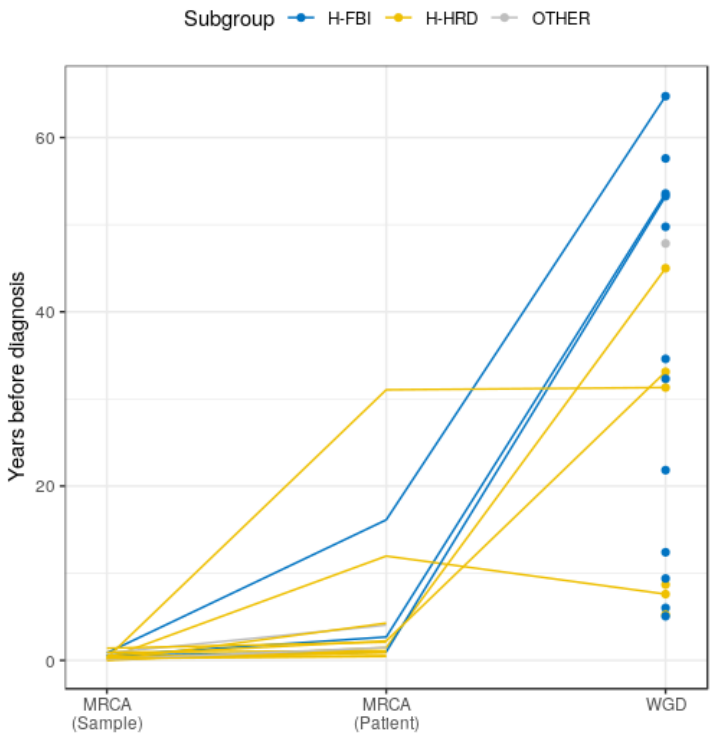


Figure 13.9: Occurrence of `WGD`, `MRCA-PID` and `MRCA-SAMPLE` in time sequence for each patient. Each dot represent the median timing of an event in one patient, and the lines connect three events observed in the same patient.

13.4 TUMOR EVOLUTION IN INDIVIDUAL PATIENTS

As described in Section 8.10.4, the multi-sample design of HIPO59 cohort allows us to look at tumor developmental trajectory with finer time granularity. Based on the results in Section 13.2, Section 8.10.4.1 re-classified small variants into two to four tumor epochs depending on which of the four scenarios in Figure 8.5 a sample belongs to. The objection of this section is to integrate the timing of major events, re-classified small variants and sample phylogeny tree in order to reconstruct the tumor evolution in individual patients.

13.4.1 *Potential cancer-associated variants in refined tumor epochs*

Given re-classified small variants, this section focus on those that are likely associated with tumorigenesis. These include all TP53 variants, deleterious or LOH variants in 42 DDR-related genes profiled in Figure 11.4, as well as deleterious variants in 717 cancer-associated genes in CGC[178] (see Section 6.3). The emergence of these variants in refined tumor epochs are described in Figure 13.10 for WGD-negative cases and in Figure 13.11 for WGD-positive tumors. Note that data from two patients were not shown here, where H059-41N6F7 might had an unusual temporal order of `WGD` and `MRCA-PID`, and for H059-LABYUN there were no relevant variants identified in the low purity tumors (<0.2 for both samples).

Single	Germline	Clonal	Subclonal
H059-E3Z5MP	WRN*	BRCA2, TP53*	
H059-P8MX2J	CSMD3*, RFC2	ATM, TP53*	
H059-QTH8	BRCA1, C17orf70*	NOTCH2, POU5F1, TP53*	MLLT4
H059-U6DA	BRCA1	ERCC4*, TP53*	ERBB3
H059-YYNAEG	BRCA1	NF1*, TP53*	
H059-PH6WVA		TP53*	
H059-VDKDQX	ATM, FANCM, MLH3, RAD1		

Multi	Germline	Clonal (truncal)	Clonal (branch)	Subclonal
H059-28C2CC		ARHGEF10, GLI1, TP53		
H059-8Y1SE7	CUL3*	FAT3*, TP53*		
H059-DQNU	RAD18*	TP53*	REV3L*	
H059-ESPXYL		BIRC6, DDB1, NF1, TP53*		
H059-N8J8	BRCA1	TP53*		
H059-NV4KXQ	DCLRE1B*, POLI*, RAD51C	RAD17*, TP53*	RAD17*	
H059-QQBCPM		RABGGTB*, TP53		
H059-YKP3	CDK12*	CDH10, MLLT6, TP53*		
H059-ASG5U9	BRCA1	TP53*		
H059-RHVSYD	TDP1*		NF1	
H059-TM8F		ARID1A, KDM5C, KMT2D, PTEN		

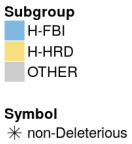


Figure 13.10: The emergence of DDR- and cancer-associated variants in tumor evolutionary trajectory are profiled for patients without WGD occurrence (Scenario "Single" and "Multi" in Figure 8.5). Patient IDs are colored according to subgroups. Genes marked with symbol * are DDR-associated variants that are not deleterious but accompanied by LOH.

WGD_Single	Germline	Clonal (before WGD)	Clonal (after WGD)	Subclonal
H059-3ZK0	LIG3*	TP53*		PAX8
H059-4PVFGF	ERCC2	PPFIBP1		
H059-CFDW82	APEX1, ENDOV*	TP53*	CSMD3	
H059-DGCF		TP53*		
H059-FD17WF			DDB1, TP53*	
H059-MRAP5C	NBN, PARP4*	BRIP1*, TP53*		
H059-RWF1GY	BRIP1*, PALB2*		FAT3*, ROBO2	
H059-U4U5X5	TREX2*	TP53, TP53*		
H059-UGNMF3			NF2, TP53"	
H059-Y22QC3	BRIP1*	CREBBP	TP53*	
H059-9BFZJ8	BRCA1		TP53"	
H059-H9Q5W6	RECQL5*, WEE1*	BRCA2, IL7R, TP53*		BRAF, FGFR4, SPO11, TGFBR2
H059-M3SDDT	FANCG	BRCA1, TP53*		
H059-DSGWDE	FANCC	TP53*		

WGD_Multi	Germline	Truncal (before WGD)	Truncal (after WGD)	Branch (clonal)	Subclonal
H059-0EJ9		TP53		ACKR3	
H059-3DCX	SMARCC1*, UVSSA*	TP53			
H059-5DFS		TP53*			
H059-1LUEUK	DNAH1*	TP53*	DNAH1, WEE1*		
H059-6GM3	ATM, BRCA1	TP53*		BIRC6, HERC2	
H059-F9BQHA	FAT3*, FLNA*, RAD51C*, RNF8*	FAT3*, PMS2*, TP53*	FAT3*, FLNA*, KMT2C, NPM1, POLH, PTPRT	KMT2D, ZEB1	
H059-D096	BRCA1*	TP53			
H059-N4GQ	BLM, FAT3*	BLM, TP53	CNTNAP2	ARID1A	LHFP

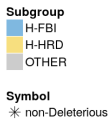


Figure 13.11: The emergence of DDR- and cancer-associated variants in tumor evolutionary trajectory are profiled for patients with WGD occurrence (Scenario "Single(WGD)" and "Multi(WGD)" in Figure 8.5). Patient IDs are colored according to subgroups. Genes marked with symbol * are DDR-associated variants that are not deleterious but accompanied by LOH. TP53 marked with " are not neither deleterious nor accompanied by LOH.

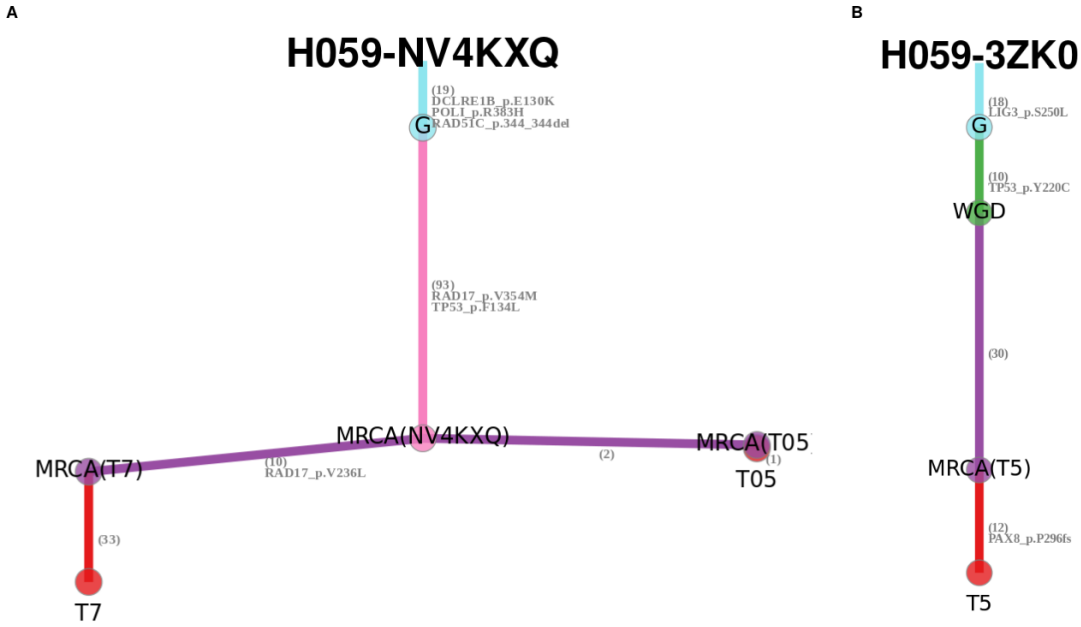


Figure 13.12: Two examples of tumor evolutionary trees for individual patients. Reconstructed sample phylogeny trees integrating timing of major events and small variants. The two examples shown are patient H059-NV4KXQ (A) and H059-3ZK0 (B).

13.4.2 *Reconstructed tumor evolutionary process in individual patients*

The refined timing of small variants, together with the timing of major events (see Figure 13.9), can then be integrated into the tumor phylogeny in Figure 12.4 and profiled for each patient. Two examples shown in Figure 13.12 are sample phylogeny trees for patient H059-NV4KXQ and H059-3ZK0.

In patient H059-NV4KXQ, the MRCA of two samples emerged 1 to 3 years before diagnosis. The samples shared rare germline missense variants in POLI and DCLRE1B, both exhibited LOH. Another shared rare germline deleterious variant in RAD51C was observed with wild type allele lost in tumor7 but not in tumor05. There were also shared somatic missense variants in TP53 and RAD17. In terms of TP53, wild type allele lost was observed in both samples. As to RAD17, a second somatic hit occurred in tumor7 led to lost of wild type allele; while in tumor05 the wild type allele was possibly retained.

In patient H059-3ZK0, `WGD` was estimated to had occurred 12.4 years before diagnosis. One rare germline missense variant were detected in a DDR-related gene LIG3 and was accompanied by LOH. In the somatic setting, a missense variant in TP53 occurred before `WGD` and also had the wild type allele lost. In another cancer-associated gene PAX8, a deleterious variants emerged subclonally.

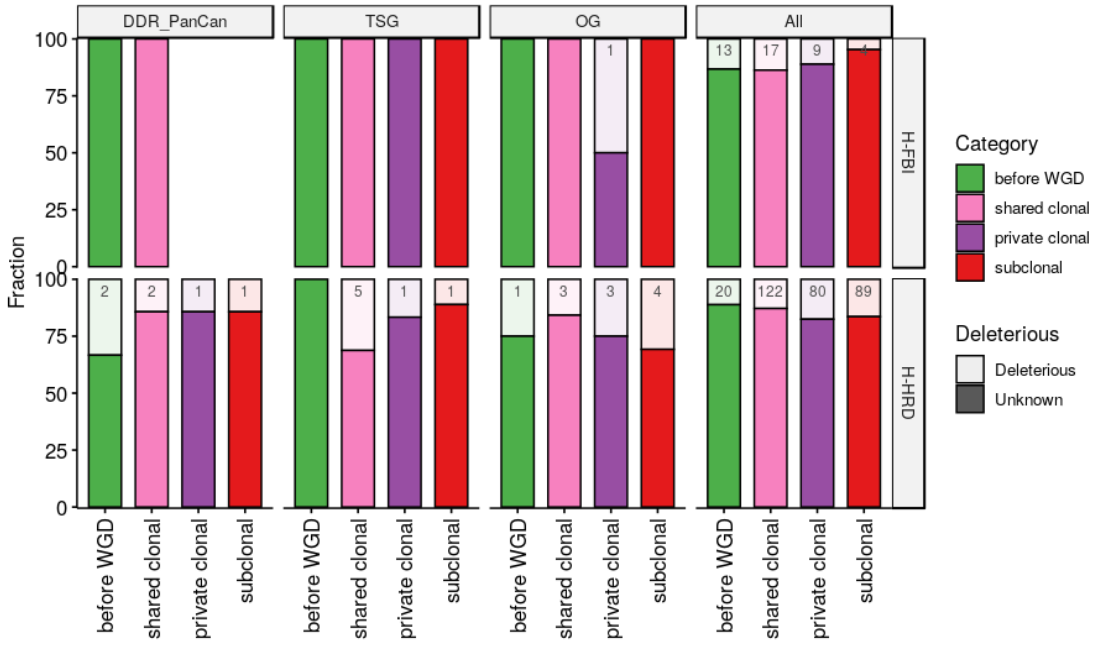


Figure 13.13: Roles of DDR-related genes, TSGs and OGs during tumorigenesis in new tumor epochs. Proportions of deleterious variants were profiled temporally and stratified into columns by gene sets (DDR-related genes, TSGs, OGs and all variants) as well as into rows by HGSC genomic subgroup (H-FBI and H-HRD). The frequency bars are colored by refined tumor epochs, with the lighter portion representing the fraction of deleterious variants and labeled with the number of deleterious variants.

13.4.3 *Role of DDR genes, OGs and TSGs in refined tumor epochs*

With the new timing classification, 2,988 small variants from the three gene sets were timed and 407 of them are deleterious variants. This amounts to a 47.6% increase of variant number compared to the previous timing scheme in Section 13.2.3. The temporal trend of deleterious variant fraction in new tumor epochs is shown in Figure 13.13. As compared to the previous timing scheme (see Figure 13.6), a similar trend is again observed for H-HRD cases, where deleterious variants in `DDR_PanCan` and `TSG` were more likely to be found in the early stages in tumor evolution. On the other hand, deleterious variants in `OG` more likely emerge in the later stage of tumorigenesis, consistent with the overall trend in "All".

13.5 TEMPORAL CHANGE IN MUTATIONAL PROCESS ACTIVITIES DURING TUMOR EVOLUTION

Figure 11.9 previously showed that, at the time of diagnosis, samples from the two HGSC genomic subgroups exhibited different level of genomic footprints for AC1 and AC3. This section aims to dissect the mutational process activities in different tumor epochs. This is done by performing mutational signature analysis separately for variants

in different tumor epochs. In the beginning, mutational process activities were profiled in the original four-epochs identified in Section 13.2. Later on, the activities were profiled with finer time granularity using re-classified small variants (see Section 13.4), which was enabled by the multi-sample experimental design of HIPO59.

Given the chronological time of major events estimated in Section 13.3, the absolute exposure of mutational signatures (Figure 11.5) can be converted into signature rate (SNVs/yr), an estimate of the number of mutations contributed by a specific signature per year. This signature rate is profiled in Figure 13.14 for WGD-positive samples. The left panel shows again that when HGSC is stratified by genomic subgroup, H-FBI had higher AC1 and lower AC3 compared to H-HRD.

The right panel then estimates the signature rate in three tumor epochs, which shows the temporal change of signature rate from early clonal, late clonal to subclonal epochs. Both AC1 and AC3 rate increase along with tumor evolution as modeled (see Section 8.10.3.1), notably H-FBI has greater rate increase in AC1 whereas H-HRD showed greater rate increase in AC3. When comparing the two subgroups within each epoch using Welch t-test, significant difference was observed for AC1 in late clonal (p-value=0.0081) and subclonal (p-value=0.031) epochs, and for AC3 in early clonal (p-value=3.4e-6) and subclonal (p-value=0.014) epochs. To note, despite a borderline significance in late clonal epoch (p-value=0.077), AC3 showed large difference in the median rate in H-FBI (14.5 SNVs/yr) and H-HRD (65.6 SNVs/yr). As a reference, AC1 is significantly different in late clonal epoch, with a median rate of 11.8 SNVs/yr in H-FBI and 7.35 SNVs/yr in H-HRD.

Next, we would like to know whether AC3 signature rate had different temporal changes between H-FBI and H-HRD subgroup. Having observed the difference in AC1 rates acceleration along tumor development, a direct comparison of temporal change in AC3 rate may not be adequate due to their baseline rate differences. To address this question, a normalized new measure is used instead. Specifically, Figure 13.15 looks at relative ratio of AC3 rate to AC1 rate in each epoch, which approximates the relative activity of AC3 to AC1 across tumor development. As the multi-sample design allows for refined tumor epochs (see Section 8.10.4), there will be two to four timepoints reconstructed depend on which of the four scenarios a sample belongs to in Figure 8.5.

As a result, temporal change of relative rate are profiled in Figure 13.15(A) for HIPO59 samples using different scenarios when possible. For example, WGD-positive cases in multi-sample set would have its tumor evolution reconstructed with both "Multi" and "WGD_Multi" scenarios. On the other hand, in Figure 13.15(B) the median of relative rate are taken from multi-sample sets to represent each unique patient, with a single line connecting different timepoints from the same patient. Taken together, it shows that AC3 activity in H-HRD tumors tend to arise earlier and to a greater extent than in H-FBI tumors. Furthermore, it also demonstrates the power of multi-sample design to reconstruct tumor evolution with better time resolution.

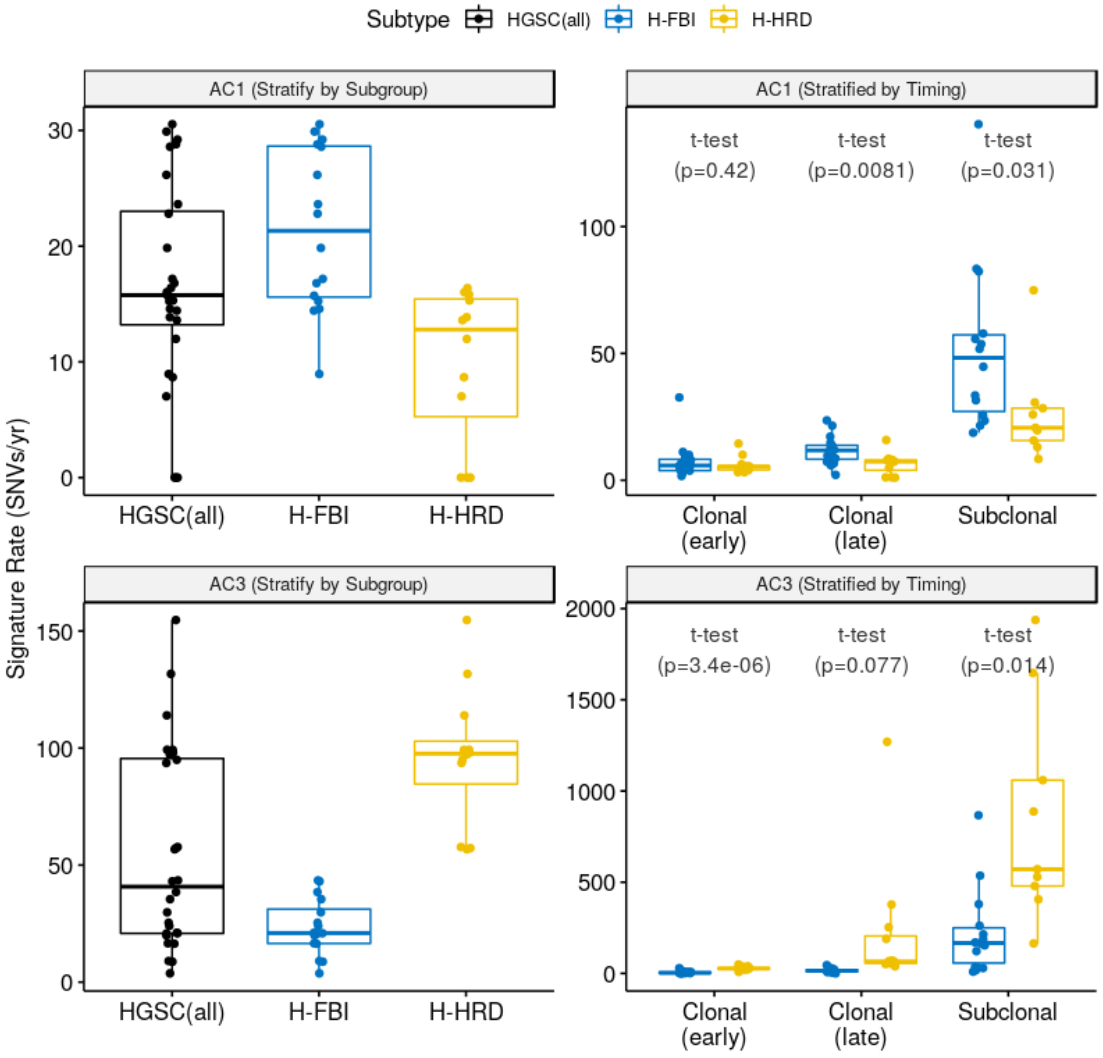


Figure 13.14: Mutation rate of mutational signature AC1 and AC3 are stratified by HGSC genomic subgroups (left panel) and further by tumor epochs (right panel).

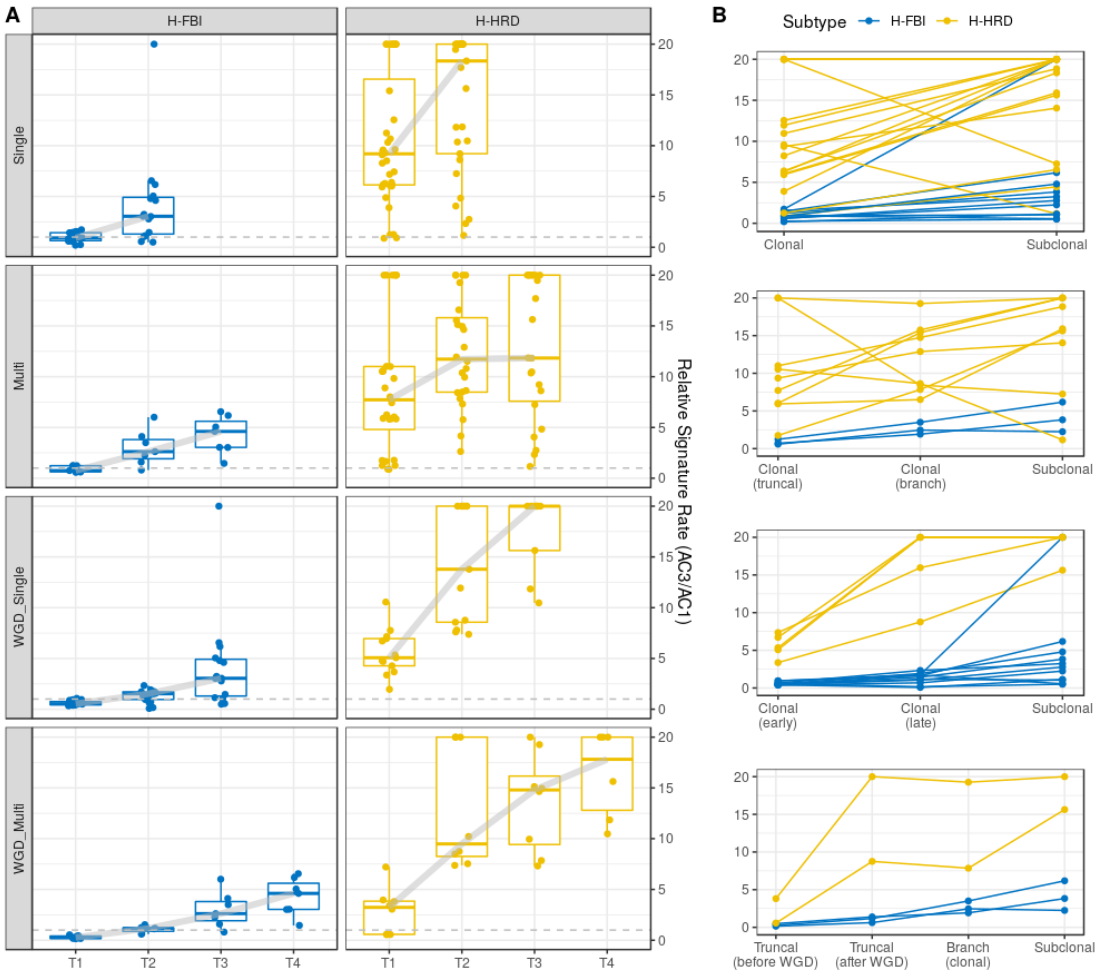


Figure 13.15: Temporal change of mutational signature 3 activity along tumor evolution. For each of the four scenarios (rows), the relative rate (AC3/AC1) are profiled in y-axis for each sample in (A) and for each patient in (B). For visualization purpose, relative rate above 20 are set as 20.

Part IV

DISCUSSION AND CONCLUSION

14 DISCUSSION

This thesis started with characterizing the germline and somatic alteration landscapes of ovarian cancer. Based on HIPO59 cohort, the germline variants in DDRs, somatic variants in SMGs, gene breakage events in SMGs, and focal SCNAs were profiled and compared with public cohorts. Within the HGSC histotype, inter-patient comparisons further revealed intrinsic differences between Shah-2017 genomic subgroups[132]. Supporting evidences showed differences in several layers, including gene alterations in HRR pathway, DNA footprints reflective of HRR deficiency, as well as temporal activity change in HRR during tumor development. These collectively suggest the existence of intrinsic subgroups characterized by different extent and onset timing of HRR defect.

On the other hand, IPH was investigated using spatially separated tumors collected from the same patient. It was shown that IPH status can be consistently estimated by different types of somatic alterations. Moreover, a higher level of IPH may be suggestive of better prognosis. Lastly, for each of the individual patient, an evolutionary tree of all related samples was reconstructed, where potential driving events in DDR genes or cancer-associated genes were marked on different tree branches corresponding to different tumor epochs. In addition, major events like `WGD`, MRCA of the patient (`MRCA-PID`) and MRCA of the sample (`MRCA-SAMPLE`) were also incorporated in the tree, with their chronological time of emergence estimated. These findings conform with current knowledge about ovarian cancer, and additionally provides novel and interesting insight into the sub-classification and clinical management of HGSC.

14.1 GERMLINE AND SOMATIC ALTERATION LANDSCAPES OF OVARIAN CANCER

14.1.1 *The majority of ovarian cancers are fueled by CIN*

Compared to other solid malignant neoplasms, HGSCs were found to have similar tumor mutational burden to other cancer types while few SMGs with mostly low prevalences, together with high level of aneuploidy shaped by recurrent focal SCNAs with high prevalences. In the TCGA-OV cohort, all SMGs, excluding the ubiquitous TP53 mutations, collectively affected only 12% of all patients; whereas recurrent focal SCNAs were found in 94.8% of the samples. Noteworthily, SMGs reported by large HGSC cohorts are more often tumor suppressors and as shown in the ICGC-AU-OV study, their inactivation frequencies can increase to three-fold when SVs were taken into consideration[131].

These observations were recapitulated in the HGSC subset of the HIPO59 cohort. TP53 was the only SMG identified and was affected in most patients. Among all reported SMGs, inclusion of gene breakage events led to more than two-fold increase in inactivation frequencies for NF1 (from 6% to 21%), RB1 (from 0% to 6%) and CDK12

(from 3% to 9%). Notably, NF1 seems to be more often affected by SVs that result in gene breakages than those result in gene dosage change. In terms of focal SCNAs, recurrent peaks identified in HIPO59 had high concordance with those found in TCGA-OV. In HIPO59, 75% of significantly recurrent arm-level gains or losses, as well as 87% of subregions containing nominated genes were also found significant in the TCGA cohort. Taken together, with HIPO59 it was able to recover 1 out of 9 SMGs, as well as 27 out of 102 gene-containing recurrent focal SCNAs identified in TCGA-OV study[130]. The shortage comes from an insufficient power given the smaller sample size of HIPO59 (n=33) compared to TCGA-OV (n=316) cohort, as well as the low prevalence of SMGs which imposed further stringency in detecting recurrent mutations and indels.

From the perspective of functional convergence, a study integrated germline truncation variants and somatic mutations in the TCGA-OV cohort at pathway and network level, and found the most prominent set of altered genes lie around the DDR pathway[205]. In contrast, RB1 pathway, RAS-PI3K pathway and MAPK pathway were more often targeted by CNAs or gene breakage events induced by SVs[130, 131, 205]. Given the fact that DDR defect can be a source of CIN, it is reasonable to hypothesize that the mutagenic events in the early stage are more likely small variants and worked on DDR pathway, which lead to CIN and fueled the cells to achieve additional hallmarks of cancer.

In particular, copy number deletions seem to be a prominent feature as more recurrent losses than gains were observed in HGSC. In HIPO59, the number of chromosome arm-level recurrent losses were three times higher than that of gains. When it comes to focal SCNAs, there were also 1.5 times more recurrent losses observed. Using the driver gene nomination procedure (see Section 8.4.2), novel targets for deletions can be identified despite HIPO59 being an underpowered study. These include deletion of BRCA2 and WRN in the HRR pathway, deletion of PIK3R1 in the RAS-PI3K pathway, and deletion of FBXW7 in RB1 pathway, deletion in SWI/SNF chromatin remodeling complex components ARID1A and ARID1B, as well as genes in 14 other peaks in Table 9.3.

Together, mutations and indels are known to more often target TP53 and the DDR pathway, we further highlight the non-negligible contribution of other alteration mechanisms acting on more tumor suppressor genes, as well as a crucial role of chromosomal instability in shaping the development of HGSCs.

14.1.2 *Germline variants in DDR genes and predisposition to ovarian cancer*

Hereditary ovarian cancer is predominantly attributed to germline pathogenic variants in BRCA1 and BRCA2 genes, hence the long established role of HRR defects in the genetic predisposition for OC. Despite the fact that hereditary cases accounts for only a small subset[60], there is a significant heritability (39%) estimated for OC[63]. As more moderate to low penetrance genes are being identified (see Table 3.2), all of them play roles in repairing DNA damage. Although contemporary studies have limited power to uncover the full spectrum of genetic associations, especially since OC is a heterogeneous disease[40], they suggest a strong connection between DDR

and ovarian cancer predisposition. Furthermore, sporadic OCs overall show complex karyotypic abnormalities. Insights from other cancer predisposition syndromes suggest a link between this CIN phenotype and germline defect in DDR genes. Therefore, in this thesis two approaches were taken to explore potential hereditary cancer predisposition as a result of DDR defect.

The first approach revealed that in the HIPO59 cohort, there is an excess burden of germline rare variants in the DDR pathway. When breaking it down into different subpathways, HRR, FA and NHEJ stood out as preferentially involved. This aligns well with two previous studies re-analyzing the TCGA-OV cohort, including the above-mentioned one looking at pathway convergence of germline and somatic variants[205], as well as the other that examined DDR pathway defect resulting from three somatic alteration mechanisms. In the latter study, the authors found that small variants, CNAs and epigenetic silencing were enriched in NHEJ and HRR axes[19]. These observations collectively suggest that there might be more low penetrance variants leading to suboptimal function that compromised the DDR system, thereby confer predisposition to ovarian cancer, and that among nine DDR sub-categories, HRR, FA and NHEJ axes are of particular importance to OC development. Given that DDR defect is mainly acquired by, while not restricted to, both germline and somatic small variants, it is tempting to speculate that it plays a role in the early stage of tumorigenesis.

The second approach aimed to identify the accountable hereditary factor, if any, for each individual patient. As the prevalence and disease penetrance of a variant vary by its genomic location in a gene[206], the pathogenicity of germline variants were determined by human geneticists instead of by in silico prediction. The results showed that one-third of HIPO59 patients harbor germline pathogenic variants in DDR pathway, and many of the disrupted genes have not been causally associated with OC. Similar phenomena were observed in the TCGA-OV cohort. In addition to showing that germline rare truncations enriched in BRCA1, BRCA2 and PALB2 as expected, scientists further noticed other germline truncation variants in NF1, MAP3K4, CDKN2B and MLL3[205]. Noteworthily, findings in HIPO59 highlights two other important aspects in ovarian cancer predisposition.

First of all, heredity is also an influential factor in other histotypes, as we found 30% of HGSC and 44% of other histotypes harbors germline pathogenic variants. The majority of these cases have a defect in the HRR pathway, including 8 out of 10 HGSC patients and 2 out of 4 OTHER histotype cases. Among these 10 cases with HRR defect, 7 of them have the variants in BRCA1, with its manifestation not restricted to HGSC histotype. To add, the frequency of non-HGSC histotype in BRCA-associated tumors observed in HIPO59 (1 out of 7) is similar to the 14% previously reported[207]. On the other hand, the four cases with defect in non-HRR axes can show low chromosomal instability. Specifically, H059-4PVFGF (HGSC) and H059-VDKDQX (OTHER) show low wGII score of 13.4% and 0.2%, respectively. Although the other two cases had higher wGII score (both > 60%), their CIN phenotype can be attributed to other factors, including somatic BRCA1 mutation in H059-M3SDDT (HGSC) and WGD in H059-DSGWDE (OTHER).

Secondly, although only 2 out of 8 affected genes (BRCA1 and RAD51C) have been shown to be causally associated with OC, the rest are associated with cancer

predisposition syndromes. Four of them are either confirmed (ATM and NBN) or candidate (ERCC2 and FANCC) predisposition genes to *HBOC syndrome*, and the other two genes (FANCG and BLM) cause the autosomal recessive diseases *Fanconi Anemia* and *Bloom Syndrome*, respectively. This observation highlights the potential importance of DNA repair syndromes in ovarian cancer predisposition, which can trigger cascade testing, or can have impact in cancer surveillance in patients with these syndromes.

This cohort is well-suited for investigating germline predisposition in individual patients. With the aid of tumor-normal WGS, it helps us more confidently determine the germline and somatic status of variants, and also unveils unexplored inherited factors without gene selection bias. In-house ClinicalWorkflow further performs rare susceptibility variant discovery by considering their frequencies in non-cancer populations. Despite the comprehensive assay, limitation lies in the interpretation of the genome and pathogenicity can be unambiguously evaluated only in a very small fraction of the genome. As such, human geneticist evaluated 67.3% of the germline variants as either GUS or VUS. Furthermore, the evaluation is a manual process and only a manageable amount of prioritized variants were sent for review. We note that some germline non-missense variants in DDR pathway were not evaluated, these include 19 variants in 17 genes (APEX1, APTX, BRCC3, MDC1, MSH3, NUDT18, POLG, POLM, RAD1, RAD51B, RAD9A, RFC2, RFC4, RNF4, SOX4, WDR48 and XRCC5).

Nevertheless, our results revealed novel clinically relevant germline pathogenic variants other than in BRCA genes. This information can be beneficial for patients and their families, which would have gone otherwise unnoticed given their lack of established association with OC. The observed excess of rare germline variants in DDR pathways further support the hypothesis that BRCA genes may not be the only contributing factors. Together with the 29 common susceptibility alleles reported as of 2017[208], there might be a polygenic effect of germline variants, where many low penetrance variants collectively exert an non-negligible effect on ovarian cancer predisposition.

14.2 UNVEILING THE DICHOTOMY IN HIGH-GRADE SEROUS OVARIAN CARCINOMA

14.2.1 *Reproducibility and robustness of HGSC subtypes*

There has been an extensive interest in discovering HGSC molecular subtypes, especially based on transcriptome data. Although clinical grade classification assays had been developed for expression-based subtyping[162, 163], it has been observed that the subtyping is influenced by tumor microenvironment[164, 165] and also varies spatially[162, 165] and temporally[168]. Therefore, its utility as a robust subtyping tool is subject to limitations owing to this dynamic behavior. In this thesis two prominent classifications were applied to the HIPO59 cohort, where TDP subgroup[169] possibly informs biological mechanisms and Shah-2017 subgroup[132] is with potential prognostic relevance.

The Shah-2017 method[132] used 20 genomic surrogates as readout for DNA repair mechanisms, based on an assumption that these features might inform class discovery. With this approach, the authors classified four histotypes in OV133 cohort into 7 clusters. This revealed two genomic subgroups within HGSC histotype, H-HRD and H-FBI. Among the 20 features used, those discriminative between different clusters can be found in Table S8 in the original publication[132]. The classification of HGSC was implemented in the thesis as an adjusted method Shah18, which yielded consistent result when applied to 94% of HGSCs belonging to H-FBI (41%) and H-HRD (53%) clusters. It is important to note some potential consequences of the simplification made in Shah18.

First of all, like the remaining 6% of the OV133 HGSCs residing in other clusters, one would expect two HGSC cases in HIPO59 intrinsically more similar to other clusters while not detected by Shah18 method. Their existence can be revealed by looking at the cluster-discriminating features. For example, H059-F9BQHA had higher "S.MMR" and "Frameshift" features and probably belongs to the "E-MSI" cluster. Some samples showing high "S.APOBEC" feature would have been classified into "C-APOBEC" cluster. In fact, implementing the cross-histotype 7-cluster system would be a non-trivial task. An unsupervised approach directly pooling the entire OV133 and entire HIPO59 cohort would not yield the original 7 clusters. Instead, a more supervised and sophisticated approach should be designed for routinely implementing the 7-cluster classification, if needed. In reality, even with well-established subtypes, from its discovery to a robust single-patient tool can be another field of research, therefore the compromise made in Shah18.

Secondly, there were 11 discriminative features between H-HRD and H-FBI, and one of them ("BalancedRearrangement") was not included in Shah18. To examine the importance of these 20 features to the 2-cluster stratification, a PCA was done on the 94% of OV133 HGSCs belonging to either H-HRD or H-FBI (data not shown). It was observed that PC1 explained 21.6% of the variance, and the contribution of "BalancedRearrangement" to PC1 ranked 7th among all input features and accounted for 7.1% of PC1. As a comparison, when another PCA was performed on the same patients with two features eliminated as did in Shah18 (data not shown), PC1 explained a similar fraction (22.6%) of the variance. Indeed, Section 10.2.1 also showed that the two missing input features did not have large impact on the classification result.

Lastly, prognostic differences between H-HRD and H-FBI were not observed in HIPO59. This can be due to the simplification from 7-cluster to 2-cluster classification, where samples belonging to other clusters were not excluded in the survival analysis. This might be especially impactful on the analysis of a small cohort.

Up to this point, it was shown that despite losing some discriminative input information, Shah18 was able to capture most of the inter-patient variability and reproduce HGSC subgrouping. This helped us cluster HIPO59 HGSC patients into two genomic subgroups, corresponding to the H-HRD and H-FBI reported in the literature. As to the minority of cases that should have gone into other clusters, they can be partially identified by discriminative features.

Mounting evidences show that TDs of different sizes are associated with specific biological mechanisms[23, 102, 157, 158]. Based on these characteristics, the TDP classification stratify TDs into five size classes (class 0: <1.6kb, class 1: 1.64-51kb, class 2: 51-622kb, class 3: 622kb-6.2Mb, class 4: >6.2Mb) to assess their corresponding features in each patient. In this thesis the algorithm was implemented and almost perfectly reproduced the result in the original publication.

However, when applied to the HIPO59 cohort, the distribution of active TD classes was a bit different from that in public cohorts (see Figure 10.3). In particular, 82.9% of major peaks found in HIPO59 representative samples fell in class 1, class 2 and class 3 categories, this is considerably less than the reported 95% in the original publication. The discrepancy is mainly due to the increased frequency of class 4 TDs in HIPO59. The mechanism associated with class 4 TDs was not known due to its rarity in the public cohort, nor can we conclude whether this is due to different SV callers unless we have re-processed their raw data.

When comparing the three major TD classes and corresponding gene inactivations (see Figure 9.10), BRCA1-mutated tumors account for half of the patients having class 1 TD activation. However, this does not include H059-DGCF, the patient with BRCA1 gene breakage event. With a manual check, the breakage was induced by copy number event, however there are possibly still two intact alleles retained and potentially maintained its normal function.

On the other hand, all tumors with potential CDK12 disruption did not show class 3 TD activation. This applies to even the case (H059-YKP3) having a rare germline missense variant accompanied by LOH in CDK12 (p.R379H). After manual inspection, one of these patients (H059-ESPXYL) seems to have a class 3 peak, while this can not be easily resolved due to its overlap with a neighboring class 2 peak. Upon checking their relationship with class 4 TD, one-sided Fisher's exact test suggests a lack of evidence of association (p-value=0.6) despite two of these patients (H059-H9Q5W6, H059-YYNAEG) showed class 4 TD activation. To this point, it is not yet clear which tumor had genuine insufficiency in CDK12 function, therefore a further confirmation on CDK12 functional status would help to resolve its association with different TDP subgroups.

In terms of subtyping variation in related samples, Shah-2017 seems invariant to anatomical position as all related samples from the same patient were classified into the same subgroup. This invariance was also observed in the multi-sample cohort published by Dr. Shah's group[165]. On the other hand, TDP subgroup assignment can be different for related samples from the same patient. Interestingly, 71% of the inconsistent subgroup are due to activation of class 2 TDs, which potentially suggest its role in the later stage of tumor development in these cases.

Together, the two DNA-based classification systems were robustly reproduced in HIPO59, where Shah-2017 identified genomic heterogeneity and TDP subgroup revealed distinct mechanisms in histologically homogeneous histotype.

14.2.2 *New evidences corroborating intrinsic subtypes of HGSC*

Shah-2017 classification[132] segregated HGSCs into two genomic subgroups. H-HRD is associated with BRCA-linked tumors and shows footprints of HRR deficiency; on the other hand, H-FBI is enriched for foldback inversions and over-represented with CCNE1 focal amplification. In the supplementary data the authors provided more subgroup-specific features. GISTIC analysis (Table S10 in [132]) showed also recurrent focal copy number deletion in PTEN for H-FBI; whereas frequent gains in MECOM, MYC, CCND1 and recurrent deletions in RB1 occurred in H-HRD. Structural variations affecting NF1 happened more often in H-HRD whereas RAD51B was disrupted more often in H-FBI. Lastly, the mutation load is higher in H-HRD, however gene alterations in BROCA gene panel (Table S9 in [132]) showed no association with either group. This thesis attempted to further fill gaps in the understanding of genomic subtype-specific features, including DDR defect landscape, comprehensive mutational processes footprints, as well as more mechanistic details.

First of all, driving events potentially causing DDR defect were found enriched in H-HRD when compared to H-FBI, and some known predisposing genes tend to associate with specific subtype. The enrichment analysis included germline pathogenic variants and truncal somatic deleterious variants, following the assumption that small variants may be the main cause of the DDR defect in the early stage of tumorigenesis. Moreover, a mutual exclusivity between germline and somatic deleterious events was observed in H-HRD but not in H-FBI. Although we do see 75% of driving events in H-HRD coming from the well-known TP53, BRCA1 and BRCA2 genes, evidences of biallelic inactivation provided more confidence that the less known drivers are true. In H-HRD, the four cases with less known drivers were H059-ESPXYL (DDB1), H059-P8MX2J (ATM), H059-NV4KXQ (RAD51C) and H059-F9BQHA (POLH). Except for DDB1, all other three cases had lost the wild type allele of the driver gene in at least one of the samples. In terms of H-FBI, 50% of the driving events came from TP53, and the rest three cases had wild type allele retained, including H059-MRAP5C (NBN), H059-4PVFGF (ERCC2) and H059-FD17WF (DDB1).

When taking all variants into considerations, interesting subgroup-specific connections popped up for some ovarian cancer predisposing genes. It was noticed that three patients with BRIP1 variants and one patient with PALB2 variant (all with wild type lost) were exclusive to H-FBI subgroup, and two patients with RAD51C variants (wild type lost observed in all three samples from H059-F9BQHA, as well as tumor7 from H059-NV4KXQ) were restricted to H-HRD subgroup. Although the case number is too small to make a conclusion, the only BRIP1 somatic variant found in OV133 was documented in supplementary data (Table S9 in [132]) to had occurred in a H-FBI case.

Secondly, more comprehensive DDR footprints confirmed HRR deficiency as a common feature of H-HRD cases. In the Shah-2017 class discovery, the authors reported 11 discriminative features for H-HRD and H-FBI (Table S8 in [132]). These features, described below with those of higher importance highlighted, include higher **S.AGE**, **S.MMR**, **S.POLE**, **Foldback.Inversion** and **Inversion** in H-FBI, as well as elevated **S.BC**, S.HRD, S.APOBEC, BalancedRearrangement, Nonsynonymous, CN.Loss in H-

HRD. Among them, the only feature reflecting HRD footprint is S.HRD, the SBS Signature 3. This suggested that S.HRD had significantly different activities in the two subgroups; however it is not distinguishing enough to separate the subgroups.

In this thesis multiple levels of genomic footprints were analyzed, all of them found more genomic scars from HRR deficiency in H-HRD cases compared to H-FBI cases. Indel signatures showed higher abundance of NHEJ footprint (ID6 and ID8) in H-HRD, suggesting an active alternative repair mechanism when HRR is incompetent. Rearrangement signatures also showed higher RS3 (BRCA1 inactivation) and RS5 (BRCA1 or BRCA2 inactivation) in H-HRD. More importantly, CNA-based genomic signature showed higher LOH score and HRD score in H-HRD. This analysis also pointed out that TAI score may be a non-discriminative feature between two genomic subgroups, therefore probably not an important component in the composite HRD score. When TAI score was excluded from HRD score, the AUC was observed to increase from 0.8954 to 0.8995. The result further suggests that an assay combining only LOH and LST score may outperform the current two clinically used genomic scar assays, which focus on LOH score alone or on composite HRD score.

Collectively, these data support that HRR deficiency is a common feature of H-HRD cases by demonstrating its footprints appeared in mutations, indels, structure variations and copy number profiles. More importantly, the result connected genomic subgroups to existing surrogate biomarkers for PARP inhibitor response used in the clinical trials. Together with the chemosensitivity observed in H-HRD group in OV133 cohort, it might be possible that genomic subgroup can itself serve as a robust biomarker for predicting chemosensitivity as well as PARP inhibitor treatment response. Among the various readout for HRD, the thesis had shown extensive evidences in the DNA footprints; however, additional functional assays or in vivo models will be helpful in dissecting more details and confirm its clinical relevance.

Thirdly, in both OV133 cohort and HIPO59 cohort, S.HRD was observed in both genomic subgroups despite with a lower fraction in H-FBI. This prompt the question whether targeting HRD would be effective in the H-FBI subgroup. As DNA footprint is reflective of ongoing and historical mutational process, in the thesis this question was addressed by reconstructing the temporal changes in mutational process activities. The result showed that the process generating S.HRD had in general higher relative activity (compared to SBS Signature 1) in H-HRD and its footprint was already observed in the truncal part of the sample phylogeny tree, suggesting an early onset in the evolutionary process and likely a driving role in these tumors. On the other hand, this process had a lower relative activity in H-FBI tumors and its onset was found more obvious only in the branch or subclonal part of the sample phylogeny tree, suggesting a later onset in tumor evolution. This delineated the mechanistic difference of HRD in the two subgroups despite its footprint found in both. In H-FBI, this process is probably a late or subclonal event thus not serving as the driving force, also less likely is it essential for all tumor cells.

This hypothesis goes along with the observation that PARP inhibitor generates benefit in patients with BRCA genes inactivation or with high LOH score[106] compared to the rest that are without. However, this does not explain another subset of platinum-

sensitive patients reported to benefit from PARP inhibitor[105] that are with lower HRD score. If this implies a subset of H-FBI tumors actually showing chemosensitivity and somehow respond to PARP inhibitor, it would be very interesting to deep-sequence their tumors and find out their associated features as well as what made them sensitive to the treatment. However, it could also be that these were actually H-HRD tumors that were not properly detected by the HRD score assay. Overall, the mechanistic difference of HRD revealed here provided an opportunity to better answer and formulate these questions.

Fourthly, in the above-mentioned TDP classification, three classes of TDs inform different biological mechanisms. Putting the two classifications together, class 2 TDs were further examined since they were reported to be associated with CCNE1 pathway activation. For patients with H-FBI tumors, there were 9 (69%) out of 13 cases exhibiting major TD peak in class 2, despite that 3 of them were TDP-negative. Notably, class 2 TD peaks were consistently observed across different deposits of all three multi-sample patients, which probably signifies a truncal activation. There were two more FBI tumors possibly showing similar feature, one TDP-negative case (H059-FD17WF) seems to have a class 2 peak after manual inspection, and another TDP-negative case (H059-Y22QC3) had major peak identified in class 3 while falls very close to the boundary to class 2. In contrast, 8 (40%) out of 20 patients with H-HRD tumors had major peak in class 2. Four of them were with multiple samples and showed inconsistent class 2 TDs across different samples, indicative of a branch event.

When it comes to class 1 TDs, 16 (80%) out of 20 H-HRD patients had major TD peak in class 1, even if 4 of them were TDP-negative. Among them, 7 out of 8 multi-sample patients showed consistent peak in class 1. By contrast, only 3 (23%) out of 13 H-FBI cases had class 1 peak and one of them is possibly false positive after manual inspection.

Together, the TDP subgroup classification suggested that H-HRD tumors mostly show feature of BRCA1 deficiency and it is likely an early and truncal event. Some of them may have CCNE1 pathway activation but it is more likely a late and branch event. In contrast, H-FBI tumors often have the characteristic of CCNE1 pathway activation and possibly as an early and truncal event in these tumors. This is supported by previous observation that CCNE1 dysregulation can be detected already in STIC lesions, and in preclinical models CCNE1 expression imparted malignant features to p53-compromised untransformed cells[209].

Lastly, our data show that WGD existed in all H-FBI tumors, whereas only 7 (35%) out of 20 H-HRD patients showed WGD. Also, H-HRD patients showed younger age at diagnosis, which aligns with the fact that S.AGE (SBS Signature 1) being a discriminative feature for the two subgroups, as well as the reported younger age in BRCA mutation carriers[91, 93].

Together, these findings are consistent with the picture of a dichotomy in HGSC histotype, and further provided more mechanistic underpinnings of these two genomic subgroups.

14.2.3 *A new perspective to interpret biomarkers and previous knowledge*

Given the evidences so far, it is tempting to speculate that two mutually exclusive subgroups constitute the majority of HGSCs, where BRCA-associated tumors and CCNE1-amplified tumors are two representatives of them. Researchers initially identified subgroups by HRR gene alterations or by CCNE1 amplification status[130, 131], and later on by genomic surrogates[132]. In the end these are just different ways of uncovering the same intrinsic subtypes.

Since the TCGA study observed the mutual exclusivity between CCNE1-amplified tumors and BRCA-associated tumors[130], more studies had confirmed this observation[131] and tried to explain this phenomenon. Some of these studies provide important clues for the dichotomy described here. In 2013, the Bowtell laboratory performed a dependency screen analysis on CCNE1-amplified cell lines to explore the vulnerabilities in these cells. The authors identified BRCA1 and other DDR genes, among others, to be essential for the survival of these cell lines[210]. Later in 2014, Dr. Ronny Drapkin's group [209] further showed that CCNE1-overexpressing cells upregulate some HRR component to deal with DNA replication stress. These findings possibly suggested that DDR genes in H-FBI tumors play essential an role in replication fork protection, therefore tumor cells harboring both defect would be less viable and have been depleted during tumorigenesis, which results in the dichotomy observed. Nonetheless, further investigation into the mechanisms leading to the death of these cells would likely provide more insight in targeting the H-FBI tumors.

Due to the specific features associated with subtypes, it is also important to discuss prognostic biomarkers in the context of the inherent structure, so as to reduce confounding effects. If it was the genomic subgroup that harbors true prognostic implications, then it seems reasonable to have found CCNE1 amplification as a poor prognostic biomarker and BRCA mutations as a better prognostic biomarker, as well as finding genomic scars reflective of subgroup-specific process being biomarker for better prognosis. Whereas if it was the BRCA status that harbors true prognostic information, it also makes sense that CCNE1 amplification and genomic subgroup both shown prognostic implication. This confounding effect was notable when the TCGA-OV study reported no survival disadvantage for CCNE1-amplified patients when restricted to BRCA genes wild type cases[130]. The authors further re-evaluate previous studies[149, 150] and proposed that previously reported survival difference for CCNE1 can possibly be explained by better survival of BRCA-associated cases (Fig. S8.15 in [130]). In the end, a multivariable model including all these factors would be able to properly address this question and identify which of these variables harbored true prognostic relevance.

Similarly, some of previously found associations can also be explained by this dichotomy. For example, amplification or transcription of 8q24 (containing MYC) were found associated with tumors with HRD or germline BRCA1 mutations[130, 131, 211], which is consistent with MYC amplification being one of subgroup-specific feature in H-HRD. Also, our finding of ubiquitous WGD in H-FBI is in line with a previous report suggesting polyploidy as a specific feature for CCNE1-amplified tumors[212].

In summary, the dichotomy provides a new perspective to interpret contemporary knowledge about HGSC, and a re-interpretation of this information helps scientists to formulate further experiments addressing questions about subtype-specific pathogenesis mechanisms or targeting strategies. These hopefully ultimately lead to more solid understanding about the biology of HGSC.

14.3 TUMOR HETEROGENEITY AND ITS PROGNOSTIC IMPLICATION

In Chapter 12, pre-treatment IPH in ovarian cancer was studied, it was also demonstrated that the proposed quantitative measures can potentially stratify patients into subgroups with different survival outcome. Before we start, it is important to note that it was inter-tumor heterogeneity investigated here, which does not fully correspond to the frequently discussed intra-tumor heterogeneity. As a tumor bulk does not always consist of a single tumor clone, the heterogeneity measured here are between two tumor samples, or precisely, between two tumor ecosystems.

Specifically, the analysis started with proposing a general method to quantify IPH based on somatic alterations. Three derived measures are based on either small variants (`HET_MutIndel`), structural variations (`HET_SV`) or copy number profiles (`HET_CNA`), and all of them gave consistent estimations of sample heterogeneity. These three measures can further stratify patients into three groups (High, Medium, Low) by the degree of heterogeneity. When restricted to HGSCs, patients in higher IPH group showed a trend toward better overall survival.

To note, when repeating the analysis, I found the heterogeneity group assignment changed due to the random seeding step in the k-means clustering, which can ultimately affect the prognostic association. This instability in classifier can be due to the small sample size (n=15), but in the case of high discriminative input features it would not be obvious. When looking into the input measures, the classifier was observed to have mainly relied on information from small variants and structural variations. This is likely due to the fact that `HET_MutIndel` and `HET_SV` both followed a trimodal distribution, indicating an inherent structure of three subgroups. Instead, `HET_CNA` showed a bimodal distribution and is therefore not able to distinguish the three subgroups well. In this sense, an alternative measure better capturing the inherent three subgroups, if it exists, may be helpful in improving the classifier stability.

The relationship between heterogeneity and survival observed here might seem counterintuitive to the common belief that higher intra-tumor heterogeneity is more likely associated with worse outcome. Nonetheless, they are related yet different concepts. In fact, our observation might go along with a previous finding by Dr. James Brenton's group. In 2015, the group conducted a study on a cohort of 135 samples derived from 14 patients, where Schwarz *et al.* quantified inter-tumor heterogeneity using solely a copy number-based measure. The authors further proposed a Clonal Expansion index (CE index), and found that patients with higher CE index showed worse survival [213]. Given that CE index reflects spatial clustering of samples in the mutational landscape, a higher CE index would suggest the existence of a group of genetically similar samples observed in the patient, which correspond to the concept of

low heterogeneity group in our classification. Therefore, the low heterogeneity group identified in HIPO59 may be indicative of tumors with higher clonal expansion potential and thus led to a trend of worse prognosis.

In summary, a variable degree of heterogeneity was observed in pre-treatment samples of OC, which agrees with a previous finding that further suggested this high genomic diversity arose early as it already existed in fallopian tube lesion[214]. A general method of summarizing IPH was proposed here, and the fact that it gave consistent measurements across three different types of alterations (small variants, CNAs and SVs) supports its compatibility with a wider variety of high-throughput profiling platforms. The observed relationship of heterogeneity and prognosis further suggests its capability of capturing the clonal expansion phenomena in patient. Nonetheless, a readily utility in the clinic requires further validation of the concept in a larger cohort, and a standardized protocol ensuring proper sampling of locations and robust sequencing assays. Another interesting observation in our study was that most non-HGSCs constitute the medium heterogeneity group and the majority of low heterogeneity group were from HGSCs. However, whether this association is related to intrinsic histotype differences is not conclusive at the moment.

14.4 TUMOR EVOLUTION IN OVARIAN CANCER

14.4.1 *Inferring tumor evolution at the time of diagnosis*

An understanding of tumor progression requires knowledge of temporal acquisition of alterations. A conventional approach toward answering this question is to study samples acquired at different tumor developmental stages. Nonetheless, precursor or dysplastic lesions are sometimes not clinically identifiable or difficult to obtain, especially in the case of ovarian carcinoma where its asymptomatic nature precludes collection of early-stage disease tissues. Despite that prophylactic surgery allowed for acquisition of precursor lesions in the fallopian tube, a special protocol (SEE-FIM) is required to increase the scant chance of discovering such lesions in high-risk individuals. In this thesis, an alternative approach was taken to study the carcinogenesis sequence.

Given samples taken at the time of diagnosis, computational methods were used to reconstruct the tumor evolutionary trajectory in each patient. Specifically, the multi-sample design allows for sample phylogeny tree construction and `MRCA-PID` identification based on somatic variants found in all related samples. Small variants were assigned to either the trunk or the branch part of the tree, given its relative position to the `MRCA-PID`. Additionally, variant timing technique helps stratify small variants into different tumor epochs. When integrated, small variants could be segregated into (at most) four refined tumor epochs based on their emergence relative to three major events, namely `WGD`, `MRCA-PID` and `MRCA-SAMPLE`. These refined tumor epochs can then correspond to the branches in the tree. Lastly, the real-world time when the three major events emerged was estimated by modeling the mutation accumulation process during tumor development. This information was put together in the sample phylogeny tree, with potential driving events in DDR gene and deleterious variants in cancer-associated

genes further marked on the tree branches. As a result, the tree provide information of potential early and late driving events for tumors in each individual, and also gives an idea of when the major events had occurred.

14.4.2 *Major events in the tumor evolution*

In the majority of the patients, tumors followed the order of a very early acquisition of `WGD` (a median latency of 34.6 years), a variable latency of `MRCA-PID` ranging from 0.2 to 33.5 years, and eventually a relatively shorter latency (a median of 0.4 year) of `MRCA-SAMPLE`. The observation that WGD usually occurred very early in ovarian cancer had been noted in the recent work of Gerstung *et al.*. In the PCAWG cohort, they estimated that OCs had a median latency of 14.1 years for `WGD`, as well as a similar median latency of 0.27 year for `MRCA-SAMPLE`. This can be put together with another study comparing fallopian tube lesions, ovarian cancers and metastases in 2017. The authors estimated that the average time between STICs and ovarian cancer in 9 patients was 6.5 years, and the time between ovarian cancer and metastases was on average 2 years[215]. Although the estimated timepoints were not exactly the same, there seems to be a bit discrepancy in the time scale. In fact, this is not surprising and one need to understand the variability in the time estimates in order to avoid over-interpreting these estimates.

It is important to note that the real-world time estimation is based on the mapping between the chronological clock to the molecular clock. This transformation relies on an important assumption that CpG>TpG mutations are a readout of a clock-like mutational process. Although a correlation between these mutations and age at diagnosis was confirmed in our cohort, the strength of the correlation is rather weak. This can be on one hand due to other active mutational processes also generating CpG>TpG mutations, and on the other due to a heterogeneous rate at which mutations were accumulated over time. Following the approach of Gerstung *et al.*, the CpG>TpG mutation accumulation was modeled here by two phases of linear increase, where in the later phase an accelerated rate was adopted. The choice of parameters used in the model affects the ultimate real-world time estimate, and therefore result in the discrepancies. In the case of the work from Labidi-Galy *et al.*, the authors took a different approach and modeled the mutation accumulation with a constant mutation rate.

Moreover, in the chronological clock, an accurate age at sample acquisition is also important for a correct inference. Nonetheless, all interval debulking samples were assumed acquired at the time of diagnosis, which obviously would yield slightly biased estimates. In the end, complete clinical information and a concordant modeling process using reasonable parameters would give comparable and meaningful result of time estimates. In any case, these previous work and our work provide a glimpse of tumor evolution in real-world time, which offer important information that can be linked to observations in the clinic and also provide implications for early detection and therapeutic planning of the disease.

14.4.3 *Nominating driving events in each patient*

The importance of germline and somatic alterations is usually assessed based on their predicted functional consequences or documented role in the literature. In fact, many missense variants, albeit VUS, can still be functionally important. This is exemplified by the TP53 variants detected in HIPO59, where we found 28 (77.8%) of 36 variants predicted non-deleterious. For tumor suppressor genes, biallelic inactivation provides further evidence on their essentiality in tumorigenesis. Take the re-analysis of TCGA-OV cohort as an example, researchers defined LOH by increased variant allele frequency, and found that germline truncation variants accompanied by LOH occurred in 100% of BRCA1 cases and 76% in BRCA2 cases[205].

In this thesis, biallelic inactivation status was more sophisticatedly determined by integrating germline and somatic small variants, as well as copy number status. In HIPO59, biallelic inactivation in known drivers was confirmed in 34 (94.4%) out of 36 patients with TP53 somatic variants, 6 (75%) out of 8 patients with BRCA1 variants, and 2 (66.7%) out of 3 patients with BRCA2 variants.

In non-HGSC cases, biallelic inactivation information led to the identification of two promising events not previously associated with ovarian cancer. In H059-TM8F, two somatic hits in PTEN was observed. The two variants probably targeted different alleles, given that the missense one is expressed in RNAseq and the other variant is a stop-gain variant. If they would have been on the same allele, this transcript would likely would go nonsense-mediated decay and not being detected or expressed. Interestingly, the ClinVar database documented that the stopgain variant (rs121909219) is found in several *Cowden syndrome* patients as germline variants, and also as somatic variants in other solid tumors or neoplasms of the large intestine, lung, breast, ovary and brain. This again emphasized the importance of paired tumor-normal sequencing, as the same germline pathogenic variant can be sometimes acquired in the somatic setting. The second patient, H059-N4GQ, was targeted multiple times in the BLM gene, with a first hit with germline pathogenic variant followed by a second hit with somatic deleterious variant. They are likely to be on different alleles according to the incompatible variant allele frequencies observed in tumor (>0.65 for the somatic variant, and <0.1 for the germline variant). It is possible that BLM underwent second somatic hits in early tumor development, while later on the allele with somatic hit took over by an copy number LOH event.

Despite the success in some cases, biallelic inactivation status can be difficult to determine in some other cases with low tumor purity or with inaccurate inference of local copy number. Two of three cases (H059-9BFZJ8, H059-UGNMF3) with TP53 or BRCA1 variants not accompanied by LOH had relatively low purity (<=15%). Moreover, somatic alterations leading to gene inactivation can be of any forms, including but not restricted to copy number deletion and small variants. This is exemplified by reported germline large deletions in BRCA1 in 3 TCGA-OV cases[205], as well as reported epigenetic silencing of BRCA1, FANCF and RAD51C[130, 216, 217]. To address this question, we note below that two additional facts, including an observable footprint and

the timing when the variant was acquired, are both important for helping us further confirm or rule out potential driver events in DDR genes.

In terms of footprints, its utility is first illustrated by using SBS Signature 3, class 1 TD activity to interrogate functional loss of BRCA genes. This helped confirm the driver role of BRCA1 in three cases showing high SBS Signature 3 and active class 1 TDs, including two HGSCs (H059-9BFZJ8 and H059-M3SDDT) with deleterious variants whereas lacking evidence for biallelic inactivation, as well as one non-HGSC case (H059-ASG5U9) with germline pathogenic variant accompanied with LOH. On the other hand, lower SBS Signature 3, TDP negativity or lack of major peak in class 1 can help preclude the driver role of BRCA1 in one non-HGSC case (H059-D096) with germline benign variant accompanied by LOH, as well as two HGSC cases (H059-DGCF, H059-3DCX) with potential gene breakage events. Similarly, a lower SBS Signature 3 helped exclude the driver role of BRCA2 in one HGSC case (H059-4PVFGF) with germline unknown variant and retained the wild type allele. Notably, footprints are imprinted by both historical and on-going processes, it does not discriminate an early or late onset unless dissected as done in the thesis.

In terms of variant acquisition timing, one can first confirm its validity by checking the timing of somatic variants in known driver genes. Not surprisingly, most of them were found in the earliest tumor epoch (see Figure 8.5) in each corresponding patient. These include two BRCA2 somatic variants found in "Clonal" and "Clonal (before WGD)" epochs, one BRCA1 somatic variant found in "Clonal (before WGD)" epoch, and 32 (88.9%) out of 36 TP53 somatic variants identified in either "Clonal", "Clonal (truncal)", "Clonal (before WGD)" or "Truncal (before WGD)" epochs. In the four exceptions (H059-9BFZJ8, H059-FD17WF, H059-UGNMF3, H059-Y22QC3), TP53 variants were identified in "Clonal (after WGD)" epoch. This largely goes well with the pathological observations that p53 signature is the earliest precursor lesion for HGSC. It is also important to note that, a variant seemingly being a branch event can be possibly a truncal event. This can be on one hand due to the low purity thereby low detection power in some of the samples, and on the other due to a lost of the variant during tumor development as a result of later copy number deletion events in the branch.

With this in mind, when we go back to the two previously mentioned patients with biallelic inactivation in PTEN and BLM, their somatic variants were found all in the earliest tumor epoch (see Figure 8.5), where BLM was found in "Truncal (before WGD)" and both PTEN variants were found in "Clonal (truncal)". More genes with supported early driving role can be found in Figure 13.10 and Figure 13.11. For example, the only case with somatic variant in BRIP1 also had the variant found in "Clonal (before WGD)" epoch. Among the three patients harboring NF1 somatic variants, two HGSC cases were restricted to H-HRD and biallelically inactivated, and the variants were found in the earliest tumor epochs ("Clonal" and "Clonal (truncal)"). The third one occurred in a non-HGSC case and found as a branch event in "Clonal (branch)" epoch.

The last case discussed here perfectly illustrates the importance of looking at both footprints and variant timing, as well as the benefit brought by multi-sample design. Among the three hypermutators, H059-F9BQHA harbors somatic pathogenic frameshift insertion in POLH (p.C16fs). Despite wild type lost was observed in only one of the

three related samples, all samples showed footprint in the corresponding SBS Signature 9 (Polymerase eta signature). Both evidences confirmed that genome instability induced by compromised POLH was probably the source of excess mutations found in this patient. Intriguingly, footprints reflecting MMR defect were also observed in these samples, including AC26, ID1 and ID2 signatures. A somatic missense variant in PMS2 (p.L323R) was found to be likely accompanied by LOH (after manually correcting the copy number) and probably responsible for these signatures. When integrating the variant timing information, it was surprising to find that, although both somatic variants were found in the truncal part of the tree, the one in PMS2 might had occurred earlier and before WGD, while the one in POLH occurred later and after WGD. These suggest that genome instability induced by MMR defect was not only the source of excess burden of indels in this patient, but probably also an earlier driving force. However, whether the earlier MMR defect caused the later insertion in POLH is not known. Notably, although both variants were acquired in the somatic setting, both POLH and PMS2 are associated with cancer predisposition syndromes. Specifically, POLH has been associated with autosomal recessive disease *Xeroderma pigmentosum variant* (*XPV*)[218], and PMS2 has been associated with *Lynch Syndrome*[219], a hereditary syndrome associated with ovarian cancer[53].

It is therefore interesting to further digest the role of variants acquired early in other patients, and to incorporate more footprints reflective of corresponding defects. Moreover, the observation that monoallelic inactivation in POLH and BRCA1 were sufficient to generate footprints, if were true, further leads to the speculation that biallelic inactivation may not be required for some genes to promote tumorigenesis. In that case, the information of haplosufficiency for individual genes would be helpful for refining the process nominating drivers in individual patient.

14.4.4 *Filling in details in the current knowledge about HGSC tumor evolution*

In HGSC, DDR-related genes were identified as the main target for small variants either at germline or somatic setting[205]. Our data added the notion that the earliest events, the germline variants, were enriched in three DDR subpathways in ovarian cancer regardless of histotype. When further considering the HGSC dichotomy, early driving events in DDR were observed most frequently in H-HRD (80%), followed by non-HGSCs (66.7%) and least prevalent in H-FBI (46%). Together they show that DDR defect widely existed already before MRCA-PID, especially for the H-HRD subgroup.

Refined tumor epochs can arrange small variants along the tumor evolutionary trajectory with finer time resolution and can increase the number of timed small variants by 47.6%. Our data suggested that TP53 is a ubiquitous and early event, supporting the notion of earliest precursor being p53 signature. Nonetheless, additional driving force is required for them to progress to malignancies, as germline TP53 mutation carriers are not significantly predisposed to ovarian cancer risk.

Candidates in the earliest tumor epoch (before WGD) suggested that, in H-HRD subgroup, known drivers BRCA1 and BRCA2, as well as some DDR genes, may serve as the additional driving force due to their very early emergence. A temporal trend

of deleterious variant abundance highlighted this phenomena in H-HRD by showing that DDR genes were most often disrupted in the earliest epoch compared to later timepoints (see Figure 13.13). These events are followed by WGD, which is itself another early event in ovarian cancer and associated with higher IPH and higher level of CIN, especially for non-HGSCs (see Figure 13.2). In the temporal trend of H-HRD, TSGs also seemed to be early targets, where variants most abundantly shown in the second epoch (shared clonal) compared to other epochs. Later on, more pathways are targeted via copy number changes which facilitate the selection process favoring cells achieving hallmarks of cancer.

It is therefore reasonable to hypothesize that in H-HRD subgroup, DNA repair defect and TP53 dysfunction are acquired mainly by germline or somatic small variants and play roles in the initial stage of tumorigenesis. The succeeding WGD is also an early event that can occur decades before diagnosis. TSG disruption occurrs afterwards but before `MRCA-PID`, whose emergence spans a variable range of time and depends on each individual. Also subsequent to WGD is CIN, which facilitates diversification of tumors and fuels tumor evolution by continuously providing substrates for selection process. On the other hand, what we know about H-FBI is comparably rather limited. Nonetheless they also share the scenario of very early TP53 mutation and WGD, and a later widespread CIN.

To note, it is not yet known whether DDR defect has to precede TP53 mutation, and further pathological studies on precursor lesions will be more suitable for answering this question. It is also unknown whether CIN follows TSG disruption or they are independent events, a further anchoring of copy number events in the phylogeny tree can provide more hints in this aspect. In terms of H-FBI, looking into other early events outside DDR genes or cancer-associated genes, as well as including copy number events, may provide additional insights into their pathogenesis.

At the moment, there has been intense interest in including copy number events in the timing scheme; however it is not yet readily applicable due to some limitations. Although molecular time estimate for copy number gains are available, they showed very wide confidence intervals due to the shorter copy number segments generated by high level of CIN. Therefore, more sophisticated methods are required to properly incorporate CNAs into the evolutionary trajectory.

Up to this point, one can posit that genomic subgroups may inform early bifurcation of carcinogenesis pathways as the two subgroups only shared the initial TP53 ubiquity and the eventual CIN phenotype. It was also proposed copy number change can lead to expression subtypes specification[220], indicating their emergence as a late diversification event. Together this further contrasts the utility of genomic subgroup and expression subtypes, where the former may inform earlier driving events that are usually more effectively treatable when targeted.

15 CONCLUSIONS AND OUTLOOK

This thesis started with providing an overview of ovarian cancer at the molecular level. Chapter 9 used public data (TCGA-OV) to compare the disease with other cancer types, and later on combined them with the result of in-house data (HIPO59) for highlighting key disrupted targets in different pathways.

A question of the utmost interest in ovarian cancer treatment is to identify patients potentially having chemosensitivity or responding to PARP inhibitors. In this sense, Chapter 10 and Chapter 11 put a particular focus on evaluating prognostic biomarkers and assessing HRD phenotype in HIPO59 patients. Specifically, prognostic biomarkers including BRCA genes inactivation, CCNE1 amplification status, LOH score, HRD score and Shah-2017 genomic subgroup (H-HRD and H-FBI) were evaluated. Moreover, additional evidences for HRD were examined and cover TDP subgroup, defects in the DDR pathway, as well as its consequent genomic scars in terms of mutations, indels and structural variations.

Given current knowledge of two mutually exclusive groups existing in HGSC, I further hypothesized that this concept may be extended to become a HGSC dichotomy. The abovementioned evidences, when put together, support this view by showing their differences in the extent and onset timing of HRD as well as CCNE1 pathway activation. Specifically, HRD is a common feature acquired early in H-HRD cases; while in H-FBI tumors, CCNE1 pathway activation is often an early event. More importantly, the fact that they showed differences in surrogate biomarkers for PARP inhibitor response is suggestive of its clinical relevance.

However, it is of note that a minor subset of cases might stand outside this dichotomy. For example, tumors featured with MMR, APOBEC or CDK12 inactivation can constitute this rare subset, which is not well-represented in HIPO59 due to its small sample size (n=33). In addition, the view of dichotomy does not preclude a finer substructure given the heterogeneity of the disease.

The multi-sample design of HIPO59 provides an opportunity to study tumor heterogeneity and tumor evolution in silico. By proposing a general method for summarizing IPH, Chapter 12 demonstrated a variable degree of IPH in the pre-treatment OC samples. This variability was especially obvious for HGSCs. More intriguingly, a higher degree of IPH in a HGSC patient suggested a trend toward better prognosis. As the IPH measure proposed here can be concordantly summarized from either small variants, CNAs or SVs, it is possible to design an optimal protocol to study this phenomena in a larger cohort.

Combining phylogeny analysis and variant timing technique, Chapter 13 arranged small variants into four tumor epochs along the tumor evolutionary trajectory, and attached them to the reconstructed sample phylogeny tree. For individual patient, this informs about potential early and late driving events, the emergence of major events, as

well as the evolutionary relationship of their related samples. Collectively, they provide a glimpse of ovarian cancer carcinogenesis in real-world time.

Although HGSCs are known to have ubiquitously early TP53 mutation and CIN phenotype at diagnosis, temporally dissected events further suggested an early bifurcation of carcinogenesis pathways in the HGSC dichotomy. Specifically, some H-HRD cases were observed to have acquired TP53 dysfunction and DDR defects from either germline or before they acquired WGD, which is with a median latency of 21.7 years. Subsequent defects in TSG accrued before the emergence of MRCA between related samples (MRCA-PID), whose latency spans a wide range from 0.4 year to 30 years, and eventually present a common CIN phenotype at diagnosis. On the other hand, H-FBI also acquired TP53 aberrations in the first epoch, they all developed WGD (median latency 33.5 years) and eventually present CIN at the time of diagnosis. Nonetheless, little evidences about an early DDR defect was observed. The same was suggested in the temporal dissection of DNA footprints, where an early and stronger onset of DDR defect was observed in H-HRD, while a later and weaker onset was observed in H-FBI. Together with the abovementioned trend showing an early and truncal CCNE1 activation in H-FBI but not in H-HRD, they suggest that the common CIN phenotype might not be achieved by the same mechanism.

THERAPEUTIC IMPLICATIONS

The findings in the thesis are, albeit preliminary, important from some therapeutic aspects.

First of all, the carcinogenesis pathway and footprints revealed in H-HRD strongly suggest a ubiquitous HRD in this group, regardless of their BRCA gene status. In this sense, the HGSC dichotomy can be used for molecularly stratifying patients in clinical trials, where their potential of informing PARP inhibitor treatment response can be better determined.

Secondly, given the link between hereditary defects in DDR pathway, it is worthwhile considering expanding the genetic testing to a broader panel covering more hereditary DNA repair-deficiency syndromes. This information can benefit their family members and reduce cancer incidence in their families.

Thirdly, tumors related to hereditary or acquired DDR defect can show vulnerabilities pertinent to the corresponding cancer predisposition syndrome. For example, *FA* patients are very sensitive to ICL-inducing agents, which provides the rationale for the use of platinum-based chemotherapies. On the other hand, *Lynch Syndrome* patients are with MMR deficiency and may benefit from immunotherapy.

Fourthly, the prognostic implication of IPH status, if holds true, can serve as a measure for interrogating clonal expansion potential of the observed tumors. With the summarizing measure proposed in the thesis, it is possible to develop a standardized protocol for evaluation in a larger cohort.

Lastly, the early and late events identified in each patient can facilitate the practice of personalized oncology. For example, germline variants can trigger cascade testing,

early driving events are informative of therapeutic choices, and truncal variants are good targets for developing disease monitoring assays during treatment course.

FUTURE DIRECTIONS

The preliminary evidences shown here also provide important insights for future researches.

First of all, the concept of HGSC dichotomy is compatible with contemporary findings about HGSC and provides a novel viewpoint for re-interpreting them. This prompts the need for future research designs to take this factor into considerations. It can also arouse more interest in revealing more subtype-specific features, and eventually lead to better understanding about subtype-specific pathogenesis and vulnerabilities.

Secondly, developing clinical grade assays would be the next step to evaluate the relevance and utility of IPH status and genomic subgroups in clinical trials. A standardized assay can also facilitate comparisons between future studies and enable an integration into routine practice.

Last but not the least, it is also important to better characterize the minor subset outside the HGSC dichotomy, and a finer substructure under this dichotomy is also worthwhile further dissection.

Overall, by maximizing the knowledge learned from HIPO59 patients, I hope it would help to bring more hopes to other patients, improve the well-being of them and their families, and meanwhile advance the understanding about ovarian cancer.

Part V

APPENDIX

A

SOFTWARES, CODES AND SUPPLEMENTAL FILES

SOFTWARES AND WORKFLOWS

The majority of the analyses performed were performed in the R Environment (see Appendix A). Analyses using other softwares are listed below:

- DKFZ workflow - AlignmentAndQC Workflow version 1.2.73-1, SNVCalling Workflow version 1.2.166-1, ClinicalWorkflow version 1.1, IndelCalling Workflow version 1.2.177, SophiaWorkflow version 1.2.16, ACEseqWorkflow version 1.2.8-4
- GISTIC version 2.0
- MutSigCV version 1.41
- PHYLIP (Phylogeny Inference Package) software version 3.6

R ENVIRONMENT AND PACKAGES

- R version 3.5.3 (2019-03-11), `x86_64-pc-linux-gnu`
- Running under: `Ubuntu 16.04.6 LTS`
- R packages: bigmemory 4.5.33, Biobase 2.42.0, BiocGenerics 0.28.0, BiocParallel 1.16.0, Biostrings 2.50.0, circlize 0.4.5, ComplexHeatmap 2.5.3, data.table 1.13.6, DelayedArray 0.8.0, dplyr 1.0.0, factoextra 1.0.5, forcats 0.3.0, futile.logger 1.4.3, gdata 2.18.0, GenomeInfoDb 1.18.0, GenomicRanges 1.34.0, ggplot2 3.3.2, ggpubr 0.4.0.999, gt 0.2.1, hrbrthemes 0.8.0, igraph 1.2.4, IRanges 2.16.0, maftools 1.8.10, matrixStats 0.54.0, pheatmap 1.0.12, png 0.1-7, purrr 0.3.2, RColorBrewer 1.1-2, readr 1.3.1, ROCit 2.1.1, Rsamtools 1.34.0, rstatix 0.6.0, S4Vectors 0.20.1, stringr 1.4.0, SummarizedExperiment 1.12.0, tibble 3.0.3, tidyr 1.1.2, tidyverse 1.2.1, VariantAnnotation 1.28.1, VennDiagram 1.6.20, XVector 0.22.0, YAPSA 1.8.0

SUPPLEMENTAL FILES

Results pertinent to the analyses can be found in the Compact disc along with the thesis and are listed below:

- `Geneticist review result.xlsx`: geneticist review result
- `HIP059_Subtype_TDP.txt`: TDP assignment for each HGSC sample, see Section 10.1.2

- `HIPO59_Tumor_Evolution.txt`: , see Section 13.3
- `Tumor_Heterogeneity_PID.txt`: heterogeneity scores (`HET_MutIndel`, `HET_SV`, `HET_CNA`) and heterogeneity group (High, Median, Low) of each patient, see Chapter 12
- `Tumor_Heterogeneity_SAMPLE.txt`: heterogeneity scores (`HET_MutIndel`, `HET_SV`, `HET_CNA`) and heterogeneity group (High, Median, Low) of each sample, see Chapter 12
- Related code used for the analyses are included in the Code directory and are described here:
  - `PCAWG-final.R`: the variant timing in this thesis follows the procedures in this code from Gerstung's tutorial, see Section 8.10
  - `Script_Methods.r`: key codes for reproducing main Specific Tasks in Chapter 8.

B

SUPPLEMENTARY DATA

B.1 ENUMERATE RECURRENT GISTIC PEAKS

Cohort	Number of Recurrent Peaks	Contribution of Top-5 Recurrent Peaks (%)
BLCA	37	8q22.3amp (24), 1q23.3amp (19.4), 6p22.3amp (16.9), 5p15.33amp (15.7), 8q24.21amp (15.2)
BRCA	28	8q24.21amp (31), 1q21.3amp (27), 1q44amp (26.5), 11q13.3amp (18.4), 8p11.23amp (17.9)
COAD	22	20q13.12amp (43.5), 20q12amp (41.5), 20q11.21amp (40.1), 13q12.13amp (26.2), 13q22.1amp (25.1)
GBM	24	7p11.2amp (54.2), 7q11.21amp (21.3), 7q21.2amp (19.9), 7q31.2amp (19.8), 12q14.1amp (15.1)
HNSC	27	11q13.3amp (26.6), 3q26.33amp (24.5), 8q24.21amp (17), 5p15.33amp (11.3), 8q11.21amp (10.9)
KIRC	8	5q35.1amp (5.1), 7q31.2amp (3.2), 3q26.32amp (1.3), 1q32.1amp (1.1), 8q24.22amp (0.9)
LAML	4	11q23.3amp (5.8), 21q22.2amp (4.2), 1p33amp (1), 20q11.21amp (0.5)
LUAD	28	5p15.33amp (21.9), 8q24.21amp (17.2), 1q21.3amp (16.3), 5p13.1amp (16.3), 14q13.3amp (14.9)
LUSC	29	3q26.33amp (53.7), 5p15.33amp (27.5), 5p12amp (24), 8p11.23amp (20.4), 8q24.21amp (16.2)
OV	31	8q24.21amp (49.9), 3q26.2amp (44.6), 19q12amp (26.9), 20q13.33amp (24.5), 12p12.1amp (22.1)
READ	20	20q11.23amp (54.5), 20q11.21amp (53.9), 20q13.33amp (53.9), 13q12.13amp (38.2), 13q12.2amp (36.4)
UCEC	48	1q22amp (21.9), 1q21.3amp (20.6), 8q24.21amp (16.9), 1q42.3amp (16.9), 8q24.21amp (16.1)

Table B.1: Top five recurrent focal SCNAs identified by GISTIC in TCGA Pan-Cancer cohort.

B.2 REPRESENTATIVE SAMPLES

H059-0EJ9 (tumor6), H059-1LUEUK (tumor0), H059-28C2CC (tumor00), H059-3DCX (tumor72), H059-3ZK0 (tumor5), H059-41N6F7 (tumor5), H059-4PVFGF (tumor71), H059-5DFS (tumor1), H059-6GM3 (tumor6), H059-8Y1SE7 (tumor7), H059-9BFZJ8 (tumor5), H059-CFDW82 (tumor6), H059-DGCF (tumor7), H059-DQNU (tumor07), H059-E3Z5MP (tumor7), H059-ESPXYL (tumor8), H059-F9BQHA (tumor4), H059-FD17WF (tumor2), H059-H9Q5W6 (tumor5), H059-M3SDDT (tumor8), H059-MRAP5C (tumor4), H059-N8J8 (tumor32), H059-NV4KXQ (tumor7), H059-P8MX2J (metastasis4), H059-QQBCPM (tumor8), H059-QTH8 (tumor-interval-debulking-surgery5), H059-RWF1GY (tumor7), H059-U4U5X5 (tumor7), H059-U6DA (tumor6), H059-UGNMF3 (tumor7), H059-Y22QC3 (tumor7), H059-YKP3 (tumor71), H059-YYNAEG (tumor0)

B.3 HIPO59 MUTSIGCV RESULT

Gene	N_nonsilent, N_silent, N_noncoding	n_nonsilent, n_silent, n_noncoding	p	q
TP53	101937,28149,0	31,0,0	2.22e-15	4.19e-11
PABPC3	145035,42867,0	2,0,0	3.20e-03	1
ELSPBP1	56694,12210,0	2,0,0	3.24e-03	1
KIAA0391	139029,36795,0	4,0,0	1.10e-02	1
TMEM132C	254232,78309,0	2,0,0	1.17e-02	1
STH	29799,8910,0	2,0,0	1.49e-02	1
DACH1	164505,50028,0	2,0,0	1.51e-02	1
DLG3	218229,60159,0	3,0,0	1.54e-02	1
WSCD2	133518,37356,0	2,0,0	1.59e-02	1

Table B.2: Top 10 recurrently mutated genes identified by MutSigCV in HIPO59 cohort.

B.4 GENE STRUCTURE OF REPORTED SIGNIFICANTLY MUTATED GENES

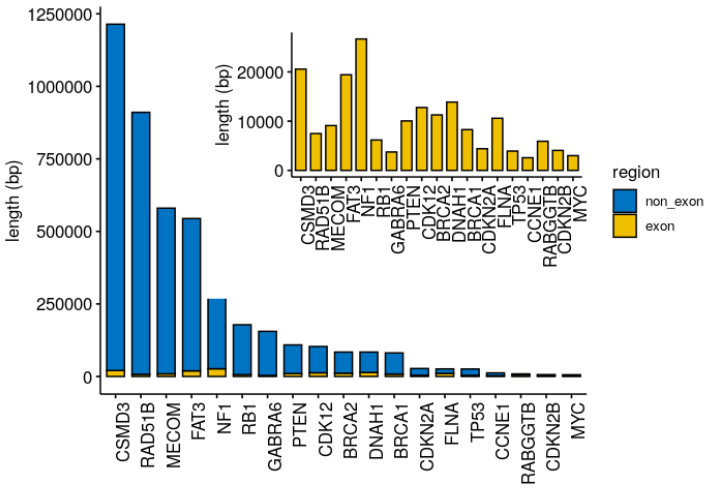


Figure B.1: Gene size of reported significantly mutated genes (SMGs).

B.5 GERMLINE PATHOGENIC VARIANTS ARE FOUND IN NOT ONLY BRCA GENES

Small variants were sent for pathogenicity review and the result from geneticist can be found in the Compact disc along with the thesis (**Geneticist review result.xlsx**). The file contains 49 germline variants and 25 somatic variants selected according to (see Section 8.6.1). Note that a somatic variant in PMS1 was not evaluated due to historical workflow version change.

B.6 RARE GERMLINE VARIANTS ARE ENRICHED IN DDR PATHWAYS

Gene Set	Set Size	p-value (HIPO59)	p-value (HGSC)
DDR	276	0.00022	0.000163
Homology-dependent recombination	88	0.00223	0.00123
Others	65	0.0162	0.0715
Fanconi Anemia	41	0.0251	0.0205
Non-homologous End Joining	23	0.0383	0.0287
Base Excision Repair	47	0.157	0.125
Translesion Synthesis	20	0.184	0.0782
Nucleotide pools	5	0.440	0.741
Mismatch Repair	24	0.548	0.335
Nucleotide Excision Repair	51	0.617	0.434

Table B.3: DDR pathway enrichment analysis. One-sided Fisher's exact test was performed on rare germline variants from the entire cohort (**HIPO59**) or from HGSC patients (**HGSC**).

B.7 GERMLINE AND SOMATIC LANDSCAPE OF DDR PATHWAYS

Patient	Subgroup	Germline pathogenic	Somatic truncal deleterious
H059-0EJ9	H-FBI		TP53(p.V157fs)
H059-3DCX	H-FBI		TP53(p.V173fs)
H059-3ZK0	H-FBI		
H059-4PVFGF	H-FBI	ERCC2(p.Q638X)	
H059-5DFS	H-FBI		
H059-CFDW82	H-FBI		
H059-DGCF	H-FBI		
H059-FD17WF	H-FBI		DDB1(p.E447X)
H059-MRAP5C	H-FBI	NBN(p.P381fs)	
H059-RWF1GY	H-FBI		
H059-U4U5X5	H-FBI		TP53(p.M66fs)
H059-UGNMF3	H-FBI		
H059-Y22QC3	H-FBI		
H059-1LUEUK	H-HRD		
H059-28C2CC	H-HRD		TP53(p.L289fs)
H059-41N6F7	H-HRD		TP53(p.P27fs)
H059-6GM3	H-HRD	ATM(p.E1978X), BRCA1(p.Q1756fs)	
H059-8Y1SE7	H-HRD		
H059-9BFZJ8	H-HRD	BRCA1(p.E1161fs)	
H059-DQNU	H-HRD		
H059-E3Z5MP	H-HRD		BRCA2(p.E2220X)
H059-ESPXYL	H-HRD		DDB1(p.M276fs)
H059-F9BQHA	H-HRD		POLH(p.C16fs)
H059-H9Q5W6	H-HRD		BRCA2(p.P655fs), SPO11(wholegene)
H059-M3SDDT	H-HRD	FANCG(p.E105X)	BRCA1(p.H1284fs)
H059-N8J8	H-HRD	BRCA1(p.D1162fs)	
H059-NV4KXQ	H-HRD	RAD51C(p.344_344del)	
H059-P8MX2J	H-HRD		ATM(p.E96fs)
H059-QQBCPM	H-HRD		TP53(p.Q104X)
H059-QTH8	H-HRD	BRCA1(p.Q1756fs)	
H059-U6DA	H-HRD	BRCA1(p.Q1756fs)	
H059-YKP3	H-HRD		
H059-YYNAEG	H-HRD	BRCA1(p.Q1756fs)	
H059-ASG5U9	OTHER	BRCA1(p.E1257fs)	
H059-D096	OTHER		TP53(p.E271X)
H059-DSGWDE	OTHER	FANCC(splicing)	
H059-LABYUN	OTHER		
H059-N4GQ	OTHER	BLM(p.Y974fs)	BLM(p.V286fs), TP53(p.P89fs)
H059-PH6WVA	OTHER		
H059-RHVSYD	OTHER		
H059-TM8F	OTHER		PTEN(p.R233X), PTEN(p.S170I)
H059-VDKDQX	OTHER	ATM(p.R2486X)	

Table B.4: Potential driving events in DDR pathways. DDR pathway gens with germline pathogenic or truncal somatic deleterious events were considered as potential driving events.

B.8 VALIDATE THE ADJUSTED SHAH-2017 METHODOLOGY

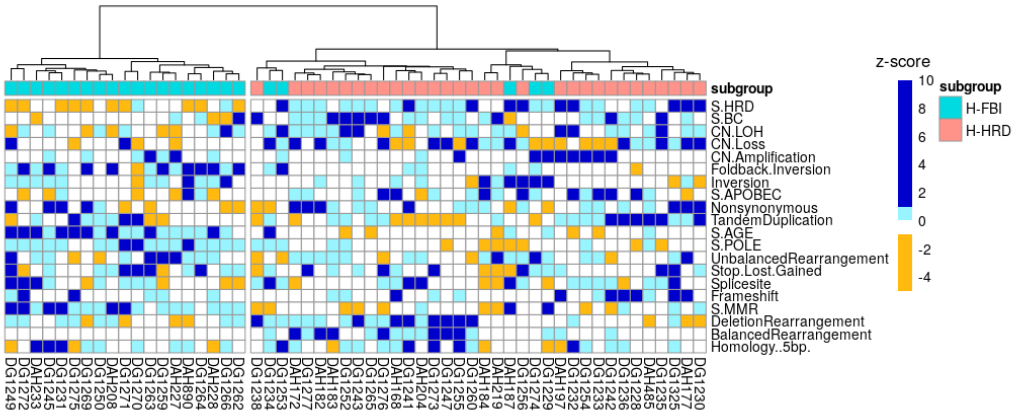


Figure B.2: Stratify the HGSC subset of OV133 cohort based on 20 genomic features.

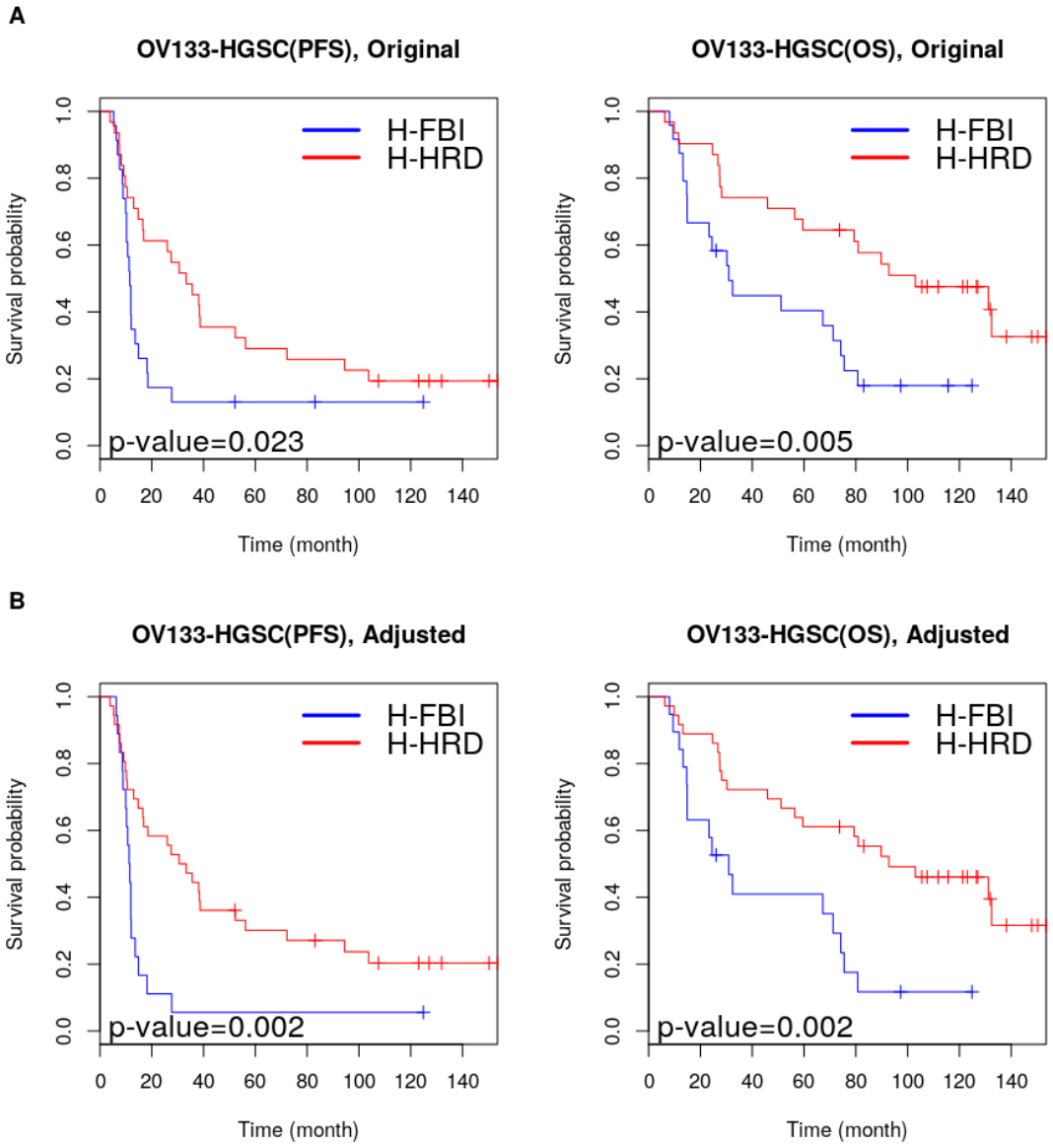


Figure B.3: Prognostic value of genomic subgroups based on 20 genomic features. Subgroups derived from original method are compared in (A). The original method was adjusted to using only HGSC samples for cluster discovery. New subgroups results from the adjusted method are compared in (B).

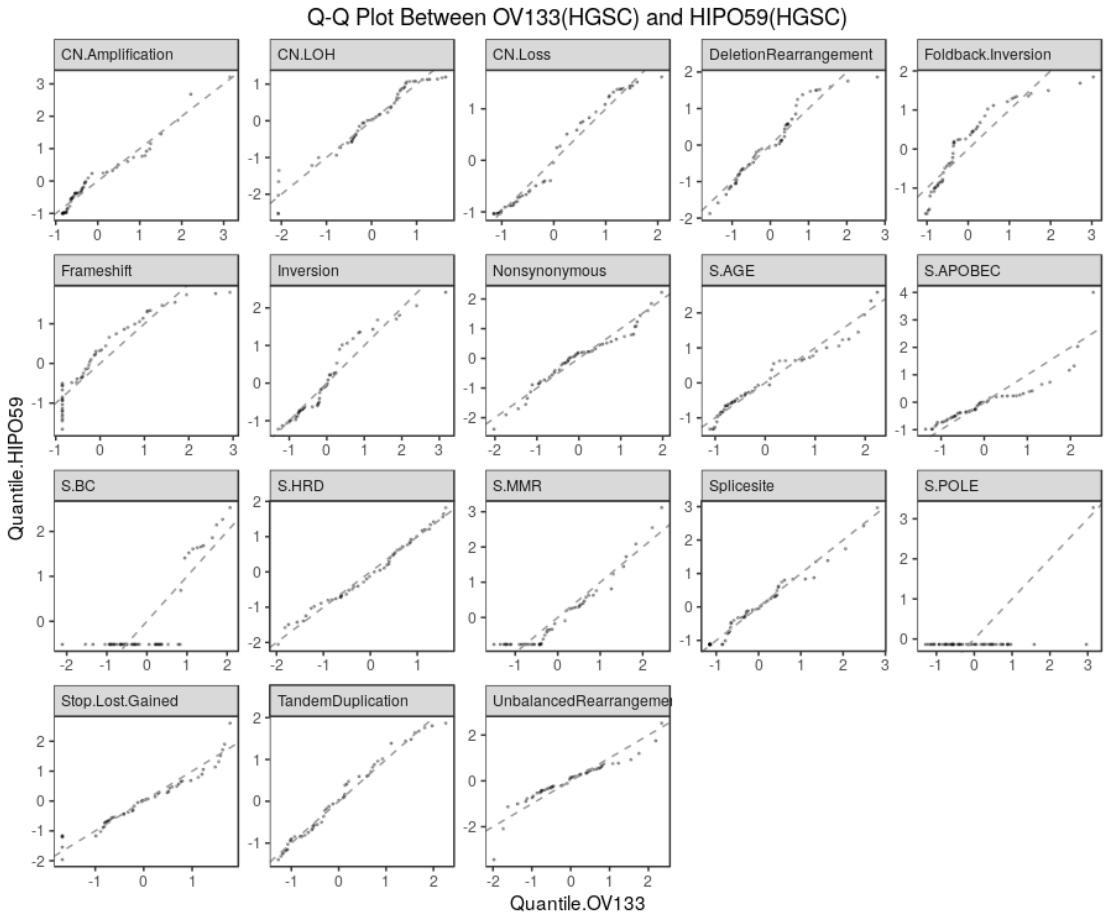


Figure B.4: Quantile-quantile plot for the distribution of each of the 18 genomic features in OV133 and HIPO59 cohort.

B.9 SIGNATURE ANALYSIS

Signature	Patient	Average Fraction	Signature Description
AC3	41	47.4	defect DNA DSB homologous recombination repair
AC8	40	19.3	unknown
AC1	30	24	spontaneous deamination
AC5	19	23	unknown
AC16	25	14.7	unknown
AC12	4	10.7	unknown
AC9	4	10.1	POL eta and SHM
AC13	3	11.8	APOBEC
AC26	1	21	defect DNA MMR
AC2	2	7.47	APOBEC
AC18	1	10	unknown
AC19	1	9.45	unknown
ID6	35	39.98	DSB repair by NHEJ; defective HRR
ID12	35	30.6	unknown
ID1	40	14.12	Replication slippage, sometimes defective DNA MMR
ID8	31	15.7	DSB repair by NHEJ
ID9	21	15.74	unknown
ID2	42	7.43	Replication slippage, sometimes defective DNA MMR
ID5	15	19.32	unknown
ID3	6	11.71	Tobacco smoking
ID4	1	24.11	unknown
Br.RS2	42	33.8	
Br.RS5	41	32.2	
Br.RS3	18	29.1	
Br.RS1	27	18.6	
Br.RS4	15	15.8	
Br.RS6	12	13.8	

Table B.5: Active signatures in mutational, indel and rearrangement signature analysis. The third column summarizes average fraction of each signature among patients with active signatures (second column). In each aberration type, the signatures are ordered by their contribution to the entire cohort.

B.10 GENOMIC FOOTPRINTS OF DDR DEFECT

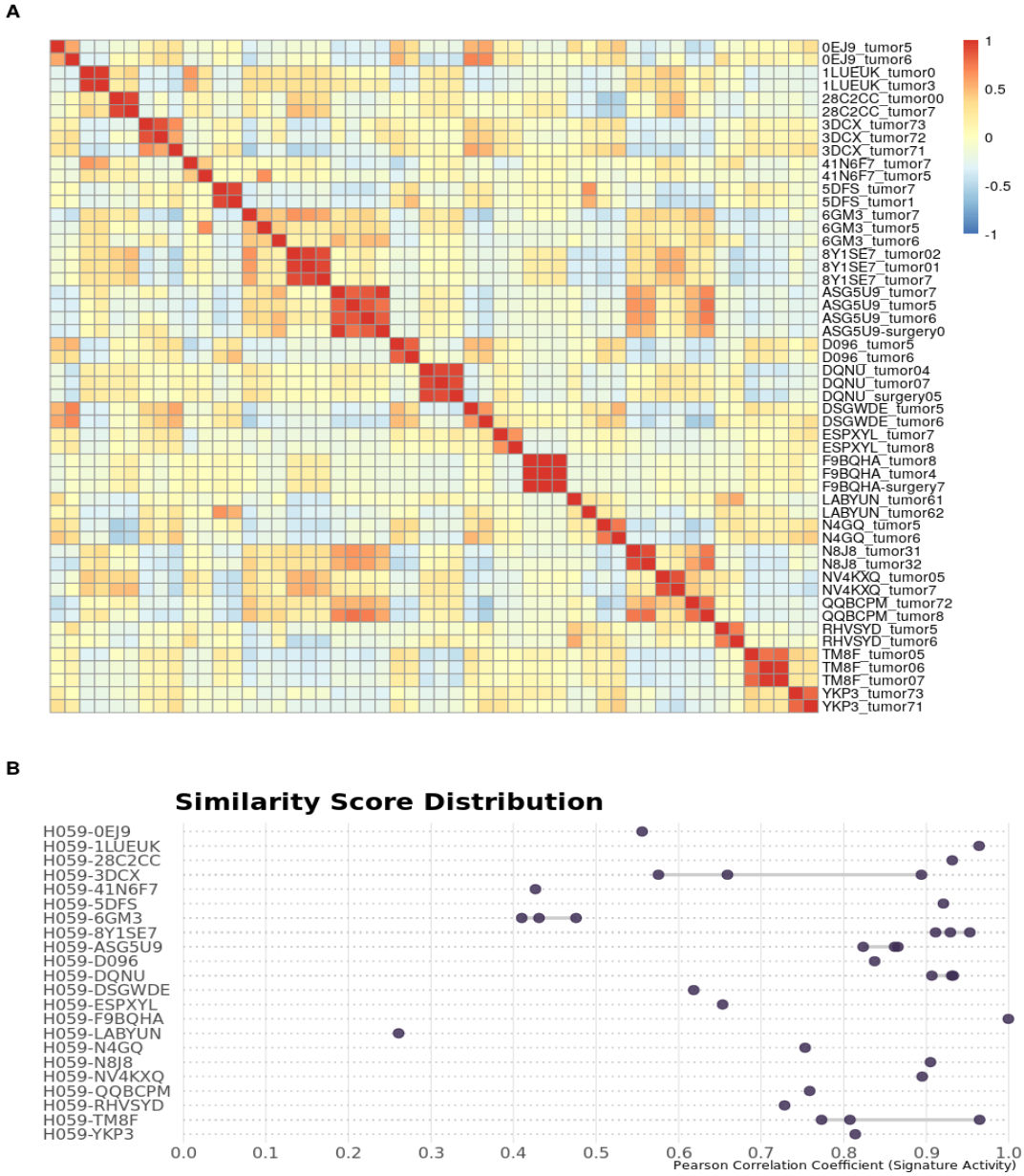


Figure B.5: Pair-wise sample comparison based on mutational signatures, indel signatures and rearrangement signatures. In (A) similarity between all sample pairs are shown, where samples arranged in columns and rows are in the same order. Each cell is one Pearson correlation coefficient of corresponding sample pair. (B) shows the degree of sample similarity in each patient, where a data point represent one score between a pair of related samples from this patient. All comparisons from the same patient are connected with a line to visualize the range of similarity score.

signature	statistic	df	p	p.adj
AC1	9.4	25.36	9.71E-10	4.37E-08
AC3	-7.16	49.34	3.59E-09	1.62E-07
AC4	-3.59	35	1.00E-03	4.50E-02
AC5	4.17	26.42	2.90E-04	1.31E-02
AC6	-5.32	47.21	2.84E-06	1.28E-04
AC25	-4.92	35.29	1.99E-05	8.95E-04
ID6	-6.26	45.45	1.22E-07	5.49E-06
ID8	-4.51	44.75	4.65E-05	2.09E-03
ID9	3.83	23.46	8.38E-04	3.77E-02
ID12	6.45	28.17	5.33E-07	2.40E-05
Br.RS3	-4.54	39.21	5.23E-05	2.35E-03
Br.RS5	-3.98	37.84	3.02E-04	1.36E-02

Table B.6: Testing difference in signature activities between HGSC genomic subgroups. Samples are divided into group 1 (17 H-FBI tumors) and group 2 (36 H-HRD tumors), and a Welch t-test was performed for each of all 54 signatures from three types of signature sets. The significance level (column p) was adjusted for bonferroni correction (column p.adj) and only those with p.adj < 0.05 are shown in the table.

score	statistic	df	p	p.adj
LOH	-3.69	29.00	9.31E-04	3.72E-03
LST	-5.69	25.29	6.11E-06	2.44E-05
TAI	-0.65	27.03	5.21E-01	1.00E2
total	-4.73	26.94	6.41E-05	2.56E-04

Table B.7: Testing difference in HRD score and its three component scores between HGSC genomic subgroups. For each patient, the median score from multiple related samples are used. Patients are divided into group 1 (13 H-FBI patients) and group 2 (20 H-HRD patients), and Welch t-test was performed between subgroups. The significance level (column p) was adjusted for bonferroni correction (column p.adj).

B.11 TUMOR HETEROGENEITY

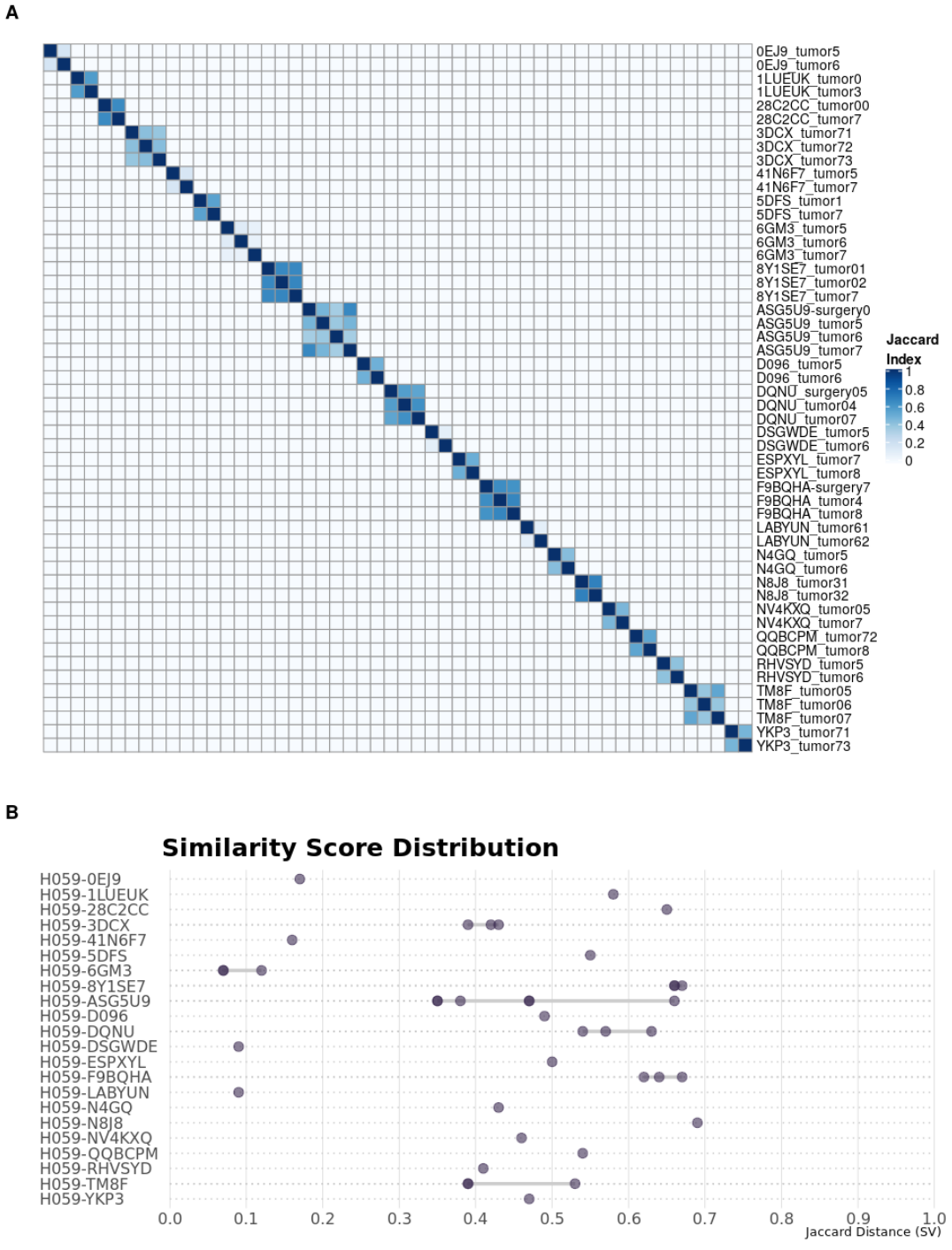


Figure B.6: Pair-wise sample comparisons based on structural variants.

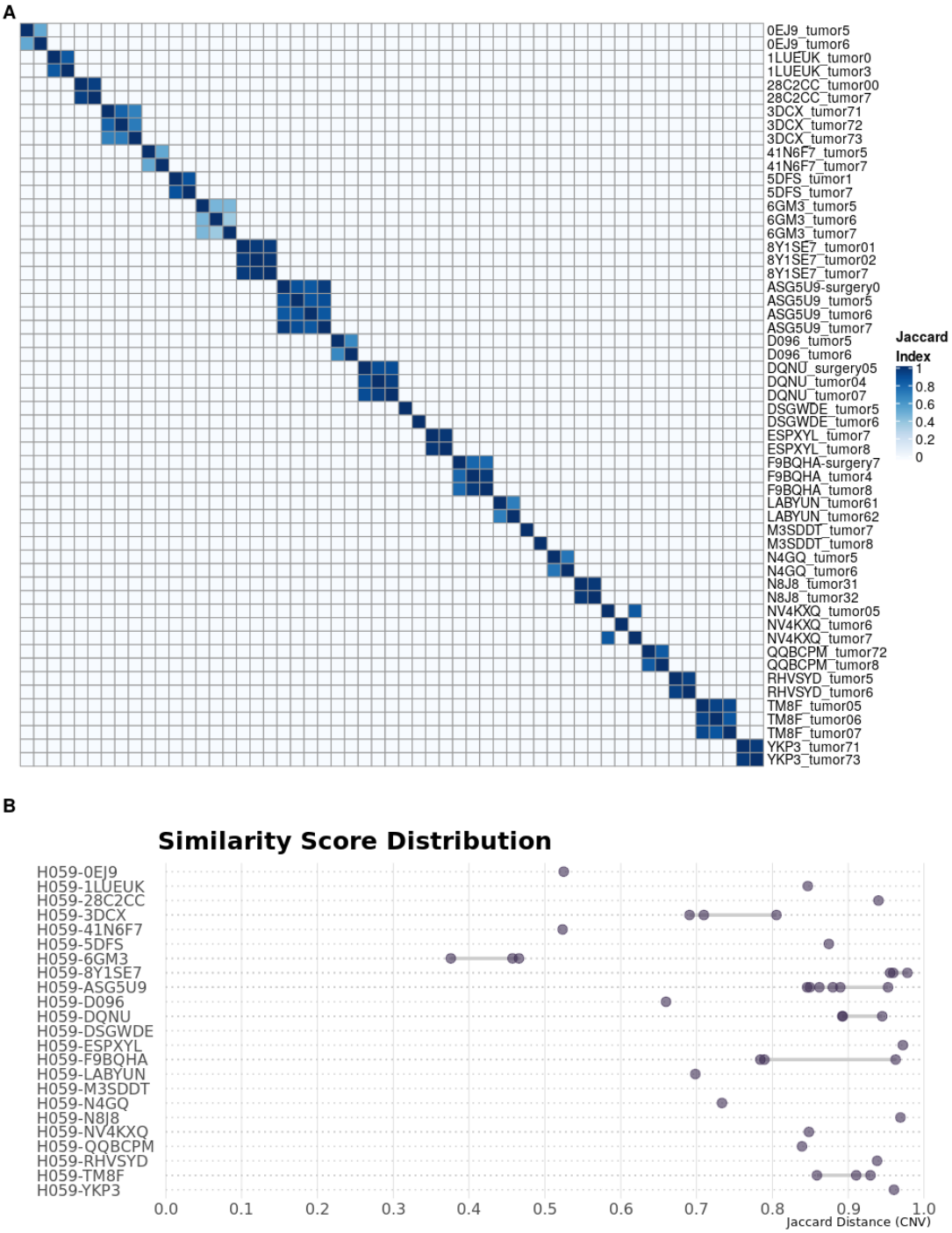


Figure B.7: Pair-wise sample comparisons based on copy number profiles.

B.12 TUMOR EVOLUTION

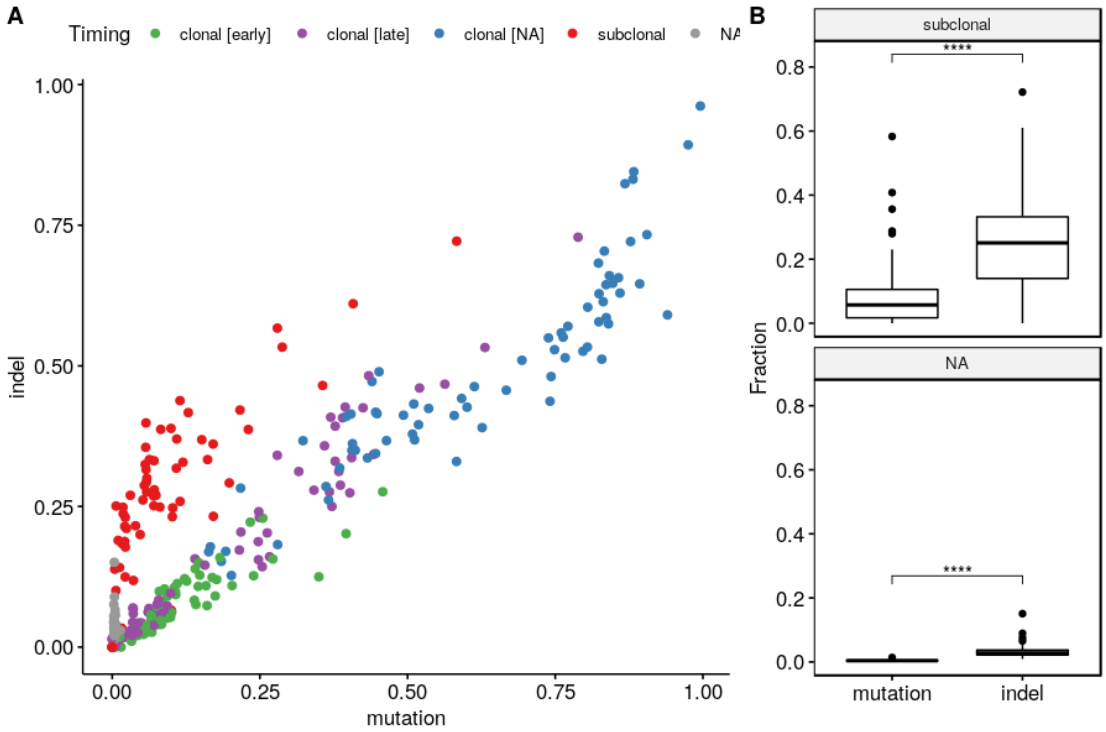


Figure B.8: Sample-wise timing class compositions in mutations and indels are profiled in (A). Each data point represent one fraction of a specific timing class observed in one sample and is colored according to timing classes. Welch t-test with bonferroni correction was used for identifying the difference within all timing classes. Significant p-value after adjustment was observed in clonal [early] (p.adj=3.55e-2), clonal [NA] (p.adj=1.45e-3), subclonal (p.adj=4.43e-10) and NA (p.adj=8.20e-17). The most significant composition differences between mutation and indel lie in two categories and are shown in (B).

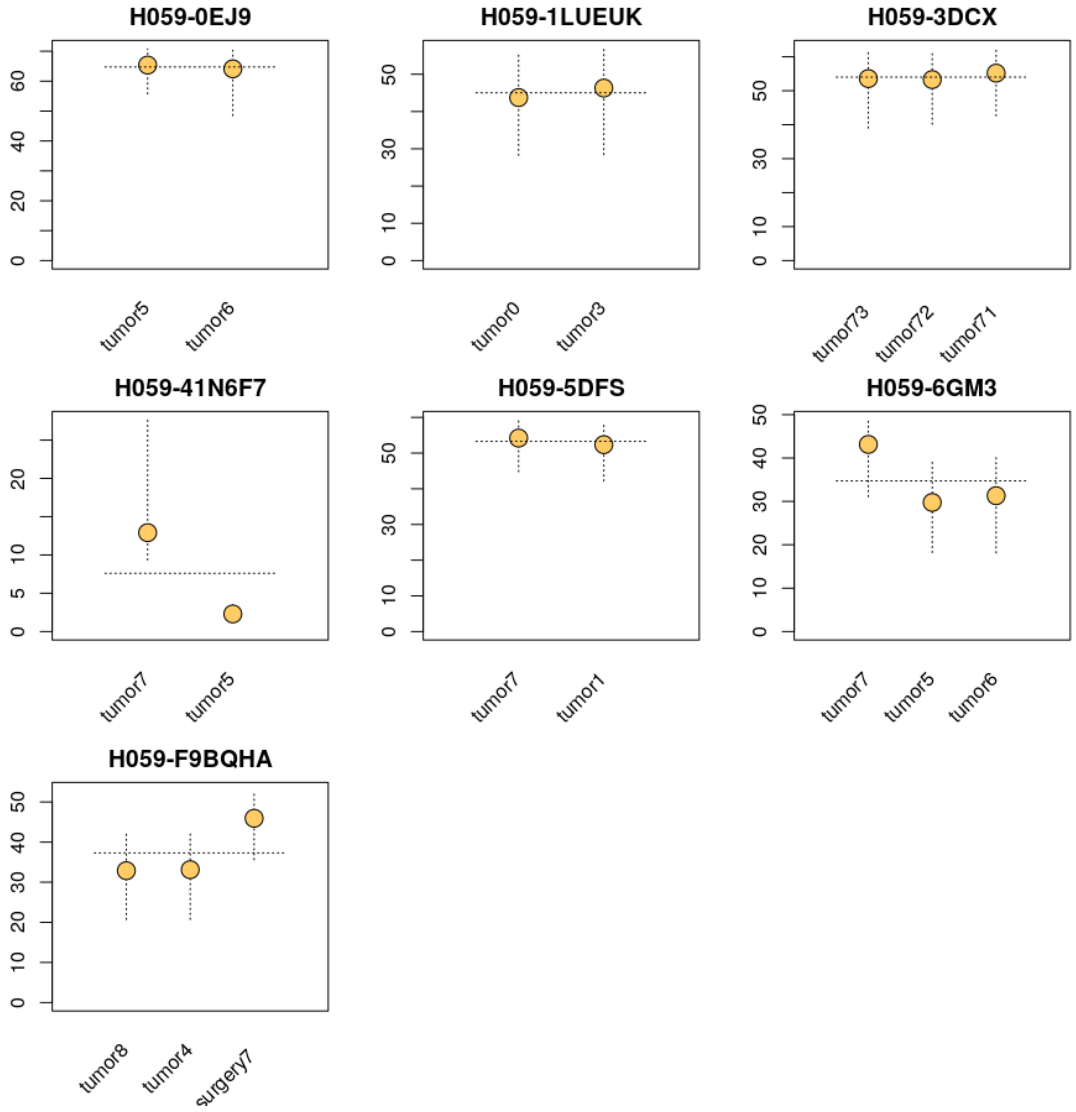


Figure B.9: Chronological time of WGD in multi-sample sets. Each estimate is shown with 80% CI.

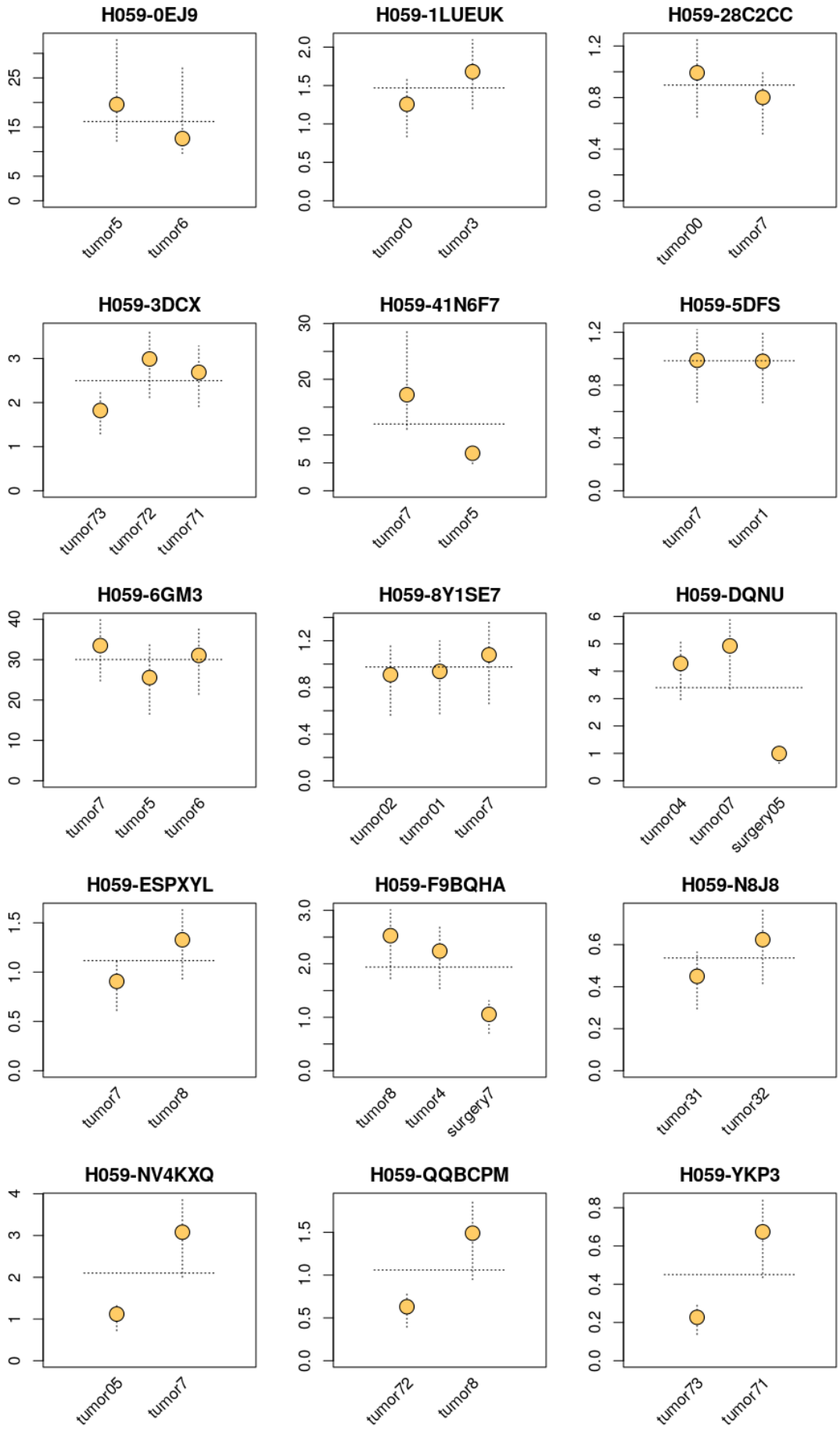


Figure B.10: Timing of MRCA-PID in multi-sample sets. Each estimate is shown with 80% CI.

C

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70. doi:10.1016/s0092-8674(00)81683-9 (2000) (cit. on p. 3).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–74. doi:10.1016/j.cell.2011.02.013 (2011) (cit. on p. 3).
3. Nagy, R., Sweet, K. & Eng, C. Highly penetrant hereditary cancer syndromes. *Oncogene* **23,** 6445–70. doi:10.1038/sj.onc.1207714 (2004) (cit. on p. 6).
4. Cleaver, J. E. Xeroderma pigmentosum: a human disease in which an initial stage of DNA repair is defective. *Proc Natl Acad Sci U S A* **63,** 428–35. doi:10.1073/pnas.63.2.428 (1969) (cit. on p. 6).
5. Lehmann, A. R., McGibbon, D. & Stefanini, M. Xeroderma pigmentosum. *Orphanet J Rare Dis* **6,** 70. doi:10.1186/1750-1172-6-70 (2011) (cit. on p. 6).
6. Knies, K. *et al.* Biallelic mutations in the ubiquitin ligase RFWD3 cause Fanconi anemia. *J Clin Invest* **127,** 3013–3027. doi:10.1172/JCI92069 (2017) (cit. on p. 6).
7. Nalepa, G. & Clapp, D. W. Fanconi anaemia and cancer: an intricate relationship. *Nat Rev Cancer* **18,** 168–185. doi:10.1038/nrc.2017.116 (2018) (cit. on p. 6).
8. Cunniff, C., Bassetti, J. A. & Ellis, N. A. Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol Syndromol* **8,** 4–23. doi:10.1159/000452082 (2017) (cit. on p. 6).
9. Patel, D. S., Misenko, S. M., Her, J. & Bunting, S. F. BLM helicase regulates DNA repair by counteracting RAD51 loading at DNA double-strand break sites. *J Cell Biol* **216,** 3521–3534. doi:10.1083/jcb.201703144 (2017) (cit. on p. 6).
10. Gray, M. D. *et al.* The Werner syndrome protein is a DNA helicase. *Nat Genet* **17,** 100–3. doi:10.1038/ng0997-100 (1997) (cit. on p. 6).
11. Siitonen, H. A. *et al.* The mutation spectrum in RECQL4 diseases. *Eur J Hum Genet* **17,** 151–8. doi:10.1038/ejhg.2008.154 (2009) (cit. on p. 6).
12. Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75,** 1215–25. doi:10.1016/0092-8674(93)90330-s (1993) (cit. on p. 7).
13. Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75,** 1027–38. doi:10.1016/0092-8674(93)90546-3 (1993) (cit. on p. 7).
14. Bronner, C. E. *et al.* Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* **368,** 258–61. doi:10.1038/368258a0 (1994) (cit. on p. 7).
15. Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263,** 1625–9. doi:10.1126/science.8128251 (1994) (cit. on p. 7).

16. Sieber, O. M. *et al.* Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med* **348,** 791–9. doi:10.1056/NEJMoa025283 (2003) (cit. on p. 7).

17. Hirsch, S., Gieldon, L., Sutter, C., Dikow, N. & Schaaf, C. P. Germline testing for homologous recombination repair genes-opportunities and challenges. *Genes Chromosomes Cancer.* doi:10.1002/gcc.22900 (2020) (cit. on p. 7).

18. Vogt, S. *et al.* Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. *Gastroenterology* **137,** 1976-85 e1–10. doi:10.1053/j.gastro.2009.08.052 (2009) (cit. on p. 7).

19. Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep* **23,** 239–254 e6. doi:10.1016/j.celrep.2018.03.076 (2018) (cit. on pp. 7, 43, 139).

20. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3,** 246–59. doi:10.1016/j.celrep.2012.12.008 (2013) (cit. on p. 9).

21. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA-Replication as a Basis of Malignant Changes. *Cancer Research* **34,** 2311–2321. https://pubmed.ncbi.nlm.nih.gov/4136142/ (1974) (cit. on p. 9).

22. Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nat Rev Cancer* **11,** 450–7. doi:10.1038/nrc3063 (2011) (cit. on p. 9).

23. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462,** 1005–10. doi:10.1038/nature08645 (2009) (cit. on pp. 9, 30, 142).

24. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–21. doi:10.1038/nature12477 (2013) (cit. on pp. 10, 47, 70).

25. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47,** 1402–7. doi:10.1038/ng.3441 (2015) (cit. on pp. 10, 63, 123).

26. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54. doi:10.1038/nature17676 (2016) (cit. on pp. 10, 11, 57, 102).

27. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578,** 94–101. doi:10.1038/s41586-020-1943-3 (2020) (cit. on p. 10).

28. Birkbak, N. J. *et al.* Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov* **2,** 366–375. doi:10.1158/2159-8290.CD-11-0206 (2012) (cit. on p. 12).

29. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br J Cancer* **107,** 1776–82. doi:10.1038/bjc.2012.451 (2012) (cit. on p. 12).

30. Popova, T. *et al.* Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res* **72,** 5454–62. doi:10.1158/0008-5472.CAN-12-1470 (2012) (cit. on p. 13).

31. Devouassoux-Shisheboran, M. & Genestie, C. Pathobiology of ovarian carcinomas. *Chin J Cancer* **34,** 50–5. doi:10.5732/cjc.014.10273 (2015) (cit. on p. 15).

32. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68,** 394–424. doi:10.3322/caac.21492 (2018) (cit. on pp. 15, 16).

33. Torre, L. A. *et al.* Ovarian cancer statistics, 2018. *CA Cancer J Clin* **68,** 284–296. doi:10.3322/caac.21456 (2018) (cit. on p. 15).

34. Lowe, K. A. *et al.* An international assessment of ovarian cancer incidence and mortality. *Gynecol Oncol* **130,** 107–14. doi:10.1016/j.ygyno.2013.03.026 (2013) (cit. on p. 15).

35. Yang, H. P. *et al.* Ovarian cancer incidence trends in relation to changing patterns of menopausal hormone therapy use in the United States. *J Clin Oncol* **31,** 2146–51. doi:10.1200/JCO.2012.45.5758 (2013) (cit. on p. 15).

36. Zhang, Y. *et al.* Global patterns and trends in ovarian cancer incidence: age, period and birth cohort analysis. *BMC Cancer* **19,** 984. doi:10.1186/s12885-019-6139-6 (2019) (cit. on p. 15).

37. Sung, P. L. *et al.* Global distribution pattern of histological subtypes of epithelial ovarian cancer: a database analysis and systematic review. *Gynecol Oncol* **133,** 147–54. doi:10.1016/j.ygyno.2014.02.016 (2014) (cit. on p. 15).

38. Permuth-Wey, J. & Sellers, T. A. Epidemiology of ovarian cancer. *Methods Mol Biol* **472,** 413–37. doi:10.1007/978-1-60327-492-0_20 (2009) (cit. on p. 15).

39. World Health Organization, W. *Global Health Observatory data - Life expectancy* 2016. http://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text. Online; accessed April 19, 2020 (cit. on p. 16).

40. Prat, J. Ovarian carcinomas: five distinct diseases with different origins, genetic alterations, and clinicopathological features. *Virchows Arch* **460,** 237–49. doi:10.1007/s00428-012-1203-5 (2012) (cit. on pp. 16, 23, 31, 138).

41. Malpica, A. *et al.* Grading ovarian serous carcinoma using a two-tier system. *Am J Surg Pathol* **28,** 496–504. doi:10.1097/00000478-200404000-00009 (2004) (cit. on p. 16).

42. Sant, M. *et al.* EUROCARE-4. Survival of cancer patients diagnosed in 1995-1999. Results and commentary. *Eur J Cancer* **45,** 931–91. doi:10.1016/j.ejca.2008.11.018 (2009) (cit. on p. 16).

43. National Cancer Institute, N. *SEER Cancer Statistics Review, 1975-2013* 2016. https://seer.cancer.gov/archive/csr/1975_2013/results_merged/topic_survival.pdf. Online; accessed April 19, 2020 (cit. on pp. 16, 18).

44. Flaum, N., Crosbie, E. J., Edmondson, R. J., Smith, M. J. & Evans, D. G. Epithelial ovarian cancer risk: A review of the current genetic landscape. *Clin Genet* **97,** 54–63. doi:10.1111/cge.13566 (2020) (cit. on pp. 17, 18).

45. Peres, L. C. *et al.* Invasive Epithelial Ovarian Cancer Survival by Histotype and Disease Stage. *J Natl Cancer Inst* **111,** 60–68. doi:10.1093/jnci/djy071 (2019) (cit. on p. 17).

46. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975-2014, Featuring Survival. *J Natl Cancer Inst* **109.** doi:10.1093/jnci/djx030 (2017) (cit. on pp. 17, 18).

47. Verdecchia, A. *et al.* Survival trends in European cancer patients diagnosed from 1988 to 1999. *Eur J Cancer* **45,** 1042–66. doi:10.1016/j.ejca.2008.11.029 (2009) (cit. on p. 17).

48. Sant, M. *et al.* Survival of women with cancers of breast and genital organs in Europe 1999-2007: Results of the EUROCARE-5 study. *Eur J Cancer* **51,** 2191–2205. doi:10.1016/j.ejca.2015.07.022 (2015) (cit. on p. 17).

49. Bewtra, C., Watson, P., Conway, T., Read-Hippee, C. & Lynch, H. T. Hereditary ovarian cancer: a clinicopathological study. *Int J Gynecol Pathol* **11,** 180–7. doi:10.1097/00004347-199207000-00003 (1992) (cit. on p. 17).

50. Stratton, J. F., Pharoah, P., Smith, S. K., Easton, D. & Ponder, B. A. A systematic review and meta-analysis of family history and risk of ovarian cancer. *Br J Obstet Gynaecol* **105,** 493–9. doi:10.1111/j.1471-0528.1998.tb10148.x (1998) (cit. on p. 18).

51. Auranen, A. *et al.* Cancer incidence in the first-degree relatives of ovarian cancer patients. *Br J Cancer* **74,** 280–4. doi:10.1038/bjc.1996.352 (1996) (cit. on p. 18).

52. Carlson, K. J., Skates, S. J. & Singer, D. E. Screening for ovarian cancer. *Ann Intern Med* **121,** 124–32. doi:10.7326/0003-4819-121-2-199407150-00009 (1994) (cit. on p. 18).

53. Garg, K., Karnezis, A. N. & Rabban, J. T. Uncommon hereditary gynaecological tumour syndromes: pathological features in tumours that may predict risk for a germline mutation. *Pathology* **50,** 238–256. doi:10.1016/j.pathol.2017.10.009 (2018) (cit. on pp. 18, 152).

54. Lheureux, S., Gourley, C., Vergote, I. & Oza, A. M. Epithelial ovarian cancer. *The Lancet* **393,** 1240–1253. doi:10.1016/s0140-6736(18)32552-2 (2019) (cit. on p. 18).

55. Kurian, A. W., Balise, R. R., McGuire, V. & Whittemore, A. S. Histologic types of epithelial ovarian cancer: have they different risk factors? *Gynecol Oncol* **96,** 520–30. doi:10.1016/j.ygyno.2004.10.037 (2005) (cit. on p. 18).

56. Lynch, H. T., Bewtra, C. & Lynch, J. F. Familial ovarian carcinoma. Clinical nuances. *The American Journal of Medicine* **81,** 1073–1076. doi:10.1016/0002-9343(86)90411-0 (1986) (cit. on p. 18).

57. Boyd, J. Specific keynote: hereditary ovarian cancer: what we know. *Gynecol Oncol* **88,** S8–10, discussion S11–3. doi:10.1006/gyno.2002.6674 (2003) (cit. on pp. 18, 19).

58. Steichengersdorf, E. *et al.* Familial Site-Specific Ovarian-Cancer Is Linked to Brca1 on 17q12-21. *American Journal of Human Genetics* **55,** 870–875. https://www.ncbi.nlm.nih.gov/pubmed/7977348 (1994) (cit. on p. 18).

59. Liede, A. *et al.* Is hereditary site-specific ovarian cancer a distinct genetic condition? *American Journal of Medical Genetics* **75,** 55–58. doi:10.1002/(sici)1096-8628(19980106)75:1<55::aid-ajmg12>3.0.co;2-r (1998) (cit. on p. 18).

60. Greggi, S. *et al.* Analysis of 138 consecutive ovarian cancer patients: Incidence and characteristics of familial cases. *Gynecologic Oncology* **39,** 300–304. doi:10.1016/0090-8258(90)90256-k (1990) (cit. on pp. 19, 138).

61. Alsop, K. *et al.* BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. *J Clin Oncol* **30,** 2654–63. doi:10.1200/JCO.2011.39.8545 (2012) (cit. on pp. 19, 21).

62. Norquist, B. M. *et al.* Inherited Mutations in Women With Ovarian Carcinoma. *JAMA Oncol* **2,** 482–90. doi:10.1001/jamaoncol.2015.5495 (2016) (cit. on pp. 19, 20).

63. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315,** 68–76. doi:10.1001/jama.2015.17703 (2016) (cit. on pp. 19, 138).

64. Lenoir, G. Familial breast-ovarian cancer locus on chromosome 17q12-q23. *The Lancet* **338,** 82–83. doi:10.1016/0140-6736(91)90076-2 (1991) (cit. on p. 19).

65. Risch, H. A. *et al.* Population BRCA1 and BRCA2 mutation frequencies and cancer penetrances: a kin-cohort study in Ontario, Canada. *J Natl Cancer Inst* **98,** 1694–706. doi:10.1093/jnci/djj465 (2006) (cit. on p. 20).

66. Chen, S. & Parmigiani, G. Meta-analysis of BRCA1 and BRCA2 penetrance. *J Clin Oncol* **25,** 1329–33. doi:10.1200/JCO.2006.09.1066 (2007) (cit. on p. 20).

67. Kuchenbaecker, K. B. *et al.* Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317,** 2402–2416. doi:10.1001/jama.2017.7112 (2017) (cit. on p. 20).

68. Song, H. *et al.* Contribution of Germline Mutations in the RAD51B, RAD51C, and RAD51D Genes to Ovarian Cancer in the Population. *J Clin Oncol* **33,** 2901–7. doi:10.1200/JCO.2015.61.2408 (2015) (cit. on p. 20).

69. Loveday, C. *et al.* Germline RAD51C mutations confer susceptibility to ovarian cancer. *Nat Genet* **44,** 475–6, author reply 476. doi:10.1038/ng.2224 (2012) (cit. on p. 20).

70. Slavin, T. P. *et al.* The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. *NPJ Breast Cancer* **3,** 22. doi:10.1038/s41523-017-0024-8 (2017) (cit. on p. 20).

71. Loveday, C. *et al.* Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat Genet* **43,** 879–882. doi:10.1038/ng.893 (2011) (cit. on p. 20).

72. Ramus, S. J. *et al.* Germline Mutations in the BRIP1, BARD1, PALB2, and NBN Genes in Women With Ovarian Cancer. *J Natl Cancer Inst* **107.** doi:10.1093/jnci/djv214 (2015) (cit. on p. 20).

73. Yang, X. *et al.* Cancer Risks Associated With Germline PALB2 Pathogenic Variants: An International Study of 524 Families. *J Clin Oncol* **38,** 674–685. doi:10.1200/JCO.19.01907 (2020) (cit. on p. 20).

74. Pal, T. *et al.* Frequency of mutations in mismatch repair genes in a population-based study of women with ovarian cancer. *Br J Cancer* **107,** 1783–90. doi:10.1038/bjc.2012.452 (2012) (cit. on p. 20).

75. Vasen, H. F. *et al.* Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis. *Gastroenterology* **110,** 1020–7. doi:10.1053/gast.1996.v110.pm8612988 (1996) (cit. on p. 20).

76. Vasen, H. F. *et al.* MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families. *J Clin Oncol* **19,** 4074–80. doi:10.1200/JCO.2001.19.20.4074 (2001) (cit. on p. 20).

77. Bonadona, V. *et al.* Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA* **305,** 2304–10. doi:10.1001/jama.2011.743 (2011) (cit. on p. 20).

78. Lu, H. M. *et al.* Association of Breast and Ovarian Cancers With Predisposition Genes Identified by Large-Scale Sequencing. *JAMA Oncol* **5,** 51–57. doi:10.1001/jamaoncol.2018.2956 (2019) (cit. on p. 20).

79. Morgan, M., Boyd, J., Drapkin, R. & Seiden, M. V. *Cancers arising in the ovary* 1592–1613 (Elsevier, 2014) (cit. on p. 19).

80. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250,** 1684–9. doi:10.1126/science.2270482 (1990) (cit. on p. 19).

81. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265,** 2088–90. doi:10.1126/science.8091231 (1994) (cit. on p. 19).

82. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71. doi:10.1126/science.7545954 (1994) (cit. on p. 19).

83. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–92. doi:10.1038/378789a0 (1995) (cit. on p. 19).

84. Wu, J., Lu, L. Y. & Yu, X. The role of BRCA1 in DNA damage response. *Protein Cell* **1,** 117–23. doi:10.1007/s13238-010-0010-5 (2010) (cit. on p. 19).

85. Gowen, L. C., Johnson, B. L., Latour, A. M., Sulik, K. K. & Koller, B. H. Brca1 deficiency results in early embryonic lethality characterized by neuroepithelial abnormalities. *Nat Genet* **12,** 191–4. doi:10.1038/ng0296-191 (1996) (cit. on p. 19).

86. Sharan, S. K. *et al.* Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. *Nature* **386,** 804–10. doi:10.1038/386804a0 (1997) (cit. on p. 19).

87. Howlett, N. G. *et al.* Biallelic inactivation of BRCA2 in Fanconi anemia. *Science* **297,** 606–9. doi:10.1126/science.1073834 (2002) (cit. on p. 19).

88. Kobayashi, H., Ohno, S., Sasaki, Y. & Matsuura, M. Hereditary breast and ovarian cancer susceptibility genes (review). *Oncol Rep* **30,** 1019–29. doi:10.3892/or.2013.2541 (2013) (cit. on p. 19).

89. Nagy, R., Sweet, K. & Eng, C. Highly penetrant hereditary cancer syndromes. *Oncogene* **23,** 6445–70. doi:10.1038/sj.onc.1207714 (2004) (cit. on p. 19).

90. Ewald, I. P. *et al.* Genomic rearrangements in BRCA1 and BRCA2: A literature review. *Genet Mol Biol* **32,** 437–46. doi:10.1590/S1415-47572009005000049 (2009) (cit. on p. 21).

91. Boyd, J. *et al.* Clinicopathologic features of BRCA-linked and sporadic ovarian cancer. *JAMA* **283,** 2260–5. doi:10.1001/jama.283.17.2260 (2000) (cit. on pp. 21, 145).

92. Lakhani, S. R. *et al.* Pathology of ovarian cancers in BRCA1 and BRCA2 carriers. *Clin Cancer Res* **10,** 2473–81. doi:10.1158/1078-0432.ccr-1029-3 (2004) (cit. on p. 21).

93. Cass, I. *et al.* Improved survival in women with BRCA-associated ovarian carcinoma. *Cancer* **97,** 2187–95. doi:10.1002/cncr.11310 (2003) (cit. on pp. 21, 145).

94. Rubin, S. C. *et al.* Clinical and pathological features of ovarian cancer in women with germ-line mutations of BRCA1. *N Engl J Med* **335,** 1413–6. doi:10.1056/NEJM199611073351901 (1996) (cit. on pp. 21, 29).

95. Lancaster, J. M., Powell, C. B., Chen, L. M., Richardson, D. L. & Committee, S. G. O. C. P. Society of Gynecologic Oncology statement on risk assessment for inherited gynecologic cancer predispositions. *Gynecol Oncol* **136,** 3–7. doi:10.1016/j.ygyno.2014.09.009 (2015) (cit. on p. 21).

96. Walker, J. L. *et al.* Society of Gynecologic Oncology recommendations for the prevention of ovarian cancer. *Cancer* **121,** 2108–20. doi:10.1002/cncr.29321 (2015) (cit. on p. 21).

97. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61,** 759–767. doi:10.1016/0092-8674(90)90186-i (1990) (cit. on p. 21).

98. Singer, G., Kurman, R. J., Chang, H.-W., Cho, S. K. R. & Shih, I.-M. Diverse Tumorigenic Pathways in Ovarian Serous Carcinoma. *The American Journal of Pathology* **160,** 1223–1228. doi:10.1016/s0002-9440(10)62549-7 (2002) (cit. on p. 21).

99. Shih, I.-M. & Kurman, R. J. Ovarian Tumorigenesis. *The American Journal of Pathology* **164,** 1511–1518. doi:10.1016/s0002-9440(10)63708-x (2004) (cit. on pp. 21, 24).

100. Kurman, R. J. & Shih Ie, M. The Dualistic Model of Ovarian Carcinogenesis: Revisited, Revised, and Expanded. *Am J Pathol* **186,** 733–47. doi:10.1016/j.ajpath.2015.11.011 (2016) (cit. on pp. 22, 24).

101. Lheureux, S., Braunstein, M. & Oza, A. M. Epithelial ovarian cancer: Evolution of management in the era of precision medicine. *CA Cancer J Clin* **69,** 280–304. doi:10.3322/caac.21559 (2019) (cit. on p. 23).

102. Ng, C. K. *et al.* The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J Pathol* **226,** 703–12. doi:10.1002/path.3980 (2012) (cit. on pp. 23, 30, 31, 142).

103. Vaughan, S. *et al.* Rethinking ovarian cancer: recommendations for improving outcomes. *Nat Rev Cancer* **11,** 719–25. doi:10.1038/nrc3144 (2011) (cit. on pp. 23, 31, 35).

104. Gelmon, K. A. *et al.* Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre, open-label, non-randomised study. *The Lancet Oncology* **12,** 852–861. doi:10.1016/s1470-2045(11)70214-5 (2011) (cit. on pp. 23, 30).

105. Mirza, M. R. *et al.* Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. *N Engl J Med* **375,** 2154–2164. doi:10.1056/NEJMoa1611310 (2016) (cit. on pp. 23, 30, 145).

106. Swisher, E. M. *et al.* Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part 1): an international, multicentre, open-label, phase 2 trial. *The Lancet Oncology* **18,** 75–87. doi:10.1016/s1470-2045(16)30559-9 (2017) (cit. on pp. 23, 30, 144).

107. Fathalla, M. F. Incessant Ovulation—a Factor in Ovarian Neoplasia ? *The Lancet* **298,** 163. doi:10.1016/s0140-6736(71)92335-x (1971) (cit. on p. 23).

108. RJ, K., ML, C., CS, H. & RH, Y. in *WHO Classification of Tumours, 4th Edition* (WHO Press, 2014) (cit. on p. 24).

109. Piek, J. M. *et al.* Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer. *J Pathol* **195,** 451–6. doi:10.1002/path.1000 (2001) (cit. on p. 24).

110. Paley, P. J. *et al.* Occult cancer of the fallopian tube in BRCA-1 germline mutation carriers at prophylactic oophorectomy: a case for recommending hysterectomy at surgical prophylaxis. *Gynecol Oncol* **80,** 176–80. doi:10.1006/gyno.2000.6071 (2001) (cit. on p. 24).

111. Leeper, K. *et al.* Pathologic findings in prophylactic oophorectomy specimens in high-risk women. *Gynecol Oncol* **87,** 52–6. doi:10.1006/gyno.2002.6779 (2002) (cit. on p. 24).

112. Kauff, N. D. *et al.* Risk-reducing salpingo-oophorectomy in women with a BRCA1 or BRCA2 mutation. *N Engl J Med* **346,** 1609–15. doi:10.1056/NEJMoa020119 (2002) (cit. on p. 24).

113. Rebbeck, T. R. *et al.* Prophylactic oophorectomy in carriers of BRCA1 or BRCA2 mutations. *N Engl J Med* **346,** 1616–22. doi:10.1056/NEJMoa012158 (2002) (cit. on p. 24).

114. Piek, J. M. J. *et al.* BRCA1/2-related ovarian cancers are of tubal origin: a hypothesis. *Gynecologic Oncology* **90,** 491. doi:10.1016/s0090-8258(03)00365-2 (2003) (cit. on p. 24).

115. Finch, A. *et al.* Clinical and pathologic findings of prophylactic salpingo-oophorectomies in 159 BRCA1 and BRCA2 carriers. *Gynecol Oncol* **100,** 58–64. doi:10.1016/j.ygyno.2005.06.065 (2006) (cit. on p. 24).

116. Medeiros, F. *et al.* The tubal fimbria is a preferred site for early adenocarcinoma in women with familial ovarian cancer syndrome. *Am J Surg Pathol* **30,** 230–6. doi:10.1097/01.pas.0000180854.28831.77 (2006) (cit. on p. 24).

117. Kindelberger, D. W. *et al.* Intraepithelial carcinoma of the fimbria and pelvic serous carcinoma: Evidence for a causal relationship. *Am J Surg Pathol* **31,** 161–9. doi:10.1097/01.pas.0000213335.40358.47 (2007) (cit. on p. 24).

118. Lee, Y. *et al.* A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *J Pathol* **211,** 26–35. doi:10.1002/path.2091 (2007) (cit. on p. 24).

119. Kuhn, E. *et al.* TP53 mutations in serous tubal intraepithelial carcinoma and concurrent pelvic high-grade serous carcinoma–evidence supporting the clonal relationship of the two lesions. *J Pathol* **226,** 421–6. doi:10.1002/path.3023 (2012) (cit. on p. 24).

120. Folkins, A. K. *et al.* A candidate precursor to pelvic serous cancer (p53 signature) and its prevalence in ovaries and fallopian tubes from women with BRCA mutations. *Gynecol Oncol* **109,** 168–73. doi:10.1016/j.ygyno.2008.01.012 (2008) (cit. on p. 24).

121. Mehra, K. K. *et al.* The impact of tissue block sampling on the detection of p53 signatures in fallopian tubes from women with BRCA 1 or 2 mutations (BRCA+) and controls. *Mod Pathol* **24,** 152–6. doi:10.1038/modpathol.2010.171 (2011) (cit. on p. 24).

122. Soong, T. R. *et al.* Evidence for lineage continuity between early serous proliferations (ESPs) in the Fallopian tube and disseminated high-grade serous carcinomas. *J Pathol* **246,** 344–351. doi:10.1002/path.5145 (2018) (cit. on pp. 24, 25).

123. Soong, T. R., Kolin, D. L., Teschan, N. J. & Crum, C. P. Back to the Future? The Fallopian Tube, Precursor Escape and a Dualistic Model of High-Grade Serous Carcinogenesis. *Cancers (Basel)* **10.** doi:10.3390/cancers10120468 (2018) (cit. on p. 24).

124. McCluggage, W. G., Hirschowitz, L., Gilks, C. B., Wilkinson, N. & Singh, N. The Fallopian Tube Origin and Primary Site Assignment in Extrauterine High-grade Serous Carcinoma: Findings of a Survey of Pathologists and Clinicians. *Int J Gynecol Pathol* **36,** 230–239. doi:10.1097/PGP.0000000000000336 (2017) (cit. on p. 24).

125. Kaldawy, A. *et al.* Low-grade serous ovarian cancer: A review. *Gynecol Oncol* **143,** 433–438. doi:10.1016/j.ygyno.2016.08.320 (2016) (cit. on p. 25).

126. Perren, T. J. Mucinous epithelial ovarian carcinoma. *Ann Oncol* **27 Suppl 1,** i53–i57. doi:10.1093/annonc/mdw087 (2016) (cit. on p. 25).

127. McConechy, M. K. *et al.* Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles. *Mod Pathol* **27,** 128–34. doi:10.1038/modpathol.2013.107 (2014) (cit. on p. 25).

128. Kuroda, T. & Kohno, T. Precision medicine for ovarian clear cell carcinoma based on gene alterations. *Int J Clin Oncol* **25,** 419–424. doi:10.1007/s10147-020-01622-z (2020) (cit. on p. 25).

129. Gorringe, K. L. *et al.* High-resolution single nucleotide polymorphism array analysis of epithelial ovarian cancer reveals numerous microdeletions and amplifications. *Clin Cancer Res* **13,** 4731–9. doi:10.1158/1078-0432.CCR-07-0502 (2007) (cit. on p. 25).

130. Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–15. doi:10.1038/nature10166 (2011) (cit. on pp. 25, 27–29, 31, 32, 40, 48, 55, 73, 74, 93, 138, 146, 150).

131. Patch, A. M. *et al.* Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521,** 489–94. doi:10.1038/nature14410 (2015) (cit. on pp. 26, 27, 29, 32, 40, 48, 74, 76, 137, 138, 146).

132. Wang, Y. K. *et al.* Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nat Genet* **49,** 856–865. doi:10.1038/ng.3849 (2017) (cit. on pp. 26, 27, 30, 32–34, 36, 41, 48, 89, 137, 140, 141, 143, 146).

133. Esteller, M. *et al.* Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst* **92,** 564–9. doi:10.1093/jnci/92.7.564 (2000) (cit. on p. 27).

134. Marks, J. R. *et al.* Overexpression and mutation of p53 in epithelial ovarian cancer. *Cancer Res* **51,** 2979–84. http://www.ncbi.nlm.nih.gov/pubmed/2032235 (1991) (cit. on p. 27).

135. Skilling, J. S., Sood, A., Niemann, T., Lager, D. J. & Buller, R. E. An abundance of p53 null mutations in ovarian carcinoma. *Oncogene* **13,** 117–23. http://www.ncbi.nlm.nih.gov/pubmed/8700537 (1996) (cit. on p. 27).

136. Yemelyanova, A. *et al.* Immunohistochemical staining patterns of p53 can serve as a surrogate marker for TP53 mutations in ovarian carcinoma: an immunohistochemical and nucleotide sequencing analysis. *Mod Pathol* **24,** 1248–53. doi:10.1038/modpathol.2011.85 (2011) (cit. on p. 27).

137. Ahmed, A. A. *et al.* Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *J Pathol* **221,** 49–56. doi:10.1002/path.2696 (2010) (cit. on p. 28).

138. Vang, R. *et al.* Molecular Alterations of TP53 are a Defining Feature of Ovarian High-Grade Serous Carcinoma: A Rereview of Cases Lacking TP53 Mutations in The Cancer Genome Atlas Ovarian Study. *Int J Gynecol Pathol* **35,** 48–55. doi:10.1097/PGP.0000000000000207 (2016) (cit. on p. 28).

139. Li, S. B., Schwartz, P. E., Lee, W. H. & Yang-Feng, T. L. Allele loss at the retinoblastoma locus in human ovarian cancer. *J Natl Cancer Inst* **83,** 637–40. doi:10.1093/jnci/83.9.637 (1991) (cit. on p. 28).

140. Hashiguchi, Y. *et al.* Combined analysis of p53 and RB pathways in epithelial ovarian cancer. *Hum Pathol* **32,** 988–96. doi:10.1053/hupa.2001.27115 (2001) (cit. on p. 28).

141. Cheaib, B., Auguste, A. & Leary, A. The PI3K/Akt/mTOR pathway in ovarian cancer: therapeutic opportunities and challenges. *Chin J Cancer* **34,** 4–16. doi:10.5732/cjc.014.10289 (2015) (cit. on p. 28).

142. Phillips-Chavez, C., Watson, M., Coward, J. & Schloss, J. A systematic literature review assessing if genetic biomarkers are predictors for platinum-based chemotherapy response in ovarian cancer patients. *Eur J Clin Pharmacol* **76,** 1059–1074. doi:10.1007/s00228-020-02874-4 (2020) (cit. on p. 28).

143. Tothill, R. W. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* **14,** 5198–208. doi:10.1158/1078-0432.CCR-08-0196 (2008) (cit. on pp. 28, 31).

144. Vencken, P. *et al.* Chemosensitivity and outcome of BRCA1- and BRCA2-associated ovarian cancer patients after first-line chemotherapy compared with sporadic ovarian cancer patients. *Ann Oncol* **22,** 1346–1352. doi:10.1093/annonc/mdq628 (2011) (cit. on p. 29).

145. Bolton, K. L. *et al.* Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. *JAMA* **307,** 382–90. doi:10.1001/jama.2012.20 (2012) (cit. on p. 29).

146. Zhong, Q., Peng, H. L., Zhao, X., Zhang, L. & Hwang, W. T. Effects of BRCA1- and BRCA2-related mutations on ovarian and breast cancer survival: a meta-analysis. *Clin Cancer Res* **21,** 211–20. doi:10.1158/1078-0432.CCR-14-1816 (2015) (cit. on p. 29).

147. Neff, R. T., Senter, L. & Salani, R. BRCA mutation in ovarian cancer: testing, implications and treatment considerations. *Ther Adv Med Oncol* **9,** 519–531. doi:10.1177/1758834017714993 (2017) (cit. on p. 29).

148. Farley, J. *et al.* Cyclin E expression is a significant predictor of survival in advanced, suboptimally debulked ovarian epithelial cancers: A Gynecologic Oncology Group study. *Cancer Research* **63,** 1235–1241. https://pubmed.ncbi.nlm.nih.gov/12649182/ (2003) (cit. on p. 29).

149. Etemadmoghadam, D. *et al.* Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. *Clin Cancer Res* **15,** 1417–27. doi:10.1158/1078-0432.CCR-08-1564 (2009) (cit. on pp. 29, 146).

150. Nakayama, N. *et al.* Gene amplification CCNE1 is related to poor survival and potential therapeutic target in ovarian cancer. *Cancer* **116,** 2621–34. doi:10.1002/cncr.24987 (2010) (cit. on pp. 29, 146).

151. Mayr, D. *et al.* Analysis of gene amplification and prognostic markers in ovarian cancer using comparative genomic hybridization for microarrays and immunohistochemical analysis for tissue microarrays. *Am J Clin Pathol* **126,** 101–9. doi:10.1309/n6x5mb24bp42kp20 (2006) (cit. on p. 29).

152. Sapoznik, S., Aviel-Ronen, S., Bahar-Shany, K., Zadok, O. & Levanon, K. CCNE1 expression in high grade serous carcinoma does not correlate with chemoresistance. *Oncotarget* **8,** 62240–62247. doi:10.18632/oncotarget.19272 (2017) (cit. on p. 29).

153. Aziz, D. *et al.* 19q12 amplified and non-amplified subsets of high grade serous ovarian cancer with overexpression of cyclin E1 differ in their molecular drivers and clinical outcomes. *Gynecol Oncol* **151,** 327–336. doi:10.1016/j.ygyno.2018.08.039 (2018) (cit. on p. 29).

154. Hoppe, M. M., Sundar, R., Tan, D. S. P. & Jeyasekharan, A. D. Biomarkers for Homologous Recombination Deficiency in Cancer. *J Natl Cancer Inst* **110,** 704–713. doi:10.1093/jnci/djy085 (2018) (cit. on p. 30).

155. Wang, Z. C. *et al.* Profiles of genomic instability in high-grade serous ovarian cancer predict treatment outcome. *Clin Cancer Res* **18,** 5806–15. doi:10.1158/1078-0432.CCR-12-0857 (2012) (cit. on p. 30).

156. Lin, K. *et al.* 2701 Quantification of genomic loss of heterozygosity enables prospective selection of ovarian cancer patients who may derive benefit from the PARP inhibitor rucaparib. *European Journal of Cancer* **51,** S531–S532. doi:10.1016/s0959-8049(16)31469-1 (2015) (cit. on p. 30).

157. McBride, D. J. *et al.* Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. *J Pathol* **227,** 446–55. doi:10.1002/path.4042 (2012) (cit. on pp. 30, 142).

158. Popova, T. *et al.* Ovarian Cancers Harboring Inactivating Mutations in CDK12 Display a Distinct Genomic Instability Pattern Characterized by Large Tandem Duplications. *Cancer Res* **76,** 1882–91. doi:10.1158/0008-5472.CAN-15-2128 (2016) (cit. on pp. 31, 142).

159. Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev* **25,** 2158–72. doi:10.1101/gad.16962311 (2011) (cit. on p. 31).

160. Bajrami, I. *et al.* Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res* **74,** 287–97. doi:10.1158/0008-5472.CAN-13-2541 (2014) (cit. on p. 31).

161. Waldron, L., Riester, M. & Birrer, M. Molecular subtypes of high-grade serous ovarian cancer: the holy grail? *J Natl Cancer Inst* **106.** doi:10.1093/jnci/dju297 (2014) (cit. on p. 31).

162. Leong, H. S. *et al.* Efficient molecular subtype classification of high-grade serous ovarian cancer. *J Pathol* **236,** 272–7. doi:10.1002/path.4536 (2015) (cit. on pp. 31, 140).

163. Talhouk, A. *et al.* Development and Validation of the Gene Expression Predictor of High-grade Serous Ovarian Carcinoma Molecular SubTYPE (PrOTYPE). *Clin Cancer Res* **26,** 5411–5423. doi:10.1158/1078-0432.CCR-20-0103 (2020) (cit. on pp. 31, 140).

164. Schwede, M. *et al.* The Impact of Stroma Admixture on Molecular Subtypes and Prognostic Gene Signatures in Serous Ovarian Cancer. *Cancer Epidemiol Biomarkers Prev* **29,** 509–519. doi:10.1158/1055-9965.EPI-18-1359 (2020) (cit. on pp. 31, 140).

165. Zhang, A. W. *et al.* Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. *Cell* **173,** 1755–1769 e22. doi:10.1016/j.cell.2018.03.073 (2018) (cit. on pp. 31, 140, 142).

166. Verhaak, R. G. *et al.* Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J Clin Invest* **123,** 517–25. doi:10.1172/JCI65833 (2013) (cit. on p. 31).

167. Tan, T. Z. *et al.* Decoding transcriptomic intra-tumour heterogeneity to guide personalised medicine in ovarian cancer. *J Pathol* **247,** 305–319. doi:10.1002/path.5191 (2019) (cit. on p. 31).

168. Braicu, E. I. *et al.* Dynamic of molecular subtypes of high-grade serous ovarian cancer in paired primary and relapsed biopsies. *Journal of Clinical Oncology* **37,** e17091–e17091. doi:10.1200/JCO.2019.37.15_suppl.e17091 (2019) (cit. on pp. 32, 140).

169. Menghi, F. *et al.* The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* **34,** 197–210 e5. doi:10.1016/j.ccell.2018.06.008 (2018) (cit. on pp. 32, 36, 50, 85, 140).

170. Menghi, F. *et al.* The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci U S A* **113,** E2373–82. doi:10.1073/pnas.1520010113 (2016) (cit. on pp. 32, 51, 85).

171. Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45,** 1113–20. doi:10.1038/ng.2764 (2013) (cit. on p. 39).

172. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12,** R41. doi:10.1186/gb-2011-12-4-r41 (2011) (cit. on pp. 39, 49, 72).

173. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218. doi:10.1038/nature12213 (2013) (cit. on pp. 40, 47, 70, 74).

174. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173,** 371–385 e18. doi:10.1016/j.cell.2018.02.060 (2018) (cit. on p. 40).

175. The International Cancer Genome Consortium, I. *ICGC Data Portal* 2019. https://dcc.icgc.org/. Online; accessed Nov 26, 2019 (cit. on pp. 40, 85).

176. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **10,** 1081–2. doi:10.1038/nmeth.2642 (2013) (cit. on p. 40).

177. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92. doi:10.4161/fly.19695 (2012) (cit. on p. 41).

178. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43,** D805–11. doi:10.1093/nar/gku1075 (2015) (cit. on pp. 43, 50, 81, 120, 126).

179. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.* https://arxiv.org/abs/1303.3997 (2013) (cit. on p. 45).

180. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9. doi:10.1093/bioinformatics/btp352 (2009) (cit. on p. 45).

181. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–93. doi:10.1093/bioinformatics/btr509 (2011) (cit. on p. 45).

182. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46,** 912–918. doi:10.1038/ng.3036 (2014) (cit. on p. 45).

183. Yung, C. K. *et al.* Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *bioRxiv.* doi:10.1101/161638 (2017) (cit. on p. 45).

184. Toprak, U. *Integrative Analysis of Omics Datasets* Thesis (2019). doi:10.11588/heidok.00027429 (cit. on pp. 45, 52).

185. Kleinheinz, K. *et al.* ACEseq – allele specific copy number estimation from whole genome sequencing. *bioRxiv.* doi:10.1101/210807 (2017) (cit. on p. 45).

186. Endesfelder, D. *et al.* Chromosomal instability selects gene copy-number variants encoding core regulators of proliferation in ER+ breast cancer. *Cancer Res* **74,** 4853–4863. doi:10.1158/0008-5472.CAN-13-2664 (2014) (cit. on p. 47).

187. Institute, B. *Cancer Genome Analysis (CGA) - MutSig* 2013. https://software.broadinstitute.org/cancer/cga/mutsig. Online; accessed Nov 5, 2019 (cit. on p. 48).

188. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45,** 1134–40. doi:10.1038/ng.2760 (2013) (cit. on p. 49).

189. Daniel Huebschmann Zuguang Gu, M. S. *YAPSA: Yet Another Package for Signature Analysis* 2015. https://bioconductor.org/packages/YAPSA/. R package version 1.0.0 (cit. on pp. 52, 57).

190. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164. doi:10.1093/nar/gkq603 (2010) (cit. on p. 52).

191. Kolde, R. *pheatmap: Pretty heatmaps* 2015. https://CRAN.R-project.org/package=pheatmap. R package version 1.0.12 (cit. on p. 52).

192. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17,** 405–24. doi:10.1038/gim.2015.30 (2015) (cit. on pp. 54, 94).

193. Hubschmann, D. *et al.* Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosomes Cancer.* doi:10.1002/gcc.22918 (2020) (cit. on p. 57).

194. Zhao, E. Y. *et al.* Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. *Clin Cancer Res* **23,** 7521–7530. doi:10.1158/1078-0432.CCR-17-1941 (2017) (cit. on p. 57).

195. Catalogue Of Somatic Mutations In Cancer, C. *Mutational Signatures (v2 - March 2015) - Single Base Substitution (SBS) Signatures* 2015. https://cancer.sanger.ac.uk/cosmic/signatures_v2.tt. Online; accessed Nov 26, 2019 (cit. on pp. 57, 101).

196. Catalogue Of Somatic Mutations In Cancer, C. *Mutational Signatures (v3.1 - June 2020) - Small Insertion and Deletion (ID) Signatures* 2020. https://cancer.sanger.ac.uk/cosmic/signatures/ID/index.tt. Online; accessed Nov 26, 2019 (cit. on pp. 57, 102).

197. Schlesner, D. H., Jopp-Saile, L., Andresen, C., Gu, Z. & Matthias. *YAPSA: Yet Another Package for Signature Analysis* 2020. https://bioconductor.org/packages/YAPSA/. Online; accessed R package version 1.14.0 (cit. on p. 57).

198. Felsenstein, J. *PHYLIP (Phylogeny Inference Package)* 2005. https://evolution.genetics.washington.edu/phylip.html. Online; accessed version 3.6 (cit. on p. 60).

199. Gerstung, M., Santiago Gonzalez, o. b. o. t. P.-1. E. & Group, H. W. *Supplementary code: The evolutionary history of 2,658 cancers* 2020. https://gerstung-lab.github.io/PCAWG-11/. Online; accessed Feb 05, 2020 (cit. on pp. 60, 120).

200. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578,** 122–128. doi:10.1038/s41586-019-1907-7 (2020) (cit. on pp. 60, 62–64, 149).

201. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–5. doi:10.1038/ng.2653 (2013) (cit. on p. 74).

202. Perez-Fidalgo, J. A. *et al.* Aurora kinases in ovarian cancer. *ESMO Open* **5.** doi:10.1136/esmoopen-2020-000718 (2020) (cit. on p. 82).

203. Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10,** e1004494. doi:10.1371/journal.pgen.1004494 (2014) (cit. on p. 95).

204. Thompson, E. R. *et al.* Reevaluation of the BRCA2 truncating allele c.9976A > T (p.Lys3326Ter) in a familial breast cancer context. *Sci Rep* **5,** 14800. doi:10.1038/srep14800 (2015) (cit. on p. 102).

205. Kanchi, K. L. *et al.* Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* **5,** 3156. doi:10.1038/ncomms4156 (2014) (cit. on pp. 138, 139, 150, 152).

206. Levy-Lahad, E. *et al.* Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. *Am J Hum Genet* **60,** 1059–67. http://www.ncbi.nlm.nih.gov/pubmed/9150153 (1997) (cit. on p. 139).

207. Lynch, H. T. *et al.* Hereditary ovarian carcinoma: heterogeneity, molecular genetics, pathology, and management. *Mol Oncol* **3,** 97–137. doi:10.1016/j.molonc.2009.02.004 (2009) (cit. on p. 139).

208. Reid, B. M., Permuth, J. B. & Sellers, T. A. Epidemiology of ovarian cancer: a review. *Cancer Biol Med* **14,** 9–32. doi:10.20892/j.issn.2095-3941.2016.0084 (2017) (cit. on p. 140).

209. Karst, A. M. *et al.* Cyclin E1 deregulation occurs early in secretory cell transformation to promote formation of fallopian tube-derived high-grade serous ovarian cancers. *Cancer Res* **74,** 1141–52. doi:10.1158/0008-5472.CAN-13-2247 (2014) (cit. on pp. 145, 146).

210. Etemadmoghadam, D. *et al.* Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc Natl Acad Sci U S A* **110,** 19489–94. doi:10.1073/pnas.1314302110 (2013) (cit. on p. 146).

211. George, J. *et al.* Nonequivalent gene expression and copy number alterations in high-grade serous ovarian cancers with BRCA1 and BRCA2 mutations. *Clin Cancer Res* **19,** 3474–84. doi:10.1158/1078-0432.CCR-13-0066 (2013) (cit. on p. 146).

212. Etemadmoghadam, D. *et al.* Resistance to CDK2 inhibitors is associated with selection of polyploid cells in CCNE1-amplified ovarian cancer. *Clin Cancer Res* **19,** 5960–71. doi:10.1158/1078-0432.CCR-13-1337 (2013) (cit. on p. 146).

213. Schwarz, R. F. *et al.* Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med* **12,** e1001789. doi:10.1371/journal.pmed.1001789 (2015) (cit. on p. 147).

214. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol* **231,** 21–34. doi:10.1002/path.4230 (2013) (cit. on p. 148).

215. Labidi-Galy, S. I. *et al.* High grade serous ovarian carcinomas originate in the fallopian tube. *Nat Commun* **8,** 1093. doi:10.1038/s41467-017-00962-1 (2017) (cit. on p. 149).

216. Esteller, M. *et al.* Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst* **92,** 564–9. doi:10.1093/jnci/92.7.564 (2000) (cit. on p. 150).

217. Lim, S. L. *et al.* Promoter hypermethylation of FANCF and outcome in advanced ovarian cancer. *Br J Cancer* **98,** 1452–6. doi:10.1038/sj.bjc.6604325 (2008) (cit. on p. 150).

218. Masutani, C. *et al.* The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase eta. *Nature* **399,** 700–4. doi:10.1038/21447 (1999) (cit. on p. 152).

219. Nicolaides, N. C. *et al.* Mutations of two PMS homologues in hereditary non-polyposis colon cancer. *Nature* **371,** 75–80. doi:10.1038/371075a0 (1994) (cit. on p. 152).

220. Bowtell, D. D. The genesis and evolution of high-grade serous ovarian cancer. *Nat Rev Cancer* **10,** 803–8. doi:10.1038/nrc2946 (2010) (cit. on p. 153).

ACRONYMS

ASR Age-standardised rate

ASCN allele-specific copy number

AI allelic imbalance

AUC area under the ROC curve

BLC basal-like breast carcinoma

BER base-excision repair

CGC Cancer Gene Census

CF Carrier frequency

CIN chromosomal instability

CCC clear-cell carcinoma

CE index Clonal Expansion index

CN copy number

CNA copy number alteration

DEL deletion

DR Direct repair

DDR DNA damage response

DSB double-strand break

ESP early serous proliferation

EC endometrioid carcinoma

EOC epithelial ovarian cancer

EMT epithelial-to-mesenchymal transition

FTE fallopian tube epithelium

FA *Fanconi anaemia*

FBI foldback inversion

FLOH fraction of genome with LOH

GUS gene of unknown significance

GDAC Genome Data Analysis Center

HBOC syndrome *Hereditary Breast and Ovarian Cancer syndrome*

`HET_CNA` heterogeneity index based on copy number profiles

`HET_MutIndel` heterogeneity index based on mutation and indels

HET_SV heterogeneity index based on structural variations

HOC Hereditary ovarian cancer

HGSC high-grade serous carcinoma

HRD homologous recombination deficiency

HRR homologous recombination repair

HDI Human Development Index

ID Signature Indel Signature

TRX translocation

ICGC International Cancer Genome Consortium

ICL inter-strand cross-link

indel insertion/deletions

IPH intra-patient heterogeneity

INV inversion

K-M Kaplan-Meier

LST large-scale state transition

LOH loss of heterozygosity

LGSC low-grade serous carcinoma

MMMT malignant mixed mesodermal tumor

MPSC micropapillary serous carcinoma

MSI microsatellite instability

MMR mismatch repair

MRCA most recent common ancestor

MC mucinous carcinoma

MAP *MUTYH-associated polyposis*

ND near-diploid

NER nucleotide excision repair

OR Odds ratio

OG oncogene

OC ovarian cancer

OSE ovarian surface epithelium

OS overall survival

MRCA-PID most recent common ancestor between samples

MRCA-SAMPLE most recent common ancestor in the sample

PARP Poly(ADP-Ribose) Polymerase

PC principal component
PCA principal component analysis
PFS progression-free survival
ROC receiver operating characteristic
SMG significantly mutated gene
SBS Signature Single Base Substitution Signature
SCNA somatic copy number alteration
SV structural variation
TD tandem duplication
TDP **Tandem Duplicator Phenotype**
TAI telomeric allelic imbalance
TCGA The Cancer Genome Atlas
TIC tubal intraepithelial carcinoma
TSG tumor suppressor gene
UV ultraviolet
VAF variant allele frequency
VUS variant of unknown significance
wGII weighted genome integrity index
WGD whole genome duplication
WGS whole genome sequencing
XP *Xeroderma pigmentosum*
XPV *Xeroderma pigmentosum variant*

D

PUBLICATION AND POSTER PRESENTATION

PEER REVIEWED JOURNAL

Priya Chudasama, Sadaf S. Mughal, Mathijs A. Sanders, Daniel Hübschmann, Inn Chung, Katharina I. Deeg, Siao-Han Wong, Sophie Rabe, Mario Hlevnjak, Marc Zapatka, Aurélie Ernst, Kortine Kleinheinz, Matthias Schlesner, Lina Sieverling, Barbara Klink, Evelin Schröck, Remco M. Hoogenboezem, Bernd Kasper, Christoph E. Heilig, Gerlinde Egerer, Stephan Wolf, Christof von Kalle, Roland Eils, Albrecht Stenzinger, Wilko Weichert, Hanno Glimm, Stefan Gröschel, Hans-Georg Kopp, Georg Omlor, Burkhard Lehner, Sebastian Bauer, Simon Schimmack, Alexis Ulrich, Gunhild Mechtersheimer, Karsten Rippe, Benedikt Brors, Barbara Hutter, Marcus Renner, Peter Hohenberger, Claudia Scholl & Stefan Fröhling, **Integrative genomic and transcriptomic analysis of leiomyosarcoma.** Nat Commun 9, 144 (2018)

MANUSCRIPT

Induction of autoreactive Tregs through promiscuous gene expression by bone marrow-resident APCs

Chih-Yeh Chen, Felix Klug, Siao-Han Wong, Franziska Durst, Sheena Pinto, Tomoyoshi Yamano, Dania Riegel, Michael Delacher, Maria Dinkelacker, Charles Imbusch, Abdelrahman Mahmoud, Roman Kurilov, Miodrag Gužvić, Claudia Gebhard, Guido Wabnitz, Valentina Volpin, Ayse Nur Menevse, Yvonne Samstag, Pärt Peterson, Michael Rehli, Slava Stamova, Maria Xydia, Christoph Klein, Mark Anderson, Christian Schmidl, Markus Feuerer, Benedikt Brors, Ludger Klein, Bruno Kyewski, and Philipp Beckhove

POSTER PRESENTATION

Structure-Informed Variant Prioritization in Personalized Oncology

Siao-Han Wong, Benedikt Brors
(Sep. 2019, German Conference on Bioinformatics 2019 Heidelberg)
(Jul. 2019, ISMB/ECCB 2019, track: 3DSIG COSI)

A Structural Approach for Prioritizing Variants in Personalized Oncology

Siao-Han Wong, Benedikt Brors
(Mar. 2019, AACR Annual Meeting 2019)
(Jun. 2018, International PhD Student Cancer Conference, with travel grant)

Integration of omics data in personalized oncology scenario

Siao-Han Wong, Benedikt Brors
(Dec. 2016, 2016 DKFZ PhD Poster Presentation)
(Nov. 2016, EMBO: From Functional Genomics to Systems Biology)

An Integrative Network Approach for Drug Target Prioritization in Personalized Oncology

Siao-Han Wong, Benedikt Brors
(May 2016, DKFZ Conference: Cancer Systems Genetics)

Network Analysis of Genetic Aberrations in Personalized Oncology

Siao-Han Wong, Benedikt Brors
(Nov. 2015, EMBL Stanford Conference: Personalised Health)

Annotation of Clinical Impact for Somatic Mutations in Cancer

Siao-Han Wong, Benedikt Brors
(Nov. 2015, EMBL Conference: Cancer Genomics)
(Jul. 2015, DKFZ PhD Retreat 2015)

E

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. Benedikt Brors for giving me this opportunity to pursue my scientific interests in his group.

Thank you for providing this friendly environment and strong infrastructure for doing research, ensuring that we have all the resources needed, and supporting us with full freedom to explore all directions in science.

Also to my Thesis Advisory Committee (TAC) members, Prof. Dr. Stefan Fröhling and Dr. Wolfgang Huber, thank you for your feedback, suggestions and help in the past years.

I appreciate the support from the DKFZ Graduate Office, DKFZ Human Resources Office and Uni-HD Faculty of Biosciences, who provide ample training opportunities, help me fulfill all the requirements for working here and for my graduation. You've made it much less complicated for an international student studying here.

I thank the HIPO59 working group, whom I enjoy collaborate with in this project; and thank all the enrolled patients for their contribution to cancer research.

And I remember the times with all the previous and current colleagues in ABI, ODCF, BODA, CRG and (previous) CO groups, It has been a great pleasure working closely together, and I thank all of you for providing kind assistance throughout these years. I also enjoyed all the activities we've done together, be it cooking, baking, hiking, canoeing, various celebrations and festivals..., oh, and doing science of course :)

To Barbara Hutter: your magic box is the secret to making all fancy PhD hats ;) and I am happy to have those months we walk to the tumor board together, also these visits are meaningful to me as it is the reason I am here.

To my two forever office mates Lina Sieverling and Charles Imbusch: thank you for everything. You are my best support in the group and in my life, and I am happy that we share so many good memories together :)

I would also like to thank Roma Kurilov, Pitithat Puranachot, Lars Feuerbach, Lina Sieverling, Naveed Ishaque, Qi Wang, whom (we together) organized many nice events for all group members.

There are many of you participating in my daily life and I will not continue the enumeration here, since I just found the list grow too long and it might be as difficult as it is to bring my thesis to an end. I thank all who I have not named here for the care, love and accompany I got form you during these years, you all made it more and more difficult for me to say goodbye <3

During my life in Heidelberg, I have encountered interesting people and many amazing friends. Although many of you have already moved to other cities around the world, I feel lucky that we have met here and grateful for the memorable moments we create together.

To Chih-Yeh, Calvin, Li-Ling, Ling-Shih, Ya-Yun, Wan-Ching: I thank our secret cluster for comprehensive support since the first day, and for those good old times we (complain) and explore the culture and Europe together.

To my dearest flat mates Bouchra and Justyna: thank you for being my second family here. Your support and care throughout the ups and downs in our PhD lives have been meaningful to me. I enjoy the moments at the kitchen, the corridor talks and all the little things we do together.

For same reason I have to stop myself from putting more names. To the many of you whom I met since the start of our studies, and to all the friends I met outside DKFZ: thank you all for your friendship and I enjoyed the time we spent together. What I learned from you influenced my perspective on life and toward things, and transformed me in many ways. All of you made these years a fruitful, unique and memorable experience in my life.

I also thank people who have been providing me with remote support from Taiwan.

To my grandparents, parents and siblings: you prepared me as who I am, and gave me full support to explore the world and fulfill my curiosity. Thank you for backing me up whenever and wherever, and I really wish I were by your side at all the moments I have missed. Love you all <3

To all my friends: how amazing it is that you came into my life and stayed ever since. Although I may not manage to see all of you in person every year, our reunions have been the highlight of the year for me and I am grateful that we are always there for each other :)

Special thanks goes to people who have inspired me throughout this journey in science.

My experience in Dr. Konan Peck's group in Academia Sinica sparked my interest in cancer research. Your insights and passion for research have been passed down to every one of us. Thank you and miss you.

I thank Tzu-Hung for your orientation in doing research and I learned from you many real sides of having it as a career.

To Emily: thank you for everything. Especially what I learned from you developed my insights and interest in answering clinical questions.

I thank my current and previous supervisors and all scientists I have encountered, who have inspired me a lot with our discussions, their advices, kind sharing, talks and their work. The good qualities I see in everyone of you positively influenced me.

I am definitely lucky to have met so many kind people in my life, thank all of you who provide(d) me love, care, support, inspiration and hope throughout these years. All of you have made me who I am today.

Special thanks to those who saw my tears and weakness, thank you for being with me when I fell in all miserable ways.

Finally, I thank myself for giving up and not giving up, and kept holding on. This experience helped me know myself a bit better, and I am looking forward to the journey ahead :)

This document was typeset using `classicthesis` style developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. It is available for LaTeX and LyX at

https://bitbucket.org/amiede/classicthesis/